STATISTICAL INFERENCE AND THE SUM OF SQUARES METHOD

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Samuel Hopkins

August 2018

STATISTICAL INFERENCE AND THE SUM OF SQUARES METHOD

Samuel Hopkins, Ph.D.

Cornell University 2018

Statistical inference on high-dimensional and noisy data is a central concern of modern computer science. Often, the main challenges are inherently computational: the problems are well understood from a purely statistical perspective, but key statistical primitives – likelihood ratios, Bayes-optimal estimators, etc. – are intractable to compute on large and high-dimensional data sets.

We develop a unified approach to algorithm design for statistical inference based on the Sum of Squares method, a powerful tool for convex programming with low-degree polynomials, which generalizes linear programming and spectral algorithms. We use this approach to design algorithms for a range of high-dimensional statistical problems, improving on state-of-the-art provable guarantees in several well-studied settings: clustering data from high-dimensional mixture models, community detection in random graphs, and more.

We also prove computational lower bounds for some statistical problems, including the long-studied *planted clique* problem. Our lower bounds provide new strong evidence for the existence of information-computation gaps – that is, statistical problems which are solvable given infinite computational resources, but not by efficient algorithms. In particular, we prove new lower bounds against the powerful Sum of Squares hierarchy of semidefinite programs, via a new *pseudocalibration* technique. Because the Sum of Squares hierarchy has provable guarantees matching those of most known techniques in algorithm design for inference, our lower bounds strongly suggest that the problems we study are

intractable for polynomial-time algorithms. At very least, improving existing algorithms for them would require a major breakthrough.

We show that polynomial-size semidefinite programs from the Sum of Squares hierarchy cannot refute the existence of cliques of size much less than $\sqrt{n}$ in n-node random graphs. Additionally, we prove a lower bound for sparse principal component analysis (PCA), showing that subexponential-size Sum of Squares semidefinite programs are needed to improve on the provable guarantees of existing spectral algorithms for sparse PCA.

Our approach to algorithms and lower bounds suggests a new method to chart the edge of algorithmic tractability for statistical inference. We propose a classification of Bayesian inference problems according to solvability by algorithms which compute only simple statistics of input data – triangle counts of graphs, top eigenvalues of matrices, etc. Our classification accurately predicts suspected information-computation gaps for many well-studied problems, including planted clique, planted constraint satisfaction, community detection in stochastic block models, component analysis problems, and more. This type of classification is novel for inference problems, and represents the first approach to trace the information-computation gaps in of all these problems to the same underlying mathematical structure.

BIOGRAPHICAL SKETCH

Sam Hopkins was born in Seattle, WA. He obtained a B.S. in mathematics and computer science from the University of Washington. After obtaining his Ph.D. from Cornell, he will join the University of California at Berkeley as a Miller Fellow.

ACKNOWLEDGEMENTS

Too many people are owed thanks to be listed here, but I will do the best that I can. First and foremost, I could not have asked for a better advisor than David Steurer. David was extraordinarily generous with his time and energy, and was an endless source of ideas and optimism. I cannot thank him enough for the countless hours of discussion and mathematics.

Boaz Barak hosted me for a summer at Microsoft Research New England and became like a second advisor. Thank you, Boaz, for sharing your limitless exuberance about all things SoS (and the rest of theoretical computer science besides).

I have had the great luck to work with many outstanding collaborators during the last five years, without whom the papers in this thesis would never have been written. Many thanks to Jon Kelner, Pravesh Kothari, Jerry Li, Ankur Moitra, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and Jonathan Shi. Over the years several of you have become close personal friends, for which I am very grateful. Thanks in particular to Pravesh and Tselil, who stood with me through countless and frantic all-night combinatorics sessions when deadlines were near: I remember them (mostly) fondly. And thanks, Tselil, for heroic feedback on the early chapters of this thesis.

The UC Berkeley theory group became a second home over the last few years. Thanks to Prasad Raghavendra and the Simons Institute for hosting me for several semesters, and to Tselil, Jonah, Ben, Aviad, and Alex for all the shenanigans.

Bobby Kleinberg and Dexter Kozen sat on my committee, and were wonderful sources of advice and interesting conversation throughout my PhD. Èva Tardos's door was always open to talk about math, academia, and life. Jennifer Chayes and Christian Borgs also hosted me for a summer at Microsoft Research New

England, and greatly improved my understanding of sparse random graphs (which play a central role in some parts of this thesis). Thanks to all of them.

The other students in the Cornell theory group made it a pleasure to come to the office: thanks especially to Thodoris, Rahmtin, and Daniel for sharing what must have been the better part of a thousand salads from the Terrace cafe. I was lucky enough to find a surrogate family of Ithacan friends: thanks in particular to Jonathan, Daniel, Molly, Mark, and Eoin for making 109 Lake Street such a pleasant place to live.

Flora, thank you for your patience, encouragement, and companionship, and for sharing Ollie from time to time.

And thanks to my family, to whom I owe it all.

The work in this thesis was supported by grants and fellowships from the National Science Foundation, the Simons Foundation, Microsoft, and Cornell University.

CONTENTS

CHAPTER 1

INTRODUCTION

Algorithms to extract useful information from noisy and high-dimensional data are a central concern of modern computer science. Progress in the design of such algorithms in the last 20 years has revolutionized statistics and artificial intelligence, enabling ever-more-sophisticated inferences based on ever-larger data sets—from scientific instruments, sensor networks, mobile phones, and (above all) the Internet. It is hard to overstate their breadth and depth of application.

Designing these algorithms remains more art than science. Exploiting the unprecedented size of data sets now available often involves overcoming serious challenges. Chiefly:

1. (*Computational intractability*) Approaches to hypothesis testing and inference suggested by classical statistics rarely scale well. Efficient algorithms—with running times which are polynomial or (even better) linear in data-set size—do not even exist for high-dimensional versions of many basic tasks.

   For example, computation of a likelihood ratio or a maximum-likelihood estimator often appears to require brute-force enumeration of sets which have size exponential in ambient dimension or number of samples, both large.

2. (*Data sparsity*) The classical way to make an inference problem easier is to gather more data: for simple tasks like estimating a population mean, more data always increases the accuracy of the standard estimator, in this case the sample mean. However, for many problems we consider, there is

not just one parameter, like the population mean, that we wish to estimate: instead, the complexity of information to be inferred grows with the data set size. In particular, the amount of data *per bit of information to be inferred* often remains constant.

For example, social network graphs have millions or billions of nodes but often constant average degree: inferring a label (say, Republican or Democrat) for every node of such a graph from just the graph structure requires inferring one bit of information per vertex from just a constant number of edges per vertex.

These challenges are intimately related: computational intractability often originates with data sparsity. Inferring $O(1)$ bits of information from $N \to \infty$ noisy samples – estimating a population mean from a growing number of samples, for example – is well studied and often computationally tractable. Inferring $\Omega(N)$ bits from N samples is a different endeavor. In this thesis we will most often be concerned with the latter sort of inference problem.

While data sparsity is a key source of computational intractability, it is not the only one. Even simple tasks, like binary hypothesis testing, can lack obvious efficient algorithms: although binary hypothesis testing has been studied since the birth of modern statistics, the classical statistical methods are often not computationally tractable in high-dimensional settings. Indeed, some binary hypothesis testing problems likely lack efficient algorithms altogether.

In spite of the challenges, theorists and practitioners have discovered clever and surprising algorithms for some noisy, high-dimensional, and data-sparse inference problems, while others seem to resist such solutions, despite intensive effort. The line between efficiently and not efficiently solvable remains murky,

and in large part investigated only on a problem-by-problem and algorithm-by-algorithm basis. The aim of this thesis is to develop systematic and principled approaches to answer the central question:

> *Which noisy and high-dimensional statistical inference problems admit computationally-efficient algorithms, and which do not?*

A satisfactory theory of algorithms for statistical inference should be able to *predict* (without exhaustive study of numerous potential algorithms) whether a given inference problem is computationally tractable. It should offer evidence, as rigorous as possible, for the correctness of its prediction. And, it should *explain* what makes an inference problem tractable or intractable, by identifying problem-and algorithm-independent mathematical structures which track the limits of efficient computation.

Besides addressing core theoretical questions, this kind of theory has the potential for wide impact on algorithm design for inference. By identifying the boundaries of algorithmic tractability, we may hope for new algorithms solving challenging inference problems where current techniques do not yet approach the limits of efficient computation. On the other hand, providing evidence for intractability gives algorithm designers a principled way to decide when to *stop* designing algorithms and reformulate their problems or devote their efforts elsewhere.

At present the prevailing mathematical approach to algorithmic statistical inference is to focus on one problem and one tailor-made algorithm at a time. While successful on a problem-by-problem basis, this approach falls short of addressing our main question in several ways. First, it does not suggest a way

to predict (in)tractability of a newly-presented inference problem, except by exhaustive attempts to design an algorithm.[1] Should these attempts fail, it offers no avenue to produce rigorous evidence of intractability. And, without connecting algorithms for one problem to those for the next, it does not suggest any problem-independent mathematical structures which could trace out the tractable/intractable divide.

In this thesis we begin by unifying many algorithms for various inference problems under one algorithmic umbrella – *the Sum of Squares method*, or *SoS*. At heart, SoS is a method to design convex relaxations for polynomial optimization problems. In this thesis, we develop an approach to the use of SoS for inference. We observe that the performance guarantees of a wide range of existing inference algorithms – in particular spectral and convex-programming methods – are captured by algorithms using SoS. By developing extensive tools for the analysis of SoS algorithms we also obtain new polynomial-time provable guarantees for a number of inference problems.

We then study what makes inference problems *intractable* for SoS algorithms, since the broad algorithmic power of SoS makes SoS-intractability a strong proxy for polynomial-time intractability. We develop *pseudocalibration*, which is the first problem-independent approach to proving lower bounds against convex-programming-based algorithms for inference problems.

Taken together, our algorithms and lower bounds point to a simple problem-

[1] To the extent that existing methods do offer such predictions, they usually rely on heuristics carried over from worst-case complexity and algorithm design. For example, it has long been understood that matrix problems are algorithmically easier to solve than tensor problems; this is just as true in statistical as in worst-case settings [87, 92]. But ideas rooted in worst-case complexity do not predict or explain phenomena which appear inherently statistical: for example, they do not explain why some matrix-based inference problems appear computationally harder than others, or predict which are the easy ones.

and algorithm-independent criterion for tractability of a broad class of Bayesian inference problems. This class includes long-studied high-dimensional hypothesis testing and hidden variable problems involving dense subgraph detection (e.g. planted clique), component analysis (e.g. sparse PCA), and community detection in large networks (e.g. the stochastic block model). Prior to this work, even making plausible non-rigorous guesses about the performance of SoS for such problems demanded substantial area-specific expertise and creativity; our work largely reduces this task to rote computation.

We develop a meta-theory of SoS algorithms for such Bayesian problems – inference problems where hidden variables are distributed according to a prior – which hinges on a deep and unexpected connection between algorithms based on convex programs – such as the SoS method – and those based on *simple statistics* – subgraph counts in graphs, top eigenvalues of matrices, and the like. Along the way, we develop a substantial new technical toolkit for rigorous analysis of the SoS method in inference settings.

Finally, we contribute to algorithm design for inference in problems which do not fit the Bayesian mold, which is broad but not all-encompassing. We use SoS to design a new algorithm for learning high-dimensional mixture models, a core data clustering problem in statistics which dates to foundational work of Pearson in the 1890s. In some (well-studied) regimes, our algorithm is the first to improve on the guarantees of naïve greedy clustering methods in polynomial time. We also use SoS to design algorithms for high-dimensional inference which tolerate many adversarially-chosen outliers: this *robust statistics* setting captures learning in the presence of model misspecification as well as malicious data tampering.

We present a problem with the following structure:

1. Unify existing algorithms for statistical inference with provable guarantees under a single umbrella (the SoS method),
2. Employ the resulting technical insight to design new algorithms with provable guarantees for challenging inference problems, and
3. Probe the limits of efficient computation by investigating what makes these algorithms break down.

This roughly parallels developments which have been key to the maturation of several other fields in theoretical computer science.

For example, the unification of approximation algorithms for constraint satisfaction problems (CSPs) under the umbrella of semidefinite programming (SDP) led to the revolutionary discovery that the mathematical structures which make SDPs fail to solve CSPs (within certain approximation ratios) are exactly what is needed to prove CSPs are NP-hard to approximate (within those same approximation ratios), under the (increasingly plausible) unique games conjecture [152]. Subject to the resolution of the unique games conjecture, this led to the solution of almost every long-standing problem in the field and yielded exactly the sort of predictive and explanatory theory whose beginnings we aim to develop here. It is auspicious (and not entirely coincidental) that our algorithmic umbrella – the SoS method – is a natural generalization of the SDP method for CSPs.

Organization In the remainder of this introduction, we discuss some of the main theorems in the thesis (Section 1.1) and some of the mathematical themes which run through it (Section 1.2). In the subsequent few chapters we describe in more detail many of the meta-theoretical ideas in this thesis: simple statistics, the SoS method, proofs-to-algorithms, and pseudocalibration. The rest of the

thesis has two parts: Part I focuses on algorithms and Part II focuses on lower bounds and the pseudocalibration method

1.1 Highlights of Results

In the course of developing the broad theory outlined above, we obtain many new algorithms and lower bounds for challenging inference problems. We are not yet prepared to describe all of the main results of this thesis rigorously, but we will take a short tour of some of the highlights.

A Nearly-Tight SoS Lower Bound for Planted Clique Detecting dense subgraphs in large random graphs is a core problem in fields from biology (DNA transcription networks, neuron connectivity graphs) to social networks (detecting communities) to economics (computation of Nash equilibria, networks of financial derivatives) to cryptography (hiding secret keys by hiding dense subgraphs). [82, 148, 130, 97, 86, 20, 16, 100, 15].

The *planted clique* problem is a simple and long-studied mathematical model for this task – dense subgraphs are assumed to be cliques, and random graphs are assumed to have independent edges which each appear with probability $\frac{1}{2}$ (as in the Erdős-Rényi $G(n, \frac{1}{2})$ model). Despite its simplicity, the problem has prompted many interesting developments in algorithm design, and it displays our first example of the *information-computation gap* phenomenon, which will concern us throughout this thesis.[2]

[2]It is also closely related to two old questions in the theory of algorithms for random graphs: (1) What is the size of the largest clique which may be found (with high probability) in polynomial time in a graph $G \sim G(n, \frac{1}{2})$? (2) What is the tightest upper bound which can be *certified* in

For every $k(n) \in \mathbb{N}$, the planted k-clique problem (phrased as a hypothesis testing problem) is the following. Given an n-node graph G, distinguish the following two hypotheses:

> H_0 : G was sampled from the Erdős-Rényi model $G(n, \frac{1}{2})$, with independent edges each appearing with probability $\frac{1}{2}$.
>
> H_1 : G was sampled by first choosing a set S of k out of n vertices at random, adding every edge between vertices in S to form a clique, then sampling the rest of the edges independently as in $G(n, \frac{1}{2})$.

The hypotheses H_0, H_1 are statistically indistinguishable (meaning there is simply not enough information in a single graph G to distinguish them) for any $k < (2 - \varepsilon) \log n$ (where $\varepsilon > 0$ is any positive constant). The size of the maximum clique in $G \sim G(n, \frac{1}{2})$ is $(2 \pm o(1)) \log n$ with high probability, meaning that if $k > (2 + \varepsilon) \log n$ a brute-force algorithm enumerating all subsets of vertices of size $(2 + \varepsilon) \log n$ will successfully distinguish H_0 and H_1. This algorithm, however, requires quasipolynomial time; that is, time $n^{O(\log n)}$.

The smallest k for which H_0 and H_1 are known to be distinguishable in polynomial time is $k = c\sqrt{n}$ for any constant $c > 0$: notice that this value of k is *exponentially greater* than the smallest such k achievable in quasipolynomial time, via the brute-force algorithm. This is an *information-computation gap*, with an apparent *algorithmic threshold* for polynomial-time algorithms at $k = \Theta(\sqrt{n})$.

Rigorously establishing the existence of such an algorithmic threshold takes two kinds of evidence.

---

polynomial time on the size of the largest clique in $G \sim G(n, \frac{1}{2})$? That is, what is the least $c = c(n)$ such that there exists a polynomial-time algorithm outputting a number $ALG(G)$ such that $ALG(G) \geqslant \textsc{max clique}(G)$ for every G, and $\mathbb{E}_{G(n,\frac{1}{2})} ALG(G) = c$?

First, to demonstrate tractability above the threshold, there should be an algorithm. In this case, an algorithm based on maximum eigenvalues of submatrices of the adjacency matrix of the graph G distinguishes H_0 and H_1 for any $k \geqslant \Omega(\sqrt{n})$ [9].[3]

Second, to demonstrate intractability below the threshold, there should be a lower bound: some evidence that no polynomial-time algorithm distinguishes H_0 and H_1 from a single sample G when $k < o(\sqrt{n})$. Since we do not expect to prove $P \neq NP$, the evidence must either be *conditional* – that is, dependent on some other unproven conjecture – or apply only to a subset of polynomial-time algorithms, ideally one as large as possible.

While conditional evidence – usually based on the conjecture $P \neq NP$ – has been very successful in establishing *worst-case* computational intractability, the resulting *reduction-based* theory does not seem readily adaptable to the inference setting. The difficulty is that combinatorial reductions between problems, which form the core of the theory of NP-completeness, produce problem instances with detailed combinatorial structure. (In the jargon, the instances are full of *gadgets*.) This makes them unlike the instances of an inference problem, which, as in planted clique, are largely random. *In particular, there is no known conditional lower bound for the planted problem based on any standard complexity hypothesis, and none appears to be in sight.*

We give the strongest-yet unconditional evidence that polynomial-time algorithms do not distinguish H_0, H_1 for $k \ll \sqrt{n}$, by ruling out SoS-based algorithms

[3]This algorithm relies on the classical fact from random matrix theory that the maximum eigenvalue of a symmetric $n \times n$ matrix with ± 1 iid entries is $O(\sqrt{n})$; this is one view on the origin of the $\Theta(\sqrt{n})$ threshold. Later on we will identify some alternative characterizations of this threshold.

for such k.[4] The natural application of SoS to planted clique yields a hierarchy of semidefinite-programming (SDP) relaxations of the MAX CLIQUE problem; the d-th SDP in the hierarchy has $n^{O(d)}$ variables and is solvable in time $n^{O(d)}$. Since MAX CLIQUE(G) does distinguish H_0, H_1 for $k > (2+\varepsilon)\log n$, trying to solve MAX CLIQUE by convex relaxation is a reasonable approach to the problem. This approach succeeds for any $k \geqslant \Omega(\sqrt{n})$, because the objective value $\mathrm{SoS}_d(G)$ of the d-th SDP on $G \sim G(n, \frac{1}{2})$ is at most $\sqrt{n}/2^{O(d)}$, with high probability [71]. We prove a strong converse, matching the upper bound up to a subpolynomial factor:

Theorem 1.1.1 (SoS Lower Bound for Planted Clique)**.** *For every $d > 0$ and $G \sim G(n, \frac{1}{2})$, with high probability $\mathrm{SoS}_d(G) \geqslant \sqrt{n}/2^{O(\sqrt{d/\log n})}$.*

To prove this lower bound we introduce the *pseudocalibration* technique, which marks a major technical development in our ability to prove lower bounds against large convex programs for inference problems. (As we will see, in a rigorous sense previous techniques to prove lower bounds for SoS algorithms could not prove a tight lower bound for planted clique.) It reveals a fascinating connection between the SoS SDP on the one hand and simple graph statistics – in this case, counts of small subgraphs – on the other. In particular, it shows that SoS algorithms fail to distinguish H_0, H_1 when $k \ll \sqrt{n}$ *because no simple subgraph-counting algorithm can distinguish H_0, H_1 for such k*.

This result almost tightly characterizes the values of k for which the planted-k-clique problem is solved in polynomial time by SoS. We are able to extend the ideas to prove similar characterizations for two other important hypothesis testing problems: *sparse principal component analysis* and *single-spiked tensors*, also known

[4] The strength of our lower bound stems from the power of SoS as an algorithm: in particular, SoS generalizes all the spectral and convex programming methods for which planted clique lower bounds are known [71].

as *tensor principal component analysis*.[5] In fact, the pseudocalibration technique unifies all existing lower bounds against SoS algorithms for statistical problems of which the author is aware.

Each of these lower bounds (and nearly-matching algorithms) demonstrate that the SoS algorithms can solve a particular hypothesis testing problem exactly when it is also solved by algorithms based on *simple statistics:* more formally, by tests which are computed by low-degree polynomials in the problem instance. (For instance, subgraph counts in a graph G are low-degree polynomials in the entries of the adjacency matrix of G.)

We discuss pseudocalibration and the challenges of proving Theorem 1.1.1 in Chapter 4. We prove Theorem 1.1.1 in Chapter 11.

SoS Algorithms for Binary Hypothesis Testing are Only as Good as Low-Degree Spectral Methods This prompts the question: is there a more general phenomenon at work, characterizing SoS algorithms for hypothesis testing in terms of simple statistics? This remains an enticing question. We make significant progress towards resolving it, by characterizing SoS algorithms for a wide range of binary hypothesis testing problems in terms of simple *matrix* statistics, a type of spectral method.

This result is somewhat meta-theoretical, so for now we state an informal version and leave the formal definitions till later.

Theorem 1.1.2 (SoS versus simple matrix statistics for binary hypothesis testing, informal)**.** *Suppose that* (H_0, H_1) *are a pair of distributions on* $\mathbb{R}^n$*, forming a hypothesis testing problem where the goal is to distinguish a sample* $x \sim H_0$ *from* $x \sim H_1$*. Further*

[5] These problems are defined in Section 11.7 and Chapter 6.

suppose that (1) H_0 is a product distribution $H_0 = \nu^{\otimes n}$, and (2) there is a polynomial-time SoS algorithm to distinguish H_0 from H_1', where H_1' is the following noisy version of H_1: to sample $x' \sim H_1'$, first sample $x \sim H_1$, then obtain x' from x by resampling a random 0.1*-fraction of its coordinates from μ.*

Then there is $m \leqslant n^{O(1)}$ and a collection of m^2 polynomials $Q_{ij} : \mathbb{R}^n \to \mathbb{R}$ of degree $O(\log n)$ such that the value of the largest eigenvalue of the matrix Q_{ij} distinguishes H_0 and H_1.

This theorem applies to nearly any hypothesis testing problem where the null model H_0 is a product distribution and the problem is *noise-robust:* the noise-robustness is captured by the replacement of the alternative hypothesis H_1 with a noisy version H_1'. In particular, the theorem applies to classic high-dimensional problems like planted clique, sparse principal component analysis, community detection, random constraint satisfaction, and more. It is also the other main use of the pseudocalibration technique in the thesis. We prove it in Chapter 12.

Detecting Overlapping Communities in Very Sparse Random Graphs *Community detection* is the problem of finding collections of similar nodes in networks. Communities can of course represent social groups or like individuals in social networks [140], but community detection finds broad application across the sciences: for instance in image segmentation [163] and in identification of biological pathways from protein-protein interaction graphs [123, 50]. (See e.g. [1] for further references.)

The *stochastic block model* (SBM) is the canonical family of generative models for random graphs with hidden community structure. The model in its most elementary form was independently invented in several communities – statistics

and machine learning, theoretical computer science, and mathematics. Studied since at least the 1980s, the SBM has been the site of a number of recent developments identifying a particularly stark information-computation gap.

Consider the following hypothesis testing problem, parameterized by $d, k \in \mathbb{N}$ and $\varepsilon > 0$. (Of course eventually one wishes to cluster nodes by community in addition to testing whether communities exist, but hypothesis testing is almost always a good starting place.)

> H_0 : G was sampled from the Erdős-Rényi model $G(n, \frac{d}{n})$, with independent edges each appearing with probability $\frac{d}{n}$.
>
> H_1 : G was sampled by first choosing a random element σ_i of $[k]$ for $i \in [n]$, then independently choosing to include each potential edge $\{i, j\} \subseteq \binom{n}{2}$ with probability $(1 - \varepsilon/k)\frac{d}{n}$ if $\sigma_i \neq \sigma_j$, and otherwise with probability $(1 + (1 - 1/k)\varepsilon)\frac{d}{n}$.[6]

The distribution specified by H_1 is called the k-community stochastic block model.

A remarkable analysis of this hypothesis testing problem in the regime $d = \Theta(1)$ (which is perhaps the most plausible for applications real-world networks, which tend to be very sparse) using non-rigorous tools from statistical physics [59] suggested an enticing conjecture:

Conjecture 1.1.3. *There is a polynomial-time algorithm to distinguish H_0, H_1 with probability $1 - o(1)$ if and only if $d \geqslant (1 + \delta)k^2/\varepsilon^2$ for some $\delta \geqslant \Omega(1)$.*

Since in exponential time it is possible to distinguish H_0, H_1 for $d \geqslant \frac{C \log k}{k\varepsilon^2}$ for some universal constant C [25], Conjecture 1.1.3 predicts an information-

[6] The coefficients $(1 - \varepsilon/k)$ and $(1 + (1 - 1/k)\varepsilon)$ are chosen so that the graph models described in H_0 and H_1 have the same expected number of edges.

computation gap for the k-community stochastic block model. Rather remarkable is the predicted sharpness of the computational threshold: the conjecture claims that the algorithmic tractability of the hypothesis testing problem differs qualitatively for $d = 1.01k^2/\varepsilon^2$ and $d = 0.99k^2/\varepsilon^2$. For historical reasons, in the block model context $d = k^2/\varepsilon^2$ is called the "Kesten-Stigum threshold".

A series of algorithmic innovations in the block model has culminated with a proof of one side of the conjecture: namely, that there is a polynomial-time algorithm to detect communities for any $d > (1+\delta)k^2/\varepsilon^2$ [2, 42, 111, 137, 136, 125]. These algorithms are impressive technical achievements, arising from detailed study of the particulars of the k-community block model. On one hand, their model-specificity allows them to detect communities up to the Kesten-Stigum threshold. But on the other hand, it is not immediately clear what these algorithms and the sharpness of the Kesten-Stigum threshold say about inference more broadly: for example, what feature of the block model makes for a sharp computational threshold, while other problems it is possible to trade off statistical noise and running time (e.g. detecting a planted clique of size $\sqrt{n}/2^d$ in time $n^{O(d)}$). Finally, as usual, model-specific algorithm design leaves little avenue to offer rigorous evidence for the other half of the conjecture: that when $d < k^2/\varepsilon^2$ no polynomial-time algorithm distinguishes H_0, H_1 with probability $1 - o(1)$.

We develop an approach to algorithm design for community detection in the stochastic block model based explicitly on the combination of simple statistics and the SoS method. Like the physics-inspired methods, our approach solves the hypothesis testing problem (and the attendant vertex-clustering/labelling problem) for any d, k, ε above the Kesten-Stigum threshold. (We also show that the physics-inspired algorithms themselves can be thought of as simple-statistics-

based.) Furthermore, our approach is sufficiently model-independent that it extends easily to design an algorithm for a substantially more complicated (yet more realistic) version of the block model, in which each node may participate in several communities simultaneously [6].

Finding rigorous explanations for this and similar non-rigorous predictions of statistical physics is a significant project in itself. Related predictions and rigorous confirmations thereof have led recently to breakthroughs in long-standing questions on random constraint satisfaction problems [69, 70, 164].

We design an algorithm for the following hypothesis testing problem (and, perhaps more interestingly, for its estimation variant, where the goal is to recover the latent community structure in the H_1 case).

> H_0 : G was sampled from the Erdős-Rényi model $G(n, \frac{d}{n})$, with independent edges each appearing with probability $\frac{d}{n}$.
>
> H_1 : G was sampled by first choosing t random elements $\sigma_1 = \{\sigma_{i1}, \ldots, \sigma_{it}\}$ of $[k]$ for $i \in [n]$, then independently choosing to include each potential edge $\{i, j\} \subseteq \binom{n}{2}$ with probability $(1+\varepsilon(|\sigma_i \cap \sigma_j| - t/k))\frac{d}{n}$.

The random graph model described by H_1 is called the *mixed-membership* stochastic block model [6]. We prove the following theorem, capturing a Kesten-Stigum-like sharp threshold for the mixed-membership block model.

Theorem 1.1.4. *For every $\delta > 0$, if $d \geqslant (1+\delta)k^2(1+\alpha)^2/\varepsilon^2$, where α is defined by $\frac{k(\alpha+1)}{k+\alpha} = t$, so $\alpha \approx t - 1$, there is a polynomial-time algorithm to distinguish H_0, H_1 with probability $1 - o(1)$.*

Our theorem generalizes previous work on the k-community block model,

captured by the limiting case $t = 1$. Our algorithm works when $d \geqslant (1+\delta)k^2(1+\alpha)^2/\varepsilon^2$ *because for such* d*, there is a simple statistic which distinguishes* H_0, H_1. This represents an appealing converse to our SoS lower bounds and the pseudocalibration technique, which can rule out polynomial-time algorithms based on SoS exactly when simple statistics fail.[7] We show that the Kesten-Stigum threshold can be explained by such simple statistics.

Theorem 1.1.5. *Let* $\mathbb{R}[x_{ij}]_{\leqslant D}$ *be the polynomials of degree at most* D *in* n^2 *variables* $\{x_{ij}\}_{i,j\in[n]}$ *with real coefficients. Let* $\mathcal{S}_{D,d} \subseteq \mathbb{R}[x_{ij}]_{\leqslant D}$ *be the set of simple statistics of degree* D *for hypothesis tests with respect to* $H_0 = G(n, d/n)$*:*

$$\mathcal{S}_D = \{p \in \mathbb{R}[x_{ij}]_{\leqslant D} \,:\, \mathop{\mathbb{E}}_{G(n,d/n)} p(G) = 0, \mathop{\mathbb{E}}_{G(n,d/n)} p(G)^2 = 1\}\,.$$

Then

$$\max_{p\in\mathcal{S}_D} \mathop{\mathbb{E}}_{G\sim H_1(k,\varepsilon)} p(G) = \begin{cases} O(1) \text{ if } d < k^2/\varepsilon^2, D < n^{0.01} \\ n^{\Omega(1)} \text{ if } d > k^2/\varepsilon^2, D \geqslant O(\log n) \end{cases}\,.$$

We discuss this theorem and a similar result we prove for the estimation problem in the block model (rather than the hypothesis testing problem), in Chapter 8. In Chapter 2 we will discuss the degree-D polynomials from this theorem statement at some length. For now we move on to the last stop on our tour of highlights.

Learning High-Dimensional Mixture Models *Clustering* is a central concern in modern statistics and computer science, especially clustering of noisy and

[7] As of the writing of this thesis it remains an enticing open problem to apply the pseudocalibration technique to the block model setting. Though there seems to be no inherent reason why pseudocalibration could not be applied to prove an SoS lower bound, sparse random graphs pose many mathematical difficulties, preventing proof techniques developed in the planted clique setting from applying directly to the stochastic block model. See Appendix A.

high-dimensional data. We study one of the main problems in the field, which dates back at least to Pearson in the 1890s: learning in mixture models.

Suppose that $X_1, \dots, X_n \in \mathbb{R}^d$ (perhaps representing images, or people, or proteins, etc.) are *heterogeneous* in that they contain representatives from several distinct underlying populations, jumbled together in one data set. Another way of putting it is: $X_1, \dots, X_n$ break up into a set of k clusters, according to which underlying population they came from. Under what conditions on $X_1, \dots, X_n$ and the underlying populations is it possible to find the clustering, and to estimate interesting statistical information about the underlying k populations, such as the true mean of each?

Modern data is often high-dimensional (if each X_i is an image, it may contain hundreds of thousands of pixels, for example), and there may be many distinct populations. So our goal is to design algorithms that recover clusters and population means given $n \leqslant \text{poly}(k, d)$ samples and running in time $\text{poly}(k, d)$.

A mixture model is a generative process for this problem. Specifically, suppose $\mathcal{D}_1, \dots, \mathcal{D}_k$ are some probability distributions supported on $\mathbb{R}^d$. The (uniform) mixture model on $\mathcal{D}_1, \dots, \mathcal{D}_k$, denoted $\frac{1}{k} \sum_{i \leqslant k} \mathcal{D}_i$, is the probability distribution which first samples i uniformly from $[k]$, then outputs a sample $X \sim \mathcal{D}_i$.

Without some additional assumptions on the distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$, accurate estimation of their means $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ from $\text{poly}(k, d)$ samples is information-theoretically impossible, which should not be surprising, given the wealth of possible distributions $\mathcal{D}_i$. Even if each $\mathcal{D}_i$ is a (spherical) Gaussian (in which case the mixture is called a *mixture of Gaussians*) with mean μ_i, that is $\mathcal{D}_i = \mathcal{N}(\mu_i, \text{Id})$, if the population means μ_i are too close together (at, say,

Euclidean distance $1/\mathrm{poly}(k)$), k Gaussians can conspire to be indistinguishable from just a single Gaussian unless $n \geqslant 2^{\Omega(k)}$ [132].

However, a pair of populations with means so close that exponentially-many samples are needed to distinguish them as distinct populations is a dubious modeling choice. It is therefore only mildly restrictive to require that $\|\mu_i - \mu_j\| \geqslant \Delta$ for some Δ large enough that $n \leqslant \mathrm{poly}(k, d)$ samples $X_1, \ldots, X_n$ suffice (at least, ignoring computation time) to estimate $\mu_1, \ldots, \mu_k$. For this purpose, in the Gaussian setting, $\Delta = \Theta(\sqrt{\log k})$ suffices [158]. However, until the work presented in this thesis, polynomial-time algorithms were known to estimate $\mu_1, \ldots, \mu_k$ only under the assumption that $\Delta \geqslant \min(d, k)^{1/4}$; for smaller Δ the only available algorithms required exponential time [174].

We show the following theorem. (In fact, we show a more general version tolerating a wide range of underlying distributions $\mathcal{D}_i$.)

Theorem 1.1.6 (Special case of main theorem on mixture models, informal)**.** *For every $\varepsilon > 0$, if $\Delta \geqslant k^{\varepsilon}$ then there is an algorithm with running time $\mathrm{poly}(k, d)$ which estimates $\mu_1, \ldots, \mu_k$ up to $1/\mathrm{poly}(k)$ error from $n \leqslant \mathrm{poly}(k, d)$ samples $X_1, \ldots, X_n \sim \frac{1}{k}\sum_{i \leqslant k} \mathcal{D}_i$, where $\mathcal{D}_i = \mathcal{N}(\mu_i, \Sigma_i)$ and $\Sigma_i \preceq \mathrm{Id}$. Additionally, for some universal constant C, if $\Delta \geqslant C\sqrt{\log k}$, then there is an algorithm with quasipolynomial time, requiring quasipolynomially-many samples, to estimate $\mu_1, \ldots, \mu_k$ up to $1/\mathrm{poly}(k)$ errors.*

This theorem represents the first improvement in the parameter Δ tolerated by polynomial-time algorithms in nearly 20 years, for a wide range of d and k, despite substantial attention in the literature [179, 53, 17, 174, 113, 4, 73, 101, 35, 132, 94, 14, 39, 54, 168, 84, 180, 76, 120, 158, 55].

The mixture model setting differs from those we have discussed so far in several ways. The other inference tasks we described so far, planted clique and the stochastic block model, were cast as binary hypothesis testing problems, while we have described the mixture models problem as one of *parameter estimation*. Of course, both of the other problems have estimation versions: identifying the vertices in a k-clique in the planted clique problem, or partitioning a sparse random graph into communities in the stochastic block model.

In general, the "simple statistics+SoS" theory we have been alluding to applies equally well to the estimation variants of planted clique, stochastic block model, and similar problems. Crucially, these estimation tasks are *Bayesian*. Bayesian estimation problems, like planted clique and the block model, come with a known (*prior*) probability distribution on the hidden variables to be estimated, and the goal is to design an algorithm which solves the estimation problem with high probability over choice of hidden variables from this prior. While Bayesian problems cover a wide range of interesting statistical settings, good priors are not always available, and there is long-running philosophical debate about when it is appropriate to employ prior-based statistical reasoning.

Learning in mixture models, as we have defined it here, requires designing an algorithm that tolerates adversarial choice of hidden variables, subject to some deterministic conditions – in this case the hidden variables are $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ (or more generally the distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$), and the condition is that $\|\mu_i - \mu_j\| \geqslant \Delta$. The algorithm should succeed with high probability over $X_1, \dots, X_n$, *for any such choice of* $\mu_1, \dots, \mu_k$.

Our algorithm still uses the SoS method and incorporates simple statistics, but requires substantial additional innovations. In particular, we employ the

proofs-to-algorithms method, which we lay out in Section 3.5. The success of the SoS method even for this kind of prior-free estimation problem suggests its power as a unifying tool beyond the Bayesian setting. We prove Theorem 1.1.6 in Chapter 9.

1.2 Themes

As important as the main results in this thesis are the threads that bind them into the start of a coherent theory of algorithms for statistical inference. We address a few of those themes now.

Simple proofs beget polynomial-time algorithms Our starting point is an approach to algorithm design for statistical inference based on *simple proofs* and convex programs. Imagine binary hypothesis testing with a null hypothesis H_0 and an alternative H_1, both specified by probability distributions on (for example) $\mathbb{R}^n$.

Under the assumption that both hypotheses are equally likely to be true, from a statistical point of view the optimal hypothesis test is the *likelihood ratio test* – given $x \in \mathbb{R}^n$, if $\mathbb{P}_{H_0}(x) < \mathbb{P}_{H_1}(x)$ the test outputs ALTERNATIVE, and otherwise the test outputs NULL. The trouble is that for many interesting alternative models, efficient algorithms to evaluate $\mathbb{P}_{H_1}(x)$ may or may not exist – for example, if x is the adjacency matrix of a graph and H_1 is the hypothesis that x was sampled from a distribution on graphs containing planted dense subgraphs, then computing $\mathbb{P}_{H_1}(x)$ amounts to detecting dense subgraphs of a random graph.

To prove that it is possible to test H_1 against H_0 given arbitrary computational

resources, one shows that

> (∗) $\mathbb{P}_{H_1}(x) \gg \mathbb{P}_{H_0}(x)$ for typical $x \sim H_1$ and $\mathbb{P}_{H_0}(x) \gg \mathbb{P}_{H_1}(x)$ for typical $x \sim H_0$.

In general such a proof says little about efficient algorithms. But if, to prove (∗), one shows that for typical $x \sim H_0$ there is a simple enough *witness* w_x to the inequality $\mathbb{P}_{H_1}(x) \ll \mathbb{P}_{H_0}(x)$, then often one has also designed a polynomial-time algorithm for (one-sided) hypothesis testing: given x, the algorithm searches for the witness w_x. This strategy works if, for example, there is a convex set $W_x \subseteq \mathbb{R}^{n^{O(1)}}$ with a polynomial-time separation oracle such that any $w \in W_x$ certifies $\mathbb{P}_{H_1}(x) \ll \mathbb{P}_{H_0}(x)$.

The SoS method offers a powerful strategy for such proofs of statements like (∗). In particular, it offers a restricted proof system which automatically comes with simple witnesses: if an inequality $f \leqslant g$ is provable in the SoS proof system, then there is always a simple witness, which we call an *SoS proof*. For each f, g (thinking of $f = \mathbb{P}_{H_1}(x)$ and $g = \mathbb{P}_{H_2}(x)$, the set of simple witnesses to $f \leqslant g$ forms a convex set $W_{f,g}$; testing whether this set is empty via convex programming yields algorithms for hypothesis testing.

We demonstrate this approach to hypothesis testing in Chapter 6, and we discuss in Section 3.5 how it may be extended from hypothesis testing to hidden-variable estimation problems, an extension we use in Chapter 9 to design algorithms for clustering (Theorem 1.1.6) and outlier-robust estimation.

SoS proofs and low-degree matrix polynomials Next, we investigate just how simple the witnesses w_x and the algorithms to find them can be. Inspecting the

SoS proofs constructed for a wide range of inference problems, we observe that the full power of convex programming is often unnecessary. We show that SoS proofs w_x can frequently be boiled down to maximum eigenvalues of *low-degree matrix-valued polynomials*: that is, $\lambda_{\max}(M(x))$ where $M : \mathbb{R}^n \to \mathbb{R}^{n^{O(1)} \times n^{O(1)}}$ is a symmetric matrix whose entries are n-variate polynomials in x of degree $O(1)$.

These matrix polynomials give rise to non-traditional spectral algorithms. Spectral methods typically focus on a few very specific matrices, like adjacency matrices of graphs, or covariance matrices of vector-valued data, whose entries are linear or quadratic functions of input data x By contrast, our methods use a variety of polynomials of higher degree, in many cases obtaining provably stronger statistical guarantees than the traditional approaches. (Expert readers may already know that SoS proofs themselves involve inequalities among low-degree polynomials: *it is important not to conflate the latter polynomials, in some symbolic variables y, with low degree polynomials in the problem instance x.*)

This simplification of SoS proofs for certain inference problems offers an avenue to design truly fast algorithms – having practical, nearly-linear running times – whose guarantees match those achieved by slower convex-programming methods. In Chapters 6 and 7 we design such algorithms for hypothesis testing and hidden-variable estimation in *spiked tensor models* and for detecting *hidden sparse vectors*. The latter algorithm, from Chapter 7, demonstrates another important principle: the spectra of simple, low-degree matrix polynomials like $M(x)$ above can leverage mathematical structure in hidden variables which is more typically exploited only by tailor-made algorithms. For example, *sparsity* in hidden variables is often leveraged in algorithm design by using some form of ℓ_1 regularization, such as in the LASSO method, and by thresholding – setting

small coordinates to 0. Here we achieve essentially the same guarantees with low-degree polynomials.

We investigate in Chapter 12 just how far this kind of simplification can be pushed by proving a meta-theorem on SoS proofs for hypothesis testing and spectra of such low-degree matrix polynomials – Theorem 1.1.2. This theorem applies to all of the *Bayesian* problems considered in this thesis – that is, every problem except those in Chapter 9, where we are reminded that in settings where hidden variables are chosen adversarially instead of sampled from a nice prior distribution, the full power of SoS convex programming holds major advantages over simpler algorithms.

Simple statistics The next theme, which we explore in detail in Chapter 2, arises by a further simplification of the kind of hypothesis testing algorithms under study. In particular, we consider the bold hypothesis that *simple statistics* are just as power full a class of hypothesis tests as SoS proofs and convex programs. A simple statistic is a low-degree *scalar-valued* polynomials $p(x) : \mathbb{R}^n \to \mathbb{R}$, by contrast to the matrix-valued polynomials above.

Because of the power method for computation of maximum eigenvalues, the maximum (magnitude) eigenvalue of an $n \times n$ symmetric matrix M (with a big enough spectral gap) is a degree $O(\log n)$ polynomial in its entries, so the degree-$O(1)$ spectral methods described above are largely captured by degree $O(\log n)$ simple statistics.[8]

In Chapter 8, we study community detection and estimation in sparse random

[8] In fact, only a subtle distinction between the maximum *magnitude* eigenvalue $\max_i |\lambda_i|$ and and the maximum eigenvalue $\max_i \lambda_i$ prevents them from being captured outright.

graphs with this hypothesis in mind. We show that the most powerful existing algorithms for these sparse graph problems, which were originally inspired by seemingly far-afield ideas from statistical physics (message-passing algorithms and the replica method [58]), are captured by simple statistics. We prove Theorem 1.1.4 by designing an algorithm for mixed-membership community detection built explicitly around simple statistics.

By the end of Part I we will have seen in three separate settings that simple statistics capture much of the power of the strongest-known polynomial time algorithms – including those based on SoS – for inference problems that touch three major themes in high-dimensional statistics: non-convex optimization (the spiked tensor model, Chapter 6), detection of sparse hidden variables (the planted sparse vector problem, Chapter 7), and sparsity in the data itself (sparse random graphs and community detection, Chapter 8).

In Chapter 4 and Part II we develop a technique, *pseudocalibration*, which offers some explanation as to why simple statistics are so powerful. We have just described a series of ideas in algorithm design which starts with SoS proofs and ends with algorithms based on simple statistics. If simple statistics are optimal among polynomial-time algorithms for (some) high-dimensional hypothesis testing problems,[9] we should aim to prove converse results: when algorithms based on simple statistics break down, so do (seemingly) more powerful SoS algorithms. While pseudocalibration remains a technically challenging technique, we are able to use it to prove our lower bound for planted clique (Theorem 1.1.1) and our reduction from SoS to low-degree spectral algorithms (Theorem 1.1.2). We discuss the idea further, and why was needed to overcome previous barriers

[9] There are also simple-statistics-based algorithms for parameter estimation problems; see Chapter 8 for example.

to provable lower bounds for SoS-based algorithms, in Chapter 4.

CHAPTER 2

SIMPLE STATISTICS

In this chapter we introduce the simplest version of one of our central theses:

> For *Bayesian* inference problems, the success or failure of algorithms based on *simple statistics* tracks the bounds of algorithmic tractability in general.

This idea will show up throughout the thesis, and can be applied to simple and complex statistical problems. In this chapter we will focus on binary (simple versus simple) hypothesis testing.

2.1 Basics

We start with a couple definitions. Let Ω be a (finite or infinite) alphabet – for example $\{0, 1\}$ or $\mathbb{R}$. We always assume Ω is a group. The following definition is a small modification of the usual notion of degree of a polynomial function $f : \Omega^n \to \mathbb{R}$.

Definition 2.1.1 (Coordinate Degree)**.** Let $n \in \mathbb{N}$ and $f : \Omega^n \to \mathbb{R}$. The coordinate degree of f is the minimum D such that there is a collection of functions $f_i : \Omega^n \to \mathbb{R}$ where $f = \sum f_i$ and each $f_i(x)$ depends on at most D coordinates.

For finite Ω, the notion of coordinate degree agrees, up to factors of $|\Omega|$, with the usual notion of the degree of a function $f : \Omega^n \to \mathbb{R}$ from discrete Fourier

analysis. [1]

Definition 2.1.2 (Simple Statistic)**.** Let $n \in \mathbb{N}$ and let ν be a probability distribution distribution on Ω^n. A function $f : \Omega^n \to \mathbb{R}$ is a D-simple statistic (with respect to ν) for some $D \in \mathbb{N}$ it has coordinate degree at most D and it is normalized with respect to ν – formally, $\mathbb{E}_{x\sim\nu} f(x) = 0$ and $\mathbb{E}_{x\sim\nu} f(x)^2 = 1$.

Simple statistics are everywhere. For example, consider the problem of detecting the presence of an unusually-dense subgraph in a dense random graph. Often, large graphs which contain dense subgraphs also contain a higher-than-expected number of small dense graphs: triangles, 4-cliques, and so on, and these subgraph counts can be used to distinguish such graphs from a null model. The function $f(G) =$ (# of triangles in G) has coordinate degree 3 (as a function of the *edges* of G), so after centering and normalization it is a 3-simple statistic.

A more sophisticated example is the maximum eigenvalue of the adjacency matrix of a graph G, which is also often used to detect anomalous structures in a graph – sparse cuts, dense subgraphs, and more. The maximum eigenvalue of an $n \times n$ symmetric matrix M can be approximated as $(\operatorname{Tr} M^\ell)^{1/\ell}$, for $\ell = O(\log n)$. For any distribution ν, normalizing the function $\operatorname{Tr} M^\ell$ with respect to ν makes it an ℓ-simple statistic.

Common *non*-examples of D-simple statistics (for small D) are values of optimization problems involving a graph G. For example, MAX CLIQUE(G) is not a low coordinate-degree polynomial in the edges of G.

[1] On the other hand, if $\Omega = \mathbb{R}$, the function $f(x) = \sum_{i,j\in[n]} e^{x_i+x_j}$ has coordinate degree 2, even though the function e^x is not a polynomial function. As a result, *coordinate degree is invariant under one-to-one coordinate-wise transformations of* Ω. This is important, since soon we will use coordinate degree as a proxy for computational complexity: the computational complexity of a problem is unchanged by (computably) invertible coordinate-wise transformation of inputs, and proxies for computational complexity should share this property.

Even the optimal values of *convex* programs are not *a priori* simple statistics, despite often being computable in polynomial time. For example, consider the standard linear programming or semidefinite programming relaxation of MAX CLIQUE(G). Computing this value involves running a simplex, ellipsiod, or interior-point algorithm: there is no clear reason for the output of such an algorithm to be expressible as a low degree function.

Nonetheless, *one of the main themes of this thesis is an intimate relationship between simple statistics and convex programs for inference.* This phenomenon largely does not appear in classical worst-case analysis of convex-programming algorithms; it appears to be unique to inference and other average-case settings where inputs to algorithms are drawn from nicely structured probability distributions.

Let ν and μ be distributions distribution on Ω^n. These induce a hypothesis testing problem: for $X \in \Omega^n$, decide between the following two hypotheses:

$$H_0 : X \sim \nu$$
$$H_1 : X \sim \mu\,.$$

H_0 is called the *null hypothesis*, and H_1 is called the *alternative hypothesis*. We will sometimes abuse notation by conflating ν with H_0 and μ with H_1.

A *test* is a function $t : \Omega^n \to \{0,1\}$. We say it is a successful hypothesis test for the pair (H_0, H_1) if, for a uniformly random bit b, when X is sampled according to H_b, we have $\mathbb{E}_b \mathbb{P}_{X\sim H_b}(t(X) \neq b) \leqslant o(1)$ [96].[2]

[2]In this account we are conflating so-called *type 1* and *type 2* errors; if one prefers to avoid the assumption that X is drawn according to a uniform choice of H_0 or H_1, one may study separately the error quantities $\mathbb{P}_{X\sim H_0}(t(X) = 1)$ and $\mathbb{P}_{X\sim H_1}(t(X) = 0)$, but these will play a limited role in this thesis.

We are interested in the question: *what is the minimal running time $T(n)$ such that there is a successful test successful test t computable in time T?*

From the perspective of classical statistics, the question of whether there exists a successful hypothesis test (irrespective of running time) for the pair (H_0, H_1) is completely resolved by the Neyman-Pearson lemma:

Lemma 2.1.3 (Neyman-Pearson (simplified) [141])**.** *For a fixed pair of null and alternative hypotheses H_0, H_1, the text t minimizing $\mathbb{E}_b \mathbb{P}_{X\sim H_b}(t(X) \neq b)$ is the* likelihood ratio test*:*

$$t(X) = \begin{cases} 1 \text{ if } \frac{\mathbb{P}_{H_1}(X)}{\mathbb{P}_{H_0}(X)} > 1 \\ 0 \text{ otherwise.} \end{cases}$$

The Neyman-Pearson lemma and its *likelihood ratio test* reduce the question of whether or not (H_0, H_1) has a successful hypothesis test to the problem of analyzing one canonical test. But very often the running time of naïve algorithms to compute the likelihood ratio $\mathbb{P}_{H_1}(X)/\mathbb{P}_{H_0}(X)$ is exponential in n. Is there a similarly canonical test when we restrict attention to efficient algorithms?

We propose that the *optimal D-simple statistic* is the canonical object to analyze when studying tests with running time approximately $n^{O(D)}$. Since simple statistics need not be 0/1-valued, we introduce a measure of the success of a simple statistic at solving a hypothesis testing problem.

Definition 2.1.4 (Success for simple statistics)**.** Let n, Ω be as above. A D-simple statistic f (with respect to H_0) is successful for the hypothesis testing problem (H_0, H_1) if $\mathbb{E}_{X\sim H_1} f(X) \to \infty$ as $n \to \infty$. (Recall that by definition, $\mathbb{E}_{H_0} f(X) = 0, \mathbb{E}_{H_0} f(X)^2 = 1$.

Hypothesis 2.1.5 (Optimality of Simple Statistics, Informal)**.** *For sufficiently-nice*

hypothesis testing problems (H_0, H_1)*, there is a successful test with running time* $n^{\tilde{O}(D)}$ *if and only if* (H_0, H_1) *admits a successful* D*-simple statistic.*

This hypothesis is not meant to be interpreted literally (hence the nebulous "sufficiently nice"), and indeed it is plainly falsifiable (in small ways) for some canonical problems.[3] Its purpose is heuristic, to provide a plausible first answer to coarse-grained questions like: for each particular choice of null and alternative hypotheses (H_0, H_1), does every successful test for (H_0, H_1) require $2^{n^{\Omega(1)}}$ time or is there a successful quasipolynomial-time or polynomial-time test?

As we will see later in this chapter, there is a simple characterization the *optimal* D-simple statistic for a hypothesis testing problem (H_0, H_1) (in terms of maximizing $\mathbb{E}_{H_1} f$, not necessarily in terms of running time or other measures of optimality) which allows the smallest D for which there is a successful D-simple statistic to be determined by some straightforward computations, for nice-enough hypothesis testing problems. Together with Hypothesis 2.1.5 these computations yield predictions for the computational complexities of many canonical high-dimensional inference problems: planted clique, sparse principal component analysis, community detection, random constraint satisfaction, and more. *Remarkably, these predictions align with proven algorithms and lower bounds for all of these problems, in polynomial-time, quasipolynomial time, and subexponential time regimes, despite the fact that those algorithms and lower bounds were historically developed with problem-specific techniques.*

We state a more formal conjecture below, but let us describe the idea behind

[3] For example, for the planted clique problem, there is an important $(\log n)$-simple statistic involving the top eigenvalue of the adjacency matrix; since this top eigenvalue can be computed in polynomial time it really corresponds to an algorithm with running time $2^{O(\log n)} = n^{O(1)}$ rather than $n^{O(\log n)}$.

Hypothesis 2.1.5.

First of all, in the case $|\Omega| = O(1)$, there is an $n^{O(D)}$-size basis (of monomials) for the functions of coordinate degree D. So, a D-simple statistic can be computed by an $n^{O(D)}$-size circuit by evaluating every degree-D monomial. Recall that the definition of a D-simple statistic requires that $\mathbb{E}_{x\sim H_0} f(x) = 0, \mathbb{E}_{x\sim H_0} f(x)^2 = 1)$. So the gap between the expectation of f under H_1 and under H_0 grows without bound as $n \to \infty$, while its standard deviation under H_0 remains equal to 1. Thus we would expect that there is a threshold value $c(n)$ such that the test which outputs 1 if and only if $f(x) > c(n)$ is successful. (The existence of such a threshold would follow from mild additional concentration assumptions on f, in particular a bound on $\mathbb{E}_\mu f(x)^2$; this kind of concentration is never an obstacle in practice.)

Since in typical n-dimensional settings evaluation of a likelihood ratio naïvely requires $2^{\Omega(n)}$ time, any successful test with subexponential running time is already interesting. For polynomial-time algorithms we will study D-simple statistics for $D = O(1)$ or $D = O(\log n)$.

The other side of the hypothesis is much bolder. If (H_0, H_1) do not admit a successful D-simple statistic for some small D, it is not at all obvious that there should be no successful algorithm. A main contribution of this thesis is a growing mass of evidence, primarily in the form of Sum of Squares lower bounds, that this is indeed the case.

In the remainder of this chapter we accomplish the following. First, we will present one formalization of the lower-bound side of Hypothesis 2.1.5. Then, we characterize the optimal D-simple statistic in terms of the likelihood ratio. Finally,

we make things more concrete by deriving a simple formula for the optimal D-simple statistic for the planted clique problem. We will see the predictions of Hypothesis 2.1.5 borne out as successful $O(\log n)$-simple statistics exist only to detect planted cliques of size at least $\approx \sqrt{n}$.

2.2 Conjecture: Bounded Almost-Independence Fools P

Hypothesis 2.1.5 is heuristic. In this section we formalize one side of the hypothesis in a conjecture: nonexistence of successful $(\log n)^{1.01}$-simple statistics implies nonexistence of polynomial time algorithms. That is: *super-logarithmic almost-independence fools polynomial time.*

Because this conjecture is stronger than $P \neq NP$ it is too much to hope for a proof. However, we can aim for as long as possible a list of examples: classes of hypothesis testing problems (H_0, H_1) and candidate polynomial-time tests which can be shown unsuccessful if H_1 is $(\log n)^{1.01}$-wise independent. The pseudocalibration technique, presented later in this thesis, shows results of this nature for SoS-based hypothesis testing algorithms.

The mathematical setting we construct for the conjecture is designed to capture many inference problems from the high-dimensional zoo, including among others planted clique, densest-k-subgraph, random constraint satisfaction, community detection, and sparse principal component analysis. For all of these problems it (essentially) correctly predicts what extensive literature suggests to be the parameter regimes which are hard for polynomial time algorithms.

To state a formal conjecture we must impose some appropriate constraints on

the hypothesis testing problems considered, for which we need a few definitions.

Definition 2.2.1 (Noise operator)**.** Let Ω be a finite set or $\mathbb{R}$, and let ν be a product distribution on Ω^n. For any other distribution μ on Ω^n and any positive $\delta > 0$, we denote by $T_\delta \mu$ the distribution on Ω^n which is given by first sampling $x \sim \mu$, then sampling $y \sim \nu$, then independently for each coordinate i, replacing x_i with y_i with probability δ.

This kind of noise operation has the effect of destroying algebraic structure which may be present in $x \sim \mu$ and exploitable by efficient algorithms. For example, if μ is a distribution on satisfiable systems of linear equations mod 2, and ν is primarily supported on unsatisfiable such systems, it is possible to distinguish these distributions via Gaussian elimination. This becomes impossible when considering $T_\delta \mu$ versus ν instead.

Definition 2.2.2 (S_n-symmetry)**.** Let Ω be a finite set or $\mathbb{R}$, and k be a fixed integer. Let $N = \binom{n}{k}$. We say a distribution μ on Ω is S_n-invariant if for every $\pi \in S_n$ and $x \in \Omega^N$ we have $\mathbb{P}_\mu(x) = \mathbb{P}_\mu(\pi \cdot x)$, where π acts by permuting coordinates.

In the special case $k = 2$ and $\Omega = \{0, 1\}$, we may think of Ω^N as the space of n-node undirected graphs. S_n-symmetry of a distribution μ on Ω^N amounts to exchangability of μ as a random graph model.

Now we state the main definition needed for the conjecture: a slight strengthening of the assumption in Hypothesis 2.1.5.

Definition 2.2.3 (Bounded almost independence)**.** Let Ω be a finite set or $\mathbb{R}$, and let k be a fixed integer. Let $N = \binom{n}{k}$. Let ν be a product distribution on Ω^N. Let μ be another distribution on Ω^N. We say μ is D-wise almost independent (with respect to ν) if for every D-simple statistic f, we have $\mathbb{E}_{x \sim \mu} f(x) = O(1)$.

If $\mathbb{E}_{x\sim\mu} f(x) = O(1)$ is replaced with $\mathbb{E}_{x\sim\mu} f(x) = 0$ we would recover the usual definition of D-wise independence.

Finally, we state the main conjecture of this section. Like most complexity-theoretic conjectures we do not expect to prove it, as it is much stronger than, say, $P \neq NP$, but we may hope to amass evidence in its favor.

Conjecture 2.2.4 (Super-logarithmic almost independence fools polynomial time)**.** *Let Ω be a finite set or $\mathbb{R}$, and let k be a fixed integer. Let $N = \binom{n}{k}$. Let ν be a product distribution on Ω^N. Let μ be another distribution on Ω^N. Suppose that μ is S_n-invariant and $(\log n)^{1+\Omega(1)}$-wise almost independent with respect to ν. Then no polynomial-time computable test distinguishes $T_\delta \mu$ and ν with probability $1 - o(1)$, for any $\delta > 0$. Formally, for all $\delta > 0$ and every polynomial-time computable $t : \Omega^N \to \{0, 1\}$ there exits $\delta' > 0$ such that for every large enough n,*

$$\frac{1}{2} \mathop{\mathbb{P}}_{x\sim\nu}(t(x) = 0) + \frac{1}{2} \mathop{\mathbb{P}}_{x\sim\mu}(t(x) = 1) \leqslant 1 - \delta' .$$

If $\Omega = \mathbb{R}$ and ν is standard Gaussian in every coordinate, we also conjecture that T_δ can be replaced with the usual Ornstein-Uhlenbeck noise operator U_δ (which applies a small amount of noise to every coordinate, rather than resampling a small fraction of coordinates).

We assume S_n-invariance in Conjecture 2.2.4 primarily because all of the canonical inference problems we study to support special cases of Conjecture 2.2.4 are S_n-invariant. That said, we are not aware of any counterexample to the conjecture when the S_n-invariance condition is dropped.

For simplicity, in Conjecture 2.2.4 we have restricted attention to the polynomial-time regime. However, there are counterparts for other running times. For example, an analogous conjecture saying that (n^c)-wise almost-independence

fools $2^{n^{o(c)}}$-time algorithms and applying the result to random constraint satisfaction (CSP) would correctly predict lower bounds against subexponential-time SoS algorithms for random CSPs (for which matching algorithms are known) [81, 153, 107]. Together with Theorem 1.1.5 on simple statistics for the k-community stochastic block model, Conjecture 2.2.4 implies Theorem 1.1.3 on the information-computation gap of the block model.

2.3 The Low-Degree Likelihood Ratio

We turn to the question: given a hypothesis testing problem (H_0, H_1), how does one determine whether there is successful D-simple statistic? We will see that this amounts to same question as: how close to a low-coordinate-degree function is the likelihood ratio?

Fix a pair of distributions ν, μ on Ω^n and keep in mind the hypothesis testing problem where ν is the null distribution and μ is the alternative. Any distribution ν on Ω^n induces an inner product of functions $f, g : \Omega^n \to \mathbb{R}$ given by $\langle f, g \rangle_\nu = \mathbb{E}_{x \sim \nu} f(x)g(x)$. We define $f^{\leqslant_\nu D}$ to be the projection of a function f to the span of coordinate-degree-D functions, where the projection is orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle_\nu$. When clear from context, we often drop the subscript ν.

Our main theorem in this section shows that to decide whether or not any D-simple statistic f achieves $\mathbb{E}_{x \sim \mu} f(x) \to \infty$ it suffices to check one canonical statistic: the low-degree likelihood ratio.

Theorem 2.3.1 (Optimality of low-degree likelihood ratio)**.** *Let $LR(x) : \Omega^n \to \mathbb{R}$ (which stands for **l**ikelihood **r**atio) be given by $LR(x) = \frac{\mathbb{P}_\mu(x)}{\mathbb{P}_\nu(x)}$. For every D, the optimal*

D-simple statistic is the centered, low-degree likelihood ratio $LR^{\leqslant D} - 1$. *Formally,*

$$\arg \max_{f \text{ } D\text{-simple}} \mathop{\mathbb{E}}_{x \sim \nu} f(x) = LR^{\leqslant D} - 1$$

and

$$\max_{f \text{ } D\text{-simple}} \mathop{\mathbb{E}}_{\mu} f(x) = \langle LR^{\leqslant D} - 1, LR^{\leqslant D} - 1 \rangle^{1/2} = \|LR^{\leqslant D} - 1\| .$$

The centering $LR(x) - 1$ comes from the simple observation that $\mathbb{E}_\nu LR(x) = \sum_x \mathbb{P}_\mu(x) = 1$.

This theorem is an appealing counterpart to the Neyman-Pearson lemma for D-simple statistics: the Neyman-Pearson lemma says that the optimal hypothesis test is given by thresholding the value of $LR(x)$, and Theorem 2.3.1 shows that the projections of the likelihood ratio are optimal D-simple statistics.

The proof of Theorem 2.3.1 amounts to some simple linear algebra, using the definition of orthogonal projection and the observation that for any function f,

$$\mathop{\mathbb{E}}_{x \sim \mu} f(x) = \mathop{\mathbb{E}}_{x \sim \nu} LR(x) f(x) = \langle LR(x), f(x) \rangle_\nu .$$

We leave the details to the reader.

Theorem 2.3.1 offers an avenue to determine, for a concrete pair of distributions ν, μ and particular D, whether or not there is a successful D-simple statistic. Fix some D, and suppose that $f_0, \ldots, f_m : \Omega^n \to \mathbb{R}$ are an orthonormal basis for the coordinate-degree D functions (with respect to $\langle \cdot, \cdot \rangle_\nu$), and that $f_0(x) = 1$ is the constant function. That is, $\langle f_i, f_j \rangle_\nu = \mathbb{E}_{x \sim \nu} f_i(x) f_j(x) = \delta_{ij}$. Then by measuring the norm of $LR^{\leqslant D} - 1$ in this basis, we find

$$\|LR^{\leqslant D} - 1\|^2 = \sum_{1 \leqslant i \leqslant m} \langle f_i, LR^{\leqslant D} - 1 \rangle^2 .$$

By definition,

$$\langle f_i, LR^{\leqslant D} - 1 \rangle = \mathop{\mathbb{E}}_{x \sim \nu} (LR^{\leqslant D}(x) - 1) \cdot f_i(x) = \mathop{\mathbb{E}}_{x \sim \nu} LR(x) f_i(x)$$

because $LR - (LR^{\leqslant D} - 1)$ is orthogonal to f_i by hypothesis. Again by definition, $\mathbb{E}_{x \sim \nu} LR(x) f_i(x) = \mathbb{E}_{x \sim \mu} f_i(x)$. Together with Theorem 2.3.1, we arrive at:

$$\max_{f \ D\text{-simple}} \mathop{\mathbb{E}}_{\mu} f = \left(\sum_{1 \leqslant i \leqslant m} (\mathop{\mathbb{E}}_{\mu} f_i)^2 \right)^{1/2}. \tag{2.3.1}$$

By (2.3.1), to decide whether or not there is a successful D-simple statistic for the pair (ν, μ), one only needs to be able to compute $\mathbb{E}_{\mu} f_i$ for some orthonormal basis functions f_i, themselves depending only on the null distribution ν. For a variety of null distributions, in particular for product distributions, such functions f_i are simply Fourier bases, which can be constructed by well-known tools [142].

2.4 Example: Planted Clique

To see all these ideas in action, we will work out an example. Recall the *planted clique* problem, involving an n-node graph G:

> H_0 : G was sampled from the Erdős-Rényi model $G(n, \frac{1}{2})$, with independent edges each appearing with probability $\frac{1}{2}$.
>
> H_1 : G was sampled adding each vertex of G to a set S independently with probability k/n, adding every edge between vertices in S to form a clique, then sampling the rest of the edges independently as in $G(n, \frac{1}{2})$.

Denote by ν the null distribution $G(n, 1/2)$ and μ_k the alternative distribution H_1. In this section we prove the following lemma about optimal D-simple

statistics for planted clique. A refined version of this lemma is the first step in a (much more complicated) SoS lower bound for the planted clique problem, presented later in this thesis.

Lemma 2.4.1. *For every $\varepsilon > 0$, if $k = n^{1/2-\varepsilon}$, then*

$$\max_{f \text{ } (C \log n)\text{-simple}} \mathop{\mathbb{E}}_{G \sim \mu_k} f(G) \leqslant O(1)$$

for every $C > 0$. On the other hand, if $k \geqslant 1.01\sqrt{n}$, then there is $C > 0$ such that

$$\max_{f \text{ } (C \log n)\text{-simple}} \mathop{\mathbb{E}}_{G \sim \mu_k} f(G) \to \infty .$$

More refined versions of this lemma allow $1.01\sqrt{n}$ to be relaxed to $\Omega(\sqrt{n})$, and to treat k in the interval $[n^{1/2-\varepsilon}, \Omega(\sqrt{n})]$ (though some questions about what precisely happens in this interval remain open for such k).

Proof. From Boolean Fourier analysis [142] we recall that the functions $\{\chi_\alpha(G) = \prod_{\{i,j\} \in \alpha} (2G_{ij} - 1)\}_{\alpha \subseteq \binom{n}{2}, |\alpha| \leqslant D}$ form an orthonormal basis for the degree-D functions $f : \{0,1\}^{\binom{n}{2}} \to \mathbb{R}$. (Here G_{ij} is the 0/1 indicator for the presence of the edge ij in G.) So by (2.3.1), we just need to compute $\mathbb{E}_{G \sim \mu} \chi_\alpha(G)$ for each such α.

Fix $\alpha \subseteq \binom{n}{2}$. Consider the process of sampling a graph $G \sim \mu$ by first sampling the clique vertices $S \subseteq [n]$. Conditioned on S the edges of G become independent, so $\mathbb{E}\, \chi_\alpha(G) = \mathbb{E}_S \prod_{ij \in \alpha} \mathbb{E}[(2G_{ij} - 1) \mid S]$. If i or j is not in S, then the edge i, j is included in G with probability $1/2$, so $\mathbb{E}[(2G_{ij} - 1) \mid \{i, j\} \not\subset S] = 0$.

Thus, χ_α has nonzero conditional expectation only if all of $V(\alpha) \stackrel{\text{def}}{=} \{i \in [n] : \text{ exists } j \in [n] \text{ s.t. } ij \in \alpha\}$ is in S. This occurs with probability precisely $(k/n)^{|V(\alpha)|}$. And, if $V(\alpha) \subseteq S$, every edge with endpoints in $V(\alpha)$ appears in G, so the conditional expectation of $\chi_\alpha(G)$ is 1. We find that $\mathbb{E}_\mu \chi_\alpha(G) = (k/n)^{|V(\alpha)|}$.

Now we need to estimate $\sum_{0<|\alpha|\leqslant D}(\mathbb{E}_\mu\,\chi_\alpha(G))^2 = \sum_{|\alpha|\leqslant D}(k/n)^{2|V(\alpha)|}$. We start with the upper bound, when $k = n^{1/2-\varepsilon}$ for some $\varepsilon > 0$ and $D = C\log n$ for some $C > 0$. Every α with $|\alpha| \leqslant D$ has $|V(\alpha)| \leqslant 2D = 2C\log n$. And, for every $t \leqslant 2C\log n$ there are at most $n^t t^{\min(2C\log n, 2t^2)}$ sets α with $|V(\alpha)| = t$ and $|\alpha| \leqslant C\log n$. So,

$$\sum_{0<|\alpha|\leqslant D}(\mathop{\mathbb{E}}_{\mu}\chi_\alpha(G))^2 \leqslant \sum_{t\leqslant\sqrt{C\log n}} n^{-2\varepsilon t}\cdot t^{2t^2} + \sum_{\sqrt{C\log n}\leqslant t\leqslant 2C\log n} n^{-\varepsilon t}\cdot t^{C\log n}\,.$$

Standard manipulations bound both the above sums by $O(1)$.

On the other hand, if $k > 1.01\sqrt{n}$, then just by considering the contributions to the sum from α's which form a cycle it is easy to show that $\sum_{|\alpha|\leqslant 100\log n}(\mathbb{E}_\mu\,\chi_\alpha(G))^2 \to \infty$. □

CHAPTER 3

THE SOS METHOD FOR ALGORITHM DESIGN

The next goal is to describe the SoS method. As a prerequisite, we introduce two statistical tasks generalizing hypothesis testing: *refutation* and *estimation*.

We need to extend the setting from the last chapter to include *hidden variables*. Let Ω, Σ be (finite or infinite) sets, and $n, m \in \mathbb{N}$. Let μ be a probability distribution on $\Omega^n \times \Sigma^m$, and ν a distribution on Ω^n. We think of μ as a distribution on pairs $\{y, x\}$ where $y \in \Omega^n$ and $x \in \Sigma^m$. If we project μ onto the marginal distribution on y we recover the hypothesis testing settings of the last chapter. Instead, we will consider algorithmic tasks in which an algorithm sees a sample y and accomplishes some task related to x.

3.1 Refutation

Definition 3.1.1. An α-refutation algorithm for (μ, ν) takes input $y \in \Omega^n$ and outputs a number $A(y)$ such that $A(y) \geqslant \max_{x \in \Sigma^m} \log \mathbb{P}_\mu(y, x)$ and $\mathbb{P}_{y \sim \nu}(A(y) > \alpha) \leqslant o(1)$. Notice that the second probability is over $y \sim \nu$, even though $A(y)$ is an upper bound on probabilities related to μ. Informally, and for the right choices of α, the algorithm A is certifying that typical $y \sim \nu$ is extremely unlikely to have come from μ.

Often, $\max_{x \in \Sigma^m} \log \mathbb{P}_\mu(y, x)$ corresponds to a natural combinatorial or analytic property of y, and the refutation problem requires certifying that y does not have some combinatorial or analytic structure. Some examples may be helpful.

Example 3.1.2 (Planted k-clique, refutation version)**.** Recall the planted k-clique

alternative distribution μ on n-node graphs G. First include each vertex independently in a set S with probability k/n, then sample a random graph with a clique on S and the rest of the edges independent as in $G(n, 1/2)$. Here, the hidden variable $x \in \{0,1\}^n$ is the indicator vector for S, and the observed variable $y \in \{0,1\}^{\binom{n}{2}}$ is the (adjacency matrix of the) graph G. We have

$$\mathop{\mathbb{P}}_{\mu}(G, S) = \mathop{\mathbb{P}}_{\mu}(S) \cdot \mathop{\mathbb{P}}_{\mu}(G \mid S) = \left(\frac{k}{n}\right)^{|S|} \left(1 - \frac{k}{n}\right)^{n-|S|} \left(\frac{1}{2}\right)^{\binom{n}{2} - \binom{|S|}{2}} .$$

if S is a G-clique and 0 otherwise. So when S is a G-clique,

$$\log \mathop{\mathbb{P}}_{\mu}(G, S) = \binom{|S|}{2} \log 2 - |S| \log\left(\frac{n}{k}\right) - |S| \log(1 - k/n) + f(n, k)$$

for some function $f(n, k)$ not depending on S, and otherwise $\log \mathbb{P}_{\mu}(G, X) = -\infty$. For $|S| \gg \log n$ and some constant C,

$$|S|^2/C + f(n, k) \leqslant \log \mathop{\mathbb{P}}_{\mu}(G, S) \leqslant C|S|^2 + f(n, k) ,$$

so the refutation problem is (for such $|S|$) equivalent to certifying upper bounds on the size of the maximum clique in $G \sim G(n, \frac{1}{2})$ (the latter appears because $G(n, \frac{1}{2})$ is typically the null distribution for planted clique).

Relation to Hypothesis Testing A refutation algorithm for distributions ν, μ can also be used to solve the hypothesis testing problem, assuming that there is an efficient algorithm to compute $\mathbb{P}_{\nu}(y)$. Often ν – the distribution of the null hypothesis – is a product distribution, or is uniform over some set of known size, making this task trivial.

In the case of planted clique this connection is intuitively clear. Graphs from μ (typically) contain cliques of size $k - O(\sqrt{k})$ and graphs from $G(n, 1/2)$ typically contain no clique larger than $2.1 \cdot \log n$. Any refutation algorithm which

successfully certifies that graphs from $G(n, 1/2)$ do not contain cliques of size $0.9k$ also indicates by its success or failure which distribution its input came from.

It is possible to derive a reduction from hypothesis testing to refutation in an very generic setting (at least for finite Ω and Σ), via the following familiar variational formula from the theory of exponential families [177]:

$$\log \mathop{\mathbb{P}}_{\mu}(y) = \log \sum_{x \in \Sigma^m} \mathop{\mathbb{P}}_{\mu}(x, y) = \max_{\mu' \in \Delta_{\Sigma^m}} \mathop{\mathbb{E}}_{x \sim \mu'} \log \mathop{\mathbb{P}}_{\mu}(x, y) + H(\mu')$$

where Δ_{Σ^m} is the set of distributions on Σ^m and H is the Shannon entropy. For any pair of distributions μ, ν, the associated refutation problem requires certifying an upper bound on $\max_x \log \mathbb{P}_\mu(x, y)$ given $y \sim \nu$.[1]

3.2 Estimation

Once again, let $\mu = \{x, y\}$ be a joint probability distribution on $\Sigma^m \times \Omega^n$. Let $\ell : \Sigma^m \times \Sigma^m \to [0, 1]$ be a *loss function*.

Definition 3.2.1. An α-estimation algorithm for the pair (μ, ℓ) takes input $y \in \Omega^n$ and returns $\widehat{x} \in \Sigma^m$ such that $\mathbb{E}_x[\ell(x, \widehat{x}) \,|\, y] \leqslant \alpha$ ..

As a simple example we can return to planted clique, where the natural loss function on a pair of subsets $x, \widehat{x} \subseteq [n]$ is the 0/1 loss: $\ell(x, \widehat{x}) = 0$ if $x = x'$ and otherwise $\ell(x, \widehat{x}) = 1$. An α-estimation algorithm for planted clique returns a subset of vertices, and by Markov's inequality it finds the correct planted clique with probability at least $1 - \alpha$.

[1] This reduction is loose only for hypothesis testing problems where the entropy term $H(\mu')$ and the likelihood term $\mathbb{P}_\mu(x, y)$ are of comparable magnitues, and so $\log \mathbb{P}_\mu(y)$ is not well approximated by $\max_x \log \mathbb{P}_\mu(x, y)$. Such situations do arise – for example in the sparse stochastic block model, where refutation and hypothesis testing are less closely related than for most other problems in this thesis.

Relation to Hypothesis Testing Like refutation, estimation is typically more difficult than hypothesis testing. In hypothesis testing, the goal is to test the y-marginal of $\mu = \{x, y\}$ against a null distribution ν on Ω^n. Generally, the corresponding estimation problem for μ involves a loss function ℓ such that with high probability over $y \sim \mu$ any $\widehat{x}$ achieving small loss is also a polynomial-time-verifiable witness to the inequality $\mathbb{P}_\mu(y) \gg \mathbb{P}_\nu(y)$.

Again, it is useful to keep examples in mind: in the planted clique problem, a successful estimation algorithm recovers a clique of size much more than $2 \log n$ from graphs from the alternative distribution. Any y-clique x of super-logarithmic size provides a strong certifiable lower bound on the alternative probability $\mathbb{P}_{H_1}(y) \geqslant \mathbb{P}_{H_1}(y, x)$.

3.3 SoS Proofs and Refutation

We start by introducing SoS as a method to design refutation algorithms. As in the last section, suppose μ is a probability distribution on $\Omega^n \times \Sigma^m$. We make three additional mild assumptions:

- $\Sigma \subseteq \mathbb{R}$ is a low-degree algebraic set. That is, there are some low-degree polynomials $p_1, \ldots, p_k \in \mathbb{R}[x]$ such that $\Sigma = \{x \in \mathbb{R} : p_i(x) = 0\}$. For concreteness, $\deg p_i \leqslant d$ for some $d \in \mathbb{N}$.
- For every $y \in \Omega^n$, the set $\{x \in \Sigma^m : \mathbb{P}_\mu(y, x) > 0\}$ is also a low degree algebraic set. That is, there are some polynomials $p_1, \ldots, p_k \in \mathbb{R}[x_1, \ldots, x_n]$ such that

$$\{x \in \Sigma^m : \mathop{\mathbb{P}}_{\mu}(y, x) > 0\} = \{x \in \Sigma^m : p_1(x) = 0, \ldots, p_k(x) = 0\}.$$

and $\deg p_i(x) \leqslant d$.

- For x such that $\mathbb{P}_\mu(y, x) > 0$, the value $\log \mathbb{P}_\mu(y, x)$ can be computed by a low degree polynomial. Formally, there is a polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ such that $\deg p \leqslant d$ and for all $x \in \Sigma^m$ such that $\log \mathbb{P}_\mu(y, x)$ is finite, $p(x) = \log \mathbb{P}_\mu(y, x)$.

Most of the inference problems in this thesis fit these assumptions. Using planted clique as our example, we check that:

- The hidden variables lie in $\{0, 1\}^n$, which is defined by the degree $d = 2$ polynomial equations $x_i^2 - x_i = 0$.
- $\mathbb{P}_\mu(y, x)$ is only nonzero if x is (the indicator vector of) a y-clique. The set of y-cliques is defined by the degree-2 equations $\{x \in \{0, 1\}^n : x_i x_j = 0 \text{ if } i \not\sim j \text{ in } y\}$.
- As we saw in Example 3.1.2, when x is the indicator vector of a y-clique, $\log \mathbb{P}_\mu(y, x)$ is a degree 2 polynomial in the number of nonzero entries in x, which is itself the degree one polynomial $\sum_{i \in [n]} x_i$.

Under these assumptions, we can describe a canonical SoS refutation algorithm. The starting point is the following polynomial optimization problem. Suppose $p_1, \ldots, p_k$ is a collection of degree d polynomials defining the set $\{x \in \Sigma^m : \log \mathbb{P}_\mu(y, x) \text{ is finite}\}$. Let $p(x) = \log \mathbb{P}_\mu(y, x)$ for x for which this quantity is finite, and assume $\deg p \leqslant d$. The refutation problem for $y \in \Omega^n$ is precisely the problem

$$\max p(x) \text{ such that } p_1(x) = 0, \ldots, p_k(x) = 0\,. \tag{3.3.1}$$

An SoS proof is a short certificate of an upper bound on the value of (3.3.1). An SoS proof of degree d that (3.3.1) $\leqslant c$ is a collection of polynomials $q_1, \ldots, q_k \in \mathbb{R}[x_1, \ldots, x_m]$ such that $\deg q_i \cdot p_i \leqslant d$ for all $i \leqslant k$, and collections of polynomials $s_1, \ldots, s_r, s'_1, \ldots, s'_r \in \mathbb{R}[x_1, \ldots, x_m]$, such that $\deg s_i \leqslant d/2$ and $\deg s'_i(x)^2 p(x) \leqslant d$, with the property that

$$-1 = \sum_{i \leqslant k} p_i(x) \cdot q_i(x) + \sum_{i \leqslant r} s_i(x)^2 + (p(x) - c) \sum_{i \leqslant r} s'_i(x)^2 \tag{3.3.2}$$

If such polynomials exist, then clearly $p(x) \leqslant c$ for all $x \in \Sigma^m$, since for every such x the first two terms on the right-hand side are nonnegative, so if the last term were nonnegative for any x there would be a contradiction. Furthermore, it is algorithmically easy to verify that (3.3.2) holds for given q_i, s_i, s'_i (in $(km)^{O(d)}$ time), by expanding all the polynomials in a fixed basis (the monomial basis, for example).

The Positivstellensatz of Krivine and Stengle says that for every $p_1, \ldots, p_m, p, c$, there exists x such that $p_i(x) = 0$ but $p(x) > c$, or there is an SoS proof that $p(x) \leqslant c$ if $p_i(x) = 0$ for all i [110, 167]. This result says nothing, however, about the degree of that SoS proof. For most applications in theoretical computer science and combinatorial optimization, degree-n proofs capture the power of all SoS proofs for an n-variable problem; in particular this is true in the case of polynomial optimization over the n-dimensional boolean hypercube.

We will be most interested in proofs whose degree d is constant compared to the number of variables x_i and constraints p_i. SoS proofs are useful for algorithm design because *there is an efficient algorithm to find an SoS proof $q_1, \ldots, q_k, s_1, \ldots, s_r$, if it exists.*[2] Using semidefinite programming, given $c, p, p_1, \ldots, p_k$ it is possible to decide in time $(km)^{O(d)}$ whether there is a degree-d SoS proof which certifies

[2]Strictly speaking, this is true only up to small numerical errors, and only under mild niceness

$p(x) \leqslant c$, for x as in (3.3.2). We typically think of our refutation problem instances as having size $m^{\Theta(1)}$, so if $k \leqslant m^{O(1)}$ and $d \leqslant O(1)$ this kind of algorithm runs in polynomial time.

We defer a (more) formal discussion of SoS proofs and algorithms for finding them to Section 3.5 and Chapter 5, and instead turn to an example.

Example 3.3.1 ($O(\sqrt{n})$ refutation for planted clique using SoS)**.** As we discussed in Example 3.1.2, the refutation problem for planted clique is equivalent (up to small factors) to certifying upper bounds on the size of the maximum clique in a graph $G \sim G(n, \frac{1}{2})$. If we work out the polynomial program (as in (3.3.1)) for this task, we find it takes the form

$$\max \left(\sum_{i \in [n]} x_i \right)^2 \text{ such that } x_i^2 - x_i = 0 \text{ and } x_i x_j = 0 \text{ if } i \not\sim j \,. \tag{3.3.3}$$

We will construct a constant-degree SoS proof that this maximum is at most $O(\sqrt{n})$. A tighter analysis is possible, showing that the maximum is at most $\sqrt{n}/2^d$ for degree $O(d)$ SoS [71].

Our main tool is the classic fact from random matrix theory that with high probability over $G \sim G(n, 1/2)$, the adjacency matrix A of G satisfies $(2\sqrt{n} - C)\mathrm{Id} + \frac{1}{2}J \succeq A$ for any constant C and large-enough n, where Id is the $n \times n$ identity matrix, J is the $n \times n$ matrix of all-1s, and the relation $\succeq$ denotes that the left-hand side minus the right-hand side is a positive semidefinite (PSD) matrix. Since any PSD matrix $M \succeq 0$ can be written as $M = \sum_{i \leqslant r} v_i v_i^\top$ for some vectors $v_1, \ldots, v_r$ (here $r \leqslant n$ is the rank of M), we find that there are vectors v_i such that

$$(2\sqrt{n} - C)\mathrm{Id} + \frac{1}{2}J - A = \sum_{i \leqslant r} v_i v_i^\top$$

conditions on $p, p_1, \ldots, p_k$. This kind of niceness will always be satisfied in this thesis. See [143, 156] for details.

and hence if we consider the polynomial $\langle x, ((2\sqrt{n} - C)\mathrm{Id} + \frac{1}{2}J - A)x\rangle$, there are polynomials $v_i(x) = \langle v_i, x\rangle$ such that

$$\langle x, ((2\sqrt{n} - C)\mathrm{Id} + \frac{1}{2}J - A)x\rangle = \sum_{i \leqslant r} v_i(x)^2 .$$

To turn these polynomials into an SoS proof takes just a few more manipulations. As an aside, we do not wish to give the impression that constructing SoS proofs is at all magical: one of main appeals of the SoS paradigm is that SoS proofs can usually be constructed in a modular and composable way, mimicking the usual proposition/lemma/theorem structure of mathematical proofs in general. The best examples in this thesis of the latter are in Section 3.5 and Chapter 9. Here, for brevity and because the SoS proof is fairly simple, we opt for a one-shot style of construction.

We start by noting that $(\sum_i x_i)^2$ can be expanded as

$$\left(\sum_i x_i\right)^2 = \sum_{ij} x_i x_j = \sum_{i \sim j} x_i x_j + \sum_{i \nsim j} x_i x_i + \sum_i x_i^2 .$$

Next, we use that $\langle x, Ax\rangle = \sum_{i\sim j} x_i x_j = (2\sqrt{n} - 10)\sum_i x_i^2 + \frac{1}{2}\sum_{ij} x_i x_j - \sum_{i \leqslant r} v_i(x)^2$ for some linear functions $v_i(x)$ to obtain that

$$\left(\sum_i x_i\right)^2 = (2\sqrt{n} - 9)\sum_i x_i^2 + \frac{1}{2}\sum_{ij} x_i x_j - \sum_{i \leqslant r} v_i(x)^2 + \sum_{i \nsim j} x_i x_i$$

which rearranges to

$$\frac{1}{2}\left(\sum_i x_i\right)^2 = (2\sqrt{n} - 9)\sum_i x_i^2 - \sum_{i \leqslant r} v_i(x)^2 + \sum_{i \nsim j} x_i x_i .$$

Thus,

$$\left(\frac{1}{2}\sum_i x_i + 1\right)\left(\sum_i x_i - 4\sqrt{n}\right) = \frac{1}{2}\left(\sum_i x_i\right)^2 - (2\sqrt{n} - 1)\sum_i x_i - 4\sqrt{n}$$

$$
\begin{aligned}
&= (2\sqrt{n}-9)\sum_i x_i^2 - \sum_{i\leqslant r} v_i(x)^2 + \sum_{i\neq j} x_i x_i \\
&\quad - (2\sqrt{n}-1)\sum_i x_i - 4\sqrt{n} \\
&= -8\sum_i x_i^2 - \sum_{i\leqslant r} v_i(x)^2 + \sum_{i\neq j} x_i x_j - 4\sqrt{n} + r(x)
\end{aligned}
$$

for some polynomial $r(x) = \sum r_i(x)(x_i^2 - x_i)$ with $\deg r_i(x) \leqslant O(1)$. This rearranges to

$$
-1 = \frac{1}{4\sqrt{n}}\left[\left(\frac{1}{2}\sum_i x_i^2 + 1\right)\left(\sum_i x_i - 4\sqrt{n}\right) + r'(x) + \sum_{i\sim j} x_i x_j - s(x)\right]
$$

for $s(x) = \sum s_i(x)^2$ a constant-degree sum-of-squares polynomial and $r'(x) = \sum r_i'(x)(x_i^2 - x_i)$ another constant-degree polynomial. This is an SoS proof that (3.3.3) has maximum value at most $4\sqrt{n}$.

Relation to simple statistics Much of this thesis, particularly Part II, is concerned with the relationship between SoS proofs and simple statistics. We are not prepared here to explore this relationship deeply, but we can already note one key theme: *the polynomials in the SoS proof we produced for planted clique either have coefficients which are themselves low-degree in (the adjacency matrix of) the instance G, or come from a PSD factorization of a matrix whose entries are such low-degree polynomials.*

This kind of SoS proof will surface again and again. Eventually, when we prove Theorem 12.1.5 in Chapter 12, we begin to unravel why this kind of SoS proof composed of simple statistics is so universal. Many interesting questions about just how universal this sort of SoS proof is remain open.

More examples For another simple SoS refutation algorithm, see Chapter 6. Numerous refutation algorithms in the literature can be phrased as SoS refutation algorithms – indeed, our planted clique example above was originally a spectral algorithm due to [9] – and this is crucial to the ability of the SoS method to unify existing approaches to algorithm design in statistical inference. A partial list of examples includes matrix completion [157], random CSPs [7, 153], decomposition of random overcomplete tensors [77], sparse PCA [62], and many more.

3.4 Pseudodistributions and Estimation

Next we discuss design of estimation algorithms using the SoS method. For this purpose, we need to introduce *pseudodistributions*, which are dual objects to SoS proofs.

Let $\mathbb{R}[x_1, \ldots, x_n]_{\leqslant d}$ be the polynomials of degree at most d in variables $x_1, \ldots, x_n$, with real cofficients. A degree-d pseudodistribution $\tilde{\mathbb{E}}$ on variables $x_1, \ldots, x_n$ is a linear map from to $\mathbb{R}$ which is *normalized* and *positive semidefinite*. Formally, $\tilde{\mathbb{E}}[1] = 1$, where on the left side we write 1 for the polynomial which takes the constant value 1, and $\tilde{\mathbb{E}}\, p(x)^2 \geqslant 0$ for every $p \in \mathbb{R}[x_1, \ldots, x_n]_{\leqslant d/2}$. We sometimes use the word *pseudoexpectation* instead of pseudodistribution; for us these are interchangeable.

We say a pseudodistribution satisfies a constraint $p = 0$ if for every polynomial q such that $\deg(p \cdot q) \leqslant d$ we have $\tilde{\mathbb{E}}\, q \cdot q = 0$. A pseudodistribution satisfies an inequality $p \geqslant 0$ if for every polynomial $s(x) = \sum s_i(x)^2$ which is a sum of squares with $\deg s \cdot p \leqslant d$ it holds that $\tilde{\mathbb{E}}\, p \cdot s \geqslant 0$. (We defer to Chapter 5 the meaning of a pseudodistribution satisfying a set of inequalities $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$.)

Pseudodistributions are dual to SoS proofs, in the following sense: for polynomials $p_1, \ldots, p_m, p$ and $c \in \mathbb{R}$, if there is a pseudodistribution satisfying $p_i(x) = 0, p(x) - c \geqslant 0$, then there is no SoS proof of $p(x) \leqslant c$ subject to $p_i(x) = 0$. To see this, just apply the pseudodistribution to the left and right-hand sides of any putative SoS proof:

$$\tilde{\mathbb{E}}[-1] = \tilde{\mathbb{E}}[s(x)(p(x) - c)] + \sum_i \tilde{\mathbb{E}}[r_i(x)p_i(x)] + \sum \tilde{\mathbb{E}}[s_i(x)^2] \,.$$

By linearity, $\tilde{\mathbb{E}}[-1] = -\tilde{\mathbb{E}}[1] = -1$, but by PSD-ness, the right-hand side is non-negative, which is a contradiction. In fact, under mild conditions on $p, p_1, \ldots, p_m$ there is a strong duality between pseudodistributions and SoS proofs: for technical details we refer the reader to [34].

Most importantly, as in the case of SoS proofs, under mild conditions on $p, p_1, \ldots, p_m$, there is an $(nm)^{O(d)}$-time algorithm to find a degree-d pseudodistribution satisfying $p(x) - c \geqslant 0, p_i(x) = 0$, if it exists. (Indeed, since pseudodistributions and SoS proofs are convex duals, there is an algorithm which either returns an SoS refutation or a pseudodistribution.) Again we defer further discussion of this algorithm to Chapter 5, and turn to the matter of designing SoS-based estimation algorithms.

Returning to the optimization formulation (3.3.1) from the last section, for now we are going to consider estimation problems where small loss ℓ is achieved by any x which is even approximately optimal for (3.3.1). Formally, these are problems where maximum likelihood estimation is a good strategy for achieving small loss – in Section 3.5 and Chapters 8 and 9 we will discuss ways to use the SoS method to design estimation algorithms in more complicated settings where maximum likelihood may not be the right approach.

The simplest SoS algorithms for estimation under a distribution μ fit the following mold. Given a problem instance $y \in \Omega^n$, using semidefinite programming, find the pseudodistribution in variables $x_1, \ldots, x_n$ which maximizes $\tilde{\mathbb{E}} \log_\mu \mathbb{P}(y, x)$ (remember that throughout we are assuming that $\log \mu \mathbb{P}(y, x)$ becomes a polynomial, under some low-degree polynomial constraints of the form $p_i(x) = 0$). Then, *round* that pseudodistribution to obtain an estimator $\widehat{x}$ for x.

Designing a rounding scheme sometimes requires creative algorithm design – indeed rounding algorithms for semidefinite programming even in simple settings like constraint satisfaction has been the subject of much study [78, 33, 154]. For the inference problems we consider in this thesis, however, there are usually straightforward rounding schemes. (One exception to this theme is when we need to estimate several exchangeable hidden variables at once – when the posterior distribution of x given y contains too much symmetry rounding can become more difficult. We develop some general-purpose tools for rounding in this situation in Chapter 8.)

For concreteness, we continue our planted clique example from previous sections.

Example 3.4.1 (Finding an $O(\sqrt{n})$-sized planted clique)**.** Suppose μ is the alternative/planted distribution for planted k-clique: remember that this is a distribution on n-node graphs G with planted k-cliques S. Given a graph G, our estimation algorithm is:

1. Find the degree-$O(1)$ pseudodistribution on variables $x_1, \ldots, x_n$ maximizing $\tilde{\mathbb{E}} \sum_{i \in [n]} x_i$ which satisfies $x_i^2 - x_i = 0$ for all $i \in [n]$ and $x_i x_j = 0$ for all i, j nonadjacent in G.

2. Output $\tilde{\mathbb{E}}\, x$

Our algorithm outputs a vector $\tilde{\mathbb{E}}\, x \in \mathbb{R}^n$ which (we will show) is close in an ℓ_2 sense to the indicator vector of the planted clique in G. Given such a vector, producing the set of vertices forming the clique is straightforward, so we leave out the details.

The key fact we use to analyze this algorithm is that with high probability over G and S,

$$\left(\sum_{i \notin S} x_i\right)^2 = \sum_{i \neq j} p_{ij}(x) x_i x_j + \sum_{i \in [n]} r_i(x)(x_i^2 - x_i) + O(\sqrt{n}) \sum_{i \notin S} x_i - s(x)$$

for some $O(1)$-degree polynomials p_{ij}, r_i, s, where s is a sum of squares. (The proof follows similar manipulations as our refutation example Example 3.3.1, using that the graph $G \setminus S$ is distributed as $G(n - |S|, 1/2)$.)

The other fact we use is a version of *pseudoexpectation Cauchy-Schwarz*, which says that if p is a degree-$d/2$ polynomial and $\tilde{\mathbb{E}}$ is a degree-d pseudoexpectation, then $(\tilde{\mathbb{E}}\, p(x))^2 \leqslant \tilde{\mathbb{E}}\, p(x)^2$. We prove this and other basic facts about pseudoexpectations in Chapter 5.

With these in hand, analyzing our algorithm is now straightforward. Let x_S be the indicator vector of the planted clique in a graph G, and suppose that $k = |S| \geqslant C\sqrt{n}$ for a big enough constant C. Because $\tilde{\mathbb{E}} \sum_i x_i$ is maximal, we know $\tilde{\mathbb{E}} \sum_i x_i \geqslant k$. Also,

$$\left(\tilde{\mathbb{E}} \sum_{i \notin S} x_i\right)^2 \leqslant \tilde{\mathbb{E}} \left(\sum_{i \notin S} x_i\right)^2 \leqslant O(\sqrt{n})\, \tilde{\mathbb{E}} \sum_{i \notin S} x_i$$

using positive semidefinite-ness and the expansion above for $\left(\sum_{i \notin S} x_i\right)^2$. This

rearranges to $\tilde{\mathbb{E}} \sum_{i \notin S} x_i \leqslant O(\sqrt{n})$. Now we can put it together:

$$\langle x_S, \tilde{\mathbb{E}}\, x \rangle = \tilde{\mathbb{E}} \sum_{i \in S} x_i = \tilde{\mathbb{E}} \sum_{i \in [n]} x_i - \tilde{\mathbb{E}} \sum_{i \notin S} x_i \geqslant \tilde{\mathbb{E}} \sum_{i \in [n]} x_i - O(\sqrt{n})\,.$$

On the other hand, $\|x_S\| = \sqrt{k}$, and $\|\tilde{\mathbb{E}}\, x\| = \sqrt{\sum_i (\tilde{\mathbb{E}}\, x_i)^2} \leqslant \sqrt{\sum_i \tilde{\mathbb{E}}\, x_i}$. So all together, x_S and $\tilde{\mathbb{E}}\, x$ have Euclidean correlation close to 1 so long as $k \gg \sqrt{n}$. That is,

$$\frac{\langle x_S, \tilde{\mathbb{E}}\, x \rangle}{\|x_S\| \|\tilde{\mathbb{E}}\, x\|} \geqslant 1 - \frac{O(\sqrt{n})}{k}\,.$$

3.5 Proofs to Algorithms

Maximum-likelihood estimation is not the only way to solve estimation problems. In this section we describe a very flexible and general-purpose method for designing SoS-based estimation algorithms, called *proofs to algorithms*. To illustrate it, we design and analyze an algorithm for a non-Bayesian inference problem, *robust mean estimation*; a more sophisticated version of this algorithm appears in Chapter 9. For us, *non-Bayesian* means that there is an adversary in the picture, who chooses hidden variables and perhaps some observed.

Our setting is estimation as before, except now there is no need for a prior distribution. So, we imagine there are hidden variables, or *parameters* $x \in \Sigma^m$, and observed variables $y \in \Omega^n$, and for every x there is a distribution μ_x on Ω^n. Now the goal is to design an algorithm which for any x, given a sample $y \sim \mu_x$ estimates x with respect to some loss function ℓ.[3]

Proofs-to-algorithms starts with the concept of *statistical identifiability*. In order for a successful estimation algorithm to exist for the estimation problem specified

[3]Traditional notation calls the parameters θ; we have opted here to maintain consistency with earlier chapters.

by the distributions μ_x and loss function ℓ, it must be information-theoretically possible (that is, given infinite computational resources) to successfully recover $\widehat{x}$ with $\ell(x, \widehat{x})$ small given $y \sim \mu_x$, with high probability over y.

That is, there must exist a map $\widehat{x}(y)$ such that for all x, with high probability over $y \sim \mu_x$, $\ell(x, \widehat{x})$ is sufficiently small. This would be violated if, for instance, there exist x_0, x_1 such that $\mu_{x_0} = \mu_{x_1}$ but no x simultaneously makes $\ell(x, x_0)$ and $\ell(x, x_1)$ small. If such a map $\widehat{x}$ does exist, we say that x is (approximately) identifiable from y.

As with the likelihood ratio test and hypothesis testing, proving identifiability for an inference problem μ_x, ℓ usually says nothing about efficient algorithms for estimating x given y. The insight in proofs-to-algorithms is: *if identifiability can be proved in a restricted proof system, then x can be estimated from y by a computationally efficient algorithm*.

Of course, the restricted proof system we have in mind is the low degree SoS proof system. In Chapter 5 we will formalize SoS as a proof system, and in designing our algorithms in Chapter 9 we will actually proceed by designing formal proofs in the system.

For now, though, we hope that the following example clarifies the proofs-to-algorithms idea. When reading it, keep in mind that the plan is to start by proving identifiability in as simple a fashion as possible, then adapt that proof to SoS, and design an algorithm around the resulting SoS proof. The final algorithm still uses semidefinite programming to find a pseudodistribution $\tilde{\mathbb{E}}$ and then rounds it, but unlike in the last section, the convex program it employs is not constructed by relaxing the maximum-likelihood problem.

Example 3.5.1 (Robust mean estimation, simplified version in one dimension)**.** As our example, we study a simplified version of the robust mean estimation problem, treated in detail later in Chapter 9. Let $\mathcal{D}$ be a probability distribution on $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq \mathrm{Id}$. Let $\varepsilon > 0$. Samples $x_1, \dots, x_{(1-\varepsilon)n} \sim \mathcal{D}$ are generated and handed to an adversary, who may add εn arbitrary samples of her own and then scramble the order of the samples. We call the resulting samples $x_1, \dots, x_n$ *ε-corrupted*. The goal is to estimate μ given the ε-corrupted samples.

To illustrate the proofs-to-algorithms method, we treat this problem in dimension $d = 1$ for now. Of course we are eventually interested in the large-d version, since this is where the problem becomes computationally challenging; we discuss this more in Chapter 9. When we do so, we employ a more formal framework to construct an SoS proof of identifiability, but here we trade directness for generality.

The first step is to prove that the mean μ is identifiable from $x_1, \dots, x_n$. The goal is to do so using only simple inequalities, so that the proof will ultimately be captured by low-degree SoS. For this we prove the following lemma.

Lemma 3.5.2. *Suppose $n \to \infty$ and $X_1, \dots, X_n$ are ε-corrupted from a distribution $\mathcal{D}$ with mean μ and covariance $\Sigma \preceq \mathrm{Id}$. With probability at least 0.99, if $S \subseteq [n]$ is any set of size $|S| = (1-\varepsilon)n$ with bounded empirical variance*

$$\mathop{\mathbb{E}}_{i \sim S} (X_i - \mu_S)^2 \leqslant 2$$

where $\mu_S = \mathbb{E}_{i \sim S} X_i$, then $|\mu_S - \mu| \leqslant O(\sqrt{\varepsilon})$. Furthermore, if $T \subseteq [n]$ is the subset of non-adversarial samples then T has bounded empirical variance with probability at least 0.99.

Lemma 3.5.2 implies that μ is identifiable up to $O(\sqrt{\varepsilon})$ error: it says that finding a subset of samples S with bounded empirical variance suffices to estimate μ, and with high probability such a bounded empirical variance subset exists. As an aside, $O(\sqrt{\varepsilon})$ error is not too impressive in this one-dimensional context, but in Chapter 9 we show that this lemma generalizes to any d, with error remaining $O(\sqrt{\varepsilon})$, independent of d. (We also show how to improve the error to $\varepsilon^{1-\delta}$ for any constant $\delta > 0$.)

Proof of Lemma 3.5.2. The fact that the set of non-adversarial samples T has bounded empirical variance follows from standard concentration. Furthermore, for $n \to \infty$, the empirical mean μ_T of the non-tampered samples is close to the ground-truth mean: $|\mu_T - \mu| \leqslant o(1)$.

Let $S \subseteq [n]$ with $|S| \geqslant (1-\varepsilon)n$ be any other set of samples with bounded empirical variance. It will suffice to show that $|\mu_T - \mu_S| \leqslant O(\sqrt{\varepsilon})$.

Since $|S \cap T| \geqslant 1 - 2\varepsilon$, there is a coupling of the random variables X, which is a random draw from S, and X', which is a random draw from T, such that X, X' are equal with probability at least $1 - 2\varepsilon$. We expand $|\mu_S - \mu_T|$ in terms of this coupling and apply two inequalities: Cauchy-Schwarz followed by the triangle inequality.

$$\begin{aligned}
|\mu_S - \mu_T| &= |\,\mathbb{E}\,\mathbf{1}_{X \neq X'} \cdot (X - X')| \\
&\leqslant \left(\mathbb{E}\,\mathbf{1}^2_{X \neq X'}\right)^{1/2} \cdot \left(\mathbb{E}(X - X')^2\right)^{1/2} \\
&\leqslant \sqrt{2\varepsilon} \cdot \left(\mathbb{E}(X - X')^2\right)^{1/2} \\
&\leqslant \sqrt{2\varepsilon} \cdot \left[\left(\mathbb{E}(X - \mu_S)^2\right)^{1/2} + \left(\mathbb{E}(X' - \mu_T)^2\right)^{1/2} + |\mu_S - \mu_T|\right].
\end{aligned}$$

This rearranges to

$$|\mu_S - \mu_T| \leqslant \frac{\sqrt{2\varepsilon}}{1-\sqrt{2\varepsilon}} \cdot \left[\left(\mathbb{E}(X-\mu_S)^2\right)^{1/2} + \left(\mathbb{E}(X'-\mu_T)^2\right)^{1/2} \right] \leqslant \frac{2\sqrt{2\varepsilon}}{1-\sqrt{2\varepsilon}} \leqslant O(\sqrt{\varepsilon})$$

using bounded variance. □

Next, we encode subsets of samples of samples with bounded empirical variance as the solutions to a system of polynomial equations and inequalities. Our variables are $w_1, \dots, w_n$; we think of w_i as the 0/1 indicator of the presence of sample i in a set S. Subsets of $(1-\varepsilon)n$ samples with bounded variance are in one-to-one correspondence with solutions to the following:

$$w_i^2 - w_i = 0 \text{ for all } i \in [n] \tag{3.5.1}$$

$$\sum_{i\in[n]} w_i = (1-\varepsilon)n \tag{3.5.2}$$

$$\frac{1}{(1-\varepsilon)n} \sum_{i\in[n]} w_i (X_i - \mu(w))^2 \leqslant 2 \tag{3.5.3}$$

where $\mu(w)$ is shorthand for the polynomial $\mu(w) = \frac{1}{(1-\varepsilon)n} \sum_{i\in[n]} w_i X_i$.

Lemma 3.5.3. *There is a constant d such that with probability 0.99 over ε-corrupted samples $X_1, \dots, X_n$, every degree-d pseudodistribution on variables $w_1, \dots, w_n$ which satisfies (3.5.1), (3.5.2), and (3.5.3) has $|\tilde{\mathbb{E}}\, \mu(w) - \mu| \leqslant O(\sqrt{\varepsilon})$.*

Since a pseudoexpectation $\tilde{\mathbb{E}}$ as described by the lemma can be found by semidefinite programming in polynomial time (should it exist), this gives a polynomial time algorithm for one-dimensional robust mean estimation.

The proof of Lemma 3.5.3 follows the proof of Lemma 3.5.2, but uses SoS/pseudodistribution versions of the key inequalities; to do this throughout the proof we have to phrase the quantities used in the proof of Lemma 3.5.2 as polynomials. This is what we mean by "making a proof SoS".

Most steps in this kind of proof can be accomplished either "in the primal" by reasoning about pseudoexpectations or "in the dual" by reasoning about SoS proofs and polynomials. We have chosen what is convenient for each step but these choices are mostly arbitrary.

We will need a few facts before we start the proof. The proofs of these are all either in Chapter 5 or they are simple algebraic exercises.

1. *SoS Cauchy Schwarz* For indeterminates $x_1, \ldots, x_n, y_1, \ldots y_n$ there is a degree-4 sum of squares polynomial $s(x)$ such that

$$\left(\sum_{i \leqslant n} x_i y_i\right)^2 + s(x) = \left(\sum_{i \leqslant n} x_i^2\right)\left(\sum_{i \leqslant n} y_i^2\right).$$

2. *Pseudoexpectation Triangle Inequality* There is a constant C such that for every pseudoexpectation $\tilde{\mathbb{E}}$ of degree d and polynomials p, q, r of degree at most $d/2$,

$$\tilde{\mathbb{E}}(p+q+r)^2 \leqslant C(\tilde{\mathbb{E}}\, p^2 + \tilde{\mathbb{E}}\, q^2 + \tilde{\mathbb{E}}\, r^2).$$

3. For any set $T \subseteq [n]$ of size $|T| = (1-\varepsilon)n$,

$$\sum_{i \in [n]} (w_i - \mathbf{1}_{i \in T})^4 = O(\varepsilon n) - s(x) + r(x)$$

where $r(x) = r_0(x)(\sum_{i \in [n]} w_i - (1-\varepsilon)n) + \sum r_i(x)(w_i^2 - w_i)$ for some polynomials $r_0, \ldots, r_n$ of constant degree at most 8 and $s(x)$ a sum of squares of constant degree.

Proof of Lemma 3.5.3. As before, we start by expanding

$$[(1-\varepsilon)n]^2\, \tilde{\mathbb{E}}(\mu(w) - \mu_T)^2 = \tilde{\mathbb{E}}\left(\sum_{i \in [n]} (w_i - \mathbf{1}_{i \in T})^2 \cdot (w_i - \mathbf{1}_{i \in T}) X_i\right)^2$$

Here, $(w_i - \mathbf{1}_{i\in T})^2$ is playing the role of $\mathbf{1}_{X\neq X'}$ from the proof of Lemma 3.5.2, and $(w_i - \mathbf{1}_{i\in T})X_i$ plays the roll of $X - X'$. The equation is true because $\tilde{\mathbb{E}}$ satisfies $w_i^2 - w_i = 0$, and hence $\tilde{\mathbb{E}}(w_i - \mathbf{1}_{i\in T})^3 p(x) = \tilde{\mathbb{E}}(w_i - \mathbf{1}_{i\in T})p(x)$ for any polynomial p. By SoS Cauchy-Schwarz, the above is at most

$$\tilde{\mathbb{E}}\left(\sum_{i\in[n]}(w_i - \mathbf{1}_{i\in T})^4\right)\left(\sum_{i\in[n]}(w_i - \mathbf{1}_{i\in T})^2 X_i^2\right). \tag{3.5.4}$$

By our fact above, $\sum_{i\in[n]}(w_i - \mathbf{1}_{i\in T})^4 = O(\varepsilon n) - s(x) + r(x)$, where $r(x) = r_0(x)(\sum_{i\in[n]} w_i - (1-\varepsilon)n) + \sum r_i(x)(w_i^2 - w_i)$ for some constant-degree polynomials $r_0, \ldots, r_n$ and $s(x)$ is a constant-degree sum of squares. So, the above is at most

$$O(\varepsilon n)\cdot\tilde{\mathbb{E}}\sum_{i\in[n]}(w_i - \mathbf{1}_{i\in T})^2 X_i^2$$

We can expand as

$$(w_i - \mathbf{1}_{i\in T})X_i = w_i(X_i - \mu(w)) - \mathbf{1}_{i\in T}(X_i - \mu_T) + (1-\varepsilon)n(\mu(w) - \mu_T)$$

and apply the pseudoexpectation triangle inequality to conclude that

$$\begin{aligned}\tilde{\mathbb{E}}\sum_{i\in[n]}(w_i - \mathbf{1}_{i\in T})^2 X_i^2 \leqslant & O(1)\cdot\sum_{i\in[n]}\tilde{\mathbb{E}}\, w_i(X_i - \mu(w))^2 + \tilde{\mathbb{E}}\,\mathbf{1}_{i\in T}(X_i - \mu_T)^2 \\ & + O(n)\cdot\tilde{\mathbb{E}}(\mu(w) - \mu_T)^2\end{aligned}$$

The conclusion follows by rearranging and using the bounded variance assumption, which implies $\tilde{\mathbb{E}}\sum w_i(X_i - \mu(w))^2 \leqslant 2(1-\varepsilon)n$. □

CHAPTER 4

SOS LOWER BOUNDS AND PSEUDOCALIBRATION

The last topic we address before beginning the mathematical body of this thesis is the *pseudocalibration* technique for proving lower bounds against SoS refutation algorithms, which we study in detail in Part II. Because of the broad algorithmic power of SoS, lower bounds like these are among the strongest sorts of evidence presently available for the existence of information-computation gaps. At heart, pseudocalibration is a tool for proving integrality gaps of large convex programs, a project of interest beyond the statistical inference setting. Most currently-known lower bounds against high-degree SoS programs – in constraint satisfaction [81], combinatorial optimization [172], and inference – were either initially proved via pseudocalibration or can be retroactively interpreted as such.[1] It is even useful for proving lower bounds against convex programs beyond SoS, such as for extension complexity of CSPs [44].

The main idea behind pseudocalibration is: *hypothesis testing problems which lack successful $O(d \log n)$-simple statistics have refutation versions which are hard for degree-d SoS*. That is, pseudocalibration leverages the existence of a hypothesis testing problem which is hard for very simplistic algorithms – D-simple statistics – to prove refutation lower bounds against a much more sophisticated algorithm, the SoS method. The success of pseudocalibration offers the most important evidence we have in favor of Conjecture 2.2.4 about hardness of D-wise almost-independent distributions.

[1] A major exception to this rule are SoS lower bounds proved via a *symmetry reduction* strategy, such as Grigoriev's lower bound for the knapsack problem [80] and Potechin's Turàn problem lower bounds [149]. These lower bounds are have a different flavor from the ones we consider in this chapter: they are usually proved for very particular polynomial optimization problems rather than distributions over polynomial optimization problems, and they are usually only interesting for constant or slightly super-constant SoS degree d.

Much about pseudocalibration remains mysterious. As we see in this chapter, it offers a recipe for the first key step in proving an SoS lower bound, suggesting how to construct a witness – in this case a pseudodistribution – to the non-existence of good SoS proofs/refutations. However, we lack satisfying explanations of why this recipe works, and we lack general-purpose proof techniques to analyze the complicated mathematical objects it produces, which are random matrices whose entries have complex combinatorial dependencies. In Part II we are nonetheless able to use pseudocalibration to prove two of the major results in this thesis, on SoS refutations for planted clique and on equivalence of SoS and spectral methods for hypothesis testing. But many open problems remain, some of which we discuss in Appendix A.

To put pseudocalibration in context requires a little mathematical background and a short, opinionated history of lower bounds for convex-programming-based refutation algorithms.

4.1 Background and History

Proving lower bounds for SoS refutation algorithms requires ruling out the existence strong SoS-provable bounds on the optimum values of random optimization problems. Two good examples to keep in mind are the refutation versions of planted clique and planted 3-XOR. We have already discussed planted clique at some length; the next example describes 3-XOR.

Example 4.1.1 (3-XOR, refutation version)**.** Refutation of random 3-XOR instances is the most fundamental CSP refutation problem. It can be obtained via a hypothesis testing problem as in Chapter 3; here we jump to the conclusion and state the problem directly.

Let $x_1, \dots, x_n$ be Boolean variables. Let $\Delta = \Delta(n) \geqslant 0$, and let $S \subseteq \binom{n}{3}$ be a uniformly random collection of Δn triples. For each triple $ijk \in S$, let a_{ijk} be a uniformly random bit. The refutation problem is to certify an upper bound on

$$\frac{1}{\Delta n} \max_{x \in \{\pm 1\}^n} \sum_{ijk \in S} \mathbf{1}_{a_{ijk} = x_i x_j x_k} \ .$$

It is not hard to show by standard Chernoff bounds that for any sufficiently-large constant $\Delta = \Delta(\varepsilon)$, the true maximum over $x \in \{\pm 1\}^n$ is at most $1/2 + \varepsilon$. (For each fixed S, a a random $x \in \{\pm 1\}^n$ satisfies half the clauses $a_{ijk} = x_i x_j x_k$ in expectation.)

Since $\mathbf{1}_{a_{ijk} = x_i x_j x_k}$ is a polynomial in x_i, x_j, x_k,

$$\mathbf{1}_{a_{ijk} = x_i x_j x_k}(x) = \frac{a_{ijk} x_i x_j x_k + 1}{2} \ ,$$

we obtain a polynomial optimization problem with degree-3 polynomials and constraints $x_i^2 = 1$, which define the set $\{\pm 1\}^n$.

Using duality of pseudodistributions and SoS proofs, producing a degree-d pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\{p_1(x) = 0, \dots, p_m(x) = 0\}$ with $\tilde{\mathbb{E}}\, p(x) \geqslant c$ rules out degree-d SoS proofs certifying $\max_{p_1(x), \dots, p_m(x) = 0} p(x) \leqslant c - \varepsilon$ for any $\varepsilon > 0$.[2] Typically, SoS lower bounds for refutation problems are proved in just this way, producing for each problem instance (e.g. graph G for planted clique, or clauses S, α for 3-XOR) a pseudodistribution $\tilde{\mathbb{E}}$.[3]

[2] There is a small subtlety here regarding whether whether this just requires $\tilde{\mathbb{E}}\, p(x) \geqslant c$ or instead requires $\tilde{\mathbb{E}}\, p(x) s(x) \geqslant c\, \tilde{\mathbb{E}}\, s(x)$ for any SoS polynomial $s(x)$. At the level of discussion in this chapter this distinction is not important; for details see Chapter 5.

[3] In Chapter 12 we develop a technique which avoids the need to produce one pseudodistribution per instance, instead producing an object which acts sufficiently like a pseudodistribution on average. Most of the discussion in this chapter applies to the latter style of lower bound proof as well.

Relation to integrality gaps for linear and semidefinite programs In the end, constructing a pseudoexpectation as above amounts to proving that that the semidefinite programs (SDPs) underlying the SoS method suffer from large integrality gaps. The SDPs are relaxations of polynomial optimization problems like $\max_{p_1(x),\ldots,p_m(x)=0} p(x)$ whose solutions are (isomorphic to) pseudodistributions. Consider our running example of planted clique, where the SDP would be a relaxation of:

$$\max \sum_{i\in[n]} x_i \text{ such that } x_i^2 = x_i, x_i x_j = 0 \text{ for } i \nsim j \tag{4.1.1}$$

for some graph G. In this case, the convex program whose solutions are degree-d pseudodistributions is an SDP over matrices indexed by $\binom{n}{\leqslant d}$, the subsets of $[n]$ of size at most d, whose intended integral solutions are d-th tensor powers $x^{\otimes d}$ of indicator vectors of G-cliques:

$$\begin{aligned}
&\max \sum_{i\in n} X_{\{i\}\{i\}} \\
&\text{such that } X_{S,T} = X_{U,V} \text{ if } S \cup T = U \cup V \\
&\text{and } X_{\emptyset,\emptyset} = 1 \\
&\text{and } X_{S,T} = 0 \text{ if } S \cup T \text{ is not a } G\text{-clique} \\
&\text{and } X \succeq 0\,.
\end{aligned}$$

The integrality gap perspective on SoS lower bounds says that the goal is to prove for $d = O(1)$ that with high probability over G this SDP has a feasible solution X^* with $\sum_{i\in[n]} X^*_{\{i\},\{i\}} \geqslant \Omega(\sqrt{n})$, even though no integral solution has objective value more than $(2+\varepsilon)\log n$.

Integrality gaps for convex relaxations, in particular for linear programs, have been studied for decades. The gaps need here differ from what is typically

produced in two major ways. First, the problem instances are not under our control: they are sampled from a distribution which is specified by the refutation problem at hand ($G \sim G(n, 1/2)$ for planted clique, random clauses S, a for 3-XOR). Second, even for moderately large d, the constraints of the SDPs to which we must produce feasible solutions are extremely complicated: while the linear programs studied in combinatorial optimization might have $O(n)$ or $O(n^2)$ constraints for a combinatorial problem over $\{\pm 1\}^n$, our SDPs have $n^{O(d)}$ variables and a similar number of linear constraints. These constraints may interact with each other and with the constraint $X \succeq 0$ in surprising and subtle ways. Together, these differences mean that a more systematic approach is needed to prove SoS lower bounds for refutation problems than is needed to prove more elementary LP or SDP integrality gaps.

Prior SoS Lower Bounds: CSPs Prior to our work, the only tight SoS lower bounds for SoS degrees $d > 4$ were for random 3-XOR and other CSPs and CSP-like problems [81, 160, 172, 28]. This theorem of Grigoriev (later independently re-disocvered by Schoenebeck) is representative:

Theorem 4.1.2 ([81, 160])**.** *For a large enough constant Δ, with high probability over S, a as above, there exists a degree-$\Omega(n)$ pseudoexpectation on variables $x_1, \ldots, x_n$ satisfying $\{x_i^2 = x_i\}$ such that $\frac{1}{\Delta n} \tilde{\mathbb{E}} \sum_{ijk \in S} \mathbf{1}_{a_{ijk} = x_i x_j x_k} = 1$.*

The theorem is tight up to the constant hidden by the $\Omega(n)$, because degree-n SoS certifies every true inequality in n Boolean variables, and even degree-4 SoS certifies $\frac{1}{\Delta n} \sum_{ijk \in S} \tilde{\mathbb{E}} \mathbf{1}_{a_{ijk} = x_i x_j x_k} \leqslant 1$.

The proofs of Theorem 4.1.2 and other lower bounds for CSP refutation (such as [28, 107]) crucially use *locality* of constraints like $a_{ijk} x_i x_j x_k = 1$. The constraints

are local in that each involves only 3 variables at once, but there is a stronger notion of locality: roughly, in a random 3-XOR instance, the value of each variable strongly affects a small number of other nearby variables (nearby in, say, the clause-variable incidence graph) and does not affect the remaining variables at all.

This idea can be made formal by two (equivalent) routes. One path is Grigoriev's original connection between SoS and the width of resolution derivations, together with the fact that expansion of a random constraint graph ensures that values for most variables are not derivable in low width from values from most others. The other path is via simple statistics and the observation (provable by a calculation that involves, unsurprisingly, low-width resolution proofs) that even in a distribution over random 3-XOR instances for which there exists a satisfying assignment x, for most pairs of variables x_i, x_j no function of low-degree in S, a, x_i correctly outputs the value of x_j.

Pursuing this further would take us too far afield here: instead, we turn to the breakdown of this kind of locality for planted clique and other dense problems.

4.2 The Challenge of Non-Locality

The success of SoS in solving some inference problems which appeared beyond the reach of other methods,[4] together with its strength as an algorithm in related domains[5] generated around 2013 a great deal of interest in the question: can SoS proofs of constant degree refute the existence of $o(\sqrt{n})$-size cliques in

[4] Dictionary learning and planted sparse vector recovery, for example [31, 30].

[5] Particularly combinatorial optimization and for problems surrounding the unique games conjecture [27].

$G \sim G(n, 1/2)$?[6]

Prior refutation lower bounds for SoS in the CSP setting were often proved by starting with pseudodistribution constructions which successfully proved lower bounds against weaker proof systems than SoS (such as the Sherali-Adams and Lovasz-Schrijver proof systems) and strengthening their analyses. Two groups of researchers – Meka, Potechin, and Wigderson, and, separately, Deshpande and Montanari – took this route, attempting to adapt a previous lower bound due to Feige and Krauthgamer for weaker, Lovasz-Schrijver proofs [71]. They made substantial headway, but did not succeed in proving a tight lower bound:

Theorem 4.2.1 (Meka, Potechin, Wigderson [128])**.** *For every $d \in \mathbb{N}$, with high probability over $G \sim G(n, 1/2)$ there is a degree-d pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\{x_i^2 - x_i = 0, x_i x_j = 0$ if $x_i \not\sim x_j$ in $G\}$ with $\tilde{\mathbb{E}} \sum_{i \in [n]} x_i \geqslant n^{1/d}/(\log n)^{O(1)}$.*[7]

This theorem rules out the possibility that constant-degree SoS can refute existence of sub-polynomial-size cliques, but it does not finish the story. Any substantial improvement to Theorem 4.2.1 comes up against a major obstacle: the pseudodistribution construction of Feige and Krauthgamer fails the positivity constraint $\tilde{\mathbb{E}}\, p(x)^2 \geqslant 0$ when parameters are set so that $\tilde{\mathbb{E}} \sum x_i \gg n^{1/(d-1)}$.

Breakdown of the Feige-Krauthgamer Witness Feige and Krauthgamer suggested the following natural potential pseudodistribution $\tilde{\mathbb{E}}$, which respects all the local consequences of the constraints $x_i x_j = 0$ for $i \not\sim j$ in G. Given a graph G and a clique-size parameter $k = k(n)$, for each $S \in \binom{n}{\leqslant d}$, set $\tilde{\mathbb{E}}_G\, x^S = (k/n)^{|S|}$

[6] Strictly speaking this question remains unresolved: even in light of our main result Theorem 1.1.1 it remains possible even in light of results in this thesis that constant degree SoS could refute existence of cliques of size roughly $\sqrt{n}/2^{\sqrt{\log n}}$.

[7] Deshpande and Montanari [63] improved this result quantitatively for the case $d = 4$, but still came short of a tight lower bound. See more discussion in [89].

if S is a G-clique and otherwise $\tilde{\mathbb{E}}_G x^S = 0$. Notice that $\tilde{\mathbb{E}} \sum x_i = k$, and clearly $\tilde{\mathbb{E}}\, p(x) x_i x_j = 0$ if $i \not\sim j$ in G.

Kelner gave the following counter-example to positivity of $\tilde{\mathbb{E}}$ when $k \gg n^{1/3}$ for $d \geqslant 4$ [104], which shows that the Feige-Krauthgamer construction fails to account for some non-local consequences of the constraints $x_i x_j = 0$. For simplicity we describe the counterexample for $d \geqslant 6$.

Fact 4.2.2 (Kelner [104]). *Let G be an n-node graph. For every $i \in [n]$, let $r_i \in \{\pm 1\}^n$ be the vector with $r_i(j) = 1$ if $i \sim j$, $r_i(j) = -1$ if $i \not\sim j$, and $r_i(i) = 1$. Every degree-6 pseudoexpectation $\tilde{\mathbb{E}}$ which satisfies $\{x_i^2 = x_i, x_i x_j = 0 \text{ if } i \not\sim j\}$ has $\tilde{\mathbb{E}} \sum_{i \in [n]} \langle r_i, x \rangle^4 \geqslant \tilde{\mathbb{E}} (\sum_{i \in [n]} x_i)^5$.*

Proof. We start by considering $\sum_{i \in [n]} x_i \langle r_i, x \rangle^4$. Expanding, $\sum_{i \in [n]} x_i \langle r_i, x \rangle^4 = \sum_{i,s,t,u,v \in [n]} x_i x_s x_t x_u x_v r_i(s) r_i(t) r_i(u) r_i(v)$. Since $\tilde{\mathbb{E}}$ satisfies $x_i x_j = 0$ for $i \not\sim j$, for every i, s, t, u, v which is not a G-clique, we have

$$\tilde{\mathbb{E}}\, x_i x_s x_t x_u x_v r_i(s) r_i(t) r_i(u) r_i(v) = \tilde{\mathbb{E}}\, x_i x_s x_t x_u x_v \,.$$

On the other hand, if $\{i, s, t, u, v\}$ is a G-clique, then $r_i(s) r_i(t) r_i(u) r_i(v) = 1$. So all together,

$$\tilde{\mathbb{E}} \sum_{i,s,t,u,v \in [n]} x_i \langle r_i, x \rangle^4 = \tilde{\mathbb{E}} \sum_{i,s,t,u,v \in [n]} x_i x_s x_t x_u x_v = \tilde{\mathbb{E}} \left(\sum_i x_i \right)^5 .$$

Now, since $\tilde{\mathbb{E}}$ satisfies $x_i^2 - x_i = 0$, we claim that

$$\tilde{\mathbb{E}} \langle r_i, x \rangle^4 \geqslant \tilde{\mathbb{E}}\, x_i \langle r_i, x \rangle^4$$

for all i, since $(1 - x) =_{x_i^2 = x_i} (1 - x)^2$. This finishes the proof. □

The proof of Fact 4.2.2 constructs an SoS proof whose coefficients are simple statistics which derives a nontrivial relationship between two polynomials which

each involve all the variables $x_1, \ldots, x_n$. (Note that when the polynomial $\langle r_i, x\rangle^4$ is evaluated at the indicator vector x of a clique S, it counts the number of 4-cliques in S with all vertices adjacent to i.) The candidate pseudoexpectation construction of Feige and Krauthgamer does not account for this sort of non-local SoS proof:

Fact 4.2.3. *Let $\tilde{\mathbb{E}}_G$ be the Feige-Krauthgamer functional for a graph G. If $d = 4$ and $k \gg n^{1/3}$, then $\mathbb{E}_{G\sim G(n,1/2)} \tilde{\mathbb{E}}_G \sum_{i\in[n]} \langle r_i, x\rangle \ll \mathbb{E}_{G\sim G(n,1/2)} \tilde{\mathbb{E}}_G \left(\sum_{i\in[n]} x_i\right)^5$.*

The proof is a straightforward calculation, so we leave it out. An intuitive explanation of the whole problem is that degree-6 SoS proves that 4-cliques which are contained in a phantom k-clique also participate, on average, in more 5-cliques than a typical 4-clique in $G(n, 1/2)$, but the Feige-Krauthgamer construction is not accounting for this.

Together, Kelner's observations show that the Feige-Krauthgamer functional cannot be used to prove a tight SoS lower bound for planted clique: it must fail the positivity requirement. With a little additional work, it is possible to conclude that there is a constant-degree polynomial $q(x)$ whose coefficients are constant-degree functions of the entries of the adjacency matrix of the graph G, such that $\mathbb{E}_{G(n,1/2)} \tilde{\mathbb{E}}\, q(x)^2 \ll 0$.

The difficulty runs deeper than just one polynomial such as $\sum_{i\in[n]} \langle r_i, x\rangle^4$ or just the planted clique problem – as one might imagine there are many other SoS-provable inequalities between nontrivial, non-local polynomials, even just for planted clique. Other dense problems – component analysis problems, for example – suffer from similar difficulties. There are several interpretations of the origin of this difficulty – see e.g. the introduction of [29]. For now we move on to the solution proposed by pseudocalibration.

4.3 The Pseudocalibration Recipe

Pseudocalibration is a recipe which turns a nice-enough *alternative distribution* into a candidate pseudoexpectation for use in proving a refutation lower bound. Imagine we are faced with a refutation problem (perhaps it came from a hypothesis testing problem as in Chapter 3, but it may also stand on its own): for some distribution ν on Ω^N and some polynomials $p, p_1, \ldots, p_m$ of degree at most d in variables $y_1, \ldots, y_N, x_1, \ldots, x_n$, we are interested degree-d SoS-certifiable upper bounds on

$$\max_{x \in \mathbb{R}^n} p(y, x) \quad \text{s.t} \quad p_1(y, x) = 0, \ldots, p_m(y, x) = 0 \tag{4.3.1}$$

when $y \sim \nu$. (We could have inequalities $p_i(y, x) \geqslant 0$ but we avoid this here for simplicity.)

If we would like to show that degree-d SoS with high probability (or in expectation) *fails* to certify an upper bound of $c - \varepsilon$ on (4.3.1), we need to invent for each y a degree-d pseudodistribution $\tilde{\mathbb{E}}_y$ in variables $x_1, \ldots, x_n$ which satisfies $\{p_1(y, x) = 0, \ldots, p_m(y, x) = 0\}$ and has $\tilde{\mathbb{E}}_y\, p(y, x) \geqslant c$ with high probability. As Section 4.2 made clear, in many interesting cases there is no obvious choice for $\tilde{\mathbb{E}}_y$.

Now hypothesis testing enters the picture. Thinking of ν as a null model, suppose that μ is another (alternative) distribution on Ω^N supported on y for which there exists x such that (4.3.1) is satisfied by y, x and $p(y, x) \geqslant c$. By abuse of notation, we also write μ for the joint distribution on $\Omega^N \times \mathbb{R}^n$ supported on pairs y, x where x is such a satisfying solution. (In most of the literature on pseudocalibration, the distribution μ is called the *planted* distribution.) In the case of planted clique, we will take μ to be the usual planted k-clique distribution.

The distribution μ induces the following map from Ω^N to linear functionals on degree-d polynomials in x: for $y \in \Omega^N$, let $\Lambda(y) : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ be the operator such that

$$\Lambda(y)[q(x)] = \frac{\mathbb{P}_\mu(y)}{\mathbb{P}_\nu(y)} \cdot \mathop{\mathbb{E}}_{\mu}[q(x) \mid y] \,.$$

Notice the appearance of the likelihood ratio $\frac{\mathbb{P}_\mu(y)}{\mathbb{P}_\nu(y)}$, familiar from hypothesis testing.

Λ has some initially promising properties which make it look like a good candidate for a pseudodistribution. First of all, since x in the support of the conditional distribution $\{x \mid y\}_\mu$ always satisfies $p_1(y,x) = 0, \ldots, p_m(y,x) = 0$, for each y we will have $\Lambda(y)[p_i(y,x)q(x)] = 0$ for every q. Furthermore, $\Lambda(y)[q(x)^2] \geqslant 0$, again because $\mathbb{E}_\mu[q(x)^2 \mid y] \geqslant 0$. All of these follow from one key observation: for any polynomial $q(y,x)$,

$$\mathop{\mathbb{E}}_{\nu} \Lambda(y)[q(y,x)] = \mathop{\mathbb{E}}_{x,y\sim\mu} q(y,x) \,. \tag{4.3.2}$$

But Λ has a fatal flaw, as might have been guessed when the likelihood ratio appeared. In short, if the lower bound we hope to prove is interesting, then for most $y \sim \nu$ there should not actually be x which makes the value of (4.3.1) greater than c. On the other hand, μ is supported on y which do have such x. The consequence is that the function $\frac{\mathbb{P}_\mu(y)}{\mathbb{P}_\nu(y)}$ takes nonzero values only on a small amount of Ω^N, as measured by ν. This makes $\Lambda(y)$ a poorly-behaved random variable – one way to see this is that $\Lambda(y)[1]$, that is $\Lambda(y)$ applied to the constant polynomial, is usually either zero or much larger than 1. Normalizing $\Lambda(y)$ will not help, as this will destroy $\Lambda(y)[p(y,x)] \geqslant c$.

The key insight is that to eliminate the challenge imposed by the sort of SoS proof from Section 4.2, whose coefficients are low degree polynomials in y, we

do not need (4.3.2) to hold for every polynomial $q(y, x)$ – only for those with degree at most d in x and degree at most $D \gg d$ in y. This opens an avenue to address the difficulty with Λ's poor behavior. If $\Lambda'(y)$ is any map from Ω^N to linear functionals which agrees *in expectation* with $\Lambda(y)$ for polynomials which are degree-d in x and degree-D in y, then

$$\mathop{\mathbb{E}}_{\nu} \Lambda'(y)[q(y, x)] = \mathop{\mathbb{E}}_{\mu} q(y, x) \tag{4.3.3}$$

and so, in particular $\mathbb{E}_\nu \Lambda'(y)[r(y, x)^2] \geqslant 0$ for any $r(y, x)$ which is degree at most $d/2$ in x and at most $D/2$ in y.[8]

Now we have to make a choice: how to choose Λ' satisfying (4.3.3)? One canonical possibility, recalling the problem with Λ that it took large values on a small portion of Ω^N and small ones elsewhere, is to choose $\Lambda'(y)$ to minimize the variance of $\Lambda'(y)[q(x)]$ with respect to $y \sim \nu$. With a small amount of linear algebra, it is possible to see that the minimizing choice is to take $\Lambda'(y)[q(x)] = (\Lambda(y)[q(x)])^{\leqslant D}$, where the superscript $\leqslant D$ denotes the orthogonal projection (with respect to ν) of the function $y \mapsto \Lambda(y)[q(x)]$ to the span of degree-D polynomials in $y_1, \ldots, y_N$.

Let us see what effect this choice has on the variance of Λ'. As a test case, consider $\Lambda'(y)[1]$, which caused so much trouble previously. *It is* $\Lambda'(y)[1] = \left(\frac{\mathbb{P}_\mu(y)}{\mathbb{P}_\nu(y)}\right)^{\leqslant D}(y)$ *– the low-degree likelihood ratio from Chapter 2.* And by definition, if there are no successful D-simple statistics to distinguish μ from ν, then the variance of $\Lambda'(y)[1]$ is $O(1)$. Since $\mathbb{E}_\nu \Lambda'(y)[1] = 1$, if the variance is $O(1)$ then there is some hope we have solved the issues with Λ.[9] *This is the first hint of*

[8] The authors of [29], adopting naming conventions from Bayesian statistics, called (4.3.3) being *calibrated to* μ with respect to simple tests q; together with "pseudo" from *pseudo*distribution this is the origin of the name *pseudocalibration*.

[9] In all our uses of this technique, we are able to adjust problem parameters – for example adjust the choice of clique size k in our planted clique lower bound – to make this variance $o(1)$.

fundamental connection between failure of simple statistics for hypothesis testing and SoS lower bounds for refutation

Of course, defining such Λ' does not yet show an SoS lower bound – just that the difficulties from Section 4.2 may no longer be problematic. One needs to check that Λ' satisfies the properties of a pseudoexpectation: that it is nonnegative, and that $\Lambda'(y)[p_i(y,x)q(x)] = 0$, or adjust Λ' accordingly if it does not. In Chapters 11 and 12 we show two different approaches to working formally with Λ'.

The merit of pseudocalibration is to offer a plausible guess – a canonical starting point – for SoS refutation lower bounds. This starting point, the degree-D operator Λ', is used in some form in every SoS refutation lower bound we know how to prove – for constraint satisfaction, component analysis, and planted clique, making it appear to be a unifying approach to SoS lower bounds. But there is a lot left to do: we do not have canonical proof techniques for analyzing Λ', nor do we know beyond guesswork how to choose a good planted distribution μ or the right threshold D. See Appendix A for more that we do not (yet) know about pseudocalibration.

CHAPTER 5

PRELIMINARIES

SoS Proofs and the SoS Algorithm

SoS proofs, pseudodistributions, and duality theorems are covered extensively in other texts, so we will be brief here. For the most complete treatment which takes a similar perspective to ours here, see [34].

Definition 5.0.1 (SoS polynomial)**.** A polynomial $p \in \mathbb{R}[x]$ is an SoS (sometimes we just write "p is SoS") if it can be expressed as $p(x) = \sum_{i \leqslant m} q_i(x)^2$ for some $q_1, \ldots, q_m \in \mathbb{R}[x]$.

Definition 5.0.2 (SoS refutation and SoS proof)**.** Let $p_1, \ldots, p_m \in \mathbb{R}[x_1, \ldots, x_n]_{\leqslant d}$ be polynomials with real coefficients of degree at most d. We say there is a degree d SoS refutation of the system $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$ if there exist SoS polynomials $\{r_S(x)\}_{S \subseteq [m]}$ such that

$$-1 = \sum_{S \subseteq [m]} r_S(x) \cdot \prod_{i \in S} p_i(x)$$

and each of the polynomials $r_S(x) \cdot \prod_{i \in S} p_i(x)$ has degree at most d.

We say that there is a degree d SoS proof of $q(x) \geqslant 0$ from $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$ if there exist SoS polynomials $\{r_S\}_{S \subseteq [m]}$ such that $q(x) = \sum_{S \subseteq [m]} r_S(x) \prod_{i \in S} p_i(x)$, and each of the polynomials $r_S(x) \cdot \prod_{i \in S} p_i(x)$ has degree at most d. If this is the case, we write

$$\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\} \vdash_d q \geqslant 0 \,.$$

We will often use without mention various valid deduction rules for SoS proofs – see [34]. We often include equations as well as inequalities in our SoS proofs; $p(x) = 0$ is shorthand for $\{p(x) \geqslant 0, -p(x) \geqslant 0\}$.

Definition 5.0.3 (Pseudoexpectation)**.** A degree-d pseudoexpectation $\tilde{\mathbb{E}}$ on variables $x = x_1, \ldots, x_n$ is a linear map $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ which is normalized and nonnegative:

$$\tilde{\mathbb{E}}[1] = 1 \text{ and } \tilde{\mathbb{E}}[p(x)^2] \geqslant 0\,.$$

Definition 5.0.4 (Satisfying inequalities)**.** A degree-d pseudoexpectation $\tilde{\mathbb{E}}$ satisfies $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$ if for every SoS polynomial q and every $S \subseteq [m]$ such that $q(x) \prod_{i \in S} p_i(x)$ has degree at most d, we have $\tilde{\mathbb{E}}\, q(x) \prod_{i \in S} p_i(x) \geqslant 0$.

See [34] for a proof of the following theorem.

Theorem 5.0.5 (Duality)**.** *Suppose $\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\}$ with $p_i \in \mathbb{R}[x]_{\leqslant d}$ contains the inequality $\|x\|^2 \leqslant M$ for some $M \in \mathbb{R}$. For every $q \in \mathbb{R}[x]_{\leqslant d}$, either*

- *there is for every $\varepsilon > 0$ either a degree-d SoS proof $\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\} \vdash_d q \geqslant -\varepsilon$, or*
- *there is a degree-d pseudoexpectation $\tilde{\mathbb{E}}$ satisfying $\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\}$ with $\tilde{\mathbb{E}}\, q(x) \leqslant 0$.*

The main utility of SoS proofs and pseudoexpectations comes from the fact that for many systems $p_1, \ldots, p_m$ which arise in theoretical computer science, the above duality theorem can be made algorithmic: that is, there is generally an $(mn)^{O(d)}$-time algorithm to decide, given q, which of the alternatives obtains. The reason, roughly, is that the set of SoS using axioms $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$ is

an $(mn)^{O(d)}$-variable semidefinite program, whose dual is the set of pseudoexpectations satisfying $\{p_1(x) \geqslant 0, \ldots, p_m(x) \geqslant 0\}$. Because of wide flexibility in designing numerically-unstable families of polynomials $p_1, \ldots, p_m$, it is possible for these semidefinite programs not to be solvable by known, efficient algorithms [143, 156].

However, such issues do not arise for any of the SoS programs we use in this paper – when necessary we occasionally note why this is the case in the text. Hence, we often use without mention the fact that for nice-enough systems of inequalities $\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\}$, there is a polynomial-time algorithm to decide whether or not there exists an SoS refutation of $\{p_1 \geqslant 0, \ldots, p_m \geqslant 0\}$.

5.0.1 Useful SoS Inequalities

We record some broadly useful SoS inequalities. More specialized statements can be found in Chapter 10.

Lemma 5.0.6. *For indeterminates $x_1, \ldots, x_n, y_1, \ldots, y_n$,*

$$\vdash_4 \langle x, y\rangle^2 \leqslant \|x\|^2\|y\|^2 .$$

Proof. Following [121], Lemma A.1, we see that the difference between the right and left hand sides is

$$\|x\|^2\|y\|^2 - \langle x, y\rangle^2 = \sum_{ij}(x_i y_j - x_j y_i)^2$$

which is a sum of squares. □

Lemma 5.0.7. *For indeterminates $x_1, \ldots, x_n$ and an $n \times n$ matrix M,*

$$\vdash_2 \langle x, Mx\rangle \leqslant \|M\| \cdot \|x\|^2 .$$

Proof. The matrix $\|M\| \cdot \mathrm{Id} - M \succeq 0$ is PSD, so the difference between right-and left-hand sides is a sum of squares. □

Concentration and Matrix Concentration

We will frequently use without comment standard concentration inequalities such as Markov's, Chebyshev's, and Chernoff-style bounds – we refer the reader to [43] for a thorough introduction.

We will also frequently use matrix concentration inequalities. When we use specialized results we state them in the text, one workhorse which we state here is the matrix Bernstein inequality. This statement is borrowed from Theorem 1.6.2 of Tropp [170].

Theorem 5.0.8 (Matrix Bernstein)**.** *Let $S_1, \dots, S_m$ be independent square random matrices with dimension n. Assume that each matrix has bounded deviation from its mean: $\|S_i - \mathbb{E}\, S_i\| \leqslant R$ for all i. Form the sum $Z = \sum_i S_i$ and introduce a variance parameter*

$$\sigma^2 = \max\{\|\,\mathbb{E}(Z - \mathbb{E}\, Z)(Z - \mathbb{E}\, Z)^T\|, \|\,\mathbb{E}(Z - \mathbb{E}\, Z)^T(Z - \mathbb{E}\, Z)\|\}\,.$$

Then

$$\mathbb{P}\{\|Z - \mathbb{E}\, Z\| \geqslant t\} \leqslant 2n \exp\left(\frac{t^2/2}{\sigma^2 + Rt/3}\right) \qquad \text{for all } t \geqslant 0\,.$$

Part I

Algorithms from Low-Degree Polynomials

CHAPTER 6

CASE STUDY: THE SPIKED TENSOR MODEL

In this chapter we will take a detailed look at one example set of Bayesian inference problems, coming from the *spiked tensor model*. The spiked tensor model is of considerable interest in its own right: it is a model for higher-order versions of principal component analysis (tensor PCA), it appears in statistical physics as the spherical p-spin model, and it is an simple testing ground to study algorithms for tensor problems which can often be extended to solve more sophisticated problems like tensor decomposition and tensor completion [121, 150]. It is also a dense version of one of the classic random models in theoretical computer science: random constraint satisfaction.

Most of the central themes of this thesis appear in (relatively) simple form in this chapter. We will see an example of polynomial-time algorithm design for refutation and estimation problems via construction of an SoS proof which itself uses ideas from random matrix theory. We will see that spiked tensor models have an information-computation gap, as evidenced by SoS lower bounds we prove via the pseudocalibration method. And, we will see that that the algorithms design and lower bounds we prove are consistent with our hypothesis on optimality of simple statistics, when we analyze (via routine calculations) which spiked tensor models permit successful simple statistics.

Definition 6.0.1 (Single-spike k-tensor model)**.** Let $k, n \in \mathbb{N}$ and $\lambda = \lambda(n) \geqslant 0$, and let P be a distribution on unit vectors in $\mathbb{R}^n$. The single-spike tensor model is a probability distribution over pairs $\{v, T\}$ where $v \in \mathbb{R}^n$ and T is a k-tensor in $(\mathbb{R}^n)^{\otimes k}$. The marginal distribution of v is P, and to sample T given v, first sample $G \in (\mathbb{R}^n)^{\otimes k}$ with iid standard Gaussian entries (subject to symmetry: formally,

for every multiset $\alpha \subset [n]$ of size k sample a Gaussian $G_\alpha \sim \mathcal{N}(0,1)$), and let

$$T = \lambda \cdot v^{\otimes k} + G\,.$$

We call λ the *signal strength* or *signal-to-noise ratio* (SNR).

Two particularly interesting choices of the prior P uniform (Haar) distribution on the unit sphere and the Rademacher prior, uniform over $\{\pm 1/\sqrt{n}\}$.

A 2-tensor is a matrix, and for the case $k = 2$ the single-spike model becomes the long-studied spiked GOE (for "Gaussian Orthogonal Ensemble") model from random matrix theory. As often happens in algorithms, the $k = 2$ case differs substantially from $k \geqslant 3$, because many problems which can be solved with eigenvalue computations when $k = 2$ have no obvious analogous algorithms when $k \geqslant 3$. For this chapter it is best to think of $k \geqslant 3$; in fact $k = 3$ and $k = 4$ are perhaps the best cases to keep in mind, as once they are understood the behavior of larger k is easy to work out.

If we introduce a null model, we get a hypothesis testing problem for every P, λ, k, n. Given a tensor $T \in (\mathbb{R}^n)^{\otimes k}$, distinguish the following:

> H_0 : T was sampled (symmetrically) with iid entries from $\mathcal{N}(0,1)$
>
> H_1 : T was sampled from the single-spike k-tensor model with signal strength λ and prior P.

For each fixed k, P, n, as λ increases the distributions specified by H_0 and H_1 become less alike, so performing hypothesis testing becomes easier. One way to phrase the main question about efficient algorithms for hypothesis testing is the following.

Question 6.0.2 (Hypothesis testing for the single-spike tensor model)**.** Fix a prior P, an integer $k \in \mathbb{N}$, and a running time $t : \mathbb{N} \to \mathbb{N}$.[1] Let $b \in \{0,1\}$ be a uniformly random bit. Suppose a tensor T is drawn from the H_b. What is the least λ such that there exists an algorithm $A(T)$ running in time $t(n)$ such that $\mathbb{P}_{b,T}(A(T) = b) \geqslant 1 - o(1)$?

There are also *certification/refutation* and estimation problems associated with the spiked tensor model.

Question 6.0.3 (Refutation for the single-spike tensor model)**.** Fix $k \in \mathbb{N}$ and a running time $t(n)$. What is the least $\alpha = \alpha(n)$ such that there exists an algorithm $A(T)$ running in time $t(n)$ which outputs a real number such that $A(T) \geqslant \max_{\|x\|=1} \langle T, x^{\otimes k} \rangle$ for every T, and $\mathbb{P}_{H_0}(A(T) > \alpha) \leqslant o(1)$?

Question 6.0.4 (Estimation for the single-spike tensor model)**.** Fix a prior P, an integer $k \in \mathbb{N}$, and a running time $t(n)$. What is the smallest λ such that there exists an algorithm $A(T)$ which outputs a unit vector in $\mathbb{R}^n$ and runs in time $t(n)$, with $\mathbb{E}_{x,T\sim H_1} \langle x, A(T) \rangle \geqslant 1 - o(1)$ (if k is odd) or $\mathbb{E}_{x,T\sim H_1} \langle x, A(T) \rangle^2 \geqslant 1 - o(1)$ (if k is even)?

When the running times $t(n)$ are 2^{Cn} for large-enough C, for most interesting priors P we enter by-now well-understood statistical territory. The reason is that in large-enough exponential time a few key primitives can be computed, including:

[1] Whenever we talk abstractly about a time function in this thesis, we mean to restrict attention to "nice" time functions, to avoid unnecessary mathematical pathologies. See e.g [26], lecture 11, or any modern introduction to theoretical computer science. Talking about running times also implicitly requires a machine model with respect to which running time is defined; throughout the thesis the real RAM model is a good choice.

- the likelihood ratio $\mathbb{P}_{H_1}(T)/\mathbb{P}_{H_0}(T)$ (needed for the statistically-optimal hypothesis test, the *likelihood ratio test*),
- the maximum-likelihood and maximum-a-posteriori estimators of the hidden spike, respectively $\operatorname{argmax}_{x\in\mathbb{R}^n} \mathbb{P}_{H_1}(T \mid x)$ and $\operatorname{argmax}_{x\in\mathbb{R}^n} \mathbb{P}_{H_1}(x \mid T)$,
- moments of the posterior distribution, $\mathbb{E}_{H_1}[x \mid T]$, $\mathbb{E}_{H_1}[xx^\top \mid T]$, etc.

Consequently, answers to all of the above questions are essentially understood when $t(n) \geqslant 2^{Cn}$ for big-enough C, at least for most reasonable priors P, including uniform (Haar-distributed) and Rademacher [147, 19]. For example, consider the following rather precise theorem of Perry, Wein, and Bandeira, addressing the hypothesis testing question.

Theorem 6.0.5 (Theorem 1.4 in [147])**.** *Suppose $t(n) \geqslant 2^{Cn}$ for large-enough C. Fix any $\varepsilon > 0$ and let P be the Rademacher prior. Let $\lambda(n,k)$ be the minimum SNR for successful hypothesis testing. Then*

$$\lim_{k\to\infty} \frac{\lim_{n\to\infty} \lambda \cdot n^{-1/2}}{2\sqrt{k! \log k}} = 1\,.$$

The theorem says that the minimal λ is $2\sqrt{k! \log k} \cdot n^{1/2}$, to leading order in both k and n. Analogous results are known for certifying and estimation. In particular, for some (known and explicit) constant $C(k)$, it is known that $\mathbb{E}_{T\sim H_0} \max_{\|x\|=1} \langle T, x^{\otimes k}\rangle = (C(k) \pm o(1))\sqrt{n}$.

When $t(n) \leqslant 2^{o(n)}$, however, the game changes entirely, since straightforward algorithms statistically-optimal hypothesis testing and estimation no longer exist: lack of running time, rather than lack of statistical information, becomes the main roadblock. The emerging picture for running times from subexponential to polynomial is:

Hypothesis 6.0.6. *For every $k \in \mathbb{N}$ and $1 \geqslant \varepsilon \geqslant 0$, spiked tensor problems with $\lambda = n^{\frac{k}{4}-\varepsilon\cdot\left(\frac{k}{4}-\frac{1}{2}\right)}$ are solvable in time $t(n) = 2^{\tilde{O}(n^{\varepsilon})}$, and not faster.*

In particular, when $k = 3$ and $t(n) = \mathrm{poly}(n)$, spiked tensor problems appear to be solvable if and only if $\lambda \geqslant \Omega(n^{3/4})$, despite their exponential-time solvability for $\lambda = \Theta(\sqrt{n})$ – this is another information-computation gap. In Section 6.5 below we will see that Hypothesis 6.0.6 is a special case for spiked tensor models of Hypothesis 2.1.5 on simple statistics and efficient algorithms.

6.1 Main Results

This thesis contributes several theorems in support of Hypothesis 6.0.6. (We also discuss some related works which fill out the picture.) The algorithms and analyses we describe are our first formal example of the connection between convex programs and simple statistics: they all involve the construction either of *dual certificates, a.k.a. SoS proofs*, or *primal solutions, a.k.a. pseudodistributions* which are themselves *composed of simple statistics.*

The first theorem concerns polynomial-time algorithms.

Theorem 6.1.1. *For every $k \in \mathbb{N}$, there are polynomial-time SoS algorithms for $O(n^{\frac{k}{4}}(\log n)^{1/4})$-refutation and for estimation with $\lambda = \omega(n^{\frac{k}{4}}(\log n)^{1/4}$ for spiked k-tensor problems (with any prior P).*

The second theorem concerns lower bounds for SoS-based refutation algorithms.

Theorem 6.1.2. *There is an absolute constant $\varepsilon^* > 0$ such that for every $e \in [0, \varepsilon^*]$, with high probability over a Gaussian 3-tensor $T \sim H_0$, the minimum degree of an SoS proof certifying $\max_{\|x\|^2=1}\langle T, x^{\otimes 3}\rangle \leqslant n^{3/4-\varepsilon}$ is $n^{\Omega(\varepsilon)}$.*

The third theorem concerns the question: are the signal-to-noise ratios achieved by polynomial-time SoS algorithms achievable by algorithms with practical polynomial running times? We answer this question affirmatively for the single-spike tensor model, giving linear time algorithms for estimation and hypothesis testing, and a subquadratic-time algorithm for refutation, all inspired by our polynomial-time SoS algorithms. For simplicity, we just treat the case $k = 3$.

Theorem 6.1.3. *There is an $n^4(\log n)^{O(1)}$-time $O(n^{3/4}(\log n)^{1/4})$-refutation refutation in the spiked tensor model (where the input size is n^3). And, for every $\varepsilon > 0$, if $\lambda > n^{3/4}(\log n)/\varepsilon$, there is a linear-time estimation algorithm which given $T \sim H_1$ outputs a unit vector v' such that $\langle v, v' \rangle \geqslant 1 - O(\varepsilon)$ with high probability.*

The last original theorem of this chapter (which is proved by a straightforward calculation) establishes the simple statistics picture for the spiked tensor model. It shows that successful D-simple statistics for small D appear only when $\lambda \geqslant \Omega(n^{3/4})$, and furthermore captures a SNR vs complexity tradeoff: as D increases, smaller λ's admit D-simple statistics. This theorem demonstrates that Hypothesis 6.0.6 and Theorems 6.1.1 to 6.1.3 and 6.1.5 which support it are predicted by Hypothesis 2.1.5 on the optimality of simple statistics for inference problems. Recall the definition of D-simple statistics from Chapter 2.

Theorem 6.1.4. *Let ν be the distribution on $\mathbb{R}^{\binom{n}{3}}$ which places a standard Gaussian in each entry; that is, ν is the null distribution H_0 for the spiked tensor problem. Let μ be the spiked tensor distribution with Rademacher prior and SNR λ. For every $D \in \mathbb{N}$, if $\frac{\lambda}{n^{3/4}} \leqslant \frac{1}{D^{O(1)}}$, then every D-simple statistic f has $\mathbb{E}_{T\sim\mu} f(T) \leqslant O(1)$. On the other hand, there is a constant c such that if $\frac{\lambda}{n^{3/4}} \geqslant 1/D^c$ and $D = \omega(1)$ then there is a D-simple statistic f with $\lim_{n\to\infty} \mathbb{E}_{T\sim\mu} f(T) = \infty$.*

Finally, the following theorem, which constitutes related work proved in-

dependently by Bhattiprolu-Guruswami-Lee and Raghavendra-Rao-Schramm, essentially completes the picture, by giving an SoS algorithm matching the lower bound in Theorem 6.1.2 (up to constants in the exponent of the SoS degree d[2]).

Theorem 6.1.5 (High degree SoS refutation algorithms for 3-tensors, Bhattiprolu-Guruswami-Lee version [40, 153])**.** *For every $d \leqslant n$, degree-d SoS certifies that*

$$\max_{\|x\|=1} \langle T, x^{\otimes 3} \rangle \leqslant \frac{2^{O(d)} \cdot (\log n)^{O(1)} \cdot n^{3/4}}{d^{1/4}}$$

with high probability for a 3-tensor $T \sim H_0$ with iid standard Gaussian entries.

6.2 Overview of Proofs

The proofs of our algorithmic results Theorem 6.1.1 and Theorem 6.1.3 both boil down to analyzing the spectrum of carefully-constructed random matrices whose entries are low degree functions of an input tensor T. In the case of Theorem 6.1.1 these random matrices serve as dual certificates to an SoS semidefinite program (that is, they are SoS proofs), while in the case of Theorem 6.1.3 the matrices are directly constructed by the algorithm.

In both cases, the main innovation over prior work is the appearance of matrices which are not simple *flattenings* of an input tensor T: the entries of these matrices are nontrivial degree-2 functions of T. This is the key to obtaining the results in Theorem 6.1.1 for *odd* values of k, where there is good reason to believe that simpler matrices cannot provide as strong refutation guarantees [92]. Of course, not just any degree-2 matrix-valued function of T has a spectrum suitable for refutation or estimation: constructing an SoS-provable upper bound

[2]We suspect that Theorem 6.1.5 has the right constants in the exponent, and Theorem 6.1.2 could in theory be tightened.

on $\max_{\|x\|=1}\langle T, x^{\otimes 3}\rangle$ in the proof of Theorem 6.1.1 leads us to the right random matrix.

The proof of Theorem 6.1.2 is much more technical, and we do not give all the details here. (A nearly identical argument is presented in full in Chapter 11.) The strategy to show that with high probability no degree-d SoS proof certifies $\max_{\|x\|=1}\langle T, x^{\otimes 3}\rangle \leqslant c$ for some $c \geqslant 0$ and T with Gaussian entries goes via duality. That is, we show that with high probability over T there exists a degree-d pseudodistribution which satisfies $\{\|x\|^2 = 1\}$ and has $\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle \geqslant c$.

The difficulty in this strategy twofold. First, there is no obvious construction of a linear map $\mathcal{L}_T : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ which could be a pseudodistribution as described above. (We discuss in detail why naïve constructions of such maps fail in a similar context in Chapter 11 so we will not duplicate the effort here.) The pseudocalibration technique enters the picture here: it leverages the *non*-existence of successful $O(d)$-simple statistics to construct a good pseudoexpectation candidate.

The second difficulty is proving that this candidate $\mathcal{L}$ is in fact a pseudoexpectation (with high probability over T). We associate to $\mathcal{L}_T$ a *moment matrix* $M_T \in \mathbb{R}^{n^d \times n^d}$ – the goal will be to show that with high probability of T, $M_T \succeq 0$. M_T is a random matrix, and its entries are degree $O(d)$ polynomials in the entries of the tensor T.

Showing that $M_T \succeq 0$ is a technical challenge, for two reasons. First, its entries are not independent random variables, so standard theorems on the its spectrum do not apply, and neither do the standard matrix concentration tools we use for Theorem 6.1.1 and Theorem 6.1.3. Second, proving that $M \succeq 0$ requires controlling *all* of the eigenvalues of M, while most matrix concentration tools are

designed only to control the maximum eigenvalue of M.

In the remainder of the chapter, we prove Theorem 6.1.1, Theorem 6.1.3 and Theorem 6.1.4. First, to give some context for the proofs of Theorems 6.1.1 and 6.1.3 we describe some straightforward but suboptimal refutation algorithms for the spiked tensor model.

6.2.1 Elementary Refutation Algorithms for Spiked Tensors

Focusing on the refutation problem and the case $k = 3$, we describe some elementary polynomial-time refutation algorithms for the spiked tensor model, as baselines for comparison. The goal is, given a Gaussian random 3-tensor T, output an upper bound on $\max_{\|x\|=1} \langle T, x^{\otimes 3} \rangle$.

The 2-norm bound As a first attempt, observe that for unit vectors $x \in \mathbb{R}^n$,

$$\langle T, x^{\otimes 3} \rangle \leqslant \|T\| \cdot \|x\|^3 = \left(\sum_{ijk \in [n]} T_{ijk}^2 \right)^{1/2},$$

and the 2-norm of T is clearly polynomial-time computable. For T with $\mathcal{N}(0,1)$ entries, $\|T\| \approx n^{3/2}$ with high probability, so this gives an $n^{3/2}$-refutation algorithm. Since $\langle T, x^{\otimes 3} \rangle^2 \preceq \|T\|^2 \|x\|^6$, this bound is certifiable by degree-6 SoS.

The spectral bound An *spectral* improvement on the 2-norm bound is possible as follows.

$$\langle T, x^{\otimes 3} \rangle \leqslant \max_{y \in \mathbb{R}^n, z \in \mathbb{R}^{n^2}} \frac{y^\top T z}{\|y\| \|z\|} = \sigma_{\max}(T)$$

where in the second expression we abuse notation and use T also for some $n \times n^2$ matrix "flattening" of T. For T with $\mathcal{N}(0,1)$ entries, the maximum singular value

of this matrix is $\Theta(n)$ with high probability [175]. The spectral upper bound is also certifiable by constant degree SoS proofs.

For even k, an analogous spectral bound does obtain the guarantees of Theorem 6.3.1 [159]. The failure of this bound for odd k is because of the fact that the maximum singular value of a random rectangular matrix scales (to first order) as the square root of the longer dimension [175].

For 3-tensors neither 2-norm nor the spectral bound matches the $n^{3/4}(\log n)^{O(1)}$-certification we eventually show is achieved by constant-degree SoS.

6.3 SoS Algorithms for Spiked Tensors

In this section we prove Theorem 6.1.1. For brevity we prove the case $k = 3$; for details concerning larger k we refer the reader to [92].

6.3.1 Refutation

We start with refutation.

Theorem 6.3.1 (Formal version of Theorem 6.1.1, refutation)**.** *Let $T \in (\mathbb{R}^n)^{\otimes 3}$ have iid entries from $\mathcal{N}(0,1)$. Then $\vdash_6 \langle T, x^{\otimes 3}\rangle^2 \leqslant O(n^{3/2}(\log n)^{1/2}) \cdot \|x\|^6$ with probability $1 - o(1)$.*

As a corollary of Theorem 6.3.1, if T is drawn from the *symmetrized* Gaussian distribution (so e.g. $T_{ijk} = T_{kji}$) the conclusion of the theorem still holds, because

such T can be decomposed as a sum $T = \sum_{\pi \in S_3} \pi \cdot A + E$, where A is a tensor with iid entries from $\mathcal{N}(0, 1/|S_3|)$ and E is a sparse error tensor. The details may be found in [92].

Remark 6.3.2 (Connection to simple statistics). The statement of Theorem 6.3.1 says that with high probability over T as in Theorem 6.3.1 there exist polynomials $q_1, \ldots, q_m$ with

$$\|x\|^6 O(n^{3/2}(\log n)^{1/2}) - \langle T, x^{\otimes 3}\rangle^2 = \sum_{i \in [m]} q_i(x)^2$$

where $\deg_x q_i(x)^2 \leqslant 6$. If the left side of the above is expanded in the monomial basis in x, each coefficient is an $O(1)$-degree function in T. Our proof of Theorem 6.3.1 will establish that each $q_i(x)$ can be taken to depend on T in the following way: $q_i(x) = \langle v_i, x^{\otimes 3}\rangle$, where $v_i = v_i(T)$ is an eigenvector of a matrix whose entries are $O(1)$-degree functions of T; that is, up to normalization, a matrix of simple statistics.

We will use the following lemmas to prove Theorem 6.3.1.

Lemma 6.3.3. *Let $M \in \mathbb{R}^{n \times n}$ have iid entries from $\mathcal{N}(0, 1)$. Then $\vdash_4 \mathbb{E}_M \langle x, Mx\rangle^2 \leqslant \|x\|^4$.*

Proof. By expanding the polynomial $\langle x, Mx\rangle^2$, we find

$$\mathop{\mathbb{E}}_M \langle x, Mx\rangle^2 = \sum_{ijkl \in [n]} (\mathbb{E}\, M_{ij} M_{kl}) \cdot x_i x_j x_k x_\ell = \sum_{ij} x_i^2 x_j^2 = \|x\|^4$$

because $\mathbb{E}\, M_{ij} M_{kl} = 0$ unless $ij = kl$, and in that case $\mathbb{E}\, M_{ij}^2 = 1$. □

Lemma 6.3.4. *Let $M_1, \ldots, M_n$ be $n \times n$ matrices, each with iid entries from $\mathcal{N}(0, 1)$. Then*

$$\left\| \sum_{i \in [n]} M_i \otimes M_i - \mathbb{E} \sum_{i \in [n]} M_i \otimes M_i \right\| \leqslant O(n^{3/2}(\log n)^{1/2})$$

with probability $1 - o(1)$, where $\|\cdot\|$ denotes the spectral norm.

Lemma 6.3.4 follows from standard matrix concentration arguments (in this case the matrix Bernsetein inequality), which can be found in Section 6.7.

Proof of Theorem 6.3.1. Let $T_i \in \mathbb{R}^{n\times n}$ denote the i-th matrix *slice* of the tensor T; that is $(T_i)_{jk} = T_{ijk}$. By SoS Cauchy-Schwarz (Lemma 5.0.6),

$$\vdash_6 \langle T, x^{\otimes 3}\rangle^2 = \left(\sum_{i\in[n]} x_i\langle x, T_i x\rangle\right)^2 \leqslant \left(\sum_{i\in[n]} x_i^2\right)\left(\sum_{i\in[n]}\langle x, T_i x\rangle^2\right) = \|x\|^2 \sum_{i\in[n]}\langle x, T_i x\rangle^2 .$$

By Lemma 6.3.3,

$$\vdash_4 \sum_{i\in[n]}\langle x, T_i x\rangle^2 = n\|x\|^4 + \left\langle x^{\otimes 2}\left(\sum_{i\in[n]} T_i\otimes T_i - \mathbb{E}\, T_i\otimes T_i\right)x^{\otimes 2}\right\rangle .$$

And by Lemma 6.3.4 together with Lemma 5.0.7,

$$\begin{aligned}\vdash_4 \sum_{i\in[n]}\left\langle x^{\otimes 2}\left(\sum_{i\in[n]} T_i\otimes T_i - \mathbb{E}\, T_i\otimes T_i\right)x^{\otimes 2}\right\rangle &\leqslant \|x\|^4\cdot\left\|\sum_{i\in[n]} T_i\otimes T_i - \mathbb{E}\, T_i\otimes T_i\right\| \\ &\leqslant O(n^{3/2}(\log n)^{1/2})\cdot\|x\|^4\end{aligned}$$

with high probability, so putting the inequalities together finishes the proof. □

6.3.2 Estimation

Although estimation and refutation are formally of incomparable difficulty, refutation algorithms which use convex programs can generally be transformed into estimation algorithms, as in this estimation algorithm for the single-spike tensor model.

Algorithm 6.3.5 (Single-spike tensor model, estimation)**.** On input a 3-tensor $T \in (\mathbb{R}^n)^{\otimes 3}$, using semidefinite programming, find the degree-6 pseudo-distribution $\{x\}$ satisfying $\{\|x\|^2 = 1\}$ which maximizes $\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle$. Output $\tilde{\mathbb{E}}\, x/\|\tilde{\mathbb{E}}\, x\|$.

Theorem 6.3.6 (Formal version of Theorem 6.1.1, estimation)**.** *Let $T' \in (\mathbb{R}^n)^{\otimes 3}$ have iid entries from $\mathcal{N}(0,1)$ and let v be a unit vector. Suppose $\lambda \geqslant n^{3/4}(\log n)^{1/4}/\varepsilon$. With probability $1 - o(1)$ over T', on input $T = \lambda v^{\otimes 3} + T'$, Algorithm 6.3.5 returns a unit vector v' with $\langle v, v' \rangle \geqslant 1 - O(\varepsilon)$.*

Proof of Theorem 6.3.6. By Theorem 6.3.1, we may assume that for every degree-6 pseudoexpectation $\tilde{\mathbb{E}}$ which satisfies $\{\|x\|^2 = 1\}$ it holds that $\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle \leqslant O(n^{3/4}(\log n)^{1/4})$. Let $\tilde{\mathbb{E}}$ denote the pseudoexpectation found by Algorithm 6.3.5. Since it is maximal,

$$\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle \geqslant \langle T, v^{\otimes 3}\rangle = \lambda + \langle T', v^{\otimes 3}\rangle \geqslant \lambda - O(n^{3/4}(\log n)^{1/4})$$

since the point-mass distribution with all mass on v is itself a pseudodistribution. On the other hand,

$$\tilde{\mathbb{E}}\langle T, x^{\otimes 3}\rangle = \lambda\, \tilde{\mathbb{E}}\langle v, x\rangle^3 + \tilde{\mathbb{E}}\langle T', x^{\otimes 3}\rangle \leqslant \lambda\, \tilde{\mathbb{E}}\langle v, x\rangle^3 + O(n^{3/4}(\log n)^{1/4})\,.$$

Putting these together and rearranging,

$$\tilde{\mathbb{E}}\langle v, x\rangle^3 \geqslant 1 - \frac{O(n^{3/4}(\log n)^{1/4}}{\lambda} \geqslant 1 - O(\varepsilon)\,.$$

By Lemma 10.0.3, which relates $\langle v, x\rangle^3$ to $\langle v, x\rangle$,

$$\langle v, \tilde{\mathbb{E}}\, x\rangle \;=\; \tilde{\mathbb{E}}\langle v, x\rangle \;\geqslant\; 1 - O(2\varepsilon) = 1 - O(\varepsilon)\,.$$

Since $\|\tilde{\mathbb{E}}\, x\| \leqslant (\tilde{\mathbb{E}}\, \|x\|^2)^{1/2} = 1$, this completes the proof. □

6.4 Spectral Algorithms for Spiked Tensors

In this section we prove Theorem 6.1.3. The key idea behind the proof is to use the fact that the dual certificates used in the last section are actually explict, low

degree matrix-valued functions of an input tensor T. By evaluating the spectra of these functions directly we design algorithms with guarantees matching those achievable by polynomial-time SoS, but without incurring the running-time cost of a generic SDP solver.

The catch, in the case of estimation, is that the dual certificate constructions from the previous section apply only to T from the null model H_0. It will take some additional analysis to show that the same matrix-valued functions also allow us to perform estimation when T is from the alternative/spiked model H_1.

Running times for this style of algorithm hit a natural barrier at the time required for spectral computationss on matrices of dimensions matching those of the dual SDP certificates which the algorithm constructs. Since the algorithm in Theorem 6.3.1 uses a degree-6 SoS SDP, we might not expect to beat the time required for a matrix-vector multiply in n^3 dimensions. For our refutation algorithm we will be able to use an $n^2 \times n^2$ matrix rather than an $n^3 \times n^3$ one, because although we have used degree-6 SoS in Theorem 6.3.1, the heart of the SoS proof is really degree-4. This leads to the $\tilde{O}(n^4)$ running time for refutation.

To obtain a linear-time algorithm for estimation, we construct a smaller matrix-valued polynomial whose spectrum is not useful for refutation, but is nonetheless useful for estimation. While this matrix is not difficult to construct for the spiked tensor model, in the next chapter we will see another estimation problem (the *planted sparse vector* problem) where an analogous but less obvious construction is needed.

We prove Theorem 6.1.3 in two parts.

Proof of Theorem 6.1.3, refutation. Let $M = \mathbb{E}\, A \otimes A$, where A is an $n \times n$ matrix

with iid entries from $\mathcal{N}(0,1)$. Since the SoS proof system is sound, the proof of Theorem 6.3.1 also shows that for any tensor T,

$$\max_{\|x\|=1} \langle T, x^{\otimes 3} \rangle \leqslant \left(n + \left\| \sum_{i \in [n]} T_i \otimes T_i - n \cdot M \right\| \right)^{1/2}$$

and that the latter quantity is at most $O(n^{3/4}(\log n)^{1/4})$ with probability $1 - o(1)$ for random T. Since constructing the matrix $\sum_{i\in[n]} T_i \otimes T_i - n \cdot M$ can be done in time $O(n^4)$ and is spectral norm can be computed in time $O(n^4(\log n)^{O(1)})$, the proof is complete. □

Linear-Time Algorithm Spiked Tensor Estimation

Algorithm 6.4.1. Input: $\mathbf{T} = \lambda \cdot v^{\otimes 3} + \mathbf{A}$.

- Compute the partial trace $M := \mathrm{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i = \sum_i \mathrm{Tr}(T_i) \cdot T_i \in \mathbb{R}^{n \times n}$, where T_i are the first-mode slices of $\mathbf{T}$.
- Output the top eigenvector v' of M with sign chosen such that $\langle T, (v')^{\otimes 3} \rangle \geqslant 0$.

Theorem 6.4.2. *When* $\mathbf{A}$ *has iid standard Gaussian entries and* $\lambda \geqslant n^{3/4} \log n/\varepsilon$, *Theorem 6.4.1 recovers* v' *with* $\langle v, v' \rangle \geqslant 1 - O(\varepsilon^2 + n^{-1/4}\varepsilon)$ *with high probability over* $\mathbf{A}$.

Theorem 6.4.3. *Theorem 6.4.1 can be implemented in linear time and sublinear space.*

Theorem 6.4.2 is proved by routine matrix concentration.

To implement the algorithm in linear time it is enough to show that the (sublinear-sized) matrix M has constant spectral gap; then a standard application of the matrix power method computes the top eigenvector.

Lemma 6.4.4. *For any v, with high probability over $\mathbf{A}$, the following occur:*

$$\left\| \sum_i \operatorname{Tr}(A_i) \cdot A_i \right\| \leqslant O(n^{3/2} \log^2 n)$$
$$\left\| \sum_i v(i) \cdot A_i \right\| \leqslant O(\sqrt{n} \log n)$$
$$\left\| \sum_i \operatorname{Tr}(A_i) v(i) \cdot vv^\top \right\| \leqslant O(\sqrt{n} \log n)\,.$$

All the matrices in the lemma are sums of independent matrices, so the proof is routine applications of the Matrix Bernstein inequality Theorem 5.0.8; we omit it for brevity.

Proof of Theorem 6.4.2. We expand the partial trace $\operatorname{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i$.

$$\begin{aligned}
\operatorname{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i &= \sum_i \operatorname{Tr}(T_i) \cdot T_i \\
&= \sum_i \operatorname{Tr}(\lambda \cdot v(i) vv^\top + A_i) \cdot (\lambda \cdot v(i) vv^\top + A_i) \\
&= \sum_i (\lambda v(i) \|v\|^2 + \operatorname{Tr}(A_i)) \cdot (\lambda \cdot v(i) vv^\top + A_i) \\
&= \lambda^2 vv^\top + \lambda \left(\sum_i v(i) \cdot A_i + \sum_i \operatorname{Tr}(A_i) v(i) vv^\top \right) + \sum_i \operatorname{Tr}(A_i) \cdot A_i\,.
\end{aligned}$$

Applying Lemma 6.4.4 and the triangle inequality, we see that

$$\begin{aligned}
&\left\| \lambda \left(\sum_i v(i) \cdot A_i + \sum_i \operatorname{Tr}(A_i) v(i) vv^\top \right) + \sum_i \operatorname{Tr}(A_i) \cdot A_i \right\| \\
&\leqslant O(n^{3/2} (\log n)^2 + \lambda \sqrt{n} \log n)
\end{aligned}$$

with high probability. Thus, for $\lambda = n^{3/4} \log n / \varepsilon$, the matrix M/λ^2 satisfies

$$\frac{M}{\lambda^2} = vv^\top + E$$

where $\|E\| \leqslant O(\varepsilon^2 + n^{-1/4}\varepsilon)$, and the result follows by standard manipulations. □

Proof of Theorem 6.4.3. Carrying over the expansion of the partial trace from above, the matrix $\mathrm{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i$ has spectral gap $\Omega(1/\varepsilon)$ and so the matrix power method finds the top eigenvector in $O(\log(n/\varepsilon))$ iterations. This matrix has dimension $n \times n$, so a single iteration takes $O(n^2)$ time, which is sublinear in the input size n^3. Finally, to construct $\mathrm{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i$ we use

$$\mathrm{Tr}_{\mathbb{R}^n} \sum_i T_i \otimes T_i = \sum_i \mathrm{Tr}(T_i) \cdot T_i$$

and note that to construct the right-hand side it is enough to examine each entry of $\mathbf{T}$ just $O(1)$ times and perform $O(n^3)$ additions. At no point do we need to store more than $O(n^2)$ matrix entries at the same time. □

6.5 Spiked Tensors and Simple Statistics

In this section we prove Theorem 6.1.4. To prove the theorem we need a combinatorial fact.

Fact 6.5.1. *Let $\mathcal{H}_{t,s,D}$ be the number of 3-uniform multi-hypergraphs on n vertices with exactly t hyperedges, at most D* unique *hyperedges, all even-degree nodes, and exactly s nodes of nonzero degree. Then (1) if $s > 3t/2$ or $s > 3D$ we have $\mathcal{H}_{t,s} = 0$, (2) for all s, t we have $\mathcal{H}_{s,t} \leqslant n^s t^{3s}$, and (3) $\mathcal{H}_{t,3t/2} \geqslant t^{\Omega(t)}$*

Proof. We start with (1) and (2). To choose such a hypergraph, we first specify s nodes. There are at most n^s choices. Then there are at most t^{3s} choices of t hyperedges. Furthermore, the even-degree requirement ensures that every node of nonzero degree has degree at least 2, so there can be at most $3t/2$ nodes of nonzero degree.

Finally, (3) follows from standard estimates on the number of t-edge graphs.

□

Proof of Theorem 6.1.4. We recall that the Hermite polynomials are an orthonormal basis for the square-integrable functions of a multivariate Gaussian. (For background on the Hermite polynomials, see e.g. [142].) Let U_D be the set of multi-indices α over $\binom{n}{3}$ such that the Hermite polynomial H_α has coordinate degree at most D. Equivalently, U_D are the multi-indices α over $[n]^3$ such that at most D entries of α are nonzero.

Using (2.3.1) from Chapter 2, we just need to compute

$$\sum_{\alpha \in U_D, \alpha \neq 0} \left(\mathop{\mathbb{E}}_{T \sim \mu} H_\alpha(T) \right)^2 . \tag{6.5.1}$$

It is an elementary fact about univariate Hermite polynomials h_i that if $g \sim \mathcal{N}(0,1)$, for every $c \in \mathbb{R}$ we have $\mathbb{E}\, h_i(g + c) = c^i$. It follows that

$$\mathop{\mathbb{E}}_{T \sim \mu} H_\alpha(T) = \left(\frac{\lambda}{n^{3/2}} \right)^{|\alpha|}$$

if α, viewed as a 3-uniform multi-hypergraph, has all even vertex degrees, and otherwise $\mathbb{E}_{T \sim \mu} H_\alpha(T) = 0$. Here $|\alpha| = \sum_{\beta \in \binom{n}{3}} \alpha_\beta$ is equivalently the total number of hyperedges in α.

What is the contribution to the sum in (6.5.1) from α with $|\alpha| = t$? Such α's are in one-to-one correspondence with 3-uniform multi-hypergraphs on n vertices with exactly t hyperedges, all of whose nodes have even degree and which have at most D nodes of nonzero degree. Let $\mathcal{H}_{t,s}$ be the number of such multi-hypergraphs where exactly s nodes have nonzero degree. Then

$$(6.5.1) = \sum_{t=1}^{\infty} \left(\frac{\lambda}{n^{3/2}} \right)^{2t} \cdot \sum_{s \leqslant 3D} \mathcal{H}_{t,s} ,$$

where we have used Fact 6.5.1 to remove the terms for $s > 3D$. By Fact 6.5.1,

$$\begin{aligned}
\sum_{t=1}^{\infty}\left(\frac{\lambda}{n^{3/2}}\right)^{2t}\cdot\sum_{s\leqslant 3D}\mathcal{H}_{t,s} &\leqslant \sum_{t=1}^{3D}\left(\frac{\lambda}{n^{3/2}}\right)^{2t}\cdot\sum_{s\leqslant 3t/2} n^s t^{3s} + \sum_{t=3D+1}^{\infty}\left(\frac{\lambda}{n^{3/2}}\right)^{2t}\cdot\sum_{s\leqslant 3D} n^s t^{3s} \\
&\leqslant \sum_{t=1}^{3D}\left(\frac{\lambda}{n^{3/2}}\right)^{2t}\cdot t^{O(t)} n^{3t/2} + \sum_{t=3D+1}^{\infty}\left(\frac{\lambda}{n^{3/2}}\right)^{2t} 3D n^{3t/2} t^{3D} \\
&\leqslant \sum_{t=1}^{3D}\left(\frac{\lambda^2}{n^{3/2}}\right)^{t}\cdot t^{O(t)} + D\cdot\sum_{t=3D+1}^{\infty}\left(\frac{\lambda^2}{n^{3/2}}\right)^{t}\cdot t^{3D} \\
&\leqslant \sum_{t=1}^{3D}\left(\frac{\lambda^2 D}{n^{3/2}}\right)^{t} + 3D\cdot\sum_{t=3D+1}^{\infty}\left(\frac{\lambda^2}{n^{3/2}}\right)^{t}\cdot t^{3D}
\end{aligned}$$

Standard manipulations show that the above is $O(1)$ so long as $\frac{\lambda^2}{n^{3/2}} \leqslant D/(\log D)^{O(1)}$. The lower bound follows along similar lines, using (3) from Fact 6.5.1. □

6.6 SoS Lower Bounds for Spiked Tensors

In this section we give an overview of the proof of Theorem 6.1.2. The techniques involved in proving the main lemmas are almost technically identical to to prove nearly-tight SoS lower bounds for planted clique [29]: these arguments are presented in Chapter 11.

To state a formal version of Theorem 6.1.2 PCA it is useful to define a Boolean version of the problem. For technical convenience we actually prove an SoS lower bound for this problem; then standard techniques (see [88]) allow us to prove the main theorem for Gaussian tensors.

Problem 6.6.1 (Spiked k-Tensor, signal-strength λ, boolean version)**.** Distinguish the following two distributions on $\Omega_k \stackrel{\text{def}}{=} \{\pm 1\}^{\binom{n}{k}}$.

- *the uniform distribution ν:* $A \sim \Omega$ chosen uniformly at random.
- *the planted distribution μ:* Choose $v \sim \{\pm 1\}^n$ and let $B = v^{\otimes k}$. Sample A by rerandomizing every coordinate of B with probability $1 - \lambda n^{-k/2}$.

We show that the natural SoS relaxation of this problem suffers from a large integrality gap, when λ is slightly less than $n^{k/4}$, even when the degree of the SoS relaxation is $n^{\Omega(1)}$. (When $\lambda \gg n^{k/4-\varepsilon}$, algorithms with running time $2^{n^{O(\varepsilon)}}$ are known for $k = O(1)$.

Theorem 6.6.2 (Formal version of Theorem 6.1.2)**.** *Let $k = O(1)$. For $A \in \Omega_k$, let*

$$SoS_d(A) \stackrel{\text{def}}{=} \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle x^{\otimes k}, A\rangle \text{ s.t. } \tilde{\mathbb{E}} \text{ is a degree-}d \text{ pseudoexpectation satisfying } \{\|x\|^2 = 1\}\,.$$

There is a constant c so that for every small enough $\varepsilon > 0$, if $d \leqslant n^{c\cdot\varepsilon}$, then for large enough n,

$$\mathop{\mathbb{P}}_{A\sim\Omega}\{SoS_d(A) \geqslant n^{k/4-\varepsilon}\} \geqslant 1 - o(1)$$

and

$$\mathop{\mathbb{E}}_{A\sim\Omega} SoS_d(A) \geqslant n^{k/4-\varepsilon}\,.$$

Moreover, the latter also holds for A with iid entries from $\mathcal{N}(0,1)$.[3]

To prove the theorem we will exhibit for a typical sample A from the uniform distribution a degree $n^{\Omega(\varepsilon)}$ pseudodistribution $\tilde{\mathbb{E}}$ which satisfies $\{\|x\|^2 = 1\}$ and has $\tilde{\mathbb{E}}\langle x^{\otimes k}, A\rangle \geqslant n^{k/4-\varepsilon}$. The following lemma ensures that the pseudodistribution we exhibit will be PSD. The function $\Lambda^{\leqslant D}$ in the lemma *is the pseudocalibrated candidate pseudoexpectation.*

Lemma 6.6.3. *Let $d \in \mathbb{N}$ and let $N_d = \sum_{s\leqslant d} n(n-1)\cdots(n-(s-1))$ be the number of $\leqslant d$-tuples with unique entries from $[n]$. There is a constant ε^* independent of n such*

[3]For technical reasons we do not prove a tail bound type statement for Gaussian A, but we conjecture that this is also true.

that for any $\varepsilon < \varepsilon^*$ also independent of n, the following is true. Let $\lambda = n^{k/4-\varepsilon}$. Let $\mu(A)$ be the density of the following distribution (with respect to the uniform distribution on $\Omega = \{\pm 1\}^{\binom{n}{k}}$).

The Planted Distribution: *Choose $v \sim \{\pm 1\}^n$ uniformly. Let $B = v^{\otimes k}$. Sample A by*

- *replacing every coordinate of B with a random draw from $\{\pm 1\}$ independently with probability $1 - \lambda n^{-k/2}$,*
- *then choosing a subset $S \subseteq [n]$ by including every coordinate with probability $n^{-\varepsilon}$,*
- *then replacing every entry of B with some index outside S independently with a uniform draw from $\{\pm 1\}$.*

Let $\Lambda : \Omega \to \mathbb{R}^{N_d \times N_d}$ be the following function

$$\Lambda(A) = \mu(A) \cdot \mathop{\mathbb{E}}_{v|A} v^{\otimes \leqslant 2d}$$

Here we abuse notation and denote by $x^{\leqslant \otimes 2d}$ the matrix indexed by tuples of length $\leqslant d$ with unique entries *from $[n]$. For $D \in \mathbb{N}$, let $\Lambda^{\leqslant D}$ be the projection of Λ into the degree-D real-valued polynomials on $\{\pm 1\}^{\binom{n}{k}}$. There is a universal constant C so that if $Cd/\varepsilon < D < n^{\varepsilon/C}$, then for large enough n*

$$\mathop{\mathbb{P}}_{A \sim \Omega} \{\Lambda^{\leqslant D}(A) \succeq 0\} \geqslant 1 - o(1)\,.$$

For a tensor A, the moment matrix of the pseudodistribution we exhibit will be $\Lambda^{\leqslant D}(A)$. We will need it to satisfy the constraint $\{\|x\|^2 = 1\}$. This follows from the following general lemma. (The lemma is much more general than what we state here, and uses only the vector space structures of space of real matrices and matrix-valued functions.)

Lemma 6.6.4. *Let $k \in \mathbb{N}$. Let V be a linear subspace of $\mathbb{R}^{N \times M}$. Let $\Omega = \{\pm 1\}^{\binom{n}{k}}$. Let $\Lambda : \Omega \to V$. Let $\Lambda^{\leqslant D}$ be the entrywise orthogonal projection of Λ to polynomials of degree at most D. Then for every $A \in \Omega$, the matrix $\Lambda^{\leqslant D}(A) \in V$.*

Proof. The function Λ is an element of the vector space $\mathbb{R}^{N \times M} \otimes \mathbb{R}^{\Omega}$. The projection $\Pi_V : \mathbb{R}^{N \times M} \to V$ and the projection $\Pi_{\leqslant D}$ from $\mathbb{R}^{\Omega}$ to the degree-D polynomials commute as projections on $\mathbb{R}^{N \times M} \otimes \mathbb{R}^{\Omega}$, since they act on separate tensor coordinates. It follows that $\Lambda^{\leqslant D} \in V \otimes (\mathbb{R}^{\Omega})^{\leqslant D}$ takes values in V. □

Last, we will require a couple of scalar functions of $\Lambda^{\leqslant D}$ to be well concentrated.

Lemma 6.6.5. *Let $\Lambda, d, \varepsilon, D$ be as in Lemma 6.6.3. The function $\Lambda^{\leqslant D}$ satisfies*

- $\mathbb{P}_{A \sim \Omega}\{\Lambda^{\leqslant D}_{\emptyset,\emptyset}(A) = 1 \pm o(1)\} \geqslant 1 - o(1)$ *(Here $\Lambda_{\emptyset,\emptyset} = 1$ is the upper-left-most entry of Λ.)*
- $\mathbb{P}_{A \sim \Omega}\{\langle \Lambda^{\leqslant D}(A), A \rangle = (1 \pm o(1)) \cdot n^{3k/4-\varepsilon}\} \geqslant 1 - o(1)$ *(Here we are abusing notation to write $\langle \Lambda^{\leqslant D}(A), A \rangle$ for the inner product of the part of $\Lambda^{\leqslant D}$ indexed by monomials of degree k and A.)*

The Boolean case of Theorem 6.6.2 follows from combining the lemmas. The Gaussian case can be proved in a black-box fashion from the Boolean case by the following proposition.

Proposition 6.6.6. *Let $k \in \mathbb{N}$ and let $A \sim \{\pm 1\}^{\binom{n}{k}}$ be a symmetric random Boolean tensor. Suppose that for every $A \in \{\pm 1\}^{\binom{n}{k}}$ there is a degree-d pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\{\|x\|^2 = 1\}$ such that*

$$\mathop{\mathbb{E}}_{A} \tilde{\mathbb{E}} \langle x^{\otimes k}, A \rangle = C \,.$$

Let $T \sim \mathcal{N}(0,1)^{\binom{n}{k}}$ be a Gaussian random tensor. Then

$$\mathop{\mathbb{E}}_{T} \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}} \langle x^{\otimes k}, T \rangle \geqslant \Omega(C)$$

where the maximization is over pseudodistributions of degree d which satisfy $\{\|x\|^2 = 1\}$.

Proof. For a tensor $T \in (\mathbb{R}^n)^{\otimes k}$, let $A(T)$ have entries $A(T)_\alpha = \text{sign}(T_\alpha)$. Now consider

$$\mathop{\mathbb{E}}_{T} \tilde{\mathbb{E}}_{A(T)} \langle x^{\otimes k}, T \rangle = \sum_{\alpha} \mathop{\mathbb{E}}_{T} \tilde{\mathbb{E}}_{A(T)} \, x^\alpha T_\alpha$$

where α ranges over multi-indices of size k over $[n]$. We rearrange each term above to

$$\mathop{\mathbb{E}}_{A(T)} (\tilde{\mathbb{E}}_{A(T)} \, x^\alpha) \cdot \mathop{\mathbb{E}}_{T_\alpha \,|\, A(T)} T_\alpha = \mathop{\mathbb{E}}_{A(T)} (\tilde{\mathbb{E}}_{A(T)} \, x^\alpha) \cdot A(T)_\alpha \cdot \mathbb{E}\,|g|$$

where $g \sim \mathcal{N}(0,1)$. Since $\mathbb{E}\,|g|$ is a constant independent of n, all of this is

$$\Omega(1) \cdot \sum_{\alpha} \mathop{\mathbb{E}}_{A} \tilde{\mathbb{E}}_{A} \, x^\alpha \cdot A_\alpha = C \,. \qquad \square$$

6.7 Matrix Concentration Bounds for Spiked Tensors

Theorem 6.7.1 (Restatement of Lemma 6.3.4)**.** *Let $c \in \{1,2\}$ and $d \geqslant 1$ an integer. Let $A_1, \ldots, A_{n^c}$ be iid random matrices in $\{\pm 1\}^{n^d \times n^d}$ or with independent entries from $\mathcal{N}(0,1)$. Then, with probability $1 - O(n^{-100})$,*

$$\left\| \sum_{i \in [n^c]} A_i \otimes A_i - \mathbb{E}\, A_i \otimes A_i \right\| \lesssim \sqrt{d}\, n^{(2d+c)/2} \cdot (\log n)^{1/2} \,.$$

Now we prove Theorem 6.7.1. Let $A_1, \ldots, A_{n^c}$ be as in Theorem 6.7.1. We first need to get a handle on their norms individually, for which we need the following lemma.

Lemma 6.7.2. *Let A be a random matrix in $\{\pm 1\}^{n^d \times n^d}$ or with independent entries from $\mathcal{N}(0,1)$. For all $t \geqslant 1$, the probability of the event $\{\|A\| > t n^{d/2}\}$ is at most $2^{-t^2 n^d / K}$ for some absolute constant K.*

Proof. The subgaussian norm of the rows of A is constant and they are identically and isotropically distributed. Hence Theorem 5.39 of [175] applies to give the result. □

Since the norms of the matrices $A_1, \dots, A_{n^c}$ are concentrated around $n^{d/2}$ (by Theorem 6.7.2), it will be enough to prove Theorem 6.7.1 after truncating the matrices $A_1, \dots, A_{n^c}$. For $t \geqslant 1$, define iid random matrices $A'_1, \dots, A'_{n^c}$ such that

$$A'_i \stackrel{\text{def}}{=} \begin{cases} A_i & \text{if } \|A_i\| \leqslant t n^{d/2}, \\ 0 & \text{otherwise} \end{cases}$$

for some t to be chosen later. Theorem 6.7.2 allows us to show that the random matrices $A_i \otimes A_i$ and $A'_i \otimes A'_i$ have almost the same expectation. For the remainder of this section, let K be the absolute constant from Theorem 6.7.2.

Lemma 6.7.3. *For every $i \in [n^c]$ and all $t \geqslant 1$, the expectations of $A_i \otimes A_i$ and $A'_i \otimes A'_i$ satisfy*

$$\left\| \mathbb{E}[A_i \otimes A_i] - \mathbb{E}[A'_i \otimes A'_i] \right\| \leqslant O(1) \cdot 2^{-t n^d / K}.$$

Proof. Using Jensen's inequality and that $A_i = A'_i$ unless $\|A_i\| > t n^{d/2}$, we have

$$\begin{aligned} \| \mathbb{E}\, A_i \otimes A_i - A'_i \otimes A'_i \| &\leqslant \mathbb{E}\, \|A_i \otimes A_i - A'_i \otimes A'_i\| \quad \text{Jensen's inequality} \\ &= \int_{t n^{d/2}}^{\infty} \mathbb{P}(\|A_i\| \geqslant \sqrt{s})\, ds \quad \text{since } A_i = A'_i \text{ unless } \|A_i\| \geqslant t n^{d/2} \\ &\leqslant \int_{t n^{d/2}}^{\infty} 2^{-s/K}\, ds \quad \text{by Theorem 6.7.2} \end{aligned}$$

$$\leqslant \sum_{i=0}^{\infty} 2^{-tn^{d/2}/K} \cdot 2^{-i/K} \quad \text{discretizing the integral}$$

$$= O(2^{-tn^{d/2}/K}) \quad \text{as desired.} \qquad \square$$

Lemma 6.7.4. *Let $B'_1, \dots, B'_{n^c}$ be i.i.d. matrices such that $B'_i = A'_i \otimes A'_i - \mathbb{E}[A'_i \otimes A'_i]$. Then for every $C \geqslant 1$ with $C \leqslant 3t^2 n^{c/2}$,*

$$\mathbb{P}\left\{\left\|\sum_{i\in[n^c]} B'_i\right\| > C \cdot n^{(2d+c)/2}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{6t^4}\right).$$

Proof. For $R = 2t^2 n^d$, the random matrices $B'_1, \dots, B'_{n^c}$ satisfy $\{\|B'_i\| \leqslant R\}$ with probability 1. Therefore, by the Bernstein bound for non-symmetric matrices [170, Theorem 1.6], Theorem 5.0.8,

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n^c} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-s^2/2}{\sigma^2 + Rs/3}\right),$$

where $\sigma^2 = \max\{\|\sum_i \mathbb{E}\, B'_i (B'_i)^\top\|, \|\sum_i \mathbb{E}(B'_i)^\top B'_i\|\} \leqslant n^c \cdot R^2$. For $s = C \cdot n^{(2d+c)/2}$, the probability is bounded by

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2 \cdot n^{(2d+c)}/2}{4t^4 \cdot n^{2d+c} + 2t^2 C \cdot n^{(4d+c)/2}/3}\right).$$

Since our parameters satisfy $t^2 C \cdot n^{(4d+c)/2}/3 \leqslant t^4 n^{(2d+c)}$, this probability is bounded by

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{6t^4}\right). \qquad \square$$

At this point, we have all components of the proof of Theorem 6.7.1.

Proof of Theorem 6.7.1. By Theorem 6.7.4,

$$\mathbb{P}\left\{\left\|\sum_i A'_i \otimes A'_i - \sum_i \mathbb{E}[A'_i \otimes A'_i]\right\| > C \cdot n^{(2d+c)/2}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right).$$

At the same time, by Theorem 6.7.2 and a union bound,

$$\mathbb{P}\left\{A_1 = A'_1, \ldots, A_n = A'_{n^c}\right\} \geqslant 1 - n^c \cdot 2^{-t^2 n^d/K}.$$

By Theorem 6.7.3 and triangle inequality,

$$\left\|\sum_i \mathbb{E}[A_i \otimes A_i] - \sum_i \mathbb{E}[A'_i \otimes A'_i]\right\| \leqslant n^c \cdot 2^{-tn^d/K}.$$

Together, these bounds imply

$$\mathbb{P}\left\{\left\|\sum_i A_i \otimes A_i - \sum_i \mathbb{E}[A_i \otimes A_i]\right\| > C \cdot n^{(2d+c)/2} + n^c \cdot 2^{-tn^d/K}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right) + n^c \cdot 2^{-t^2 n^d/K}.$$

We choose $t = 1$ and $C = 100\sqrt{2Kd \log n}$ and assume that n is large enough so that $C \cdot n^{(2d+c)/2} \geqslant n^c \cdot 2^{-tn^d/K}$ and $2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right) \geqslant n^c \cdot 2^{-t^2 n^d/K}$. Then the probability satisfies

$$\mathbb{P}\left\{\left\|\sum_i A_i \otimes A_i - \sum_i \mathbb{E}[A_i \otimes A_i]\right\| > 20n^{(2d+c)/2}\sqrt{2Kd \log n}\right\} \leqslant 4n^{-100}. \qquad \square$$

6.8 Chapter Notes

Attributions. The original results presented in this chapter first appeared in [92, 91, 88], joint works with (collectively) Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, Jonathan Shi, and David Steurer.

Tensor principal component analysis and origins of the spiked tensor model. The spiked tensor model was introduced to the computer science literature by Montanari and Richard [159] as a statistical model for *tensor principal component*

analysis. (Consequently in recent literature the problems considered in this chapter often go under the name "tensor PCA".)

Principal component analysis (PCA), the process of identifying a direction of largest possible variance from a matrix of pairwise correlations, is among the most basic tools for data analysis in a wide range of disciplines. In recent years, variants of PCA have been proposed that promise to give better statistical guarantees for many applications. These variants include restricting directions to the nonnegative orthant (nonnegative matrix factorization) or to directions that are sparse linear combinations of a fixed basis ("sparse principal component analysis").

Often we have access to not only pairwise but also higher-order correlations. In this case, analog of PCA is to find a direction with largest possible third moment or other higher-order moment (higher-order PCA or tensor PCA), a problem which amounts to optimizing a random degree-3 or higher polynomial.. The spiked tensor model offers a simple average-case setting in which to study this problem: so long as $\lambda \gg \sqrt{n}$, the spike is the unique maximum of the degree-k random polynomial $\langle T, x^{\otimes k} \rangle$.

Montanari and Richard showed a version of Theorem 6.1.1 which applies only to even k (and offers sub-optimal guarantees for odd k).

Relation to random constraint satisfaction. The spiked and Gaussian tensor models are related to dense random constraint satisfaction problems (CSPs), most closely random k-XOR (sometimes called k-LIN). (See [7, 153] for definitions.) The spiked k-tensor model with SNR λ behaves like random k-XOR with about λ^2 clauses, and all of the theorems stated in this section have analogues for random

k-XOR. Random k-XOR is fundamental among random CSPs in the dense regime (many more clauses than variables) because refutation and estimation algorithms for random/planted k-XOR can be transformed into algorithms for any other random CSP [7].

The proofs of algorithmic results for k-XOR have a similar flavor to the proofs we presented here, although some matrix concentration arguments must be changed to accommodate sparse random matrices. SoS lower bounds for k-XOR and other CSPs can also be proved using pseudocalibration (though this is a reinterpretation of arguments originally constructed with other ideas) [81, 107, 28]) – however, the technical arguments for PSDness of the associated moment matrix differ substantially.

Further related work Random tensors make appearances in fields besides theoretical computer sicnece. As one example, they appear in statistical physics in study of *spherical p-spin models*. While Gaussian tensors (corresponding to our null model H_0 in this chapter) have been studied for some time [129], spiked models have just recently entered consideration [18, 117].

Simultaneously with [92] (in which Theorem 6.1.1 first appears), Barak and Moitra used similar ideas to design an SoS algorithm for a related tensor problem: *tensor completion* [32]. Subsequent papers built (substantially) on these ideas to solve a number of tensor-based inference problems with the SoS method – some ideas from these papers are presented in later chapters of this thesis [91, 150, 134, 77, 121].

The lower bound Theorem 6.1.2, proved in [88] using ideas from [29], was anticipated by a simpler lower bound proved in [92] which rules out only to

degree-4 SoS proofs to improve on $n^{3/4}$-refutation for the spiked tensor model.

CHAPTER 7

DETECTING SPARSE VECTORS

Estimation of sparse structures is a key tool in high-dimensional statistics – the so-called "bet on sparsity" principle suggests that for many high-dimensional estimation tasks, the ground truth to be estimated is sparse in some basis [74]. Leveraging this sparsity has led to some of the most impressive accomplishments of machine learning and signal processing, like compressed sensing and matrix completion. Many algorithms which exploit sparse structure in data employ some form of ℓ_1 regularization, as in LASSO regression, or thresholding technique, as in sparse principal component analysis [62].

In this chapter, we give a spectral algorithm for the *planted sparse vector* problem. Our algorithm has performance guarantees similar to those of the SoS method, which offers the best guarantees for this problem known to be achievable in polynomial time.

For simplicity, we study a hypothesis testing version of the problem. The input is a basis for a d-dimensional subspace V of $\mathbb{R}^n$, and the goal is to test the following two hypotheses:

$H_0 : V$ is a uniformly random d-dimensional subspace.

$H_1 : V$ is the span of a uniformly random $(d-1)$-dimensional subspace and a random vector $v \in \mathbb{R}^n$ with εn nonzero coordinates.

Estimation and refutation versions of this problem are also of interest – we refer the reader to [30, 92] – but the main ideas are captured by our hypothesis testing algorithm.

Like the spectral algorithms for spiked tensors in the last chapter, our algorithm here is based on spectra of matrices whose entries are low-degree polynomials in the problem input (viewed as usual as a matrix or tensor). Together with our spectral algorithms for spiked tensors, this sets it apart from traditional spectral methods, which typically focus on a small number of canonical matrices – adjacency and Laplacian matrices of graphs, for example – whose entries are typically linear functions of problem inputs.

Designing algorithms using these high(er) degree matrix-valued functions expands the algorithm-design possibilities but also introduces new technical challenges, especially because the entries of the resulting random matrices have complex dependencies. As with spiked tensors, we will be guided by an SoS-based algorithm and its analysis. And again we will obtain an algoirithm with nearly linear running time, whose guarantees match or nearly-match those obtained by SoS. The SoS algorithm in question is due to work of Barak, Kelner, and Steurer [30].

7.1 Main Result

Detecting a planted sparse vector problem gets more difficult as the subspace dimension $d = d(n)$ grows by comparison to the ambient dimension n, and as the vector v becomes less sparse. Our algorithm for this problem runs in nearly linear time in the input size, and matches the best guarantees known to be achievable by polynomial-time algorithms, up to a polylogarithmic factor in the subspace dimension [30]. In what follows, we use the notation $g(n) \leqslant \tilde{O}(f(n))$ to denote that there is a constant C such that for every large-enough n, $g(n) \leqslant f(n)(\log(n))^C$.

Theorem 7.1.1 (Planted sparse vector in nearly-linear time)**.** *For every $\varepsilon > 0$ there exists an algorithm which, given d orthonormal vectors $v_1, \ldots, v_d$ in $\mathbb{R}^n$, outputs* NULL *or* ALTERNATIVE*, solving the planted sparse vector hypothesis testing problem correctly with probability $1 - o(1)$ so long as $d \leqslant \sqrt{n}/(\log n)^{O(1)}$. The running time of the algorithm is $\tilde{O}(nd)$.*

7.2 Algorithm Overview

Sparsity, p-norms, and leverage scores We will describe the connections between our algorithm and SoS below, as well as the ideas necessary to obtain near-linear running time. First we describe the algorithm directly and see how simple statistics can leverage sparsity.

Recall that for a vector $v \in \mathbb{R}^n$ and $p \in \mathbb{N}$, the p norm is given by $\|v\|_p = \left(\sum_{i \leqslant n} |v_i|^p\right)^{1/p}$. If v is chosen as a unit vector, $\|v\|_2 = 1$, then its p norms for $p \neq 2$ reveal information about its sparsity. In particular, we extensively use the idea that if v has only εn nonzero entries, then its 4-norm is at least $(\varepsilon n)^{-1}$; compare this to the 4-norm of a random unit vector w in $\mathbb{R}^n$, which is about n^{-1}.

Our algorithm will attempt to detect, given a subspace V of $\mathbb{R}^n$, whether V contains a unit vector v of 4-norm $(\varepsilon n)^{-1}$. But how to access this information with only simple statistics?

The algorithm recieves as input d orthonormal vectors $v_1, \ldots, v_d$ which span a subspace V. Let us first imagine a simpler situation: the algorithm recieves a matrix $V = (v_1, \ldots, v_d)$ where $v_2, \ldots, v_d \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ and v_1 either is also random from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ or is a unit vector with at most εn nonzero entries. (Of

course this hypothesis testing problem is very easy, but we will gain some insight from it nonetheless.)

Our first observation is that the distribution of 2-norms of the *rows* of the matrix V are affected by the presence of the sparse vector v. If v is a typical εn-sparse vector, its nonzero entries have magnitude roughly $\sqrt{\varepsilon n}$. If $v(i)$ is nonzero, then $\|a_i\|^2 \approx \frac{d-1}{n} + \frac{1}{\varepsilon n}$, while if $v(i)$ is zero, $\|a_i\|^2 \approx \frac{d}{n}$. (The quantities $\|a_i\|$ are well studied – in this context they are called *leverage scores*.)

While this difference may seem small, it is enough to noticeably affect the distribution of $\|a_i\|^2$ across $1 \leqslant i \leqslant n$. A direct computation shows that for an appropriate choice of $\alpha \approx d/n$, the polynomial

$$f(V) = \sum_{i \leqslant n} (\|a_i\|^2 - \alpha)\|a_i\|^2$$

has expectation zero and variance $O(d^3/n^3)$ if $v_1 \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id})$, but has expectation $\Omega(\varepsilon n)^{-1}$ if v_1 is εn-sparse. This makes it a successful simple statistic (after normalization) so long as $d \ll n^{1/3}$.

Finally, $f(V)$ depends only on the norms $\|a_i\|^2$, which are invariant under rotations of $\mathrm{Span}\{v_1, \ldots, v_d\}$, so even if given VR for some $d \times d$ orthogonal matrix R this still yields a successful simple statistic. While the existence of a sparse vector is apparent from the matrix V, for, say, random R the matrix VR is almost a random orthonormal basis of $\mathrm{Span}\{v_1, \ldots, v_d\}$. (Eventually we show that any orthogonal basis suffices.)

Our main result improves over this initial simple statistic to tolerate larger subspace dimension – $d \leqslant \sqrt{n}/(\log n)^{O(1)}$ – by employing the top eigenvalue of the matrix $\sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) a_i a_i^{\mathsf{T}}$. This matrix has the advantage that its spectrum can also be used to solve the estimation problem; see [91].

Next, we describe how one might invent the matrix $\sum_{i \leqslant n}(\|a_i\|^2 - \frac{d}{n})a_i a_i^\top$ by studying the SoS algorithm of [30] and its analysis.

From degree-4 SoS to nearly-linear time: compressing matrices with partial traces The SoS algorithm of [30] which guides the development of our algorithm for recovery of planted sparse vectors uses degree-4 SoS – the underlying SDP involves $n^2 \times n^2$ matrices. Our algorithms will avoid the use of semidefinite programming; we will also have to avoid doing even elementary linear-algebraic operations in n^2 dimensions (like matrix-vector multiplication, for example) to obtain a nearly-linear running time. We already did this once in Chapter 6 in designing our nearly-linear time algorithm; now we discuss the technique used to arrive at an $n \times n$ matrix from an $n^2 \times n^2$ one a little further.

In both the spiked tensor model from Chapter 6 and our planted sparse vector algorithm, eventually we obtain a large matrix (suggested by a degree-4 SoS algorithm) which consists of a rank-one spike obscured by random noise. We show that in some situations, this large matrix can be compressed significantly without loss in the signal by applying *partial trace* operations. In these situations, the partial trace yields a smaller, $n \times n$ matrix with the same signal-to-noise ratio as the large matrix suggested by degree-4 SoS, even in situations when lower degree sum-of-squares approaches are known to fail (as for the planted sparse vector and spiked tensor problems).[1]

The partial trace $\mathrm{Tr}_{\mathbb{R}^d} \colon \mathbb{R}^{d^2 \times d^2} \to \mathbb{R}^{d \times d}$ is the linear operator that satisfies $\mathrm{Tr}_{\mathbb{R}^d} A \otimes B = (\mathrm{Tr}\, A) \cdot B$ for all $A, B \in \mathbb{R}^{d \times d}$. To see how the partial trace can be used

[1] For both problems we use matrices with dimensions corresponding to degree-2 SoS programs. An argument of Spielman et al. ([165], Theorem 9) shows that degree-2 sum-of-squares can only find sparse vectors with sparsity $k \leqslant \tilde{O}(\sqrt{n})$, wherease we achieve sparsity as large as $k = \Theta(n)$. For spiked tensors, the degree-2 SoS program cannot even express the objective function.

to compress large matrices to smaller ones with little loss, consider the following problem: Given a matrix $M \in \mathbb{R}^{d^2 \times d^2}$ of the form $M = \tau \cdot (v \otimes v)(v \otimes v)^\top + A \otimes B$ for some unit vector $v \in \mathbb{R}^d$ and matrices $A, B \in \mathbb{R}^{d \times d}$, we wish to recover the vector v. (This is a simplified version of the situation in Chapter 6 and in planted sparse vector.)

It is straightforward to see that the matrix $A \otimes B$ has spectral norm $\|A \otimes B\| = \|A\| \cdot \|B\|$, and so when $\tau \gg \|A\|\|B\|$, the matrix M has a noticeable spectral gap, and the top eigenvector of M will be close to $v \otimes v$. If $|\operatorname{Tr} A| \approx \|A\|$, the matrix $\operatorname{Tr}_{\mathbb{R}^d} M = \tau \cdot vv^\top + \operatorname{Tr}(A) \cdot B$ has a matching spectral gap, and we can still recover v, but now we only need to compute the top eigenvector of a $d \times d$ (as opposed to $d^2 \times d^2$) matrix.[2]

If A is a Wigner matrix (e.g. a symmetric matrix with iid ± 1 entries), then both $\operatorname{Tr}(A), \|A\| \approx \sqrt{n}$, and the above condition is indeed met. In our average case/machine learning settings the "noise" component is not as simple as $A \otimes B$ with A a Wigner matrix. Nonetheless, we are able to ensure that the noise displays a similar behavior under partial trace operations. In Chapter 6 this trick succeeded without further work: for planted sparse vector we will have to center the matrix eigenvalue distribution in order to obtain the kinds of cancellations in the noise described here.

Partial trace operations have previously been applied for rounding SoS relaxations. Specifically, the operation of *reweighing* and *conditioning*, used in rounding algorithms for sum-of-squares such as [33, 155, 30, 31, 119], corresponds to applying a partial trace operation to the moments matrix returned by the

[2] In some of our applications, the matrix M is only represented implicitly and has size super-linear in the size of the input, but nevertheless we can compute the top eigenvector of the partial trace $\operatorname{Tr}_{\mathbb{R}^d} M$ in nearly-linear time.

sum-of-squares relaxation.

SoS analysis Recall that in this problem we are given a linear subspace U (represented by some basis) that is spanned by a k-sparse unit vector $v_0 \in \mathbb{R}^d$ and random unit vectors $v_1, \ldots, v_{d-1} \in \mathbb{R}^d$. The goal is to recover the vector v_0 approximately.

Let $A \in \mathbb{R}^{n \times d}$ be a matrix whose columns form an orthonormal basis for U. Our starting point is the polynomial $f(x) = \|Ax\|_4^4 = \sum_{i=1}^n (Ax)_i^4$. Previous work showed that for $d \ll \sqrt{n}$ the maximizer of this polynomial over the sphere corresponds to a vector close to v_0 and that degree-4 sum-of-squares is able to capture this fact [27, 30]. Indeed, typical random vectors v in $\mathbb{R}^n$ satisfy $\|v\|_4^4 \approx 1/n$ whereas our planted vector satisfies $\|v_0\|_4^4 \geqslant 1/k \gg 1/n$, and this degree-4 information is leveraged by the SoS algorithms.

The polynomial f has a convenient matrix representation $M = \sum_{i=1}^n (a_i a_i^\top)^{\otimes 2}$, where $a_1, \ldots, a_n$ are the rows of the generator matrix A. It turns out that the eigenvalues of this matrix indeed give information about the planted sparse vector v_0. In particular, the vector $x_0 \in \mathbb{R}^d$ with $Ax_0 = v_0$ witnesses that M has an eigenvalue of at least $1/k$ because M's quadratic form with the vector $x_0^{\otimes 2}$ satisfies $\langle x_0^{\otimes 2}, M x_0^{\otimes 2} \rangle = \|v_0\|_4^4 \geqslant 1/k$. If we let M' be the corresponding matrix for the subspace U without the planted sparse vector, M' turns out to have only eigenvalues of at most $O(1/n)$ up to a single spurious eigenvalue with eigenvector far from any vector of the form $x \otimes x$ [27].

It follows that in order to hypothesis test between a random subspace with a planted sparse vector (alternative hypothesis) and a completely random subspace (null hypothesis), it is enough to compute the second-largest eigenvalue of a

d^2-by-d^2 matrix (representing the 4-norm polynomial over the subspace as above).

Improvements The best running time we can hope for with this basic approach is $O(d^4)$ (the size of the matrix). Since we are interested in $d \leqslant O(\sqrt{n})$, the resulting running time $O(nd^2)$ would be subquadratic but still super-linear in the input size $n \cdot d$ (for representing a d-dimensional subspace of $\mathbb{R}^n$). To speed things up, we use the partial trace approach outlined above. We will begin by applying the partial trace approach naively, obtaining reasonable bounds, and then show that a small modification to the matrix before the partial trace operation allows us to achieve even smaller signal-to-noise ratios.

In the alternative case, we may approximate $M \approx \frac{1}{k}(x_0 x_0^\top)^{\otimes 2} + Z$, where x_0 is the vector of coefficients of v_0 in the basis representation given by A (so that $Ax_0 = v_0$), and Z is the noise matrix. Since $\|x_0\| = 1$, the partial trace operation preserves the projector $(x_0 x_0^\top)^{\otimes 2}$ in the sense that $\mathrm{Tr}_{\mathbb{R}^d}(x_0 x_0^\top)^{\otimes 2} = x_0 x_0^\top$. Hence, with our heuristic approximation for M above, we could show that the top eigenvector of $\mathrm{Tr}_{\mathbb{R}^d} M$ is close to x_0 by showing that the spectral norm bound $\|\mathrm{Tr}_{\mathbb{R}^d} Z\| \leqslant o(1/k)$.

The partial trace of our matrix $M = \sum_{i=1}^n (a_i a_i^\top) \otimes (a_i a_i^\top)$ is easy to compute directly,

$$N = \mathrm{Tr}_{\mathbb{R}^d} M = \sum_{i=1}^n \|a_i\|_2^2 \cdot a_i a_i^\top .$$

In the alternative case (random subspace with planted sparse vector), a direct computation shows that

$$\lambda_{\mathrm{alt}} \geqslant \langle x_0, N x_0 \rangle \approx \tfrac{d}{n} \cdot \left(1 + \tfrac{n}{d}\|v_0\|_4^4\right) \geqslant \tfrac{d}{n}\left(1 + \tfrac{n}{dk}\right) .$$

Hence, a natural approach to distinguish between the alternative case and null cases is to upper bound the spectral norm of N in the null case.

In order to simplify the bound on the spectral norm of N in the null case, suppose that the columns of A are iid samples from the Gaussian distribution $\mathcal{N}(0, \frac{1}{d}\mathrm{Id})$ (rather than an orthogonal basis for the random subspace)–Section 7.3.1 establishes that this simplification is legitimate. In this simplified setup, the matrix N in the null case is the sum of n iid matrices $\{\|a_i\|^2 \cdot a_i a_i^\top\}$, and we can upper bound its spectral norm λ_{null} by $d/n \cdot (1 + O(\sqrt{d/n}))$ using standard matrix concentration bounds. Hence, using the spectral norm of N, we will be able to distinguish between the null and alternative cases so long as

$$\sqrt{d/n} \ll n/(dk) \implies \lambda_{\text{null}} \ll \lambda_{\text{alt}}\,.$$

For linear sparsity $k = \varepsilon \cdot n$, this inequality is true so long as $d \ll (n/\varepsilon^2)^{1/3}$, which is somewhat worse than the bound $\sqrt{n}$ bound on the dimension that we are aiming for.

Recall that $\operatorname{Tr} B = \sum_i \lambda_i(B)$ for a symmetric matrix B. As discussed above, the partial trace approach works best when the noise behaves as the tensor of two Wigner matrices, in that there are cancellations when the eigenvalues of the noise are summed. In our case, the noise terms $(a_i a_i^\top) \otimes (a_i a_i^\top)$ do not have this property, as in fact $\operatorname{Tr} a_i a_i^\top = \|a_i\|^2 \approx d/n$. Thus, in order to improve the dimension bound, we will center the eigenvalue distribution of the noise part of the matrix. This will cause it to behave more like a Wigner matrix, in that the spectral norm of the noise will not increase after a partial trace. Consider the partial trace of a matrix of the form

$$M - \alpha \cdot \mathrm{Id} \otimes \sum_i a_i a_i^\top\,,$$

for some constant $\alpha > 0$. The partial trace of this matrix is

$$N' = \sum_{i=1}^{n} (\|a_i\|_2^2 - \alpha) \cdot a_i a_i^\top\,.$$

We choose the constant $\alpha \approx d/n$ such that our matrix N' has expectation 0 in the null case, when the subspace is completely random. In the alternative case, the Rayleigh quotient of N' at x_0 simply shifts as compared to N, and we have $\lambda_{\text{alt}} \geqslant \langle x_0, N'x_0 \rangle \approx \|v_0\|_4^4 \geqslant 1/k$. On the other hand, in the null case, this centering operation causes significant cancellations in the eigenvalues of the partial trace matrix (instead of just shifting the eigenvalues). In the null case, N' has spectral norm $\lambda_{\text{null}} \leqslant O(d/n^{3/2})$ for $d \ll \sqrt{n}$. Therefore, the spectral norm of the matrix N' allows us to distinguish between the alternative and null case as long as $d/n^{3/2} \ll 1/k$, which is satisfied as long as $k \ll n$ and $d \ll \sqrt{n}$.

7.3 Algorithm and Analysis

Theorem 7.1.1 follows immediately from the following two lemmas, together with two observations:

1. The top eigenvalue of the matrix $\sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) a_i a_i^\top$ can be computed (with sufficient accuracy) in $\tilde{O}(nd)$ time.
2. Any unit vector $v \in \mathbb{R}^n$ with εn nonzero entries has $\|v\|_4^4 = \sum_{i \leqslant n} v(i)^4 \geqslant \frac{1}{\varepsilon n}$.

The first lemma bounds the spectral norm of the matrix $\sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) a_i a_i^\mathsf{T}$ in the null case.

Lemma 7.3.1. *Let $v_1, \ldots, v_d \in \mathbb{R}^n$ be d random (Haar-distributed) orthonormal vectors. Let $V = (v_1, \ldots, v_d) \in \mathbb{R}^{n \times d}$ be the matrix whose columns are $v_1, \ldots, v_d$. Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be the rows of V. With high probability, $\left\| \sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) a_i a_i^\mathsf{T} \right\| \leqslant \max(d/n^{3/2}, d^2/n^2) \cdot (\log n)^{O(1)}$.*

The second lemma shows that the same matrix has a large eigenvalue when $v_1, \ldots, v_d$ span a sparse vector.

Lemma 7.3.2. *Suppose $v \in \mathbb{R}^n$ has at most εn nonzero entries, and $g_1, \ldots, g_{d-1} \sim \mathcal{N}(0, \mathrm{Id})$. Let $v_1, \ldots, v_d$ be an orthonormal basis for $\mathrm{Span}\{v, g_1, \ldots, g_{d-1}\}$. Let $V = (v_1, \ldots, v_d)$ and let $a_1, \ldots, a_n$ be the rows of the matrix V. With high probability, the matrix $\sum_{i \leqslant n}(\|a_i\|^2 - \frac{d}{n})a_i a_i^\top$ has an eigenvalue of magnitude at least $\Omega(\|v\|_4^4) - \tilde{O}\left(\frac{d(d+\sqrt{n})}{n^2}\right) - O(\|v\|_4^2) \cdot \frac{\sqrt{d}}{n} - \frac{1}{n}$..*

Proof of Theorem 7.1.1. The algorithm is simply to output ALTERNATIVE if the maximum eigenvalue of $\sum_{i \leqslant n}(\|a_i\|^2 - \frac{d}{n})a_i a_i^\top$ is at least $(1/\varepsilon n) - d/n$ and null otherwise. This algorithm achieves the guarantees of the theorem so long as

$$C\|v\|_4^4) - \tilde{O}\left(\frac{d(d+\sqrt{n})}{n^2}\right) - O(\|v\|_4^2) \cdot \frac{\sqrt{d}}{n} - \frac{1}{n} \gg \max(d/n^{3/2}, d^2/n^2) \cdot (\log n)^{O(1)}$$

for some small-enough constant C. Since $\|v\|_4^4 \geqslant (\varepsilon n)^{-1}$, for every small-enough ε we can choose $d \leqslant \sqrt{n}/(\log n)^{O(1)}$ so that the right-hand side is $o(1/n)$ while the left is $\Omega(\varepsilon n)^{-1}$. To implement it in nearly-linear time, it is enough to do $(\log n)^{O(1)}$ matrix-vector multiplications by the matrix $\sum_{i \leqslant n}(\|a_i\|^2 - d/n)a_i a_i^\mathsf{T}$ (using the power method with a random starting point), each of which can be accomplished in linear time. □

7.3.1 Basis Swap Lemmas

The main technical difficulty in proving Lemmas 7.3.1 and 7.3.2 is to exchange the arbitrary basis $v_1, \ldots, v_d$ for the subspace V for a nice basis, where one of the basis vectors is the sparse vector v (if it exists) and the entries of the (other) basis vectors are iid Gaussian vectors. The following lemmas will accomplish this.

In the sequel, we write w.ov.p. for "with overwhelming probability," meaning probability $1 - n^{-\omega(1)}$.

First, the most critical fact: orthogonalizing does not change the leverage scores too much, in either the null or alternative models. [3]

Lemma 7.3.3. *Let $v \in \mathbb{R}^n$ be a unit vector and let $b_1, \ldots, b_n \in \mathbb{R}^{d-1}$ be iid from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id}_{d-1})$. Let $a_i \in \mathbb{R}^d$ be given by $a_i := (v(i)\ b_i)$. Let $A := \sum_i a_i a_i^\top$. Let $c \in \mathbb{R}^{d-1}$ be given by $c := \sum_i v(i) b_i$. Then for every index $i \in [n]$, w.ov.p.,*

$$\left| \|A^{-1/2} a_i\|^2 - \|a_i\|^2 \right| \leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n}\right) \cdot \|a_i\|^2 .$$

Lemma 7.3.4. *Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be independent random vectors from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ with $d \leqslant n$ and let $A = \sum_{i=1}^n a_i a_i^\top$. Then for every index $i \in [n]$, with overwhelming probability $1 - d^{\omega(1)}$,*

$$\left| \langle a_j, A^{-1} a_j \rangle - \|a_j\|^2 \right| \leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n}\right) \cdot \|a_j\|^2 .$$

The proof uses standard concentration and matrix inversion formulas, and can be found in Section 7.3.4. We also need the following lemmas, which follow from [175], theorem 5.9 – again there is one lemma for the null model and one for the alternative.

Lemma 7.3.5. *Let $v \in \mathbb{R}^n$ be a unit vector. Let $b_1, \ldots, b_n \in \mathbb{R}^{d-1}$ be iid from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id}_{d-1})$. Let $a_i \in \mathbb{R}^d$ be given by $a_i := (v(i)\ b_i)$. Then w.ov.p. $\| \sum_{i=1}^n a_i a_i^\top - \mathrm{Id}_d \| \leqslant \tilde{O}(d/n)^{1/2}$. In particular, when $d = o(n)$, this implies that w.ov.p. $\|(\sum_{i=1}^n a_i a_i^\top)^{-1} - \mathrm{Id}_d\| \leqslant \tilde{O}(d/n)^{1/2}$ and $\|(\sum_{i=1}^n a_i a_i^\top)^{-1/2} - \mathrm{Id}_d\| \leqslant \tilde{O}(d/n)^{1/2}$.*

Fact 7.3.6 (Special case of [176], Theorem 5.58)**.** *Let $b_1, \ldots, b_n \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ be iid.*

[3] Strictly speaking the good basis does not have leverage scores since it is not orthogonal, but we can still talk about the norms of the rows of the matrix whose columns are the basis vectors.

Then with high probability,

$$\left\| \sum_{i \leqslant n} b_i b_i^\mathsf{T} - \mathrm{Id} \right\| \leqslant O\left(\sqrt{\frac{d}{n}}\right).$$

7.3.2 Analysis for null model

In this section we prove Lemma 7.3.1. The following is the key fact.

Fact 7.3.7. *Let $b_1, \dots, b_n \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id}_d)$ be independent d-dimensional Gaussian vectors, with $d \leqslant n$. With high probability,*

$$\left\| \sum (\|b_i\|^2 - \tfrac{d}{n}) b_i b_i^\mathsf{T} \right\| \leqslant \max\left(\frac{d}{n^{3/2}}, \frac{d^2}{n^2}, \frac{d\sqrt{\log n}}{n^{3/2}} \right) \cdot (\log n)^{O(1)}$$

where $\|\cdot\|$ is the spectral norm, so long as $d \ll \sqrt{n}$.

Proof. Let $B = C \max(\frac{d}{n}, \frac{\log n}{n})$ for a big-enough C we choose later. We can replace b_i with $b_i' = b_i \cdot \mathbf{1}_{\|b_i\|^2 \leqslant B}$, since by standard concentration for big-enough C, we have $\mathbb{P}(\|b_i\|^2 > C \max(d/n, \log n/n) \leqslant n^{-10}$. Then we note that the matrices $(\|b_i'\|^2 - \frac{d}{n})(b_i')(b_i')^\mathsf{T}$ are iid, so we can apply a matrix Bernstein bound. We bound the covariance of the sum:

$$\begin{aligned}
\mathbb{E} \sum_{i \leqslant n} \left(\|b_i'\|^2 - \frac{d}{n} \right)^2 \|b_i'\|^2 (b_i')(b_i')^\mathsf{T} &\preceq \mathbb{E} \sum_{i \leqslant n} \left(\|b_i'\|^2 - \frac{d}{n} \right)^2 \|b_i'\|^2 b_i b_i^\mathsf{T} \\
&\preceq B \cdot \mathbb{E} \sum_{i \leqslant n} \left(\|b_i'\|^2 - \frac{d}{n} \right)^2 b_i b_i^\mathsf{T} \\
&\preceq B \cdot O\left(\frac{d}{n^2}\right) \cdot \mathbb{E} \sum_{i \leqslant n} b_i b_i^\mathsf{T} + O(n^{-8})
\end{aligned}$$

where in the last step we have replaced b_i' by b_i and exploited that the variance of $\|b_i\|^2$ is quite small to replace $(\|b_i\|^2 - d/n)^2$ with $O(d/n^2)$. (Because of the presence of $b_i b_i^\mathsf{T}$, we are not really working with the variance per se, and

some algebra is required to confirm this step, which we have omitted; see [91].) Finally, $\mathbb{E}\, b_i b_i^\top = \frac{1}{n}\mathrm{Id}$, so we conclude that $\left\|\mathbb{E}\sum_{i\leqslant n}\left(\|b_i'\|^2 - \frac{d}{n}\right)^2 \|b_i'\|^2 (b_i')(b_i')^\mathsf{T}\right\| \leqslant Bd/n^2 + O(n^{-8})$ (the important term is the first one).

Applying Theorem 5.0.8, we conclude that for any t,

$$\mathbb{P}\left\{\left\|\sum(\|b_i\|^2 - \tfrac{d}{n})b_i b_i^\mathsf{T}\right\| > t\right\} \leqslant O(\log d)\cdot \exp\left(\frac{-\Omega(t^2)}{\frac{Bd}{n^2} + n^{-8} + t\cdot\frac{d^2}{n^2}}\right)$$

and so the conclusion follows. □

Now we can prove Lemma 7.3.1.

Proof of Lemma 7.3.1. We can sample the vectors $v_1, \ldots, v_d$ by first sampling an $n \times d$ matrix of iid Gaussians G with entries from $\mathcal{N}(0, \frac{1}{n})$, then taking $v_1, \ldots, v_d$ to be the columns of $G(G^\top G)^{-1/2}$. Let $b_1, \ldots, b_n$ be the rows of G, and $a_1, \ldots, a_n$ be the rows of $G(G^\top G)^{-1/2}$. With high probability, by Fact 7.3.7,

$$\left\|\sum_{i\leqslant n}(\|b_i\|^2 - \tfrac{d}{n})b_i b_i^\mathsf{T}\right\| \leqslant \max\left(\frac{d}{n^{3/2}}, \frac{d^2}{n^2}, \frac{d\sqrt{\log n}}{n^{3/2}}\right)\cdot(\log n)^{O(1)}\,.$$

Now we apply Lemma 7.3.4 to find that

$$\sum_{i\leqslant n}(\|b_i\|^2 - \|a_i\|^2)b_i b_i^\mathsf{T} \preceq \tilde{O}\left(\frac{d}{n}\cdot\frac{d+\sqrt{n}}{n}\right)\sum_{i\leqslant n} b_i b_i^\mathsf{T}$$

with high probability. Since with high probability $\sum_{i\leqslant n} b_i b_i^\mathsf{T} \preceq O(1)$, we conclude that

$$\left\|\sum_{i\leqslant n}(\|b_i\|^2 - \tfrac{d}{n})b_i b_i^\mathsf{T} - \sum_{i\leqslant n}(\|a_i\|^2 - \tfrac{d}{n})b_i b_i^\mathsf{T}\right\| \leqslant \tilde{O}\left(\frac{d(d+\sqrt{n})}{n^2}\right)$$

with high probability. Since $a_i = (\sum_{i\leqslant n} b_i b_i^\mathsf{T})^{-1/2} b_i$ and the matrix $(\sum_{i\leqslant n} b_i b_i^\mathsf{T}) \preceq O(1)$, the lemma follows. □

7.3.3 Analysis for alternative model

Proof of Lemma 7.3.2. Let $b_1, \dots, b_n$ be the rows of the matrix $(v, g_1, \dots, g_d)$. By rotation invariance, we can take $a_i = (\sum b_i b_i^\top)^{-1/2} b_i$. Since with high probability by Lemma 7.3.5 $\sum_{i \leqslant n} b_i b_i^\top = \mathrm{Id} \pm \tilde{O}(\sqrt{d/n})$, it will be enough to show that $\sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) b_i b_i^\top$ has a large-enough eigenvalue.

Consider the test vector $e_1 \in \mathbb{R}^d$, the first standard basis vector. We analyze the quadratic form $\sum_{i \leqslant n} (\|a_i\|^2 - \frac{d}{n}) \langle b_i, e_1 \rangle^2$. By Lemma 7.3.3, with high probability

$$\left| \sum_{i \leqslant n} (\|a_i\|^2 - \tfrac{d}{n}) \langle b_i, e_1 \rangle^2 - \sum_{i \leqslant n} (\|b_i\|^2 - \tfrac{d}{n}) \langle b_i, e_1 \rangle^2 \right| \leqslant \sum_{i \leqslant n} \tilde{O}((d+\sqrt{n})/n) \|b_i\|^2 \langle b_i, e_1 \rangle^2$$

By standard concentration, this is at most $\tilde{O}(d(d + \sqrt{n})/n^2)$ with high probability.

So with high probability,

$$\sum_{i \leqslant n} (\|a_i\|^2 - \tfrac{d}{n}) \langle b_i, e_1 \rangle^2 \geqslant \sum_{i \leqslant n} (\|b_i\|^2 - \tfrac{d}{n}) \langle b_i, e_1 \rangle^2 - \tilde{O}(d(d + \sqrt{n})/n^2)$$

Expanding $\|b_i\|^2 = \sum_{j \leqslant d} \langle b_i, e_j \rangle^2$, the first term on the right hand side equals

$$\sum_{i \leqslant n} \langle b_i, e_1 \rangle^4 + \sum_{i \leqslant n} \left(\sum_{d \geqslant j > 1} \langle b_i, e_j \rangle^2 - \tfrac{d}{n} \right) \langle b_i, e_1 \rangle^2 .$$

Now recall that $\langle b_i, e_1 \rangle = v(i)$, the i-th entry of the sparse vector. So this is equal to $\|v\|_4^4 + \sum \sum_{i \leqslant n} \left(\sum_{d \geqslant j > 1} b_{ij}^2 - \frac{d}{n} \right) v(i)^2$ where b_{ij} are iid Gaussians $\mathcal{N}(0, 1/n)$. This is a sum of independent random variables $(b_{ij}^2 - d/n) v(i)^2$; By standard concentration, it is at most $1/n + O(\sqrt{d}/n) \|v\|_4^2$ in magnitude with high probability.

Putting it together, we get that with high probability $\sum_{i \leqslant n} (\|a_i\|^2 - d/n) a_i a_i^\mathsf{T}$ has an eigenvalue of magnitude at least

$$\Omega(\|v\|_4^4) - \tilde{O}\left(\frac{d(d + \sqrt{n})}{n^2} \right) - O(\|v\|_4^2) \cdot \frac{\sqrt{d}}{n} - \frac{1}{n} .$$

□

7.3.4 Concentration bounds for basis swap

Here we prove (restatements of) Lemmas 7.3.3 and 7.3.4.

Lemma 7.3.8. *Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be independent random vectors from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ with $d \leqslant n$ and let $A = \sum_{i=1}^n a_i a_i^\top$. Then for every unit vector $x \in \mathbb{R}^d$, with overwhelming probability $1 - d^{-\omega(1)}$,*

$$\left|\langle x, A^{-1}x\rangle - \|x\|^2\right| \leqslant \tilde{O}\left(\frac{d+\sqrt{n}}{n}\right) \cdot \|x\|^2 .$$

Proof. Let $x \in \mathbb{R}^d$. By scale invariance, we may assume $\|x\| = 1$.

By standard matrix concentration bounds, the matrix $B = \mathrm{Id} - A$ has spectral norm $\|B\| \leqslant \tilde{O}(d/n)^{1/2}$ w.ov.p. [175, Corollary 5.50]. Since $A^{-1} = (\mathrm{Id} - B)^{-1} = \sum_{k=0}^\infty B^k$, the spectral norm of $A^{-1} - \mathrm{Id} - B$ is at most $\sum_{k=2}^\infty \|B\|^k$ (whenever the series converges). Hence, $\|A^{-1} - \mathrm{Id} - B\| \leqslant \tilde{O}(d/n)$ w.ov.p..

It follows that it is enough to show that $|\langle x, Bx\rangle| \leqslant \tilde{O}(1/n)^{1/2}$ w.ov.p.. The random variable $n - n\langle x, Bx\rangle = \sum_{i=1}^n \langle \sqrt{n} \cdot a_i, x\rangle^2$ is χ^2-distributed with n degrees of freedom. Thus, by standard concentration bounds, $n|\langle x, Bx\rangle| \leqslant \tilde{O}(\sqrt{n})$ w.ov.p. [115].

We conclude that with overwhelming probability $1 - d^{-\omega(1)}$,

$$\left|\langle x, A^{-1}x\rangle - \|x\|^2\right| \leqslant |\langle x, Bx\rangle| + \tilde{O}(d/n) \leqslant \tilde{O}\left(\frac{d+\sqrt{n}}{n}\right) .$$

□

Lemma 7.3.9 (Restatement of Lemma 7.3.4)**.** *Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be independent random vectors from $\mathcal{N}(0, \frac{1}{n}\mathrm{Id})$ with $d \leqslant n$ and let $A = \sum_{i=1}^n a_i a_i^\top$. Then for every index $i \in [n]$, with overwhelming probability $1 - d^{\omega(1)}$,*

$$\left|\langle a_j, A^{-1}a_j\rangle - \|a_j\|^2\right| \leqslant \tilde{O}\left(\frac{d+\sqrt{n}}{n}\right) \cdot \|a_j\|^2 .$$

Proof. Let $A_{-j} = \sum_{i \neq j} a_i a_i^\top$. By Sherman–Morrison,

$$A^{-1} = (A_{-j} + a_j a_j^\top)^{-1} = A_{-j}^{-1} - \frac{1}{1 + a_j^\top A_{-j}^{-1} a_j} A_{-j}^{-1} a_j a_j^\top A_{-j}^{-1}$$

Thus, $\langle a_j, A^{-1} a_j \rangle = \langle a_j, A_{-j}^{-1} a_j \rangle - \langle a_j, A_{-j}^{-1} a_j \rangle^2 / (1 + \langle a_j, A_{-j}^{-1} a_j \rangle)$. Since $\| \frac{n}{n-1} A_{-j} - \mathrm{Id} \| = \tilde{O}(d/n)^{1/2}$ w.ov.p., we also have $\|A_{-j}^{-1}\| \leqslant 2$ with overwhelming probability. Therefore, w.ov.p.,

$$\left| \langle a_j, A^{-1} a_j \rangle - \langle a_j, A_{-j}^{-1} a_j \rangle \right| \leqslant \langle a_j, A_{-j}^{-1} a_j \rangle^2 \leqslant 4 \|a_j\|^4 \leqslant \tilde{O}(d/n) \cdot \|a_j\|^2 .$$

At the same time, by Lemma 7.3.8, w.ov.p.,

$$\left| \langle a_j, \tfrac{n}{n-1} A_{-j}^{-1} a_j \rangle - \|a_j\|^2 \right| \leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n} \right) \cdot \|a_j\|^2 .$$

We conclude that, w.ov.p.,

$$\begin{aligned} \left| \langle a_j, A^{-1} a_j \rangle - \|a_j\|^2 \right| &\leqslant \left| \langle a_j, A^{-1} a_j \rangle - \langle a_j, A_{-j}^{-1} a_j \rangle \right| + \left| \langle a_j, A_{-j}^{-1} a_j \rangle - \tfrac{n-1}{n} \|a_j\|^2 \right| + \tfrac{1}{n} \|a_j\|^2 \\ &\leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n} \right) . \end{aligned}$$

□

Lemma 7.3.10. *Let A be a block matrix where one of the diagonal blocks is the 1×1 identity; that is,*

$$A = \begin{pmatrix} \|v\|^2 & c^\top \\ c & B \end{pmatrix} = \begin{pmatrix} 1 & c^\top \\ c & B \end{pmatrix} .$$

for some matrix B and vector c. Let x be a vector which decomposes as $x = (x(1)\ x')$ where $x(1) = \langle x, e_1 \rangle$ for e_1 the first standard basis vector.

Then

$$\langle x, A^{-1} x \rangle = \langle x', \left(B^{-1} + \frac{B^{-1} c c^\top B^{-1}}{1 - c^\top B^{-1} c} \right) x' \rangle + 2x(1) \langle \left(B^{-1} + \frac{B^{-1} c c^\top B^{-1}}{1 - c^\top B^{-1} c} \right) c, x' \rangle + (1 - c^\top B^{-1} c)^{-1} x(1)^2 .$$

Proof. By the formula for block matrix inverses,

$$A^{-1} = \begin{pmatrix} (1 - c^\top B^{-1} c)^{-1} & c^T (B - cc^\top)^{-1} \\ (B - cc^\top)^{-1} c & (B - cc^\top)^{-1} \end{pmatrix}.$$

The result follows by Sherman-Morrison applied to $(B - cc^\top)^{-1}$ and the definition of x. □

Lemma 7.3.11 (Restatement of Lemma 7.3.3)**.** *Let $v \in \mathbb{R}^n$ be a unit vector and let $b_1, \dots, b_n \in \mathbb{R}^{d-1}$ have iid entries from $\mathcal{N}(0, 1/n)$. Let $a_i \in \mathbb{R}^d$ be given by $a_i := (v(i)\ b_i)$. Let $A := \sum_i a_i a_i^T$. Let $c \in \mathbb{R}^{d-1}$ be given by $c := \sum_i v(i) b_i$. Then for every index $i \in [n]$, w.ov.p.,*

$$\left| \langle a_i, A^{-1} a_i \rangle - \|a_i\|^2 \right| \leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n} \right) \cdot \|a_i\|^2 .$$

Proof. Let $B := \sum_i b_i b_i^T$. By standard concentration, $\|B^{-1} - \mathrm{Id}\| \leqslant \tilde{O}(d/n)^{1/2}$ w.ov.p. [175, Corollary 5.50]. At the same time, since v has unit norm, the entries of c are iid samples from $\mathcal{N}(0, 1/n)$, and hence $n\|c\|^2$ is χ^2-distributed with d degrees of freedom. Thus w.ov.p. $\|c\|^2 \leqslant \frac{d}{n} + \tilde{O}(dn)^{-1/2}$. Together these imply the following useful estimates, all of which hold w.ov.p.:

$$\begin{aligned} |c^\top B^{-1} c| &\leqslant \|c\|^2 \|B^{-1}\|_{op} \leqslant \frac{d}{n} + \tilde{O}\left(\frac{d}{n}\right)^{3/2} \\ \|B^{-1} c c^\top B^{-1}\|_{op} &\leqslant \|c\|^2 \|B^{-1}\|_{op}^2 \leqslant \frac{d}{n} + \tilde{O}\left(\frac{d}{n}\right)^{3/2} \\ \left\| \frac{B^{-1} c c^\top B^{-1}}{1 - c^\top B^{-1} c} \right\|_{op} &\leqslant \frac{d}{n} + \tilde{O}\left(\frac{d}{n}\right)^{3/2}, \end{aligned}$$

where the first two use Cauchy-Schwarz and the last follows from the first two.

We turn now to the expansion of $\langle a_i, A^{-1} a_i \rangle$ offered by Lemma 7.3.10,

$$\langle a_i, A^{-1} a_i \rangle = \langle b_i, \left(B^{-1} + \frac{B^{-1} c c^\top B^{-1}}{1 - c^\top B^{-1} c} \right) b_i \rangle \tag{7.3.1}$$

$$+ 2v(i)\langle\left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)c, b_i\rangle \tag{7.3.2}$$

$$+ (1 - c^\top B^{-1}c)^{-1}v(i)^2\,. \tag{7.3.3}$$

Addressing 7.3.1 first, by the above estimates and Lemma 7.3.4 applied to $\langle b_i, B^{-1}b_i\rangle$,

$$\left|\langle b_i, \left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)b_i\rangle - \|b_i\|^2\right| \leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n}\right)\cdot\|b_i\|^2$$

w.ov.p.. For 7.3.2, we pull out the important factor of $\|c\|$ and separate $v(i)$ from b_i: w.ov.p.,

$$\begin{aligned}
\left|2v(i)\langle\left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)c, b_i\rangle\right| &= \left|2\|c\|v(i)\langle\left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)\frac{c}{\|c\|}, b_i\rangle\right| \\
&\leqslant \left|\|c\|^2\left(v(i)^2 + \langle\left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)\frac{c}{\|c\|}, b_i\rangle^2\right)\right| \\
&\leqslant \tilde{O}\left(\frac{d}{n}\right)(v(i)^2 + \|b_i\|^2) \\
&= \tilde{O}\left(\frac{d}{n}\right)\|a_i\|^2\,,
\end{aligned}$$

where the last inequality follows from our estimates above and Cauchy-Schwarz.

Finally, for 7.3.3, since $(1 - c^\top B^{-1}c) \geqslant 1 - \tilde{O}(d/n)$ w.ov.p., we have that

$$|(1 - c^\top B^{-1}c)^{-1}v(i)^2 - v(i)^2| \leqslant \tilde{O}\left(\frac{d}{n}\right)v(i)^2\,.$$

Putting it all together,

$$\begin{aligned}
\left|\langle a_i, A^{-1}a_i\rangle - \|a_i\|^2\right| &\leqslant \left|\langle b_i, \left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)b_i\rangle - \|b_i\|^2\right| \\
&\quad + \left|2v(i)\langle\left(B^{-1} + \frac{B^{-1}cc^\top B^{-1}}{1 - c^\top B^{-1}c}\right)c, b_i\rangle\right| \\
&\quad + |(1 - c^\top B^{-1}c)^{-1}v(i)^2 - v(i)^2| \\
&\leqslant \tilde{O}\left(\frac{d + \sqrt{n}}{n}\right)\cdot\|a_i\|^2\,.
\end{aligned}$$

□

Table 7.1: Comparison of algorithms for the planted sparse vector problem with ambient dimension n, subspace dimension d, and relative sparsity ε.

Reference	Technique	Runtime	Largest d	Largest ε
Demanet, Hand [60]	linear programming	poly	any	$\Omega(1/\sqrt{d})$
Barak, Kelner, Steurer [30]	SoS, general SDP	poly	$\Omega(\sqrt{n})$	$\Omega(1)$
Qu, Sun, Wright [151]	alternating minimization	$\tilde{O}(n^2d^5)$	$\Omega(n^{1/4})$	$\Omega(1)$
this work	SoS, partial traces	$\tilde{O}(nd)$	$\tilde{\Omega}(\sqrt{n})$	$\Omega(1)$

7.4 Chapter Notes

Attributions The material in this chapter is adapted from [91], joint work with Tselil Schramm, Jonathan Shi, and David Steurer.

History and Related Work The problem of finding a sparse vector planted in a random linear subspace was introduced by Spielman, Wang, and Wright as a way of learning sparse dictionaries [165]. Subsequent works have found further applications and begun studying the problem in its own right [60, 30, 151]. Several kinds of algorithms have been proposed for this problem based on linear programming (LP), basic semidefinite programming (SDP), sum-of-squares, and non-convex gradient descent (alternating directions method).

An inherent limitation of simpler convex methods (LP and basic SDP) [165, 56] is that they require the relative sparsity of the planted vector to be polynomial in the subspace dimension (less than $n/\sqrt{d}$ non-zero coordinates).

Sum-of-squares and non-convex methods do not share this limitation. They can recover planted vectors with constant relative sparsity even if the subspace has polynomial dimension (up to dimension $O(n^{1/2})$ for sum-of-squares [30] and up to $O(n^{1/4})$ for non-convex methods [151]).

CHAPTER 8

SIMPLE STATISTICS, SOS, AND SHARP THRESHOLDS: ALGORITHMS AND LOWER BOUNDS FOR COMMUNITY DETECTION

In this chapter, we make a detailed study of community detection and stochastic block models from the perspective of simple statistics and the SoS method. We design a new algorithm for estimating latent community structure in very sparse (constant average degree) graphs with overlapping communities. Along the way, we develop a theory of estimation via simple statistics to complement our hypothesis testing theory from Chapter 2, and we make progress in explaining one of the most striking information-computation gaps: the sharp easy-to-hard phase transition threshold in stochastic block models known as the *Kesten-Stigum threshold*.

Our overlapping community recovery algorithm also uses the SoS method in a novel way in the inference context – instead of using a relaxation of a maximum likelihood problem as described in Chapter 3, we use SoS to de-noise an initial estimate of the hidden variables. For this we design a new SoS algorithm for tensor decomposition in a high-noise regime, which we believe is of independent interest. This approach to using SoS for estimation removes some of the creativity involved in constructing dual witnesses/SoS proofs as we did in Chapter 6.

Stochastic block models are probability distributions on graphs with latent community structure. In the simplest case, the block model is a distribution μ on pairs (G, x) where $x \in \{0,1\}^n$ is a partition of n vertices into two communities, and G is a random graph with edges more likely to appear between vertices on the same side of the parition than between vertices on opposite sides. Richer models support more than two communities and allow nodes to participate in multiple

communities. As usual, block models induce hypothesis testing, estimation, and refutation problems – in this chapter we focus on estimation, where the goal is to estimate the hidden community structure given access to the graph.

In the constant-average-degree regime, we show that the best previous algorithms can be interpreted in terms of simple statistics [125, 138, 3]. Moreover, for the case of richer community structures like multiple communities and especially overlapping communities, our simple-statistics-plus-SoS estimation method achieves significantly stronger recovery guarantees.[1]

Chapter Overview In the remainder of this introduction we discuss our results and their relation to previous work in more detail. In Section 9.2 (Techniques) we describe a theory of estimation via simple statistics, and we illustrate the idea by recovering a famous result in the theory of spiked random matrices with a much simplified proof: the Baik-Ben-Arous-Peché push-out effect [22]. In Section 8.3 (Warmup) we re-prove (up to some loss in the running time) the result of Mossel-Neeman-Sly on the two-community block model as an application of our meta-algorithm, again with very simple proofs. In Section 8.4 (Matrix estimation) we re-interpret the best existing results on the block model, due to Abbe and Sandon, as applications of our meta-algorithm.

In Section 8.5 (Tensor estimation) we apply our meta-algorithm to the mixed-membership block model. Following that, in Section 8.6 (Lower bounds) we prove that no algorithm captured by our meta-algorithm can recover communities in the block model past the Kesten-Stigum threshold.

[1] If we represent the community structure by k vectors $y_1, \dots, y_k \in \{0,1\}^n$ that indicate community memberships, then previous algorithms [3] do not aim to recover these vectors but, roughly speaking, only a random linear combination of them. While for some settings it is in fact impossible to estimate the individual vectors, we show that in many settings it is possible to estimate them (in particular for symmetric block models).

In Section 8.7 (Tensor decomposition), which can be read independently of much of the rest of the paper, we give a new algorithm for tensor decomposition and prove its correctness; this algorithm is used by our meta-algorithm as a black box.

8.1 Results

8.1.1 Bayesian estimation via simple statistics

We first describe an approach to estimation algorithm design hat is enough to capture the best known algorithms for the stochastic block model with k disjoint communities, which we now define. Let $\varepsilon, d > 0$. Draw y uniformly from $[k]^n$. For each pair $i \neq j$, add the edge $\{i, j\}$ to a graph on n vertices with probability $(1+(1-\frac{1}{k})\varepsilon)\frac{d}{n}$ if $y_i = y_j$ and $(1-\frac{\varepsilon}{k})\frac{d}{n}$ otherwise. The resulting graph has expected average degree d.

A series of recent works has explored the problem of estimating y in these models for the sparsest-possible graphs. The emerging picture, first conjectured via techniques from statistical physics in the work [58], is that in the k-community block model it is possible to recover a nontrivial estimate of y via a polynomial time algorithm if and only if $d = (1+\delta)\frac{k^2}{\varepsilon^2}$ for $\delta \geqslant \Omega(1)$. This is called the Kesten-Stigum threshold. The algorithmic side of this conjecture was confirmed by [125, 138] for $k = 2$ and [3] for general k.

By taking a simple statistics perspective, we are able to design an algorithm with similarly-precise guarantees for a more complex estimation problem, which

is more realistic for real-world networks: the *mixed-membership* block model [6] which we now define informally. Let $\alpha \geqslant 0$ be an overlap parameter. Draw y from $\binom{k}{t}^n$, where $t = \frac{k(\alpha+1)}{k+\alpha} \approx \alpha + 1$; that is for each of n nodes pick a set S_j of roughly $\alpha + 1$ communities.[2] For each pair i, j, add an edge to the graph with probability $(1 + (\frac{|S_i \cap S_j|}{t^2} - \frac{1}{k})\varepsilon)\frac{d}{n}$. (That is, with probability which increases as i and j participate in more communities together.) In the limit $\alpha \to 0$ this becomes the k-community block model.

Returning to the estimation problems in general, (but keeping in mind the block model), let $p(x, y)$ be a joint probability distribution over observable variables $x \in \mathbb{R}^n$ and hidden variables $y \in \mathbb{R}^m$. Nature draws (x, y) from the distribution p, we observe x and our goal is to provide an estimate $\widehat{y}(x)$ for y, with quality measured by some loss function ℓ. Often the mean square error $\mathbb{E}_{p(x,y)} \left\|\widehat{y}(x) - y\right\|^2$ is a reasonable loss function. For this measure, the information-theoretically optimal estimate is the mean of the posterior distribution $\widehat{y}(x) = \mathbb{E}_{p(y|x)}\, y$. This approach has two issues that we address in the current work.

The first issue is that, as we have discussed before, naively computing the mean of the posterior distribution takes time exponential in the dimension of y. For example, if $y \in \{\pm 1\}^m$, then $\mathbb{E}_{p(y|x)}\, y = \sum_{y \in \{\pm 1\}^m} y \cdot p(y \mid x)$; there are 2^m terms in this sum. There are many well-known algorithmic approaches that aim to address this issue or related ones, for example, belief propagation [75, 145] or expectation maximization [61]. While these approaches appear to work well in practice, they are notoriously difficult to analyze.

[2] In actuality one draws for each node $i \in [n]$ a probability vector $\sigma_i \in \Delta_{k-1}$ from the Dirichlet distribution with parameter α; we describe a nearly-equivalent model here for the sake of simplicity—see Section 8.1.2 for details. Our guarantees for recovery in the mixed-membership model also apply to the model here because it has the same second moments as the Dirichlet distribution.

Our strategy is to analytically determine a low-degree polynomial $f(x)$ so that $\mathbb{E}_{p(x,y)} \left\| f(x) - y \right\|^2$ is as small as possible and use the fact that low-degree polynomials can be evaluated efficiently (even for high dimensions n).[3] Up to normalization, f is simple statistic. Because the maximum eigenvector of an n-dimensional linear operator with a spectral gap is an $O(\log n)$-degree polynomial of its entries, this approach captures spectral properties of linear operators whose entries are low-degree polynomials of observable variables x. Examples of such operators include adjacency matrices (when x is a graph), empirical covariance matrices (when x is a list of vectors), as well as more sophisticated objects such as linearized belief propagation operators (e.g., [2]) and the Hashimoto non-backtracking operator.

The second issue is that even if we could compute the posterior mean exactly, it may not contain any information about the hidden variable y and the mean square error is not the right measure to assess the quality of the estimator. This situation typically arises if there are symmetries in the posterior distribution. For example, in the stochastic block model with two communities we have $\mathbb{E}_{p(y|x)} y = 0$ regardless of the observations x because $p(y \mid x) = p(-y|x)$. A simple way to resolve this issue is to estimate higher-order moments of the hidden variables. For stochastic block models with disjoint communities, the second moment $\mathbb{E}_{p(y|x)} yy^{\mathsf{T}}$ would suffice. (For overlapping communities, we need third moments $\mathbb{E}_{p(y|x)} y^{\otimes 3}$ due to more substantial symmetries.)

For now, we think of y as an m-dimensional vector and x as an n-dimensional vector (in the blockmodel on N nodes, this would correspond to $m \approx kN$ and $n = N^2$). Our estimation algorithms follow a two-step strategy:

[3] Our polynomials typically have logarithmic degree and naive evaluation takes time $n^{O(\log n)}$. However, we show that under mild conditions it is possible to approximately evaluate these polynomials in polynomial time using the idea of color coding [11].

1. Given $x \sim p(x|y)$, evaluate a fixed, low-degree polynomial $P(x)$ taking values in $(\mathbb{R}^m)^{\otimes \ell}$. (Usually ℓ is 2 or 3.)
2. Apply a robust eigenvector or SDP-based algorithm (if $\ell = 2$), or a robust tensor decomposition algorithm (if $\ell = 3$ or higher) to P to obtain an estimator $\widehat{y}$ for y.

The polynomial $P(x)$ should be an optimal low-degree approximation to $y^{\otimes \ell}$, in the following sense: if n is sufficiently large that some low-degree polynomial $Q(x)$ has constant correlation with $y^{\otimes \ell}$

$$\mathop{\mathbb{E}}_{x,y} \langle Q, y^{\otimes \ell} \rangle \geqslant \Omega(1) \cdot (\mathop{\mathbb{E}}_{x} \|Q\|^2)^{1/2} (\mathbb{E} \|y^{\otimes \ell}\|^2)^{1/2} ,$$

then P has this guarantee. (The inner products and norms are all Euclidean.)

A prerequisite for applying this approach to a particular inference problem $p(x, y)$ is that it be possible to estimate y given $\mathbb{E}\left[y^{\otimes \ell} \mid x\right]$ for some constant ℓ. For such a problem, the optimal Bayesian inference procedure (ignoring computational constraints) can be captured by computing $F(x) = \mathbb{E}\left[y^{\otimes \ell} \mid x\right]$, then using it to estimate y. When $p(x, y)$ is such that it is information-theoretically possible to estimate y from x, these posterior moments will generally satisfy $\mathbb{E}\langle F(x), y^{\otimes \ell} \rangle \geqslant \Omega(1) \cdot (\mathbb{E} \|F(x)\|^2)^{1/2} (\mathbb{E} \|y^{\otimes \ell}\|^2)^{1/2}$, for some constant ℓ. Our observation is that when F is approximately a low-degree function, this estimation procedure can be carried out via an efficient algorithm.

Matrix estimation and prior results for block models In the case $\ell = 2$, where one uses the covariance $\mathbb{E}\left[yy^{\mathsf{T}} \mid x\right]$ to estimate y, the preceding discussion is captured by the following theorem.

Theorem 8.1.1 (Bayesian estimation meta-theorem—2nd moment)**.** *Let $\delta > 0$ and $p(x, y)$ be a distribution over vectors $x \in \{0, 1\}^n$ and unit vectors $y \in \mathbb{R}^d$. Assume*

$p(x) \geqslant 2^{-n^{O(1)}}$ *for all* $x \in \{0,1\}^n$.[4] *Suppose there exists a matrix-valued degree-*D *polynomial* $P(x)$ *such that*

$$\mathop{\mathbb{E}}_{p(x,y)} \langle P(x), yy^{\mathsf{T}} \rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|_F^2 \right)^{1/2} . \tag{8.1.1}$$

Then, there exists $\delta' \geqslant \delta^{O(1)} > 0$ *and an estimator* $\widehat{y}(x)$ *computable by a circuit of size* $n^{O(D)}$ *such that*

$$\mathop{\mathbb{E}}_{p(x,y)} \langle \widehat{y}(x), y \rangle^2 \geqslant \delta' . \tag{8.1.2}$$

To apply this theorem to the previously-discussed setting of samples $x_1, \ldots, x_N$ generated from $p(x \mid y)$, assume the samples $x_1, \ldots, x_N$ are in some fixed way packaged into a single n-length vector x.

One curious aspect of the theorem statement is that it yields a nonuniform algorithm—a family of circuits—rather than a uniform algorithm. If the coefficients of the polynomial P can themselves be computed in polynomial time, then the conclusion of the algorithm is that an $n^{O(D)}$-time algorithm exists with the same guarantees.

As previously mentioned, the meta-algorithm has a parameter D, the degree of the polynomial P. If $D = n$, then whenever it is information-theoretically possible to estimate y from $\mathbb{E}[yy^\top \mid x]$, the meta-algorithm can do so (in exponential time). This follows from the fact that every function in n Boolean variables is a polynomial of degree at most n. It is also notable that, while a degree D polynomial can be evaluated by an $n^{O(D)}$-size circuit, some degree-D polynomials can be evaluated by much smaller circuits. We exploit such polynomials for the block model (computable via *color coding*), obtaining $n^{O(1)}$-time algorithms from degree $\log n$

[4] This mild condition on the marginal distribution of x allows us to rule out pathological situations where a low-degree polynomial in x may be hard to evaluate accurately enough because of coefficients with super-polynomial bit-complexity.

polynomials. By using very particular polynomials, which can be computed via powers of *non-backtracking operators*, previous works on the block model are able to give algorithms with near-linear running times [138, 3].[5]

Using the appropriate polynomial P, this theorem captures the best known guarantees for partial recovery in the k-community stochastic block model. Via the same polynomial, applied in the mixed-membership setting, it also yields our first nontrivial algorithm for the mixed-membership model. However, as we discuss later, the recovery guarantees are weak compared to our main theorem.

Recalling the ε, d, k block model from the previous section, let $y \in \mathbb{R}^n$ be the centered indicator vector of, say, community 1.

Theorem 8.1.2 (Implicit in [125, 138, 3], special case of our main theorem, Theorem 8.1.4)**.** *Let $\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d}$. If x is sampled according to the n-node, k-community, ε-biased, α-mixed-membership block model with average degree d and y is the centered indicator vector of community 1, there is a $n \times n$-matrix valued polynomial P of degree $O(\log n)/\delta^{O(1)}$ such that*

$$\mathop{\mathbb{E}}_{x} \langle P(x), yy^\mathsf{T} \rangle \geqslant \left(\frac{\delta}{k(\alpha+1)}\right)^{O(1)} (\mathbb{E}\,\|P(x)\|^2)^{1/2} (\mathbb{E}\,\|yy^\mathsf{T}\|^2)^{1/2}\,.$$

Together with Theorem 8.1.1, up to questions of $n^{O(\log n)}$ versus $n^{O(1)}$ running times, when $\alpha \to 0$ this captures the previous best efficient algorithms for the k-community block model. (Once one has a unit vector correlated with y, it is not

[5] In this work we choose to work with *self-avoiding* walks rather than non-backtracking ones; while the corresponding polynomials cannot to our knowledge be evaluated in near-linear time, the analysis of these polynomials is much simpler than the analysis needed to understand non-backtracking walks. This helps to make the analysis of our algorithms much simpler than what is required by previous works, at the cost of large polynomial running times. It is an interesting question to reduce the running times of our algorithm for the mixed-membership block model to near-linear via non-backtracking walks, but since our aim here is to distinguish what is computable in polynomial time versus, say, exponential time, we do not pursue that improvement here.

hard to approximately identify the vertices in community 1.) While the previous works [125, 138, 3] did not consider the mixed-membership blockmodel, this theorem is easily obtained using techniques present in those works (as we show in our meta-algorithm, in Section 8.4).[6]

Symmetries in the posterior, tensor estimation, and improved error guarantees We turn next to our main theorem on the mixed-membership model, which offers substantial improvement on the correlation which can be obtained via Theorem 8.1.2. The matrix-based algorithm discussed above, Theorem 8.1.2, contains a curious asymmetry; namely the arbitrary choice of community 1. The block model distributions are symmetric under relabeling of the communities, which means that any estimator $P(x)$ of yy^{T} is also an estimator of $y'y'^{\mathsf{T}}$, where y' is the centered indicator of community $j > 1$. Since one wants to estimate all the vectors $y_1, \ldots, y_k$ (with y_i corresponding to the i-th community), it is more natural to consider the polynomial P to be an estimator of the matrix $M = \sum_{i\in[k]} y_i y_i^{\mathsf{T}}$.[7] Unsurprisingly, P is a better estimator of M than it is of y_1. In fact, with the same notation as in the theorems,

$$\mathop{\mathbb{E}}_{x,y} \langle P(x), M(y)\rangle \geqslant \delta^{O(1)} (\mathbb{E}\,\|P(x)\|^2)^{1/2} (\mathbb{E}\,\|M(y)\|^2)^{1/2}\,,$$

removing the $k^{O(1)}$ factor in the denominator. This guarantee is stronger: now the error in the estimator depends only on the distance δ of the parameters $\varepsilon, d, k, \alpha$

[6] In fact, if one is willing to lose an additional 2^{-k} in the correlation obtained in this theorem, one can obtain a similar result for the mixed-membership model by reducing it to the disjoint-communities with $K \approx 2^k$ communities, one for each subset of k communities. This works when each node participates in a subset of communities; if one uses the Dirichlet version of the mixed-membership model then suitable discretization would be necessary.

[7] In more general versions of the blockmodel studied in [3], where each pair i, j of communities may have a different edge probability Q_{ij} it is not always possible to estimate all of $y_1, \ldots, y_k$. We view it as an interesting open problem to extract as much information about $y_1, \ldots, y_k$ as possible in that setting; the guarantee of [3] amounts, roughly, to finidng a single vector in the linear span of $y_1, \ldots, y_k$.

from the critical threshold $\frac{k^2(\alpha+1)^2}{\varepsilon^2 d} = 1$ rather than additionally on k.

If given the matrix M exactly, one way to extract an estimator $\widehat{y_i}$ for some y_i is just to sample a random unit vector in the span of the top k eigenvectors of M. Such an estimator $\widehat{y_i}$ would have $\mathbb{E}\langle \widehat{y_i}, y_i\rangle^2 \geqslant \frac{1}{k^{O(1)}}\|y_i\|$, recovering the guarantees of the theorems above but not offering an estimator $\widehat{y_i}$ whose distance to y_i depends only on the distance δ above the critical threshold. Indeed, without exploiting additional structure of the vectors y_i is unclear how to remove this $1/k^{O(1)}$ factor. As a thought experiment, if one had the matrix $M' = \sum_{i \leqslant k} a_i a_i^{\mathsf{T}}$, where $a_1, \ldots, a_k$ were random unit vectors, then $a_1, \ldots, a_k$ would be nearly orthonormal and one could learn essentially only their linear span. (From the linear span it is only possible to find $\widehat{a_i}$ with correlation $\langle \widehat{a_i}, a_i\rangle^2 \geqslant 1/k^{O(1)}$.)

In the interest of generality we would like to avoid using such additional structure: while in the disjoint-community model the vectors y_i have disjoint support (after un-centering them), no such special structure is evident in the mixed-membership setting. Indeed, when α is comparable to k, the vectors y_i are similar to independent random vectors of the appropriate norm.

To address this issue we turn to tensor methods. To illustrate the main idea simply: if $a_1, \ldots, a_k$ are orthonormal, then it is possible to recover $a_1, \ldots, a_k$ exactly from the 3-tensor $T = \sum_{i \leqslant k} a_i^{\otimes 3}$. More abstractly, the meta-algorithm which uses 3rd moments is able to estimate hidden variables whose posterior distributions have a high degree of symmetry, without errors which worsen as the posteriors become more symmetric.

Theorem 8.1.3 (Bayesian estimation meta-theorem—3rd moment)**.** *Let $p(x, y_1, \ldots, y_k)$ be a joint distribution over vectors $x \in \{0,1\}^n$ and exchangable,*[8]

[8] Here, exchangeable means that for every $x \in \{0,1\}^n$ and every permutation $\pi : [k] \to [k]$,

orthonormal[9] *vectors* $y_1, \dots, y_k \in \mathbb{R}^d$. *Assume the marginal distribution of* x *satisfies* $p(x) \geqslant 2^{-n^{O(1)}}$ *for all* $x \in \{0,1\}^n$.[10] *Suppose there exists a tensor-valued degree-*D *polynomial* $P(x)$ *such that*

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \langle P(x), \sum_{i=1}^{k} y_i^{\otimes 3} \rangle \geqslant \delta \cdot \left(\mathop{\mathbb{E}}_{p(x)} \|P(x)\|^2 \right)^{1/2} \cdot \sqrt{k}\,. \tag{8.1.3}$$

(Here, $\|\cdot\|$ *is the norm induced by the inner product* $\langle \cdot, \cdot \rangle$. *The factor* $\sqrt{k}$ *normalizes the inequality because* $\|\sum_{i=1}^{k} y_i^{\otimes 3}\| = \sqrt{k}$ *by orthonormality.) Then, there exists* $\delta' \geqslant \delta^{O(1)} > 0$ *and a circuit of size* $n^{O(D)}$ *that given* $x \in \{0,1\}^n$ *outputs a list of unit vectors* $z_1, \dots, z_m$ *with* $m \leqslant n^{\mathrm{poly}(1/\delta)}$ *so that*

$$\mathop{\mathbb{E}}_{p(x,y_1,\dots,y_k)} \mathop{\mathbb{E}}_{i \sim [k]} \max_{j \in [m]} \langle y_i, z_j \rangle^2 \geqslant \delta'\,. \tag{8.1.4}$$

That the meta-algorithm captured by this theorem outputs a list of $n^{1/\mathrm{poly}(\delta)}$ vectors rather than just k vectors is an artifact of the algorithmic difficulty of multilinear algebra as compared to linear algebra. However, in most Bayesian estimation problems it is possible by using a very small number of additional samples (amounting to a low-order additive term in the total sample complexity) to cross-validate the vectors in the list $z_1, \dots, z_m$ and throw out those which are not correlated with some $y_1, \dots, y_k$. Our eventual algorithm for tensor decomposition (see Section 8.1.3 and Section 8.7) bakes this step in by assuming access to an oracle which evaluates the function $v \mapsto \sum_{i \in [k]} \langle v, y_i \rangle^4$.

A key component of the algorithm underlying Theorem 8.1.3 is a new algorithm for very robust orthogonal tensor decomposition.[11] Previous algorithms

we have $p(y_1, \dots, y_k \mid x) = p(y_{\pi(1)}, \dots, y_{\pi(k)} \mid x)$.

[9] Here, we say the vector-valued random variables $y_1, \dots, y_k$ are orthonormal if with probability 1 over the distribution p we have $\langle y_i, y_j \rangle = 0$ for all $i \neq j$ and $\|y_i\|^2 = 1$.

[10] As in the previous theorem, this mild condition on the marginal distribution of x allows us to rule out pathological situations where a low-degree polynomial in x may be hard to evaluate accurately enough because of coefficients with super-polynomial bit-complexity.

[11] An orthogonal 3-tensor is $\sum_{i=1}^{m} a_i^{\otimes 3}$, where $a_1, \dots, a_m$ are orthonormal.

for tensor decomposition require that the input tensor is close (in an appropriate norm) to only one orthogonal tensor. By contrast, our tensor decomposition algorithm is able to operate on a tensor T which is just $\delta \ll 1$ correlated to the orthogonal tensor $\sum y_i^{\otimes 3}$, and in particular might also be δ-correlated with $1/\delta$ other orthogonal tensors. If one views tensor decomposition as a *decoding* task, taking a tensor T and decoding it into its rank-one components, then our guarantees are analogous to list-decoding. Our algorithm in this setting involves a novel entropy-maximization program which, among other things, ensures that given a tensor T which for example is δ-correlated with two distinct orthogonal tensors A and B, the algorithm produces a list of vectors correlated with both the components of A and those of B.

Applying this meta-theorem (plus a simple cross-validation scheme to prune the vectors in the $n^{1/\mathrm{poly}(\delta)}$-length list) to the mixed-membership block model (and its special case, the k-disjoint-communities block model) yields the following theorem. (See Section 8.1.2 for formal statements.)

Theorem 8.1.4 (Main theorem on the mixed-membership block model, informal)**.** *Let $\varepsilon, d, k, \alpha$ be paramters of the mixed-membership block model, and let $\delta = 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} \geqslant \Omega(1)$. Let y_i be the centered indicator vector of the i-th community. There is an $n^{1/\mathrm{poly}(\delta)}$-time algorithm which, given a sample x from the $\varepsilon, d, k, \alpha$ block model, recovers vectors $\widehat{y}_1(x), \ldots, \widehat{y}_k(x)$ such that there is a permutation $\pi : [k] \to [k]$ with*

$$\mathbb{E}\langle \widehat{y}_{\pi(i)}, y_i \rangle^2 \geqslant \delta^{O(1)} (\mathbb{E} \|\widehat{y}_{\pi(i)}\|^2)^{1/2} (\mathbb{E} \|y_i\|^2)^{1/2} .$$

The eventual goal, as we discuss in Section 8.1.2, is to label each vertex by a probability vector τ_i which is correlated with the underlying label σ_i, but given the $\widehat{y}$ vectors from this theorem this is easily accomplished.

Comparison to the method of moments Another approach for designing statistical estimators for provable guarantees is the method of moments. Typically one considers parameters θ (which need not have a prior distribution $p(\theta)$) and iid samples $x_1, \dots, x_n \sim p(x|\theta)$. Generally one shows that the moments of the distribution $\{x|\theta\}$ are related to some function of θ: for example perhaps $\mathbb{E}[xx^\top \mid \theta] = f(\theta)$. Then one uses the samples x_i to estimate the moment $M = \mathbb{E}[xx^\top \mid \theta]$, and finally to estimate θ by $f^{-1}(M)$.

While the method of moments is quite flexible, for the high-noise problems we consider here it is not clear that it can achieve optimal sample complexity. For example, in our algorithms (and existing sample-optimal algorithms for the block model) it is important to exploit the flexibility to compute any polynomial of the samples jointly—given n samples our algorithms can evaluate a polynomial $P(x_1, \dots, x_n)$, and P often will not be an empirical average of some simpler function like $\sum_{i \leqslant n} q(x_i)$. The best algorithm for the mixed-membership block model before our work uses the method of moments and consequently requires much denser graphs than our method [12].

8.1.2 Detecting overlapping communities

We turn now to discuss our results for stochastic block models in more detail and compare them to the existing literature.

The stochastic block model is a widely studied (family of) model(s) of random graphs containing latent community structure. It is most common to study the block model in the sparse graph setting: many large real-world networks are sparse, and the sparse graph setting is nearly always more mathematically

challenging than the dense setting. A series of recent works has for the first time obtained algorithms which recover communities in block model graphs under (conjecturally) optimal sparsity conditions. For an excellent survey, see [1].

Such sharp results remain limited to relatively simple versions of the block model; where, in particular, each vertex is assigned a single community in an iid fashion. A separate line of work has developed more sophisticated and realistic random graph models with latent community structure, with the goal of greater applicability to real-life networks. The mixed-membership stochastic block model [6] is one such natural extension of the stochastic block model that allows for communities to overlap, as they do in large networks found in the wild.

In addition to the number of vertices n, the average degree d, the correlation parameter ε, and the number of communities k, this model has an overlap parameter $\alpha \geqslant 0$ that controls how many communities a typical vertex participates in. Roughly speaking, the model generates an n-vertex graph by choosing k communities as random vertex subsets of size $(1+\alpha)n/k$ and choosing $dn/2$ random edges, favoring pairs of vertices that have many communities in common.

Definition 8.1.5 (Mixed-membership stochastic block model)**.** The mixed-membership stochastic block model $\mathrm{SBM}(n, d, \varepsilon, k, \alpha)$ is the following distribution over n-vertex graphs G and k-dimensional probability vectors $\sigma_1, \ldots, \sigma_n$ for the vertices:

- draw $\sigma_1, \ldots, \sigma_n$ independently from $\mathrm{Dir}(\alpha)$ the symmetric k-dimensional Dirichlet distribution with parameter $\alpha \geqslant 0$,[12]
- for every potential edge $\{i, j\}$, add it to G with probability $\frac{d}{n} \cdot \Big(1 + \big(\langle \sigma_i, \sigma_j \rangle -$

[12]In the symmetric k-dimensional Dirichlet distribution with parameter $\alpha > 0$, the probability of a probability vector σ is proportional to $\prod_{t=1}^{k} \sigma(t)^{\alpha/k-1}$. By passing to the limit, we define $\mathrm{Dir}(0)$ to be the uniform distribution over the coordinate vectors $\mathbf{1}_1, \ldots, \mathbf{1}_k$.

$\frac{1}{k}\big)\varepsilon\Big)$.

Due to symmetry, $\langle \sigma_i, \sigma_j \rangle$ has expected value $\frac{1}{k}$, which means that the expected degree of every vertex in this graph is d. In the limit $\alpha \to 0$, the Dirichlet distribution is equivalent to the uniform distribution over coordinate vectors $\mathbf{1}_1, \ldots, \mathbf{1}_k$ and the model becomes SBM(n, d, ε, k), the stochastic block model with k *disjoint* communities. For $\alpha = k$, the Dirichlet distribution is uniform over the open $(k-1)$-simplex [178]. For general values of α, a probability vector from Dir(α) turns out to have expected collision probability $(1 - \frac{1}{k})\frac{1}{\alpha+1} + \frac{1}{k}$, which means that we can think of the probability vector being concentrated on about $\alpha + 1$ coordinates.[13] This property of the Dirichlet distribution is what determines the threshold for our algorithm. Correspondingly, our algorithm and analysis extends to a large class of distributions over probability vectors that share this property.

Measuring correlation with community structures In the constant-average-degree regime of the block model, recovering the label of every vertex correctly is information-theoretically impossible. For example, no information is present in a typical sample about the label of any isolated vertex, and in a typical sample a constant fraction of the vertices are isolated. Instead, at least in the k-disjoint-community setting, normally one looks to label vertices by labels $1, \ldots, k$ so that (up to a global permutation), this labeling has positive correlation with the true community labels.

When the communities are disjoint, one can measure such correlation using the sizes of $|S_j \cap \widehat{S}_j|$, where $S_j \subseteq [n]$ is the set of nodes in community j and $\widehat{S}_j$ is

[13] When k and α are comparable in magnitude, it is important to interpret this more accurately as $(\alpha + 1) \cdot \frac{k}{k+\alpha}$ coordinates.

an estimated set of nodes in community j. The original definition of *overlap*, the typical measure of labeling-accuracy in the constant-degree regime, takes this approach [58].

For present purposes this definition must be somewhat adapted, since in the mixed-membership block model there is no longer a good notion of a discrete set of nodes S_j for each community $j \in [k]$. We define a smoother notion of correlation with underlying community labels to accommodate that the labels σ_i are vectors in Δ_{k-1}. In discrete settings, for example when $\alpha \to 0$ (in which case one recovers the k-disjoint-community model), or more generally when each σ_i is the uniform distribution over some number of communities, our correlation measure recovers the usual notion of overlap.

Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ and $\tau = (\tau_1, \ldots, \tau_n)$ be labelings of the vertices $1, \ldots, n$ by by k-dimensional probability vectors. We define the *correlation* $\mathrm{corr}(\sigma, \tau)$ as

$$\max_{\pi} \mathop{\mathbb{E}}_{i \sim n} \langle \sigma_i, \tau_{\pi(i)} \rangle - \frac{1}{k} \tag{8.1.5}$$

where π ranges over permutations of the k underlying communities. This notion of correlation is closely related to the *overlap* of the distributions σ_i, τ_i.

To illustrate this notion of correlation, consider the case of disjoint communities (i.e., $\alpha = 0$), where the ground-truth labels τ_i are indicator vectors in k dimensions. Then, if $\mathbb{E}_i \langle \sigma_i, \tau_{\pi(i)} \rangle - \frac{1}{k} > \delta$, by looking at the large coordinates of σ_i it is possible to correctly identify the community memberships of a $\delta^{O(1)} + \frac{1}{k}$ fraction of the vertices, which is a $\delta^{O(1)}$ fraction more than would be identified by randomly assigning labels to the vertices without looking at the graph.

When the ground truth labels τ_i are spread over more than one coordinate—say, for example, they are uniform over t coordinates—the best recovery algorithm

cannot find σ's with correlation better than

$$\mathrm{corr}(\sigma, \tau) = \frac{1}{t} - \frac{1}{k},$$

which is achieved by $\sigma = \tau$. This is because in this case τ has collision probability $\langle \tau, \tau \rangle = \frac{1}{t}$.

Main result for mixed-membership models The following theorem gives a precise bound on the number of edges that allows us to find in polynomial time a labeling of the vertices of an n-node mixed membership block model having nontrivial correlation with the true underlying labels. Here, the parameters $d, \varepsilon, k, \alpha$ of the mixed-membership stochastic block model may even depend on the number of vertices n.

Theorem 8.1.6 (Mixed-membership SBM—significant correlation)**.** *Let $d, \varepsilon, k, \alpha$ be such that $k \leqslant n^{o(1)}$, $\alpha \leqslant n^{o(1)}$, and $\varepsilon^2 d \leqslant n^{o(1)}$. Suppose $\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0$. (Equivalently for small δ, suppose $\varepsilon^2 d \geqslant (1+\delta) \cdot k^2(\alpha+1)^2$.) Then, there exists $\delta' \geqslant \delta^{O(1)} > 0$ and an $n^{1/\mathrm{poly}(\delta)}$-time algorithm that given an n-vertex graph G outputs $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ such that*

$$\mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \mathrm{corr}(\sigma, \tau) \geqslant \delta' \cdot \left(\frac{1}{t} - \frac{1}{k}\right) \tag{8.1.6}$$

where $t = (\alpha+1) \cdot \frac{k}{k+\alpha}$ (samples from the α, k Dirichlet distribution are roughly uniform over t out of k coordinates). In particular, as $\delta \to 1$ we have $\mathbb{E}\,\mathrm{corr}(\sigma, \tau) \to \frac{1}{t} - \frac{1}{k}$, while $\mathbb{E}\,\mathrm{corr}(\sigma, \sigma) = \frac{1}{t} - \frac{1}{k}$.

Note that in the above theorem, the correlation δ' that our algorithm achieves depends only on δ (the distance to the threshold) and in particular is independent of n and k (aside from, for the latter, the dependence on k via δ). For disjoint

communities ($\alpha = 0$), our algorithm achieves constant correlation with the planted labeling if $\varepsilon^2 d/k^2$ is bounded away from 1 from below.

We conjecture that the threshold achieved by our algorithm is best-possible for polynomial-time algorithms. Concretely, if $d, \varepsilon, k, \alpha$ are constants such that $\varepsilon^2 d < k^2(\alpha+1)^2$, then we conjecture that for every polynomial-time algorithm that given a graph G outputs $\tau_1, \dots, \tau_n \in \Delta_{k-1}$,

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{(G,\sigma)\sim \mathrm{SBM}(n,d,\varepsilon,k,\alpha)} \mathrm{corr}(\sigma,\tau) = 0\,. \tag{8.1.7}$$

This conjecture is a natural extension of a conjecture for disjoint communities ($\alpha = 0$), which says that beyond the Kesten–Stigum threshold, i.e., $\varepsilon^2 d < k^2$, no polynomial-time algorithm can achieve correlation bounded away from 0 with the true labeling [135]. For large enough values of k, this conjecture predicts a computation-information gap because the condition $\varepsilon^2 d \geqslant \Omega(k \log k)$ is enough for achieving constant correlation information-theoretically (and in fact by a simple exponential-time algorithm). We discuss these ideas further in Section 8.1.4.

Comparison with previous matrix-based algorithms We offer a reinterpretation in the simple statistics framework of the algorithms of Mossel-Neeman-Sly and Abbe-Sandon. This will permit us to compare our algorithm for the mixed-membership model with what could be achieved by the methods in these prior works, and to point out one respect in which our algorithm improves on previous ones even for the disjoint-communities block model. The result we discuss here is a slightly generalized version of Theorem 8.1.2.

Let $\mathcal{U}$ be a (possibly infinite or continuous) universe of labels, and let W

assign to every $x, y \in \mathcal{U}$ a nonnegative real number $W(x,y) = W(y,x) \geqslant 0$. Let μ be a probability distribution on $\mathcal{U}$, which induces the inner product of functions $f, g : \mathcal{U} \to \mathbb{R}$ given by $\langle f, g \rangle = \mathbb{E}_{x \sim \mu} f(x)g(x)$. The function W can be considered as linear operator on $\{f : \mathcal{U} \to \mathbb{R}\}$, and under mild assumptions it has eigenvalues $\lambda_1, \lambda_2, \ldots$ with respect to the inner product $\langle \cdot, \cdot \rangle$.

The pair μ, W along with an average degree parameter d induce a generalized stochastic block model, where labels for nodes are drawn from μ and an edge between a pair of nodes with labels x and y is present with probability $\frac{d}{n} \cdot W(x,y)$. When $\mathcal{U}$ is Δ_{k-1} and μ is the Dirichlet distribution, this captures the mixed-membership block model.

Assume $\lambda_1 = 1$ and that μ and W are sufficiently *nice* (see Section 8.4 for all the details). Then one can rephrase results of Abbe and Sandon in this setting as follows.

Theorem 8.1.7 (Implicit in [3])**.** *Suppose the operator W has eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r$ (each possibly with higher multiplicity) and $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{d\lambda_2^2} > 0$. Let Π be the projector to the second eigenspace of the operator W. For types $x_1, \ldots, x_n \sim \mathcal{U}$, let $A \in \mathbb{R}^{n \times n}$ be the random matrix $A_{ij} = \Pi(x_i, x_j)$, where we abuse notation and think of $\Pi \colon \mathcal{U} \times \mathcal{U} \to \mathbb{R}$. There is an algorithm with running time $n^{\mathrm{poly}(1/\delta)}$ which outputs an $n \times n$ matrix P such that for $x, G \sim G(n, d, W, \mu)$,*

$$\mathop{\mathbb{E}}_{x,G} \operatorname{Tr} P \cdot A \geqslant \delta^{O(1)} \cdot (\mathop{\mathbb{E}}_{x,G} \|A\|^2)^{1/2} (\mathop{\mathbb{E}}_{x,G} \|P\|^2)^{1/2}.$$

In one way or another, existing algorithms for the block model in the constant-degree regime are all based on estimating the random matrix A from the above theorem, then extracting from an estimator for A some labeling of vertices by communities. In our mixed-membership setting, one may show that the matrix

A is $\sum_{s\in[k]} v_s v_s^\mathsf{T}$, where $v_s \in \mathbb{R}^n$ has entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Furthermore, as we show in Section 8.4, the condition $d\lambda_2^2 > 1$ translates for the mixed-membership model to the condition $\varepsilon^2 d > k(\alpha+1)^2$, which means that under the same hypotheses as our main theorem on the mixed-membership model it is possible in polynomial time to evaluate a constant-correlation estimator for $\sum_{s\in[k]} v_s v_s^\mathsf{T}$. As we discussed in Section 8.1.1, however, extracting estimates for $v_1, \ldots, v_k$ (or, almost equivalently, estimates for $\sigma_1, \ldots, \sigma_n$) from this matrix seems to incur an inherent $1/k$ loss in the correlation. Thus, the final guarantee one could obtain for the mixed-membership block model using the techniques in previous work would be estimates $\tau_1, \ldots, \tau_n$ for $\sigma_1, \ldots, \sigma_n$ such that $\mathrm{corr}(\sigma, \tau) \geqslant \left(\frac{\delta}{k}\right)^{O(1)}$.[14] We avoid this loss in our main theorem via tensor methods.

Although this $1/k$ multiplicative loss in the correlation with the underlying labeling is not inherent in the disjoint-community setting (roughly speaking this is because the matrix A is a 0/1 block-diagonal matrix), previous algorithms nonetheless incur such loss. (In part this is related to the generality of the work of Abbe and Sandon: they aim to allow W where A might only have rank one, while in our settings A always has rank $k-1$. For low-rank A this $1/k$ loss is probably necessary for polynomial time algorithms.)

Thus our main theorem on the mixed membership model offers an improvement on the guarantees in the previous literature even for the disjoint-communities setting: when W only has entries $1-\varepsilon$ and ε we obtain a labeling of the vertices whose correlation with the underlying labeling depends only on

[14] In fact, it is not clear one can obtain even this guarantee using strictly matrix methods. Strictly speaking, in estimating, say, v_1 using the above described matrix method, one obtains a unit vector v such that $\langle v, v_1\rangle^2 \geqslant \Omega(1) \cdot \|v_1\|^2$. Without knowing whether v or $-v$ is the correct vector it is not clear how to transform estimates for the v_s's to estimates for the σ's. However, matrix methods cannot distinguish between v_s and $-v_s$. In our main algorithm we avoid this issue because the 3rd moments $\sum v_s^{\otimes 3}$ are not sign-invariant.

δ. This allows the number k of communities to grow with n without incurring any loss in the correlation (so long as the average degree of the graph grows accordingly).

For further discussion of the these results and a proof of the above theorem, see Section 8.4.

Comparison to previous tensor algorithm for mixed-membership models Above we discussed a reinterpretation (allowing a continuous space $\mathcal{U}$ of labels) of existing algorithms for the constant-average-degree block model which would give an algorithm for the mixed-membership model, and discussed the advantages of our algorithm over this one. Now we turn to algorithms in the literature which are specifically designed for stochastic block models with overlapping communities.

The best such algorithm requires $\varepsilon^2 d \geqslant O(\log n)^{O(1)} \cdot k^2(\alpha+1)^2$ [12]. Our bound saves the $O(\log n)^{O(1)}$ factor. (This situation is analogous to the standard block model, where simpler algorithms based on eigenvectors of the adjacency matrix require the graph degree to be logarithmic.) Notably, like ours this algorithm is based on estimating the tensor $T = \sum_{s\in[k]} v_s^{\otimes 3}$, where $v_s \in \mathbb{R}^n$ has entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. However, the algorithm differs from ours in two key respects.

1. The algorithm [12] estimates the tensor T using a 3-tensor analogue of a high power of the adjacency matrix of an input graph, while we use self-avoiding walks (which are rather like a tensor analogue of the nonbacktracking operator).
2. The tensor decomposition algorithm used in [12] to decompose the (estima-

tor for the) tensor T tolerates much less error than our tensor decomposition algorithm; the result is that a higher-degree graph is needed in order to obtain a better estimator for the tensor T.

The setting considered by [12] does allow a more sophisticated version of the Dirichlet distribution than we allow, in which different communities have different sizes. It is an interesting open problem to extend the guarantees of our algorithm to that setting.

8.1.3 Low-correlation tensor decomposition

Tensor decomposition is the following problem. For some unit vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ and a constant k (often $k = 3$ or 4), one is given the tensor $T = \sum_{i=1}^{m} a_i^{\otimes k} + E$, where E is some error tensor. The goal is to recover vectors $b_1, \ldots, b_m \in \mathbb{R}^n$ which are as close as possible to $a_1, \ldots, a_m$.

Tensor decomposition has become a common primitive used by algorithms for parameter learning and estimation problems [51, 13, 76, 79, 31, 121, 161]. In the simplest examples, the hidden variables are orthogonal vectors $a_1, \ldots, a_m$ and there is a simple function of the observed variables which estimates the tensor $\sum_{i \leqslant m} a_i^{\otimes k}$ (often an empirical k-th moment of observed variables suffices). Applying a tensor decomposition algorithm to such an estimate yields estimates of the vectors $a_1, \ldots, a_m$.

We focus on the case that $a_1, \ldots, a_m$ are orthonormal. Algorithms for this case are already useful for a variety of learning problems, and it is often possible to reduce more complicated problems to the orthonormal case using a small amount

of side information about $a_1, \ldots, a_m$ (in particular their covariance $\sum_{i=1}^m a_i a_i^\mathsf{T}$). In this setting the critical question is: how much error E (and measured in what way) can the tensor decomposition algorithm tolerate and still produce useful outputs $b_1, \ldots, b_m$?

When we use tensor decomposition in our meta-algorithm, the error E will be incurred when estimating $\sum_{i=1}^m a_i^{\otimes k}$ from observable samples. Using more samples would decrease the magnitude of $T - \sum_{i=1}^m a_i^{\otimes k}$, but because our goal is to obtain algorithms with optimal sample complexity we need a tensor decomposition algorithm which is robust to greater errors than those in the existing literature.

Our main theorem on tensor decomposition is the following.

Theorem 8.1.8 (Informal)**.** *For every $\delta > 0$, there is a randomized algorithm with running time $n^{1/\mathrm{poly}(\delta)}$ that given a 3-tensor $T \in (\mathbb{R}^n)^{\otimes 3}$ and a parameter k outputs $n^{\mathrm{poly}(1/\delta)}$ unit vectors $b_1, \ldots, b_m$ with the following property: if T satisfies $\langle T, \sum_{i=1}^k a_i^{\otimes 3} \rangle \geqslant \delta \cdot \|T\| \cdot \sqrt{k}$ for some orthonormal vectors $a_1, \ldots, a_k$, then*

$$\mathop{\mathbb{E}}_{i \sim [k]} \max_{j \in [m]} \langle a_i, b_j \rangle^2 \geqslant \delta^{O(1)} .$$

Furthermore, if the algorithm is allowed to make $n^{1/\mathrm{poly}(\delta)}$ calls to an oracle $\mathcal{O}$ which correctly answers queries of the form "does the unit vector v satisfy $\sum_{i=1}^m \langle a_i, v \rangle^4 \geqslant \delta^{O(1)}$?", then it outputs just k orthonormal vectors, $b_1, \ldots, b_k$ such that there is a permutation $\pi : [k] \to [k]$ with

$$\mathop{\mathbb{E}}_{i \in [k]} \langle a_i, b_{\pi(i)} \rangle^2 \geqslant \delta^{O(1)} .$$

(These guarantees hold in expectation over the randomness used in the decomposition algorithm.)

(For a more formal statement, and in particular the formal requirements for the oracle $\mathcal{O}$, see Section 8.7.)

Rescaling T as necessary, one may reinterpret the condition $\langle T, \sum_{i=1}^{k} a_i^{\otimes 3} \rangle \geqslant \delta \cdot \|T\| \cdot \sqrt{k}$ as $T = \sum_{i=1}^{k} a_i^{\otimes 3} + E$, where $\langle E, \sum_{i=1}^{m} a_i^{\otimes 3} \rangle = 0$ and $\|E\| \leqslant \sqrt{k}/\delta$ and $\|\cdot\|$ is the Euclidean norm. In particular, E may have Euclidean norm which is a large constant factor $1/\delta$ larger than the Euclidean norm of the tensor $\sum_{i=1}^{m} a_i^{\otimes 3}$ that the algorithm is trying to decompose! (One way such error could arise is if T is actually correlated with $1/\delta$ unrelated orthogonal tensors; our algorithm in that case ensures that the list of outputs vectors is correlated with every one of these orthogonal tensors.)

In all previous algorithms of which we are aware (even for the case of orthogonal $a_1, \ldots, a_m$), the error E must have spectral norm (after flattening to an $n^2 \times n^2$ matrix) at most ε for $\varepsilon < \frac{1}{2}$,[15] or E must have Euclidean norm at most $\varepsilon\sqrt{m}$ [161]. The second requirement is strictly stronger than ours (thus our algorithm has weaker requirements and so stronger guarantees). The first, on the spectral norm of E when flattened to a matrix, is incomparable to the condition in our theorem. However, when E satisfies such a spectral bound it is possible to decompose T using (sophisticated) spectral methods [121, 161]. In our setting such methods seem unable to avoid producing only vectors b which are correlated with E but not with any $a_1, \ldots, a_m$. In other words, such methods would *overfit to the error* E. To avoid this, our algorithm uses a novel maximum-entropy convex program (see Section 8.7 for details).

One a priori unusual requirement of our tensor decomposition algorithm is access to the oracle $\mathcal{O}$. In any tensor decomposition setting where E satisfies $\|E\|_{inj} = \max_{\|x\|=1} \langle E, x^{\otimes 3} \rangle \leqslant o(1)$, the oracle $\mathcal{O}$ can be implemented just be evaluating $\langle T, v^{\otimes 3} \rangle = \sum_{i=1}^{k} \langle a_i, v \rangle^3 + o(1)$. All previous works on tensor decomposition of which we are aware either assume that the injective norm $\|E\|_{inj}$ is bounded as

[15] Or, mildly more generally, E should have SoS norm less than ε [121].

above, or (as in [161]) can accomplish this with a small amount of preprocessing on the tensor T. Our setting allows, for example, $E = 100 \cdot v^{\otimes 3}$ for some unit vector v, and does not appear to admit the possibility of such preprocessing, hence the need for an auxiliary implementation of O. In our learning applications we are able to implement O by straightforward holdout set/cross-validation methods.

8.1.4 Information-computation gaps and concrete lower bounds

In this work, we show an unconditional lower bound for simple statistics for the stochastic block model with k communities at the Kesten–Stigum threshold. For $k \geqslant 4$, this threshold is bounded away from the information-theoretic threshold [2]. In this way, our lower bounds gives evidence for an inherent gap between the information-theoretical and computational thresholds. Recall the definitions of simple statistics from Chapter 2.

Theorem 8.1.9. *Let d, ε, k be constants. Then,*

$$\max_{p \in \mathbb{R}[x]_{\leqslant \ell}} \frac{\mathbb{E}_{x \sim \mathrm{SBM}(n,d,\varepsilon,k)} p(x)}{\left(\mathbb{E}_{x \sim G(n,d/n)} p(x)^2\right)^{1/2}} = \begin{cases} \geqslant n^{\Omega(1)} & \text{if } \varepsilon^2 d > k^2,\ \ell \geqslant O(\log n) \\ \leqslant O(1) & \text{if } \varepsilon^2 d < k^2, \ell \leqslant n^{0.01} \end{cases} \tag{8.1.8}$$

The difference between $O(1)$ and $n^{\Omega(1)}$ shows the Kesten-Stigum phase transition is visible to simple statistics, offering the means to connect the Kesten-Stigum threshold to other information-computation gaps in this thesis (for spiked tensors, planted clique, and so on).

We also study low degree polynomials for estimation (as opposed to hypothesis testing against $G(n, d/n)$) at the Kesten-Stigum threshold. For this we prove the following theorem.

Theorem 8.1.10. *Let $d, \varepsilon, k, \delta$ be constants such that $\varepsilon^2 d < (1-\delta)k^2$. Let $f : \{0,1\}^{n \times n} \to \mathbb{R}$ be any function, let $i, j \in [n]$ be distinct. Then if f satisfies $\mathbb{E}_{x \sim G(n, \frac{d}{n})} f(x) = 0$ and is correlated with the indicator $\mathbf{1}_{\sigma_i = \sigma_j}$ that i and j are in the same community in the following sense:*

$$\frac{\mathbb{E}_{x \sim SBM(n,d,\varepsilon,k)} f(x)(\mathbf{1}_{\sigma_i=\sigma_j} - \frac{1}{k})}{(\mathbb{E}_{x \sim G(n,\frac{d}{n})} f(x)^2)^{1/2}} \geqslant \Omega(1)$$

then $\deg f \geqslant n^{c(d,\varepsilon,k)}$ *for some* $c(d, \varepsilon, k) > 0$.

There is one subtle difference between the polynomials ruled out by this theorem and those which could be used by our algorithmic techniques. Namely, this theorem rules out any f whose correlation with the indicator $\mathbf{1}_{\sigma_i=\sigma_j}$ is large *compared to f's standard deviation under* $G(n, d/n)$, whereas our meta-algorithm needs a polynomial f where this correlation is large compared to f's standard deviation under the block model. In implementing our approach for the block model and for other problems, we have found that these two measures of standard deviation are always equal (up to low-order additive terms) for the polynomials which turn out to provide sample-optimal constant-correlation estimators of hidden variables.

Interesting open problems are to prove a version of the above theorem where standard deviation is measured according to the block model and to formalize the idea that $\mathbb{E}_{SBM} f(x)^2$ should be related to $\mathbb{E}_{G(n,d/n)} f(x)^2$ for good estimators f. It would also be quite interesting to see how large the function $c(d, \varepsilon, k)$ can be made: the above theorem shows that when $d < (1-\delta)k^2/\varepsilon^2$ the degree of any good estimator of $\mathbf{1}_{\sigma_i=\sigma_j}$ must be polynomial in n—perhaps it must be linear, or even quadratic in n.

8.2 Techniques

To illustrate the idea of low-degree estimators for posterior moments, let's first consider the most basic stochastic block model with $k = 2$ disjoint communities ($\alpha = 0$). (Our discussion will be similar to the analysis in [138].) Let $y \in \{\pm 1\}^n$ be chosen uniformly at random and let $x \in \{0,1\}^{n \times n}$ be the adjacency matrix of a graph such that for every pair $i < j \in [n]$, we have $x_{ij} = 1$ with probability $(1 + \varepsilon y_i y_j)\frac{d}{n}$. Our goal is to find a matrix-valued low-degree polynomial $P(x)$ that correlates with yy^T. It turns out to be sufficient to construct for every pair $i, j \in [n]$ a low-degree polynomial that correlates with $y_i y_j$.

The linear polynomial $p_{ij}(x) = \frac{n}{\varepsilon d}\left(x_{ij} - \frac{d}{n}\right)$ is an unbiased estimator for $y_i y_j$ in the sense that $\mathbb{E}[p_{ij}(x) \mid y] = y_i y_j$. By itself, this estimator is not particular useful because its variance $\mathbb{E}\, p_{ij}(x)^2 \approx \frac{n}{\varepsilon^2 d}$ is much larger than the quantity $y_i y_j$ we are trying to estimate. However, if we let $\alpha \subseteq [n]^2$ be a length-ℓ path between i and j (in the complete graph), then we can combine the unbiased estimators along the path α and obtain a polynomial

$$p_\alpha(x) = \prod_{ab \in \alpha} p_{ab}(x) \tag{8.2.1}$$

that is still an unbiased estimator $\mathbb{E}[p_\alpha(x) \mid y_i, y_j] = \mathbb{E}\left[\prod_{ab \in \alpha} y_a y_b \mid y_i, y_j\right] = y_i y_j$. This estimator has much higher variance $\mathbb{E}\, p_\alpha(x)^2 \approx (\frac{n}{\varepsilon^2 d})^\ell$. But we can hope to reduce this variance by averaging over all such paths. The number of such paths is roughly $n^{\ell-1}$ (because there are $\ell - 1$ intermediate vertices to choose). Hence, if these estimators $\{p_\alpha(x)\}_\alpha$ were pairwise independent, this averaging would reduce the variance by a multiplicative factor $n^{\ell-1}$, giving us a final variance of $(\frac{n}{\varepsilon^2 d})^\ell \cdot n^{1-\ell} = (\frac{1}{\varepsilon^2 d})^\ell \cdot n$. We can see that above the Kesten–Stigum threshold, i.e., $\varepsilon^2 d \geqslant 1 + \delta$ for $\delta > 0$, this heuristic variance bound $(\frac{1}{\varepsilon^2 d})^\ell \cdot n \leqslant 1$ is

good enough for estimating the quantity $y_i \cdot y_j$ for paths of length $\ell \geqslant \log_{1+\delta} n$.

Two steps remain to turn this heuristic argument into a polynomial-time algorithm for estimating the matrix yy^T. First, it turns out to be important to consider only paths that are self-avoiding. As we will see next, estimators from such paths are pairwise independent enough to make our heuristic variance bound go through. Second, a naive evaluation of the final polynomial takes quasi-polynomial time because it has logarithmic degree (and a quasi-polynomial number of non-zero coefficients in the monomial basis). We describe the high-level ideas for avoiding quasi-polynomial running time later in this section (Section 8.2.5).

8.2.1 Approximately pairwise-independent estimators

Let $\mathrm{SAW}_\ell(i, j)$ be the set of self-avoiding walks $\alpha \subseteq [n]^2$ of length ℓ between i and j. Consider the unbiased estimator $p(x) = \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x)$ for $y_i y_j$. Above the Kesten–Stigum threshold and for $\ell \geqslant O(\log n)$, we can use the following lemma to show that $p(x)$ has variance $O(1)$ and achieves constant correlation with $z = y_i y_j$. We remark that the previous heuristic variance bound corresponds to the contribution of the terms with $\alpha = \beta$ in the left-hand side of Eq. (8.2.2).

Lemma 8.2.1 (Constant-correlation estimator)**.** *Let (x, z) be distributed over $\{0,1\}^n \times \mathbb{R}$. Let $\{p_\alpha\}_{\alpha \in \mathcal{I}}$ be a collection of real-valued n-variate polynomials with the following properties:*

1. *unbiased estimators: $\mathbb{E}[p_\alpha(x) \mid z] = z$ for every $\alpha \in \mathcal{I}$*

2. *approximate pairwise independence: for* $\delta > 0$,

$$\sum_{\alpha,\beta\in\mathcal{I}} \mathbb{E}\, p_\alpha(x)\cdot p_\beta(x) \leqslant \tfrac{1}{\delta^2}\cdot|\mathcal{I}|^2\,\mathbb{E}\, z^2 \tag{8.2.2}$$

Then, the polynomial $p = \frac{1}{|\mathcal{I}|}\sum_{\alpha\in\mathcal{I}} p_\alpha$ *satisfies* $\mathbb{E}\, p(x)\cdot z \geqslant \delta\cdot\left(\mathbb{E}\, p(x)^2\cdot\mathbb{E}\, z^2\right)^{1/2}$.

Remark 8.2.2. In applying the lemma we often substitute for Eq. (8.2.2) the equivalent condition

$$\mathbb{E}\, z^2\cdot\sum_{\alpha,\beta\in\mathcal{I}}\mathbb{E}\, p_\alpha(x)\cdot p_\beta(x) \leqslant \frac{1}{\delta^2}\cdot\sum_{\alpha,\beta\in\mathcal{I}}(\mathbb{E}\, p_\alpha(x)z)\cdot(\mathbb{E}\, p_\beta(x)z)$$

which is conveniently invariant to rescaling of the p_α's.

Proof. Since the polynomial p is an unbiased estimator for z, we have $\mathbb{E}\, p(x)z = \mathbb{E}\, z^2$. By Eq. (8.2.2), $\mathbb{E}\, p(x)^2 \leqslant (1/\delta^2)\cdot\mathbb{E}\, z^2$. Taken together, we obtain the desired conclusion. □

In Section 8.3.1, we present the short combinatorial argument that shows that above the Kesten–Stigum bound the estimators for self-avoiding walks satisfy the conditions Eq. (8.2.2) of the lemma.

We remark that if instead of self-avoiding walks we were to average over all length-ℓ walks between i and j, then the polynomial $p(x)$ computes up to scaling nothing but the (i,j)-entry of the ℓ-th power of the centered adjacency $x - \frac{d}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. For $\ell \approx \log n$, the ℓ-th power of this matrix converges to vv^T, where v is the top eigenvector of the centered adjacency matrix. For constant degree $d = O(\log n)$, it is well-known that this eigenvector fails to provide a good approximation to the true labeling. In particular, the corresponding polynomial fails to satisfy the conditions of Lemma 8.2.1 close to the Kesten–Stigum threshold.

8.2.2 Low-degree estimators for higher-order moments

Let's turn to the general mixed-membership stochastic block model SBM$(n, d, \varepsilon, k, \alpha_0)$. Let (G, σ) be graph G and community structure $\sigma = (\sigma_1, \ldots, \sigma_n)$ drawn from this model. Recall that $\sigma_1, \ldots, \sigma_n$ are k-dimensional probability vectors, each roughly uniform over $\alpha_0 + 1$ of the coordinates. Let $x \in \{0,1\}^{n \times n}$ be the adjacency matrix of G and let $y_1, \ldots, y_k \in \mathbb{R}^n$ be centered community indicator vectors, so that $(y_s)_i = (\sigma_i)_s - \frac{1}{k}$.

It's instructive to see that, unlike for disjoint communities, second moments are not that useful for overlapping communities. As a thought experiment suppose we are given the matrix $\sum_{s=1}^{k} (y_s)(y_s)^\intercal$ (which we can estimate using the path polynomials described earlier).

In case of disjoint communities, this matrix allows us to "read off" the community structure directly (because two vertices are in the same community if and only if the entry in the matrix is $1 - O(1/k)$).

For overlapping communities (say the extreme case $\alpha_0 \gg k$ for simplicity), we can think of each σ_i as a random perturbation of the uniform distribution so that $(\sigma_i)_s = (1 + \xi_{i,s})\frac{1}{k}$ for iid Gaussians $\{\xi_{i,s}\}$ with small variance. Then, the centered community indicator vectors $y_1, \ldots, y_k$ are iid centered, spherical Gaussian vectors. In particular, the covariance matrix $\sum_{s=1}^{k} y_s y_s^\intercal$ essentially only determines the subspace spanned by the vectors $y_1, \ldots, y_k$ but not the vectors themselves. (This phenomenon is sometimes called the "rotation problem" for matrix factorizations.)

In contrast, classical factor analysis results show that if we were given the third moment tensor $\sum_{s=1}^{k} y_s^{\otimes 3}$, we could efficiently reconstruct the vectors $y_1, \ldots, y_k$

[85, 118]. This fact is the reason for aiming to estimate higher order moments in order to recover overlapping communities.

In the same way that a single edge $x_{i,j} - \frac{d}{n}$ gives an unbiased estimator for the (i, j)-entry of the second moment matrix, a 3-star $(x_{i,c} - \frac{d}{n})(x_{j,c} - \frac{d}{n})(x_{k,c} - \frac{d}{n})$ gives an unbiased estimator for the (i, j, k)-entry of the third moment tensor $\sum_{s=1}^{k} y_s^{\otimes 3}$. This observation is key for the previous best algorithm for mixed-membership community detection [12]. However, even after averaging over all possible centers c, the variance of this estimator is far too large for sparse graphs. In order to decrease this variance, previous algorithms [12] project the tensor to the top eigenspace of the centered adjacency matrix of the graph. In terms of polynomial estimators this projection corresponds to averaging over all length-ℓ-armed 3-stars[16] for $\ell = \log n$. Even for disjoint communities, this polynomial estimator would fail to achieve the Kesten–Stigum bound.

In order to improve the quality of this polynomial estimator, informed by the shape of threshold-achieving estimator for second moments, we average only over such long-armed 3-stars that are self-avoiding. We show that the resulting estimator achieves constant correlation with the desired third moment tensor precisely up to the Kesten–Stigum bound (Section 8.5.2).

8.2.3 Correlation-preserving projection

A recurring theme in our algorithms is that we can compute an approximation vector P that is correlated with some unknown ground-truth vector Y in the Euclidean sense $\langle P, Y \rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$, where the norm $\|\cdot\|$ is induced by the inner

[16] A length-ℓ-armed 3-star between $i, j, k \in [n]$ consists of three length-ℓ walks between i, j, k and a common center $c \in [n]$

product $\langle \cdot, \cdot \rangle$. (Typically, we obtain P by evaluating a low-degree polynomial in the observable variables and Y is the second or third moment of the hidden variables.)

In this situation, we often seek to improve the quality of the approximation P—not in the sense of increasing the correlation, but in the sense of finding a new approximation Q that is "more similar" to Y while roughly preserving the correlation, so that $\langle Q, Y \rangle \geqslant \delta^{O(1)} \cdot \|Q\| \cdot \|Y\|$. As a concrete example, we may know that Y is a positive semidefinite matrix with all-ones on the diagonal and our goal is to take an arbitrary matrix P correlated with Y and compute a new matrix Q that is still correlated with Y but in addition is positive semidefinite and has all-ones on the diagonal. More generally, we may know that Y is contained in some convex set C and the goal is "project" P into the set C while preserving the correlation. We note that the perhaps most natural choice of Q as the vector closest to P in C does not work in general. (For example, if $Y = (1, 0)$, $C = \{(a, b) \mid a \leqslant 1\}$, and $P = (\delta \cdot M, M)$, then the closest vector to P in C is $(1, M)$, which has poor correlation with Y for large M.)

Theorem 8.2.3 (Correlation-preserving projection)**.** *Let C be a convex set and $Y \in C$. Let P be a vector with $\langle P, Y \rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$. Then, if we let Q be the vector that minimizes $\|Q\|$ subject to $Q \in C$ and $\langle P, Q \rangle \geqslant \delta \cdot \|P\| \cdot \|Y\|$, we have*

$$\langle Q, Y \rangle \geqslant \delta/2 \cdot \|Q\| \cdot \|Y\| . \tag{8.2.3}$$

Furthermore, Q satisfies $\|Q\| \geqslant \delta \|Y\|$.

Proof. By construction, Q is the Euclidean projection of 0 into the set $C' := \{Q \in C \mid \langle P, Q \rangle \geqslant \delta \|P\| \cdot \|Y\|\}$. It's a basic geometric fact (sometimes called Pythagorean inequality) that a Euclidean projection into a set decreases distances

to points into the set. Therefore, $\|Y - Q\|^2 \leqslant \|Y - 0\|^2$ (using that $Y \in C'$). Thus, $\langle Y, Q\rangle \geqslant \|Q\|^2/2$. On the other hand, $\langle P, Q\rangle \geqslant \delta\|P\| \cdot \|Y\|$ means that $\|Q\| \geqslant \delta\|Y\|$ by Cauchy–Schwarz. We conclude $\langle Y, Q\rangle \geqslant \delta/2 \cdot \|Y\| \cdot \|Q\|$. □

In our applications the convex set C typically consists of probability distributions or similar objects (for example, quantum analogues like density matrices or pseudo-distributions—the sum-of-squares analogue of distributions). Then, the norm minimization in Theorem 8.2.3 can be viewed as maximizing the Rényi entropy of the distribution Q. From this perspective, maximizing the entropy within the set C' ensures that the correlation with Y is not lost.

8.2.4 Low-correlation tensor decomposition

Earlier we described how to efficiently compute a 3-tensor P that has correlation $\delta > 0$ with a 3-tensor $\sum_{i=1}^k y_i^{\otimes 3}$, where $y_1, \ldots, y_k$ are unknown orthonormal vectors we want to estimate (Section 8.2.2). Here, the correlation δ depends on how far we are from the threshold and may be minuscule (say 0.001).

It remains to decompose the tensor P into a short list of vectors L so as to ensure that $\mathbb{E}_{i\in[k]} \max_{\widehat{y}\in L} \langle \widehat{y}, y\rangle \geqslant \delta^{O(1)}$. (Ideally of course $|L| = k$. In the block model context this guarantee requires a small amount of additional work to cross-validate vectors in a larger list.) To the best of our knowledge, previous tensor decomposition algorithms do not achieve this kind of guarantee and require that the correlation of P with the orthogonal tensor $\sum_{i=1}^k y_i^{\otimes 3}$ is close to 1 (sometimes even within polynomial factors $1/n^{O(1)}$).

In the current work, we achieve this guarantee building on previous sum-

of-squares based tensor decomposition algorithms [31, 121]. These algorithms optimize over moments of pseudo-distributions (a generalization of probability distributions) and then apply Jennrich's classical tensor decomposition algorithms to these "pseudo-moments". The advantage of this approach is that it provably works even in situations where Jennrich's algorithm fails when applied to the original tensor.

As a thought experiment, suppose we are able to find pseudo-moments M that are correlated with the orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$. Extending previous techniques [121], we show that Jennrich's algorithm applied to M is able to recover vectors that have constant correlation with a constant fraction of the vectors $y_1, \ldots, y_k$.

A priori it is not clear how to find such pseudo-moments M because we don't know the orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$, we only know a 3-tensor P that is slightly correlated with it. Here, the correlation-preserving projection discussed in the previous section comes in: by Theorem 8.2.3 we can efficiently project P into the set of pseudo-moments in a way that preserves correlation. In this way, we obtain pseudo-moments M that are correlated with the unknown orthogonal tensor $\sum_{i=1}^{k} y_i^{\otimes 3}$.

When P is a 3-tensor as above, we encounter technical difficulties inherent to odd-order tensors. (This is a common phenomenon in the tensor-algorithms literature.) To avoid these difficulties we give a simple algorithm, again using the correlation-preserving projection idea, to lift a 3-tensor P which is δ-correlated with an orthogonal tensor A to a 4-tensor P' which is $\delta^{O(1)}$-correlated with an appropriate orthogonal 4-tensor. See Section 8.7.2.

8.2.5 From quasi-polynomial time to polynomial time

In this section, we describe how to evaluate certain logarithmic-degree polynomials in polynomial-time (as opposed to quasi-polynomial time). The idea is to use color coding [11].[17]

For a coloring $c\colon [n] \to [\ell]$ and a subgraph $\alpha \subseteq [n]^2$ on ℓ vertices, let $F_{c,\alpha} = \frac{\ell^\ell}{\ell!} \cdot \mathbf{1}_{c(\alpha)=[\ell]}$ be a scaled indicator variable of the event that α is colorful.

Theorem 8.2.4 (Evaluating colorful-path polynomials)**.** *There exists a $n^{O(1)} \cdot \exp(\ell)$-time algorithm that given vertices $i, j \in [n]$, a coloring $c\colon [n] \to [\ell]$ and an adjacency matrix $x \in \{0,1\}^{n \times n}$ evaluates the polynomial*

$$p_c(x) := \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x) \cdot F_{c,\alpha}\,. \tag{8.2.4}$$

(Here, $p_\alpha \propto \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$ is the polynomial in Eq. (8.2.1).)

Proof. We can reduce this problem to computing the ℓ-th power of the following $n \cdot 2^\ell$-by-$n \cdot 2^\ell$ matrix: The rows and columns are indexed by pairs (a, S) of vertices $a \in [n]$ and color sets $S \subseteq [\ell]$. The entry for column (a, S) and row (b, T) is equal to $x_{ab} - \frac{d}{n}$ if $T = S \cup \{c(a)\}$ and 0 otherwise. If we compute the ℓ-th power of this matrix, then the entry for column $(i, \emptyset)$ and row $(j, [\ell])$ is the sum over all colorful ℓ-paths from i to j. □

For a fixed coloring c, the polynomial p_c does not provide a good approximation for the polynomial $p(x) := \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x)$. In order to get a good approximation, we will choose random colorings and average over them.

[17] We thank Avi Wigderson for suggesting that color coding may be helpful in this context.

If we let c be a random coloring, then by construction $\mathbb{E}_c F_{c,\alpha} = 1$ for every simple ℓ-path α. Therefore, $\mathbb{E}_c p_c(x) = p(x)$ for every $x \in \{0,1\}^{n \times n}$. We would like to estimate the variance of $p_c(x)$. Here, it turns out to be important to consider a typical x drawn from stochastic block model distribution SBM.

$$\mathop{\mathbb{E}}_{x \sim \mathrm{SBM}(n,d,\varepsilon)} \mathop{\mathbb{E}}_{c} p_c(x)^2 = \frac{1}{|\mathrm{SAW}_\ell(i,j)|^2} \sum_{\alpha,\beta \in \mathrm{SAW}_\ell(i,j)} \mathop{\mathbb{E}}_{c} F_{c,\alpha} \cdot F_{c,\beta} \cdot \mathop{\mathbb{E}}_{x \sim \mathrm{SBM}} p_\alpha(x) p_\beta(x) \tag{8.2.5}$$

$$\leqslant e^{2\ell} \cdot \frac{1}{|\mathrm{SAW}_\ell(i,j)|} \sum_{\alpha,\beta \in \mathrm{SAW}_\ell(i,j)} \left| \mathop{\mathbb{E}}_{x} p_\alpha(x) p_\beta(x) \right| . \tag{8.2.6}$$

For the last step, we use that $\mathbb{E}_c F_{c,\alpha}^2 \leqslant e^{2\ell}$ (because $\ell^\ell / \ell! \leqslant e^\ell$).

The right-hand side of Eq. (8.2.6) corresponds precisely to our notion of approximate pairwise independence in Lemma 8.2.1. Therefore, if we are within the Kesten–Stigum bound, $\varepsilon^2 d \geqslant 1 + \delta$, the right-hand side of Eq. (8.2.6) is bounded by $e^{2\ell} \cdot 1/\delta^{O(1)}$.

We conclude that with high probability over x, the variance of $p_c(x)$ for random c is bounded by $e^{O(\ell)}$. It follows that by averaging over $e^{O(\ell)}$ random colorings we obtain a low-variance estimator for $p(x)$.

8.2.6 Illustration: push-out effect in spiked Wigner matrices

We turn to a first demonstration of our meta-algorithm beyond the stochastic block model: deriving the critical signal-to-noise ratio for (Gaussian) Wigner matrices (i.e. symmetric matrices with iid entries) with rank-one spikes. This section demonstrates the use of Theorem 8.1.1; more sophisticated versions of the same ideas (for example our 3rd-moment meta-theorem, Theorem 8.1.3) will be used in the course of our block model algorithms.

Consider the following Bayesian estimation problem: We are given a spiked Wigner matrix $A = \lambda vv^\mathsf{T} + W$ so that W is a random symmetric matrix with Gaussian entries $W_{ij} \sim \mathcal{N}(0, \frac{1}{n})$ and $v \sim \mathcal{N}(0, \frac{1}{n}\mathrm{Id})$. The goal is to estimate v, i.e., compute a unit vector $\widehat{v}$ so that $\langle v, \widehat{v}\rangle^2 \geqslant \Omega(1)$. Since the spectral norm of a Wigner matrix satisfies $\mathbb{E}\|W\| = \sqrt{2}$, it follows that for $\lambda > \sqrt{2}$, the top eigenvector $\widehat{v}$ of A satisfies $\langle v, \widehat{v}\rangle^2 \geqslant \Omega(1)$. However, it turns out that we can estimate the spike v even for smaller values of λ: a remarkable property of spiked Wigner matrices is that as soon as $\lambda > 1$, the top eigenvector $\widehat{v}$ becomes correlated with the spike v [22]. (This property is sometimes called the "pushout effect".)

Unfortunately known proofs of this property a quite involved. In the following, we apply Theorem 8.1.1 to give an alternative proof of the fact that it is possible to efficiently estimate the spike v as soon as $\lambda > 1$. Our algorithm is more involved and less efficient than computing the top eigenvector of A. The advantage is that its analysis is substantially simpler compared to previous analyses.

Theorem 8.2.5 (implicit in [22]). *If $\lambda = 1 + \delta$ for some $1 > \delta > 0$, there is a degree $\delta^{-O(1)} \cdot \log n$ matrix-valued polynomial $f(A) = \{f_{ij}(A)\}_{ij \leqslant n}$ such that*

$$\frac{\mathbb{E}_{W,v} \operatorname{Tr} f(A) vv^\top}{(\mathbb{E}\,\|f(A)\|_F^2)^{1/2} \cdot (\mathbb{E}\,\|vv^\top\|_F^2)^{1/2}} \geqslant \delta^{O(1)}\,.$$

Together with Theorem 8.1.1, the above theorem gives an algorithm with running time $n^{\log n/\delta^{O(1)}}$ to find $\widehat{v}$ with nontrivial $\mathbb{E}\langle \widehat{v}, v\rangle^2$.[18]

The analysis of [22] establishes the above theorem for the polynomial $f(A) = A^\ell$ with $\ell = \delta^{-O(1)} \cdot \log n$. Our proof chooses a different polynomial, which

[18] While this algorithm is much slower than the eigenvector-based algorithm—even after using color coding to improve the $n^{\log n/\delta^{O(1)}}$ running time to $n^{1/\delta^{O(1)}}$—the latter requires many sophisticated innovations and ideas from random matrix theory. This algorithm, by contrast, can be derived and analyzed with our meta-theorem, little innovation required.

affords a substantially simpler analysis.

Proof of Theorem 8.2.5. For $\alpha \subseteq \binom{n}{2}$, let $\chi_\alpha(A) = \prod_{\{i,j\}\in\alpha} A_{ij}$. Let $L = \log n/\delta^C$ for C a large enough constant. For $ij \in [n]$, let $SAW_{ij}(L)$ be the collection of all self-avoiding paths from i to j in the complete graph on n vertices. Observe that $\frac{n^{L-1}}{\lambda^L}\chi_\alpha$ for $\alpha \in SAW_{ij}(L)$ is an unbiased estimator of $v_i v_j$:

$$\mathbb{E}\left[\chi_\alpha(A) \mid v_i, v_j\right] = \mathop{\mathbb{E}}_{v}\left[\prod_{k\ell\in\alpha}\mathop{\mathbb{E}}_{W}(W_{k\ell} + \lambda v_k v_\ell) \mid v_i, v_j\right] = \lambda^L v_i v_j \,\mathbb{E}\prod_{k\in\alpha\setminus\{i,j\}} v_k^2 = \frac{\lambda^L}{n^{L-1}}\cdot v_i v_j\,.$$

We further claim that the collection $\{\frac{n^{L-1}}{\lambda^L}\chi_\alpha\}_{\alpha\in SAW_{ij}(L)}$ is approximately pairwise independent in the sense of Lemma 8.2.1. To show this we must check that

$$\frac{n^{2(L-1)}}{\lambda^{2L}}\sum_{\alpha,\beta}\mathbb{E}\,\chi_\alpha\chi_\beta \leqslant \frac{1}{\delta^2}|SAW_{ij}(L)|^2\,\mathbb{E}\,v_i^2 v_j^2 = \frac{1}{\delta^2}|SAW_{ij}(L)|^2\cdot\frac{1}{n^2}\,.$$

The dominant contributers to the sum are α, β which intersect only on the vertices i and j. In that case,

$$\frac{n^{2(L-1)}}{\lambda^{2L}}\,\mathbb{E}\,\chi_\alpha\chi_\beta = n^{2(L-1)}\,\mathbb{E}\prod_{k\in\alpha\cup\beta} v_k^2 = \mathbb{E}\,v_i^2 v_j^2\,.$$

The only other terms which might contribute to the same order are α, β such that $\alpha\cap\beta$ is a union of two paths, one starting at i and one at j. If the lengths of these paths are t and t', respectively, and $t' + t' < L$, then

$$\frac{n^{2(L-1)}}{\lambda^{2L}}\,\mathbb{E}\,\chi_\alpha\chi_\beta = \frac{n^{2(L-1)}}{\lambda^{2(t+t')}}\mathop{\mathbb{E}}_{v}\left[\prod_{(k,\ell)\in\alpha\cap\beta}(\mathop{\mathbb{E}}_{W}A_{k\ell}^2)\cdot\prod_{(k,\ell)\in\alpha\triangle\beta} v_k v_\ell\right] = \frac{n^{t+t'}}{\lambda^{t+t'}}\cdot(1+O(\lambda^2/n))^{t+t'}$$

where we have used that $\mathbb{E}\left[A_{k\ell}^2 \mid v_k, v_\ell\right] = \frac{1}{n}(1 + O(\lambda^2/n))\cdot\mathbb{E}\,v_i^2 v_j^2$.

There are at most $|SAW_{ij}(L)|^2/n^{t+t'}$ choices for such pairs α, β, so long as $t + t' < L$. If $t + t' = L$, then there are n times more choices than the above bound. All together,

$$\frac{n^{2(L-1)}}{\lambda^{2L}}\sum_{\alpha,\beta\in SAW_{ij}(L)}\mathbb{E}\,\chi_\alpha\chi_\beta \leqslant |SAW_{ij}(L)|\cdot\left(\left(\sum_{t=0}^{L}\frac{1}{\lambda^t}\right)^2 + \frac{n}{\lambda^L}\right)\cdot\mathbb{E}\,v_i^2 v_j^2 \leqslant \frac{1+o(1)}{1-1/\lambda}\cdot|SAW_{ij}(L)|\cdot\mathbb{E}\,v_i^2 v_j^2$$

where we have used that $\lambda = 1+\delta > 1$ and chosen C large enough that $n/\lambda^L \leqslant 1/n$. Rewriting in terms of $\delta = \lambda - 1$ and applying Lemma 8.2.1 finishes the proof. □

8.3 Warmup: stochastic block model with two communities

We demonstrate our meta-algorithm by applying it to the two-community stochastic block model. The algorithm achieves here the same threshold for partial recovery as the best previous algorithms [136, 125], which is also known to be the information-theoretic threshold [139].

While the original works involved a great deal of ingenuity, the merit of our techniques is to provide a simple and automatic way to discover and analyze an algorithm achieving the same guarantees.

Definition 8.3.1 (Two-community stochastic block model)**.** For parameters $\varepsilon, d > 0$, let $\mathrm{SBM}(n, d, \varepsilon)$ be the following distribution on pairs (x, y) where $x \in \{0,1\}^{\binom{n}{2}}$ is the adjacency matrix of an n-vertex graph and $y \in \{\pm 1\}^n$ is a labeling of the n vertices. First, sample $y \sim \{\pm 1\}^n$ uniformly. Then, independently for every pair $i < j$, add the edge $\{i, j\}$ with probability $(1+\varepsilon)\frac{d}{n}$ if $y_i = y_j$ and with probability $(1-\varepsilon)\frac{d}{n}$ if $y_i \neq y_j$.

The following theorem gives the best bounds for polynomial-time algorithms for partial recovery in this model. (We remark that the algorithms in [136, 124] actually run in time close to linear. In this work, we content ourselves with coarser running time bounds.)

Theorem 8.3.2 ([136, 124])**.** *Let $\varepsilon \in \mathbb{R}$, $d \in \mathbb{N}$ with $\delta := 1 - \frac{1}{\varepsilon^2 d}$ and $d \leqslant n^{o(1)}$. Then, there exists a randomized polynomial-time algorithm A that given a graph $x \in \{0,1\}^{\binom{n}{2}}$*

outputs a labeling $\tilde{y}(x)$ such that for all sufficiently large $n \geqslant n_0(\varepsilon, d)$,

$$\mathop{\mathbb{E}}_{(x,y)\sim \mathrm{SBM}(n,d,\varepsilon)} \langle \tilde{y}(x), y\rangle^2 \geqslant \delta^{O(1)} \cdot n^2 \,.$$

Here, the factor n^2 in the conclusion of the theorem normalizes the vectors $\tilde{y}(x)$ and y because $\|\tilde{y}(x)\|^2 \cdot \|y\|^2 = n^2$.

In the remainder of this section, we will prove the above theorem by specializing our meta-algorithm for two-community stochastic block model. For simplicity, we will here only analyze a version of algorithm that runs in quasi-polynomial time. See Section 8.2.5 for how to improve the running time to $n^{1/\mathrm{poly}(\delta)}$.

Algorithm 8.3.3. For a given n-vertex graph $x \in \{0,1\}^{\binom{n}{2}}$ with average degree d and some parameter $\delta > 0$, execute the following steps:[19]

1. evaluate the following matrix-valued polynomial $P(x) = (P_{ij}(x))$

$$P_{ij}(x) \coloneqq \sum_{\alpha \in \mathrm{SAW}_\ell(i,j)} p_\alpha(x) \,. \tag{8.3.1}$$

   Here as in Section 9.2, $\mathrm{SAW}_\ell(i,j) \subseteq \binom{n}{2}^\ell$ consists of all sets of vertex pairs that form a simple (self-avoiding) path between i and j of length $\ell = \Theta(\log n)/\delta^{O(1)}$.[20] The polynomial p_α is a product of centered edge indicators, so that $p_\alpha(x) = \prod_{ab\in\alpha} \left(x_{ab} - \frac{d}{n}\right)$.[21]

2. compute a matrix Y with minimum Frobenius norm satisfying the constraints

$$\left\{ \begin{array}{r} \mathrm{diag}(Y) = \mathbf{1} \\ \frac{1}{\|P(x)\|_F \cdot n} \cdot \langle P(x), Y\rangle \geqslant \delta' \\ Y \succeq 0 \end{array} \right\} . \tag{8.3.2}$$

[19] The right choice of δ' will depend in a simple way on the parameters ε and d.

[20] In particular, the paths in $\mathrm{SAW}_\ell(i,j)$ are not necessarily paths in the graph x but in the complete graph on n vertices.

[21] Up to scaling, this polynomial is a d/n-biased Fourier character of sparse Erdős-Rényi graph.

and output a vector $\tilde{y} \in \{\pm 1\}^n$ obtained by taking coordinate-wise signs of a centered Gaussian vector with covariance Y.[22]

The matrix $P(x)$ is essentially the same as the matrix based on self-avoiding walks analyzed in [136]. The main departure from previous algorithms lies in the second step of our algorithm.

As stated, the first step of the algorithm takes quasi-polynomial because it involves a sum over n^ℓ terms (for $\ell = \Theta(\log n)/\delta^{O(1)}$). In prior works this running time is improved by using non-backtracking paths instead of self-avoiding paths. Non-backtracking paths can be counted in $n^{O(1)}$ time using matrix multiplication, but relating the non-backtracking path polynomial to the self-avoiding path polynomial requires intensive moment-method calculations. An alternative, described in Section 8.2.5, is to compute the self-avoiding path polynomial P using color-coding, requiring time $n^{O(1)+1/\delta^{O(1)}}$, still polynomial time for any constant $\delta > 0$.

The second step of the algorithm is a convex optimization problem over an explicitly represented spectrahedron. Therefore, this step can be carried out in polynomial time.

We break the analysis of the algorithm into two parts corresponding to the following lemmas. The first lemma shows that if $\varepsilon^2 d > 1$ then the matrix $P(x)$ has constant correlation with yy^T for $(x, y) \sim \mathrm{SBM}(n, d, \varepsilon)$ and n sufficiently large. (Notice that this the main preconditon to apply meta-Theorem 8.1.1.)

Lemma 8.3.4 (Low-degree estimator for posterior second moment)**.** *Let $\varepsilon \in \mathbb{R}$ and $d \in \mathbb{N}$, and assume $d = n^{o(1)}$. If $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{\varepsilon^2 d} > 0$ and $n > n_0(\varepsilon, d, \delta)$ is sufficiently*

[22] In other words, we apply the hyperplane rounding algorithm of Goemans and Williamson.

large, then the matrix-valued polynomial $P(x)$ in Eq. (8.3.1) satisfies

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathrm{SBM}(n,d,\varepsilon)} \langle P(x), yy^{\mathsf{T}}\rangle \geqslant \delta^{O(1)} \cdot \left(\mathop{\mathbb{E}}_{x\sim\mathrm{SBM}(n,d,\varepsilon)} \|P(x)\|_F^2\right)^{1/2} \cdot n \tag{8.3.3}$$

(Here, the factor n in the conclusion normalizes the matrix yy^{T} because $\|yy^{\mathsf{T}}\|_F = n$.)

By application of Markov's inequality to the conclusion of this theorem one shows that with P has $\Omega(1)$-correlation with $yy^{\top}$ with $\Omega(1)$-probability. As we have noted several times, the same theorem would hold if we replaced P, an average over self-avoiding walk polynomials, with an average over nonbacktracking walk polynomials. This would have the advantage that the resulting polynomial can be evaluated in $n^{O(1)}$ time (i.e. with running time independent of δ), rather than $n^{O(\log n)/\mathrm{poly}(\delta)}$ for P (which can be improved to $n^{\mathrm{poly}(1/\delta)}$ via color coding), but at the cost of complicating the moment-method analysis. Since we are aiming for the simplest possible proofs here we use P as is.

The second lemma shows that given a matrix P that has constant correlation with yy^{T} for an unknown labeling $y \in \{\pm 1\}^n$, we can efficiently compute a labeling $\tilde{y} \in \{\pm 1\}^n$ that has constant correlation with y. We remark that for this particular situation simpler and faster algorithms work (e.g., choose a random vector in the span of the top $1/\delta^{O(1)}$ eigenvectors of P); these are captured by the meta-Theorem 8.1.1, which we could use in place of the next lemma. (We are presenting this lemma, which involves a more complex and slower algorithm, in order to have a self-contained analysis in this warmup and because it illustrates a simple form of a semidefinite programming technique that is important for our tensor decomposition algorithm, which we use for overlapping communities.)

Lemma 8.3.5 (Partial recovery from posterior moment estimate)**.** *Let $P \in \mathbb{R}^{n\times n}$ be a matrix and $y \in \{\pm 1\}^n$ be a vector with $\delta' := \frac{1}{\|P\|\cdot n}\langle P, yy^{\mathsf{T}}\rangle$. Let Y be the matrix*

of minimum Frobenius such that $Y \succeq 0$, $\mathrm{diag} Y = \mathbf{1}$, *and* $\frac{1}{\|P\| \cdot n} \langle Y, P \rangle \geqslant \delta'$ *(i.e., the constraints Eq. (8.3.2)). Then, the vector* $\tilde{y}$ *obtained by taking coordinate-wise signs of a Gaussian vector with mean* 0 *and covariance* Y *satisfies*

$$\mathbb{E} \langle \tilde{y}, y \rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2 .$$

(Here, the factor n^2 *in the conclusion normalizes the vectors* $\tilde{y}, y$ *because* $\|\tilde{y}\|^2 \cdot \|y\|^2 = n^2$*.)*

Proof. By Theorem 8.2.3, the matrix Y satisfis $\langle Y, yy^\mathsf{T} \rangle \geqslant (\delta'/2)\|Y\| \cdot \|y\|^2$ and $\|Y\| \geqslant \delta \cdot \|y\|^2$. In particular, $\langle Y, yy^\mathsf{T} \rangle \geqslant \delta^2 n^2/2$. The analysis of rounding algorithm for the Grothendieck problem on psd matrices [10], shows that $\mathbb{E} \langle \tilde{y}, y \rangle^2 \geqslant \frac{2}{\pi} \langle Y, yy^\mathsf{T} \rangle \geqslant \Omega(\delta^2) \cdot n^2$. (Here, we use that yy^T is a psd matrix.) □

Taken together, the above lemmas imply a quasi-polynomial time algorithm for partial recovery in $\mathrm{SBM}(n, d, \varepsilon)$ when $\varepsilon^2 d > 1$.

Proof of Theorem 8.3.2 (quasi-polynomial time version). Let $(x, y) \sim \mathrm{SBM}(n, d, \varepsilon)$ with $\delta := 1 - 1/\varepsilon^2 d > 0$. Run Algorithm 8.3.3 on x with the parameter δ' chosen as $\frac{1}{10}$ times the correlation factor in the conclusion of Lemma 8.3.4.

Then, by Lemma 8.3.4, $\mathbb{E}_{(x,y) \sim \mathrm{SBM}(n,d,\varepsilon)} \langle P(x), yy^\mathsf{T} \rangle \geqslant 10\delta' \cdot \mathbb{E}_{x \sim \mathrm{SBM}(n,d,\varepsilon)} \|P(x)\| \cdot n$. By a variant of Markov inequality Theorem 8.8.1, the matrix $P(x)$ satisfies with constant probability $\langle P(x), yy^\mathsf{T} \rangle \geqslant \delta' \cdot \|P(x)\| \cdot n$. In this event, by Lemma 8.3.5, the final labeling $\tilde{y}$ satisfies $\mathbb{E}_{\tilde{y}} \langle \tilde{y}, y \rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2$. Since this event has constant probability, the total expected correlation satisfies $\mathbb{E}_{(x,y) \sim \mathrm{SBM}(n,d,\varepsilon)} \langle \tilde{y}(x), y \rangle^2 \geqslant \Omega(\delta')^2 \cdot n^2$ as desired. □

It remains to prove Lemma 8.3.4.

8.3.1 Low-degree estimate for posterior second moment

We will apply Lemma 8.2.1 to prove Lemma 8.3.4. The next two lemmas verify that the conditions of that lemma hold; they immediately imply Lemma 8.3.4.

Lemma 8.3.6 (Unbiased estimators for $y_i y_j$)**.** *For $i, j \in [n]$ distinct, let $\mathrm{SAW}_\ell(i,j)$ be the set of all simple paths from i to j in the complete graph on n vertices of length ℓ. Let x_{ij} be the ij-th entry of the adjacency matrix of $G \sim \mathrm{SBM}(n,d,\varepsilon)$, and for $\alpha \in \mathrm{SAW}_\ell(i,j)$, let $p_\alpha(x) = \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$. Then for any $y_i, y_j \in \{\pm 1\}$ and $\alpha \in \mathrm{SAW}_\ell(i,j)$,*

$$\left(\frac{n}{\varepsilon d}\right)^\ell \mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = y_i y_j \,.$$

Thus, each simple path α from i to j in the complete graph provides an unbiased estimator $(n/\varepsilon d)^\ell p_\alpha(x)$ of $y_i y_j$. It is straightforward to compute that each has variance $\left(\frac{n}{\varepsilon^2 d}\right)^\ell$. If they were pairwise independent, they could be averaged to give an estimator with variance $\frac{1}{|\mathrm{SAW}_\ell(i,j)|} \cdot \left(\frac{n}{\varepsilon^2 d}\right)^\ell = n(\varepsilon^2 d)^{-\ell}$, since there are $n^{\ell-1}$ simple paths from i to j. If ℓ is logarithmic in n, this becomes small. The estimators are not strictly pairwise independent, but they do satisfy an approximate pairwise independence property which will be enough for us.

Lemma 8.3.7 (Approximate conditional independence)**.** *Suppose $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{\varepsilon^2 d} \geqslant \Omega(1)$ and $d = n^{o(1)}$. For $i, j \in [n]$ distinct, let $\mathrm{SAW}_\ell(i,j)$ be the set of all simple paths from i to j in the complete graph on n vertices of length $\ell = \Theta(\log n)/\delta^C$ for a large-enough constant C. Let x_{ij} be the ij-th entry of the adjacency matrix of $G \sim \mathrm{SBM}(n,d,\varepsilon)$. Let $p_\alpha(x) = \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$. Then*

$$\mathbb{E}\, y_i^2 y_j^2 \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(x) p_\beta(x) \leqslant \delta^{-O(1)} \cdot \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \left(\mathbb{E}\, p_\alpha(x) y_i y_j\right)\left(\mathbb{E}\, p_\beta(x) y_i y_j\right) \,.$$

To prove the lemmas we will use the following fact; the proof is straightforward.

Fact 8.3.8. *For $x, y \sim$ SBM, the entries of x are all independent conditioned on y, and a, b distinct,*

$$\mathbb{E}\left[x_{ab} - \tfrac{d}{n} \mid y_a, y_b\right] = \frac{\varepsilon d}{n} \cdot y_a y_b \quad \text{and} \quad \mathbb{E}\left[\left(x_{ab} - \tfrac{d}{n}\right)^2 \mid y_a, y_b\right] = \frac{d}{n}\left(1 + \varepsilon y_a y_b + O(d/n)\right).$$

We can prove both of the lemmas.

Proof of Lemma 8.3.6. We condition on y and expand the expectation.

$$\mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha} \mathbb{E}[x_{ab} - \tfrac{d}{n} \mid y]\right] = \left(\frac{\varepsilon d}{n}\right)^{\ell} \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha} y_a y_b\right] \quad \text{by Fact 8.3.8.}$$

Because α is a path from i to j, every index $a \in [n]$ except for i and j appears exactly twice in the product. So, removing the conditioning on y_a for all $a \neq i, j$, we obtain $\mathbb{E}\left[p_\alpha(x) \mid y_i y_j\right] = \left(\frac{\varepsilon d}{n}\right)^{\ell} \cdot y_i y_j$ as desired. □

The proof of Lemma 8.3.7 is the heart of the proof, and will use the crucial assumption $\varepsilon^2 d > 1$.

Proof of Lemma 8.3.7. Let $\alpha, \beta \in \mathrm{SAW}_\ell(i, j)$, and suppose that α and β share r edges. Let $\alpha \triangle \beta$ denote the symmetric difference of α and β. Then

$$\begin{aligned}
\mathbb{E}\, p_\alpha(x) p_\beta(x) &= \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha\cap\beta} \mathop{\mathbb{E}}_{x}\left[(x_{ab} - \tfrac{d}{n})^2 \,|\, y_a, y_b\right] \cdot \prod_{ab\in\alpha\triangle\beta} \mathop{\mathbb{E}}_{x}\left[x_{ab} - \tfrac{d}{n} \,|\, y_a, y_b\right]\right] \\
&= \left(\frac{d}{n}\right)^{2\ell - r} \varepsilon^{2\ell - 2r} \mathop{\mathbb{E}}_{y}\left[\prod_{ab\in\alpha\cap\beta} (1 + \varepsilon y_a y_b + O(d/n)) \cdot \prod_{ab\in\alpha\triangle\beta} y_a y_b\right]
\end{aligned}$$

using Fact 8.3.8 in the second step. Since α and β are paths, the graph $\alpha \triangle \beta$ has all even degrees, so $\prod_{ab\in\alpha\triangle\beta} y_a y_b = 1$. Furthermore, any subgraph of $\alpha \cap \beta$ contains some odd-degree vertex. So $\mathbb{E}_y \prod_{ab\in\alpha\cap\beta}(1 + \varepsilon y_a y_b + O(d/n)) = (1 + O(d/n))^r$. All in all, we obtain

$$\mathbb{E}\, p_\alpha(x) p_\beta(x) = \left(\frac{d}{n}\right)^{2\ell - r} \varepsilon^{2\ell - 2r} (1 + O(d/n))^r \tag{8.3.4}$$

Suppose $r < \ell$. Paths α, β sharing r edges must share at least r vertices. If they share exactly r vertices, then the shared vertices must form paths in α and β beginning at i and j. Since each path has length ℓ and therefore contains $\ell - 1$ vertices in addition to i and j, there are at most $r \cdot n^{2(\ell-1)-r}$ such pairs α, β (the multiplicative factor r comes because the shared paths starting from i and j could have lengths between 0 and r). Other pairs α, β share r edges but s vertices for some $s > r$. For each s and r, there are at most $n^{2(\ell-1)-s}\ell^{O(s-r)}$ such pairs, because the shared edges must occur as at most $s - r$ paths. Furthermore, $\ell^{O(s-r)}n^{-(s-r)} \leqslant n^{-\Omega(1)}$ when $s > r$. Putting all of this together,

$$\begin{aligned}\sum_{\alpha,\beta\in \mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(x)p_\beta(x) &\leqslant n^{-2} \cdot \left[\sum_{r=0}^{\ell-1} d^{2\ell-r}\varepsilon^{2\ell-2r}(1+O(d/n))^r \left(r + n^{-\Omega(1)}\right) + (\varepsilon^2 d)^\ell \cdot n\right] \\ &= n^{-2} \cdot (1+n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\sum_{r=0}^{\ell} r \cdot (\varepsilon^2 d)^{-r} + (\varepsilon^2 d)^{-\ell} \cdot n\right),\end{aligned}$$

The additive factor of $(\varepsilon^2 d)^\ell n$ in the first line comes from the case $r = \ell$ (i.e., $\alpha = \beta$), where there are $n^{\ell-1}$ paths. In the second line we have used the assumption that $d \ll n$ to simplify the expression. Finally, by convergence of the series $\sum_{m=0}^{\infty} m \cdot z^m$ for $|z| < 1$, and the choice of ℓ logarithmic in n, this is at most

$$(1+n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\frac{1}{1-\frac{1}{\varepsilon^2 d}}\right)^{O(1)}.$$

So, now our goal is to show that

$$\sum_{\alpha,\beta\in \mathrm{SAW}_\ell(i,j)} (\mathbb{E}\, p_\alpha(x)y_i y_j)(\mathbb{E}\, p_\beta(x)y_i y_j) \geqslant n^{-2} \cdot (1+n^{-\Omega(1)}) \cdot (\varepsilon d)^{2\ell} \cdot \left(\frac{1}{1-\frac{1}{\varepsilon^2 d}}\right)^{O(1)}.$$

Each term in the left-hand sum is $(\varepsilon d/n)^{2\ell}$ (by Lemma 8.3.6) and there are $\Omega(n^{2\ell-2})$ such terms, so the left-hand side of the above is at least $\Omega((\varepsilon d)^{2\ell}/n^2)$. This proves the Lemma. □

8.4 Matrix estimation for generalized block models

In this section we phrase a result essentially due to Abbe and Sandon [3] (and closely related to results by Bordenave et al [42]) in somewhat more general terms. This turns out to be enough to capture an algorithm to estimate a pairwise-vertex-similarity matrix in the $d, k, \alpha, \varepsilon$ mixed-membership block model when $\varepsilon^2 d > k^2(\alpha+1)^2$.

Let $\mathcal{U}$ be a universe of labels, endowed with some base measure ν, such that $\int 1 \cdot d\nu = 1$. Let μ be a probability distribution on $\mathcal{U}$, with a density relative to ν. (We abuse notation by conflating μ and its associated density). Let $W : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_+$ be a bounded nonnegative function with $W(x, y) = W(y, x)$ for every $x, y \in \mathcal{U}$. Consider a random graph model $G(n, d, W, \mu)$ sampled as follows. For each of n vertices, draw a label $x_i \sim \mu$ independently. Then for each pair $ij \in [n]^2$, independently add the edge (i, j) to the graph with probability $\frac{d}{n} W(x_i, x_j)$. (This captures the W-random graph models used in literature on graphons.)

Let $\mathcal{F}$ denote the space of square-integrable functions $f : \mathcal{U} \to \mathbb{R}$, endowed with the inner product $\langle f, g \rangle = \mathbb{E}_{x\sim\mu} f(x)g(x)$. That is, $f \in \mathcal{F}$ if $\mathbb{E}_{x\sim\mu} f(x)^2$ exists.

We assume throughout that

1. (Stochasticity) For every $x \in \mathcal{U}$, the average $\mathbb{E}_{y\sim\mu} W(x, y) = 1$.
2. (Finite rank) W has a finite-rank decomposition $W(x, y) = \sum_{i \leqslant r} \lambda_i f_i(x) f_i(y)$ where $\lambda_i \in \mathbb{R}$ and $f_i : \mathcal{U} \to \mathbb{R}$. The values λ_i are the eigenvalues of W with respect to the inner product generated by μ. The eigenfunctions are orthonormal with respect to the μ inner product. Notice that the

assumptions on W imply that its top eigenfunction $f_1(x)$ is the constant function, with eigenvalue $\lambda_1 = 1$.

3. (Niceness I) Certain rational moments of μ^{-1} exist; that is $\mathbb{E}_{x\sim\mu}\,\mu(x)^{-t}$ exists for $t = -3/2, -2$.

4. (Niceness II) W and μ are nice enough that $W(x,y) \leqslant 1/\sqrt{\mu(x)\mu(y)}$ and $|\overline{W}(x,y)| \leqslant \lambda_2/\sqrt{\mu(x)\mu(y)}$ for every $x, y \in \mathcal{U}$, where $\overline{W}(x,y) = W(x,y) - 1$. (Notice that in the case of discrete W and μ this is always true, and for smooth enough W and μ it is true via a δ-function argument.)

The function W induces a Markov operator $W\colon \mathcal{F} \to \mathcal{F}$. If $f \in \mathcal{F}$, then

$$(Wf)(x) = \mathop{\mathbb{E}}_{y\sim\mu} W(x,y)f(y)\,.$$

(We abuse notation by conflating the function W and the Markov operator W.)

Theorem 8.4.1 (Implicit in [3])**.** *Suppose the operator W has eigenvalues $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_r$ (each possibly with higher multiplicity) and $\delta \stackrel{\text{def}}{=} 1 - \frac{1}{d\lambda_2^2} > 0$. Let Π be the projector to the second eigenspace of the operator W. For types $x_1, \ldots, x_n \sim \mu$, let $A \in \mathbb{R}^{n\times n}$ be the random matrix $A_{ij} = \Pi(x_i, x_j)$, where we abuse notation and think of $\Pi\colon \mathcal{U}\times\mathcal{U} \to \mathbb{R}$. There is an algorithm with running time $n^{\mathrm{poly}(1/\delta)}$ which outputs an $n \times n$ matrix P such that for $x, G \sim G(n, d, W, \mu)$,*

$$\mathop{\mathbb{E}}_{x,G} \operatorname{Tr} P\cdot A \geqslant \delta^{O(1)}\cdot (\mathop{\mathbb{E}}_{x,G}\|A\|^2)^{1/2}(\mathop{\mathbb{E}}_{x,G}\|P\|^2)^{1/2}\,.$$

When $\mathcal{U}$ is discrete with k elements one recovers the usual k-community stochastic block model, and the condition $\lambda_2^2 > 1$ matches the Kesten-Stigum condition in that setting. When $\lambda_2^2 > 1 + \delta$, the guarantees of Abbe and Sandon can be obtained by applying the above theorem to obtain an estimator P for the matrix $M = \sum_{s\in[k]} v_s v_s^\top$, where v_s is the centered indicator vector of community

s. The estimator P will have at least $\delta^{O(1)}/k$ correlation with M, and a random vector in the span of the top $k/\delta^{O(1)}$ eigenvectors of M will have correlation $(\delta/k)^{O(1)}$ with some v_s. Thresholding that vector leads to the guarantees of Abbe and Sandon for the k-community block model, with one difference: Abbe and Sandon's algorithm runs in $O(n \log n)$ time, much faster than the $n^{\mathrm{poly}(1/\delta)}$ running time outlined above. In essence, they achieve this by computing an estimator P' for M which counts only non-backtracking paths in G (the estimator P counts *self-avoiding* paths).

In Section 8.4.1 we prove a corollary of Theorem 8.4.1. This yields the algorithm discussed Theorem 8.1.2 for the mixed-membership blockmodel. As discussed before, the quantitative recovery guarantees of this algorithm are weaker than those of our final algorithm, whose recovery accuracy depends only on the distance δ of the signal-to-noise ratio of the mixed-membership blockmodel to 1. In Section 8.4.2 we prove Theorem 8.4.1.

8.4.1 Matrix estimation for the mixed-membership model

We turn to the mixed-membership model and show that Theorem 8.4.1 yields an algorithm for partial recovery in the mixed-membership block model. However, the correlation of the vectors output by this algorithm with the underlying community memberships depends both on the signal-to-noise ratio and the number k of communiteis. (In particular, when k is super-constant this algorithm no longer solves the partial recovery task.)

Definition 8.4.2 (Mixed-Membership Block Model)**.** Let $G(n, d, \varepsilon, \alpha, k)$ be the following random graph ensemble. For each node $i \in [n]$, sample a probability

vector $\sigma_i \in \mathbb{R}^k_{\geqslant 0}$ with $\sum_{t\in[k]} \sigma_i(t) = 1$ according to the following (simplified) Dirichlet distribution.

$$\mathbb{P}(\sigma) \propto \prod_{t\in[k]} \sigma_i(t)^{\alpha/k-1}$$

For each pair of vertices $i, i' \in [n]$, sample communities $t \sim \sigma_i$ and $t' \sim \sigma_{i'}$. If $t = t'$, add the edge $\{i, i'\}$ to G with probability $\frac{d}{n}(1 + (1 - \frac{1}{k})\varepsilon)$. If $t \neq t'$, add the edge $\{i, i'\}$ to G with probability $\frac{d}{n}(1 - \frac{\varepsilon}{k})$. (For simplicity, throughout this paper we consider only the case that the communities have equal sizes and the connectivity matrix has just two unique entries.)

Theorem 8.4.3 (Constant-degree partial recovery for mixed-membership block model, k-dependent error)**.** *For every $\delta > 0$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$ and for $s \in [k]$ let $v_s \in \mathbb{R}^n$ be given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$.

The algorithm outputs a vector x such that $\mathbb{E}\langle x, v_1\rangle^2 \geqslant \delta'\|x\|^2\|v_1\|^2$, for some $\delta' \geqslant (\delta/k)^{O(1)}$.[23]

Ideally one would prefer an algorithm which outputs $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ with $\mathrm{corr}(\sigma, \tau) \geqslant \delta'/(\alpha+1)$. If one knew that $\langle x, v_1\rangle \geqslant \delta'\|x\|\|v\|$ rather than merely the guarantee on $\langle x, v_1\rangle^2$ (which does not include a guarantee on the sign of x), then this could be accomplished by correlation-preserving projection, Theorem 8.2.3. The tensor methods we use in our final algorithm for the mixed-membership model are able to obtain a guarantee on $\langle x, v_1\rangle$ and hence can output probability vectors $\tau_1, \ldots, \tau_n$.[24]

[23]The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

[24]Such a guarantee could be obtained here by using a cross-validation scheme on x to choose

To prove Theorem 8.4.3 we will apply Theorem 8.4.1 and then a simple spectral rounding algorithm; the next two lemmas capture these two steps.

Lemma 8.4.4 (Mixed-membership block model, matrix estimation)**.** *If $\mathcal{U}$ is the $(k-1)$-simplex, μ is the α, k Dirichlet distribution, and $W(\sigma, \sigma') = 1 - \frac{\varepsilon}{k} + \varepsilon\langle\sigma, \sigma'\rangle$, then $G(n, d, W, \mu)$ is the mixed-membership block model with parameters $k, d, \alpha, \varepsilon$. In this case, the second eigenvalue of W has multiplicity $k-1$ and has value $\lambda_2 = \frac{\varepsilon}{k(\alpha+1)}$.*

Proof. The first part of the claim follows from the definitions. For the second part, note that W has the following decomposition

$$W(\sigma, \tau) = 1 + \sum_{i \leqslant k} \varepsilon(\sigma_i - \tfrac{1}{k})(\tau_i - \tfrac{1}{k})\,.$$

The functions $\sigma \mapsto \sigma_i - \frac{1}{k}$ are all orthogonal to the constant function $\sigma \mapsto 1$ with respect to μ; i.e.

$$\mathop{\mathbb{E}}_{\sigma\sim\mu} 1 \cdot (\sigma_i - \tfrac{1}{k}) = 0$$

because $\mathbb{E}\,\sigma_i = \frac{1}{k}$.

It will be enough to test the above Rayleigh quotient

$$\frac{\mathbb{E}_{\sigma\sim\mu} f(\sigma) \cdot (Wf)(\sigma)}{\mathbb{E}_{\sigma\sim\mu} f(\sigma)^2}$$

with any function $f(\sigma)$ in the span of the functions $\sigma \mapsto \sigma_i - \frac{1}{k}$. If we pick $f(\sigma) = \sigma_1 - \frac{1}{k}$ the remaining calculation is routine, using only the second moments of the Dirichlet distribution (see Fact 8.4.5 below). □

Fact 8.4.5 (Special case of Fact 8.8.3)**.** *Let $\sigma \in \mathbb{R}^k$ be distributed according to the α, k Dirichlet distribution. Let $\tilde{\sigma} = \sigma - \frac{1}{k} \cdot 1$ be centered. Then $\mathbb{E}(\tilde{\sigma})(\tilde{\sigma})^\top = \frac{1}{k(\alpha+1)} \cdot \Pi$ where Π is the projector to the complement of the all-1s vector in $\mathbb{R}^k$.*

between x and $-x$. Since we are focused on what can be accomplished by matrix estimation methods generally we leave this to the reader.

We analyze a simple rounding algorithm.

Lemma 8.4.6. *Let $M = \sum_{i=1}^{k} v_i v_i^\top$ be an $n \times n$ symmetric rank-k PSD matrix. Let $P \in \mathbb{R}^{n \times n}$ be another symmetric matrix such that $\langle P, M \rangle \geqslant \delta \|P\| \|M\|$ (where $\|\cdot\|$ is the Frobenious norm). Then for at least one vector v among $v_1, \ldots, v_k$, a random unit vector x in the span of the top $(k/\delta)^{O(1)}$ eigenvectors of P satisfies*

$$\mathbb{E}\langle x, v \rangle^2 \geqslant (\delta/k)^{O(1)} \|v\|^2 .$$

Now we can prove Theorem 8.4.3.

Proof of Theorem 8.4.3. Lemma 8.4.4 shows that the conditions of Theorem 8.4.1 hold, and hence (via color coding) there is an $n^{\mathrm{poly}(1/\delta)}$ time algorithm to compute a matrix P such that $\langle P, M \rangle \geqslant \delta^{O(1)} \|P\| \|M\|$ with probability at least $\delta^{O(1)}$, where $M = \sum_{s \in [k]} v_s v_s^\top$. (The reader may check that the matrix A of Theorem 8.4.1 is in this case the matrix M described here.)

Applying Lemma 8.4.6 shows that a random unit vector x in the span of the top $(k/\delta)^{O(1)}$ eigenvectors of P satisfies $\langle x, v \rangle^2 \geqslant (\delta/k)^{O(1)} \|v\|^2$, where $v \in \mathbb{R}^n$ has entries $v_i = \sigma_i(1)$. (The choice of 1 is without loss of generality.) □

8.4.2 Proof of Theorem 8.4.1

Definition 8.4.7. For a pair of functions $A, B : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$, we denote by AB their product, whose entries are $(AB)(x, y) = \mathbb{E}_{z \sim \mu} A(x, z) B(z, y)$.

The strategy to prove Theorem 8.4.1 will as usual be to apply Lemma 8.2.1. We check the conditions of that Lemma in the following Lemmas, deferring their proofs till the end of this section.

Lemma 8.4.8. *Let G_{ij} be the $0/1$ indicator for the presence of edge $i \sim j$ in a graph G. As usual, let $\mathrm{SAW}_\ell(i,j)$ be the collection of simple paths of length ℓ in the complete graph on n vertices from i to j.*

Let $x, G \sim G(n,d,W,\mu)$. Let $\alpha \in \mathrm{SAW}_\ell(i,j)$. Let $p_\alpha(G) = \prod_{ab\in\alpha}(G_{ab} - \frac{d}{n})$. Let $\overline{W}(x,y) = W(x,y) - 1$. Then

$$\mathbb{E}\left[p_\alpha(G) \mid x_i, x_j\right] = \left(\frac{d}{n}\right)^\ell \overline{W}^{\ell-1}(x_i,x_j)$$

Lemma 8.4.9. *With the same notation as in Lemma 8.4.8, as long as $\ell \geqslant C\log n/\delta^{O(1)}$ for a large-enough constant C,*

$$\left(\frac{n}{d}\right)^{2\ell} \sum_{\alpha,\beta\in\mathrm{SAW}_\ell(i,j)} \mathbb{E}\, p_\alpha(G)p_\beta(G) \leqslant \delta^{-O(1)} \cdot |\mathrm{SAW}_\ell(i,j)|^2 \cdot \mathbb{E}\,\overline{W}^{\ell-1}(x_i,x_j)^2\,.$$

(The constant C depends on W and the moments of μ.)

Proof of Theorem 8.4.1. Let $B_{ij} = \lambda_2^{-(\ell-1)}\overline{W}^{\ell-1}(x_i,x_j)$. By Lemma 8.4.8, Lemma 8.4.9, and Lemma 8.2.1, there is matrix polynomial $P(G)$, computable to $1/\mathrm{poly}(n)$-accuracy in time $n^{\mathrm{poly}(1/\delta)}$ by color coding, such that

$$\mathbb{E}\,\mathrm{Tr}\,PB^T \geqslant \delta^{O(1)}(\mathbb{E}\,\|P\|^2)^{1/2}(\mathbb{E}\,\|B\|^2)^{1/2}\,.$$

At the same time, $B - A$ has entries

$$(B-A)_{ij} = \sum_{3\leqslant t\leqslant r}\left(\frac{\lambda_t}{\lambda_2}\right)^{\ell-1}\Pi_t(x_i,x_j)$$

where the Π_t projects to the t-th eigenspace of W. Since W is bounded, choosing ℓ a large enough multiple of $\log n$ ensures that $\mathbb{E}\,\|B-A\|^2 \leqslant n^{-100}\,\mathbb{E}\,\|B\|^2$, so the theorem now follows by standard manipulations. □

8.4.3 Proofs of Lemmas

Proof of Lemma 8.4.8. As usual, we simply expand p, obtaining

$$\begin{aligned}\mathbb{E}\left[p_\alpha(G) \mid x_i, x_j\right] &= \mathop{\mathbb{E}}_{x}\left[\prod_{ab\in\alpha}\frac{d}{n}\cdot(W_{x_a,x_b}-1) \mid x_i, x_j\right] \\ &= \left(\frac{d}{n}\right)^{\ell}\cdot\overline{W}^{\ell-1}(x_i,x_j)\,.\end{aligned}$$

□

We will need some small facts to help in proving Lemma 8.4.9.

Fact 8.4.10. *If $\ell - t \geqslant C\log n$ for large enough $C = C(W)$, then*

$$\lambda_2^{2t}\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell-t}(x,y)^2 \leqslant (1+o(1))\cdot\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell}(x,y)^2\,.$$

Also, for any $t \leqslant \ell$,

$$\lambda_2^{2t}\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell-t}(x,y)^2 \leqslant r\cdot\mathop{\mathbb{E}}_{x,y\sim\mu}\overline{W}^{\ell}(x,y)^2\,.$$

where r is the rank of W.

Proof. Using the eigendecomposition of $\overline{W}$, we have that $\mathbb{E}_{x,y\sim\mu}\overline{W}^{\ell-t}(x,y)^2 = \sum_{2\leqslant i\leqslant r}\lambda_i^{2(\ell-t)}$ and similarly $\mathbb{E}_{x,y\sim\mu}\overline{W}^{\ell}(x,y)^2 = \sum_{2\leqslant i\leqslant r}\lambda_i^{2\ell}$. If $i > 2$, then

$$\lambda_2^{2t}\lambda_i^{2(\ell-t)} = \lambda_2^{2\ell}(\lambda_i/\lambda_2)^{2(\ell-t)} \leqslant \lambda_2^{2\ell}/n$$

by our assumption that $\ell - t \geqslant C\log n$ for large enough C. This finishes the proof of the first claim; the second one is similar. □

Proof of Lemma 8.4.9. Pairs α, β which share only the vertices i, j each contribute exactly $\mathbb{E}\,\overline{W}^{\ell-1}(x_i,x_j)^2$ to the left-hand side, by Lemma 8.4.8. Consider next the contribution of α, β whose shared edges form paths originating at i and j. Suppose there are t such shared edges. Then

$$\mathbb{E}\,p_\alpha p_\beta = \left(\frac{d}{n}\right)^{2\ell-t}\mathop{\mathbb{E}}_{x}\prod_{ab\in\alpha\triangle\beta}\overline{W}(x_a,x_b)\cdot\prod_{ab\in\alpha\cap\beta}(W(x_a,x_b)+O(d/n))$$

$$= \left(\frac{d}{n}\right)^{2\ell-t} (1 + O(d/n))^t \, \mathbb{E}\, \overline{W}^{2(\ell-t-1)}(x,y)^2 \,,$$

where for the second equality we used the assumption $\mathbb{E}_{x\sim\mu} W(x,y) = 1$ for every y.

If $\ell - t > C \log n$ for the constant in Fact 8.4.10, then this is at most $(1 + o(1)) \left(\frac{d}{n}\right)^{2\ell-t} \lambda_2^{-2t} \, \mathbb{E}\, \overline{W}^{2(\ell-1)}(x,y)^2$, and for every $t \leqslant \ell$ it is at most $r \cdot \left(\frac{d}{n}\right)^{2\ell-t} \lambda_2^{-2t} \, \mathbb{E}\, \overline{W}^{2(\ell-1)}(x,y)^2$.

There are at most $|\mathrm{SAW}_\ell(i,j)|^2/n^t \cdot t$ choices for such pairs α, β, except when $t = \ell$, in which case there are $|\mathrm{SAW}_\ell(i,j)|^2/n^{t-1}$ choices. So the total contribution from such α, β is at most

$$\begin{aligned}
&|\mathrm{SAW}_\ell(i,j)|^2 \cdot \mathop{\mathbb{E}}_{x,y} \overline{W}^{\ell-1}(x,y)^2 \cdot \left(\sum_{t \leqslant \ell/2} t d^{-t} \lambda_2^{-2t} + nr \cdot \sum_{\ell \geqslant t > \ell/2} t d^{-t} \lambda_2^{-2t} \right) \\
&\leqslant \delta^{-O(1)} |\mathrm{SAW}_\ell(i,j)|^2 \mathop{\mathbb{E}}_{x,y} \overline{W}^{\ell-1}(x,y)^2 \,.
\end{aligned}$$

It remains to handle pairs α, β which share t vertices and s edges for $t > s$. If $t, s \leqslant \ell - 2$, then there are only $n^{2(\ell-1)-s} \ell^{O(t-s)}$ choices for such a pair α, β. The contribution of each such pair we bound as follows

$$\mathbb{E}\, p_\alpha p_\beta = \left(\frac{d}{n}\right)^{2\ell-s} \mathbb{E} \prod_{ab \in \alpha \cap \beta} \mathbb{E}(G_{ab} - \tfrac{d}{n})^2 \cdot \prod_{ab \in \alpha \triangle \beta} \overline{W}_{x_a, x_b} \,.$$

Now, $\mathbb{E}\left[(G_{ab} - \frac{d}{n})^2 \mid x\right] = \frac{d}{n}(W(x_a, x_b) + O(d/n))$ by straightforward calculations, so the above is

$$\begin{aligned}
&(1 + O(d/n))^s \left(\frac{d}{n}\right)^{2\ell-s} \mathop{\mathbb{E}}_{x} \prod_{ab \in \alpha \cap \beta} W(x_a, x_b) \cdot \prod_{ab \in \alpha \triangle \beta} \overline{W}(x_a, x_b) \\
&\leqslant (1 + O(d/n))^s \left(\frac{d}{n}\right)^{2\ell-s} \lambda_2^{2\ell-s} \prod_{a \in \alpha \cup \beta} \mu(x_a)^{-\deg_{\alpha,\beta}(a)/2}
\end{aligned}$$

where $\deg_{\alpha,\beta}(a)$ is the degree of the vertex a in the graph $\alpha \cup \beta$. Any degree-2 vertices simply contribute 1 in the above, since $\mathbb{E}_{x\sim\mu} 1/\mu(x) = 1$. There are at most $t - s$ vertices of higher degree; they may have degree at most 4. They each contribute at most some number $C = C(\mu)$ by the niceness assumptions on μ. So the above is at most

$$(1 + o(1))\left(\frac{d}{n}\right)^{2\ell-s} \lambda_2^{2\ell-s} \exp\{O(t-s)\}\,.$$

Putting things together as in Lemma 8.3.7 finishes the proof. □

Proof of Lemma 8.4.6. By averaging, there is some v among $v_1, \ldots, v_k$ such that

$$\langle P, vv^\top\rangle \geqslant \frac{\delta}{k} \cdot \|P\| \cdot \|M\| \geqslant \frac{\delta}{k} \cdot \|P\| \cdot \|vv^\top\|$$

where the second inequality uses $M \succeq 0$. Renormalizing, we may assume $\|P\|$ has Frobenious norm 1 and v is a unit vector; in this case we obtain $\langle v, Pv\rangle \geqslant \delta/k$. Writing out the eigendecomposition of P, let $P = \sum_{i=1}^n \lambda_i u_i u_i^\top$ and we get

$$\sum_{i=1}^n \lambda_i \langle v, u_i\rangle^2 \geqslant \delta/k$$

By Cauchy-Schwarz,

$$\sum_{i=1}^n \lambda_i \langle v, u_i\rangle^2 \leqslant \left(\sum_{i=1}^n \lambda_i^2 \langle v, u_i\rangle^2\right)^{1/2}$$

and hence $\sum_{i=1}^n \lambda_i^2 \langle v, u_i\rangle^2 \geqslant (\delta/k)^2$, while $\sum_{i=1}^n \lambda_i^2 = 1$. Now the Lemma follows from Markov's inequality. □

8.5 Tensor estimation for mixed-membership block models

8.5.1 Main theorem and algorithm

Theorem 8.5.1 (Constant-degree partial recovery for mixed-membership block model)**.** *There is a constant C such that the following holds. Let $G(n, d, \varepsilon, k, \alpha)$ be the mixed-membership block model. For every $\delta \in (0,1)$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$. Let $t = (\alpha+1)\cdot\frac{k}{k+\alpha}$ (samples from the α, k Dirichlet distribution are approximately uniform over t coordinates). Given G, the algorithm outputs probability vectors $\tau_1, \dots, \tau_n \in \Delta_{k-1}$ such that

$$\mathbb{E}\,\mathrm{corr}(\sigma, \tau) \geqslant \delta^C \left(\frac{1}{t} - \frac{1}{k}\right)\,.$$

(Recall the definition of correlation from (8.1.5).)[25]

Let $c \in (0,1)$ be a small-enough constant. Let $C(c) \in \mathbb{N}$ be a large-enough constant (different from the constant in the theorem statement above). There are three important parameter regimes:

1. Large δ, when $\delta \in [1-c, 1)$.
2. Small δ, when $\delta \in (1-c, 1/k^{1/C})$. This is the main regime of interest. In particular when $k(n) \to \infty$ this contains most values of δ.
3. Tiny δ, when $\delta \in (0, 1/k^{1/C}]$. (This regime only makes sense when $k(n) \leqslant O(1)$.)

[25] The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

Let G_{input} be an n-node graph.

Algorithm 8.5.2 (Main algorithm for mixed-membership model)**.** Let $\eta > 0$ be chosen so that $1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d(1-\eta)} \geqslant \delta^2$ and $o(1) \geqslant \eta \geqslant n^{-\gamma}$ for every constant $\gamma > 0$. (This guarantees that enough edges remain in the input after choosing a holdout set.)

1. Select a partition of $[n]$ into A and $\overline{A}$ at random with $|\overline{A}| = \eta n$. Let $G = A \cap G_{\text{input}}$
2. If δ is large, run Algorithm 8.5.5 on (G_{input}, G, A).
3. If δ is small, run Algorithm 8.5.4 on (G_{input}, G, A).
4. If δ is tiny, run Algorithm 8.5.3 on (G_{input}, G, A).

Algorithm 8.5.3 (Tiny δ)**.**

1. Run the algorithm from Theorem 8.4.3 on G with parameters $(1-\eta)d, k, \varepsilon, \alpha$ to obtain a vector $x \in \mathbb{R}^{n-\eta n}$.
2. Evaluate the quantities $s_x^{(3)} = S_3(G_{\text{input}} \setminus G, x)$ and $s_x^{(4)} = S_4(G_{\text{input}} \setminus G, x)$, the polynomials from Lemma 8.5.8. If $s_x^{(4)} < C(n, \alpha, k, \varepsilon, d, \eta)$, output random labels $\tau_1, \ldots, \tau_n$. (The scalar $C(n, \alpha, k, \varepsilon, d, \eta)$ depends in a simple way on the parameters.)
3. If $s_x^{(3)} < 0$, replace x by $-x$.
4. Run the cleanup algorithm from Lemma 8.5.11 on the vector x, padded with zeros to make a length n vector. Output the resulting $\tau_1, \ldots, \tau_n$.

Algorithm 8.5.4 (Small δ)**.**

1. Using color coding, evaluate the degree-$\log n/\text{poly}(\delta)$ polynomial $P(G) =$

$(P_{ijk}(G))$ from Lemma 8.5.6. (This takes time $n^{\mathrm{poly}(1/\delta)}$.)

2. Run the 3-tensor to 4-tensor lifting algorithm (Theorem 8.7.14) on $P(G)$ to obtain a 4-tensor T.
3. Run the low-correlation tensor decomposition algorithm (Corollary 8.7.3) on T, implementing the cross-validation oracle $\mathcal{O}$ as follows. For each query $x \in \mathbb{R}^{n-\eta n}$, compute $s_x^{(4)} = S_4(G_{\text{input}} \setminus G, x)$, the quantity from Lemma 8.5.9. If $s_x^{(4)} > C(n, d, k, \varepsilon, \alpha, \eta)$ (distinct from the C above, again depending in a simple way on the parameters), output YES, otherwise output NO. The tensor decomposition algorithm returns unit vectors $x_1, \ldots, x_k \in \mathbb{R}^{n-\eta n}$.
4. For each $x_1, \ldots, x_k$, compute $s_i^{(3)} = S_3(G_{\text{input}} \setminus G, x_i)$ and $s_i^{(4)} = S_4(G_{\text{input}} \setminus G, x_i)$. For any x_i for which $s_i^{(4)} < C(n, d, k, \varepsilon, \alpha, \eta)$, replace x_i with a uniformly random unit vector. For any x_i for which $s_i^{(3)} < 0$, replace x_i with $-x_i$.
5. Run the algorithm from Lemma 8.5.10 on $(x_1, \ldots, x_k)$ (padded with zeros to make an $n \times k$ matrix) and output the resulting $\tau_1, \ldots, \tau_n$.

Algorithm 8.5.5 (Large δ)**.**

1. Using color coding, evaluate the degree-$\log n/\mathrm{poly}(\delta)$ polynomial $P(G) = (P_{ijk}(G))$ from Lemma 8.5.6. (This takes time $n^{\mathrm{poly}(1/\delta)}$.)
2. Run the 3-tensor to 4-tensor lifting algorithm (Theorem 8.7.14) on $P(G)$ to obtain a 4-tensor T.
3. Run the low-correlation tensor decomposition algorithm on T, obtaining unit vectors $x_1, \ldots, x_k$.
4. For each x_i, compute the quantity $s_i^{(4)} = S_4(G_{\text{input}} \setminus G, x_i)$ from Lemma 8.5.9. If $s_i^{(4)} < C(n, d, k, \varepsilon, \alpha, \eta)$, replace x_i with a uniformly random unit vector. (The scalar threshold $C(n, d, k, \varepsilon, \alpha, \eta)$ depends in a simple way on the

parameters.)

5. For each x_i, compute the quantity $s_i^{(3)} = S_3(G_{\text{input}} \setminus G, x_i)$ from Lemma 8.5.9. If $s_i^{(3)} < 0$, replace x_i with $-x_i$.
6. Run the algorithm from Lemma 8.5.10 on the matrix $x = (x_1, \ldots, x_k)$ and output the resulting $\tau_1, \ldots, \tau_n$.

We will analyze each of these algorithms separately, but we state the main lemmas together because many are shared among tiny, small, and large δ cases. Two of the algorithms use the low-correlation tensor decomposition algorithm as a black box; Corollary 8.7.3 in Section 8.7 captures the guarantees of that algorithm.

The first thing we need is Theorem 8.4.3, which describes a second-moment based algorithm used as a subroutine by Algorithm 8.5.3. (This subroutine was already analyzed in Section 8.4.)

Theorem (Restatement of Theorem 8.4.3)**.** *For every $\delta > 0$ and $d(n), \varepsilon(n), k(n), \alpha(n)$, there is an algorithm with running time $n^{O(1)+1/\delta^{O(1)}}$ with the following guarantees when*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}.$$

Let $\sigma, G \sim G(n, d, \varepsilon, k, \alpha)$ and for $s \in [k]$ let $v_s \in \mathbb{R}^n$ be given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$.

The algorithm outputs a vector x such that $\mathbb{E}\langle x, v_1\rangle^2 \geqslant \delta' \|x\|^2 \|v_1\|^2$, for some $\delta' \geqslant (\delta/k)^{O(1)}$.[26]

The proofs of all the lemmas that follow can be found later in this section.

[26] The requirement $\varepsilon^2 d \leqslant n^{o(1)}$ is for technical convenience only; as $\varepsilon^2 d$ increases the recovery problem only becomes easier.

Next, we state the tensor estimation lemma used to analyze the tensor P computed in Algorithm 8.5.4 and Algorithm 8.5.5.

Lemma 8.5.6. *Suppose*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad \varepsilon^2 d \leqslant n^{1-\Omega(1)} \text{ and } k, \alpha \leqslant n^{o(1)}.$$

For a collection $\sigma_1, \ldots, \sigma_n$ of probability vectors, let $V(\sigma) = \sum_{s\in[k]} v_s^{\otimes 3}$, where the vectors $v_s \in \mathbb{R}^n$ have entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_s \in \mathbb{R}^n$ have entries $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$. (Note that $\mathbb{E}\langle w_s, w_t\rangle = 0$ for $s \neq t$.) Let $W(\sigma) = \sum_{s\in[k]} w_s^{\otimes 3}$.

If $G \sim G(n, d, \varepsilon, \alpha, k)$, there is a degree $O(\log n/\delta^{O(1)})$ polynomial $P(G) \in (\mathbb{R}^n)^{\otimes 3}$ such that

$$\frac{\mathbb{E}_{\sigma,G}\langle P(G), W(\sigma)\rangle}{\left(\mathbb{E}_{\sigma,G}\|P(G)\|^2\right)^{1/2} \cdot \left(\mathbb{E}_{\sigma,G}\|W(\sigma)\|^2\right)^{1/2}} \geqslant \delta^{O(1)}$$

Furthermore, P can be evaluated up to $(1 + 1/\mathrm{poly}(n))$ multiplicative error (whp) in time $n^{\mathrm{poly}(1/\delta)}$.

Two of our algorithms use the low-correlation tensor decomposition algorithm of Corollary 8.7.3. That corollary describes an algorithm which recovers an underlying orthogonal tensor, but the tensor W is not quite orthogonal. The following lemma, proved via standard matrix concentration, captures the notion that W is close to orthogonal.

Lemma 8.5.7. *Let $\sigma_1, \ldots, \sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $w_s \in \mathbb{R}^n$ be given by $w_s(i) = \sigma_i(s) - \frac{1}{k}(1 - 1/\sqrt{\alpha+1})$. Then as long as $k, \alpha \leqslant n^{o(1)}$, with high probability*

$$(1 + n^{-\Omega(1)}) \cdot \mathrm{Id} \preceq \frac{1}{k}\sum_{s=1}^{k} \frac{w_s w_s^\top}{\mathbb{E}\|w_s\|^2} \preceq (1 + n^{-\Omega(1)}) \cdot \mathrm{Id}\,.$$

All of the algorithms perform some cross-validation using the holdout set $\overline{A}$. The next two lemmas offer what we need to analyze the cross-validations.

Lemma 8.5.8. *Let n_0, n_1 satisfy $n_0 + n_1 = n$. Let $A \subseteq [n]$ have size $|A| = n_1 \geqslant n^{\Omega(1)}$. Let $k = k(n), d = d(n), \varepsilon = \varepsilon(n), \alpha = \alpha(n) > 0$ and $\alpha, k, \varepsilon^2 d \leqslant n^{o(1)}$. Let $\sigma \in \Delta_{k-1}^{n_0}$. Let $v_s \in \mathbb{R}^{n_0}$ have entries $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $\tau_1, \ldots, \tau_{n_1}$ be iid from the α, k Dirichlet distribution.*

Let G be a random bipartite graph on vertex sets $A, [n] \setminus A$, with edges distributed according to the $n, d, \varepsilon, k, \alpha$ mixed-membership model with labels σ, τ. Let $x \in \mathbb{R}^{n_0}$. For $a \in A$, let $P_a(G, x)$ be the expression

$$P_a(G, x) = \sum_{ijk \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n})(G_{ak} - \tfrac{d}{n}) x_i x_j x_k \,.$$

Let $S_3(G, x)$ be

$$S_3(G, x) = \sum_{a \in A} P_a(G, x) \,.$$

There is a number $C = C(n, d, k, \varepsilon, \alpha, n_1)$ such that

$$\mathop{\mathbb{P}}_{G,\tau} \left\{ \left| C \cdot S_3(G, x) - \sum_{s \in [k]} \frac{\langle v_s, x \rangle^3}{\|v_s\|^3} \right| > n^{-\Omega(1)} \right\} \leqslant \exp(-n^{\Omega(1)}) \,.$$

Similarly, there are scalars $C(n, d, k, \varepsilon, \alpha, n_1), C'(n, d, k, \varepsilon, \alpha, n_1)$ such that the following holds. For $a \in A$, let

$$Q_a(G, x) = \sum_{ijk\ell \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n})(G_{ak} - \tfrac{d}{n})(G_{a\ell} - \tfrac{d}{n}) x_i x_j x_k x_\ell \,.$$

and let

$$R_a(G, x) = \sum_{ij \in \overline{A} \text{ distinct}} (G_{ai} - \tfrac{d}{n})(G_{aj} - \tfrac{d}{n}) x_i x_j \,.$$

Finally let

$$S_4(G, x) = C \cdot \sum_{a \in A} Q_a(G, x) - C' \cdot \left(\sum_{a \in A} R_a(G, x) \right)^2 \,.$$

Then

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|S_4(G,x) - \sum_{s\in[k]} \frac{\langle v_s, x\rangle^4}{\|v_s\|^4}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)})\,.$$

Lemma 8.5.9. *Under the same hypotheses as Lemma 8.5.8, there are $S_3(G,x), S_4(G,x)$, polynomials of degree 3 and 4, respectively, in x and in the edge indicators of G, such that*

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|S_4(G,x) - \sum_{s\in[k]} \frac{\langle w_s, x\rangle^4}{\|w_s\|^4}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)})\,,$$

and

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|C\cdot S_3(G,x) - \sum_{s\in[k]} \frac{\langle w_s, x\rangle^3}{\|w_s\|^3}\right| > n^{-\Omega(1)}\right\} \leqslant \exp(-n^{\Omega(1)})\,,$$

where $w_1,\dots,w_k$ are the vectors $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$.

Finally, all of the algorithms have a cleanup phase to transform n-length vectors to probability vectors $\tau_1,\dots,\tau_n \in \Delta_{k-1}$. The following lemma describes the guarantees of the cleanup algorithm used by the small and large δ algorithms, which takes as input vectors x correlated with the vectors w.

Lemma 8.5.10. *Let $\delta \in (0,1)$ and $k = k(n) \in \mathbb{N}$ and $\alpha = \alpha(n) \geqslant 0$, with $\alpha, k \leqslant n^{o(1)}$. Suppose $\delta \geqslant 1/k^{1/C}$ for a big-enough constant C. There is a $\mathrm{poly}(n)$-time algorithm with the following guarantees.*

Let $\sigma_1,\dots,\sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $v_1,\dots,v_k \in \mathbb{R}^n$ be the vectors given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_1,\dots,w_k \in \mathbb{R}^n$ be the vectors given by $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$, so that $\mathbb{E}\langle w_s, w_t\rangle = 0$ for $s \neq t$. Let $M = \sum_s w_s w_s^\top$. Let E be the event that

1. $\left\|M^{-1/2} w_s - \frac{w_s}{(\mathbb{E}\,\|w_s\|^2)^{1/2}}\right\| \leqslant \frac{1}{\mathrm{poly}\, n}$ *for every $s \in [k]$.*
2. $\|w_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|w_s\|^2)^{1/2}$ *for every $s \in [k]$.*

3. $\|v_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|v_s\|^2)^{1/2}$ *for every* $s \in [k]$.

Suppose $x_1, \ldots, x_k \in \mathbb{R}^n$ *are unit vectors such that for at least* δk *vectors* $w_1, \ldots, w_m$ *there exists* $t \in [k]$ *such that* $\langle w_s, x_t \rangle \geqslant \delta \|w_s\|$.

The algorithm takes input $x_1, \ldots, x_k$ *and when* E *happens returns probability vectors* $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ *such that*

$$\mathrm{corr}(\sigma, \tau) \geqslant \delta^{O(1)}\,\mathbb{E}\,\|v\|^2 = \delta^{O(1)} \left(\frac{1}{\alpha + 1} \cdot \frac{k + \alpha}{k} - \frac{1}{k} \right).$$

Finally, the last lemma captures the cleanup algorithm used by the tiny-δ algorithm, which takes a single vector x correlated with v_1.

Lemma 8.5.11. *Let* $\delta \in (0, 1)$ *and* $k = k(n) \in \mathbb{N}$ *and* $\alpha = \alpha(n) \geqslant 0$, *with* $\alpha, k \leqslant n^{o(1)}$. *Suppose* $\delta \leqslant k^{1/C}$ *for any constant* C. *There is a* $\mathrm{poly}(n)$*-time algorithm with the following guarantees.*

Let $\sigma_1, \ldots, \sigma_n$ *be iid draws from the* α, k *Dirichlet distribution. Let* $v_1, \ldots, v_k \in \mathbb{R}^n$ *be the vectors given by* $v_s(i) = \sigma_i(s) - \frac{1}{k}$. *Let* $x \in \mathbb{R}^n$ *be a unit vector satisfying* $\langle x, v_s \rangle \geqslant \delta \|v_s\|$ *for some* $s \in [k]$. *On input* x, *the algorithm produces* $\tau_1, \ldots, \tau_n \in \Delta_{k-1}$ *such that*

$$\mathrm{corr}(\sigma, \tau) \geqslant \left(\frac{\delta}{k} \right)^{O(1)} \cdot \mathbb{E}\,\|v\|^2 = \delta^{O(1)} \left(\frac{1}{\alpha + 1} \cdot \frac{k + \alpha}{k} - \frac{1}{k} \right).$$

so long as the event E *from Lemma 8.5.10 occurs.*

Analysis for tiny δ (Algorithm 8.5.3)

Proof of Theorem 8.5.1, tiny-δ case. Let $C \in \mathbb{N}$ and $1 \geqslant \delta > 0$ be any fixed constants. We will prove that if $k \leqslant \delta^C$ then the output of Algorithm 8.5.2 satisfies the

conclusion of Theorem 8.5.1. Let $x \in \mathbb{R}^{(1-\eta)n}$ be the output of the matrix estimation algorithm of Theorem 8.4.3. By Markov's inequality, with probability $(\delta/k)^{O(1)}$ over G and $\sigma_1, \ldots, \sigma_{(1-\eta)n}$, the vector x satisfies $\langle v, x\rangle^2 \geqslant (\delta/k)^{O(1)}\|v\|^2\|x\|^2$, where $v \in \mathbb{R}^{(1-\eta)n}$ is the vector $v(i) = \sigma_i(1) - \frac{1}{k}$. By our assumption $k \leqslant \delta^C$, this means that with probability $\delta^{O(1)}$ the vector x satisfies $\langle x, v\rangle^2 \geqslant \delta^{O(1)}\|x\|^2\|v\|^2$.

Now, the labels $\sigma_{(1-\eta)n}, \ldots, \sigma_n$ and the edges from nodes $1, \ldots, (1-\eta)n$ to nodes $(1-\eta)n, \ldots, n$ are independent of everything above. So, invoking Lemma 8.5.8, we can assume that the quantity $s_x^{(4)}$ computed by Algorithm 8.5.3 satisfies

$$\left| s_x^{(4)} - \sum_{s\in[k]} \frac{\langle v_s, x\rangle^4}{\|v_s\|^4} \right| \leqslant n^{-\Omega(1)}.$$

Now, if x satisfies $\langle v_s, x\rangle^2 \geqslant \delta^{O(1)}\|v_s\|^2$ for some v_s, then also $s_x^{(4)} \geqslant \delta^{O(1)}$. On the other hand, if $s_x^{(4)} \geqslant \delta^{O(1)}$ then there is some s such that $\langle x, v_s\rangle^2 \geqslant \frac{\delta^{O(1)}}{k}\|v_s\|^2$. So choosing the threshold C in Algorithm 8.5.3 appropriately, we have obtained that with probability $\delta^{O(1)}$ the algorithm reaches step 3 with a vector x which satisfies $\langle x, v_s\rangle^2 \geqslant \delta^{O(1)}\|v_s\|^2$, and otherwise the algorithm outputs random $\tau_1, \ldots, \tau_n$.

Step 3 is designed to check the sign of $\langle x, v_s\rangle$. Call x good if there is $s \in [k]$ such that $\langle x, v_s\rangle \geqslant \delta^{O(1)}\|v_s\|$. If $|s_x^{(3)}| \leqslant \delta^{O(1)}$ then x there are v_s, v_t such that $\langle v_s, x\rangle \geqslant \delta^{O(1)}\|v_s\|$ and $\langle v_t, x\rangle \leqslant -\delta^{O(1)}\|v_t\|$, so both x and $-x$ are good If $|s_x^{(3)}| > \delta^{O(1)}$ then clearly step 3 outputs a good vector. Since after step 3 the vector x is good, applying Lemma 8.5.11 finishes the proof in the tiny δ case. □

Analysis for small and large δ (Algorithm 8.5.4, Algorithm 8.5.5)

Proof of Theorem 8.5.1, small δ case. Let $n_0 = (1-\eta)n$ and $n_1 = \eta n$ with η as in Algorithm 8.5.2.

By Markov's inequality applied to Lemma 8.5.6, with probability $\delta^{O(1)}$ the tensor P satisfies $\langle P, W\rangle \geqslant \delta^{O(1)}\|P\|\|W\|$, where $W \in (\mathbb{R}^{n_0})^{\otimes 3}$ is as in Lemma 8.5.6. Let $M = \sum_{s\in[k]} w_s w_s^\top$, where w_s is as in Lemma 8.5.6. The vectors $M^{-1/2}w_s$ are orthonormal, and Lemma 8.5.7 guarantees that $\|\frac{w_s}{\|w_s\|} - M^{-1/2}w_s\| \leqslant n^{-\Omega(1)}$ with high probability. Let $W' = \sum_{s\in[k]}(M^{-1/2}w_s)^{\otimes 3}$ and let $W'_4 = \sum_{s\in[k]}(M^{-1/2}w_s)^{\otimes 4}$. Then also $\langle P, W'\rangle \geqslant \delta^{O(1)}\|P\|\|W'\|$. By the guarantees of the 3-to-4 lifting algorithm (Theorem 8.7.14), finally we get $\langle T, W'_4\rangle \geqslant \delta^{O(1)}\|T\|\|W'_4\|$.

In order to conclude that Algorithm 8.5.4 successfully runs the low-correlation tensor decomposition algorithm, we have to check correctness of its implementation of the cross-validation oracle. This follows from Lemma 8.5.7, Lemma 8.5.9, the size of η, and a union bound over the $\exp(k/\mathrm{poly}(\delta)) \leqslant \exp(n^{o(1)})$ queries made by the nonadaptive implementation of the low-correlation tensor decomposition algorithm, and independence of the randomness in the holdout set.

We conclude that with probability at least $\delta^{O(1)}$, the tensor decomposition algorithm returns unit vectors $x_1, \ldots, x_k \in \mathbb{R}^{n_0}$ such that a $\delta^{O(1)}$ fraction of w_s among $w_1, \ldots, w_k$ have x_t such that $\langle w_s, x_t\rangle^2 \geqslant \delta^{O(1)}\|w_s\|^2$. By the same reasoning as in the tiny δ case, using Lemma 8.5.9 after the sign-checking step the same guarantee holds with the strengthened conclusion $\langle w_s, x_t\rangle \geqslant \delta^{O(1)}\|w_s\|$. Finally, we apply Lemma 8.5.10 (along with elementary concentration arguments to show that the event E occurs with high probability) to conclude that the last step of Algorithm 8.5.4 gives $\tau_1, \ldots, \tau_n$ such that (in expectation) $\mathrm{corr}(\sigma, \tau) \geqslant \delta^{O(1)}\left(\frac{1}{\alpha+1} \cdot \frac{k}{k+\alpha} - \frac{1}{k}\right)$ as desired. □

8.5.2 Low-degree estimate for posterior third moment

In this section we prove Lemma 8.5.6. The strategy is to apply Lemma 8.2.1 to find an estimator for the 3-tensor $\sum_{s\in[k]} v_s^{\otimes 3}$. With that in hand, combining with the estimators in Section 8.4 for the second moments $\sum_{s\in[k]} v_s v_s^\top$ is enough to obtain an estimator for W, since

$$\sum_{s\in[k]} w_s^{\otimes 3} = \sum_{s\in[k]} (v_s + c\cdot 1)^{\otimes 3} \tag{8.5.1}$$

$$= \sum_{s\in[k]} v_s^{\otimes 3} + c(v_s\otimes v_s\otimes 1 + v_s\otimes 1\otimes v_s + 1\otimes v_s\otimes v_s) + 1^{\otimes 3} \tag{8.5.2}$$

where 1 is the all-1s vector, $c = \frac{1}{k\sqrt{\alpha+1}}$, and we have used that $\sum_{s\in[k]} v_s = 0$. Thus if R is a degree $n^{\mathrm{poly}(1/\delta)}$ polynomial such that

$$\langle R, \sum_{s\in[k]} v_s^{\otimes 3}\rangle \geqslant \delta^{O(1)} (\mathbb{E}\,\|R\|^2)^{1/2} (\mathbb{E}\left\|\sum_{s\in[k]} v_s^{\otimes 3}\right\|^2)^{1/2}$$

and Q is similar but estimates $\sum_{s\in[k]} v_s v_s^\top$, then R and Q can be combined according to (8.5.2) to obtain the polynomial P from the lemma statement.

Thus in the remainder of this section we focus on obtaining such a polynomial R; we change notation to call this polynomial P. The first step will be to define a collection of polynomials $\{G^\alpha\}_\alpha$ for all distinct $i, j, k \in [n]$.

Definition 8.5.12. Any $\alpha \subseteq \binom{n}{2}$ can be interpreted as a graph on some nodes in $[n]$. Such an α is a long-armed star if it consists of three self-avoiding paths, each with ℓ edges, joined at one end at a single central vertex, at the other end terminating at distinct nodes $i, j, k \in [n]$. (See figure.) Let $\mathrm{STAR}_\ell(i, j, k)$ be the set of 3-armed stars with arms of length ℓ and terminal vertices i, j, k. For any $\alpha \subseteq \binom{n}{2}$ let $G^\alpha = \prod_{ab\in\alpha}(x_{ab} - \frac{d}{n})$ be the product of centered edge indicators.

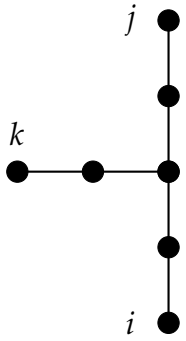


Figure 8.1: A
3-armed star with arms of length 2. We will eventually use arms of length $t \approx \log n$.

The next two lemmas check the conditions to apply Lemma 8.2.1 to the sets $\{G^\alpha\}_{\alpha \in \mathrm{STAR}_\ell(i,j,k)}$.

Lemma 8.5.13 (Unbiased Estimator)**.** *Let $i, j, k \in [n]$ all be distinct. Let $\alpha \in \mathrm{STAR}_\ell(i,j,k)$.*

For a collection of probability vectors $\sigma_1, \ldots, \sigma_k$, let $V(\sigma) = \sum_{s \in [k]} v_s^{\otimes 3}$ where $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $G \sim G(n, d, \varepsilon, \alpha_0, k)$.

$$\mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \left(\frac{1}{k(\alpha_0+1)}\right)^{3(\ell-1)} \cdot C_3 \cdot V(\sigma)_{ijk}\,.$$

Here $\alpha_0 \geqslant 0$ is the Dirichlet concentration paramter, unrelated to the graph α, and $C_3 = 1/(k^{O(1)}\alpha_0^{O(1)})$ is a constant related to third moments of the Dirichlet distribution.

Lemma 8.5.14 (Approximate conditional independence)**.** *If*

$$\delta \stackrel{\text{def}}{=} 1 - \frac{k^2(\alpha_0+1)^2}{\varepsilon^2 d} > 0 \quad \text{and} \quad k, \alpha_0 \leqslant n^{o(1)} \text{ and } \varepsilon^2 d \leqslant n^{o(1)}\,.$$

and $\ell \geqslant C \log n/\delta^{O(1)}$ for a large enough constant C, then for $G \sim G(n, d, \varepsilon, k, \alpha_0)$,

$$\mathbb{E}\left[V(\sigma)_{ijk}^2\right] \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\, G^\alpha G^\beta \leqslant 1/\delta^{O(1)} \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\left[G^\alpha V(\sigma)_{i,j,k}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{i,j,k}\right]\,.$$

Now we can prove Lemma 8.5.6.

Proof of Lemma 8.5.6. As discussed at the beginning of this section, it is enough to find an estimator for the tensor $V(\sigma)$. Lemma 8.5.13 and Lemma 8.5.14 show that Lemma 8.2.1 applies to each set of polynomials $\mathrm{STAR}_\ell(i,j,k)$. The conclusion is that for every distinct $i,j,k \in [n]$ there is a degree $\log n \operatorname{poly}(1/\delta)$ polynomial $P(G)_{ijk}$ so that

$$\frac{\mathbb{E}\, P(G)_{ijk} V(\sigma)_{ijk}}{(\mathbb{E}\, P(G)_{ijk}^2)^{1/2} \cdot (\mathbb{E}\, V(\sigma)_{ijk}^2)^{1/2}} \geqslant \Omega(1)\,.$$

One may check that the entries i,j,k for i,j,k all distinct of the tensor $V(\sigma)$ comprise nearly all of its 2-norm. That is,

$$\sum_{i,j,k \text{ distinct}} \mathbb{E}\, V(\sigma)_{i,j,k}^2 \geqslant (1 - o(1))\, \mathbb{E}\, \|V(\sigma)\|^2\,.$$

This is sufficient to conclude that the tensor-valued polynomial $P(G)$ whose (i,j,k)-th entry is $P_{i,j,k}(G)$ when i,j,k are all distinct and is 0 otherwise is a good estimator of $V(\sigma)$ (see Fact 8.8.2). Thus,

$$\frac{\mathbb{E}_{\sigma,G} \langle P(G), V(\sigma)\rangle}{\left(\mathbb{E}_{\sigma,G} \|P(G)\|^2\right)^{1/2} \cdot \left(\mathbb{E}_{\sigma,G} \|V(\sigma)\|^2\right)^{1/2}} \geqslant \Omega(1)\,. \qquad \square$$

Details of unbiased estimator

We work towards proving Lemma 8.5.13. We will need to assemble a few facts. The first will help us control moment tensors of the Dirichlet distribution. The proof can be found in the appendix.

Fact 8.5.15 (Special case of Fact 8.8.3)**.** *Let σ be distributed according to the α, k Dirichlet distribution. Let $\tilde{\sigma} = \sigma - \frac{1}{k}1$. There are numbers C_2, C_3 depending on α, k so that for every x_1, x_2, x_3 in $\mathbb{R}^k$ with $\sum_{s\in[k]} x_i(s) = 0$,*

$$\mathop{\mathbb{E}}_{\sigma} \langle \tilde{\sigma}, x_1\rangle \langle \tilde{\sigma}, x_2\rangle = C_2 \langle x_1, x_2\rangle$$

and

$$\mathop{\mathbb{E}}_{\sigma}\langle\tilde{\sigma}, x_1\rangle\langle\tilde{\sigma}, x_2\rangle\langle\tilde{\sigma}, x_3\rangle = C_3 \sum_{s\in[k]} x_1(s)x_2(s)x_3(s)\,.$$

Furthermore,

$$C_2 = \frac{1}{k(\alpha+1)} \quad \textit{and} \quad C_3 = \frac{1}{k^{O(1)}\alpha^{O(1)}}\,.$$

Now we can prove Lemma 8.5.13.

Proof of Lemma 8.5.13. For any collection of σ's and $\alpha \in \mathrm{STAR}_\ell(i,j,k)$,

$$\mathop{\mathbb{E}}_{G}[G^\alpha \mid \sigma] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \prod_{(a,b)\in\alpha} \langle\tilde{\sigma}_a, \tilde{\sigma}_b\rangle$$

Let a be the central vertex of the star α. Taking expectations over all the vertices in the arms of the star,

$$\mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell} \left(\frac{1}{k(\alpha_0+1)}\right)^{3(\ell-1)} \mathop{\mathbb{E}}_{\sigma_a}\langle\tilde{\sigma}_i, \tilde{\sigma}_a\rangle\langle\tilde{\sigma}_j, \tilde{\sigma}_a\rangle\langle\tilde{\sigma}_k, \tilde{\sigma}_a\rangle\,.$$

Finally, using the second part of Fact 8.5.15 completes the proof. □

Details of approximate conditional independence

We prove Lemma 8.5.14, first gathering some facts. In the sum $\sum_{\alpha,\beta\in\mathrm{STAR}_\ell(i,j,k)} G^\alpha G^\beta$, the terms α, β which (as graphs) share only the vertices i, j, k will not cause us any trouble, because such G^α and G^β are independent conditioned on $\sigma_i, \sigma_j, \sigma_k$.

Fact 8.5.16. *If $\alpha, \beta \in \mathrm{STAR}_\ell(i,j,k)$ share only the vertices i, j, k, then for any collection σ of probability vectors,*

$$\mathbb{E}\left[G^\alpha G^\beta \mid \sigma_i, \sigma_j, \sigma_k\right] = \mathbb{E}\left[G^\alpha \mid \sigma_i, \sigma_j, \sigma_k\right] \cdot \mathbb{E}\left[G^\beta \mid \sigma_i, \sigma_j, \sigma_k\right]\,.$$

Proof. To sample G^α, one needs to know σ_a for any $a \in [n]$ with nonzero degree in α, and similar for $b \in [n]$ and G^β. The only overlap is $\sigma_i, \sigma_j, \sigma_k$. □

The next fact is the key one. Pairs α, β which share vertices forming paths originating at i, j, and k make the next-largest contribution (after α, β sharing only i, j, k) to $\sum_{\alpha,\beta} \mathbb{E}\, G^\alpha G^\beta$.

Fact 8.5.17. *Let $i, j, k \in [n]$ be distinct. Let $V(\sigma)_{ijk}$ be as in the Lemma 8.5.14. Let $C_2 \in \mathbb{R}$ be as in Fact 8.5.15.*

Let $\alpha, \beta \in \mathrm{STAR}_\ell(i, j, k)$ share s vertices (in addition to i, j, k) for some $s \leqslant \frac{t}{2}$, and suppose the shared vertices form paths in α and β starting at i, j, and k. Then

$$\mathbb{E}\, V(\sigma)_{ijk}^2 \cdot \mathbb{E}\, G^\alpha G^\beta \leqslant \varepsilon^{-2s} \left(\frac{d}{n}\right)^{-s} (1+O(d/n))^{-s} \cdot \left(\frac{1}{k(\alpha_0+1)}\right)^{-2s} \cdot \mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right] .$$

Proof. Let $\sigma_{\alpha\cap\beta}$ be the σ's corresponding to vertices sharerd by α, β. Let i', j', k' be the last shared vertices along the paths beginning at i, j, k respectively. We expand $G^\alpha G^\beta$ and use conditional independence of the G_e's given the σ's:

$$\mathbb{E}\, G^\alpha G^\beta = \mathop{\mathbb{E}}_{\sigma_{i',j',k'}} \left[\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \,|\sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] \cdot \mathbb{E}\left[G^{\alpha\setminus\beta} \mid \sigma_{i'}\sigma_{j'}\sigma_{k'}\right] \cdot \mathbb{E}\left[G^{\beta\setminus\alpha} \mid \sigma_{i'}\sigma_{j'}\sigma_{k'}\right]\right] .$$

Both $G^{\alpha\setminus\beta}$ and $G^{\beta\setminus\alpha}$ are long-armed stars with terminal vertices i', j', k'. The arm lengths of $G^{\alpha\setminus\beta}$ total $3\ell - s$. By a similar argument to Lemma 8.5.13, $G^{\alpha\setminus\beta}$ is an unbiased estimator of $V(\sigma)_{i'j'k'}$ with

$$\mathbb{E}\left[G^{\alpha\setminus\beta} \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = \left(\frac{\varepsilon d}{n}\right)^{3\ell-s} \left(\frac{1}{k(\alpha_0+1)}\right)^{3(\ell-1)-s} \cdot C_3 \cdot V(\sigma)_{i',j',k'}$$

and the same goes for $G^{\beta\setminus\alpha}$. Furthermore,

$$\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = \left(\frac{d}{n}\right)^{|\alpha\cap\beta|} \mathbb{E}\left[\prod_{(a,b)\in\alpha\cap\beta} (1 + \varepsilon\langle\tilde{\sigma}_a, \tilde{\sigma}_b\rangle + O(d/n)) \,\middle|\, \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] .$$

By our assumption that $\alpha \cap \beta$ consists just of paths, every subset of edges in the graph $\alpha \cap \beta$ contains a vertex of degree 1. Hence, $\mathbb{E}\left[(G^{\alpha\cap\beta})^2 \mid \sigma_{i'}, \sigma_{j'}, \sigma_{k'}\right] = (1 + O(d/n))^{|\alpha\cap\beta|}(d/n)^{|\alpha\cap\beta|}$. Putting these together,

$$\mathbb{E}\, G^\alpha G^\beta = (1 + O(d/n))^s \varepsilon^{6\ell-2s} \left(\frac{d}{n}\right)^{6\ell-s} \left(\frac{1}{k(\alpha_0+1)}\right)^{6(\ell-1)-2s} C_3^2\, \mathbb{E}\, V(\sigma)_{ijk}^2$$

At the same time, one may apply Lemma 8.5.13 to $\mathbb{E}\, G^\alpha V(\sigma)_{ijk}$ to obtain

$$\mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right] = \left(\frac{\varepsilon d}{n}\right)^{6\ell} \left(\frac{1}{k(\alpha_0+1)}\right)^{6(\ell-1)} C_3^2 \cdot \left(\mathop{\mathbb{E}}_{\sigma_i,\sigma_j,\sigma_k} V(\sigma)_{ijk}^2\right)^2 .$$

The lemma follows. □

The last fact will allow us to control α, β which intersect in some way other than paths starting at i, j, k. The key idea will be that such pairs α, β must share more vertices than they do edges.

Fact 8.5.18. *Let $i, j, k \in [n]$ be distinct. Let $V(\sigma)_{ijk}$ be as in the Lemma 8.5.14. Let $C_2 \in \mathbb{R}$ be as in Fact 8.5.15. $C_2 = \frac{1}{k(\alpha_0+1)}$.*

Let $\alpha, \beta \in \mathrm{STAR}_\ell(i, j, k)$ share s vertices (in addition to i, j, k) and r edges. Then

$$\mathbb{E}\, V(\sigma)_{ijk}^2 \cdot \mathbb{E}\, G^\alpha G^\beta \leqslant \varepsilon^{-2r} \left(\frac{d}{n}\right)^{-r} \cdot C_2^{-2s} \cdot k^{O(s-r)} (1+\alpha_0)^{O(s-r)} \cdot \mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right] .$$

Proof. Expanding as usual,

$$\mathbb{E}\, G^\alpha G^\beta = \left(\frac{d}{n}\right)^{6\ell-r} \mathop{\mathbb{E}}_{\sigma} \prod_{ab\in\alpha\triangle\beta} \langle \tilde{\sigma}_a, \tilde{\sigma}_b\rangle \cdot \prod_{ab\in\alpha\cap\beta} (1 + \varepsilon\langle \tilde{\sigma}_a, \tilde{\sigma}_b\rangle + O(d/n)) .$$

Any nontrivial edge-induced subgraph of $\alpha \cap \beta$ contains a degree-1 vertex; using this to expand the second product and simplifying with $\mathbb{E}\, \tilde{\sigma}_a = 0$, the above is

$$\left(\frac{d}{n}\right)^{6\ell-r} \mathop{\mathbb{E}}_{\sigma} \prod_{ab\in\alpha\triangle\beta} \langle \tilde{\sigma}_a, \tilde{\sigma}_b\rangle \cdot (1 + O(d/n))^r .$$

For every degree-2 vertex in $\alpha \triangle \beta$ we can use Fact 8.8.3 to take the expectation. Each such vertex contributes a factor of C_2 and there are at least $3\ell - O(s-r)$ such vertices. The remaining expression will be bounded by 1. The fact follows. □

Now we can prove Lemma 8.5.14.

Proof of Lemma 8.5.14. Let us recall that our goal is to show

$$\mathbb{E}\left[V(\sigma)_{ijk}^2\right] \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\, G^\alpha G^\beta \leqslant \delta^{O(1)} \cdot \sum_{\alpha,\beta \in \mathrm{STAR}_\ell(i,j,k)} \mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right]$$

where $\delta = 1 - \frac{k^2(\alpha_0+1)^2}{\varepsilon^2 d}$. Let $c = \mathbb{E}\left[G^\alpha V(\sigma)_{ijk}\right] \cdot \mathbb{E}\left[G^\beta V(\sigma)_{ijk}\right]$. (Notice this number does not depend on α or β.) The right-hand side above simplifies to $|\mathrm{STAR}_\ell(i,j,k)|^2 \cdot c$.

On the left-hand side, what is the contribution from α, β sharing s vertices? First consider what happens with $s \leqslant t/2$ and the intersecting vertices form paths in α and β starting at i, j, k. Choosing a random pair α, β from $\mathrm{STAR}_\ell(i,j,k)$, the probability that they intersect along paths of length s_1, s_2, s_3 starting at i, j, k respectively is at most $n^{-s_1-s_2-s_3}$. There are at most $(1+s^2)$ choices for nonnegative integers s_1, s_2, s_3 with $s_1 + s_2 + s_3 = s$. By Fact 8.5.17, such terms therefore contribute at most

$$c \cdot \frac{|\mathrm{STAR}_\ell(i,j,k)|^2}{n^{-s}} \cdot \left(\varepsilon\sqrt{\tfrac{d}{n}}(1+O(d/n))\right)^{-2s} C_2^{-2s} \cdot s^2 = c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot (\varepsilon^2 d C_2^2 (1+O(d/n)))^{-s} \cdot s^2$$

where $C_2 = \frac{1}{k(\alpha_0+1)}$. By hypothesis, $\delta > 0$. Consider the sum of all such contributions for $s \leqslant t/2$; this is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot \sum_{s=0}^{t/2} (1+s^2) \cdot \left(\frac{k^2(\alpha_0+1)^2}{\varepsilon^2 d}\right)^s \leqslant \delta^{O(1)} \cdot c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 .$$

Next, consider the contribution from α, β which share s vertices in some pattern other than those considered above. Unless $\alpha = \beta$, this means α, β share at

least one more vertex than the number r of edges that they share. Suppose $\alpha \neq \beta$ and let $s - r = q$. There are $t^{O(q)}$ patterns in which such an intersection might occur, and each occurs for a random pair $\alpha, \beta \in \mathrm{STAR}_\ell(i,j,k)$ with probabilty n^{-s}. So using Fact 8.5.18, the contribution is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \cdot \sum_{q=1}^{t} \left(\frac{\varepsilon^2 d}{n}\right)^q \cdot k^{O(q)}(1+\alpha_0)^{O(q)} t^{O(q)}$$

By the hypotheses $k, \alpha = n^{o(1)}$ and $\varepsilon^2 d = n^{1-\Omega(1)}$, this is all $o(c|\mathrm{STAR}_\ell(i,j,k)|^2)$.

Finally, consider the case $\alpha = \beta$. Then, using Fact 8.5.18 again, the contribution is at most

$$c \cdot |\mathrm{STAR}_\ell(i,j,k)|^2 \left(\frac{\varepsilon^2 d}{k^2(\alpha_0+1)^2}\right)^{-t} k^{O(1)} \alpha^{O(1)}$$

which is $o(c|\mathrm{STAR}_\ell(i,j,k)|^2)$ because $t \gg \log(n)$. Putting these things together gives the lemma. □

8.5.3 Cross validation

In this section we show how to use a holdout set of vertices to cross-validate candidate community membership vectors. The arguments are all standard, using straightforward concentration inequalities. At the end we prove the first part of Lemma 8.5.8, on the estimator S_3. The proof of the second part, on S_4 is similar, using standard facts about moments of the Dirichlet distribution (see Fact 8.8.3). The proof of Lemma 8.5.9 is also similar, using the discussion in Section 8.5.2 to turn estimators for moments of the v vectors into estimators for moments of the w vectors—we leave it to the reader.

We will need a few facts to prove the lemma.

Fact 8.5.19. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 8.5.8. Let $a \in A$. There is a number $C = C(k, \alpha) \leqslant \mathrm{poly}(k, \alpha)$ such that*

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G, x) = \left(\frac{\varepsilon d}{n}\right)^3 \cdot C \cdot \sum_{ijk \in \overline{A} \text{ distinct}} \sum_{s \in [k]} \sigma_i(s)\sigma_j(s)\sigma_k(s) x_i x_j x_k \,.$$

Proof. Immediate from Fact 8.5.15. □

Fact 8.5.20. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 8.5.8. Let $a \in A$. The following variance bound holds.*

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G, x)^2 - \left(\mathop{\mathbb{E}}_{G,\tau} P_a(G, x)\right)^2 \leqslant \frac{\mathrm{poly}(k, \alpha, \varepsilon, d)}{n^3} \,.$$

Proof. Expanding $P_a(G, x)$ and using that $|\langle \sigma, \sigma' \rangle| \leqslant 1$ for any $\sigma, \sigma' \in \Delta_{k-1}$ we get

$$\mathop{\mathbb{E}}_{G,\tau} P_a(G, x)^2 \leqslant \left(\frac{d}{n}\right)^6 \sum_{\substack{ijk \text{ distinct} \\ i'j'k' \text{ distinct}}} \left| x_i x_j x_k x_{i'} x_{j'} x_{k'} \right| \leqslant \left(\frac{d}{n}\right)^6 \cdot n^3 \cdot \|x\|^{12} \,.$$

□

Fact 8.5.21. *Let $n_0, n_1, A, k, d, \varepsilon, \alpha, \sigma, v, \tau, G, x, P$ be as in Lemma 8.5.8. Let $a \in A$. For some constant $\gamma_*(\varepsilon, d, k, \alpha)$ and every $\gamma_* > \gamma > 0$,*

$$\mathop{\mathbb{P}}_{G,\tau} \{|P_a(G, x)| > n^\gamma\} \leqslant \exp(-n^{\Omega(\gamma)})$$

Proof. The fact follows from a standard exponential tail bound on the degree of vertex a. □

We can put these facts together to prove the S_3 portion of Lemma 8.5.8 (as we discussed above, the S_4 portion and Lemma 8.5.9 are similar). The strategy will be to use the following version of Bernstein's inequality, applied to the random variables $\langle G_a, v^{\otimes 3} \rangle$. The proof of the inequality is in the appendix.

Proposition 8.5.22 (Bernstein wth tails)**.** *Let X be a random variable satisfying $\mathbb{E}\, X = 0$ and, for some numbers $R, \delta, \delta' \in \mathbb{R}$,*

$$\mathbb{P}\{|X| > R\} \leqslant \delta \text{ and } \mathbb{E}\,|X| \cdot \mathbf{1}_{|X|>R} \leqslant \delta' .$$

Let $X_1, \ldots, X_m$ be independent realizations of X. Then

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i\leqslant m} X_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1)\cdot m \cdot t^2}{\mathbb{E}\, X^2 + t \cdot R}\right) + m\delta .$$

Now we can prove Lemma 8.5.8.

Proof of Lemma 8.5.8. We apply Proposition 8.5.22 to the n_1 random variables $X_a = \left(\frac{\varepsilon d}{n}\right)^{-3} C^{-1} P_a(G, x)$ for $a \in A$, where $C = C(k, \alpha)$ is the number from Fact 8.5.20. (For each $a \in A$ these are iid over G, τ.) Take $t = n^{3/2-\gamma'}$ for a small-enough constant γ' so that $n_1 t^2/n^3 \geqslant n^{\gamma}$ for some constant γ, using the assumption $n_1 \geqslant n^{\Omega(1)}$. All together, we get

$$\mathop{\mathbb{P}}_{G,\tau}\left\{\left|\frac{1}{n_1}\sum_{a\in A} X_a - \sum_{s\in[k]}\sum_{ijk\in\overline{A}\text{ distinct}} \sigma_s(i)\sigma_s(j)\sigma_s(k)x_i x_j x_k\right| \geqslant n^{3/2-\gamma'}\right\} \leqslant \exp(n^{-\gamma'})$$

for some constants γ, γ' (possibly different from γ, γ' above) and large-enough n. For any unit $x \in \mathbb{R}^{n_0}$ and $\sigma \in \Delta_{k-1}^{n_0}$, using that $k \leqslant n^{o(1)}$ it is not hard to show via Cauchy-Schwarz that

$$\left|\sum_{s\in[k]} \langle v_s, x\rangle^3 - \sum_{s\in[k]}\sum_{ijk\in\overline{A}\text{ distinct}} \sigma_s(i)\sigma_s(j)\sigma_s(k)x_i x_j x_k\right| \leqslant n^{1+o(1)} .$$

The lemma follows. □

8.5.4 Producing probability vectors

In this section we prove Lemma 8.5.10. The proof of Lemma 8.5.11 is very similar (in fact it is somewhat easier) so we leave it to the reader.

Lemma (Restatement of Theorem 8.5.10)**.** *Let $\delta \in (0,1)$ and $k = k(n) \in \mathbb{N}$ and $\alpha = \alpha(n) \geqslant 0$, with $\alpha, k \leqslant n^{o(1)}$. Suppose $\delta \geqslant 1/k^{1/C}$ for a big-enough constant C. There is a* poly(n)*-time algorithm with the following guarantees.*

Let $\sigma_1, \dots, \sigma_n$ be iid draws from the α, k Dirichlet distribution. Let $v_1, \dots, v_k \in \mathbb{R}^n$ be the vectors given by $v_s(i) = \sigma_i(s) - \frac{1}{k}$. Let $w_1, \dots, w_k \in \mathbb{R}^n$ be the vectors given by $w_s(i) = v_s(i) + \frac{1}{k\sqrt{\alpha+1}}$, so that $\mathbb{E}\langle w_s, w_t\rangle = 0$ for $s \neq t$. Let $M = \sum_s w_s w_s^\top$. Let E be the event that

1. $\left\| M^{-1/2} w_s - \frac{w_s}{(\mathbb{E}\,\|w_s\|^2)^{1/2}} \right\| \leqslant \frac{1}{\operatorname{poly} n}$ *for every $s \in [k]$.*
2. $\|w_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|w_s\|^2)^{1/2}$ *for every $s \in [k]$.*
3. $\|v_s\| = (1 \pm 1/\mathrm{poly}(n))(\mathbb{E}\,\|v_s\|^2)^{1/2}$ *for every $s \in [k]$.*

Suppose $x_1, \dots, x_k \in \mathbb{R}^n$ are unit vectors such that for at least δk vectors $w_1, \dots, w_m$ there exists $t \in [k]$ such that $\langle w_s, x_t\rangle \geqslant \delta \|w_s\|$.

The algorithm takes input $x_1, \dots, x_k$ and when E happens returns probability vectors $\tau_1, \dots, \tau_n \in \Delta_{k-1}$ such that

$$\operatorname{corr}(\sigma, \tau) \geqslant \delta^{O(1)}\,\mathbb{E}\,\|v\|^2 = \delta^{O(1)} \left(\frac{1}{\alpha+1} \cdot \frac{k+\alpha}{k} - \frac{1}{k} \right).$$

First some preliminaries. Let $\sigma_1, \dots, \sigma_n$ be iid from the α, k Dirichlet distribution. There are two important families of vectors in $\mathbb{R}^n$. Let

$$v_s(i) = \sigma_i(s) - \frac{1}{k} \qquad w_s(i) = \sigma_i(s) - \frac{1}{k}\left(1 - \frac{1}{\sqrt{\alpha+1}}\right).$$

We will also work with a normalized version of the v vectors:

$$\overline{v}_s = \frac{v_s}{(\mathbb{E}\,\|v_s\|^2)^{1/2}}.$$

By construction, $\mathbb{E}\,\|\overline{v}_s\|^2 = 1$. Also by definition, $\sum_s v_s = \sum_s \overline{v}_s = 0$. Thus $\mathbb{E}\langle \sum_s \overline{v}_s, \sum_s \overline{v}_s\rangle = k + \sum_{s\neq t} \mathbb{E}\langle \overline{v}_s, \overline{v}_t\rangle = 0$ and so by symmetry $\mathbb{E}\langle \overline{v}_s, \overline{v}_t\rangle = \frac{-1}{k-1}$. We let

$$\overline{w}_s = \overline{v}_s + \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{k-1}}$$

so that $\mathbb{E}\langle \overline{w}_s, \overline{w}_t\rangle = 0$ for $s \neq t$. (In the facts which follow we sometimes write $\overline{v}$ as v when both normalizations are not needed; this is always noted.)

We will want the following fact; the proof is elementary.

Fact 8.5.23. *Let σ, u, v, w as above, and suppose y is an $n \times k$ matrix whose rows are in $\Delta_{k-1} - \frac{1}{k}$ (that is they are shifted probability vectors). Then $\tau = y + \frac{1}{k}$ is a matrix whose rows are probability vectors, and τ satisfies*

$$\langle \tau, \sigma\rangle \geqslant \langle y, v\rangle + \frac{n}{k}\,.$$

The following fact will be useful when δ is small but not tiny; i.e. $\delta < 1 - c$ for some fixed constant c but $\delta \gg 1/\sqrt{k}$.

Fact 8.5.24. *Suppose that $x_1, \ldots, x_k$ are unit vectors and $w_1, \ldots, w_k$ are orthonormal. Also suppose that there is $1 > \delta > 0$ such that for at least δk vectors w_s among $w_1, \ldots, w_k$ there exists a vector x_t among $x_1, \ldots, x_k$ such that $\langle w_s, x_t\rangle \geqslant \delta$. Then there is a permutation $\pi : [k] \to [k]$ such that if $x = (x_1, \ldots, x_k)$ is an $n \times k$ matrix and similarly for w,*

$$\langle x, \pi \cdot w\rangle \geqslant \left(\delta^5 - \frac{1}{\sqrt{k}}\left(\frac{1}{1-\delta^4}\right)^{1/2}\right)\|x\|\|w\|\,,$$

where $x = (x_1, \ldots, x_k)$ is an $n \times k$ matrix and similarly for w.

Proof. We will think of π as a matching of $w_1, \ldots, w_k$ to $x_1, \ldots, x_k$. Call x_t *good* for w_s if $\langle w_s, x_t\rangle \geqslant \delta$. First of all, by orthogonality of vectors $w_1, \ldots, w_k$, any

particular vector x_t is good for at most $1/\delta^2$ vectors w_s. Hence, there is a set S of $\delta^4 k$ vectors w_s such that for each w_s there exists a good x_t and all the good x_t's are distinct.

Begin by matching each $w_s \in S$ to its good x_t. Let π be the result of extending that matching randomly to a perfect matching of k to k.

We need to lower bound $\mathbb{E}\sum_{s\notin S}\langle w_s, x_{\pi(s)}\rangle$. Consider that for a particular t,

$$\mathbb{E} -\langle x_t, w_{\pi^{-1}(t)}\rangle \leqslant (\mathbb{E}\langle x_t, w_{\pi^{-1}(t)}\rangle^2)^{1/2}.$$

The distribution of $\pi^{-1}(t)$ is uniform among all $s \notin S$. So

$$\mathbb{E}\langle x_t, w_{\pi^{-1}(t)}\rangle^2 = \frac{1}{k-|S|}\sum_{s\notin S}\langle w_s, x_t\rangle^2 \leqslant \frac{1}{k}\left(\frac{1}{1-\delta^4}\right)$$

since $\sum_{s\in[k]}\langle w_s, x_t\rangle^2 \leqslant 1$. It follows that

$$\mathbb{E}\langle x_t, w_{\pi^{-1}(t)}\rangle \geqslant -\frac{1}{\sqrt{k}}\left(\frac{1}{1-\delta^4}\right)^{1/2}.$$

Therefore, $\mathbb{E}\sum_{s\notin S}\langle w_s, x_{\pi(s)}\rangle \geqslant -\sqrt{k}\left(\frac{1}{1-\delta^4}\right)^{1/2}$. Thus there is some choice of π such that $\sum_{s\notin S}\langle w_s, x_{\pi(s)}\rangle \geqslant -\sqrt{k}\left(\frac{1}{1-\delta^4}\right)^{1/2}$. Hence for this π one gets

$$\sum_{s\in[k]}\langle w_s, x_{\pi(s)}\rangle \geqslant \delta^5 k - \sqrt{k}\left(\frac{1}{1-\delta^4}\right)^{1/2} = \left(\delta^5 - \frac{1}{\sqrt{k}}\left(\frac{1}{1-\delta^4}\right)^{1/2}\right)\|x\|\|w\|. \qquad \square$$

The next fact serves the same purpose as the previous one but in the large δ case (i.e. δ close to 1).

Fact 8.5.25. *Under the same hypotheses as Fact 8.5.24, letting $\delta = 1-\varepsilon$ for some $\varepsilon > 0$, there is a permutation $\pi : [k] \to [k]$ such that $\langle x, \pi\cdot w\rangle \geqslant (1-9\varepsilon)\|x\|\|w\|$.*

Proof. As in the proof of Fact 8.5.24, we construct a matching π by first matching a set S of at least $\delta^4 k \geqslant (1-4\varepsilon)k$ vectors w_s to corresponding x_t. Then we match

the remaining vectors arbitrarily. For any s, t we know $\langle w_s, x_t \rangle \geqslant -1$. So the result is

$$\langle x, \pi \cdot w \rangle \geqslant (1 - 5\varepsilon)k - 4\varepsilon k = (1 - 9\varepsilon)k = (1 - 9\varepsilon)\|x\|\|w\| \,. \qquad \square$$

We will also want a way to translate a matrix correlated with w to one correlated with v, so that we can apply Fact 8.5.23.

Fact 8.5.26. *Suppose v is an $n \times k$ matrix whose rows are centered probability vectors and $w = v + c$ is a coordinate-wise additive shift of v. Suppose y is also an $n \times k$ matrix whose rows are centered probability vectors shifted by c in each coordinate (so $y - c$ is a matrix of centered probability vectors). Then the shifted matrix $y - c$ satisfies*

$$\langle y - c, v \rangle \geqslant \langle y, w \rangle - c^2 nk \,.$$

Proof. By definition, $\langle y - c, v \rangle = \langle y, v \rangle$. Since $v = w - c$, we get

$$\langle y - c, v \rangle = \langle y, v \rangle = \langle y, w \rangle - c\langle y, 1 \rangle = \langle y, w \rangle - c^2 nk \,. \qquad \square$$

Proof of Lemma 8.5.10. First assume $\delta < 1 - c$ for any small constant c. Let π be the permutation guaranteed by Fact 8.5.24 applied to the vectors $x_1, \ldots, x_k$ and $M^{-1/2}w_1, \ldots, M^{-1/2}w_k$. (Without loss of generality reorder the vectors so that π is the identity permutation.) Since $1 - c \geqslant \delta \geqslant 1/k^{1/C}$ for big-enough C and small-enough c (which are independent of n, k) and the guarantee of Fact 8.5.24, by event E we get that

$$\langle x, w \rangle \geqslant \delta^{O(1)}\|x\|\|w\| \,.$$

So by taking a correlation-preserving projection of x into the set of matrices whose rows are shifted probability vectors, we get a matrix y with the guarantee

$$\langle y, w \rangle \geqslant \delta^{O(1)}\|y\|\|w\| \quad \text{and} \quad \|y\| \geqslant \delta^{O(1)}\|w\| \,.$$

Applying Fact 8.5.26, we obtain

$$\langle y - c, v\rangle \geqslant \langle y, w\rangle - c^2 nk = \langle y, w\rangle - \frac{\mathbb{E}\,\|w\|^2}{k}$$

where $c = \frac{1}{k\sqrt{\alpha+1}}$. Putting things together and using $\mathbb{E}\,\|v\|^2 \leqslant \mathbb{E}\,\|w\|^2$ and the event E, we get

$$\langle y - c, v\rangle \geqslant \delta^{O(1)}\,\mathbb{E}\,\|v\|^2\,.$$

So applying Fact 8.5.23 finishes the proof in this case.

Now suppose $\delta \geqslant 1 - c$ for a small-enough constant c. Then using event E and Fact 8.5.25, there is π such that $\langle x, w\rangle \geqslant (1 - O(c))\|x\|(\mathbb{E}\,\|w\|^2)$ (where again we have without loss of generality reordered the vectors so that π is the identity permutation). Now taking the Euclidean projection of $x \cdot \frac{(\mathbb{E}\,\|w\|^2)^{1/2}}{\|x\|}$ into the $n \times k$ matrices whose rows are centered probability vectors shifted entrywise by $c = \frac{1}{k\sqrt{\alpha+1}}$, we get a matrix y which again satisfies $\langle y, w\rangle \geqslant (1 - O(c))\|y\|\|w\|$ and $\|y\| \geqslant (1 - O(c))\|w\|$, so (using event E), $\langle y, w\rangle \geqslant (1 - O(c))\,\mathbb{E}\,\|w\|^2$. Removing the contribution from $\langle y, 1\rangle$, this implies that $\langle y - c, v\rangle \geqslant (1 - O(c))\,\mathbb{E}\,\|v\|^2$. For c small enough, this is at least $\delta^{O(1)}\,\mathbb{E}\,\|v\|^2$. Applying Fact 8.5.23 finishes the proof. □

8.5.5 Remaining lemmas

We provide sketches of the proofs of Lemma 8.5.7 and Lemma 8.5.10, since the proofs of these lemmas use only standard techniques.

Proof sketch of Lemma 8.5.7. For $\sigma \in \mathbb{R}^k$, let $\tilde{\sigma} = \sigma - (1 - 1/\sqrt{\alpha+1})/k$. Standard calculations show that if σ is drawn from the α, k Dirichlet distribution then $\mathbb{E}\,\tilde{\sigma}\tilde{\sigma}^\top = \frac{1}{k(\alpha+1)}\mathrm{Id}$. It follows by standard matrix concentration and the assumption

$k, \alpha \leqslant n^{o(1)}$ that the eigenvalues of $\frac{1}{n}\sum_{i\leqslant n}\tilde{\sigma}_i\tilde{\sigma}_i^\top$ are all $1 \pm n^{-\Omega(1)}$, where $\sigma_1, \ldots, \sigma_n$ are iid draws from the α, k Dirichlet distribution.

For the second part of the Lemma, use the first part to show that $\left\|\frac{v_s}{\|v_s\|} - w_s'\right\| \leqslant 1/\mathrm{poly}(k)$. Then when $k \geqslant \delta^{-C}$ for large-enough C, if $\langle x, v_s\rangle^3 \geqslant \delta^{O(1)}\|v_s\|^3$ it follows that also $\langle x, w_s\rangle \geqslant \delta^{O(1)} - 1/\mathrm{poly}(k) \geqslant \delta^{O(1)}$. The lemma follows. □

Proof sketch of Lemma 8.5.10. If $\delta < 1 - \Omega(1)$, then $\delta^2/2 \geqslant \delta^{O(1)}$, so the Lemma follows from standard concentration and Theorem 8.2.3 on correlation-preserving projection. On the other hand, if $\delta \geqslant 1 - o(1)$, then $\|v' - \tilde{\sigma}\| \leqslant o(1) \cdot \|\tilde{\sigma}\|$, so the same is also true for the projection of v' into $(\tilde{\Delta}_{k-1})^n$ by convexity and the lemma follows. □

8.6 Lower bounds against low-degree polynomials at the Kesten-Stigum threshold

In this section we prove two lower bounds for k-community partial recovery algorithms based on low-degree polynomials.

8.6.1 Low-degree Fourier spectrum of the k-community block model

Theorem 8.6.1. *Let d, ε, k be constants. Let $\mu : \{0,1\}^{n\times n} \to \mathbb{R}$ be the relative density of $SBM(n, d, \varepsilon, k)$ with respect to $G(n, \frac{d}{n})$. Let $\mu^{\leqslant \ell}$ be the projection of μ to the degree-ℓ*

polynomials with respect to the norm induced by $G(n, \frac{d}{n})$.[27] *For any constant* $\delta > 0$,

$$\|\mu^{\leqslant \ell}\| \text{ is } \begin{cases} \geqslant n^{\Omega(1)} \text{ if } \varepsilon^2 d > (1+\delta)k^2, \ell \geqslant O(\log n) \\ \leqslant O_{\delta,k,\varepsilon,d}(1) \text{ if } \varepsilon^2 d < (1-\delta)k^2, \ell < n^{0.01} \end{cases}.$$

This proves Theorem 8.1.9 (see discussion following statement of that theorem). To prove the theorem we need the following lemmas.

Lemma 8.6.2. *Let* $\chi_\alpha : \{0,1\}^{n \times n} \to \mathbb{R}$ *be the* $\frac{d}{n}$*-biased Fourier character. If* $\alpha \subseteq \binom{n}{2}$*, considered as a graph on* n *vertices, has any degree-one vertex, then*

$$\mathop{\mathbb{E}}_{G \sim SBM(n,d,\varepsilon,k)} \chi_\alpha(G) = 0$$

The proof follows from calculations very similar to those in Section 8.5, so we omit it.

Proof of Theorem 8.6.1. The bound $\|\mu^{\leqslant \ell}\| \geqslant n^{\Omega(1)}$ when $\varepsilon^2 d > (1+\delta)k^2$ and $\ell \gg \log(n)$, follows from almost identical calculations to Section 8.5,[28] so we omit this argument and focus on the regime $\varepsilon^2 d < (1-\delta)k^2$.

By definition and elementary Fourier analysis,

$$\|\mu^{\leqslant \ell}\|^2 = \sum_{\alpha \subseteq \binom{n}{2}, |\alpha| \leqslant \ell} \widehat{\mu}(\alpha)^2 \tag{8.6.1}$$

Also by definition,

$$\widehat{\mu}(\alpha) = \mathop{\mathbb{E}}_{G \sim G(n, \frac{d}{n})} \mu(G)\chi_\alpha(G) = \mathop{\mathbb{E}}_{G \sim SBM(n,d,\varepsilon,k)} \chi_\alpha$$

[27] That is, $\|f\| = (\mathbb{E}_{G \sim G(n, \frac{d}{n})} f(G)^2)^{1/2}$.

[28] The calculations in Section 8.5 are performed for long-armed stars; to prove the present result the analogous calculations should be performed for cycles of logarithmic lengh. Similar calculations also appear in many previous works.

where $\{\chi_\alpha\}$ are the $\frac{d}{n}$-biased Fourier characters. Thus, using Lemma 8.6.2 we may restrict attion to the contribution of those $\alpha \subseteq \binom{n}{2}$ with $|\alpha| \leqslant \ell$ and containing no degree-1 vertices.

Fix such an α, and suppose it has $C(\alpha)$ connected components and $V_2(\alpha)$ vertices of degree 2 (considered again as a graph on $[n]$). Fact 8.6.3 (following this proof) together with routine computations shows that

$$\begin{aligned}\left(\mathop{\mathbb{E}}_{G\sim SBM(n,d,\varepsilon,k)} \chi_\alpha(G)\right)^2 &\leqslant \left((1+O(\tfrac{d}{n}))\varepsilon^2\tfrac{d}{n}\right)^{|\alpha|} k^{-2(V_2(\alpha)-C(\alpha))} \\ &\leqslant \left(1+O(\tfrac{d}{n})\right)^{|\alpha|}\cdot n^{-|\alpha|}\cdot(1-\delta)^{|\alpha|}\cdot k^{2(|\alpha|-V_2(\alpha)+C(\alpha))}.\end{aligned}$$

Let $c(\alpha) = \left(1+O(\frac{d}{n})\right)^{|\alpha|}\cdot n^{-|\alpha|}\cdot(1-\delta)^{|\alpha|}\cdot k^{2(|\alpha|-V_2(\alpha)+C(\alpha))}$ be this upper bound on the contribution of α to the right-hand side of (8.6.1). It will be enough to bound

$$(*) \quad \stackrel{\text{def}}{=} \sum_{\substack{\alpha\subseteq\binom{n}{2}\\ |\alpha|\leqslant\ell\\ \alpha \text{ has no degree 1 nodes}}} c(\alpha)$$

Given any α as in the sum, we may partition it into two vertex-disjoint subgraphs, α_0 and α_1, where α_0 is a union of cycles and no connected component of α_1 is a cycle, such that $\alpha = \alpha_0 \cup \alpha_1$. Thus,

$$(*) \leqslant \left(\sum_{\alpha_0} c(\alpha_0)\right)\left(\sum_{\alpha_1} c(\alpha_1)\right)$$

where α_0 ranges over unions of cycles with $|\alpha_0| \leqslant \ell$ and α_1 ranges over graphs on $[n]$ with at most ℓ where all degrees are at least 2 and containing no connected component which is a cycle. Lemmas 8.6.4 and 8.6.5, which follow, the terms above as $O(1)$, which finishes the proof. □

Fact 8.6.3. *Let U be a connected graph where all vertices have degree at least 2, and let t be the number of degree 2 vertices in U. For each vertex v of U let $\sigma_v \in \mathbb{R}^k$ be a*

uniformly random standard basis vector. Let $\tilde{\sigma}_v = \sigma_v - \frac{1}{k} \cdot 1$. *Then*

$$\left| \mathbb{E} \prod_{(u,v)\in U} \langle \tilde{\sigma}_v, \tilde{\sigma}_u \rangle \right| \leqslant k^{-t+1}$$

Proof. The covariance $\mathbb{E}\, \tilde{\sigma}\tilde{\sigma}^\top = \frac{1}{k}\Pi \in \mathbb{R}^{k\times k}$, where Π is the projector to the orthogonal complement of the all-1's vector. Consider marginalizing out the degree-2 vertices v one by on. Until reaching the last degree-2 vertex, each marginalization gives a factor of $1/k$:

$$\mathbb{E} \prod_{(u,v)\in U} \langle \tilde{\sigma}_v, \tilde{\sigma}_u \rangle = k^{-t}\, \mathbb{E} \prod_{(u,v)\in U'} \langle \tilde{\sigma}_v, \tilde{\sigma}_u \rangle$$

where U' is the graph obtained from U by iteratively replacing every degree 2 vertex but for one with an edge connecting its neighbors. (The last degree-2 vertex may lead to a self-loop.) Since $|\langle \tilde{\sigma}_u, \tilde{\sigma}_v \rangle| \leqslant 1$, we are done. □

Lemma 8.6.4. *For* $\alpha \subseteq \binom{n}{2}$, *let* $V(\alpha)$ *be the number of vertices in* α, *let* $C(\alpha)$ *be the number of connected components in* α. *For constants* ε, d, k, *let* $c(\alpha) \stackrel{\text{def}}{=} \left(1 + O(\frac{d}{n})\right)^{|\alpha|} \cdot n^{-|\alpha|} \cdot (1-\delta)^{|\alpha|} \cdot k^{2(|\alpha| - V_2(\alpha) + C(\alpha))}$ *Let* $\ell \leqslant n^{0.01}$ *and*

$$U = \left\{ \alpha \subseteq \binom{n}{2} \,:\, \alpha \text{ has all degrees} \geqslant 2, \text{ has no connected components which are cycles}, |\alpha| \leqslant \ell \right\}.$$

Then

$$\sum_{\alpha \in U} c(\alpha) \leqslant O(1)\,.$$

Proof. We will use a coding argument to bound the number of $\alpha \in U$ with V vertices, E edges, and C connected components. We claim that any such α is uniquely specified by the following encoding.

To encode α, start by picking an arbitrary vertex v_1 in α. List the vertices $v_1, \ldots, v_{|V|}$ of α, each requiring $\log n$ bits, starting from v_1, using the following rules to pick v_i.

1. If v_{i-1} has a neighbor not yet appearing in the list $v_1, \ldots, v_{i-1}$, let v_i be any such neighbor.
2. Otherwise, if v_{i-1} has a neighbor v_j which
   (a) appears in the list $v_1, \ldots, v_{i-1}$ and
   (b) for which either $j = 1$ or v_{j-1} is not adjacent to v_j in α, and
   (c) for which if $j \neq i'$ for $i' \leqslant i - 1$ being the minimal index such that $v_{i'}, \ldots, v_{i-1}$ is a path in α (i.e. $v_j, \ldots, v_{i-1}$ are not a cycle in α)

   then reorder the list as follows. Remove vertices $v_j, \ldots, v_{j'}$ where j' is the greatest index so that all edges $v_\ell, v_{\ell+1}$ exist in α for $j \leqslant \ell \leqslant j'$. Also remove vertices $v_{i'}, \ldots, v_{i-1}$ where i' is analogously the minimal index such that edes $v_\ell, v_{\ell+1}$ exist in α for $i' \leqslant \ell \leqslant i - 1$. Then, append the list $v_{j'}, v_{j'-1}, \ldots, v_j, v_{i-1}, \ldots, v_{i'}$. By construction, all of these vertices appear in a path in α. The new list retains the invariant that every vertex either preceeds a neighbor in α or has no neighbors in α which have not previous appeared in the list.
3. Otherwise, let v_i be an arbitrary vertex in α in the same connected component as v_{i-1}, if some such vertices has not yet appeared in the list.
4. Otherwise, let v_i be an arbitrary vertex of α not yet appearing among $v_1, \ldots, v_{i-1}$.

After the list of vertices, append to the encoding the following information. First, a list of the R (for **r**emoved) pairs v_i, v_{i+1} for which there is not an edge (v_i, v_{i+1}) in α. This uses $2R \log V$ bits. Last, a list of the edges in α which are not among the pairs v_i, v_{i+1} (each edge encoded using $2 \log V$ bits).

We argue that the number R of removed pairs (and hence the length of their list in the encoding) is not too great. In particular, we claim $R \leqslant 2(E - V)$. In fact, this is true connected-component-wise in α. To see it, proceed as follows.

Fix a connected component β of α. Let v_t be the first vertex in β to appear in the list $v_1, \ldots, v_{|V|}$. Proceeding in increasing order down the list from v_t, let $(v_{r_1}, v_{r_1+1}), (v_{r_2}, v_{r_2+1}), \ldots$ be the pairs encountered (before leaving β) which do not correspond to edges in α (and hence will later appear in the list of removed pairs).

Construct a sequence of subgraphs β_j of β as follows. The graph β_1 is the line on vertices $v_t, \ldots, v_{r_1}$. To construct the graph β_j, start from β_{j-1} and add the line from $v_{r_{j-1}+1}$ to v_{r_j} (by definition all these edges appear in β). Since v_{r_j} must have at least degree 2, it has a neighbor u_j in β among the vertices v_a for $a < r_j$ aside from v_{r_j-1}. (If v_{r_j} had a neighbor not yet appearing in the list, then v_{r_j+1} would have been that neighbor, contrary to assumption.) Choose any such neighbor and add it to β_j; this finishes construction of the graph β_j. For later use, note that either adding the edge to u_j turns $\beta_j \setminus b_{j-1}$ into a cycle or u_j is not itself among the v_r's, since otherwise in constructing the list we would have done a reordering operation.

In each of the graphs β_j, the number of edges is equal to the number of vertices. To obtain β, we must add $E_\beta - V_\beta$ edges (where E_β is the number of edges in β and V_β is the number of vertices). We claim that in so doing at least one half of a distinct such edge must be added per β_j; we prove this via a charging scheme. As noted above, each graph $\beta_j \setminus \beta_{j-1}$ either contains $v_{r_{j-1}}$ as a degree-1 vertex or it forms cycle. If it contains a degree-1 vertex, by construction this vertex is not $u_{j'}$ for any $j' > j$, otherwise we would have reordered. So charge β_j to the edge

which must be added to fix the degree-1 vertex.

In the cycle case, either some edge among the $E_\beta - V_\beta$ additional edges is added incident to the cycle (in which case we charge β_j to this edge), or some $u_{j'}$ for $j' > j$ is in $\beta_j \setminus \beta_{j-1}$. If the latter, then $\beta_{j'} \setminus \beta_{j'-1}$ contains a degree-1 vertex and $\beta_j \setminus \beta j - 1$ can be charged to the edge which fixes that degree 1 vertex. Every additional edge was charged at most twice. Thus, $R \leqslant 2(E - V)$

It is not hard to check that α can be uniquely decoded from the encoding previously described. The final result of this encoding scheme is that each α can be encoded with at most $V \log n + 6(E - V) \log V$ bits, and so there are at most $n^V \cdot V^{6(E-V)}$ choices for α. The contribution of such α to $\sum_{\alpha \in U} c(\alpha)$ is thus at most

$$n^{-(E-V)} V^{6(E-V)} (1 - \delta/2)^E k^{2(E-V_2+C)}$$

There are at least $V - (E - V)$ degree-2 vertices V_2, so $E - V_2 \leqslant E - (V - (E - V)) \leqslant 2(E - V)$. Furthermore, $C \leqslant E - V$, since no connected component of α is a cycle. All in all, this is at most $(V^6 k^6/n)^{E-V}(1 - \delta/2)^E$. So as long as $k, V \leqslant n^{0.01}$, we obtain that this contributes at most $n^{-(E-V)/2}(1 - \delta/2)^E$. Summing across all $V, E \leqslant n^{0.01}$ with $E \geqslant V + 1$, the lemma follows. □

Lemma 8.6.5. *For $\alpha \subseteq \binom{n}{2}$, let $V(\alpha)$ be the number of vertices in α, let $C(\alpha)$ be the number of connected components in α. For constants $1 > \delta > 0$ and k, let $c(\alpha) \stackrel{\text{def}}{=} \left(1 + O(\frac{d}{n})\right)^{|\alpha|} \cdot n^{-|\alpha|} \cdot (1 - \delta)^{|\alpha|} \cdot k^{2(|\alpha| - V(\alpha) + C(\alpha))}$ Let $\ell \leqslant n^{\xi/k^2}$ for some $\xi > 0$ (allowing $\xi \leqslant o(1)$) and*

$$U = \left\{ \alpha \subseteq \binom{n}{2} \,:\, \alpha \text{ is a union of cycles} \right\} .$$

Then

$$\sum_{\alpha \in U} c(\alpha) \leqslant \exp(k^2 \cdot O_\delta(1)) .$$

Proof. Let U_t be the set of α which are unions of t-cycles (we exclude the empty α). Let $c_t = \sum_{\alpha \in U_t} c_\alpha$. Then

$$\sum_{\alpha \in U} c(\alpha) \leqslant \prod_{t \leqslant \ell} (1 + c_t)\,.$$

Count the $\alpha \in U_t$ which contain exactly p cycles of length t by first choosing a list of pt vertices—there are n^{pt} choices. In doing so we will count each alpha $p!t^p$ times, since each of the p cycles can be rotated and the cycles can themselves be exchanged. All in all, there are at most $n^{pt}/(p!t^p)$ such α, and they contribute at most

$$\frac{c(\alpha) n^{pt}}{p!t^p} \leqslant \frac{(1-\delta/2)^{pt} k^{2p}}{p!t^p}\,.$$

for large enough n. Summing over all $p > 0$, we get

$$c_t \leqslant \sum_{p \geqslant 1} \left(\frac{(1-\delta/2)k^2}{t} \right)^p / p! = \exp\left(\frac{(1-\delta/2)k^2}{t} \right) - 1$$

So

$$\prod_{t \leqslant \ell} (1 + c_t) \leqslant \exp\left(\sum_{t \geqslant 0} \frac{(1-\delta/2)k^2}{t} \right) \leqslant \exp(k^2 \cdot O_\delta(1))\,.$$

□

8.6.2 Lower bound for estimating communities

Theorem 8.6.6. *Let $d, \varepsilon, k, \delta$ be constants such that $\varepsilon^2 d < (1-\delta)k^2$. Let $f : \{0,1\}^{n \times n} \to \mathbb{R}$ be any function, let $i, j \in [n]$ be distinct. Then if f satisfies $\mathbb{E}_{G \sim G(n, \frac{d}{n})} f(G) = 0$ and is correlated with the indicator $\mathbf{1}_{\sigma_i = \sigma_j}$ that i and j are in the same community in the following sense:*

$$\frac{\mathbb{E}_{G \sim SBM(n,d,\varepsilon,k)} f(G)(\mathbf{1}_{\sigma_i = \sigma_j} - \frac{1}{k})}{(\mathbb{E}_{G \sim G(n, \frac{d}{n})} f(G)^2)^{1/2}} \geqslant \Omega(1)$$

then $\deg f \geqslant n^{c(d,\varepsilon,k)}$ *for some* $c(d, \varepsilon, k) > 0$.

Proof. Let $g(G) = \mu(G)\,\mathbb{E}[\mathbf{1}_{\sigma_i=\sigma_j} - \frac{1}{k} \mid G]$, where $\mu(G)$ is the relative density of $SBM(n, d, \varepsilon, k)$. Standard Fourier analysis shows that the optimal degree-ℓ choice for such f to maximize the above correlation is $g^{\leqslant \ell}$, the orthogonal projection of g to the degree-ℓ polynomials with respect to the measure $G(n, \frac{d}{n})$, and the correlation is at most $\|g^{\leqslant \ell}\|$. It suffices to show that for some constant $c(d, \varepsilon, k)$, if $\ell < n^{c(d,\varepsilon,k)}$ then $\|g^{\leqslant \ell}\| \leqslant o(1)$.

For this we expand g in the Fourier basis, noting that

$$\widehat{g}(\alpha) = \mathop{\mathbb{E}}_{\sigma, G \sim SBM(n,d,\varepsilon,k)} \langle \tilde{\sigma}_i, \tilde{\sigma}_j \rangle \chi_\alpha(G)$$

where as usual $\tilde{\sigma}_i = \sigma_i - \frac{1}{k} \cdot 1$ is the centered indicator of i's community. By-now routine computations show that

$$\widehat{g}(\alpha)^2 \leqslant \left((1 + O(d/n))\varepsilon^2 \tfrac{d}{n}\right)^{|\alpha|} \cdot \left(\mathbb{E}\langle \tilde{\sigma}_i, \tilde{\sigma}_j \rangle \cdot \prod_{(k,\ell)\in\alpha} \langle \tilde{\sigma}_i, \tilde{\sigma}_j \rangle \right)^2$$

We assume that $(i, j) \notin \alpha$; it is not hard to check that such α's dominate the norm $\|g^{\leqslant \ell}\|$. If some vertex aside from i, j in α has degree 1 then this is zero. Similarly, if i or j does not appear in α then this is zero. Otherwise,

$$\widehat{g}(\alpha)^2 \leqslant ((1 + O(d/n)))^{|\alpha|}\, n^{-|\alpha|} (1 - \delta)^{|\alpha|} k^{2(|\alpha| - V(\alpha) + C(\alpha)}$$

where as usual $V(\alpha)$ is the number of vertices in α and $C(\alpha)$ is the number of connected components in α. Let $\beta(\alpha)$ be the connected component of α containing i and j (if they are not in the same component the arguments are mostly unchanged). Then we can bound

$$\|g^{\leqslant \ell}\|^2 = \sum_{|\alpha| \leqslant \ell} \widehat{g}(\alpha)^2 \leqslant \|\mu^{\leqslant \ell}\|^2 \cdot \sum_{\beta} ((1 + O(d/n)))^{|\beta|}\, n^{-|\beta|} (1 - \delta)^{|\beta|} k^{2(|\beta| - V(\beta) + 1}$$

where β ranges over connected graphs with vertices from $[n]$, at most ℓ edges, every vertex except i and j having degree at least 2, and containing i and j

with degree at least 1. There are at most $n^{V-2}V^{O(E-V)}$ such graphs containing at V vertices aside from i and j and E edges (by an analogous argument as in Lemma 8.6.4). The total contribution from such β is therefore at most

$$\frac{k^{2(E-V+1)}V^{O(E-V)}}{n^{E-V+2}}$$

Summing over V and E, we get

$$\sum_{\beta}((1+O(d/n)))^{|\beta|}\, n^{-|\beta|}(1-\delta)^{|\beta|}k^{2(|\beta|-V(\beta)+1} \leqslant n^{-\Omega(1)}$$

so long as $\ell \leqslant n^c$ for small enough c. Using Theorem 8.6.1 to bound $\|\mu^{\leqslant \ell}\|$ finishes the proof. □

8.7 Tensor decomposition from constant correlation

Problem 8.7.1 (Orthogonal n-dimensional 4-tensor decomposition from constant correlation)**.** Let $a_1, \ldots, a_m \in \mathbb{R}^n$ be orthonormal, and let $A = \sum_{i=1}^m a_i^{\otimes 4}$. Let $B \in (\mathbb{R}^n)^{\otimes 4}$ satisfy $\frac{\langle A,B\rangle}{\|A\|\|B\|} \geqslant \delta = \Omega(1)$.
Let $\mathcal{O}$ be an oracle such that for any unit $v \in \mathbb{R}^n$,

$$\mathcal{O}(v) = \begin{cases} \textbf{YES} \text{ if } \sum_{i=1}^m \langle a_i, v\rangle^4 \geqslant \delta^{O(1)} \\ \textbf{NO} \text{ otherwise} \end{cases}$$

Input: The tensor B, and if $\delta < 0.01$, access to the oracle $\mathcal{O}$.

Goal: Output orthonormal vectors $b_1, \ldots, b_m$ so that there is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} \cdot m$ where for every $i \in S$ there is $j \leqslant m$ with $\langle b_j, a_i\rangle^2 \geqslant \delta^{O(1)}$.

We will give an $n^{1/\delta^{O(1)}}$-time algorithm (hence using at most $n^{1/\delta^{O(1)}}$ oracle calls) for this problem based on a maximum-entropy Sum-of-Squares relaxation.

The main theorem is the following; the subsequent corollary arrives at the final algorithm.

Theorem 8.7.2. *Let A, B and $a_1, \ldots, a_m$ and $\delta \leqslant 0.01$ be as in Problem 8.7.1. Let $v_1, \ldots, v_r$ for $r \leqslant \delta^4 m$ be orthonormal vectors. There is a randomized algorithm ALG with running time $n^{O(1)}$ which takes input $B, v_1, \ldots, v_r$ and outputs a unit vector v, orthogonal to $v_1, \ldots, v_r$, with the following guarantee. There is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} \cdot m$ so that for $i \in S$,*

$$\mathbb{P}\left\{\langle v, a_i\rangle^2 \geqslant \delta^{O(1)}\right\} \geqslant n^{-1/\mathrm{poly}(\delta)}.$$

The following corollary captures the overall algorithm for tensor decomposition, using the oracle $\mathcal{O}$ to filter the output of the algorithm of Theorem 8.7.2.

Corollary 8.7.3. *Let $a_1, \ldots, a_n, A, B, \delta$ be as in Theorem 8.7.2 and $\mathcal{O}$ as in Problem 8.7.1. There is a $n^{\mathrm{poly}(1/\delta)}$-time algorithm which takes the tensor B as input and returns $b_1, \ldots, b_m$ such that with high probability there is a set $S \subseteq [m]$ of size $|S| \geqslant \delta^{O(1)} m$ which has the guarantee that for all $i \in S$ there is $j \leqslant m$ with $\langle a_i, b_j\rangle^2 \geqslant \delta^{O(1)}$. If $\delta \leqslant 1 - \Omega(1)$, the algorithm makes $n^{1/\mathrm{poly}(\delta)}$ adaptive queries to the oracle $\mathcal{O}$.*

The algorithm can also be implemented with nonadaptive queries as follows. Once the input B and the random coins of the algorithm are fixed, there is a list of at most $n^{\mathrm{poly}(k/\delta)}$. Query the oracle $\mathcal{O}$ nonadaptively on all these vectors and assemble the answers into a lookup table; then the decomposition algorithm can be run using access only to the lookup table.

Proof of Corollary 8.7.3. If $\delta \geqslant 1 - \varepsilon^*$ for a small enough constant ε^* then the tensor decomposition algorithm of Schramm and Steurer has the appropriate guarantees. (See Theorem 4.4 and Lemma 4.9 in [161]. This algorithm has several advantages,

including that it does not need to solve any semidefinite program, but it cannot handle the high-error regime we need to address here.)

From here on we assume $\delta \leqslant 0.01 < 1 - \varepsilon^*$. (Otherwise, we can replace δ with $\delta^C \leqslant 0.01$ for large enough C.) Our algorithm is as follows.

Algorithm 8.7.4 (Constant-correlation tensor decomposition)**.**

1. Let V be an empty set of vectors.
2. For rounds $1, \ldots, T = \delta^{O(1)}m$, do:
   (a) Use the algorithm of Theorem 8.7.2 on the tensor B to generate $w_1, \ldots, w_t$, where $t = n^{1/\delta^{O(1)}}$.
   (b) Call $\mathcal{O}$ on successive vectors $w_1, \ldots, w_t$, and let w be the first for which it outputs **YES**. (If no such vector exists, the algorithm halts and outputs random orthonormal vectors $b_1, \ldots, b_m$.)
   (c) Add w to V.
3. Let $b_1, \ldots, b_{m-|V|}$ be random orthonormal vectors, orthogonal to each $v \in V$.
4. Output $\{b_1, \ldots, b_{m-|V|}\} \cup V$.

Choosing $t = n^{1/\delta^{O(1)}}$ large enough, and $T = \delta^{O(1)}m$ small enough, by Theorem 8.7.2 with high probability in every round $1, \ldots, T$ there is some w among $w_1, \ldots, w_t$ for which $\mathcal{O}$ outputs **YES**. Suppose that occurs. In this case, the algorithm outputs (along with some random vectors b_i) a set of vectors V which are orthonormal, and each $v \in V$ satisfies $\langle v, a_i \rangle \geqslant \delta^{O(1)}$ for some a_i; say that this a_i is *covered* by v. Each a_i can be covered at most $1/\delta^{O(1)}$ times, by orthonormality of the set V. So, at least $\delta^{O(1)}|V| = \delta^{O(1)}m$ vectors are covered at least once, which proves the corollary. □

We turn to the proof of Theorem 8.7.2. We will use the following lemmas,

whose proofs are later in this section. The problem is already interesting when the list $v_1, \dots, v_r$ is empty, and we encourange the reader to understand this case first.

The first lemma says that a pseudodistribution of high entropy (in the 2-norm sense[29]) which is correlated with the tensor B must also be nontrivially correlated with A.

Lemma 8.7.5. *Let A, B be as in Problem 8.7.1. Let $v_1, \dots, v_r \in \mathbb{R}^n$ be orthonormal, with $r \leqslant \delta^4 m$. Suppose $\tilde{\mathbb{E}}$ is the degree-4 pseudodistribution solving*

$$\min \| \tilde{\mathbb{E}}\, x^{\otimes 4} \|_F \tag{8.7.1}$$

$$\text{s.t. } \tilde{\mathbb{E}} \text{ satisfies } \{\|x\|^2 \leqslant 1, \langle x, v_1\rangle = 0, \dots, \langle x, v_r\rangle = 0\}$$

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, B\rangle \geqslant \frac{\delta}{2m}$$

$$\left\| \tilde{\mathbb{E}}\, x x^\top \right\| \leqslant \tfrac{1}{m} \tag{8.7.2}$$

$$\left\| \tilde{\mathbb{E}}\, x x^\top \otimes x x^\top \right\| \leqslant \tfrac{1}{m} \tag{8.7.3}$$

Then $\tilde{\mathbb{E}} \sum_{i \leqslant m} \langle x, a_i\rangle^4 \geqslant \delta^2/8$. Furthermore, it is possible to find $\tilde{\mathbb{E}}$ in polynomial time.[30]

The second lemma says that given a high-entropy (in the spectral sense of [121]) pseudodistribution $\tilde{\mathbb{E}}$ having nontrivial correlation with some $a \in \mathbb{R}^n$, contracting $\tilde{\mathbb{E}}$ with a yields a matrix whose quadratic form is large at a and which does not have too many large eigenvalues.

Lemma 8.7.6. *Let $a_1, \dots, a_m \in \mathbb{R}^n$ be orthonormal.*

Let $\tilde{\mathbb{E}}$ be a degree-4 pseudoexpectation such that

[29] For a distribution μ finitely-supported on a family of orthonormal vectors, the Frobenious norm $\| \mathbb{E}_{x\sim\mu}\, x^{\otimes k}\|$ is closely related to the collision probability of μ, itself closely related to the order-2 case of Rényi entropy.

[30] Up to inverse-polynomial error, which we ignore here. See [121] for the ideas needed to show polynomial-time solvability.

1. $\tilde{\mathbb{E}}$ *satisfies* $\{\|x\|^2 \leqslant 1\}$
2. $\tilde{\mathbb{E}} \sum_{i \leqslant m} \langle x, a_i \rangle^4 \geqslant \delta$.
3. $\|\tilde{\mathbb{E}} xx^\top]\|_{op}, \|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\|_{op} \leqslant \frac{1}{m}$.[31]

Let $M_i \in \mathbb{R}^{n \times n}$ *be the matrix* $\tilde{\mathbb{E}} \langle x, a_i \rangle^2 xx^\top$. *For every* $i \in [m]$, *the matrix* M_i *has at most* $4/\delta$ *eigenvalues larger than* $\frac{\delta}{4m}$. *Furthermore,*

$$\mathop{\mathbb{P}}_{i \sim [m]} \left\{ \langle a_i, M_i a_i \rangle \geqslant \tfrac{\delta}{2m} \right\} \geqslant \tfrac{\delta}{2} .$$

The last lemma will help show that a random contraction of a high-entropy pseudodistribution behaves like one of the contractions from Lemma 8.7.6, with at least inverse-polynomial probability.

Lemma 8.7.7. *Let* $g \sim \mathcal{N}(0, \Sigma)$ *for some* $0 \preceq \Sigma \preceq \mathrm{Id}$ *and let* $\tilde{\mathbb{E}}$ *be a degree-4 pseudoexpectation where*

- $\tilde{\mathbb{E}}$ *satisfies* $\{\|x\|^2 \leqslant 1\}$.
- $\left\|\tilde{\mathbb{E}} xx^\top\right\| \leqslant c$.
- $\left\|\tilde{\mathbb{E}} xx^\top \otimes xx^\top\right\| \leqslant c$

Then

$$\mathop{\mathbb{E}}_{g} \left\|\tilde{\mathbb{E}} \langle g, x \rangle^2 xx^\top\right\| \leqslant O(c \cdot \log n) .$$

Now we can prove Theorem 8.7.2.

Proof of Theorem 8.7.2. The algorithm is as follows:

[31] Recall that $\|\cdot\|$ denotes the operator norm, or maximum singular value, of a matrix.

Algorithm 8.7.8 (Low-correlation tensor decomposition). 1. Use the first part of Lemma 8.7.5 to obtain a degree-4 pseudoexpectation with $\tilde{\mathbb{E}} \sum_{i\in[m]} \langle a_i, x\rangle^4 \geqslant \delta^2/4$ satisfying $\{\|x\|^2 \leqslant 1, \langle x, v_1\rangle = 0, \ldots, \langle x, v_r\rangle = 0\}$.

2. Sample a random $g \sim \mathcal{N}(0, \mathrm{Id})$ and compute the contraction $M = \tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top$.

3. Output a random unit vector b in the span of the top $\frac{32}{\delta^2}$ eigenvectors of M.

First note that for any $v \in \mathrm{Span}\{v_1, \ldots, v_r\}$, we must have $\langle v, Mv\rangle = \tilde{\mathbb{E}}\langle g, x\rangle^2\langle v, x\rangle^2 = 0$, so v lies in the kernel of M. Hence, the ouput of the algorithm will always be orthogonal to $v_1, \ldots, v_r$.

Let Π_{32/δ^2} be the projector to the top $32/\delta^2$ eigenvectors of M. For any unit vector a with $\|\Pi_{32/\delta^2} a\| \geqslant \delta^{O(1)}$, the algorithm will output b with nontrivial correlation with a. Formally, for any such a,

$$\mathop{\mathbb{E}}_{b} \langle b, a\rangle^2 \geqslant \delta^{O(1)}.$$

So, our goal is to show that for a $\delta^{O(1)}$-fraction of the vectors $a_1, \ldots, a_m$,

$$\mathop{\mathbb{P}}_{g}\{\|\Pi_{32/\delta^2} a_i\| \geqslant \delta^{O(1)}\} \geqslant n^{-1/\delta^{O(1)}}.$$

For $i \in [m]$, let $M_i = \tilde{\mathbb{E}}\langle a_i, x\rangle^2 xx^\top$. Let i be the index of some a_i so that

$$\langle a_i, M_i a_i\rangle \geqslant \tfrac{\delta^2}{16m} \text{ and } \operatorname{rank} M_i^{\geqslant \frac{\delta^2}{32m}} \leqslant \tfrac{32}{\delta^2}$$

as in Lemma 8.7.6. (There are $\Omega(\delta^2 m)$ possible choices for a_i, according to the Lemma.)

We expand the Gaussian vector g from the algorithm as

$$g = g_0 \cdot a_i + g'$$

where $g_0 \sim \mathcal{N}(0,1)$ and $\langle g', a_i \rangle = 0$. We note for later use that g' is a Gaussian vector independent of g_0 and that $\mathbb{E}(g')(g')^\top \preceq \mathrm{Id}$. Using this expansion,

$$M = g_0^2 \, \tilde{\mathbb{E}} \langle a_i, x \rangle^2 x x^\top + 2 \cdot g_0 \, \tilde{\mathbb{E}} \langle g', x \rangle \langle a_i, x \rangle x x^\top + \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \, .$$

We will show that all but the first term have small spectral norm. Addressing the middle term first, by Cauchy-Schwarz, for any unit $v \in \mathbb{R}^n$,

$$\tilde{\mathbb{E}} \langle g', x \rangle \langle a_i, x \rangle \langle v, x \rangle^2 \leqslant \left(\tilde{\mathbb{E}} \langle g', x \rangle^2 \langle x, v \rangle^2 \right)^{1/2} \left(\tilde{\mathbb{E}} \langle a_i, x \rangle^2 \langle v, x \rangle^2 \right)^{1/2} \leqslant \left\| \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \right\|^{1/2} \cdot \left(\tfrac{1}{m} \right)^{1/2} ,$$

where in the last step we have used that $\left\| \tilde{\mathbb{E}} \, x x^\top \otimes x x^\top \right\| \leqslant \frac{1}{m}$.

By Markov's inequality and Lemma 8.7.7,

$$\mathop{\mathbb{P}}_{g'} \left\{ \left\| \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \right\| > \tfrac{t \log n}{m} \right\} \leqslant O\left(\tfrac{1}{t} \right) \, .$$

Let t be a large enough constant so that

$$\mathop{\mathbb{P}}_{g'} \left\{ \left\| \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \right\| \leqslant \tfrac{t \log n}{m} \right\} \geqslant 0.9 \, .$$

For any constant c, with probability $n^{-1/\mathrm{poly}(\delta)}$, the foregoing occurs and g_0 (which is independent of g') is large enough that

$$g_0^2 \cdot \tfrac{c\delta^2}{m} > \tfrac{1}{\delta^4} \left\| M - g_0^2 M_i \right\| \, .$$

Choosing c large enough, in this case

$$M' \stackrel{\text{def}}{=} \tfrac{1}{g_0^2} M = M_i + O(\delta^6/m) \, .$$

Hence the vector a_i satisfies

$$\tfrac{1}{g_0^2} \langle a_i, M a_i \rangle \geqslant \tfrac{\delta^2}{33m}$$

This means that the projection b of a_i into the span of eigenvectors of M' with eigenvalue at least $\delta^2/60m$ has $\|b\|^2 \geqslant \delta^{O(1)}$. This finishes the proof. □

8.7.1 Proofs of Lemmas

These lemmas and their proofs use many ideas from [121]. The main difference here is that we want to contract the tensor $\tilde{\mathbb{E}}\, x^{\otimes 4}$ in 2 modes, to obtain the matrix $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top$. For us this is useful because $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top \succeq 0$. By contrast, the tools in [121] would only allow us to analyze the contraction $\tilde{\mathbb{E}}\langle h, x \otimes x\rangle xx^\top$ for $h \sim \mathcal{N}(0, \mathrm{Id}_{n^2})$.

We start with an elementary fact.

Fact 8.7.9. *Let $a_1, \dots, a_m \in \mathbb{R}^n$ be orthonormal. Let Π be the projector to a subspace of codimension at most δm. Let $A = \sum_{i=1}^m a_i^{\otimes 4}$ and $\Pi A = \sum_{i=1}^m (\Pi a_i)^{\otimes 4}$. Then $\langle A, \Pi A\rangle \geqslant (1 - O(\sqrt{\delta}))\|A\| \cdot \|\Pi A\|$.*

A useful corollary of Fact 8.7.9 is that if T is any 4-tensor satisfying $\langle T, \Pi A\rangle \geqslant \delta\|T\|\|\Pi A\|$ and Π has codimension $\ll \delta^2 m$, then $\langle T, A\rangle \geqslant \Omega(\delta)\|T\|\|A\|$.

Proof of Fact 8.7.9. We expand

$$\langle A, \Pi A\rangle = \sum_{i,j \leqslant m} \langle a_i, \Pi a_j\rangle^4 \geqslant \sum_{i,j \leqslant m} \|\Pi a_i\|^8$$

Writing Π in the a_i basis, we think of $\|\Pi a_i\|^4 = \Pi_{ii}^2$, the square of the i-th diagonal entry of Π. Since Π has codimension at most δm,

$$\operatorname{rank} \Pi = \operatorname{Tr} \Pi = \sum_{i \leqslant n} \Pi_{ii} \geqslant n - \delta m\,.$$

Furthermore, for each i, it must be that $0 \leqslant \Pi_{ii} \leqslant 1$. By Markov's inequality, at most $\sqrt{\delta} m$ diagonal entries of Π can be less than $1 - \sqrt{\delta}$ in magnitude. Hence, $\sum_{i \leqslant m} \Pi_{ii}^4 \geqslant (1 - 4\sqrt{\delta})m$. On the other hand, $\|A\|^2 = m$; this proves the fact. □

Now we can prove Lemma 8.7.5.

Proof of Lemma 8.7.5. We will appeal to Theorem 8.2.3. Let C be the convex set of all pseudo-moments $\tilde{\mathbb{E}}\, x^{\otimes 4}$ such that $\tilde{\mathbb{E}}$ is a deg-4 pseudo-distribution that satisfies the polynomial constraints $\{\|x\|^2 \leqslant 1, \langle x, v_i\rangle = 0\}$ and the operator norm conditions

$$\left\|\tilde{\mathbb{E}}\, xx^\top\right\| \leqslant \tfrac{1}{m},$$
$$\left\|\tilde{\mathbb{E}}\, xx^\top \otimes xx^\top\right\| \leqslant \tfrac{1}{m}.$$

Let Π be the projector to the orthogonal space of $v_1, \ldots, v_r$. Notice that $\frac{1}{m}\Pi A \in C$. Furthermore, $\langle B, \Pi A\rangle \geqslant \delta/2$ by Fact 8.7.9, the assumption that $r \leqslant \delta^4 m$, and the assumption $\delta \leqslant 0.01$. By Theorem 8.2.3, and Fact 8.7.9 again, the optimizer of the convex program in the Lemma satisfies $\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, \frac{1}{m}A\rangle \geqslant \frac{\delta^2}{8m}$) and the result follows. □

Proof of Lemma 8.7.6. By the assumption $\|\tilde{\mathbb{E}}\, xx^\top \otimes xx^\top\| \leqslant \frac{1}{m}$, for every a_i it must be that $\tilde{\mathbb{E}}\langle x, a_i\rangle^4 \leqslant \frac{1}{m}$. Since $\tilde{\mathbb{E}}\sum_{i=1}^m \langle x, a_i\rangle^4 \geqslant \delta$, at least $\delta m/2$ of the a_i's must satisfy $\tilde{\mathbb{E}}\langle x, a_i\rangle^4 \geqslant \frac{\delta}{2m}$. Rewritten, for any such a_i we obtain $\langle a_i, M_i a_i\rangle \geqslant \frac{\delta}{2m}$.

For any M_i,

$$\operatorname{Tr} M_i = \tilde{\mathbb{E}}\langle x, a_i\rangle^2 \|x\|^2 = \tilde{\mathbb{E}}\langle x, a_i\rangle^2 \leqslant \tfrac{1}{m}$$

because $\|\tilde{\mathbb{E}}\, xx^\top\| \leqslant \frac{1}{m}$. Also, $M_i \succeq 0$. Hence, M_i can have no more than $\frac{4}{\delta}$ eigenvalues larger than $\frac{\delta}{4m}$. □

Now we turn to the proof of Lemma 8.7.7. We will need spectral norm bounds on certain random matrices associated to the random contraction $\tilde{\mathbb{E}}\langle g, x\rangle xx^\top$. The following are closely related to Theorem 6.5 and Corollary 6.6 in [121].

Lemma 8.7.10. *Let $g \sim \mathcal{N}(0, \mathrm{Id})$ and let $\tilde{\mathbb{E}}$ be a degree-4 pseudoexpectation where*

- $\tilde{\mathbb{E}}$ *satisfies* $\{\|x\|^2 = 1\}$.
- $\left\|\tilde{\mathbb{E}}\, xx^\top\right\| \leqslant c$.
- $\left\|\tilde{\mathbb{E}}\, xx^\top \otimes xx^\top\right\| \leqslant c$

Then

$$\mathop{\mathbb{E}}_{g} \left\|\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top\right\| \leqslant O(c \cdot \log n)\,.$$

Before proving the lemma, we will need a classical decoupling inequality.

Fact 8.7.11 (Special case of Theorem 1 in [57])**.** *Let* $g, h \sim \mathcal{N}(0, \mathrm{Id}_n)$ *be independent. Let* M_{ij} *for* $i, j \in [n]$ *be a family of matrices. There is a universal constant* C *so that*

$$\mathop{\mathbb{E}}_{g} \left\|\sum_{i\neq j} g_i g_j \cdot M_{ij}\right\| \leqslant C \cdot \mathop{\mathbb{E}}_{g,h} \left\|\sum_{i\neq j} g_i h_j \cdot M_{ij}\right\|\,.$$

We will also need a theorem from [121].

Fact 8.7.12 (Corollary 6.6 in [121])**.** *Let* $T \in \mathbb{R}^p \otimes \mathbb{R}^q \otimes \mathbb{R}^r$ *be an order-3 tensor. Let* $g \sim \mathcal{N}(0, \Sigma)$ *for some* $0 \preceq \Sigma \preceq \mathrm{Id}_r$*. Then for any* $t \geqslant 0$*,*

$$\mathop{\mathbb{P}}_{g}\left\{\left\|(\mathrm{Id}\otimes\mathrm{Id}\otimes g)^\top T\right\|_{\{1\},\{2\}} \geqslant t\cdot \max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}\right\} \leqslant 2(p+q)\cdot e^{-t^2/2}\,,$$

and consequently,

$$\mathop{\mathbb{E}}_{g}\left[\left\|(\mathrm{Id}\otimes\mathrm{Id}\otimes g)^\top T\right\|_{\{1\},\{2\}}\right] \leqslant O(\log(p+q))^{1/2}\cdot \max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}$$

Proof of Lemma 8.7.10. We expand the matrix $\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top$ as

$$\tilde{\mathbb{E}}\langle g, x\rangle^2 xx^\top = \sum_{i\in[n]} g_i^2\, \tilde{\mathbb{E}}\, x_i^2 xx^\top + \sum_{i\neq j\in[n]} g_i g_j \cdot \tilde{\mathbb{E}}\, x_i x_j xx^\top\,.$$

Addressing the first term, by standard concentration, $\mathbb{E}\max_{i\in[n]} g_i^2 = O(\log n)$. So,

$$\mathop{\mathbb{E}}_{g}\left\|\sum_{i\in[n]} g_i^2\,\tilde{\mathbb{E}}\,x_i^2 xx^\top\right\| \leqslant \mathop{\mathbb{E}}_{g}\left[\max_{i\in[n]} g_i^2\cdot\left\|\tilde{\mathbb{E}}\,\|x\|^2 xx^\top\right\|\right] = O(\log n)\cdot\left\|\tilde{\mathbb{E}}\,xx^\top\right\| = O(c\cdot\log n)\,.$$

The second term we will decouple using Fact 8.7.11.

$$\mathop{\mathbb{E}}_{g}\left\|\sum_{i\neq j} g_i g_j\cdot\tilde{\mathbb{E}}\,x_i x_j xx^\top\right\| \leqslant O(1)\cdot\mathop{\mathbb{E}}_{g,h}\left\|\sum_{i\neq j} g_i h_j\cdot\tilde{\mathbb{E}}\,x_i x_j xx^\top\right\|\,.$$

We add some aditional terms to the sum; by similar reasoning to our bound on the first term they do not contribute too much to the norm.

$$\mathop{\mathbb{E}}_{g,h}\left\|\sum_{i\neq j} g_i h_j\cdot\tilde{\mathbb{E}}\,x_i x_j xx^\top\right\| \leqslant O(1)\cdot\mathop{\mathbb{E}}_{g,h}\left\|\sum_{i,j\in[n]} g_i h_j\cdot\tilde{\mathbb{E}}\,x_i x_j xx^\top\right\| + O(c\cdot\log n)\,.$$

We can rewrite the matrix in the first term on the right-hand side as

$$\sum_{i,j\in[n]} g_i h_j\cdot\tilde{\mathbb{E}}\,x_i x_j xx^\top = \tilde{\mathbb{E}}\langle g,x\rangle\langle h,x\rangle xx^\top\,.$$

Now we can apply Fact 8.7.12 twice in a row; first to g and then to h, which together with our norm bound on $\mathbb{E}\,xx^\top\otimes xx^\top$, gives

$$\mathop{\mathbb{E}}_{g,h}\left\|\tilde{\mathbb{E}}\langle g,x\rangle\langle h,x\rangle xx^\top\right\| \leqslant O(c\cdot\log n)\,.$$

Putting all of the above together, we get the lemma. □

Next we prove Lemma 8.7.7 as a corollary of Lemma 8.7.7 which applies to random contractions which are non-spherical. The proof technique is very similar to that for Fact 8.7.12.

Proof of Lemma 8.7.7. Let $h \sim \mathcal{N}(0, \mathrm{Id} - \Sigma)$ be independent of g, and define $g' = g + h$ and $g'' = g - h$, so that $g = \frac{1}{2}(g' + g'')$. It is sufficient to bound

$\mathbb{E}_{g,h} \left\| \tilde{\mathbb{E}} \langle g' + g'', x \rangle^2 x x^\top \right\|$. Expanding and applying triangle inequality,

$$\mathop{\mathbb{E}}_{g,h} \left\| \tilde{\mathbb{E}} \langle g' + g'', x \rangle^2 x x^\top \right\| \leqslant \mathop{\mathbb{E}}_{g,h} \left\| \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \right\| + 2 \mathop{\mathbb{E}}_{g,h} \left\| \tilde{\mathbb{E}} \langle g', x \rangle \langle g'', x \rangle x x^\top \right\| + \mathop{\mathbb{E}}_{g,h} \left\| \tilde{\mathbb{E}} \langle g'', x \rangle^2 x x^\top \right\| .$$

The first and last terms are $O(c \cdot \log n)$ by Lemma 8.7.10. For the middle term, consider the quadratic form of the matrix $\tilde{\mathbb{E}} \langle g', x \rangle \langle g'', x \rangle x x^\top$ on a vector $v \in \mathbb{R}^n$:

$$\tilde{\mathbb{E}} \langle g', x \rangle \langle g'', x \rangle \langle x, v \rangle^2 \leqslant \tilde{\mathbb{E}} \langle g', x \rangle^2 \langle x, v \rangle^2 + \tilde{\mathbb{E}} \langle g'', x \rangle^2 \langle x, v \rangle^2$$

by pseudoexpectation Cauchy-Schwarz. Thus for every g', g'',

$$\left\| \tilde{\mathbb{E}} \langle g', x \rangle \langle g'', x \rangle x x^\top \right\| \leqslant \left\| \tilde{\mathbb{E}} \langle g', x \rangle^2 x x^\top \right\| + \left\| \tilde{\mathbb{E}} \langle g'', x \rangle^2 x x^\top \right\| .$$

Together with Lemma 8.7.10 this concludes the proof. □

8.7.2 Lifting 3-tensors to 4-tensors

Problem 8.7.13 (3-to-4 lifting)**.** Let $a_1, \ldots, a_m \in \mathbb{R}^n$ be orthonormal. Let $A_3 = \sum_{i=1}^m a_i^{\otimes 3}$ and $A_4 = \sum_{i=1}^m a_i^{\otimes 4}$. Let $B \in \mathbb{R}^{n \times n \times n}$ satisfy $\langle B, A_3 \rangle \geqslant \delta \cdot \|A_3\| \cdot \|B\|$.

Input: The tensor B.

Goal: Output B' satisfying $\langle B', A_4 \rangle \geqslant \delta^{O(1)} \cdot \|A_4\| \cdot \|B'\|$.

Theorem 8.7.14. *There is a polynomial time algorithm, using the sum of squares method, which solves the 3-to-4 lifting problem.*

Proof. **Small δ regime: $\delta < 1 - \Omega(1)$:** The algorithm is to output the fourth moments of the optimizer of the following convex program.

$$\begin{aligned} \min_{\tilde{\mathbb{E}}} \quad & \| \tilde{\mathbb{E}}\, x^{\otimes 3} \| \\ \text{s.t.} \quad & \tilde{\mathbb{E}} \text{ is degree-4} \end{aligned}$$

$$\tilde{\mathbb{E}} \text{ satisfies } \{\|x\|^2 = 1\}$$
$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, B\rangle \geqslant \frac{\delta\|B\|}{\sqrt{m}}$$
$$\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant \frac{1}{\sqrt{m}}\,.$$

To analyze the algorithm we apply Theorem 8.2.3. Let C be the set of degree-4 pseudodistributions satisfying $\{\|x\|^2 = 1\}$ and having $\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1/\sqrt{m}$. The uniform distribution over $a_1, \ldots, a_m$, whose third and fourth moments are $\frac{1}{m}A_3$ and $\frac{1}{m}A_4$, respectively, is in C.

Let $\tilde{\mathbb{E}}$ be the pseudoexpectation solving the convex program. By Theorem 8.2.3,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \tfrac{1}{m}A_3\rangle \geqslant \frac{\delta}{2}\cdot\frac{1}{\sqrt{m}}\cdot\|\tilde{\mathbb{E}}\, x^{\otimes 3}\| \geqslant \frac{\delta^2}{2m}$$

At the same time,

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 3}, \tfrac{1}{m}A_3\rangle = \frac{1}{m}\sum_{i=1}^m \tilde{\mathbb{E}}\langle x, a_i\rangle^3 \leqslant \frac{1}{m}\left(\tilde{\mathbb{E}}\sum_{i=1}^m \langle x, a_i\rangle^4\right)^{1/2}$$

by Cauchy-Schwarz. Putting these together, we obtain

$$\langle \tilde{\mathbb{E}}\, x^{\otimes 4}, A_4\rangle = \tilde{\mathbb{E}}\sum_{i=1}^m \langle x, a_i\rangle^4 \geqslant \delta^4/4\,.$$

Finally, $\|A_4\|\cdot\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1$ (since we constrained $\|\tilde{\mathbb{E}}\, x^{\otimes 4}\| \leqslant 1/\sqrt{m}$), which finishes the proof.

Large δ regime: $\delta \geqslant 1 - o(1)$: Modify the convex program from the small-δ regime to project $(B/\|B\|)\cdot 1/\sqrt{m}$ to same convex set C. The normalization is so that

$$\left\|(B/\|B\|)\cdot 1/\sqrt{m}\right\| = \left\|\tfrac{1}{m}\cdot A_3\right\|\,.$$

The analysis is similar. □

8.8 Toolkit and Omitted Proofs

8.8.1 Probability and linear algebra tools

Fact 8.8.1. *Consider any inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ with associated norm $\|\cdot\|$. Let X and Y be jointly-distributed $\mathbb{R}^n$-valued random variables. Suppose that $\|X\|^2 \leqslant C\,\mathbb{E}\,\|X\|^2$ with probability 1, and that*

$$\frac{\mathbb{E}\langle X, Y\rangle}{(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2}} \geqslant \delta\,.$$

Then

$$\mathbb{P}\left\{\frac{\langle X, Y\rangle}{\|X\| \cdot \|Y\|} \geqslant \frac{\delta}{2}\right\} \geqslant \frac{\delta^2}{4C^2}\,.$$

Proof of Fact 8.8.1. Let $\mathbf{1}_E$ be the 0/1 indicator of an event E. Note that

$$\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle \leqslant \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] \leqslant \tfrac{\delta}{2}\,\mathbb{E}\,\|X\| \cdot \|Y\|$$

Hence,

$$\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] = \mathbb{E}\langle X, Y\rangle - \mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle \leqslant \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] \geqslant \mathbb{E}\langle X, Y\rangle - \tfrac{\delta}{2}\,\mathbb{E}\,\|X\|\cdot\|Y\|$$

At the same time,

$$\begin{aligned}
\mathbb{E}\left[\langle X, Y\rangle\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right] &\leqslant \left(\mathbb{E}\,\|X\|^2 \cdot \|Y\|^2\right)^{1/2} \cdot \left(\mathbb{E}\,\mathbf{1}_{\langle X,Y\rangle > \frac{\delta}{2}\cdot\|X\|\cdot\|Y\|}\right)^{1/2} \\
&= \left(\mathbb{E}\,\|X\|^2 \cdot \|Y\|^2\right)^{1/2} \cdot \left(\mathbb{P}\{\langle X, Y\rangle > \tfrac{\delta}{2} \cdot \|X\| \cdot \|Y\|\}\right)^{1/2} \\
&\leqslant C(\mathbb{E}\,\|X\|^2)^{1/2}(\mathbb{E}\,\|Y\|^2)^{1/2} \cdot \left(\mathbb{P}\{\langle X, Y\rangle > \tfrac{\delta}{2} \cdot \|X\| \cdot \|Y\|\}\right)^{1/2}\,.
\end{aligned}$$

Putting the inequalities together and rearranging finishes the proof. □

Proof of Proposition 8.5.22. We decompose X_i as

$$X_i = X_i\,\mathbf{1}_{|X_i| \leqslant R} + X_i\,\mathbf{1}_{|X_i| > R}\,.$$

Let $Y_i = X_i \mathbf{1}_{|X_i| \leqslant R}$. Then

$$|\mathbb{E}\, Y_i| = |\mathbb{E}\, X_i - \mathbb{E}\, X_i \mathbf{1}_{|X_i| > R}| \leqslant \delta'$$

and

$$\mathbb{V}\, Y_i \leqslant \mathbb{E}\, Y_i^2 \leqslant \mathbb{E}\, X_i^2 .$$

So we can apply Bernstein's inequality to $\frac{1}{m}\sum_{i \leqslant m} Y_i$ to obtain that

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i \leqslant m} Y_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1) \cdot m \cdot t^2}{\mathbb{E}\, X^2 + t \cdot R}\right) .$$

Now, with probability at least $1 - \delta$ we know $X_i = Y_i$, so by a union bound,

$$\mathbb{P}\left\{\left|\frac{1}{m}\sum_{i \leqslant m} X_i\right| \geqslant t + \delta'\right\} \leqslant \exp\left(\frac{-\Omega(1) \cdot m \cdot t^2}{\mathbb{E}\, X^2 + t \cdot R}\right) + m\delta . \qquad \square$$

Fact 8.8.2. *Let $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ are jointly distributed real-valued random variables. Suppose there is $S \subseteq [m]$ with $|S| \geqslant (1 - o_m(1)) \cdot m$ such that for each $i \in S$ there a degree-D polynomials p_i satisfying*

$$\frac{\mathbb{E}\, p_i(X) Y_i}{(\mathbb{E}\, Y^2)^{1/2} (\mathbb{E}\, p_i(X)^2)^{1/2}} \geqslant \delta .$$

Furthermore, suppose $\sum_{i \in S} \mathbb{E}\, Y_i^2 \geqslant (1 - o(1)) \sum_{i \in [m]} \mathbb{E}\, Y_i^2$. Let $Y \in \mathbb{R}^m$ be the vector-valued random variable with i-th coordinate Y_i, and similarly let $P(X)$ have i-th coordinate $p_i(X)$. Then

$$\frac{\mathbb{E}\langle P(X), Y\rangle}{(\mathbb{E}\, \|Y\|^2)^{1/2} \cdot (\mathbb{E}\, \|P(X)\|^2)^{1/2}} \geqslant (1 - o(1)) \cdot \delta$$

Proof. The proof is by Cauchy-Schwarz.

$$\begin{aligned}
\mathbb{E}\langle P(X), Y\rangle &= \sum_{i \in S} \mathbb{E}\, p_i(X) Y \\
&\geqslant \delta \sum_{i \in S} (\mathbb{E}\, p_i(X)^2)^{1/2} (\mathbb{E}\, Y_i^2)^{1/2} \\
&\geqslant \delta \left(\mathbb{E} \sum_{i \in S} p_i(x)^2\right)^{1/2} \cdot (1 - o(1)) \left(\sum_{i \in [m]} Y_i^2\right)^{1/2} . \qquad \square
\end{aligned}$$

8.8.2 Tools for symmetric and Dirichlet priors

Proof of Fact 8.5.15. Let X be any $\mathbb{R}^k$-valued random variable which is symmetric in distribution with respect to permutations of coordinates and satisfies $\sum_{s\in[k]} X(s) = 0$ with probability 1. (The variable $\tilde{\sigma}$ is one example.)

We prove the claim about $\mathbb{E}\langle X, x\rangle\langle X, y\rangle\langle X, z\rangle\langle X, w\rangle$; the other proofs are similar. Consider the matrix $M = \mathbb{E}(X \otimes X)(X \otimes X)^\top$. Since x, y, z, w are orthogonal to the all-1's vector, we may add $1 \otimes v$, for any $v \in \mathbb{R}^n$, to any row or column of M without affecting the statement to be proved. Adding multiples of $1 \otimes e_i$ to rows and columns appropriately makes M a block diagonal matrix, with the top block indexed by coordinates (i, i) for $i \in [k]$ and the bottom block indexed by pairs (i, j) for $i \neq j$.

The resulting top block takes the form $c\mathrm{Id} + c'J$, where J is the all-1's matrix. The bottom block will be a matrix from the Johnson scheme. Standard results on eigenvectors of the Johnson scheme (see e.g. [63] and references therein) finish the proof. The values of constants C for the Dirichlet distribution follow from the next fact. □

Fact 8.8.3. *Let $\sigma \in \mathbb{R}^k$ be distributed according to a (symmetric) Dirichlet distribution with parameter α. That is, $\mathbb{P}(\sigma) \propto \prod_{j\in[k]} \sigma^{\alpha-1}$.*

Let $\gamma \in \mathbb{N}^k$ be a k-tuple, and let $\sigma^\gamma = \prod_{j\leqslant k} \sigma_j^{\gamma_j}$. Let $|\gamma| = \sum_{j\leqslant k} \gamma_j$. Then

$$\mathbb{E}\,\sigma^\gamma = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha + |\gamma|)} \cdot \frac{\prod_{j\leqslant k} \Gamma(\alpha + \gamma_j)}{\Gamma(\alpha)^k}.$$

Furthermore, let $\tilde{\sigma} \in \mathbb{R}^k$ be given by $\tilde{\sigma}_i = \sigma_i - \frac{1}{k}$. Then

$$\mathbb{E}\,\tilde{\sigma}\tilde{\sigma}^\top = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha + 2)} \left(\frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} - \frac{\Gamma(\alpha + 1)^2}{\Gamma(\alpha)^2} \right) \cdot \Pi = \frac{1}{k(k\alpha + 1)} \cdot \Pi,$$

where $\Pi \in \mathbb{R}^{k\times k}$ is the projector to the subspace orthogonal to the all-1s vector.

Proof. We recall the density of the k-dimensional Dirichlet distribution with parameter vector $\alpha_1, \ldots, \alpha_k$. Here Γ denotes the usual Gamma function.

$$\mathbb{P}\{\sigma\} = \frac{\Gamma(\sum_{j\leqslant k} \alpha_j)}{\prod_{j\leqslant k} \Gamma(\alpha_j)} \cdot \prod_{j\leqslant k} \sigma_j^{\alpha_j - 1}.$$

In particular,

$$\frac{\Gamma(\sum_{j\leqslant k} \alpha_j)}{\prod_{j\leqslant k} \Gamma(\alpha_j)} \cdot \int \prod_{j\leqslant k} \sigma_j^{\alpha_j - 1}\, d\sigma = 1$$

where the integral is taken with respect to Lebesgue measure on $\{\sigma : \sum_{j\leqslant k} \sigma_j = 1\}$.

Using this fact we can compute the moments of the symmetric Dirichlet distribution with parameter α. We show for example how to compute second moments; the general formula can be proved along the same lines. For $s \neq t \in [k]$,

$$\begin{aligned}
\mathbb{E}\, \sigma_s \sigma_t &= \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \cdot \int \sigma_s \sigma_t \prod_{j\leqslant k} \sigma_j^{\alpha-1} \\
&= \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)} \cdot \frac{\Gamma(\alpha+1)^2}{\Gamma(\alpha)^2} \cdot \frac{\Gamma(k\alpha+2)}{\Gamma(\alpha)^{k-2}\Gamma(\alpha+1)^2} \cdot \int \sigma_s^{(\alpha+1)-1} \sigma_t^{(\alpha+1)-1} \prod_{j\neq s,t} \sigma_j^{\alpha-1} \\
&= \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)} \cdot \frac{\Gamma(\alpha+1)^2}{\Gamma(\alpha)^2}.
\end{aligned}$$

Similarly,

$$\mathbb{E}\, \sigma_s^2 = \frac{\Gamma(k\alpha)}{\Gamma(k\alpha+2)} \cdot \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}.$$

The formula for $\mathbb{E}\, \tilde{\sigma}\tilde{\sigma}^\top$ follows immediately. □

8.9 Chapter Notes

The material in this chapter is adapted from [93], joint work with David Steurer.

CHAPTER 9

BEYOND BAYESIAN INFERENCE: MIXTURE MODELS, ROBUSTNESS, AND SUM OF SQUARES PROOFS

In Chapters 6 to 8 we studied a number of inference problems: spiked tensor models, planted sparse vectors, and stochastic block models. All of these problems are Bayesian: that is, hidden variables are distributed according to prior distributions which are known to the algorithm designer. Bayesian-ness makes all these problems accessible via the simple statistics approach.

Not every inference problem is Bayesian. In this chapter we study two important non-Bayesian problems in high-dimensional statistics: learning mixture models under separation assumptions and mean estimation in the presence of adversarial data corruption. The SoS algorithms we design for these problems follow the proofs-to-algorithms methodology from Section 3.5.

9.1 Results

Both of the problems we study have a long history; for now we just note some highlights and state our main results.

Mixture models Mixture models are fundamental generative models for inhomogeneous data – data coming from multiple underlying populations. The problem of learning mixture models dates to Pearson in 1894, who invented the method of moments in order to separate a mixture of two Gaussians [146]; they have sinse become ubiquitous in data analysis across many disciplines [169, 126]. In recent years, computer scientists have devised many ingenious algorithms for

learning mixture models as it became clear that classical statistical methods (e.g. maximum likelihood estimation) often suffer from computational intractability, especially when there are many mixture components or the components are high dimensional.

A highlight of this work is a series of algorithmic results when the components of the mixture model are Gaussian [52, 53, 17, 174]. Here the main question is: how close together can the clusters be – as measured by the minimum separation Δ between cluster centers – such that there exists an algorithm to estimate $\mu_1, \ldots, \mu_k$ from samples $x_1, \ldots, x_n$ in $\text{poly}(k, d)$ time (hence also using $n = \text{poly}(k, d)$ samples)? Focusing for simplicity on spherical Gaussian components (i.e. with covariance equal to the identity matrix Id) and with number of components similar to the ambient dimension of the data (i.e. $k = d$) and uniform mixing weights (i.e. every cluster has roughly the same representation among the samples), the best result in previous work gives a $\text{poly}(k)$-time algorithm when $\Delta \geqslant k^{1/4}$.

Separation $\Delta = k^{1/4}$ represents a natural algorithmic barrier: when $\Delta \geqslant k^{1/4}$, *every pair of samples from the same cluster are closer to each other in Euclidean distance than are every pair of samples from distinct clusters (with high probability)*, while this is no longer true if $\Delta < k^{1/4}$. Thus, when $\Delta \geqslant k^{1/4}$, a simple greedy algorithm correctly clusters the samples into their components (this algorithm is sometimes called *single-linkage clustering*). On the other hand, standard information-theoretic arguments show that the means remain approximately identifiable from $\text{poly}(k, d)$ samples when Δ is as small as $O(\sqrt{\log k})$, but these methods yield only exponential-time algorithms.[1] Nonetheless, despite substantial attention, this

[1]Recent and sophisticated arguments show that the means are identifiable (albeit inefficiently) with error depending only on the number of samples and not on the separation Δ even when

$\Delta = k^{1/4}$ barrier representing the breakdown of single-linkage clustering has stood for nearly 20 years.

We prove the following main theorem, breaking the single-linkage clustering barrier.

Theorem 9.1.1 (Informal, special case for uniform mixture of spherical Gaussians)**.** *For every $\gamma > 0$ there is an algorithm with running time $(dk)^{O(1/\gamma^2)}$ using at most $n \leqslant k^{O(1)} d^{O(1/\gamma)}$ samples which, given samples $x_1, \ldots, x_n$ from a uniform mixture of k spherical Gaussians $\mathcal{N}(\mu_i, \mathrm{Id})$ in d dimensions with means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ satisfying $\|\mu_i - \mu_j\| \geqslant k^\gamma$ for each $i \neq j$, returns estimators $\widehat{\mu}_1, \ldots, \widehat{\mu}_k \in \mathbb{R}^d$ such that $\|\widehat{\mu}_i - \mu_i\| \leqslant 1/\mathrm{poly}(k)$ (with high probability).*

We pause here to make several remarks about this theorem. Our algorithm makes novel use of higher order moments of Gaussian (and sub-Gaussian) distributions. Most previous work for efficiently learning well-separated mixtures either used only second-order moment information, and required separation $\Delta \geqslant \Omega(\sqrt{k})$, or made mild use of log-concavity to improve this to $k^{1/4}$, whereas we use $O(1/\gamma)$ moments.

The guarantees of our theorem hold well beyond the Gaussian setting; the theorem applies to any mixture model with k^γ separation and whose component distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$ are what we term $O(1/\gamma)$-*explicitly bounded*. We define this notion formally below, but roughly speaking, a t-explicitly bounded distribution $\mathcal{D}$ has t-th moments obeying a subgaussian-type bound—that is, for every unit vector $u \in \mathbb{R}^d$ one has $\mathbb{E}_{Y \sim \mathcal{D}} |\langle Y, u \rangle|^t \leqslant t^{t/2}$—and there is a certain kind of *simple certificate* of this fact, namely a low-degree Sum of Squares proof. Among

$\Delta = O(\sqrt{\log k})$ [158].

other things, this means the theorem also applies to mixtures of symmetric product distributions with bounded moments.

For mixtures of distributions with sufficiently-many bounded moments (such as Gaussians), our guarantees go even further. We show that using $d^{O(\log k)^2}$ time and $d^{O(\log k)}$ samples, we can recover the means to error $1/\mathrm{poly}(k)$ even if the separation is only $C\sqrt{\log k}$ for some universal constant C. Strikingly, [158] show that any algorithm that can learn the means nontrivially given separation $o(\sqrt{\log k})$ must require super-polynomial samples and time. Our results show that just above this threshold, it is possible to learn with just quasipolynomially many samples and time.

Finally, throughout this chapter we state error guarantees roughly in terms of obtaining $\widehat{\mu}_i$ with $\|\widehat{\mu}_i - \mu_i\| \leqslant 1/\mathrm{poly}(k) \ll k^{\gamma}$, meaning that we get ℓ_2 error which is much less than the true separation. In the special case of spherical Gaussians, we note that we can use our algorithm as a warm-start to recent algorithms due to [158], and achieve error δ using $\mathrm{poly}(m, k, 1/\delta)$ additional runtime and samples for some polynomial independent of γ.

Robust mean estimation Estimators which are robust to outlying or corrupted samples have been studied in statistics at least since the 1960s [95, 171]. The model we consider in this paper is a slight generalization of Hüber's contamination model [95]. We are given $X_1, \ldots, X_n$, originally drawn iid from some unknown distribution $\mathcal{D}$, but an adversary has changed an ε fraction of these points adversarially. We call such a set of points ε-corrupted.[2] The goal of robust statistics is to recover statistics of $\mathcal{D}$ such as mean and covariance, given ε-

[2]Hüber's contamination model essentially only allows the adversary to add corrupted points, but not remove uncorrupted points.

corrupted samples from $\mathcal{D}$.

In classical robust statistics, the robust mean estimation problem is known as *robust estimation of location*, and robust covariance estimation is known as *robust estimation of scale*. Classical works consider a measure known as breakdown point, which is (informally) the fraction of samples that an adversary must corrupt before the estimator has no provable guarantees. They often design robust estimators for mean and covariance that achieve optimal error in many fundamental settings. For instance, given samples from a symmetric sub-Gaussian distribution in k dimensions such that an ε-fraction are arbitrarily corrupted, an estimator known as the Tukey median [171] achieves error $O(\varepsilon)$, which is information theoretically optimal. However, these estimators are all NP-hard to compute [99, 36] and the best known algorithms require $\exp(d)$ time in general.

For a long time, all known computationally efficient robust statistics for the mean or covariance of a d-dimensional Gaussian had error degrading polynomially with the dimension.[3] In recent work, [64, 114] gave efficient and robust estimators for these statistics which achieve substantially better error. In particular, [64] achieve error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ for estimating the mean of a Gaussian with identity covariance, and error $O(\varepsilon \log^{3/2} 1/\varepsilon)$ for robustly estimating the mean of a Gaussian with unknown variance $\Sigma \preceq I$.

Unfortunately, these results are somewhat tailored to Gaussian distributions, or require covariance very close to identity. For general sub-Gaussian distributions with unknown variance $\Sigma \preceq I$, the best known efficient algorithms achieve only $O(\varepsilon^{1/2})$ error [65, 166]. We substantially improve this, under a slightly stronger condition than sub-Gaussianity. Recall that a distribution $\mathcal{D}$ with mean μ over

[3] We remark that this was the state of affairs even for the Hüber contamination model.

$\mathbb{R}^d$ is sub-Gaussian if for every unit vector u and every $t \in \mathbb{N}$ even, the following moment bound holds:

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}} \langle u, X - \mu\rangle^t \leqslant t^{t/2} \,.$$

Informally stated, our algorithms will work under the condition that this moment bound can be certified by a low degree SoS proof, for all $s \leqslant t$. We call such distributions *t-explicitly bounded* (we are ignoring some parameters, see Definition 9.3.1 for a formal definition). This class captures many natural sub-Gaussian distributions, such as Gaussians, product distributions of sub-Gaussians, and rotations thereof (see Appendix 9.7.5). For such distributions, we show:

Theorem 9.1.2 (informal, see Theorem 9.6.1)**.** *Fix $\varepsilon > 0$ sufficiently small and let $t \geqslant 4$. Let $\mathcal{D}$ be a $O(t)$-explicitly bounded distribution over $\mathbb{R}^d$ with mean μ^*. There is an algorithm with sample complexity $d^{O(t)}(1/\varepsilon)^{O(1)}$ running time $(d^t\varepsilon)^{O(t)}$ such that given an ε-corrupted set of samples of sufficiently large size from $\mathcal{D}$, outputs μ so that with high probability $\|\mu - \mu^*\| \leqslant O(\varepsilon^{1-1/t})$.*

As with mixture models, we can push our statistical rates further, if we are willing to tolerate quasipolynomial runtime and sample complexity. In particular, we can obtain error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ with $d^{O(\log 1/\varepsilon)}$ samples and $d^{O(\log 1/\varepsilon)^2}$ time.

9.1.1 Organization

In Section 9.2 we discuss at a high level the ideas in our algorithms and SoS proofs. In Section 9.3 we give standard background on SoS proofs. Section 9.4 discusses the important properties of the family of polynomial inequalities we use in both algorithms. Section 9.5 and Section 9.6 state our algorithms formally

and analyze them. Finally, Section 9.7 describes the polynomial inequalities our algorithms employ in more detail.

9.2 Algorithm and Proof Overview

In this section we give a high-level overview of the main ideas in our algorithms. We use the proofs-to-algorithms method described in Section 3.5. The key step in desiging our algorithm is to invent proofs of identifiability for the mixture model and robust statistics settings which can be captured by low degree SoS.

For this overview, we discuss the main idea in our identifiability proofs informally, using only simple inequalities, like Cauchy-Schwarz and Hölder's inequalities. Later we formally show that the proofs are captured in low degree SoS.

We consider an idealized version of situations we encounter in both the mixture model and robust estimation settings. Let $\mu^* \in \mathbb{R}^d$. Let $X_1, \dots, X_n \in \mathbb{R}^d$ have the guarantee that for some $T \subseteq [n]$ of size $|T| = \alpha n$, the vectors $\{X_i\}_{i \in T}$ are iid samples from $\mathcal{N}(\mu^*, \mathrm{Id})$, a spherical Gaussian centered at μ^*; for the other vectors we make no assumption. The goal is to estimate the mean μ^*.

Our main idea is to leverage the fact that, as long as $\alpha n \gg d^t$, by standard concentration the empirical moments of T are sub-Gaussian, up to order t. We show that if S is any other set of αn samples with sub-Gaussian t-th moments, then if S shares even a small number of samples with T, the empirical mean of S must be close to the empirical mean of T.

Formally, let $S \subseteq [n]$ and let $\mu = \frac{1}{|S|} \sum_{i \in S} X_i$ be the empirical mean of S. For

$t \in \mathbb{N}$, consider the following crucial moment inequality, our criterion for a *good* subset S:

$$\frac{1}{|S|}\sum_{i\in S}\langle X_i - \mu, u\rangle^t \leqslant 2\cdot t^{t/2}\cdot \|u\|^t \qquad \text{for all } u \in \mathbb{R}^d\,. \tag{9.2.1}$$

This inequality says that every one-dimensional projection u of the samples in S, centered around their empirical mean, has a sub-Gaussian empirical t-th moment. (The factor 2 accounts for deviations in the t-th moments of the samples.) By standard concentration of measure, if $\alpha n \gg d^t$ the inequality holds for $S = T$. It turns out that this property can be enforced by polynomials of degree t.

We would like to show that any S which satisfies (9.2.1) has empirical mean close to μ^* using a low-degree SoS proof,. This is in fact true when $\alpha = 1 - \varepsilon$ for small ε, which is at the core of our robust estimation algorithm. However, in the mixture model setting, when $\alpha = 1/(\#\text{ of components})$, for each component j there is a subset $T_j \subseteq [n]$ of samples from component j which provides a valid solution $S = T_j$ to $\mathcal{A}$. The empirical mean of T_j is close to μ_j and hence not close to μ_i for any $i \neq j$.

We will prove something slightly weaker, which still demonstrates the main idea in our identifiability proof.

Lemma 9.2.1. *With high probability, for every $S \subseteq [n]$ which satisfies (9.2.1), if $\mu = \frac{1}{|S|}\sum_{i\in S} X_i$ is the empirical mean of samples in S, then $\|\mu - \mu^*\| \leqslant 4t^{1/2}\cdot(|T|/|S\cap T|)^{1/t}$.*

Notice that a random $S \subseteq [n]$ of size αn will have $|S \cap T| \approx \alpha^2 n$. In this case the lemma would yield the bound $\|\mu - \mu^*\| \leqslant \frac{4t^{1/2}}{\alpha^{1/t}}$. Thinking of $\alpha \ll 1/t$, this bound improves exponentially as t grows. In the d-dimensional k-component mixture model setting, one has $1/\alpha = \text{poly}(k)$, and thus the bound becomes $\|\mu - \mu^*\| \leqslant 4t^{1/2}\cdot k^{O(1/t)}$. In a mixture model where components are separated by

k^ε, such an estimate is nontrivial when $\|\mu - \mu^*\| \ll k^\varepsilon$, which requires $t = O(1/\varepsilon)$. This is the origin of the quantitative bounds in our mixture model algorithm.

We turn to the proof of Lemma 9.2.1. As we have already emphasized, the crucial point is that this proof will be accomplished using only simple inequalities, avoiding any union bound over all possible subsets S.

Proof of Lemma 9.2.1. Let w_i be the 0/1 indicator of $i \in S$. To start the argument, we expand in terms of samples:

$$\begin{aligned} |S \cap T| \cdot \|\mu - \mu^*\|^2 &= \sum_{i \in T} w_i \|\mu - \mu^*\|^2 \\ &= \sum_{i \in T} w_i \langle \mu^* - \mu, \mu^* - \mu \rangle \qquad (9.2.2) \\ &= \sum_{i \in T} w_i \left[\langle X_i - \mu, \mu^* - \mu \rangle + \langle \mu^* - X_i, \mu^* - \mu \rangle \right] . \qquad (9.2.3) \end{aligned}$$

The key term to bound is the first one; the second amounts to a deviation term. By Hölder's inequality and for even t,

$$\begin{aligned} \sum_{i \in T} w_i \langle X_i - \mu, \mu^* - \mu \rangle &\leqslant \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot \left(\sum_{i \in T} w_i \langle X_i - \mu, \mu^* - \mu \rangle^t \right)^{1/t} \\ &\leqslant \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot \left(\sum_{i \in [n]} w_i \langle X_i - \mu, \mu^* - \mu \rangle^t \right)^{1/t} \\ &\leqslant \left(\sum_{i \in T} w_i \right)^{\frac{t-1}{t}} \cdot 2t^{1/2} \cdot \|\mu^* - \mu\| \\ &= |S \cap T|^{\frac{t-1}{t}} \cdot 2t^{1/2} \cdot \|\mu^* - \mu\| . \end{aligned}$$

The second line follows by adding the samples from $[n] \setminus T$ to the sum; since t is even this only increases its value. The third line uses the moment inequality (9.2.1). The last line just uses the definition of w.

For the second, deviation term, we use Hölder's inequality again:

$$\sum_{i\in T} w_i\langle \mu^* - X_i, \mu^* - \mu\rangle \leqslant \left(\sum_{i\in T} w_i\right)^{\frac{t-1}{t}} \cdot \left(\sum_{i\in T}\langle \mu^* - X_i, \mu^* - \mu\rangle^t\right)^{1/t}.$$

The distribution of $\mu^* - X_i$ for $i \in T$ is $\mathcal{N}(0, \mathrm{Id})$. By standard matrix concentration, if $|T| = \alpha n \gg d^t$,

$$\sum_{i\in T}\left[(X_i-\mu^*)^{\otimes t/2}\right]\left[(X_i-\mu^*)^{\otimes t/2}\right]^\top \preceq 2|T| \mathop{\mathbb{E}}_{Y\sim\mathcal{N}(0,\mathrm{Id})}\left(Y^{\otimes t/2}\right)\left(Y^{\otimes t/2}\right)^\top$$

with high probability and hence, using the quadratic form at $(\mu^* - \mu)^{\otimes t/2}$,

$$\sum_{i\in T}\langle \mu^* - X_i, \mu^* - \mu\rangle^t \leqslant 2|T| t^{t/2}\cdot\|\mu^* - \mu\|^t.$$

Putting these together and simplifying constants, we have obtained that with high probability,

$$|S\cap T|\cdot\|\mu-\mu^*\|^2 \leqslant 4t^{1/2}|T|^{1/t}\cdot|S\cap T|^{(t-1)/t}\cdot\|\mu-\mu^*\|$$

which simplifies to

$$|S\cap T|^{1/t}\cdot\|\mu-\mu^*\| \leqslant 4t^{1/2}|T|^{1/t}. \qquad \square$$

Once we formalize our identifiability proofs in SoS, the rest of the algorithm design is standard. The final algorithm solves an SoS SDP and applies stanard rounding techniques.

9.3 Problem Statements

Throughout the chapter we let d be the dimensionality of the data, and we will be interested in the regime where d is at least a large constant. We also let $\|v\|$

denote the ℓ_2 norm of a vector v, and $\|M\|_F$ to denote the Frobenius norm of a matrix M; often we just write $\|M\|$. We will also give randomized algorithms for our problems that succeed with probability $1 - \text{poly}(1/k, 1/d)$; by standard techniques this probability can be boosted to $1 - \xi$ by increasing the sample and runtime complexity by a mulitplicative $\log 1/\xi$.

We now formally define the class of distributions we will consider throughout this chapter. At a high level, we will consider distributions which have bounded moments, for which there exists a low degree SoS proof of this moment bound. Formally:

Definition 9.3.1. Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ with mean μ. For $c \geqslant 1, t \in \mathbb{N}$, we say that $\mathcal{D}$ is t-explicitly bounded with variance proxy σ if for every even $s \leqslant t$ there is a degree s SoS proof of

$$\vdash_s \; E_{Y \sim \mathcal{D}_k} \langle \left(Y - \mu\right), u \rangle^s \leqslant (\sigma s)^{s/2} \|u\|^s \, .$$

Equivalently, the polynomial $p(u) = (\sigma s)^{s/2}\|u\|^s - E_{Y \sim \mathcal{D}_k} \langle \left(Y - \mu\right), u \rangle^s$ should be a sum-of-squares. In our typical use case, $\sigma = 1$, we will omit it and call the distribution t-explicitly bounded.

Throughout this paper, since all of our problems are scale invariant, we will assume without loss of generality that $\sigma = 1$. This class of distributions captures a number of natural classes of distributions. Intuitively, if u were truly a vector in $\mathbb{R}^k$ (rather than a vector of indeterminants), then this exactly captures sub-Gaussian type moment. Our requirement is simply that these types of moment bounds not only hold, but also have a SoS proof.

We remark that our results also hold for somewhat more general settings. It is not particularly important that the s-th moment bound has a degree s proof;

our techniques can tolerate degree $O(s)$ proofs. Our techniques also generally apply for weaker moment bounds. For instance, our techniques naturally extend to explicitly bounded sub-exponential type distributions in the obvious way. We omit these details for simplicity.

As we show in Appendix 9.7.5, this class still captures many interesting types of nice distributions, including Gaussians, product distributions with sub-Gaussian components, and rotations therof. With this definition in mind, we can now formally state the problems we consider in this chapter:

Learning well-separated mixture models We first define the class of mixture models for which our algorithm works:

Definition 9.3.2 (t-explicitly bounded mixture model with separation Δ)**.** Let $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ satisfy $\|\mu_i - \mu_j\| > \Delta$ for every $i \neq j$, and let $\mathcal{D}_1, \ldots, \mathcal{D}_k$ have means $\mu_1, \ldots, \mu_k$, so that each $\mathcal{D}_i$ is t-explicitly bounded. Let $\lambda_1, \ldots, \lambda_k \geqslant 0$ satisfy $\sum_{i \in [k]} \lambda_i = 1$. Together these define a mixture distribution on $\mathbb{R}^d$ by first sampling $i \sim \lambda$, then sampling $x \sim \mathcal{D}_i$.

The problem is then:

Problem 9.3.3. Let $\mathcal{D}$ be a t-explicitly bounded mixture model in $\mathbb{R}^d$ with separation Δ with k components. Given k, Δ, and n independent samples from $\mathcal{D}$, output $\widehat{\mu}_1, \ldots, \widehat{\mu}_m$ so that with probability at least 0.99, there exists a permutation $\pi : [k] \to [k]$ so that $\|\mu_i - \widehat{\mu}_{\pi(i)}\| \leqslant \delta$ for all $i = 1, \ldots, k$.

Robust mean estimation We consider the same basic model of corruption introduced in [64].

Definition 9.3.4 (ε-corruption)**.** We say a set of samples $X_1, \ldots, X_n$ is ε-corrupted from a distribution $\mathcal{D}$ if they are generated via the following process. First, n independent samples are drawn from $\mathcal{D}$. Then, an adversary changes εn of these points arbitrarily, and the altered set of points is then returned to us in an arbitrary order.

The problem we consider in this setting is the following:

Problem 9.3.5 (Robust mean estimation)**.** Let $\mathcal{D}$ be an $O(t)$-explicitly bounded distribution over $\mathbb{R}^d$ wih mean μ. Given t, ε, and an ε-corrupted set of samples from $\mathcal{D}$, output $\widehat{\mu}$ satisfying $\|\mu - \widehat{\mu}\| \leqslant O(\varepsilon^{1-1/t})$.

Gaussian distributions are explicitly bounded In Section 9.7.5 we show that product distributions (and rotations thereof) with bounded t-th moments are explicitly bounded.

Lemma 9.3.6. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ so that $\mathcal{D}$ is a rotation of a product distribution $\mathcal{D}'$ where each coordinate X with mean μ of $\mathcal{D}$ satisfies*

$$\mathbb{E}[(X-\mu)^s] \leqslant 2^{-s}\left(\frac{s}{2}\right)^{s/2}$$

Then $\mathcal{D}$ is t-explicitly bounded (with variance proxy 1).

(The factors of $\frac{1}{2}$ can be removed for many distributions, including Gaussians.)

9.4 Capturing Empirical Moments with Polynomials

To describe our algorithms we need to describe a system of polynomial equations and inequalities which capture the following problem: among $X_1, \ldots, X_n \in \mathbb{R}^d$,

find a subset of $S \subseteq [n]$ of size αn such that the empirical t-th moments obey a moment bound: $\frac{1}{\alpha n} \sum_{i \in S} \langle X_i, u \rangle^t \leqslant t^{t/2} \|u\|^t$ for every $u \in \mathbb{R}^d$.

Let $k, n \in \mathbb{N}$ and let $w = (w_1, \ldots, w_n)$, $\mu = (\mu_1, \ldots, \mu_k)$ be indeterminates. Let

1. $X_1, \ldots, X_n \in \mathbb{R}^d$
2. $\alpha \in [0, 1]$ be a number (the intention is $|S| = \alpha n$).
3. $t \in \mathbb{N}$ be a power of 2, the order of moments to control
4. $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, which will eventually be the means of a k-component mixture model, or when $k = 1$, the true mean of the distribution whose mean we robustly estimate.
5. $\tau > 0$ be some error magnitude accounting for fluctuations in the sizes of clusters (which may be safely ignored at first reading).

Definition 9.4.1. Let $\mathcal{A}$ be the following system of equations and inequalities, depending on all the parameters above.

1. $w_i^2 = w_i$ for all $i \in [n]$ (enforcing that w is a 0/1 vector, which we interpret as the indicator vector of the set S).
2. $(1 - \tau)\alpha n \leqslant \sum_{i \in [n]} w_i \leqslant (1 + \tau)\alpha n$, enforcing that $|S| \approx \alpha n$ (we will always choose $\tau = o(1)$).
3. $\mu \cdot \sum_{i \in [n]} w_i = \sum_{i \in [n]} w_i X_i$, enforcing that μ is the empirical mean of the samples in S
4. $\sum_{i \in [n]} w_i \langle X_i - \mu, \mu - \mu_j \rangle^t \leqslant 2 \cdot t^{t/2} \sum_{i \in [n]} w_i \|\mu - \mu_j\|^t$ for every μ_j among $\mu_1, \ldots, \mu_m$. This enforces that the t-th empirical moment of the samples in S is bounded *in the direction* $\mu - \mu_j$.

Notice that since we will eventually take μ_j's to be unknown parameters we are trying to estimate, the algorithm cannot make use of $\mathcal{A}$ directly, since the last family of inequalities involve the μ_j's. Later in this paper we exhibit a system of inequalities which requires the empirical t-th moments to obey a sub-Gaussian type bound in every direction, hence implying the inequalities here without requiring knowledge of the μ_j's to write down. Formally, we will show:

Lemma 9.4.2. *Let $\alpha \in [0,1]$. Let $t \in \mathbb{N}$ be a power of 2, $t \geqslant 4$.*[4] *Let $0.1 > \tau > 0$. Let $X_1, \ldots, X_n \in \mathbb{R}^d$. Let $\mathcal{D}$ be a $10t$-explicitly bounded distribution.*

There is a family $\widehat{\mathcal{A}}$ of polynomial equations and inequalities of degree $O(t)$ on variables $w = (w_1, \ldots, w_n)$, $\mu = (\mu_1, \ldots, \mu_k)$ and at most $n^{O(t)}$ other variables, whose coefficients depend on $\alpha, t, \tau, X_1, \ldots, X_n$, such that

1. *(Satisfiability) If there $S \subseteq [n]$ of size at least $(\alpha - \tau)n$ so that $\{X_i\}_{i \in S}$ is an iid set of samples from $\mathcal{D}$, and $(1-\tau)\alpha n \geqslant d^{100t}$, then for d large enough, with probability at least $1 - d^{-8}$, the system $\widehat{\mathcal{A}}$ has a solution over $\mathbb{R}$ which takes w to be the $0/1$ indicator vector of S.*
2. *(Solvability) For every $C \in \mathbb{N}$ there is an $n^{O(Ct)}$-time algorithm which, when $\widehat{\mathcal{A}}$ is satisfiable, returns a degree-Ct pseudodistribution which satisfies $\widehat{\mathcal{A}}$ (up to additive error 2^{-n}).*
3. *(Moment bounds for polynomials of μ) Let $f(\mu)$ be a length-d vector of degree-ℓ polynomials in indeterminates $\mu = (\mu_1, \ldots, \mu_k)$. $\widehat{\mathcal{A}}$ implies the following inequality and the implication has a degree $t\ell$ SoS proof.*

$$\widehat{\mathcal{A}} \vdash_{O(t\ell)} \frac{1}{\alpha n} \sum_{i \in [n]} w_i \langle X_i - \mu, f(\mu) \rangle^t \leqslant 2 \cdot t^{t/2} \|f(\mu)\|^t .$$

[4] The condition $t \geqslant 4$ is merely for technical convenience.

4. *(Booleanness)* $\widehat{\mathcal{A}}$ *includes the equations* $w_i^2 = w_i$ *for all* $i \in [n]$.

5. *(Size)* $\widehat{\mathcal{A}}$ *includes the inequalities* $(1-\tau)\alpha n \leqslant \sum w_i \leqslant (1+\tau)\alpha n$.

6. *(Empirical mean)* $\widehat{\mathcal{A}}$ *includes the equation* $\mu \cdot \sum_{i\in[n]} w_i = \sum_{i\in[n]} w_i X_i$.

In particular this implies that $\widehat{\mathcal{A}} \vdash_{O(t)} \mathcal{A}$.

The proof of Lemma 9.4.2 can be found in Section 9.7.

Remark 9.4.3 (Numerical accuracy, semidefinite programming, and other monsters). We pause here to address issues of numerical accuracy. Our final algorithms use point 2 in Lemma 9.4.2 (itself implemented using semidefinite programming) to obtain a pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\widehat{\mathcal{A}}$ approximately, up to error $\eta = 2^{-n}$ in the following sense: for every r a sum of squares and $f_1, \ldots, f_\ell \in \mathcal{A}$ with $\deg\left[r \cdot \prod f_i \leqslant Ct\right]$, one has $\tilde{\mathbb{E}}\, r \cdot \prod_{i\in\mathcal{A}} f \geqslant -\eta \cdot \|r\|$, where $\|r\|$ is ℓ_2 norm of the coefficients of r. Our main analyses of this pseudodistribution employ the implication $\widehat{\mathcal{A}} \vdash \mathcal{B}$ for another family of inequalities $\mathcal{B}$ to conclude that if $\tilde{\mathbb{E}}$ satisfies $\mathcal{A}$ then it satisfies $\mathcal{B}$, then use the latter to analyze our rounding algorithms. Because all of the polynomials eventually involved in the SoS proof $\widehat{\mathcal{A}} \vdash \mathcal{B}$ have coefficients bounded by n^B for some large constant B, it may be inferred that if $\tilde{\mathbb{E}}$ approximately satisfies $\widehat{\mathcal{A}}$ in the sense above, it also approximately satisfies $\mathcal{B}$, with some error $\eta' \leqslant 2^{-\Omega(n)}$. The latter is a sufficient for all of our rounding algorithms.

Aside from mentioning at a couple key points why our SoS proofs have bounded coefficients, we henceforth ignore all numerical issues. For further discussion of numerical accuracy and well-conditioned-ness issues in SoS, see [143, 34, 156].

9.5 Mixture Models: Algorithm and Analysis

In this section we formally describe and analyze our algorithm for mixture models. We prove the following theorem.

Theorem 9.5.1 (Main theorem on mixture models)**.** *For every large-enough $t \in \mathbb{N}$ there is an algorithm with the following guarantees. Let $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, satisfy $\|\mu_i - \mu_j\| \geqslant \Delta$. Let $\mathcal{D}_1, \ldots, \mathcal{D}_k$ be $10t$-explicitly bounded, with means $\mu_1, \ldots, \mu_k$. Let $\lambda_1, \ldots, \lambda_k \geqslant 0$ satisfy $\sum \lambda_i = 1$. Given $n \geqslant (d^t k)^{O(1)} \cdot (\max_{i \in [m]} 1/\lambda_i)^{O(1)}$ samples from the mixture model given by $\lambda_1, \ldots, \lambda_k, \mathcal{D}_1, \ldots, \mathcal{D}_k$, the algorithm runs in time $n^{O(t)}$ and with high probability returns $\{\widehat{\mu}_1, \ldots, \widehat{\mu}_k\}$ (not necessarily in that order) such that*

$$\|\mu_i - \widehat{\mu}_i\| \leqslant \frac{2^{Ct} k^C t^{t/2}}{\Delta^{t-1}}$$

for some universal constant C.

In particular, we note two regimes: if $\Delta = k^\gamma$ for a constant $\gamma > 0$, choosing $t = O(1/\gamma)$ we get that the ℓ_2 error of our estimator is $\mathrm{poly}(1/k)$ for any $O(1/\gamma)$-explicitly bounded distribution, and our estimator requires only $(dk)^{O(1)}$ samples and time. This matches the guarantees of Theorem 9.1.1.

On the other hand, if $\Delta = C'\sqrt{\log k}$ (for some universal C') then taking $t = O(\log k)$ gives error

$$\|\mu_i - \widehat{\mu}_i\| \leqslant k^{O(1)} \cdot \left(\frac{\sqrt{t}}{\Delta}\right)^t$$

which, for large-enough C' and t, can be made $1/\mathrm{poly}(k)$. Thus for $\Delta = C'\sqrt{\log k}$ and any $O(\log k)$-explicitly bounded distribuion we obtain error $1/\mathrm{poly}(k)$ with $d^{O(\log k)}$ samples and $d^{O(\log k)^2}$ time.

In this section we describe and analyze our algorithm. To avoid some technical work we analyze the uniform mixtures setting, with $\lambda_i = 1/m$. For the adaptation to the nonuniform mixture setting, see [90].

9.5.1 Algorithm and main analysis

We formally describe our mixture model algorithm now. We use the following lemma, which we prove in Section 9.5.6. The lemma says that given a matrix which is very close, in Frobenious norm, to the 0/1 indicator matrix of a partition of $[n]$ it is possible to approximately recover the partition. (The proof is standard.)

Lemma 9.5.2 (Second moment rounding, follows from Theorem 9.5.15)**.** *Let $n, m \in \mathbb{N}$ with $m \ll n$. There is a polynomial time algorithm* RoundSecondMoments *with the following guarantees. Suppose $S_1, \ldots, S_m$ partition $[n]$ into m pieces, each of size $\frac{n}{2m} \leqslant |S_i| \leqslant \frac{2n}{m}$. Let $A \in \mathbb{R}^{n \times n}$ be the 0/1 indicator matrix for the partition S; that is, $A_{ij} = 1$ if $i, j \in S_\ell$ for some ℓ and is 0 otherwise. Let $M \in \mathbb{R}^{n \times n}$ be a matrix with $\|A - M\|_F \leqslant \varepsilon n$. Given M, with probability at least $1 - \varepsilon^2 m^3$ the algorithm returns a partition $C_1, \ldots, C_m$ of $[n]$ such that up to a global permutation of $[m]$, $C_i = T_i \cup B_i$, where $T_i \subseteq S_i$ and $|T_i| \geqslant |S_i| - \varepsilon^2 m^2 n$ and $|B_i| \leqslant \varepsilon^2 m^2 n$.*

Algorithm 9.5.3 (Mixture Model Learning)**.**

1: **function** LearnMixtureMeans($t, X_1, \ldots, X_n, \delta, \tau$)

2: By semidefinite programming (see Lemma 9.4.2, item 2), find a pseudoexpectation of degree $O(t)$ which satisfies the structured subset polynomials from Lemma 9.4.2, with $\alpha = n/m$ such that $\| \tilde{\mathbb{E}}\, ww^\top \|_F$ is minimized among all such pseudoexpectations.

3: Let $M \leftarrow m \cdot \tilde{\mathbb{E}}\, ww^\top$.

4: Run the algorithm RoundSecondMoments on M to obtain a partition

$C_1, \dots, C_m$ of $[n]$.

5: Run the algorithm EstimateMean from Section 9.6 on each cluster C_i, with $\varepsilon = 2^{Ct} t^{t/2} m^4 / \Delta^t$ for some universal constant C to obtain a list of mean estimates $\widehat{\mu}_1, \dots, \widehat{\mu}_m$.

6: Output $\widehat{\mu}_1, \dots, \widehat{\mu}_m$.

7: **end function**

Remark 9.5.4 (On the use of EstimateMean). As described, LearnMixtureMeans has two phases: a clustering phase and a mean-estimation phase. The clustering phase is the heart of the algorithm; we will show that after running RoundSecondMoments the algorithm has obtained clusters $C_1, \dots, C_k$ which err from the ground-truth clustering on only a $\frac{2^{O(t)} t^{t/2} \operatorname{poly}(k)}{\Delta^t}$-fraction of points. To obtain estimates $\widehat{\mu}_i$ of the underlying means from such a clustering, one simple option is to output the empirical mean of the clusters. However, without additional pruning this risks introducing error in the mean estimates which grows with the ambient dimension d. By using the robust mean estimation algorithm instead to obtain mean estimates from the clusters we obtain errors in the mean estimates which depend only on the number of clusters k, the between-cluster separation Δ, and the number t of bounded moments.

Remark 9.5.5 (Running time). We observe that LearnMixtureMeans can be implemented in time $n^{O(t)}$. The main theorem requires $n \geqslant k^{O(1)} d^{O(t)}$, which means that the final running time of the algorithm is $(kd^t)^{O(t)}$.[5]

[5] As discussed in Section 9.4, correctness of our algorithm at the level of numerical accuracy requires that the coefficients of every polynomial in the SoS program $\widehat{\mathcal{A}}$ (and every polynomial in the SoS proofs we use to analyze $\widehat{\mathcal{A}}$) are polynomially bounded. This may not be the case if some vectors $\mu_1, \dots, \mu_m$ have norms $\|\mu_i\| \geqslant d^{\omega(1)}$. This can be fixed by naively clustering the samples $X_1, \dots, X_n$ via single-linkage clustering, then running LearnMixtureMeans on each cluster. It is routine to show that the diameter of each cluster output by a naive clustering algorithm is at most $\operatorname{poly}(d, k)$ under our assumptions, and that with high probability single-linkage clustering produces a clustering respecting the distributions $\mathcal{D}_i$. Hence, by centering each cluster before running LearnMixtureMeans we can assume that $\|\mu_i\| \leqslant \operatorname{poly}(d, k)$ for every $i \leqslant d$.

9.5.2 Proof of main theorem

In this section we prove our main theorem using the key lemmata; in the following sections we prove the lemmata.

Deterministic Conditions We recall the setup. There are k mean vectors $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$, and corresponding distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$ where $\mathcal{D}_j$ has mean μ_j. The distributions $\mathcal{D}_j$ are $10t$-explicitly bounded for a choice of t which is a power of 2. Vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ are samples from a uniform mixture of $\mathcal{D}_1, \ldots, \mathcal{D}_k$. We will prove that our algorithm succeeds under the following condition on the samples $X_1, \ldots, X_n$.

(D1) (Empirical moments) For every cluster $S_j = \{X_i : X_i \text{ is from } \mathcal{D}_j\}$, the system $\widehat{\mathcal{A}}$ from Lemma 9.4.2 with $\alpha = 1/m$ and $\tau = \Delta^{-t}$ has a solution which takes $w \in \{0,1\}^n$ to be the 0/1 indicator vector of S_j.

(D2) (Empirical means) Let $\overline{\mu}_j$ be the empirical mean of cluster S_j. The $\overline{\mu}_j$'s satisfy $\|\overline{\mu}_i - \mu_i\| \leqslant \Delta^{-t}$.

We note a few useful consequences of these conditions, especially (D1). First of all, it implies all clusters have almost the same size: $(1-\Delta^{-t}) \cdot \frac{n}{k} \leqslant |S_j| \leqslant (1+\Delta^{-t}) \cdot \frac{n}{k}$. Second, it implies that all clusters have explicitly bounded moments: for every S_j,

$$\vdash_t \frac{k}{n} \sum_{i \in S_j} \langle X_i - \overline{\mu}_j, u \rangle^t \leqslant 2 \cdot t^{t/2} \cdot \|u\|^t .$$

Lemmas The following key lemma captures our SoS identifiability proof for mixture models.

Lemma 9.5.6. *Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k$ be as in Theorem 9.5.1, with mean separation Δ. Suppose (D1), (D2) occur for samples $X_1, \dots, X_n$. Let $t \in \mathbb{N}$ be a power of two. Let $\tilde{\mathbb{E}}$ be a degree-$O(t)$ pseudoexpectation which satisfies $\mathcal{A}$ from Lemma 9.4.2 with $\alpha = 1/k$ and $\tau \leqslant \Delta^{-t}$. Then for every $j, \ell \in [k]$,*

$$\tilde{\mathbb{E}}\langle a_j, w\rangle\langle a_\ell, w\rangle \leqslant 2^{8t+8} \cdot t^{t/2} \cdot \frac{n^2}{k} \cdot \frac{1}{\Delta^t}\,.$$

The other main lemma shows that conditions (D1) and (D2) occur with high probability.

Lemma 9.5.7 (Concentration for mixture models)**.** *With notation as above, conditions (D1) and (D2) simultaneously occur with probability at least $1 - 1/d^{15}$ over samples $X_1, \dots, X_n$, so long as $n \geqslant d^{O(t)}k^{O(1)}$, for $\Delta \geqslant 1$.*

Lemma 9.5.7 follows from Lemma 9.4.2, for (D1), and standard concentration arguments for (D2). Now we can prove the main theorem.

Proof of Theorem 9.5.1 (uniform mixtures case). Suppose conditions (D1) and (D2) hold. Our goal will be to bound $\|M - A\|^2 \leqslant n \cdot \frac{2^{O(t)}t^{t/2}k^4}{\Delta^t}$, where A is the 0/1 indicator matrix for the ground truth partition $S_1, \dots, S_k$ of $X_1, \dots, X_n$ according to $\mathcal{D}_1, \dots, \mathcal{D}_k$. Then by Lemma 9.5.2, the rounding algorithm will return a partition $C_1, \dots, C_k$ of $[n]$ such that C_ℓ and S_ℓ differ by at most $n\frac{2^{O(t)}t^{t/2}k^{10}}{\Delta^t}$ points, with probability at least $1 - \frac{2^{O(t)}t^{t/2}k^{30}}{\Delta^t}$. By the guarantees of Theorem 9.6.1 regarding the algorithm EstimateMean, with high probability the resulting error in the mean estimates $\widehat{\mu}_i$ will satisfy

$$\|\mu_i - \widehat{\mu}_i\| \leqslant \sqrt{t} \cdot \left(\frac{2^{O(t)}t^{t/2}k^{10}}{\Delta^t}\right)^{\frac{t-1}{t}} \leqslant \frac{2^{O(t)} \cdot t^{t/2} \cdot k^{10}}{\Delta^{t-1}}\,.$$

We turn to the bound on $\|M - A\|^2$. First we bound $\langle \tilde{\mathbb{E}}\, ww^\top, A\rangle$. Getting started,

$$\tilde{\mathbb{E}}\left(\sum_{i\in[k]}\langle w, a_i\rangle\right)^2 = \tilde{\mathbb{E}}\left(\sum_{i\in[n]} w_i\right)^2 \geqslant (1-\Delta^{-t})^2 \cdot n^2/k^2\,.$$

By Lemma 9.5.6, choosing t later,

$$\sum_{i\neq j\in[k]} \tilde{\mathbb{E}}\langle a_i, w\rangle\langle a_j, w\rangle \leqslant n^2 2^{O(t)} t^{t/2}\cdot k\cdot \frac{1}{\Delta^t}\,.$$

Together, these imply

$$\tilde{\mathbb{E}}\sum_{i\in[k]}\langle w, a_i\rangle^2 \geqslant \frac{n^2}{k^2}\cdot\left[1 - \frac{2^{O(t)}t^{t/2}k^3}{\Delta^t}\right]\,.$$

At the same time, $\|\tilde{\mathbb{E}}\, ww^T\|_F \leqslant \frac{1}{k}\|A\|_F$ by minimality (since the uniform distribution over cluster indicators satisfies $\mathcal{A}$), and by routine calculation and assumption (D1), $\|A\|_F \leqslant \frac{n}{\sqrt{k}}(1 + O(\Delta^{-t}))$. Together, we have obtained

$$\langle M, A\rangle \geqslant \left(1 - \frac{2^{O(t)}t^{t/2}k^3}{\Delta^t}\right)\cdot \|A\|\|M\|$$

which can be rearranged to give $\|M - A\|^2 \leqslant n\cdot\frac{2^{O(t)}t^{t/2}k^4}{\Delta^t}$. □

9.5.3 Identifiability

In this section we prove Lemma 9.5.6. We use the following helpful lemmas. The first is in spirit an SoS version of Lemma 9.2.1.

Lemma 9.5.8. *Let $\mu_1,\dots,\mu_k,\mathcal{D}_1,\dots,\mathcal{D}_k,t$ be as in Theorem 9.5.1. Let $\overline{\mu}_i$ be as in (D1). Suppose (D1) occurs for samples $X_1,\dots,X_n$. Let $\mathcal{A}$ be the system from Lemma 9.4.2, with $\alpha = 1/k$ and any τ. Then*

$$\mathcal{A} \vdash_{O(t)} \langle a_j, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} \leqslant 2^{t+2}t^{t/2}\cdot\frac{n}{k}\cdot\langle a_j, w\rangle^{t-1}\cdot\|\mu - \overline{\mu}_j\|^t\,.$$

The second lemma is an SoS triangle inequality, capturing the consequences of separation of the means. The proof is standard given Fact 10.0.5.

Lemma 9.5.9. *Let $a, b \in \mathbb{R}^k$ and $t \in \mathbb{N}$ be a power of 2. Let $\Delta = \|a - b\|$. Let $u = (u_1, \dots, u_k)$ be indeterminates. Then $\vdash_t \|a - u\|^t + \|b - u\|^t \geqslant 2^{-t} \cdot \Delta^t$.*

The last lemma helps put the previous two together. Although we have phrased this lemma to concorde with the mixture model setting, we note that the proof uses nothing about mixture models and consists only of generic manipulations of pseudodistributions.

Lemma 9.5.10. *Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k, X_1, \dots, X_n$ be as in Theorem 9.5.1. Let a_j be the $0/1$ indicator for the set of samples drawn from $\mathcal{D}_j$. Suppose $\tilde{\mathbb{E}}$ is a degree-$O(t)$ pseudodistribution which satisfies*

$$\begin{aligned}
\langle a_j, w\rangle &\leqslant n \\
\langle a_\ell, w\rangle &\leqslant n \\
\|\mu - \overline{\mu}_j\|^{2t} + \|\mu - \overline{\mu}_\ell\|^{2t} &\geqslant A \\
\langle a_j, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} &\leqslant Bn\langle a_j, w\rangle^{t-1}\|\mu - \overline{\mu}_j\|^t \\
\langle a_\ell, w\rangle^t \|\mu - \overline{\mu}_\ell\|^{2t} &\leqslant Bn\langle a_\ell, w\rangle^{t-1}\|\mu - \overline{\mu}_\ell\|^t
\end{aligned}$$

for some scalars $A, B \geqslant 0$. Then

$$\tilde{\mathbb{E}}\langle a_j, w\rangle\langle a_\ell, w\rangle \leqslant \frac{2n^2 B}{\sqrt{A}}.$$

Now we have the tools to prove Lemma 9.5.6.

Proof of Lemma 9.5.6. We will verify the conditions to apply Lemma 9.5.10. By Lemma 9.5.8, when (D1) holds, the pseudoexpectation $\tilde{\mathbb{E}}$ satisfies

$$\langle a_j, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} \leqslant Bn\langle a_j, w\rangle^{t-1}\|\mu - \overline{\mu}_j\|^t$$

for $B = 4(4t)^{t/2}/k$, and similarly with j, ℓ interposed. Similarly, by separation of the empirical means, $\tilde{\mathbb{E}}$ satisfies $\|\mu - \overline{\mu}_j\|^{2t} + \|\mu - \overline{\mu}_\ell\|^{2t} \geqslant A$ for $A = 2^{-2t}\Delta^{2t}$, recalling that the empirical means are pairwise separated by at least $\Delta - 2\Delta^{-t}$. Finally, clearly $\mathcal{A} \vdash_{O(1)} \langle a_j, w\rangle \leqslant n$ and similarly for $\langle a_\ell, w\rangle$. So applying Lemma 9.5.10 we get

$$\tilde{\mathbb{E}}\langle a_j, w\rangle\langle a_\ell, w\rangle \leqslant \frac{2n^2 B}{\sqrt{A}} \leqslant \frac{n^2 2^{2t+2} t^{t/2}}{k} \cdot \frac{1}{\Delta^t}\,. \qquad \square$$

9.5.4 Proof of Lemma 9.5.8

In this subsection we prove Lemma 9.5.8. We use the following helpful lemmata. The first bounds error from samples selected from the wrong cluster using the moment inequality.

Lemma 9.5.11. *Let $j, \mathcal{A}, X_1, \ldots, X_n, \mu_j, \overline{\mu}_j$ be as in Lemma 9.5.8. Then*

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \overline{\mu}_j\rangle\right)^t \leqslant 2t^{t/2} \cdot \langle a_j, w\rangle^{t-1} \|\mu - \overline{\mu}_j\|^t\,.$$

Proof. The proof goes by Hölder's inequality followed by the moment inequality in $\mathcal{A}$. Carrying this out, by Fact 10.0.8 and evenness of t,

$$\{w_i^2 = w_i\} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \overline{\mu}_j\rangle\right)^t \leqslant \left(\sum_{i \in S_j} w_i\right)^{t-1} \cdot \left(\sum_{i \in [n]} w_i \langle \mu - X_i, \mu - \overline{\mu}_j\rangle^t\right).$$

Then, using the main inequality in $\mathcal{A}$,

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i\right)^{t-1} \cdot 2t^{t/2} \cdot \|\mu - \overline{\mu}_j\|^t = 2t^{t/2} \cdot \langle a_j, w\rangle^{t-1} \|\mu - \overline{\mu}_j\|^t\,. \qquad \square$$

The second lemma bounds error from deviations in the empirical t-th moments of the samples from the j-th cluster.

Lemma 9.5.12. *Let $\mu_1, \dots, \mu_k, \mathcal{D}_1, \dots, \mathcal{D}_k$ be as in Theorem 9.5.1. Suppose condition (D1) holds for samples $X_1, \dots, X_n$. Let $w_1, \dots, w_n$ be indeterminates. Let $u = u_1, \dots, u_d$ be an indeterminate. Then for every $j \in [k]$,*

$$\{w_i^2 = w_i\} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle X_i - \overline{\mu}_j, u \rangle \right)^t \leqslant \langle a_j, w \rangle^{t-1} \cdot 2 \cdot \frac{n}{k} \cdot \|u\|^t .$$

Proof. The first step is Hölder's inequality again:

$$\{w_i^2 = w_i\} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle X_i - \overline{\mu}_j, u \rangle \right)^t \leqslant \langle a_j, w \rangle^{t-1} \cdot \sum_{i \in S_j} \langle X_i - \overline{\mu}_j, u \rangle^t .$$

Finally, condition (D1) yields

$$\{w_i^2 = w_i\} \vdash_{O(t)} \left(\sum_{i \in S_j} w_i \langle X_i - \overline{\mu}_j, u \rangle \right)^t \leqslant \langle a_j, w \rangle^{t-1} \cdot 2 \cdot \frac{n}{k} \cdot \|u\|^t . \qquad \square$$

We can prove Lemma 9.5.8 by putting together Lemma 9.5.11 and Lemma 9.5.12.

Proof of Lemma 9.5.8. Let $j \in [k]$ be a cluster and recall $a_j \in \{0,1\}^n$ is the 0/1 indicator for the samples in cluster j. Let S_j be the samples in the j-th cluster, with empirical mean $\overline{\mu}_j$. We begin by writing $\langle a_j, w \rangle \|\mu - \overline{\mu}_j\|^2$ in terms of samples $X_1, \dots, X_n$.

$$\begin{aligned} \langle a_j, w \rangle \|\mu - \overline{\mu}_j\|^2 &= \sum_{i \in [n]} w_i \langle \mu - \overline{\mu}_j, \mu - \overline{\mu}_j \rangle \\ &= \sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \overline{\mu}_j \rangle + \sum_{i \in [n]} w_i \langle X_i - \overline{\mu}_j, \mu - \overline{\mu}_j \rangle . \end{aligned}$$

Hence, using $(a+b)^t \leqslant 2^t (a^t + b^t)$, we obtain

$$\vdash_{O(t)} \langle a_j, w \rangle^t \|\mu - \overline{\mu}_j\|^{2t} \leqslant 2^t \cdot \left(\sum_{i \in S_j} w_i \langle \mu - X_i, \mu - \overline{\mu}_j \rangle \right)^t + 2^t \cdot \left(\sum_{i \in S_j} w_i \langle X_i - \overline{\mu}_j, \mu - \overline{\mu}_j \rangle \right)^t .$$

Now using Lemma 9.5.11 and Lemma 9.5.12,

$$\mathcal{A} \vdash_{O(t)} \langle a_j, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} \leqslant 2^{t+2} t^{t/2} \cdot \frac{n}{k} \cdot \langle a_j, w\rangle^{t-1} \cdot \|\mu - \overline{\mu}_j\|^t$$

as desired. □

9.5.5 Proof of Lemma 9.5.10

We prove Lemma 9.5.10. The proof only uses standard SoS and pseudodistribution tools. The main inequality we will use is the following version of Hölder's inequality.

Fact 9.5.13 (Pseudoexpectation Hölder's, see Lemma A.4 in [30])**.** *Let p be a degree-ℓ polynomial. Let $t \in N$ and let $\tilde{\mathbb{E}}$ be a degree-$O(t\ell)$ pseudoexpectation on indeterminates x. Then*

$$\tilde{\mathbb{E}}\, p(x)^{t-2} \leqslant \left(\tilde{\mathbb{E}}\, p(x)^t\right)^{\frac{t-2}{t}} .$$

Now we can prove Lemma 9.5.10.

Proof of Lemma 9.5.10. We first establish the following inequality.

$$\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} \leqslant B^2 n^2 \cdot \tilde{\mathbb{E}}\langle a_j, w\rangle^{t-2} \langle a_\ell, w\rangle^t . \tag{9.5.1}$$

(The inequality will also hold by symmetry with j and ℓ exchanged.) This we do as follows:

$$\begin{aligned}\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t} &\leqslant Bn\, \tilde{\mathbb{E}}\langle a_j, w\rangle^{t-1} \langle a_\ell, w\rangle^t \|\mu - \overline{\mu}_j\|^t \\ &\leqslant Bn \left(\tilde{\mathbb{E}}\langle a_j, w\rangle^{t-2} \langle a_\ell, w\rangle^t\right)^{1/2} \cdot \left(\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \|\mu - \overline{\mu}_j\|^{2t}\right)^{1/2}\end{aligned}$$

where the first line is by assumption on $\tilde{\mathbb{E}}$ and the second is by pseudoexpectation Cauchy-Schwarz. Rearranging gives the inequality (9.5.1).

Now we use this to bound $\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t$. By hypothesis,

$$\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \leqslant \frac{1}{A}\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t (\|\mu - \overline{\mu}_j\|^{2t} + \|\mu - \overline{\mu}_\ell\|^{2t}),$$

which, followed by (9.5.1) gives

$$\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \leqslant \frac{1}{A} \cdot B^2 n^2 \cdot \tilde{\mathbb{E}}\left[\langle a_j, w\rangle^{t-2}\langle a_\ell, w\rangle^t + \langle a_\ell, w\rangle^{t-2}\langle a_j, w\rangle^t\right].$$

Using $\langle a_j, w\rangle, \langle a_\ell, w\rangle \leqslant n$, we obtain

$$\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \leqslant \frac{2}{A} \cdot B^2 n^4 \cdot \tilde{\mathbb{E}}\langle a_j, w\rangle^{t-2}\langle a_\ell, w\rangle^{t-2}.$$

Finally, using Fact 9.5.13, the right side is at most $2B^2n^4/A \cdot \left(\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t\right)^{(t-2)/t}$, so cancelling terms we get

$$\left(\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t\right)^{2/t} \leqslant \frac{2B^2n^4}{A}.$$

Raising both sides to the $t/2$ power gives

$$\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t \leqslant \frac{2^{t/2}B^t n^{2t}}{A^{t/2}},$$

and finally using Cauchy-Schwarz,

$$\tilde{\mathbb{E}}\langle a_j, w\rangle\langle a_\ell, w\rangle \leqslant \left(\tilde{\mathbb{E}}\langle a_j, w\rangle^t \langle a_\ell, w\rangle^t\right)^{1/t} \leqslant \frac{2n^2B}{\sqrt{A}}. \qquad \square$$

9.5.6 Rounding

In this section we state and analyze our second-moment round algorithm. As have discussed already, our SoS proofs in the mixture model setting are quite strong, meaning that the rounding algorithm is relatively naive.

The setting in this section is as follows. Let $n, m \in \mathbb{N}$ with $m \ll n$. There is a ground-truth partition of $[n]$ into m parts $S_1, \ldots, S_m$ such that $|S_i| = (1 \pm \delta)\frac{n}{m}$.

Let $A \in \mathbb{R}^{n\times n}$ be the 0/1 indicator matrix for this partition, so $A_{ij} = 1$ if $i, j \in S_\ell$ for some ℓ and is 0 otherwise. Let $M \in \mathbb{R}^{n\times n}$ be a matrix such that $\|M - A\| \leqslant \varepsilon n$, where $\|\cdot\|$ is the Frobenious norm. The algorithm takes M and outputs a partition $C_1, \ldots, C_m$ of $[m]$ which makes few errors compared to $S_1, \ldots, S_m$.

Algorithm 9.5.14.
1: **function** RoundSecondMoments($M \in \mathbb{R}^{n\times n}, E \in \mathbb{R}$)
2: Let $S = [n]$
3: Let $v_1, \ldots, v_n$ be the rows of M
4: **for** $\ell = 1, \ldots, m$ **do**
5: Choose $i \in S$ uniformly at random
6: Let
$$C_\ell = \left\{ i' \in S : \|v_i - v_{i'}\|_2 \leqslant 2\frac{n^{1/2}}{E} \right\}$$
7: Let $S \leftarrow S \setminus C_\ell$
8: **end for**
9: **return** The clusters $C_1, \ldots, C_m$.
10: **end function**

We will prove the following theorem.

Theorem 9.5.15. *With notation as before Algorithm 9.5.14 with $E = m$, with probability at least $1 - \varepsilon^2 m^3$ Algorithm 9.5.14 returns a partition $C_1, \ldots, C_m$ of $[n]$ such that (up to a permutation of $[m]$), $C_\ell = T_\ell \cup B_\ell$, where $T_\ell \subseteq S_\ell$ has size $|T_\ell| \geqslant |S_\ell| - \varepsilon^2 mn$ and $|B_\ell| \leqslant \varepsilon^2 mn$.*

To get started analyzing the algorithm, we need a definition.

Definition 9.5.16. For cluster S_j, let $a_j \in \mathbb{R}^n$ be its 0/1 indicator vector. If $i \in S_j$, we say it is *E-good* if $\|v_i - a_j\|_2 \leqslant \sqrt{n/E}$, and otherwise *$E$-bad*, where v_i is the i-th row of M. Let $I_g \subseteq [n]$ denote the set of E-good indices and I_b denote the set of

E-bad indices. (We will choose E later.) For any $j = 1, \dots, k$, let $I_{g,j} = I_g \cap S_j$ denote the set of good indices from cluster j.

We have:

Lemma 9.5.17. *Suppose E as in RoundSecondMoments satisfies $E \geqslant m/8$. Suppose that in iterations $1, \dots, m$, RoundSecondMoments has chosen only good vectors. Then, there exists a permutation $\pi : [m] \to [m]$ so that $C_\ell = I_{g,\pi(\ell)} \cup B_\ell$, where $B_\ell \subseteq I_b$ for all ℓ.*

Proof. We proceed inductively. We first prove the base case. WLOG assume that the algorithm picks v_1, and that v_1 is good, and is from component j. Then, for all $i \in I_{g,j}$, by the triangle inequality we have $\|v_i - v_1\|_2 \leqslant 2\frac{n^{1/2}}{B}$, and so $I_{g,j} \subseteq C_1$. Moreover, if $i \in I_{g,j'}$ for some $j' \neq j$, we have

$$\|v_i - v_1\|_2 \geqslant \|a'_j - a_j\|_2 - 2\frac{n^{1/2}}{E^{1/2}} \geqslant \frac{n^{1/2}}{\sqrt{m}} - 2\frac{n^{1/2}}{E^{1/2}} > 2\frac{n^{1/2}}{E^{1/2}},$$

and so in this case $i \notin C_1$. Hence $C_1 = I_{g,j} \cup B_1$ for some $B_1 \subseteq I_b$.

Inductively, suppose that if the algorithm chooses good indices in iterations $1, \dots, a-1$, then there exist distinct $j_1, \dots, j_{a-1}$ so that $C_\ell = I_{g,j_\ell} \cup B_\ell$ for $B_\ell \subseteq I_b$. We seek to prove that if the algorithm chooses a good index in iteration a, then $C_a = I_{g,j_a} \cup B_a$ for some $j_a \notin \{j_1, \dots, j_{a-1}\}$ and $B_a \subseteq I_b$. Clearly by induction this proves the Lemma. WLOG assume that the algorithm chooses v_1 in iteration a. Since by assumption 1 is good, and we have removed I_{g_ℓ} for $\ell = 1, \dots, a-1$, then $1 \in I_{g,j_a}$ for some $j_a \notin \{j_1, \dots, j_{a-1}\}$. Then, the conclusion follows from the same calculation as in the base case. □

Lemma 9.5.18. *There are at most $\varepsilon^2 En$ indices which are E-bad; i.e. $|I_b| \leqslant \varepsilon^2 En$.*

Proof. We have

$$\begin{aligned}\varepsilon^2 n^2 &\geqslant \left\| M - \sum_{i \leqslant m} a_i a_i^\top \right\|_F^2 \geqslant \sum_j \sum_{i \in S_j \text{ bad}} \|v_i - a_j\|_2^2 \\ &\geqslant \frac{n}{E} |I_b| \ ,\end{aligned}$$

from which the claim follows by simplifying. □

This in turns implies:

Lemma 9.5.19. *With probability at least* $1 - \varepsilon^2 m^3$*, the algorithm* RoundSecondMoments *chooses good indices in all* k *iterations.*

Proof. By Lemma 9.5.18, in the first iteration the probability that a bad vector is chosen is at most $\varepsilon^2 E$. Conditioned on the event that in iterations $1, \ldots, a$ the algorithm has chosen good vectors, then by Lemma 9.5.17, there is at least one j_a so that no points in I_{g,j_a} have been removed. Thus at least $(1 - \delta)n/m$ vectors remain, and in total there are at most $\varepsilon^2 En$ bad vectors, by Lemma 9.5.18. So, the probability of choosing a bad vector is at most $\varepsilon^2 Em$. Therefore, by the chain rule of conditional expectation and our assumption , the probability we never choose a bad vector is at least

$$\left(1 - \varepsilon^2 Em\right)^m$$

Choosing $E = m$ this is $(1 - \varepsilon^2 m^2)^m \geqslant 1 - \varepsilon^2 m^3$. as claimed. □

Now Theorem 9.5.15 follows from putting together the lemmas.

9.6 Robust estimation: algorithm and analysis

Our algorithm for robust estimation is very similar to our algorithm for mixture models. Suppose the underlying distribution $\mathcal{D}$, whose mean μ^* the algorithm robustly estimates, is $10t$-explicitly bounded. As a reminder, the input to the algorithm is a list of $X_1, \dots, X_n \in \mathbb{R}^d$ and a sufficiently-small $\varepsilon > 0$. The guarantee is that at least $(1-\varepsilon)n$ of the vectors were sampled according to $\mathcal{D}$, but εn of the vectors were chosen adversarially.

The algorithm solves a semidefinite program to obtain a degree $O(t)$ pseudodistribution which satisfies the system $\mathcal{A}$ from Section 9.4 with $\alpha = 1 - \varepsilon$ and $\tau = 0$. Throughout this section, we will always assume that $\mathcal{A}$ is instantiated with these parameters, and omit them for conciseness. Then the algorithm just outputs $\tilde{\mathbb{E}}\, \mu$ as its estimator for μ^*.

Our main contribution in this section is a formal description of an algorithm EstimateMean which makes these ideas rigorous, and the proof of the following theorem about its correctness:

Theorem 9.6.1. *Let $\varepsilon > 0$ sufficiently small and $t \in \mathbb{N}$. Let $\mathcal{D}$ be a $10t$-explicitly bounded distribution over $\mathbb{R}^d$ with mean μ^*. Let $X_1, \dots, X_n$ be an ε-corrupted set of samples from $\mathcal{D}$ where $n = d^{O(t)}/\varepsilon^2$. Then, given ε, t and $X_1, \dots, X_n$, the algorithm EstimateMean runs in time $d^{O(t)}$ and outputs μ so that $\|\mu - \mu^*\|_2 \leqslant O(t^{1/2}\varepsilon^{1-1/t})$, with probability at least $1 - 1/d$.*

As a remark, observe that if we set $t = 2\log 1/\varepsilon$, then the error becomes $O(\varepsilon\sqrt{\log 1/\varepsilon})$. Thus, with $n = O(d^{O(\log 1/\varepsilon)}/\varepsilon^2)$ samples and $n^{O(\log 1/\varepsilon)} = d^{O(\log 1/\varepsilon)^2}$ runtime, we achieve the same error bounds for general explicitly

bounded distributions as the best known polynomial time algorithms achieve for Gaussian mean estimation.

9.6.1 Additional Preliminaries

Throughout this section, let $[n] = S_g \cup S_b$, where S_g is the indices of the uncorrupted points, and S_b is the indices of the corrupted points, so that $|S_b| = \varepsilon n$ by assumption. Moreover, let $Y_1, \ldots, Y_n$ be iid from $\mathcal{D}$ so that $Y_i = X_i$ for all $i \in S_g$.

We now state some additional tools we will require in our algorithm.

Naive Pruning We will require the following elementary pruning algorithm, which removes all points which are very far away from the mean. We require this only to avoid some bit-complexity issues in semidefinite programming; in particular we just need to ensure that the vectors $X_1, \ldots, X_n$ used to form the SDP have polynomially-bounded norms. Formally:

Lemma 9.6.2 (Naive pruning)**.** *Let ε, t, μ^*, and $X_1, \ldots, X_n$ be as in Theorem 9.6.1. There is an algorithm NaivePrune, which given ε, t and $X_1, \ldots, X_n$, runs in time $O(\varepsilon d n^2)$, and outputs a subset $S \subseteq [n]$ so that with probability $1 - 1/d^{10}$, the following holds:*

- *No uncorrupted points are removed, that is $S_g \subseteq S$, and*
- *For all $i \in S$, we have $\|X_i - \mu^*\| \leqslant O(d)$.*

In this case, we say that NaivePrune succeeds.

This algorithm goes by straightforward outlier-removal. It is very similar the procedure described in Fact 4.18 of [64] (using bounded t-th moments instead of sub-Gaussianity), so we omit it.

Satisfiability In our algorithm, we will use the same set of polynomial equations $\widehat{\mathcal{A}}$ as in Lemma 9.4.2. However, the data we feed in does not exactly fit the assumptions in the Lemma. Specifically, because the adversary is allowed to remove an ε-fraction of good points, the resulting uncorrupted points are no longer iid from $\mathcal{D}$. Despite this, we are able to specialize Lemma 9.4.2 to this setting:

Lemma 9.6.3. *Fix $\varepsilon > 0$ sufficiently small, and let $t \in \mathbb{N}$, $t \geqslant 4$ be a power of 2. Let $\mathcal{D}$ be a $10t$-explicitly bounded distribution. Let $X_1, \dots, X_n \in \mathbb{R}^d$ be an ε-corrupted set of samples from $\mathcal{D}$, and let $\widehat{\mathcal{A}}$ be as in Lemma 9.4.2. The conclusion (1 – Satisfiability) of Lemma 9.4.2 holds, with w taken to be the $0/1$ indicator of the $(1-\varepsilon)n$ good samples among $X_1, \dots, X_n$.*

We sketch the proof of Lemma 9.6.3 in Section 9.7.4.

9.6.2 Formal Algorithm Specification

With these tools in place, we can now formally state the algorithm. The formal specification of this algorithm is given in Algorithm 9.6.4.

Algorithm 9.6.4 (Robust Mean Estimation)**.** 1: **function** EstimateMean($\varepsilon, t, \kappa, X_1, \dots, X_n$)

2: Preprocess: let $X_1, \dots, X_n \leftarrow$ NaivePrune($\varepsilon, X_1, \dots, X_n$), and let $\widehat{\mu}$ be the empirical mean

3: Let $X_i \leftarrow X_i - \widehat{\mu}$

4: By semidefinite programming, find a pseudoexpectation of degree $O(t)$ which satisfies the structured subset polynomials from Lemma 9.6.3, with $\alpha = (1-\varepsilon)n$ and $\tau = 0$.

5: **return** $\tilde{\mathbb{E}}\, \mu + \widehat{\mu}$.

6: **end function**

The first two lines of Algorithm 9.6.4 are only necessary for bit complexity reasons, since we cannot solve SDPs exactly. However, since we can solve them to doubly-exponential accuracy in polynomial time, it suffices that all the quantities are at most polynomially bounded (indeed, exponentially bounded suffices) in norm, which these two lines easily achieve. For the rest of this section, for simplicity of exposition, we will ignore these issues.

9.6.3 Deterministic conditions

With these tools in place, we may now state the deterministic conditions under which our algorithm will succeed. Throughout this section, we will condition on the following events holding simultaneously:

(E1) NaivePrune succeeds,

(E2) The conclusion of Lemma 9.6.3 holds,

(E3) We have the following concentration of the uncorrupted points:

$$\left\| \frac{1}{n} \sum_{i \in S_g} X_i - \mu^* \right\| \leqslant O(t^{1/2} \varepsilon^{1-1/t}) \text{ , and}$$

(E4) We have the following concentration of the empirical t-th moment tensor:

$$\frac{1}{n}\sum_{i\in[n]}\left[(Y_i-\mu^*)^{\otimes t/2}\right]\left[(Y_i-\mu^*)^{\otimes t/2}\right]^\top \preceq \mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[(X-\mu^*)^{\otimes t/2}\right]\left[(X-\mu^*)^{\otimes t/2}\right]^\top + 0.1\cdot \mathrm{Id}\ ,$$

for Id is the $d^{t/2}\times d^{t/2}$-sized identity matrix.

The following lemma says that with high probability, these conditions hold simultaneously:

Lemma 9.6.5. *Let ε, t, μ^*, and $X_1,\dots,X_n \in \mathbb{R}^d$ be as in Theorem 9.6.1. Then, Conditions (E1)-(E4) hold simultaneously with probability at least $1-1/d^5$.*

We defer the proof of this lemma to the Appendix.

For simplicity of notation, throughout the rest of the section, we will assume that NaivePrune does not remove any points whatsoever. Because we are conditioning on the event that it removes no uncorrupted points, it is not hard to see that this is without loss of generality.

9.6.4 Identifiability

Our main identifiability lemma is the following.

Lemma 9.6.6. *Let ε, t, μ^* and $X_1,\dots,X_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E1)–(E4). Then, we have*

$$\mathcal{A} \vdash_{O(t)} \|\mu-\mu^*\|^{2t} \leqslant O(t^{t/2})\cdot\varepsilon^{t-1}\cdot\|\mu-\mu^*\|^t\ .$$

Since this lemma is the core of our analysis for robust estimation, in the remainder of this section we prove it. The proof uses the following three lemmas

to control three sources of error in $\tilde{\mathbb{E}}\,\mu$, which we prove in Section 9.6.6. The first, Lemma 9.6.7 controls sampling error from true samples from $\mathcal{D}$.

Lemma 9.6.7. *Let ε, t, μ^* and $X_1, \dots, X_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E1)–(E4) satisfy (E1)–(E4). Then, we have*

$$\vdash_{O(t)} \left(\sum_{i \in S_g} \langle X_i - \mu^*, \mu - \mu^* \rangle \right)^t \leqslant O(\varepsilon^{t-1}) \cdot t^{t/2} \cdot n^t \cdot \|\mu - \mu^*\|^t .$$

To describe the second and third error types, we think momentarily of $w \in \mathbb{R}^n$ as the 0/1 indicator for a set S of samples whose empirical mean will be the output of the algorithm. (Of course this is not strictly true, but this is a convenient mindset in constructing SoS proofs.) The second type of error comes from the possible failure of S to capture some ε fraction of the good samples from $\mathcal{D}$. Since $\mathcal{D}$ has $O(t)$ bounded moments, if T is a set of m samples from $\mathcal{D}$, the empirical mean of any $(1-\varepsilon)m$ of them is at most $\varepsilon^{1-1/t}$-far from the true mean of $\mathcal{D}$.

Lemma 9.6.8. *Let ε, t, μ^* and $X_1, \dots, X_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E1)–(E4). Then, we have*

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S_g} (w_i - 1) \langle X_i - \mu^*, \mu - \mu^* \rangle \right)^t \leqslant 2\varepsilon^{t-1} n^t \cdot t^{t/2} \cdot \|\mu - \mu^*\|^t .$$

The third type of error is similar in spirit: it is the contribution of the original uncorrupted points that the adversary removed. Formally:

Lemma 9.6.9. *Let ε, t, μ^* and $X_1, \dots, X_n \in \mathbb{R}^d$ and $Y_1, \dots, Y_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E1)–(E4). Then, we have*

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \in S_b} \langle Y_i - \mu^*, \mu - \mu^* \rangle \right)^t \leqslant 2\varepsilon^{t-1} n^t \cdot t^{t/2} \cdot \|\mu - \mu^*\|^t .$$

Finally, the fourth type of error comes from the εn adversarially-chosen vectors. We prove this lemma by using the bounded-moments inequality in $\mathcal{A}$.

Lemma 9.6.10. *Let ε, t, μ^* and $X_1, \dots, X_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E1)–(E4). Then, we have*

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i \notin S_g} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle \right)^t \leqslant 2\varepsilon^{t-1} n^t \cdot t^{t/2} \cdot \|\mu - \mu^*\|^t .$$

With these lemmas in place, we now have the tools to prove Lemma 9.6.6.

Proof of Lemma 9.6.6. Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be as in Theorem 9.6.1. We expand the norm $\|\mu - \mu^*\|^2$ as $\langle \mu - \mu^*, \mu - \mu^* \rangle$ and rewrite $\sum_{i \in [n]} w_i \mu$ as $\sum_{i \in [n]} w_i X_i$:

$$\begin{aligned}
\sum_{i \in [n]} w_i \|\mu - \mu^*\|^2 &\overset{(a)}{=} \sum_{i \in [n]} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle \\
&\overset{(b)}{=} \sum_{i \in S_g} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle + \sum_{i \in S_b} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle \\
&\overset{(c)}{=} \sum_{i \in S_g} \langle X_i - \mu^*, \mu - \mu^* \rangle + \sum_{i \in S_g} (w_i - 1) \langle X_i - \mu^*, \mu - \mu^* \rangle \\
&\quad + \sum_{i \in S_b} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle \\
&\overset{(d)}{=} \sum_{i \in [n]} \langle X_i - \mu^*, \mu - \mu^* \rangle + \sum_{i \in S_g} (w_i - 1) \langle X_i - \mu^*, \mu - \mu^* \rangle \\
&\quad - \sum_{i \in S_b} \langle Y_i - \mu^*, \mu - \mu^* \rangle + \sum_{i \in S_b} w_i \langle X_i - \mu^*, \mu - \mu^* \rangle ,
\end{aligned}$$

where (a) follows from the mean axioms, (b) follows from splitting up the uncorrupted and the corrupted samples, (c) follows by adding and subtracting 1 to each term in S_g, and (d) follows from the assumption that $Y_i = X_i$ for all $i \in [n]$. We will rearrange the last term by adding and subtracting μ. Note the following polynomial identity:

$$\langle X_i - \mu^*, \mu - \mu^* \rangle = \langle X_i - \mu, \mu - \mu^* \rangle + \|\mu - \mu^*\|^2$$

and put it together with the above to get

$$\sum_{i\in[n]} w_i\|\mu-\mu^*\|^2 = \sum_{i\in S_g}\langle X_i-\mu^*,\mu-\mu^*\rangle + \sum_{i\in S_g}(w_i-1)\langle X_i-\mu^*,\mu-\mu^*\rangle \\ - \sum_{i\in S_b}\langle Y_i-\mu^*,\mu-\mu^*\rangle + \sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle + \sum_{i\in S_b} w_i\|\mu-\mu^*\|^2 .$$

which rearranges to

$$\sum_{i\in S_g} w_i\|\mu-\mu^*\|^2 = \sum_{i\in S_g}\langle X_i-\mu^*,\mu-\mu^*\rangle + \sum_{i\in S_g}(w_i-1)\langle X_i-\mu^*,\mu-\mu^*\rangle \\ - \sum_{i\in S_b}\langle Y_i-\mu^*,\mu-\mu^*\rangle + \sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle .$$

Now we use $\vdash_t\ (x+y+z+w)^t \leqslant \exp(t)\cdot(x^t+y^t+z^t+w^t)$ for any even t, and Lemma 9.6.7, Lemma 9.6.8, and Lemma 9.6.10 and simplify to conclude

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i\in S_g} w_i\right)^t \|\mu-\mu^*\|^{2t} \leqslant \exp(t)\cdot t^{t/2}\cdot n^t\cdot\varepsilon^{t-1}\cdot\|\mu-\mu^*\|^t .$$

Lastly, since $\mathcal{A}\vdash_2 \sum_{i\in T} w_i \geqslant (1-2\varepsilon)n$, we get

$$\mathcal{A} \vdash_{O(t)} \|\mu-\mu^*\|^{2t} \leqslant \exp(t)\cdot t^{t/2}\cdot\varepsilon^{t-1}\cdot\|\mu-\mu^*\|^t ,$$

as claimed. □

9.6.5 Rounding

The rounding phase of our algorithm is extremely simple. If $\tilde{\mathbb{E}}$ satisfies $\mathcal{A}$, we have by Lemma 9.6.6 and pseudoexpectation Cauchy-Schwarz that

$$\tilde{\mathbb{E}}\,\|\mu-\mu^*\|^{2t} \leqslant \exp(t)\cdot t^{t/2}\cdot\varepsilon^{t-1}\cdot\tilde{\mathbb{E}}\left(\|\mu-\mu^*\|^t\right) \leqslant \exp(t)\cdot t^{t/2}\cdot\varepsilon^{t-1}\cdot\tilde{\mathbb{E}}\left(\|\mu-\mu^*\|^{2t}\right)^{1/2}$$

which implies that

$$\tilde{\mathbb{E}}\,\|\mu-\mu^*\|^{2t} \leqslant \exp(t)\cdot t^t\cdot\varepsilon^{2(t-1)} . \tag{9.6.1}$$

Once this is known, analyzing $\|\tilde{\mathbb{E}}\,\mu - \mu^*\|$ is straightforward. By (9.6.1) and pseudo-Cauchy-Schwarz again,

$$\|\tilde{\mathbb{E}}[\mu] - \mu^*\|^2 \leqslant \tilde{\mathbb{E}}\,\|\mu - \mu^*\|^2 \leqslant \left(\tilde{\mathbb{E}}\,\|\mu - \mu^*\|^{2t}\right)^{1/t} \leqslant O(t \cdot \varepsilon^{2-2/t}) \,,$$

which finishes analyzing the algorithm.

9.6.6 Proofs of Lemmata 9.6.7–9.6.10

We first prove Lemma 9.6.7, which is a relatively straightforward application of SoS Cauchy Schwarz.

Proof of Lemma 9.6.7. We have

$$\begin{aligned}
\vdash_{O(t)} \left(\sum_{i\in S_g} \langle X_i - \mu^*, \mu - \mu^*\rangle\right)^t &= \left(\left\langle \sum_{i\in S_g} (X_i - \mu^*), \mu - \mu^* \right\rangle\right)^t \\
&\leqslant \left\|\sum_{i\in S_g} (X_i - \mu^*)\right\|^t \|\mu - \mu^*\|^t \\
&\leqslant \left(n \cdot O\left(\varepsilon^{1-1/t}\right) \cdot t^{1/2}\right)^t \|\mu - \mu^*\|^t \,,
\end{aligned}$$

where the last inequality follows from (E3). This completes the proof. □

Before we prove Lemmata 9.6.8–9.6.10, we prove the following lemma which we will use repeatedly:

Lemma 9.6.11. *Let ε, t, μ^* and $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be as in Theorem 9.6.1, and suppose they satisfy (E4). Then, we have*

$$\mathcal{A} \vdash_{O(t)} \sum_{i\in[n]} \langle Y_i - \mu^*, \mu - \mu^*\rangle^t \leqslant 2nt^{t/2}\|\mu - \mu^*\|^t \,.$$

Proof. We have that

$$\begin{aligned}
\vdash_t \sum_{i\in[n]} \langle Y_i - \mu^*, \mu - \mu^*\rangle^t &= \left[(\mu-\mu^*)^{\otimes 2}\right]^\top \sum_{i\in[n]} \left[(Y_i-\mu^*)^{\otimes t/2}\right]\left[(Y_i-\mu^*)^{\otimes t/2}\right]^\top \left[(\mu-\mu^*)^{\otimes 2}\right] \\
&\overset{(a)}{\leqslant} n\left(\left[(\mu-\mu^*)^{\otimes 2}\right]^\top \left(\mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[(X-\mu^*)^{\otimes t/2}\right]\left[(X-\mu^*)^{\otimes t/2}\right]^\top + 0.1\cdot \mathrm{Id}\right)\left[(\mu-\mu^*)^{\otimes 2}\right]\right) \\
&= n\cdot \mathop{\mathbb{E}}_{X\sim\mathcal{D}} \langle X-\mu^*, \mu-\mu^*\rangle^t + n\cdot 0.1\cdot \|\mu-\mu^*\|^t \\
&\overset{(b)}{\leqslant} 2n\cdot t^{t/2}\|\mu-\mu^*\|^t \,,
\end{aligned}$$

where (a) follows from (E4) and (b) follows from $10t$-explicitly boundedness. □

We now return to the proof of the remaining Lemmata.

Proof of Lemma 9.6.8. We start by applying Hölder's inequality, Fact 10.0.8, (implicitly using that $w_i^2 = w_i \vdash_2 (1-w_i)^2 = 1-w_i$), to get

$$\begin{aligned}
\mathcal{A} \vdash_{O(t)} \left(\sum_{i\in S_g}(w_i-1)\langle X_i-\mu^*, \mu-\mu^*\rangle\right)^t &= \left(\sum_{i\in S_g}(1-w_i)\langle X_i-\mu^*, \mu-\mu^*\rangle\right)^t \\
&\leqslant \left(\sum_{i\in S_g}(w_i-1)\right)^{t-1}\left(\sum_{i\in S_g}\langle X_i-\mu^*, \mu-\mu^*\rangle^t\right).
\end{aligned}$$

By Lemma 9.6.11, we have

$$\begin{aligned}
\mathcal{A} \vdash_{O(t)} \sum_{i\in S_g}\langle X_i-\mu^*, \mu-\mu^*\rangle^t &\leqslant \sum_{i\in[n]}\langle Y_i-\mu^*, \mu-\mu^*\rangle^t \\
&\leqslant 2n\cdot t^{t/2}\cdot\|\mu-\mu^*\|^t \,.
\end{aligned}$$

At the same time,

$$A \vdash_2 \sum_{i\in T}(1-w_i) = (1-\varepsilon)n - \sum_{i\in[n]} w_i + \sum_{i\notin T} w_i = \sum_{i\notin T} w_i \leqslant \varepsilon n \,.$$

So putting it together, we have

$$\mathcal{A} \vdash_{O(t)} \left(\sum_{i\in T}(w_i-1)\langle X_i-\mu^*, \mu-\mu^*\rangle\right)^t \leqslant 2(\varepsilon n)^{t-1}\cdot n\cdot t^{t/2}\cdot\|\mu-\mu^*\|^t \,,$$

as claimed. □

Proof of Lemma 9.6.9. We apply Hölder's inequality to obtain that

$$\begin{aligned}\vdash_{O(t)} \left(\sum_{i\in S_b}\langle X_i-\mu^*,\mu-\mu^*\rangle\right)^t &\leqslant |S_b|^{t-1}\sum_{i\in S_b}\langle Y_i-\mu^*,\mu-\mu^*\rangle^t\\ &\overset{(a)}{\leqslant} (\varepsilon n)^{t-1}\sum_{i\in[n]}\langle Y_i-\mu^*,\mu-\mu^*\rangle^t\\ &\overset{(b)}{\leqslant} 2(\varepsilon n)^{t-1}nt^{t/2}\|\mu-\mu^*\|^t\ ,\end{aligned}$$

where (a) follows from the assumption on the size of S_b and since the additional terms in the sum are SoS, and (b) follows follows from Lemma 9.6.11. This completes the proof. □

Proof of Lemma 9.6.10. The proof is very similar to the proof of the two previous lemmas, except that we use the moment bound inequality in $\mathcal{A}$. Getting started, by Hölder's:

$$\mathcal{A}\vdash_{O(t)} \left(\sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle\right)^t \leqslant \left(\sum_{i\in S_b} w_i\right)^{t-1}\left(\sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle^t\right)$$

By evenness of t,

$$\vdash_t \sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle^t \leqslant \sum_{i\in[n]} w_i\langle X_i-\mu,\mu-\mu^*\rangle^t\ .$$

Combining this with the moment bound in $\mathcal{A}$,

$$\mathcal{A}\vdash_{O(t)} \left(\sum_{i\in S_b} w_i\langle X_i-\mu,\mu-\mu^*\rangle\right)^t \leqslant \left(\sum_{i\in S_b} w_i\right)^{t-1}\cdot 2\cdot t^{t/2}\cdot n\cdot\|\mu-\mu^*\|^t\ .$$

Finally, clearly $\mathcal{A}\vdash_2 \sum_{i\notin T} w_i \leqslant \varepsilon n$, which finishes the proof. □

9.7 Encoding structured subset recovery with polynomials

The goal in this section is to prove Lemma 9.4.2. The eventual system $\widehat{\mathcal{A}}$ of polynomial inequalities we describe will involve inequalities among matrix-valued polynomials. We start by justifying the use of such inequalities in the SoS proof system.

9.7.1 Matrix SoS proofs

Let $x = (x_1, \ldots, x_n)$ be indeterminates. We describe a proof system which can reason about inequalities of the form $M(x) \succeq 0$, where $M(x)$ is a symmetric matrix whose entries are polynomials in x.

Let $M_1(x), \ldots, M_m(x)$ be symmetric matrix-valued polynomials of x, with $M_i(x) \in \mathbb{R}^{s_i \times s_i}$, and let $q_1(x), \ldots, q_m(x)$ be scalar polynomials. (If $s_i = 1$ then M_i is a scalar valued polynomial.) Let $M(x)$ be another matrix-valued polynomial. We write

$$\{M_1 \succeq 0, \ldots, M_m \succeq 0, q_1(x) = 0, \ldots, q_m(x) = 0\} \vdash_d M \succeq 0$$

if there are vector-valued polynomials $\{r_S^j\}_{j \leqslant N, S \subseteq [m]}$ (where the S's are multisets), a matrix B, and a matrix Q whose entries are polynomials in the ideal generated by $q_1, \ldots, q_m$, such that

$$M = B^\top \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] B + Q(x)$$

and furthermore that $\deg \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \leqslant d$ for every $S \subseteq [m]$, and $\deg Q \leqslant d$. Observe that in the case $M_1, \ldots, M_m, M$ are actually 1×1 matrices, this reduces to the usual notion of scalar-valued sum of squares proofs.

Adapting pseudodistributions to the matrix case, we say a pseudodistribution $\tilde{\mathbb{E}}$ of degree $2d$ satisfies the inequalities $\{M_1(x) \succeq 0, \ldots, M_m(x) \succeq 0\}$ if for every multiset $S \subseteq [m]$ and $p \in \mathbb{R}[x]$ such that $\deg \left[p(x)^2 \cdot (\otimes_{i \in S} M_i(x))\right] \leqslant 2d$,

$$\tilde{\mathbb{E}} \left[p(x)^2 \cdot (\otimes_{i \in S} M_i(x))\right] \succeq 0 \,.$$

For completeness, we prove the following lemmas in the appendix.

Lemma 9.7.1 (Soundness)**.** *Suppose $\tilde{\mathbb{E}}$ is a degree-$2d$ pseudodistribution which satisfies constraints $\{M_1 \succeq 0, \ldots, M_m \succeq 0\}$, and*

$$\{M_1 \succeq 0, \ldots, M_m \succeq 0\} \vdash_{2d} M \succeq 0 \,.$$

Then $\tilde{\mathbb{E}}$ satisfies $\{M_1 \succeq 0, \ldots, M_m \succeq 0, M \succeq 0\}$.

Lemma 9.7.2. *Let $f(x)$ be a degree-ℓ s-vector-valued polynomial in indeterminates x. Let $M(x)$ be a $s \times s$ matrix-valued polynomial of degree ℓ'. Then*

$$\{M \succeq 0\} \vdash_{\ell\ell'} \langle f(x), M(x) f(x) \rangle \geqslant 0 \,.$$

Polynomial-time algorithms to find pseudodistributions satisfying matrix-SoS constraints follow similar ideas as in the non-matrix case. In particular, recall that to enforce a scalar constraint $\{p(x) \geqslant 0\}$, one imposes the convex constraint $\tilde{\mathbb{E}}\, p(x)(x^{\otimes d})(x^{\otimes d})^\top \succeq 0$. Enforcing a constraint $\{M(x) \succeq 0\}$ can be accomplished similarly by adding constraints of the form $\tilde{\mathbb{E}}\, M(x) \succeq 0, \tilde{\mathbb{E}}\, M(x)p(x) \succeq 0$, etc.

9.7.2 Warmup: Gaussian moment matrix-polynomials

In this section we develop the encoding as low degree polynomials of the following properties of an n-variate vector w and a d-variate vector μ. We will not

be able to encode exactly these properties, but they will be our starting point. Let $d, n \in \mathbb{N}$, and suppose there are some vectors (a.k.a. samples) $X_1, \ldots, X_n \in \mathbb{R}^d$.

1. Boolean: $w \in \{0,1\}^n$.
2. Size: $(1-\tau)\alpha n \leqslant \sum_{i\in[n]} w_i \leqslant (1+\tau)\alpha n$.
3. Empirical mean: $\mu = \frac{1}{\sum_{i\in[n]} w_i} \sum_{i\in[n]} w_i X_i$.
4. t-th Moments: the t-th empirical moments of the vectors selected by the vector w, centered about μ, are subgaussian. That is,

$$\max_{u\in\mathbb{R}^d} \frac{1}{\alpha n} \sum_{i\in[n]} w_i \langle X_i - \mu, u\rangle^t \leqslant 2 \cdot t^{t/2} \|u\|^t .$$

The second property is already phrased as two polynomial inequalities, and the third can be rearranged to a polynomial equation. For the first, we use polynomial equations $w_i^2 = w_i$ for every $i \in [n]$. The moment constraint will be the most difficult to encode. We give two versions of this encoding: a simple one which will work when the distribution of the structured subset of samples to be recovered is Gaussian, and a more complex version which allows for any explicitly bounded distribution. For now we describe only the Gaussian version. We state some key lemmas and prove them for the Gaussian case. We carry out the general case in the following section.

To encode the bounded moment constraint, for this section we let $M(w, \mu)$ be the following matrix-valued polynomial

$$M(w,\mu) = \frac{1}{\alpha n} \sum_{i\in[n]} w_i \left[\left(X_i - \mu\right)^{\otimes t/2} \right] \left[\left(X_i - \mu\right)^{\otimes t/2} \right]^\top$$

Definition 9.7.3 (Structured subset axioms, Gaussian version)**.** For parameters $\alpha \in [0,1]$ (for the size of the subset), t (for which empirical moment to control), and

$\tau > 0$ (to account for some empirical deviations), the structured subset axioms are the following matrix-polynomial inequalities on variables $w = (w_1, \ldots, w_n)$, $\mu = (\mu_1, \ldots, \mu_d)$.

1. booleanness: $w_i^2 = w_i$ for all $i \in [n]$
2. size: $(1-\tau)\alpha n \leqslant \sum_{i\in[n]} w_i \leqslant (1+\tau)\alpha n$
3. t-th moment boundedness: $M(w,\mu) \preceq 2 \cdot \mathbb{E}_{X\sim\mathcal{N}(0,\mathrm{Id})} \left[X^{\otimes t/2}\right] \left[X^{\otimes t/2}\right]^\top$.
4. μ is the empirical mean: $\mu \cdot \sum_{i\in[n]} w_i = \sum_{i\in[n]} w_i X_i$.

Notice that in light of the last constraint, values for the variables μ are always determined by values for the variables w, so strictly speaking μ could be removed from the program. However, we find it notationally convenient to use μ. We note also that the final constraint, that μ is the empirical mean, will be used only for the robust statistics setting but seems unnecessary in the mixture model setting.

Next, we state and prove some key lemmas for this Gaussian setting, as warmups for the general setting.

Lemma 9.7.4 (Satisfiability, Gaussian case)**.** *Let $d \in \mathbb{N}$ and $\alpha = \alpha(d) > 0$. Let $t \in \mathbb{N}$. Suppose $(1-\tau)\alpha n \geqslant d^{100t}$. Let $0.1 > \tau > 0$. If $X_1, \ldots, X_n \in \mathbb{R}^d$ has a subset $S \subseteq [n]$ such that $\{X_i\}_{i\in S}$ are iid samples from $\mathcal{N}(\mu^*, \mathrm{Id})$ and $|S| \geqslant (1-\tau)\alpha n$, then with probability at least $1 - d^{-8}$ over these samples, the α, t, τ structured subset axioms are satisfiable.*

Proof. Suppose S has size exactly $(1-\tau)\alpha n$; otherwise replace S with a random subset of S of size exactly $(\alpha-\tau)n$. As a solution to the polynomials, we will take w to be the indicator vector of S and $\mu = \frac{1}{|S|}\sum_{i\in[n]} w_i X_i$. The booleanness and

size axioms are trivially satisfied. The spectral inequality

$$\frac{1}{\alpha n}\sum_{i\leqslant[n]} w_i \left[(X_i-\mu)^{\otimes t/2}\right]\left[(X_i-\mu)^{\otimes t/2}\right]^\top \preceq 2\cdot \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\mathrm{Id})}\left[X^{\otimes t/2}\right]\left[X^{\otimes t/2}\right]^\top$$

follows from concentration of the empirical mean to the true mean μ^* and standard matrix concentration (see e.g. [170]). □

The next lemma is actually a corollary of Lemma 9.7.2.

Lemma 9.7.5 (Moment bounds for polynomials of μ, Gaussian case)**.** *Let $f(\mu)$ be a length-d vector of degree-ℓ polynomials in indeterminates $\mu=(\mu_1,\dots,\mu_k)$. The t-th moment boundedness axiom implies the following inequality with a degree $t\ell$ SoS proof.*

$$\left\{M(w,\mu)\preceq 2\cdot \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\mathrm{Id})}\left[X^{\otimes t/2}\right]\left[X^{\otimes t/2}\right]^\top\right\}$$
$$\vdash_{O(t\ell)} \frac{1}{\alpha n}\sum_{i\in[n]} w_i\langle X_i-\mu, f(\mu)\rangle^t \leqslant 2\cdot \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\mathrm{Id})}\langle X, f(\mu)\rangle^t\,.$$

9.7.3 Moment polynomials for general distributions

In this section we prove Lemma 9.4.2.

We start by defining polynomial equations $\widehat{\mathcal{A}}$, for which we introduce some extra variables For every pair of multi-indices γ,ρ over $[k]$ with degree at most $t/2$, we introduce a variable $M_{\gamma,\rho}$. The idea is that $M=[M_{\gamma,\rho}]_{\gamma,\rho}$ forms an $n^{t/2}\times n^{t/2}$ matrix. By imposing equations of the form $M_{\gamma,\rho}=f_{\gamma,\rho}(w,\mu)$ for some explicit polynomials $f_{\gamma,\rho}$ of degree $O(t)$, we can ensure that

$$\langle u^{\otimes t/2}, Mu^{\otimes t/2}\rangle = 2\cdot t^{t/2}\|u\|^t - \frac{1}{\alpha n}\sum_{i\in[n]} w_i\langle X_i-\mu,u\rangle^t\,.$$

(This equation should be interpreted as an equality of polynomials in indeterminates u.) Let $\mathcal{L}$ be such a family of polynomial equations. Our final system

$\widehat{\mathcal{A}}(\alpha, t, \tau)$ of polynomial equations and inequalities follows. The important parameters are α, controlling the size of the set of samples to be selected, and t, how many moments to control. The parameter τ is present to account for random fluctuations in the sizes of the cluster one wants to recover.

Definition 9.7.6. Let $\widehat{\mathcal{A}}(\alpha, t, \tau)$ be the set of (matrix)-polynomial equations and inequalities on variables $w, \mu, M_{\gamma,\rho}$ containing the following.

1. Booleanness: $w_i^2 = w_i$ for all $i \in [n]$
2. Size: $(1-\tau)\alpha n \leqslant \sum w_i \leqslant (1+\tau)\alpha n$.
3. Empirical mean: $\mu \cdot \sum_{i\in[n]} w_i = \sum_{i\in[n]} w_i X_i$.
4. The equations $\mathcal{L}$ on M described above.
5. Positivity: $M \succeq 0$.

In the remainder of this section we prove the satisfiability and moment bounds parts of Lemma 9.4.2. To prove the lemma we will need a couple of simple facts about SoS proofs.

Fact 9.7.7. *Let $X_1, \ldots, X_m \in \mathbb{R}^d$. Let $v \in \mathbb{R}^d$ have $\|v\| \leqslant 1$. Let $Y_i = X_i + v$. Let $t \in \mathbb{N}$ be even. Suppose there is $C \in \mathbb{R}$ with $C \geqslant 1$ such that for all $s \leqslant t$,*

$$\frac{1}{m}\sum_{i\in[m]} \|X_i\|^s \leqslant C^s$$

Then

$$\vdash_t \frac{1}{m}\sum_{i\in[n]} \left[\langle X_i, u\rangle^t - \langle Y_i, u\rangle^t\right] \leqslant \left(2^t C^{t-1}\|v\|\right)\|u\|^t$$

and similarly for $\frac{1}{m}\sum_{i\in[n]} \left[\langle Y_i, u\rangle^t - \langle X_i, u\rangle^t\right]$.

Proof. Expanding $\langle Y_i, u\rangle^t$, we get

$$\langle Y_i, u\rangle^t = \langle X_i + v, u\rangle^t = \sum_{s\leqslant t} \binom{t}{s} \langle X_i, v\rangle^s \langle v, u\rangle^{t-s} .$$

So,

$$\frac{1}{m}\sum_{i\in[m]} \left[\langle X_i, u\rangle^t - \langle Y_i, u\rangle^t\right] = -\frac{1}{m}\sum_{i\in[m]}\sum_{s<t} \binom{t}{s} \langle X_i, u\rangle^s \langle v, u\rangle^{t-s} .$$

For each term, by Cauchy-Schwarz, $\vdash_t \langle X_i, u\rangle^s \langle v, u\rangle^{t-s} \leqslant \|X_i\|^s \|v\|^{t-s} \cdot \|u\|^t$. Putting these together with the hypothesis on $\frac{1}{n}\|X_i\|^s$ and counting terms finishes the proof. □

Proof of Lemma 9.4.2: Satisfiability. By taking a random subset S if necessary, we assume $|S| = (1-\tau)\alpha n = m$. We describe a solution to the system $\widehat{\mathcal{A}}$. Let w be the 0/1 indicator vector for S. Let $\mu = \frac{1}{m}\sum_{i\in S} w_i X_i$. This satisfies the Boolean-ness, size, and empirical mean axioms.

Describing the assignment to the variables $\{M_{\gamma,\rho}\}$ takes a little more work. Re-indexing and centering, let $Y_1 = X_{i_1} - \mu, \ldots, Y_m = X_{i_m} - \mu$ be centered versions of the samples in S, where $S = \{i_1, \ldots, i_m\}$ and μ remains the empirical mean $\frac{1}{m}\sum_{i\in S} X_i$. First suppose that the following SoS proof exists:

$$\vdash_t \frac{1}{\alpha n}\sum_{i\in S} \langle Y_i, u\rangle^t \leqslant 2\cdot t^{t/2}\|u\|^t .$$

Just substituting definitions, we also obtain

$$\vdash_t \frac{1}{\alpha n}\sum_{i\in[n]} w_i \langle X_i - \mu, u\rangle^t \leqslant 2\cdot t^{t/2}\|u\|^t .$$

where now w and μ are scalars, not variables, and u are the only variables remaining. The existence of this SoS proof means there is a matrix $P \in \mathbb{R}^{d^{t/2}\times d^{t/2}}$ such that $P \succeq 0$ and

$$\langle u^{\otimes t/2}, P u^{\otimes t/2}\rangle = 2t^{t/2}\|u\|^t - \frac{1}{\alpha n}\sum_{i\in[n]} w_i \langle X_i - \mu, u\rangle^t .$$

Let $M_{\gamma,\rho} = P_{\gamma,\rho}$. Then clearly $M \succeq 0$ and M, w, μ together satisfy $\mathcal{L}$.

It remains to show that the first SoS proof exists with high probability for large enough m. Since t is even and $0.1 > \tau > 0$, it is enough to show that

$$\vdash_t \frac{1}{m} \sum_{i \in [S]} \langle Y_i, u \rangle^t \leqslant 1.5 \cdot t^{t/2} \|u\|^t$$

Let $Z_i = X_i - \mu^*$, where μ^* is the true mean of $\mathcal{D}$. Let

$$a(u) = \frac{1}{m} \sum_{i \in S} \left[\langle Z_i, u \rangle^t - \langle Y_i, u \rangle^t \right] \qquad b(u) = \frac{1}{m} \sum_{i \in S} \langle Z_i, u \rangle^t - \mathop{\mathbb{E}}_{Z \sim \mathcal{D} - \mu^*} \langle Z, u \rangle^t .$$

We show that for $d \geqslant 2$,

$$\vdash_t a(u) \leqslant \tfrac{1}{4} \|u\|^t \qquad \vdash_t b(u) \leqslant \tfrac{1}{4} \|u\|^t$$

so long as the following hold

1. (bounded norms) for every $s \leqslant t$ it holds that $\frac{1}{m} \sum_{i \in [m]} \|Z_i\|^s \leqslant s^{100s} d^{s/2}$.
2. (concentration of empirical mean) $\|\mu - \mu^*\| \leqslant d^{-5t}$.
3. (bounded coefficients) For every multiindex θ of degree $|\theta| = t$, one has $\left| \frac{1}{m} \sum_{i \in [m]} Z_i^\theta - \mathbb{E}_{Z \sim \mathcal{D}} Z^\theta \right| \leqslant d^{-10t}$.

We verify in Fact 9.7.8 following this proof that these hold with high probability by standard concentration of measure, for $m \geqslant d^{100t}$ and $\mathcal{D}$ $10t$-explicitly bounded, as assumed. Together with the assumption $\vdash_t \mathbb{E}_{Z \sim \mathcal{D} - \mu^*} \langle Z, u \rangle^t \leqslant t^{t/2} \|u\|^t$, this will conclude the proof.

Starting with $a(u)$, using Fact 9.7.7, it is enough that $2^t C^{t-1} \|v\| \leqslant \frac{1}{4}$, where $v = \mu - \mu^*$ and C is such that $\frac{1}{m} \sum_{i \in [m]} \|Z_i\|^s \leqslant C^s$. By 1 and 2, we can assume $\|v\| \leqslant d^{-5t}$ and $C = t^{100} d^{1/2}$. Then the conclusion follows for $t \geqslant 3$.

We turn to $b(u)$. A typical coefficient of $b(u)$ in the monomial basis—say, the coefficient of u^θ for some multiindiex θ of degree $|\theta| = t$, looks like

$$\frac{1}{m}\sum_{i\in[m]} Y_i^\theta - \mathop{\mathbb{E}}_{Y\sim\mathcal{D}} Y^\theta\,.$$

By assumption this is at most d^{-10t} in magnitude, so the sum of squared coefficients of $b(u)$ is at most d^{-18t}. The bound on $b(u)$ for $d \geqslant 2$. □

Proof of Lemma 9.4.2: Moment bounds. As in the lemma statement, let $f(\mu)$ be a vector of degree-ℓ polynomials in μ. By positivity and Lemma 9.7.2,

$$M(w,\mu) \succcurlyeq 0 \vdash_{O(t\ell)} \langle f(\mu)^{\otimes t/2}, M(w,\mu) f(\mu)^{\otimes t/2}\rangle \succcurlyeq 0\,.$$

Using this in conjunction with the linear equations $\mathcal{L}$,

$$\widehat{\mathcal{A}} \vdash_{O(t\ell)} 2t^{t/2}\|f(\mu)\|^t - \frac{1}{\alpha n}\sum_{i\in[n]} w_i\langle X_i-\mu, f(\mu)\rangle^t \succcurlyeq 0$$

which is what we wanted to show. □

Fact 9.7.8 (Concentration for items 1, 2,3)**.** *Let $d, t \in \mathbb{N}$. Let $\mathcal{D}$ be a mean-zero distribution on $\mathbb{R}^d$ such that $\mathbb{E}\langle Z,u\rangle^s \leqslant s^s\|u\|^s$ for all $s \leqslant 10t$ for every $u \in \mathbb{R}^d$. Then for $t \geqslant 4$ and large enough d and $m \geqslant d^{100t}$, for m independent samples $Z_1,\dots,Z_m \sim \mathcal{D}$,*

1. *(bounded norms) for every $s \leqslant t$ it holds that $\frac{1}{m}\sum_{i\in[m]}\|Z_i\|^s \leqslant s^{100s}d^{s/2}$.*
2. *(concentration of empirical mean) $\left\|\frac{1}{m}\sum_{i\in[m]} Z_i\right\| \leqslant d^{-5t}$.*
3. *(bounded coefficients) For every multiindex θ of degree $|\theta| = t$, one has $\left|\frac{1}{m}\sum_{i\in[m]} Z_i^\theta - \mathbb{E}_{Z\sim\mathcal{D}} Z^\theta\right| \leqslant d^{-10t}$.*

Proof. The proofs are standard applications of central limit theorems, in particular the Berry-Esseen central limit theorem [37], since all the quantities in question

are sums of iid random variables with bounded moments. We will prove only the first statement; the others are similar.

Note that $\frac{1}{m}\sum_{i\in[m]}\|Z_i\|^s$ is a sum of iid random variables. Furthermore, by our moment bound assumption, $\mathbb{E}_{Z\sim\mathcal{D}}\|Z\|^s \leqslant s^{2s}d^{s/2}$. We will apply the Berry-Esseen central limit theorem [37]. The second and third moments $\mathbb{E}(\|Z\|^s - \mathbb{E}\|Z\|^s)^2, \mathbb{E}(\|Z\|^s - \mathbb{E}\|Z\|^s)^3$ are bounded, respectively, as $s^{O(s)}k^s$ and $s^{O(s)}d^{3s/2}$. By Berry-Esseen,

$$\mathbb{P}\left\{\frac{\sqrt{m}}{d^{s/2}}\cdot\frac{1}{m}\sum_{i\in[m]}\|Z_i\|^s > r + \frac{\sqrt{m}}{d^{s/2}}\mathbb{E}\|Z\|^s\right\} \leqslant e^{-\Omega(r^2)} + s^{O(s)}\cdot m^{-1/2}.$$

□

Finally we remark on the polynomial-time algorithm to find a pseudoexpectation satisfying $\widehat{\mathcal{A}}$. As per [34], it is just necessary to ensure that if $x = (w, \mu)$, the polynomials in $\widehat{\mathcal{A}}$ include $\|x\|^2 \leqslant M$ for some large number M. In our case the equation $\|x\|^2 \leqslant (nkm)^{O(1)}$ can be added without changing any arguments.

9.7.4 Modifications for robust estimation

We briefly sketch how the proof of Lemma 9.4.2 may be modified to prove Lemma 9.6.3. The main issue is that $\widehat{\mathcal{A}}$ of Lemma 9.4.2 is satisfiable when there exists an SoS proof

$$\vdash_t \frac{1}{(1-\varepsilon)n}\sum_{i\in[n]} w_i\langle X_i - \mu, u\rangle^t \leqslant 2t^{t/2}\|u\|^t$$

where μ is the empirical mean of X_i such that $w_i = 1$. In the proof of Lemma 9.4.2 we argued that this holds when w is the indicator for a set of iid samples from a

$10t$-explicitly bounded distribution $\mathcal{D}$. However, in the robust setting, w should be taken to be the indicator of the $(1-\varepsilon)n$ good samples remaining from such a set of iid samples after εn samples are removed by the adversary. If $Y_1, \ldots, Y_n$ are the original samples, with empirical mean μ^*, the proof of Lemma 9.4.2 (with minor modifications in constants) says that with high probability,

$$\vdash_t \frac{1}{n} \sum_{i \in [n]} \langle Y_i - \mu^*, u \rangle^t \leqslant 1.1 t^{t/2} \|u\|^t$$

For small-enough ε, this also means that

$$\vdash_t \frac{1}{(1-\varepsilon)n} \sum_{i \text{ good}} \langle X_i - \mu^*, u \rangle^t \leqslant 1.2 t^{t/2} \|u\|^t .$$

This almost implies that $\widehat{\mathcal{A}}$ is satisfiable given the ε-corrupted vectors $X_1, \ldots, X_n$ and parameter $\alpha = (1-\varepsilon)n$, except for that $\mu^* = \frac{1}{n} \sum_{i \in [n]} Y_i$ and we would like to replace it with $\mu = \frac{1}{(1-\varepsilon)n} \sum_{i \text{ good}} X_i$. This can be accomplished by noting that, as argued in Section 9.6, with high probability $\|\mu - \mu^*\| \leqslant O(t \cdot \varepsilon^{1-1/t})$.

9.7.5 Examples of explicitly bounded distributions

In this section, we show that many natural high dimensional distributions are explicitly bounded. Recall that if a univariate distribution X *sub-Gaussian* (with variancy proxy σ) with mean μ then we have the following bound on its even centered moments for $t \geqslant 4$:

$$\mathbb{E}[(X-\mu)^t] \leqslant \sigma^t \left(\frac{t}{2}\right)^{t/2} ,$$

if t is even.

More generally, we will say a univariate distribution is t-bounded with mean μ and variance proxy σ if the following general condition holds for all even

$4 \leqslant s \leqslant t$:

$$\mathbb{E}[(X-\mu)^s] \leqslant \sigma^s \left(\frac{s}{2}\right)^{s/2} .$$

The factor of $1/2$ in this expression is not important and can be ignored upon first reading.

Our main result in this section is that any rotation of products of independent t-bounded distributions with variance proxy $1/2$ is t-explicitly bounded with variance proxy 1:

Lemma 9.7.9. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ so that $\mathcal{D}$ is a rotation of a product distribution $\mathcal{D}'$ where each coordinate of $\mathcal{D}$ is a t-bounded univariate distribution with variance proxy $1/2$. Then $\mathcal{D}$ is t-explicitly bounded (with variance proxy 1).*

Proof. Since the definition of explicitly bounded is clearly rotation invariant, it suffices to show that $\mathcal{D}'$ is t-explicitly bounded. For any vector of indeterminants u, and for any $4 \leqslant s \leqslant t$ even, we have

$$\begin{aligned}
\vdash_s \mathop{\mathbb{E}}_{X\sim\mathcal{D}'} \langle X-\mu, u\rangle^s &= \mathop{\mathbb{E}}_{X\sim\mathcal{D}'} \langle X - \mathop{\mathbb{E}}_{X'\sim\mathcal{D}'} X', u\rangle^s \\
&= \mathop{\mathbb{E}}_{X\sim\mathcal{D}'} \left(\mathop{\mathbb{E}}_{X'} \langle X-X', u\rangle\right)^s \\
&\leqslant \mathop{\mathbb{E}}_{X,X'\sim\mathcal{D}'} \langle X-X', u\rangle^s ,
\end{aligned}$$

where X' is an independent copy of X, and the last line follows from SoS Cauchy-Schwarz. We then expand the resulting polynomial in the monomial basis:

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,X'\sim\mathcal{D}'} \langle X-X', u\rangle^s &= \sum_\alpha u^\alpha \mathop{\mathbb{E}}_{X,X'} (X-X')^\alpha \\
&= \sum_{\alpha \text{ even}} u^\alpha \mathop{\mathbb{E}}_{X,X'} (X-X')^\alpha ,
\end{aligned}$$

since all α with odd monomials disappear since $X - X'$ is a symmetric product distribution. By t-boundedness, all remaining coefficients are at most s^{cs}, from which we deduce

$$\vdash_s \mathop{\mathbb{E}}_{X,X' \sim \mathcal{D}'} \langle X - X', u \rangle^s \leqslant s^{s/2} \sum_{\alpha \text{ even}} u^\alpha = s^{s/2} \|u\|^s \,,$$

which proves that $\mathcal{D}'$ is t-explicitly bounded, as desired. □

As a corollary observe this trivially implies that all Guassians $\mathcal{N}(\mu, \Sigma)$ with $\Sigma \preceq I$ are t-explicitly bounded for all t.

We note that our results are tolerant to constant changes in the variance proxy (just by scaling down). In particular, this implies that our results immediately apply for all rotations of products of t-bounded distributions with a loss of at most 2.

9.8 Omitted Proofs

9.8.1 Sum of squares proofs for matrix positivity

Lemma 9.8.1 (Soundness)**.** *Suppose $\tilde{\mathbb{E}}$ is a degree-$2d$ pseudodistribution which satisfies constraints $\{M_1 \succeq 0, \ldots, M_m \succeq 0\}$, and*

$$\{M_1 \succeq 0, \ldots, M_m \succeq 0\} \vdash_{2d} M \succeq 0 \,.$$

Then $\tilde{\mathbb{E}}$ satisfies $\{M_1 \succeq 0, \ldots, M_m \succeq 0, M \succeq 0\}$.

Proof. By hypothesis, there are r_S^j and B such that

$$M = B^\top \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] B \,.$$

Now, let $T \subseteq [m]$ and p be a polynomial. Let $M' = \otimes_{i \in T} M_i$. Suppose that $\deg\left(p^2 \cdot M \otimes M'\right) \leqslant 2d$. Using the hypothesis on M, we obtain

$$\begin{aligned} p^2 \cdot M \otimes M' &= p^2 \cdot B^\top \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] B \otimes M' \\ &= (B \otimes I)^\top \left[p^2 \cdot \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] \otimes M' \right] (B \otimes I) \,. \end{aligned}$$

Applying $\tilde{\mathbb{E}}$ to the above, note that by hypothesis,

$$\tilde{\mathbb{E}} \left[p^2 \cdot \left[\sum_{S \subseteq [m]} \left(\sum_j (r_S^j(x))(r_S^j(x))^\top \right) \otimes [\otimes_{i \in S} M_i(x)] \right] \otimes M' \right] \succeq 0 \,.$$

The lemma follows by linearity. □

Lemma 9.8.2. *Let $f(x)$ be a degree-ℓ s-vector-valued polynomial in indeterminates x. Let $M(x)$ be a $s \times s$ matrix-valued polynomial of degree ℓ'. Then*

$$\{M \succeq 0\} \vdash_{\ell \ell'} \langle f(x), M(x) f(x) \rangle \geqslant 0 \,.$$

Proof. Let $u \in \mathbb{R}^{s \otimes s}$ have entries $u_{ij} = 1$ if $i = j$ and otherwise $u_{ij} = 0$. Then $\langle f(x), M(x) f(x) \rangle = u^\top (M(x) \otimes f(x) f(x)^\top) u$. □

9.8.2 Omitted Proofs from Section 9.6

Proof of Lemma 9.6.5 We will show that each event (E1)–(E4) holds with probability at least $1 - d^{-8}$. Clearly for d sufficiently large this implies the

desired guarantee. That (E1) and (E2) occur with probability $1 - d^{-8}$ follow from Lemmas 9.6.2 and 9.6.3, respectively. It now suffices to show (E3) and (E4) holds with high probability. Indeed, that (E4) holds with probability $1 - d^{-8}$ follows trivially from the same proof of Lemma 9.4.2 (it is in fact a simpler version of this fact).

Finally, we show that (E3) holds.

By basic concentration arguments (see e.g. [175]), we know that by our choice of n, with probability $1 - d^{-8}$ we have that

$$\left\| \frac{1}{n} \sum_{i \in [n]} X_i - \mu^* \right\| \leqslant \varepsilon \ . \tag{9.8.1}$$

Condition on the event that this and (E4) simultaneously hold. Recall that Y_i for $i = 1, \ldots, n$ are defined so that Y_i are iid and $Y_i = X_i$ for $i \in S_g$. By the triangle inequality, we have

$$\begin{aligned} \left\| \frac{1}{|S_g|} \sum_{i \in S_g} X_i - \mu^* \right\| &\leqslant \frac{n}{|S_g|} \left\| \frac{1}{n} \sum_{i \in [n]} Y_i - \mu^* \right\| + \frac{|S_b|}{|S_g|} \left\| \frac{1}{|S_b|} \sum_{i \in S_b} Y_i - \mu^* \right\| \\ &\overset{(a)}{\leqslant} \frac{\varepsilon}{1 - \varepsilon} + \frac{|S_b|}{|S_g|} \left\| \frac{1}{|S_b|} \sum_{i \in S_b} Y_i - \mu^* \right\| , \end{aligned} \tag{9.8.2}$$

where (a) follows from (9.8.1).

We now bound the second term in the RHS. For any unit vector $u \in \mathbb{R}^d$, by Hölder's inequality,

$$\begin{aligned} &\left\langle \sum_{i \in S_b} (Y_i - \mu^*), u \right\rangle^t \leqslant |S_b|^{t-1} \sum_{i \in S_b} \left\langle (Y_i - \mu^*), u \right\rangle^t \\ &\leqslant |S_b|^{t-1} \sum_{i \in [n]} \left\langle (Y_i - \mu^*), u \right\rangle^t \\ &= |S_b|^{t-1} \left[u^{\otimes t/2} \right]^\top \sum_{i \in [n]} \left[(Y_i - \mu^*)^{\otimes t/2} \right] \left[(Y_i - \mu^*)^{\otimes t/2} \right]^\top \left[u^{\otimes t/2} \right] \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leqslant} |S_b|^{t-1} \cdot n \cdot \left[u^{\otimes t/2}\right]^\top \left(\mathop{\mathbb{E}}_{Y\sim D} \left[(Y-\mu^*)^{\otimes t/2}\right] \left[(Y-\mu^*)^{\otimes t/2}\right]^\top + \delta \cdot \mathrm{Id} \right) \left[(Y-\mu^*)^{\otimes t/2}\right] \\
&= |S_b|^{t-1} \cdot n \cdot \left(\mathop{\mathbb{E}}_{Y\sim D} \left\langle Y-\mu^*, u \right\rangle^t + \delta \right) \\
&\leqslant |S_b|^{t-1} \cdot n \cdot (t^{t/2} + \delta) \\
&\overset{(b)}{\leqslant} 2|S_b|^{t-1} \cdot n \cdot t^{t/2} \,,
\end{aligned}
$$

where (a) follows from (E4), and (b) follows since $\delta \ll t^t$. Hence

$$
\left\| \sum_{i\in S_b} (Y_i - \mu^*) \right\| = \max_{\|u\|=1} \left\langle \sum_{i\in S_b} (Y_i - \mu^*), u \right\rangle \leqslant O(|S_b|^{1-1/t} \cdot n^{1/t} \cdot t^{1/2})
$$

Taking the t-th root on both sides and combining it with (9.8.2) yields

$$
\left\| \frac{1}{|S_g|} \sum_{i\in S_g} X_i - \mu^* \right\| \leqslant \frac{\varepsilon}{1-\varepsilon} + \frac{\varepsilon}{1-\varepsilon} (n/|S_b|)^{-1/t} \cdot t^{1/2} = O(\varepsilon^{1-1/t} \cdot t^{1/2}) \,,
$$

as claimed.

9.9 Chapter Notes

The material in this chapter originally appeared in [90], joint work with Jerry Li. Similar results appeared contemporaneously in [108, 68]. We highlight one result of [108]: that any probability distribution $\mathcal{D}$ which obeys a certain isoperimetric inequality called the *Poincaré inequality* is explicitly bounded for all t. This implies that strongly log-concave distributions are explicitly bounded; this generalizes our result on product and rotations of product distributions.

9.9.1 Related work

Mixture models The literature on mixture models is vast so we cannot attempt a full survey here. The most directly related line of work to our results studies mixtures models under mean-separation conditions, and especially mixtures of Gaussians, where the number k of components of the mixture grows with the dimension d [52, 53, 17, 174]. The culmination of these works is the algorithm of Vempala and Wang, which used spectral dimension reduction to improve on the $d^{1/4}$ separation required by previous works to $k^{1/4}$ in ℓ_2 distance for $k \leqslant d$ spherical Gaussians in d dimensions. Concretely, they show the following:

Theorem 9.9.1 ([174], informal)**.** *There is a constant $C > 0$ and an algorithm with running time* $\mathrm{poly}(k, d)$ *such that for every $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ and $\sigma_1, \ldots, \sigma_k > 0$, satisfying*

$$\|\mu_i - \mu_j\| > C \max(\sigma_i, \sigma_j) k^{1/4} \log^{1/4}(d)$$

with high probability the algorithm produces estimates $\widehat{\mu_1}, \ldots, \widehat{\mu_k}$ with $\|\mu_i - \widehat{\mu_i}\| \leqslant 1/\mathrm{poly}(k)$, given $\mathrm{poly}(k, d)$ *samples from a mixture $\frac{1}{k} \sum_{i \leqslant k} \mathcal{N}(\mu_i, \sigma_i I)$.*

The theorem extends naturally to isotropic log-concave distributions; our main theorem generalizes to distributions with explicitly bounded moments. These families of distributions are not strictly comparable.

Other works have relaxed the requirement that the underlying distributions be Gaussian [113, 4]; to second-moment moment boundedness instead of to log-concavity; these algorithms typically tolerate separation of order $\sqrt{k}$ rather than $k^{1/4}$. Our work can be thought of as a generalization of these algorithms to use boundedness of higher moments. One recent work in this spirit uses SDPs to cluster mixture models under separation assumptions [131]; the authors show

that a standard SDP relaxation of k-means achieves guarantees comparable to previously-known specially-tailored mixture model algorithms.

Information-theoretic sample complexity: Recent work of [158] considers the Gaussian mixtures problem in an information-theoretic setting: they show that there is some constant C so that if the means are pairwise separated by at least $C\sqrt{\log k}$, then the means can be recovered to arbitrary accuracy (given enough samples). They give an efficient algorithm which, warm-started with sufficiently-good estimates of the means, improves the accuracy to δ using $\mathrm{poly}(1/\delta, d, k)$ additional samples. However, their algorithm for providing this warm start requires time exponential in the dimension d. Our algorithm requires somewhat larger separation but runs in polynomial time. Thus by combining the techniques in the spherical Gaussian setting we can estimate the means with ℓ_2 error δ in polynomial time using an extra $\mathrm{poly}(1/\delta, d, k)$ samples, when the separation is at least k^{γ}, for any $\gamma > 0$.

Fixed number of Gaussians in many dimensions: Other works address parameter estimation for mixtures of $k \ll d$ Gaussians (generally $k = O(1)$ and d grows) under weak identifiability assumptions [101, 35, 132, 84]. In these works the only assumptions are that the component Gaussians are statistically distinguishable; the goal is to recover their parameters of the underlying Gaussians. It was shown in [132] that algorithms in this setting provably require $\exp(k)$ samples and running time. The question addressed in our paper is whether this lower bound is avoidable under stronger identifiability assumptions. A related line of work addresses proper learning of mixtures of Gaussians [73, 54, 168, 120], where the goal is to output a mixture of Gaussians which is close to the unknown mixture in total-variation distance, avoiding the $\exp(k)$ parameter-learning sample-complexity

lower bound. These algorithms achieve poly(k, d) sample complexity, but they all require exp(k) running time, and moreover, do not provide any guarantee that the parameters of the distributions output are close to those for the true mixture.

Tensor-decomposition methods: Another line of algorithms focus on settings where the means satisfy algebraic non-degeneracy conditions, which is the case for instance in smoothed analysis settings [94, 14, 76]. These algorithms are typically based on finding a rank-one decomposition of the empirical 3rd or 4th moment tensor of the mixture; they heavily use the special structure of these moments for Gaussian mixtures. One paper we highlight is [39], which also uses much higher moments of the distribution. They show that in the smoothed analysis setting, the ℓth moment tensor of the distribution has algebraic structure which can be algorithmically exploited to recover the means. Their main structural result holds only in the smoothed analysis setting, where samples from a mixture model on perturbed means are available.

In contrast, we do not assume any non-degeneracy conditions and use moment information only about the individual components rather than the full mixture, which always hold under separation conditions. Moreover, our algorithms do not need to know the exact structure of the 3rd or 4th moments. In general, clustering-based algorithms like ours seem more robust to modelling errors than algebraic or tensor-decomposition methods.

Expectation-maximization (EM): EM is the most popular algorithm for Gaussian mixtures in practice, but it is notoriously difficult to analyze theoretically. The works [53, 23, 55, 180] offer some theoretical guarantees for EM, but non-convergence results are a barrier to strong theoretical guarantees [179].

Robust statistics The literature on robust estimation is too large to do justice to here. There has been a long line of work on making algorithms tolerant to error in supervised settings [173, 103], especially for learning halfspaces [162, 105, 21, 67], and for problems such as PCA [45, 46, 116, 181]. See [64] for a more detailed discussion on the relationship between these questions (and others) and the model we consider here.

We consider the classical statistical notion of robustness against corruption, introduced back in the 70's in seminal works of [95, 171, 83]. Even for the mean of a Gaussian distribution, essentially all classical robust estimators are hard in the worst case to compute ([99, 36]). However, a recent flurry of work ([64, 114, 49, 66, 166]) has given new, computationally efficient, nearly optimal robust estimators for the mean and covariance of a high dimensional Gaussian distribution. Given sufficiently-many samples from a sub-Gaussian distribution with identity covariance, where an ε-fraction are arbitrarily corrupted, these algorithms can output mean estimates which achieve error at most $O(\varepsilon\sqrt{\log 1/\varepsilon})$ in ℓ_2, which is information-theoretically optimal up to the $\sqrt{\log 1/\varepsilon}$ factor. However, these mean estimation algorithms heavily rely on knowing that the covariance is equal (or very close) to the identity. When the distribution is a general sub-Gaussian distribution with unknown covariance, the best known error achieved by an efficient algorithm is $O(\varepsilon^{1/2})$ [166, 65]. Under a slightly stronger assumption, our algorithm is able to achieve $O(\varepsilon^{1-1/t})$ error in polynomial time, for arbitrarily large $t \in \mathbb{N}$, and error $O(\varepsilon\sqrt{\log 1/\varepsilon})$ in quasipolynomial time for distributions with $O(\log 1/\varepsilon)$ bounded moments.

CHAPTER 10

SOS TOOLKIT

In this chapter we record a number of useful SoS inequalities, which we have employed in analysis of the algorithms in this part of the thesis.

Lemma 10.0.1 (Pseudo-Cauchy-Schwarz, Function Version, [27])**.** *Let x, y be vector-valued polynomials. Then*

$$\langle x, y\rangle \le \frac{1}{2}(\|x\|^2 + \|y\|^2).$$

Lemma 10.0.2 (Pseudo-Cauchy-Schwarz, pseudoexpectation version)**.** *If $\tilde{\mathbb{E}}$ is a degree-d pseudoexpectation on variables x and $p, q \in \mathbb{R}[x]_{\leqslant d/2}$ then $(\tilde{\mathbb{E}}\, p(x)q(x))^2 \leqslant \tilde{\mathbb{E}}\, p(x) \cdot \tilde{\mathbb{E}}\, q(x)$.*

See [30] for the cleanest proofs.

We will need the following inequality relating $\tilde{\mathbb{E}}\langle x, v_0\rangle^3$ and $\tilde{\mathbb{E}}\langle x, v_0\rangle$ when $\tilde{\mathbb{E}}\langle x, v_0\rangle^3$ is large.

Lemma 10.0.3. *Let $\{x\}$ be a degree-4 pseudo-distribution satisfying $\{\|x\|^2 = 1\}$, and let $v_0 \in \mathbb{R}^n$ be a unit vector. Suppose that $\tilde{\mathbb{E}}\langle x, v_0\rangle^3 \geqslant 1 - \varepsilon$ for some $\varepsilon \geqslant 0$. Then $\tilde{\mathbb{E}}\langle x, v_0\rangle \geqslant 1 - 2\varepsilon$.*

Proof. Let $p(u)$ be the univariate polynomial $p(u) = 1 - 2u^3 + u$. It is easy to check that $p(u) \geqslant 0$ for $u \in [-1, 1]$. It follows from classical results about univariate polynomials that $p(u)$ then can be written as

$$p(u) = s_0(u) + s_1(u)(1 + u) + s_2(u)(1 - u)$$

for some SoS polynomials s_0, s_1, s_2 of degrees at most 2. (See [144], fact 3.2 for a precise statement and attributions.)

Now we consider

$$\tilde{\mathbb{E}}\, p(\langle x, v_0\rangle) \geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(1 + \langle x, v_0\rangle)] + \tilde{\mathbb{E}}[s_2(\langle x, v_0\rangle)(1 - \langle x, v_0\rangle)]\,.$$

We have by Lemma 10.0.1 that $\langle x, v_0\rangle \leq \frac{1}{2}(\|x\|^2 + 1)$ and also that $\langle x, v_0\rangle \geq -\frac{1}{2}(\|x\|^2 + 1)$. Multiplying the latter SoS relation by the SoS polynomial $s_1(\langle x, v_0\rangle)$ and the former by $s_2(\langle x, v_0\rangle)$, we get that

$$\begin{aligned}
\tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(1 + \langle x, v_0\rangle)] &= \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] + \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)\langle x, v_0\rangle] \\
&\geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] - \frac{1}{2}\,\tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(\|x\|^2 + 1)] \\
&\geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] - \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] \\
&\geqslant 0\,,
\end{aligned}$$

where in the second-to-last step we have used the assumption that $\{x\}$ satisfies $\{\|x\|^2 = 1\}$. A similar analysis yields

$$\tilde{\mathbb{E}}[s_2(\langle x, v_0\rangle)(1 - \langle x, v_0\rangle)] \geqslant 0\,.$$

All together, this means that $\tilde{\mathbb{E}}\, p(\langle x, v_0\rangle) \geqslant 0$. Expanding, we get $\tilde{\mathbb{E}}[1 - 2\langle x, v_0\rangle^3 + \langle x, v_0\rangle] \geqslant 0$. Rearranging yields

$$\tilde{\mathbb{E}}\langle x, v_0\rangle \geqslant 2\,\tilde{\mathbb{E}}\langle x, v_0\rangle^3 - 1 \geqslant 2(1 - \varepsilon) - 1 \geqslant 1 - 2\varepsilon\,. \qquad \square$$

We will need a bound on the pseudo-expectation of a degree-3 polynomial in terms of the operator norm of its coefficient matrix.

Lemma 10.0.4. *Let $\{x\}$ be a degree-4 pseudo-distribution. Let $M \in \mathbb{R}^{n^2 \times n}$. Then $\tilde{\mathbb{E}}\langle x^{\otimes 2}, Mx\rangle \leqslant \|M\|(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4}$.*

Proof. We begin by expanding in the monomial basis and using pseudo-Cauchy-Schwarz:

$$\tilde{\mathbb{E}}\langle x^{\otimes 2}, Mx\rangle = \tilde{\mathbb{E}} \sum_{ijk} M_{(j,k),i} x_i x_j x_k$$

$$
\begin{aligned}
&= \tilde{\mathbb{E}} \sum_i x_i \sum_{jk} M_{(j,k),i} x_j x_k \\
&\leqslant (\tilde{\mathbb{E}} \|x\|^2)^{1/2} \left[\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \right]^{1/2} \\
&\leqslant (\tilde{\mathbb{E}} \|x\|^4)^{1/4} \left[\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \right]^{1/2}
\end{aligned}
$$

We observe that MM^T is a matrix representation of $\sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2$. We know $MM^T \preceq \|M\|^2 \mathrm{Id}$, so

$$
\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \leqslant \|M\|^2 \tilde{\mathbb{E}} \|x\|^4 .
$$

Putting it together, we get $\tilde{\mathbb{E}} \langle x^{\otimes 2}, Mx \rangle \leqslant \|M\| (\tilde{\mathbb{E}} \|x\|^4)^{3/4}$ as desired. □

Fact 10.0.5. *Let x, y be n-length vectors of indeterminates. Then*

$$
\vdash_2 \|x + y\|^2 \leqslant 2\|x\|^2 + 2\|y\|^2 .
$$

Proof. The sum of squares proof of Cauchy-Schwarz implies that $\|x\|^2 + \|y\|^2 - 2\langle x, y \rangle$ is a sum of squares. Now we just expand

$$
\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \preceq 2(\|x\|^2 + \|y\|^2) .
$$

□

Fact 10.0.6. *Let $P(x) \in \mathbb{R}[x]_\ell$ be a homogeneous degree ℓ polynomial in indeterminates $x = x_1, \ldots, x_n$. Suppose that the coefficients of P are bounded in 2-norm:*

$$
\sum_{\alpha \subseteq [n]} \widehat{P}(\alpha)^2 \leqslant C .
$$

(Here $\widehat{P}(\alpha)$ are scalars such that $P(x) = \sum_\alpha \widehat{P}(\alpha) x^\alpha$.) Let $a, b \in \mathbb{N}$ be integers such that $a + b = \ell$. Then

$$\vdash_{\max(2a,2b)} P(x) \leqslant \sqrt{C}(\|x\|^{2a} + \|x\|^{2b}) .$$

Proof. Let M be a matrix whose rows and columns are indexed by multisets $S \subseteq [n]$ of sizes a and b. Thus M has four blocks: an (a, a) block, an (a, b) block, a (b, a) block, and a (b, b) block. In the (a, b) and (b, a) blocks, put matrices M_{ab}, M_{ba} such that $\langle x^{\otimes a}, M_{ab} x^{\otimes b} \rangle = \frac{1}{2} . P(x)$. In the (a, a) and (b, b) blocks, put $\sqrt{C} \cdot I$. Then, letting $z = (x^{\otimes a}, x^{\otimes b})$, we get $\langle z, Mz \rangle = \sqrt{C}(\|x\|^{2a} + \|x\|^{2b}) - P(x)$. Note that $\|M_{ab}\| \leqslant \sqrt{C}$ by hypothesis, so $M \succeq 0$, which completes the proof. □

Fact 10.0.7. *Let $u = (u_1, \ldots, u_k)$ be a vector of indeterminantes. Let D be sub-Gaussian with variancy proxy 1. Let $t \geqslant 0$ be an integer. Then we have*

$$\vdash_{2t} \mathop{\mathbb{E}}_{X \sim D} \langle X, u \rangle^{2t} \leqslant (2t)! \cdot \|u\|^{2t}$$

$$\vdash_{2t} \mathop{\mathbb{E}}_{X \sim D} \langle X, u \rangle^{2t} \geqslant -(2t)! \cdot \|u\|^{2t} .$$

Proof. Expand the polynomial in question. We have

$$\mathop{\mathbb{E}}_{X \sim D} \langle X, u \rangle^{2t} = \mathop{\mathbb{E}}_{X \sim D} \sum_\beta u^\beta \, \mathbb{E}[X^\beta] .$$

Let β range over $[k]^{2t}$

$$\vdash_{2t} \sum_\beta u^{2\beta} \, \mathbb{E}\, X^{2\beta} \leqslant (2t)! \sum_{\beta \text{ even}} u^\beta \leqslant \|u\|_2^{2t} .$$

where we have used upper bounds on the Gaussian moments $\mathbb{E}\, X^{2\beta}$ and that every term is a square in u. □

Fact 10.0.8 (SoS Hölder)**.** *Let $w_1, \ldots, w_n$ and $x_1, \ldots, x_n$ be indeterminates. Let $q \in \mathbb{N}$ be a power of 2. Then*

$$\{w_i^2 = w_i \, \forall i \in [n]\} \vdash_{O(q)} \left(\sum_{i \leqslant n} w_i x_i \right)^q \leqslant \left(\sum_{i \leqslant n} w_i \right)^{q-1} \cdot \left(\sum_{i \leqslant n} x_i^q \right)$$

and

$$\{w_i^2 = w_i \,\forall i \in [n]\} \vdash_{O(q)} \left(\sum_{i\leqslant n} w_i x_i\right)^q \leqslant \left(\sum_{i\leqslant n} w_i\right)^{q-1} \cdot \left(\sum_{i\leqslant n} w_i \cdot x_i^q\right).$$

Proof. We will only prove the first inequality. The second inequality follows since $w_i^2 = w_i \vdash_2 w_i x_i = w_i \cdot (w_i x_i)$, applying the first inequality, and observing that $w_i^2 = w_i \vdash_q w_i^q = w_i$.

Applying Cauchy-Schwarz (Lemma 5.0.6) and the axioms, we obtain to start that for any even number t,

$$\begin{aligned}\{w_i^2 = w_i \,\forall i \in [n]\} \vdash_{O(t)} \left[\left(\sum_{i\leqslant n} w_i x_i\right)^2\right]^{t/2} &= \left[\left(\sum_{i\leqslant n} w_i^2 x_i\right)^2\right]^{t/2} \\ &\leqslant \left[\left(\sum_{i\leqslant n} w_i^2\right)\left(\sum_{i\leqslant n} w_i^2 x_i^2\right)\right]^{t/2} = \left(\sum_{i\leqslant n} w_i\right)^{t/2}\left(\sum_{i\leqslant n} w_i x_i^2\right)^{t/2}.\end{aligned}$$

It follows by indution that

$$\{w_i^2 = w_i \,\forall i \in [n]\} \vdash_{O(t)} \left[\left(\sum_{i\leqslant n} w_i x_i\right)\right]^q \leqslant \left(\sum_{i\leqslant n} w_i\right)^{q-2}\left(\sum_{i\leqslant n} w_i x_i^{q/2}\right)^2.$$

Applying Lemma 5.0.6 one more time to get $\left(\sum_{i\leqslant n} w_i x_i^{q/2}\right) \leqslant \left(\sum_{i\leqslant n} w_i^2\right)\left(\sum_{i\leqslant n} x_i^q\right)$ and then the axioms $w_i^2 = w_i$ completes the proof. □

Part II

Pseudocalibration

CHAPTER 11

SOS LOWER BOUNDS FROM PSEUDOCALIBRATION: PLANTED CLIQUE AND RELATED PROBLEMS

In this chapter we prove a nearly-tight SoS lower bound for the planted clique problem. The proof introduces the pseudocalibration technique, which leverages failure of D-simple statistics to distinguish n-node graphs with planted $n^{1/2-\varepsilon}$-cliques from graphs from $G(n, \frac{1}{2})$. We discuss some prior techniques for SoS lower bounds for planted clique and the fundamental roadblocks they faced which prevented them from proving the kind of nearly-tight lower bound we present here. We will see that pseudocalibration overcomes these challenges.

11.1 Main Results

The main theorem in this chapter is the following SoS lower bound for planted clique.

Theorem 11.1.1. *There is a constant $c > 0$ such for every $d = d(n) \in \mathbb{N}$, with probability $1 - o(1)$ over G sampled from $G(n, 1/2)$ there is a degree-d pseudodistribution $\tilde{\mathbb{E}}_G$ satisfying $\{x_i^2 = x_i\}_{i\in[n]}$ and $\{x_i x_j = 0\}_{i\not\sim j \text{ in } G}$ such that $\tilde{\mathbb{E}}_G \sum_{i\in[n]} x_i \geqslant n^{1/2-c(d/\log n)^{1/2}}$.*

Theorem 11.1.1 says that with high probability over $G \sim G(n, 1/2)$ there is a degree-d pseudodistribution which satisfies constraints as thought it is supported on G-cliques of expected size $n^{1/2-c(d/\log n)^{1/2}}$, even though the largest clique in G has size $(2 + o(1)) \log n$. This means that for every constant $\varepsilon > 0$, SoS degree $\Omega(\log n)$ is required to certify that the maximum clique in a random graph G from $G(n, 1/2)$ has size less than $n^{1/2-\varepsilon}$.

It also follows, by the usual duality arguments, that if $C \ll n^{1/2} - c(d/\log n)^{1/2}$, then with high probability over $G \sim G(n, 1/2)$ every set of polynomials $q_{ij}(x)$ for $i \not\sim j$ in G and $r_i(x)$ satisfying $C - \sum_{i\in[n]} x_i =_{\{0,1\}^n} \sum_{i\not\sim j} x_i x_j q_{ij}(x) + \sum_{i\in[m]} r_i(x)^2$ have degree at least d. (The notation $=_{\{0,1\}^n}$ means that the left and right hand sides are equal when evaluated at any $x \in \{0,1\}^n$.)

Near-tightness of this theorem follows from the following result of Feige and Krauthgamer [71].

Theorem 11.1.2 ([71]). *There is a constant $c > 0$ such that for every $d = d(n) \in \mathbb{N}$, with probability $1 - o(1)$ over G from $G(n, 1/2)$ every degree-d pseudodistribution $\tilde{\mathbb{E}}_G$ satisfying $\{x_i^2 = x_i\}_{i\in[n]}$ and $\{x_i x_j = 0\}_{i\not\sim j \text{ in } G}$ has $\tilde{\mathbb{E}}_G \sum_{i\in[n]} x_i \leqslant n^{1/2 - cd/\log n}$.*

Thus, our main result is tight up to the exponent $1/2$ on the term $d/\log n$. It is an interesting (but quite technical) open problem to remove this remaining gap between upper and lower bounds.

11.1.1 Extensions to component analysis problems

The techniques which prove Theorem 11.1.1 extend to (at least) two other important refutation problems, in the spiked tensor model and the sparse spiked matrix model (a.k.a. sparse principal component analysis, or sparse PCA). We have already addressed lower bounds for the spiked tensor model in Chapter 6.

The refutation version of sparse principal component analysis asks to certify upper bounds on the maximum of the quadratic form of a random matrix over sparse vectors. By standard concentration arguments, if M has standard Gaussian entries, then $\max_{\|x\|=1, |x|_0 \leqslant k} \langle x, Mx\rangle \approx \sqrt{k \log n}$, where $|x|_0$ denotes the number

of nonzero entries of x. Polynomial-time algorithms are known which certify (with high probability) that $\max_{\|x\|=1,|x|_0\leqslant k}\langle x, Mx\rangle \leqslant \tilde{O}(\min(k,\sqrt{n}))$ [62]. We show that improving by any polynomial factor in k or n on the guarantees of those algorithms requires requires *subexponential* time.

The subexponential lower bound we prove here is beyond the reach of previous approaches to hardness of sparse PCA, which typically rely on reduction from the planted clique problem: because planted clique admits quasipolynomial time algorithms, those techniques can only show quasipolynomial hardness [38].

Further context for the following theorem is given in Section 11.7, where we also outline the proof. (The details are very similar to the proof of Theorem 11.1.1; they can be found in the full version of [88].)

Theorem 11.1.3. *If $A \in \mathbb{R}^{n\times n}$, let*

$$SoS_{d,k}(A) = \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle x, Ax\rangle \text{ s.t. } \tilde{\mathbb{E}} \text{ is degree } d \text{ and satisfies } \left\{x_i^3 = x_i, \|x\|^2 = k\right\}.$$

There are absolute constants $c, \varepsilon^* > 0$ so that for every $\rho \in (0,1)$ and $\varepsilon \in (0,\varepsilon^*)$, if $k = n^\rho$, then for $d \leqslant n^{c\cdot\varepsilon}$,

$$\mathbb{P}_{A\sim\{\pm1\}^{\binom{n}{2}}}\{SoS_{d,k}(A) \geqslant \min(n^{1/2-\varepsilon}k, n^{\rho-\varepsilon}k)\} \geqslant 1 - o(1)$$

and

$$\mathbb{E}_{A\sim\{\pm1\}^{\binom{n}{2}}} SoS_{d,k}(A) \geqslant \min(n^{1/2-\varepsilon}k, n^{\rho-\varepsilon}k).$$

Furthermore, the latter is true also if A is symmetric with iid entries from $\mathcal{N}(0,1)$.

11.2 Preliminaries

11.2.1 General Notation

- We use small Greek letters indicate constants/parameters.
- $\mathcal{P}_d^n$ denotes the linear space of all *multilinear* polynomials of degree at most d on $\{0,1\}^n$.
- We write $_Q$ for any event Q to be the 0-1 indicator of whether Q happens.
- For a subset $T \subseteq \binom{[n]}{2}$ of edges of a graph on vertex set $[n]$, we write $\mathcal{V}(T) \subseteq [n]$ to denote the vertices that have at least one edge incident on them in T.
- For a matrix $Q \in \mathbb{R}^{N \times N}$, $\|Q\|$ denotes its spectral norm (or the largest singular value) and $\|Q\|_F = \sqrt{\sum_{x,y \in [N]} Q(x,y)^2}$ denotes its Frobenius norm.
- For a graph G, let $C_q = C_q(G) = \{I \subseteq [n] : I \text{ is a } q\text{-clique in } G\}$, and let $C_{\leqslant q} = \bigcup_{q' \leqslant q} C_{d'}$. Let $C(G) = C_{\leqslant \infty}$ be the collection of all cliques in G. We count the empty set and all singletons as cliques.
- We write $\mathcal{G}(n, \frac{1}{2})$ to denote the distribution on graphs on the vertex set $[n]$ where each edge is included with probability $1/2$ independently of others.
- We say that an event E with respect to the probability distribution $\mathcal{G}(n, \frac{1}{2})$ happens *with high probability (w.h.p.)* if $\mathbb{P}[E] \geqslant 1 - \Omega(1)/n^{10 \log n}$ for large enough n.
- We write $f(n) \ll g(n)$ to mean that for every constant c there is an n_0 such that if $n \geqslant n_0$, $f(n) \leqslant Cg(n)$.

11.2.2 Graphs

We identify a graph G with its $\{-1,1\}$ adjacency matrix and write $G_e \in \{-1,1\}$ for the $\{-1,1\}$-indicator of whether $e \in [n]\times[n]$ is an edge (indicated by $G_e = +1$) in the graph G or not. When $G \sim \mathcal{G}(n, \frac{1}{2})$, G_e are independent $\{-1,1\}$-random variables.

A *graph function* is a real-valued function of the variables $G_e \in \{-1,1\}$ for $e \in \binom{[n]}{2}$. For graphs $G^1, G^2, \ldots, G^k$ on the vertex set $[n]$, we define $\Delta(G^1, G^2, \ldots, G^k)$ to be the graph G satisfying $G_e = \Pi_{i\leqslant k} G^i_e$.

Definition 11.2.1 (Vertex Separator)**.** For a graph G on $[n]$ and vertex sets $I, J \subseteq [n]$, a set of vertices $S \subseteq [n]$ is said to be a *minimal vertex separator* if S is a set of smallest possible size such that every path between I and J in G passes through some vertex of S.

Often, I and J will be allowed to intersect in which case any vertex separator must contain $I \cap J$.

Fact 11.2.2 (Menger's Theorem)**.** *For a graph G on $[n]$ and two subsets of vertices $I, J \subseteq [n]$, the maximum number of vertex disjoint paths between I and J in G is equal to the size of any minimal vertex separator between I and J in G.*

11.2.3 Fourier Analysis

Any graph function $f : G \to \mathbb{R}$ can be represented as a Fourier polynomial in the variables G_e:

$$f(G) = \sum_{W\subseteq\binom{[n]}{2}} \widehat{f}(W)\chi_W(G),$$

where $\chi_W(G)$ is the *parity* function on edges in W:

$$\chi_W(G) = \Pi_{e \in W} G_e.$$

The parity function χ_W are an orthonormal basis for functions on G under the inner product defined by $\langle f, h \rangle = \mathbb{E}_{G \sim G(n,\frac{1}{2})}[f(G)h(G)]$ for any graph functions f and h.

The following fact is easy to verify:

Fact 11.2.3. *Let G be a graph on n described by the vector $G \in \{-1,1\}^{\binom{n}{2}}$. For any subset $S \subseteq [n]$ of the vertices, we have the identity:*

$$\sum_{W \subseteq \binom{S}{2}} \chi_W(G) = \begin{cases} 2^{\binom{|S|}{2}} & \text{if } S \text{ is a clique in } G, \\ 0 & \text{otherwise.} \end{cases}$$

11.3 Definition of Pseudocalibrated $\tilde{\mathbb{E}}$ and Proof of Theorem 11.1.1

We now define our pseudo-distribution operator $\tilde{\mathbb{E}}_G$. Following the pseudo-calibration recipe, it is a well-chosen projection of a planted distribution to low-degree functions. In order to satisfy the constraints $x_i x_j = 0$ for $i \nsim j$ in G, we choose for each monomial x_S a slightly different set of low degree graph functions to project to when defining $\tilde{\mathbb{E}}_G[x_S]$.

Important Parameters The following parameters will be fixed for the rest of the paper.

- $\varepsilon \in (0, 1/2)$, which determines the size $\omega = n^{1/2-\varepsilon}$ of the planted clique.
- $d = d(n) \in \mathbb{N}$, the degree of the SoS relaxation against which we prove a lower bound.
- $\tau = \tau(n) \in \mathbb{N}$, the degree of our pseudoexpectation $\tilde{\mathbb{E}}$ as a function of $G \sim G(n, 1/2)$.

We always assume that $Cd/\varepsilon \leqslant \tau \leqslant (\varepsilon/C)\log n$ and $\varepsilon \geqslant C \log\log n/\log n$ for a sufficiently-large constant C. Eventually we will set $d = (\varepsilon/C)^2 \log n$, (this yields the parameters stated in Theorem Theorem 11.1.1, since then $n^{1/2-\varepsilon} = n^{1/2-\Omega(d/\log n)^{1/2}}$), which implies that $\varepsilon \gg \log\log n/\log n$.

11.3.1 Definition of $\tilde{\mathbb{E}}$

Since the pseudoexpectation $\tilde{\mathbb{E}}$ which we need to produce to prove Theorem 11.1.1 will satisfy $\{x_i^2 = x_i\}$, we just need to specify its *multilinear moments*: $\tilde{\mathbb{E}}[x_I]$ for $I \subseteq [n]$ and $|I| \leqslant d$. The quantity $\tilde{\mathbb{E}}[x_I]$ is a function of G_e for $e \in \binom{[n]}{2}$ and so can be written as a polynomial in G_e with coefficients $\widehat{\tilde{\mathbb{E}}[x_S]}(T)$ for each $T \subseteq \binom{[n]}{2}$. We will obtain an explicit expression for these Fourier coefficients as the low-degree Fourier coefficients of a function associated to the planted distribution $G(n, 1/2, \omega)$.

Definition 11.3.1 ($\tilde{\mathbb{E}}$ of degree d, clique-size ω, truncation τ)**.** Let $S \subseteq [n]$ be a set of vertices of size $|S| \leqslant d$. Let $T \subseteq \binom{[n]}{2}$ be a set of edges. Let $\chi_T = \prod_{e \in T} G_e$. Let

$$\widehat{\tilde{\mathbb{E}}[x_S]}(T) = \begin{cases} \mathbb{E}_{(G,x)\sim G(n,1/2,\omega)}[\chi_T(G) x_S] & \text{if } |\mathcal{V}(T) \cup S| \leqslant \tau \\ 0 & \text{otherwise}. \end{cases}$$

As usual, $\tilde{\mathbb{E}}[x_S] = \sum_{T \subseteq \binom{[n]}{2}} \widehat{\tilde{\mathbb{E}}[x_S]}(T) \cdot \chi_T(G)$.

By definition, $\tilde{\mathbb{E}}[x_S]$ is the projection to $\mathrm{Span}\{\chi_T : |V(T) \cup S| \leqslant \tau\}$ of the function $G \mapsto \frac{\mathbb{P}_{G(n,1/2,\omega)}(G)}{\mathbb{P}_{G(n,1/2)}(G)} \cdot \mathbb{E}_{x \sim G(n,1/2,\omega)}[x_S \mid G]$ which maps a graph G to the likelihood ratio G times the conditional probability that x_S is in a planted clique in G, according to the planted distribution.

The Fourier coefficients in the definition can in fact be explicitly computed easily (in fact we have seen this computation before, in Chapter 2):

Lemma 11.3.2. *Let $T \subseteq \binom{[n]}{2}$, $S \subseteq [n]$ and $\mathcal{V}(T) \subseteq [n]$ be the vertices incident to edges in T. Then*

$$\mathop{\mathbb{E}}_{(H,x) \sim G(n,1/2,\omega)} [\chi_T \cdot x_S] = \left(\tfrac{\omega}{n}\right)^{|\mathcal{V}(T) \cup S|}.$$

Proof. Throughout this proof, we suppress explicit notation for the underlying random variable which is $(H, x) \sim G(n, \frac{1}{2}, \omega)$. We claim that $\mathbb{E}[\chi_T \cdot x_S] = \mathbb{P}[x_{\mathcal{V}(T) \cup S} = 1]$. To see this, note that

$$\begin{aligned}\mathbb{E}[\chi_T \cdot x_S] = \mathbb{P}[x_{\mathcal{V}(T)\cup S} = 1] \cdot \mathbb{E}[\chi_T \cdot x_S \mid x_{\mathcal{V}(T)\cup S} = 0] \\ + (1 - \mathbb{P}[x_{\mathcal{V}(T)\cup S} = 1]) \cdot \mathbb{E}[\chi_T \cdot x_S \mid x_{\mathcal{V}(T)\cup S} = 0]. \quad (11.3.1)\end{aligned}$$

We note that the second term above is 0. It's easy to see if $x_S = 0$. Otherwise, $x_{\mathcal{V}(T)} = 0$, and there is an edge $e \in T$ but not contained in the clique x. Thus,

$$\mathbb{E}[\chi_e \chi_{T \setminus e} \cdot x_S \mid x_{\mathcal{V}(T)\cup S} = 0] = 0\,.$$

If $x_{\mathcal{V}(T)\cup S} = 1$ then $\chi_T = 1$, so $\mathbb{E}[\chi_T \cdot x_S \mid x_{\mathcal{V}(T)\cup S} = 1] = 1$. By a simple computation,

$$\mathbb{P}[x_{V(T)\cup S} = 1] = \left(\tfrac{\omega}{n}\right)^{|\mathcal{V}(T)\cup S|}. \qquad \square$$

11.3.2 $\tilde{\mathbb{E}}$ Satisfies Constraints

We now show that the $\tilde{\mathbb{E}}$ defined in the previous section satisfies all the necessary linear constraints. That is, 1) $\tilde{\mathbb{E}}[1] \approx 1$, 2) $\tilde{\mathbb{E}}[\sum_{i\in[n]} x_i] \approx \omega$, and 3) $\tilde{\mathbb{E}}[x_S] = 0$ for every $S \subseteq [n]$ which is not a clique in G.

We analyze $\tilde{\mathbb{E}}[1]$ and $\tilde{\mathbb{E}}[\sum_{i\in[n]} x_i]$ in the next lemma and include a proof based on the moment method in Section 11.6.1.

Lemma 11.3.3. *With high probability,* $\tilde{\mathbb{E}}[1] = 1 \pm n^{-\Omega(\varepsilon)}$ *and* $\tilde{\mathbb{E}}[\sum_{i\in[n]} x_i] = \omega \cdot (1 \pm n^{-\Omega(\varepsilon)})$.

The next lemma shows that $\tilde{\mathbb{E}}[x_S] = 0$.

Lemma 11.3.4. *With probability* 1, *if* $S \subseteq [n]$ *of size at most* d *is not a clique in* G, *then* $\tilde{\mathbb{E}}[x_S] = 0$.

Proof. Let $S \subseteq [n]$ have size at most d. Recall that $_{S \text{ is a clique in } G} = 2^{-\binom{|S|}{2}} \sum_{T\subseteq\binom{S}{2}} \chi_T$. Becasue the Fourier expansion of $\tilde{\mathbb{E}}[x_S]$ is truncated using the threshold $|\mathcal{V}(T) \cup S| \leqslant \tau$, two Fourier characters $\chi_T, \chi_{T'}$ have the same coefficient in $\tilde{\mathbb{E}}[x_S]$ if $T \oplus T' \subseteq \binom{S}{2}$. So we can factor $\tilde{\mathbb{E}}[x_S] =_{S \text{ is a clique in } G} \cdot f_S(G)$ for some function f_S. □

11.3.3 Proof of Main Theorem

Our main technical claim is that $\tilde{\mathbb{E}} = \tilde{\mathbb{E}}_G$ is (approximately) PSD. That is:

Lemma 11.3.5. *With high probability over* G *from* $G(n, 1/2)$, *every* $p \in \mathcal{P}_d$ *satisfies,*

$$\tilde{\mathbb{E}}_G[p(x)^2] \geqslant 0$$

It is easy to complete the proof of Theorem 11.1.1 now:

Proof of Theorem Theorem 11.1.1. By Lemma 11.3.3, Lemma 11.3.4, and Lemma 11.3.5, there is a universal C so that if $Cd/\varepsilon \leqslant \tau \leqslant (1/C)\varepsilon \log n$, (by a union bound) with high probability the following all hold:

1. $\tilde{\mathbb{E}}[1] = 1 \pm n^{-\Omega(\varepsilon)}$.
2. $\tilde{\mathbb{E}}[x_S] = 0$ for every S of size at most d not a clique in G.
3. $\tilde{\mathbb{E}}[\sum_i x_i] \geqslant (1 - n^{-\Omega(\varepsilon)})\omega$.
4. $\tilde{\mathbb{E}}[p(x)^2] \geqslant 0$ for every $p \in \mathcal{P}_d$.

Thus, choose $\varepsilon = (C^2 d/\log n)^{1/2}$ and $\tau = (1/C)\varepsilon \log n$. The operator given by $\tilde{\mathbb{E}}^*[p(x)] = \tilde{\mathbb{E}}[p(x)]/\tilde{\mathbb{E}}[1]$ is a valid degree-d pseudo-distribution with $\tilde{\mathbb{E}}^*[\sum_i x_i] \geqslant \Omega(n^{1/2-\Theta(d/\log n)^{1/2}})$ as desired. □

11.3.4 Proof Plan for Lemma 11.3.5

As is standard, we can reduce Lemma 11.3.5 to showing that the associated *moment matrix*, is positive semidefinite.

Definition 11.3.6 (Moment Matrix)**.** Let $\mathcal{M} \in \mathbb{R}^{\binom{[n]}{\leqslant d} \times \binom{[n]}{\leqslant d}}$ be given by $\mathcal{M}(I, J) = \tilde{\mathbb{E}}[x_I x_J]$.

Thus, Lemma 11.3.5 is equivalent to showing:

Lemma 11.3.7. *With high probability,* $\mathcal{M} \succeq 0$.

The proof of Lemma 11.3.7 is the most technical in this thesis, and will require significant work. At a high level our plan involves first getting an approximate factorization of the moment matrix $\mathcal{M} = \mathcal{L}\mathcal{Q}_0\mathcal{L}^\top + \mathit{error}$ for appropriately defined matrices $\mathcal{L}$ and $\mathcal{Q}_0$. This step is the key technical part of the proof – given such a factorization, our task reduces to showing that $\mathcal{Q}_0$ and $\mathcal{L}\mathcal{L}^\top$ have large enough positive eigenvalues to compensate for the error.

The first approximate factorization step will occupy us in Section 11.4. The technical work in second step involves showing upper bounds on the spectral norms of appropriately defined pieces of $\mathcal{Q}_0$ and is the content of Section 11.5.

11.4 Approximate Factorization of the Moment Matrix

11.4.1 Ribbons and Vertex Separators

In this section we get set up for the first step in the proof of Lemma 11.3.7 by setting up some definitions. *Ribbons* will play a crucial role in our analysis:

Definition 11.4.1 (Ribbon)**.** An (I, J)-ribbon $\mathcal{R}$ is a graph with edge set $W_{\mathcal{R}} \subseteq \binom{[n]}{2}$ and vertex set $V_{\mathcal{R}} \supseteq \mathcal{V}(W_{\mathcal{R}}) \cup I \cup J$, for two specially identified subsets $I, J \subseteq [n]$, each of size at most d, called the *left* and the *right ends*, respectively. We sometimes write $\mathcal{V}(\mathcal{R}) \stackrel{\text{def}}{=} V_{\mathcal{R}}$ and call $|\mathcal{V}(\mathcal{R})|$ the *size* of $\mathcal{R}$. Also, we write $\chi_{\mathcal{R}}$ for the monomial $\chi_{W_{\mathcal{R}}}$ where $W_{\mathcal{R}}$ is the edge set of the ribbon $\mathcal{R}$.

In our analysis, (I, J)-ribbons arise as the terms in the Fourier decomposition of the entry $\mathcal{M}(I, J)$ in the moment matrix. It is important to emphasize that the subsets I and J in an (I, J)-ribbon are allowed to intersect. Also $\mathcal{V}(\mathcal{R})$ can

contain vertices that are not in $\mathcal{V}(W_{\mathcal{R}})$ if there are isolated vertices in the ribbon. Ultimately, we will want to partition a ribbon into three subribbons in such a way that we can express the moment matrix as the sum of positive semidefinite matrices, and some error terms. Our partitioning will be based on minimum vertex separators.

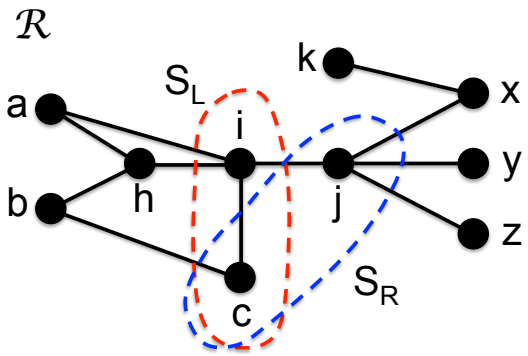
Definition 11.4.2 (Vertex Separator)**.** For an (I, J)-ribbon $\mathcal{R}$ with edge set $W_{\mathcal{R}}$, a subset $Q \subseteq \mathcal{V}(\mathcal{R})$ of vertices is a *vertex separator* if Q separates I and J in $W_{\mathcal{R}}$. A vertex separator is *minimum* if there are no other vertex separators with strictly fewer vertices. The *separator size* of $\mathcal{R}$ is the cardinality of any minimum vertex separator of $\mathcal{R}$.

The following elementary lemma establishes that a ribbon has a unique *leftmost* and *rightmost* vertex separator of minimum size. We defer its proof to Section 11.6.2.

Lemma 11.4.3 (Leftmost/Rightmost Vertex Separator)**.** *Let $\mathcal{R}$ be an (I, J)-ribbon. There is a unique minimum vertex separator S of $\mathcal{R}$ such that S separates I and Q for any vertex separator Q of $\mathcal{R}$. We call S the* leftmost *separator in $\mathcal{R}$. We define the* rightmost *separator analogously and we denote them by $S_L(\mathcal{R})$ and $S_R(\mathcal{R})$ respectively.*

We illustrate the notion of a leftmost and rightmost vertex separator in the example below.

Let $I = \{a, b, c\}$ and let $J = \{c, x, y, z\}$. The maximum number of vertex disjoint paths from I to J is 2 — for example, we could take the path $\{c\}$ and the path $\{b, h, i, j, z\}$. The leftmost and rightmost separators are $S_L = \{c, i\}$ and $S_R = \{c, j\}$ respectively. This example illustrates an important point that when I and J intersect, S_L and S_R must both contain $I \cap J$.



11.4.2 Factorization of Monomials

Our factorization of $\mathcal{M}$ will rely on an iterative argument for grouping and factoring the Fourier characters in the decomposition of $\mathcal{M}(I, J)$.

Definition 11.4.4 (Canonical Factorization)**.** Let $\mathcal{R}$ be an (I, J)-ribbon with edge set $W_{\mathcal{R}}$ and vertex set $V_{\mathcal{R}}$. Let V_ℓ be the vertices reachable from I without passing through $S_L(\mathcal{R})$, and similarly for V_r, and let $V_m = V_{\mathcal{R}} \setminus (V_\ell \cup V_r)$. Let $W_\ell \subseteq W_{\mathcal{R}}$ be given by

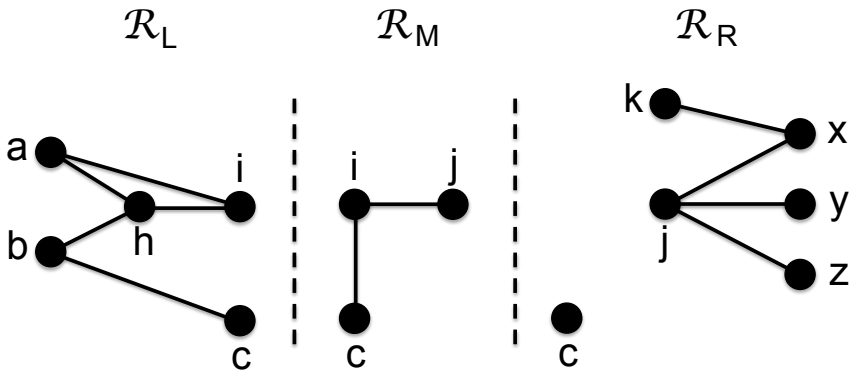
$$W_\ell = \{(u, v) \in W_{\mathcal{R}} : u \in V_\ell \text{ and } v \in V_\ell \cup S_L\}$$

and similarly for W_r. Finally, let $W_m = W_{\mathcal{R}} \setminus (W_\ell \cup W_r)$.

Let $\mathcal{R}_\ell$ be the $(I, S_L(\mathcal{R}))$-ribbon with vertex set $V_\ell \cup S_L(\mathcal{R})$ and edge set W_ℓ and similarly for $\mathcal{R}_r$. Let $\mathcal{R}_m$ be the $(S_L(\mathcal{R}), S_R(\mathcal{R}))$-ribbon with vertex set V_m and edge set W_m . The triple $(\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r)$ is the *canonical factorization* of $\mathcal{R}$.

Some facts about the canonical factorization are worth emphasizing. First, W_ℓ, W_m and W_r are disjoint and are a partition of $W_{\mathcal{R}}$ by construction. Hence $\chi_{\mathcal{R}} = \chi_{W_\ell} \cdot \chi_{W_m} \cdot \chi_{W_r}$. Second, some vertices in I may not be in V_ℓ at all. However any such vertices that are in I but not V_ℓ are necessarily in S_L and thus will be contained in $\mathcal{R}_\ell$ anyways. This is why we can say that $\mathcal{R}_\ell$ is an $(I, S_L(\mathcal{R}))$-ribbon.

The following illustrates what the canonical factorization would look like in our earlier example:



We chose this example to illustrate a subtle point. The edge (i, c) has both its endpoints in both $\mathcal{R}_\ell$ and $\mathcal{R}_m$. We could in principle choose to place it in either, but we have adopted the convention that because both of its endpoints are in S_L we place it in $\mathcal{R}_m$. In this way, there are no edges within S_L in $\mathcal{R}_\ell$ or within S_R in $\mathcal{R}_m$. Finally, note that there can be isolated vertices in $\mathcal{R}_\ell$ or $\mathcal{R}_r$ but such vertices need to be in I or J respectively.

With the definition of the canonical factorization in hand, we will collect some important properties about it that we will make use of later:

Claim 11.4.5. Let $\mathcal{R}$ be an (I, J)-ribbon with canonical factorization $(\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r)$. Then

$$|\mathcal{V}(\mathcal{R})| = |\mathcal{V}(\mathcal{R}_\ell)| + |\mathcal{V}(\mathcal{R}_m)| + |\mathcal{V}(\mathcal{R}_r)| - |S_L(\mathcal{R})| - |S_R(\mathcal{R})|.$$

Proof. It is important to note that $S_L(\mathcal{R})$ and $S_R(\mathcal{R})$ are not necessarily disjoint (indeed, this happens in the example above). Nevertheless, we know that by construction V_ℓ, V_m and V_r are disjoint and that $S_L(\mathcal{R}) \cup S_R(\mathcal{R}) \subseteq V_m$. Every vertex that appears just once in $S_L(\mathcal{R})$ and $S_R(\mathcal{R})$ appears twice in the canonical

factorization. And every vertex that is in $S_L(\mathcal{R}) \cap S_R(\mathcal{R})$ appears three times. Thus

$$\begin{aligned}|\mathcal{V}(\mathcal{R})| = |\mathcal{V}(\mathcal{R}_\ell)| + |\mathcal{V}(\mathcal{R}_m)| + |\mathcal{V}(\mathcal{R}_r)| \\ - |S_L(\mathcal{R})/S_R(\mathcal{R})| - |S_R(\mathcal{R})/S_L(\mathcal{R})| - 2|S_L(\mathcal{R}) \cap S_R(\mathcal{R})|\end{aligned}$$

which completes the proof. □

In the discussion above, we established some properties that a canonical factorization must satisfy. Next we show the reverse direction, that any collection of ribbons that satisfies the below properties must be a canonical factorization. Consider a collection of ribbons $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$, and the following list of properties:

S_ℓ, S_r **Factorization Conditions for** $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$ **(Here** $S_\ell, S_r \subseteq [n]$**.)**

1. $\mathcal{R}_0$ is an (I, S_ℓ)-ribbon with $S_L(\mathcal{R}_0) = S_R(\mathcal{R}_0) = S_\ell$, and all vertices in $\mathcal{V}(\mathcal{R}_0)$ are either reachable from I without passing through S_ℓ or are in I or S_ℓ. Finally, $\mathcal{R}_0$ has no edges between vertices in S_ℓ.
2. $\mathcal{R}_2$ is an (S_r, J)-ribbon with $S_L(\mathcal{R}_2) = S_R(\mathcal{R}_2) = S_r$, and all vertices in $\mathcal{V}(\mathcal{R}_2)$ are either reachable from J without passing through S_r or are in J or S_r. Finally, $\mathcal{R}_2$ has no edges between vertices in S_r.
3. $\mathcal{R}_1$ is an (S_ℓ, S_r)-ribbon with $S_L(\mathcal{R}_1) = S_\ell$ and $S_R(\mathcal{R}_1) = S_r$. Every vertex in $\mathcal{V}(\mathcal{R}_1) \setminus (S_\ell \cup S_r)$ has degree at least 1.
4. $W_{\mathcal{R}_0}, W_{\mathcal{R}_1}, W_{\mathcal{R}_2}$ are pairwise disjoint. Also, $V_{\mathcal{R}_0} \cap V_{\mathcal{R}_1} = S_\ell$, $V_{\mathcal{R}_1} \cap V_{\mathcal{R}_2} = S_r$, and $V_{\mathcal{R}_0} \cap V_{\mathcal{R}_2} = S_\ell \cap S_r$.

Lemma 11.4.6. *Let $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$ be ribbons. Then $(\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2)$ is the canonical factorization of the (I, J)-ribbon $\mathcal{R}$ with edge set $W_{\mathcal{R}_0} \oplus W_{\mathcal{R}_1} \oplus W_{\mathcal{R}_2}$ and vertex set*

$\mathcal{V}(\mathcal{R}_0) \cup \mathcal{V}(\mathcal{R}_1) \cup \mathcal{V}(\mathcal{R}_2)$ *if and only if the* S_ℓ, S_r *factorization conditions hold for* $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$ *for some* $S_\ell, S_r \subseteq [n]$.

Proof. If $\mathcal{R}$ is a ribbon with leftmost and rightmost vertex separators S_ℓ and S_r and canonical factorization $(\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2)$, then many of the conditions above are automatically satisfied. By construction, $W_{\mathcal{R}_0}, W_{\mathcal{R}_1}, W_{\mathcal{R}_2}$ are pairwise disjoint. Because any edge with both endpoints in S_ℓ is included in $\mathcal{R}_m$ we have that there are no edges between vertices in S_ℓ in $\mathcal{R}_0$, and similarly for $\mathcal{R}_2$. Finally suppose there is a vertex u in $\mathcal{R}_0$. If u is not reachable from I without passing through S_ℓ and is not in I or S_ℓ then it would not be included in $\mathcal{R}_0$. An identical argument holds for $\mathcal{R}_2$.

All that remains is to verify that $S_L(\mathcal{R}_0) = S_R(\mathcal{R}_0) = S_\ell$ and similarly for $\mathcal{R}_1, \mathcal{R}_2$. If $S_\ell = S_L(\mathcal{R})$ is not a minimum-size vertex separator for $\mathcal{R}_0$, then it is also not a minimum-size vertex separator for $\mathcal{R}$, which is impossible. Similarly, if it is not the leftmost separator for $\mathcal{R}_0$ then it was not the leftmost separator for $\mathcal{R}$. Since $\mathcal{R}_0$ is an (I, S_ℓ)-ribbon and S_ℓ is a minimum-size separator, it must also be the right-most minimum-size separator.

Now in the reverse direction, suppose that $\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2$ are ribbons that meet the S_ℓ, S_r factorization conditions. We claim that S_ℓ is the leftmost separator for $\mathcal{R}$. If not, then either their is a smaller vertex separator, or there is a vertex separator S'_ℓ of the same size that separates I and S_ℓ. To rule out the former case, note that since S_ℓ and S_r are both minimum vertex separators for $\mathcal{R}_1$, we must have $|S_\ell| = |S_r|$. Then it follows from the S_ℓ, S_r factorization conditions that there are $|S_\ell|$ vertex disjoint paths from I to J, but this would contradict the fact that there is a vertex separator with fewer than $|S_\ell|$ vertices. In the latter case, any other vertex separator S'_ℓ of the same size that separates I and S_ℓ would

contradict the condition $S_L(\mathcal{R}_0) = S_\ell$. An identical argument shows that S_r is the rightmost separator for $\mathcal{R}$.

Finally, by assumption all the vertices in $\mathcal{V}(\mathcal{R}_0)$ are either reachable from I without passing through S_ℓ or are in I or S_ℓ and hence would be included in $\mathcal{R}_0$. Similarly, there are no edges in $W_{\mathcal{R}_0}$ with both endpoints in S_ℓ. Thus if we were to compute the canonical factorization for $\mathcal{R}$ we would get the same set of vertices in each ribbon and the same partition of the edges. □

11.4.3 Factorization of Matrix Entries

This leads to our first factorization of the entries $\mathcal{M}(I, J)$ of $\mathcal{M}$. Unfortunately, the error terms in this first attempt will be too large. Using canonical factorizations and Claim 11.4.5, for any $I, J \subseteq [n]$ of size at most d we can write

$$\mathcal{M}(I,J) \tag{11.4.1}$$

$$= \sum_{\substack{\mathcal{R} \text{ an } (I,J)\text{-ribbon with edge set } W,\\ |\mathcal{V}(W)| \leqslant \tau \\ \text{canonical factorization } (\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r)}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R})|} \cdot \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r}$$

$$= \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell| = |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \tag{11.4.2}$$

$$\sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \subseteq \binom{[n]}{2} \\ \text{satisfying } S_\ell, S_r \text{ factorization conditions} \\ \text{and } |\mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_m)|+|\mathcal{V}(\mathcal{R}_r)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r}$$

Notice that except for the disjointness condition, the S_ℓ, S_r factorization conditions can be separated into condition 1 for $\mathcal{R}_\ell$, condition 3 for $\mathcal{R}_m$, and condition 2 for $\mathcal{R}_r$. We use this to rewrite as

$$= \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell| = |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \tag{11.4.3}$$

$$\left(\sum_{\substack{\mathcal{R}_\ell \text{ having } 1 \\ |\mathcal{V}(\mathcal{R}_\ell)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|} \chi_{\mathcal{R}_\ell} \right) \left(\sum_{\substack{\mathcal{R}_m \text{having } 3 \\ |\mathcal{V}(\mathcal{R}_m)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_m)| - \frac{|S_\ell|+|S_r|}{2}} \chi_{\mathcal{R}_m} \right) \left(\sum_{\substack{\mathcal{R}_r \text{having } 2 \\ |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_r)|} \chi_{\mathcal{R}_r} \right) \tag{11.4.4}$$

$$- \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell| = |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|-|S_r|}{2}} \underbrace{\sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \\ \text{satisfying } S_\ell, S_r \text{ conditions} \\ |\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau, \\ |\mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)| > \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_m)|+|\mathcal{V}(\mathcal{R}_r)| - \frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r}}_{\stackrel{\text{def}}{=} \xi_0(I,J), \text{ the error from ribbon size}} \tag{11.4.5}$$

$$- \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell| = |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|-|S_r|}{2}} \underbrace{\sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ 1,3,2 \text{ and not } 4 \\ |\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_m)|+|\mathcal{V}(\mathcal{R}_r)| - \frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r}}_{\stackrel{\text{def}}{=} E_0(I,J), \text{ the error from ribbon nondisjointness}} \cdot \tag{11.4.6}$$

11.4.4 Factorization of the Matrix $\mathcal{M}$

In lines 11.4.5 and 11.4.6 we have defined two error matrices, $\xi_0, E_0 \in \mathbb{R}^{\binom{[n]}{\leqslant d} \times \binom{[n]}{\leqslant d}}$. Inspired by the factorization of $\mathcal{M}(I,J)$ in line 11.4.4, we define another pair of matrices as follows:

$$Q_0 \in \mathbb{R}^{\binom{[n]}{d} \times \binom{[n]}{d}} \quad \text{given by} \quad Q_0(S_\ell, S_r) = \sum_{\substack{\mathcal{R}_m \text{ having } 3 \\ |\mathcal{V}(\mathcal{R}_m)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_m)| - \frac{|S_\ell|+|S_r|}{2}} \chi_{\mathcal{R}_m}$$

$$\mathcal{L} \in \mathbb{R}^{\binom{[n]}{d} \times \binom{[n]}{d}} \quad \text{given by} \quad \mathcal{L}(I,S) = \left(\frac{\omega}{n}\right)^{-\frac{|S|}{2}} \sum_{\substack{\mathcal{R}_\ell \text{ having } 1 \\ |\mathcal{V}(\mathcal{R}_\ell)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|} \chi_{\mathcal{R}_\ell} \,.$$

The powers of (ω/n) are split between Q_0 and $\mathcal{L}$ so that the typical of eigenvalue of Q_0 will be approximately 1 (although it will be some time before we are prepared to prove that).

The equation in lines 11.4.4, 11.4.5, and 11.4.6 can be written succinctly as

$$M = \mathcal{L} Q_0 \mathcal{L}^\top - \xi_0 - E_0 \,.$$

As we will see later, with high probability $Q_0 \succeq 0$, and thus also $\mathcal{L} Q_0 \mathcal{L}^\top \succeq 0$. So long as τ is sufficiently large, the spectral norm $\|\xi_0\|$ of the error term that accounts for ribbons whose size is too large will be negligible. However, the error E_0 does not turn out to be negligible. To overcome this we will apply a similar factorization approach to E_0 as we did for M; iterating this factorization will push down the error from ribbon nondisjointness.

We record an elementary fact about Q_0:

Lemma 11.4.7. *Let Π be the projector to $\mathrm{Span}\{e_C : C \in C_{\leqslant d}\}$. Then $Q_0 = \Pi Q_0 = Q_0 \Pi$.*

Proof. Suppose S is not a clique in G. We need to show that the row $Q_0(S, \cdot)$ is zero. For every entry $Q_0(S, S')$, notice that the Fourier coefficients $\widehat{Q_0(S,S')}(T) = \widehat{Q_0(S,S')}(T')$ if $T, T' \subseteq \binom{[n]}{2}$ disagree only on edges inside S. (That is, $T \oplus T' \subseteq \binom{S}{2}$.) This means that $Q_0(S,S') =_{S \text{ is a clique in } G} \cdot f_{S,S'}(G)$ for some function $f_{S,S'}$. □

11.4.5 Iterative Factorization of E_0

We recall now the definition of the matrix $E_0 \in \mathbb{R}^{\binom{[n]}{\leqslant d} \times \binom{[n]}{\leqslant d}}$.

$$E_0(I,J) = \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell| = |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ 1,3,2 \text{ and not } 4 \\ |\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r} \cdot$$

In what follows, we will show how to factor a slightly more general sort of matrix; this factorization will be applicable iteratively, starting with E_0.

The matrix $\mathcal{E}_c$ and its factorization

To express the family of matrices we will factor, we introduce a relaxation of our definition of ribbon and a corresponding relaxation 3* of condition 3 of the S_ℓ, S_r factorization conditions.

Definition 11.4.8 (Improper Ribbon)**.** An *improper* (I,J)*-ribbon* $\mathcal{R}$ is an (I,J)-ribbon $\mathcal{R}_0$ together with a set $\mathcal{Z}(\mathcal{R}) \subseteq [n]$ of vertices disjoint from $\mathcal{V}(\mathcal{R}_0)$. (Think of adding the vertices $\mathcal{Z}(\mathcal{R})$ to the ribbon $\mathcal{R}_0$ as degree-0 nodes.) We write $\mathcal{V}(\mathcal{R}) = \mathcal{V}(\mathcal{R}_0) \cup \mathcal{Z}(\mathcal{R})$. When we need to distinguish, we sometimes call ordinary ribbons "proper".

Every ribbon is also an improper ribbon by taking $\mathcal{Z}(\cdot) = \emptyset$, and every improper ribbon has a corresponding ribbon given by deleting its degree-0 vertices.

Relaxed Factorization Condition for ribbon $\mathcal{R}_1$ with $\mathcal{S}_\ell, \mathcal{S}_r \subseteq [n]$
3*. $\mathcal{R}_1$ is an improper (S_ℓ, S_r)-ribbon.

Let c be a $\mathbb{R}$-valued function $c(\mathcal{R})$ on (possibly improper) ribbons. Let $\mathcal{E}_c \in \mathbb{R}^{\binom{[n]}{\leqslant d}\times\binom{[n]}{\leqslant d}}$ be given by

$$\mathcal{E}_c(I,J) = \sum_{\substack{S_\ell,S_r\subseteq[n]\\ |S_\ell|,|S_r|\leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying}\\ 1,3^*,2 \text{ and not } 4\\ |\mathcal{V}(\mathcal{R}_\ell)|,|\mathcal{V}(\mathcal{R}_m)|,|\mathcal{V}(\mathcal{R}_r)|\leqslant\tau}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}_\ell}\cdot\chi_{\mathcal{R}_m}\cdot\chi_{\mathcal{R}_r}\cdot \tag{11.4.7}$$

Note that 3 is a strictly more restrictive condition than 3*. Hence we can define the function c_0 by $c_0(\mathcal{R}_m) = 1$ if $\mathcal{R}_m$ satisfies 3 and $c_0(\mathcal{R}_m) = 0$ otherwise. Then $E_0 = \mathcal{E}_{c_0}$. In this subsection, we will show how to factor any matrix of the form $\mathcal{E}_c$ as

$$\mathcal{E}_c = \mathcal{L}\,\mathcal{Q}_{c'}\,\mathcal{L}^\top - \mathcal{E}_{c'} - \xi_c$$

for some function c' on ribbons and matrices $\mathcal{Q}_{c'}, \xi_c \in \mathbb{R}^{\binom{[n]}{\leqslant d}\times\binom{[n]}{\leqslant d}}$ where $\|\xi_c\|$ is negligible with high probability.

Just as our initial factorization of $\mathcal{M}$ began with a factorization of each ribbon appearing in the Fourier expansion, our factorization of $\mathcal{E}_c$ depends on a factorization for each triple $(\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r)$ appearing in 11.4.7. Since they do not satisfy 4, there must be some vertices occurring in more than one of $\mathcal{V}(\mathcal{R}_\ell),\mathcal{V}(\mathcal{R}_m),\mathcal{V}(\mathcal{R}_\ell)$. Before, the canonical factorization depended on the leftmost and rightmost vertex separators in an (I,J)-ribbon $\mathcal{R}$ separating I from J. But now we will be interested in leftmost and rightmost separators that separate both I and J from each other and from these repeated vertices.

Definition 11.4.9 (Separating Factorization)**.** Let $\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r$ be ribbons satisfying S_ℓ, S_r factorization conditions 1, 3*, 2 but not 4, with $|\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau$. Let $\mathcal{R}$ be the (I,J)-ribbon with edge set $W_{\mathcal{R}_\ell}\oplus W_{\mathcal{R}_m}\oplus W_{\mathcal{R}_r}$ and vertex set

$\mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)$. (Thus, $\chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r} = \chi_{\mathcal{R}}$.)

Let S'_ℓ be the leftmost minimum-size vertex separator in $\mathcal{R}$ which separates I from J and any vertices appearing in more than one of $\mathcal{V}(\mathcal{R}_\ell), \mathcal{V}(\mathcal{R}_m), \mathcal{V}(\mathcal{R}_r)$. Similarly, let S'_r be the rightmost minimum-size vertex separator in $\mathcal{R}$ separating J from I and these repeated vertices. (Notice that S'_ℓ and S'_r could have different sizes.)

Let V'_ℓ be the vertices reachable from I without passing through S'_ℓ and similarly for V'_r. Let $V'_m = V_{\mathcal{R}} \setminus (V'_\ell \cup V'_r)$. Let $W'_\ell = \{(u, v) \in W_{\mathcal{R}} : u \in V_\ell, v \in V_\ell \cup S'_\ell\}$ and similarly for W'_r, and let $W'_m = W_{\mathcal{R}} \setminus (W'_\ell \cup W'_r)$.

Let $\mathcal{R}'_\ell$ be the (I, S'_ℓ)-ribbon with vertex set $V'_\ell \cup S'_\ell$ and edge set W'_ℓ and let $\mathcal{R}'_r$ be the (S'_r, J)-ribbon with vertex set $V'_r \cup S'_r$ and edge set W'_r. Finally, let $\mathcal{R}'_m$ be the improper (S'_ℓ, S'_r)-ribbon with edge set W'_m and vertex set $(\mathcal{V}(\mathcal{R}) \setminus (V'_\ell \cup V'_r)) \cup S'_\ell \cup S'_r)$.

Note that $\chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r} = \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r}$ if $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ is the separating factorization for $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$. We can use this to rewrite $\mathcal{E}_c$ as

$$\mathcal{E}_c(I, J) = \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell|, |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \\ |\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau \\ \text{separating factorization} \\ \mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, S'_\ell, S'_r}} c(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r} \tag{11.4.8}$$

Our goal is to find some coefficient function c' on (improper) ribbons and a matrix $Q_{c'}$ so that this is approximately equal to $\mathcal{L} Q_{c'} \mathcal{L}^\top - \mathcal{E}_{c'}$. For c' yet to be chosen,

we take

$$Q_{c'}(S'_\ell, S'_r) \stackrel{\text{def}}{=} \sum_{\substack{\mathcal{R}'_m \text{ having 3*} \\ |\mathcal{V}(\mathcal{R}'_m)| \leqslant \tau}} c'(\mathcal{R}'_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}'_m)| - \frac{|S'_\ell|+|S'_r|}{2}} \chi_{\mathcal{R}'_m}$$

and have that

$$\mathcal{L} Q_{c'} \mathcal{L}^\top(I,J) - \mathcal{E}_{c'}(I,J) = \sum_{\substack{S'_\ell, S'_r \subseteq [n] \\ |S'_\ell|, |S'_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S'_\ell|+|S'_r|}{2}} \sum_{\substack{\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r \text{ satisfying} \\ \text{1,3*,2, and 4} \\ |\mathcal{V}(\mathcal{R}'_\ell)|, |\mathcal{V}(\mathcal{R}'_m)|, |\mathcal{V}(\mathcal{R}'_r)| \leqslant \tau}} c'(\mathcal{R}'_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}'_\ell)|+|\mathcal{V}(\mathcal{R}'_r)|+|\mathcal{V}(\mathcal{R}'_m)| - \frac{|S'_\ell|+|S'_r|}{2}} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r} \cdot \tag{11.4.9}$$

We will compare (11.4.8) and (11.4.9) by collecting like terms, but first we handle the discrepancy in the size bounds on the ribbons with a corresponding error term ξ_c. The following matrix is similar to $\mathcal{E}_c$, but places a size bound on the ribbons in the separating factorization $|\mathcal{V}(\mathcal{R}'_\ell)|, |\mathcal{V}(\mathcal{R}'_m)|, |\mathcal{V}(\mathcal{R}'_r)| \leqslant \tau$. We define

$$\mathcal{E}'_c(I,J) = \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell|, |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ \text{1,3*,2 and not 4} \\ \text{separating factorization} \\ \mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, S'_\ell, S'_r \\ |\mathcal{V}(\mathcal{R}'_\ell)|, |\mathcal{V}(\mathcal{R}'_m)|, |\mathcal{V}(\mathcal{R}'_r)| \leqslant \tau}} c(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)| - \frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r}$$

We take $\xi_c = \mathcal{E}'_c - \mathcal{E}_c$ and we will show below that with high probability the error $\|\xi_c\|$ is negligible. Before doing this, we show that $\mathcal{E}'_c$ is exactly equal to $\mathcal{L}^\top Q_{c'} \mathcal{L}^\top - \mathcal{E}_{c'}$ for the correct choice of c'.

To collect like terms, it helps to define the following quantity $\gamma_{\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, I, J, S'_\ell, S'_r}$.

$$\gamma_{\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, I, J, S'_\ell, S'_r} \stackrel{\text{def}}{=} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ \text{1,3*,2 and not 4 for some } S_\ell, S_r \\ \text{separating factorization } \mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, S'_\ell, S'_r}} c(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_m)|+|\mathcal{V}(\mathcal{R}_r)| + \frac{|S'_\ell|+|S'_r|}{2} - |S_\ell| - |S_r|}.$$

Then we can rewrite $\mathcal{E}'_c(I,J)$ again as

$$\mathcal{E}'_c(I,J) = \sum_{\substack{S'_\ell,S'_r\subseteq[n]\\ |S'_\ell|,|S'_r|\leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S'_\ell|+|S'_r|}{2}} \sum_{\substack{\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r\\ \text{satisfying 1, 3*, 2, 4 for } S'_\ell,S'_r\\ |\mathcal{V}(\mathcal{R}'_\ell)|,|\mathcal{V}(\mathcal{R}'_m)|,|\mathcal{V}(\mathcal{R}_r)|\leqslant\tau}} \gamma_{\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r,I,J,S'_\ell,S'_r} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r}$$

We will obtain $\mathcal{E}'_c = \mathcal{L}^\top \mathcal{Q}_{c'} \mathcal{L}^\top - \mathcal{E}_{c'}$ if we define $c'(\mathcal{R}'_m)$ so that

$$c'(\mathcal{R}'_m)\left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}'_\ell)|+|\mathcal{V}(\mathcal{R}'_r)|+|\mathcal{V}(\mathcal{R}'_m)|-\frac{|S'_\ell|+|S'_r|}{2}} = \gamma_{\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r,I,J,S'_\ell,S'_r}$$

To express this in terms of the function c, we expand out $\gamma_{\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r,I,J,S'_\ell,S'_r}$. It is useful to define:

Definition 11.4.10. Let

$$r = (|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_m)|+|\mathcal{V}(\mathcal{R}_r)|-|S_\ell|-|S_r|)-(|\mathcal{V}(\mathcal{R}'_\ell)|+|\mathcal{V}(\mathcal{R}'_m)|+|\mathcal{V}(\mathcal{R}'_r)|-|S'_\ell|-|S'_r|)\,.$$

(The ribbons $\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r,\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r$ will always be clear from context.)

Note that $(\mathcal{V}(\mathcal{R}_\ell)| + |\mathcal{V}(\mathcal{R}_m)| + |\mathcal{V}(\mathcal{R}_r)| - |S_\ell| - |S_r|)$ is the total number of vertices we would have in the (I,J)-ribbon with vertex set $\mathcal{V}(\mathcal{R}_\ell)\cup\mathcal{V}(\mathcal{R}_m)\cup\mathcal{V}(\mathcal{R}_\ell)$ if $\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r$ satisfied condition 4 (which they do not!). Similarly, $(|\mathcal{V}(\mathcal{R}'_\ell)| + |\mathcal{V}(\mathcal{R}'_m)| + |\mathcal{V}(\mathcal{R}'_r)| - |S'_\ell| - |S'_r|)$ is the total number of vertices in the (I,J)-ribbon with edge set $\mathcal{W}(\mathcal{R}'_\ell)\cup\mathcal{W}(\mathcal{R}'_m)\cup\mathcal{W}(\mathcal{R}'_r)$ and vertex set $\mathcal{V}(\mathcal{R}'_\ell)\cup\mathcal{V}(\mathcal{R}'_m)\cup\mathcal{V}(\mathcal{R}'_r)$. Thus, r is the number of vertices occurring with multiplicity higher than they should in $\mathcal{V}(\mathcal{R}_\ell)\cup\mathcal{V}(\mathcal{R}_m)\cup\mathcal{V}(\mathcal{R}_r)$.

We can rewrite the γ's as

$$\gamma_{\mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r,I,J,S'_\ell,S'_r} = \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}'_\ell)|+|\mathcal{V}(\mathcal{R}'_m)|+|\mathcal{V}(\mathcal{R}'_r)|-\frac{|S'_\ell|+|S'_r|}{2}} \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying}\\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r\\ r \text{ intersections outside } S_\ell,S_r\\ \text{separating factorization } \mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r S'_\ell,S'_r}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^r .$$

Thus, we will have that $\mathcal{E}'_c = \mathcal{L}\,\mathcal{Q}_{c'}\,\mathcal{L}^\top - \mathcal{E}_{c'}$ if and only if for every (S'_ℓ, S'_r)-ribbon $\mathcal{R}'_m$ and every $\mathcal{R}'_\ell, \mathcal{R}'_r$ satisfying 1, 2,

$$c'(\mathcal{R}'_m) = \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r \\ r \text{ intersections outside } S_\ell,S_r \\ \text{separating factorization } \mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r S'_\ell,S'_r}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^r .$$

Note that for this to happen, the right hand side must be independent of $\mathcal{R}'_\ell$ and $\mathcal{R}'_r$. If this is the case, then we can define

$$c'(\mathcal{R}'_m) \stackrel{\text{def}}{=} \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r \\ r \text{ intersections outside } S_\ell,S_r \\ \text{separating factorization } \mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r S'_\ell,S'_r}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^r \text{ for some } \mathcal{R}'_\ell, \mathcal{R}'_r \text{ satisfying } 1, 2 .$$

The next claim shows that, indeed, the choice of $\mathcal{R}'_\ell, \mathcal{R}'_r$ does not matter. (This would not have been true without passing from $\mathcal{E}_c$ to $\mathcal{E}'_c$.)

Claim 11.4.11. Let $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ satisfy 1, 3*, 2, 4 for some $S'_\ell, S'_r \subseteq [n]$. Let $\mathcal{R}''_\ell$ and $\mathcal{R}''_r$ also satisfy 1 and 2, respectively, for S'_ℓ, S'_r, respectively. Then

$$\sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r \\ r \text{ intersections outside } S_\ell,S_r \\ \text{separating factorization } \mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r S'_\ell,S'_r}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^r = \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r \\ r \text{ intersections outside } S_\ell,S_r \\ \text{separating factorization } \mathcal{R}''_\ell,\mathcal{R}'_m,\mathcal{R}''_r S'_\ell,S'_r}} c(\mathcal{R}_m)\left(\frac{\omega}{n}\right)^r .$$

(Notice that the left-hand sum refers to $\mathcal{R}'_\ell, \mathcal{R}'_r$ and the right-hand one to $\mathcal{R}''_\ell, \mathcal{R}''_r$.)

Proof. We prove this by showing that there is an exact match between terms on the left hand side and terms on the right hand side. Consider a term on the left hand side. Note that the part of $\mathcal{R}_\ell$ between I and S'_ℓ must be $\mathcal{R}'_\ell$ while the part of $\mathcal{R}_\ell$ between S'_ℓ and S_ℓ becomes part of $\mathcal{R}'_m$. To shift from $\mathcal{R}'_\ell$ to $\mathcal{R}''_\ell$, we simply replace $\mathcal{R}'_\ell$ by $\mathcal{R}''_\ell$ within $\mathcal{R}_\ell$. Similarly, to shift from $\mathcal{R}'_r$ to $\mathcal{R}''_r$, we simply replace $\mathcal{R}'_r$ by $\mathcal{R}''_r$ within $\mathcal{R}_r$.

To show that this gives an exact match, we need to show that r is unaffected by these shifts. To see that shifting from $\mathcal{R}'_\ell$ to $\mathcal{R}''_\ell$ does not affect r, note that all vertices in $\mathcal{V}(\mathcal{R}'_\ell) \setminus S'_\ell$ or $\mathcal{V}(\mathcal{R}'_\ell) \setminus S'_\ell$ must appear in the corresponding $\mathcal{R}_\ell$ and cannot appear in $\mathcal{R}_m$ or $\mathcal{R}_r$. Thus, these vertices always have multiplicity 1 in $\mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)$. All other vertices (including the ones in S'_ℓ) may appear in $\mathcal{R}_m$ or $\mathcal{R}_r$ as well as $\mathcal{R}_\ell$ but whether or not they do so is unaffected by the shift so their multiplicities in $\mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)$ are unaffected by the shift and r remains the same. A similar argument holds for shifting from $\mathcal{R}'_r$ to $\mathcal{R}''_r$ □

Remark 11.4.12. For this argument, it was important to keep track of the isolated vertices in $\mathcal{R}'_m$. If we did not keep track of isolated vertices and instead had them disappear, we could have a situation where there is a vertex v which appears in $\mathcal{R}_\ell$ and $\mathcal{R}_m$ but disappears from $\mathcal{R}'_m$ and is not in S'_ℓ. Since v is no longer in $\mathcal{R}'_m$, $\mathcal{R}''_\ell$ could contain v. If so, then we cannot shift from $\mathcal{R}'_\ell$ to $\mathcal{R}''_\ell$ as this would create a copy of v to the left of S'_ℓ but v should be to the right of S'_ℓ.

Putting everything together, $\mathcal{E}'_c = \mathcal{L}\mathcal{Q}_{c'}\mathcal{L}^\top - \mathcal{E}_{c'}$. Since we defined $\xi_c = \mathcal{E}'_c - \mathcal{E}_c$, we get that $\mathcal{E}_c = \mathcal{L}\mathcal{Q}_c\mathcal{L}^\top - \mathcal{E}_{c'} - \xi_c$, as needed.

The remaining step will be to show that with high probability, the error term ξ_c has negligible norm, which we will accomplish in Section 11.5.5.

Finally, we record the following easy lemma about separating factorizations, which will be useful in the application of the foregoing to factor $\mathcal{E}_0$.

Lemma 11.4.13. *Suppose $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$ satisfy conditions 1, 3*, 2, but not 4. Let $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ be their separating factorization, with separators S'_ℓ, S'_r. Then*

$$\frac{|S'_\ell| + |S'_r|}{2} - \frac{|S_\ell| + |S_r|}{2} \geqslant \frac{1}{2}$$

Proof. We claim that $|S_\ell| + |S_r| + 1 \leqslant |S'_\ell| + |S'_r|$ By the violation of condition 4, we cannot have $S_\ell = S'_\ell$ and $S_r = S'_r$. But since S'_ℓ separates I from S_ℓ in $\mathcal{R}_\ell$ and $\mathcal{R}_\ell$ is an (I, S_ℓ)-ribbon whose rightmost vertex separator is also S_ℓ, if $S_\ell \neq S'_\ell$ then $|S_\ell| < |S'_\ell|$, and similarly for S_r and S'_r. So either $|S_\ell| < |S'_\ell|$ or $|S_r| < |S'_r|$, and since the separator sizes are integers, so the difference must be at least 1 and we are done. □

Application to E_0 and $\mathcal{M}$

We are ready to define our recursive factorization of E_0. Recall that $c_0(\mathcal{R}_m) = 1$ if $\mathcal{R}_m$ satisfies 3 and $c_0(\mathcal{R}_m) = 0$ otherwise and $E_0 = \mathcal{E}_{c_0}$. Applying the factorization above to $\mathcal{E}_{c_0}$ we obtain matrices $\xi_1 = \xi_{c_0}, Q_1$, and $\mathcal{E}_{c_1}$. Then of course we can apply the factorization again to $\mathcal{E}_{c_1}$.

Proceeding inductively, for all $i \in [1, 2d]$ let $\xi_i = \xi_{c_{i-1}}, Q_i$, and $\mathcal{E}_{c_i}$ be the matrices given by applying the factorization to $\mathcal{E}_{c_{i-1}}$ at step i.

Claim 11.4.14.

$$\mathcal{M} = \mathcal{L}(Q_0 - Q_1 + Q_2 - \ldots - Q_{2d-1} + Q_{2d})\,\mathcal{L}^\top - (\xi_0 - \xi_1 + \xi_2 - \ldots - \xi_{2d-1} + \xi_{2d})\,.$$

Proof. We have that $\mathcal{M} = \mathcal{L}(Q_0)\,\mathcal{L}^\top - \mathcal{E}_0 - \xi_0$ and $\mathcal{E}_{i-1} = \mathcal{L}\,Q_i\,\mathcal{L}^\top - \mathcal{E}_i - \xi_{c_{i-1}} = \mathcal{L}\,Q_i\,\mathcal{L}^\top - \mathcal{E}_i - \xi_i$. We prove the claim by starting with the first formula and appliying the second formula for each $i \in [1, 2d]$. At the end, we are left with an extra term $\mathcal{E}_{2d}$. We must show that $\mathcal{E}_{2d} = 0$.

To see why $\mathcal{E}_{2d} = 0$, note that every time we have a separating factorization $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ for $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$, the size of either the left separator or the right separator must increase (see Lemma 11.4.13). However, the size of these separators is

always at most d, so the only way we can do this for $2d$ steps is if we started with the empty set as the separators and increased the size of either the left or right separator by 1 each time, but not both. However, this too is impossible as if we start with the empty set as the separators, after the first step both the new left separator and the new right separator must have size at least 1. □

11.5 $\mathcal{M}$ is PSD

In this section we combine the factorization of $\mathcal{M}$ in terms of the matrices $\mathcal{L}, Q_i, \xi_i$ that we obtained in Section 11.4 with estimates on the eigenvalues of the Qs and ξs. The starting point is the following PSDness claim for Q_0.

Lemma 11.5.1. *Let $D \in \mathbb{R}^{\binom{[n]}{\leqslant d}\times\binom{[n]}{\leqslant d}}$ be the diagonal matrix with $D(S,S) = 2^{\binom{|S|}{2}}/4$ if S is a clique in G and 0 otherwise. With high probability, $Q_0 \succeq D$.*

We also need to bound $\|Q_i\|$ for $i > 0$.

Lemma 11.5.2. *Let $D \in \mathbb{R}^{\binom{[n]}{\leqslant d}\times\binom{[n]}{\leqslant d}}$ be the diagonal matrix with $D(S,S) = 2^{\binom{S}{2}}/4$ if S is a clique and is otherwise zero. With high probability, every Q_i for $i \in [1, 2d]$ satisfies*

$$\frac{-D}{8d} \preceq Q_i \preceq \frac{D}{8d}.$$

The preceding lemmas are enough to obtain $Q_0 - \ldots + Q_{2d} \succeq D/2$, but in the end we need to work with the matrix $\mathcal{L}(Q_0 - \ldots + Q_{2d})\,\mathcal{L}^\top - (\xi_0 - \ldots + \xi_{2d})$. The next two lemmas allow us to make this last step.

Lemma 11.5.3. *With high probability, $\Pi\,\mathcal{L}\,\Pi\,\mathcal{L}^\top\,\Pi \succeq \Omega(\omega/n)^{d+1} \cdot \Pi$, where as usual Π is the projector to $\mathrm{Span}\{e_C : C \in C_{\leqslant d}\}$.*

Finally, we need a bound on the ξ matrices.

Lemma 11.5.4. *With high probability,* $\|\xi_0 - \ldots + \xi_{2d}\| \leqslant n^{-16d}$.

We can now prove Lemma 11.3.7.

Proof of Lemma 11.3.7. By Claim 11.4.14,

$$\mathcal{M} = \mathcal{L}(Q_0 - Q_1 + Q_2 - \ldots - Q_{2d-1} + Q_{2d})\,\mathcal{L}^\top - (\xi_0 - \xi_1 + \xi_2 - \ldots - \xi_{2d-1} + \xi_{2d})\,.$$

By a union bound, with high probability the conclusions of Lemmas 11.5.1, 11.5.2, 11.5.3, and 11.5.4 all hold. By Lemma 11.5.1 and Lemma 11.5.2,

$$Q_0 - Q_1 + Q_2 - \ldots - Q_{2d-1} + Q_{2d} \geq \frac{D}{2} \geq \frac{\Pi}{2}\,.$$

where as usual Π is the projector to Span $e_C : C \in C_{\leqslant d}$. Thus by Lemma 11.5.3, we obtain $\mathcal{L}(Q_0 - \ldots + Q_{2d})\,\mathcal{L}^\top \geq \Omega(\omega/n)^{d+1} \cdot \Pi$. Finally, by Lemma 11.5.4 we have

$$\mathcal{M} = \Pi \cdot \mathcal{M} \cdot \Pi \geq \Omega\left(\frac{\omega}{n}\right)^{d+1} \cdot \Pi - n^{-16d} \cdot \Pi \geq 0\,. \qquad \square$$

In the next subsections, we prove the foregoing lemmas.

11.5.1 Ribbons and Spectral Norms

We will require bounds on the spectral norm of certain random matrices. Our random matrices arise out of decompositions of the moment matrix from Definition 11.3.6 and are functions of a graph G on vertex set $[n]$. Our norm bounds will hold for what we call *graphical matrices*, that are are defined to capture the matrices that are invariant under a permutation of the vertices of G and are in fact "minimal" such matrices.

We first define the *shape* of a ribbon that identifies the structure of a ribbon up to relabelling.

Definition 11.5.5 (Shape of a Ribbon)**.** For an (I, J)-ribbon $\mathcal{R}$, consider the graph U on the vertex set $[|\mathcal{V}(\mathcal{R})|]$ whose edges are

$$E(U) = \{(i, j) : \text{ there is an edge in } \mathcal{R} \text{ from the } i\text{-th to the } j\text{-th least element of } \mathcal{V}(\mathcal{R})\}.$$

(Here we are considering $\mathcal{V}(\mathcal{R})$ to have the usual ordering inherited from $[n]$.) Also, let U have two distinguished subsets of vertices A and B, where $A = \{i : \text{the } i\text{-th element of } \mathcal{V}(\mathcal{R}) \text{ is in } I\}$, and similarly for B and J. We call U the *shape* of $\mathcal{R}$ and write $\mathsf{shape}(\mathcal{R}) = U$.

We record some observations on shapes of ribbons.

- If $\mathcal{R}$ is a ribbon (not an improper ribbon), its shape satisfies the condition that every vertex outside $A \cup B$ has degree at least 1.
- If, for example, $\mathcal{R}$ is an (I, J) ribbon where $I \cap J = \{1\}$ (which must be the least element in both I and J), then in order for the (I', J')-ribbon $\mathcal{R}'$ to have the same shape as $\mathcal{R}$ it is necessary that $|I' \cap J'| = 1$. More broadly, specifying the shape of a ribbon in particular specifies the pattern of intersection of its endpoints.
- A matrix $M \in \mathbb{R}^{\binom{n}{\leqslant d} \times \binom{n}{\leqslant d}}$ whose entries are given by $M(I, J) = \sum_{\mathcal{R} \text{ an } (I,J)\text{-ribbon with shape } U} \chi_{\mathcal{R}}$ satisfies the assumptions of Lemma 11.5.8. In the following sections, our main strategy will be to decompose the matrices Q_i into matrices of this form.

We are now ready to define graphical matrices.

Definition 11.5.6 (Graphical Matrices)**.** Let U be a graph on the vertex set $[t]$ with two distinguished sets of vertices $A, B \subseteq [t]$. Let $\mathcal{T}(\mathcal{U})$ be the collection of all I, J ribbons with shape U. The graphical matrix $M \in \mathbb{R}^{\binom{[n]}{|A|} \times \binom{[n]}{|B|}}$ of shape U is defined by

$$M(I,J) = \sum_{\mathcal{R}:\mathcal{R} \text{ is an } (I,J)\text{-ribbon and } \mathsf{shape}(\mathcal{R})=U} \chi_{\mathcal{R}}.$$

Example 11.5.7. When U is a graph on 2 vertices with distinguished sets $\{1\}$ and $\{2\}$ of size 1 each and a single edge connecting vertex 1 and 2, the graphical matrix of shape U is just the standard $\{-1, 1\}$-adjacency matrix of the graph G.

The following lemma will be our main tool. It is in essence due to Medarametla and Potechin [127] and special cases of the bound have been proven and used in [89, 63]. We give a proof in the appendix for completeness.

Lemma 11.5.8. *Let U be a graph on $t \leqslant O(\log n)$ vertices, with two distinguished subsets of vertices A and B, and suppose:*

- *U admits p vertex-disjoint paths from $A \setminus B$ to $B \setminus A$.*
- *$|A \cap B| = r$.*
- *Every vertex outside $A \cup B$ has degree at least 1.*

Let $M = M(G)$ be the graphical matrix with shape U. Then, whp, $\|M\| \leqslant n^{\frac{t-p-r}{2}} \cdot 2^{O(t)} \cdot (\log n)^{O(t-r+p)}$.

Remark 11.5.9. Lemma 11.5.8 can be seen as a generalization of the standard upper bound on the spectral norm of the adjacency matrix. Example 11.5.7 shows how adjacency matrix is a graphical matrix with a shape U on 2 vertices with a single edge connecting them, thus $t = 2$, $r = 0$ and $p = 1$. Lemma 11.5.8 thus shows an upper bound of $\sqrt{n}$ poly log (n) on the spectral norm of the adjacency matrix which is tight up to a poly log (n) factor.

11.5.2 Positivity for Q_0 — Proof of Lemma 11.5.1

In this section we prove Lemma 11.5.1, which we restate here.

Lemma (Restatement of Theorem 11.5.1)**.** *Let $D \in \mathbb{R}^{\binom{[n]}{\leqslant d}\times\binom{[n]}{\leqslant d}}$ be the diagonal matrix with $D(S,S) = 2^{\binom{|S|}{2}}/4$ if S is a clique in G and 0 otherwise. With high probability, $Q_0 \succeq D$.*

Proof of Lemma 11.5.1. To begin, we split Q_0 into its diagonal Q_0^{diag} and its off-diagonal $Q_0^{\text{off-diag}}$ parts.

$$Q_0^{\text{diag}}(S_\ell, S_r) = \begin{cases} Q_0(S_\ell, S_r) \text{ if } S_\ell = S_r \\ 0 \text{ otherwise.} \end{cases} \qquad Q_0^{\text{off-diag}}(S_\ell, S_r) = \begin{cases} Q_0(S_\ell, S_r) \text{ if } S_\ell \neq S_r \\ 0 \text{ otherwise.} \end{cases}$$

Then $Q_0 = Q_0^{\text{diag}} + Q_0^{\text{off-diag}}$. Expanding Q_0^{diag},

$$Q_0^{\text{diag}}(S,S) =$$

$$2^{\binom{|S|}{2}} \cdot_{S \text{ is a clique}} \cdot \left(1 + \sum_{\substack{\mathcal{R} \text{ nonempty, having 3} \\ \text{and no edges inside } S \\ |S|<|\mathcal{R}|\leqslant\tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R})|-|S|} \cdot \chi_{\mathcal{R}} \right) = 2^{\binom{|S|}{2}} \cdot_{S \text{ is a clique}} \cdot (1 \pm n^{-\Omega(\varepsilon)})$$

for all $S \in \binom{[n]}{d}$ with high probability by a similar argument as in Lemma 11.3.3 and a union bound.

Next, we bound $\|Q_0^{\text{off-diag}}\|$ be decomposing it according to ribbon shape. Fix $s, t \leqslant \tau$. Let $U_1^{(s,t)}, \ldots, U_q^{(s,t)}$ be all the graphs on vertex set $[t]$ with two distinguished sets of vertices A, B, both of size s, with $|A \cap B| \leqslant s-1$, and where there are $s - |A \cap B|$ vertex-disjoint paths from $A \setminus B$ to $B \setminus A$. Let $M_i^{(s,t)}$ be given by

$$M_i^{(s,t)}(S_\ell, S_r) = \sum_{\mathcal{R} \text{ an } (S_\ell, S_r)\text{-ribbon with shape } U_i^{(s,t)}} \chi_{\mathcal{R}} \cdot$$

Then

$$Q_0^{\text{off-diag}} = \sum_{\substack{s \leqslant d \\ t \leqslant \tau \\ i \leqslant q}} \left(\frac{\omega}{n}\right)^{t-s} \cdot M_i^{(s,t)} .$$

We can apply Lemma 11.5.8 to conclude that with probability at least $1 - O(n^{-100 \log n})$,

$$\left\| \left(\frac{\omega}{n}\right)^{t-s} \cdot M_i^{(s,t)} \right\| \leqslant \left(\frac{\omega}{n}\right)^{t-s} \cdot n^{\frac{t-s}{2}} \cdot 2^{O(t)} \cdot (\log n)^{O(t-|A\cap B|+|A\setminus B|)} \leqslant n^{-\varepsilon(t-s)} \cdot 2^{O(t)} \cdot (\log n)^{O(t-s)} ,$$

where to conclude the bound on the exponent in $(\log n)^{O(t-|A\cap B|+|A\setminus B|)}$ we have used that $t \geqslant 2s - |A \cap B|$.

Notice that for fixed s and t, there are at most $2^{\binom{t}{2}+O(t)}$ unique shapes $U_1^{(s,t)}, \ldots, U_q^{(s,t)}$. Thus, a union bound followed by the triangle inequality, we obtain that for fixed s and t, with probability at least $1 - O(n^{-99 \log n})$,

$$\left\| \left(\frac{\omega}{n}\right)^{t-s} \sum_{i \leqslant q} M_i^{(s,t)} \right\| \leqslant 2^{\binom{t}{2}+O(t)} \cdot n^{-\varepsilon(t-s)} \cdot 2^{O(t)} \cdot (\log n)^{O(t-s)} .$$

Under our assumptions on the parameters d, τ, and ε, this is at most $2^{\binom{s}{2}}/(100\tau)$. Summing over all $t \leqslant \tau$, for a fixed s we have

$$\left\| \left(\frac{\omega}{n}\right)^{t-s} \sum_{\substack{t \leqslant \tau \\ i \leqslant q}} M_i^{(s,t)} \right\| \leqslant \frac{2^{\binom{s}{2}}}{100} .$$

Notice that the above matrix is exactly the block of $Q_0^{\text{off-diag}}$ corresponding to subsets of size s. Together with our bound on Q_0^{diag}, this proves the lemma. □

11.5.3 Norm Bounds for Q_i — Proof of Lemma 11.5.2

In this section we prove Lemma 11.5.2, restated here.

Lemma (Restatement of Theorem 11.5.2)**.** *Let* $D \in \mathbb{R}^{\binom{[n]}{\leqslant d} \times \binom{[n]}{\leqslant d}}$ *be the diagonal matrix with* $D(S,S) = 2^{\binom{S}{2}}/4$ *if* S *is a clique and is otherwise zero. With high probability, every* Q_i *for* $i \in [1, 2d]$ *satisfies*

$$\frac{-D}{8d} \preceq Q_i \preceq \frac{D}{8d}.$$

We will need to bound the coefficients $c_i(\mathcal{R}'_m)$ used to define the matrices Q_i which we set up in Section 11.4.

Lemma 11.5.10. *Let* $c_1, \dots, c_{2d}$ *be the coefficient functions defined in Section 11.4. For all improper* (S_ℓ, S_r)*-ribbons* $\mathcal{R}_m$ *admitting exactly* p *vertex-disjoint paths from* S_ℓ *to* S_r*, and all* $i \leqslant 2d$*, writing* $s = \frac{|S_\ell|+|S_r|}{2}$*,*

$$c_i(\mathcal{R}_m) \leqslant \left(\frac{\omega}{n}\right)^s \cdot n^{\frac{p-|\mathcal{Z}(\mathcal{R}_m)|-i/2}{2}+\varepsilon s}.$$

recalling that $\omega = n^{1/2-\varepsilon}$*. Furthermore, if* $\mathcal{R}_m$ *and* $\mathcal{R}'_m$ *have the same shape, then* $c_i(\mathcal{R}_m) = c_i(\mathcal{R}'_m)$*.*

With this lemma in hand we can prove Lemma 11.5.2.

Proof of Lemma 11.5.2. Fix some $0 < i \leqslant 2d$. We will use Lemma 11.5.8, which requires that we first decompose each Q_i into simpler matrices. First of all, for a proper ribbon $\mathcal{R}_m$, let

$$\tilde{c}_i(\mathcal{R}_m) = \sum_{\mathcal{R}'_m \text{ an improper ribbon whose largest proper subribbon is } \mathcal{R}_m} \left(\frac{\omega}{n}\right)^{|\mathcal{Z}(\mathcal{R}'_m)|} \cdot c_i(\mathcal{R}'_m).$$

Note that we include $\mathcal{R}_m$ itself in this sum as a proper ribbon is also an improper ribbon.

Claim 11.5.11. $\tilde{c}_i(\mathcal{R}_m) \leqslant 2(\omega/n)^s \cdot n^{\frac{p-i/2}{2}+\varepsilon s}$, where p is the number of vertex-disjoint paths from S_ℓ to S_r in $\mathcal{R}_m$.

Proof. Consider all of the improper ribbons $\mathcal{R}'_m$ with k isolated vertices whose largest proper subribbon is $\mathcal{R}_m$. For each such ribbon $\mathcal{R}'_m$, by Lemma 11.5.10, $(\omega/n)^k c_i(\mathcal{R}'_m) \leqslant \left(\frac{\omega}{n}\right)^{k+s} \cdot n^{\frac{p-k-i/2}{2}+\varepsilon s}$. There are at most n^k such improper ribbons. Adding all of their contributions together gives at most

$$\left(\frac{\omega}{\sqrt{n}}\right)^k \left(\frac{\omega}{n}\right)^s \cdot n^{\frac{p-i/2}{2}+\varepsilon s} < 2^{-k}(\omega/n)^s \cdot n^{\frac{p-i/2}{2}+\varepsilon s}$$

Summing this up over all $k \geqslant 0$ gives the result. □

Now fix $s_\ell, s_r \leqslant d$ and $t \leqslant \tau$ and let $U_1^{(s_\ell,s_r,t)}, \ldots, U_q^{(s_\ell,s_r,t)}$ be all graphs on the vertex set $[t]$ with two distinguished subsets of vertices: A of size s_ℓ and B of size s_r. Let

$$\begin{aligned} M_j^{(s_\ell,s_r,t)}(S_\ell, S_r) &= \sum_{\mathcal{R} \text{ is an } (S_\ell,S_r)\text{-ribbon with shape } U_j^{(s_\ell,s_r,t)}} \tilde{c}_i(\mathcal{R}) \cdot \left(\frac{\omega}{n}\right)^{t-s} \cdot \chi_{\mathcal{R}} \\ &= \tilde{c}_i(U_j^{(s_\ell,s_r,t)}) \sum_{\mathcal{R} \text{ is an } (S_\ell,S_r)\text{-ribbon with shape } U_j^{(s_\ell,s_r,t)}} \left(\frac{\omega}{n}\right)^{t-s} \cdot \chi_{\mathcal{R}}\,, \end{aligned}$$

where $s = \frac{s_\ell+s_r}{2}$ and we have used the fact that $\tilde{c}_i(\mathcal{R})$ depends only on the shape of $\mathcal{R}$.

Let $r = |A \cap B|$ where A, B are the distinguished sets of vertices for $U_j^{(s_\ell,s_r,t)}$, and let $\tilde{p}$ be the number of vertex-disjoint paths from $A \setminus B$ to $B \setminus A$, so that $p = r + \tilde{p}$. We can apply Lemma 11.5.8 and our bound on $\tilde{c}_i$ to get that with probability $1 - O(n^{-100\log n})$,

$$\begin{aligned} \left\| M_j^{(s_\ell,s_r,t)} \right\| &\leqslant \left(\frac{\omega}{n}\right)^{t-s} \cdot n^{\frac{\tilde{p}+r-i/2}{2}+\varepsilon s} \cdot n^{\frac{t-\tilde{p}-r}{2}} \cdot 2^{O(t)} \cdot (\log n)^{O(t-r+\tilde{p})} \\ &= n^{-\varepsilon(t-s)-i/4} \cdot 2^{O(t)} \cdot (\log n)^{O(t-r+\tilde{p})} \\ &= n^{-\varepsilon(t-s)-i/4} \cdot 2^{O(t)} \cdot (\log n)^{O(t-s)}\,, \end{aligned}$$

where in the last step we have used that $t \geqslant 2s - r$ and $\tilde{p} \leqslant s - r$.

By inspection,

$$Q_i = \sum_{\substack{s_\ell, s_r \leqslant d \\ t \leqslant \tau \\ j \leqslant q}} M_j^{(s_\ell, s_r, t)}.$$

For a fixed t there are at most $2^{\binom{t}{2}+O(t)}$ choices for U, so $q \leqslant 2^{\binom{t}{2}+O(t)}$. Now we fix s_ℓ, s_r and sum over t to obtain the block of Q_i corresponding to size-s_ℓ and size-s_r subsets. By triangle inequality and a union bound, with probability at least $1 - O(n^{-97 \log n})$,

$$\left\| \sum_{\substack{t \leqslant \tau \\ j \leqslant q}} M_j^{(s_\ell, s_r, t)} \right\| \leqslant 2^{\binom{t}{2}+O(t)} \cdot n^{-\varepsilon(t-s)-i/4} \cdot 2^{O(t)} \cdot (\log n)^{O(t-s)}.$$

From our assumptions on d, τ, and ε, this is at most $2^{\binom{s_\ell}{2}/2+\binom{s_r}{2}/2}/100d^3$.

As usual, let Π be the projector to $\mathrm{Span}\{e_C : C \in C_{\leqslant d}\}$. Note that $\Pi Q_i = Q_i \Pi = Q_i$, since $Q_i(I, J) = 0$ whenever I or J is not a clique. So, to show that $D/8d \succeq Q_i \succeq -D/8d$, it is sufficient to show that for all vectors v with $v = \Pi v$ it happens that $|v^\top Q_i v| \leqslant v^T (D/8d) v$. To see this, let v_k be the part of v indexed by cliques of size exactly k. Now,

$$\begin{aligned}
|v^\top Q_i v| &\leqslant \sum_{k_1=0}^{d} \sum_{k_2=0}^{d} \|v_{k_1}\| \left\| \sum_{\substack{t \leqslant \tau \\ j \leqslant q}} M_j^{(k_1, k_2, t)} \right\| \|v_{k_2}\| \\
&\leqslant \sum_{k_1=0}^{d} \sum_{k_2=0}^{d} \frac{1}{100d^3} \left(2^{\binom{k_1}{2}/2+\binom{k_2}{2}/2} \|v_{k_1}\| \|v_{k_2}\| \right) \\
&\leqslant \sum_{k_1=0}^{d} \sum_{k_2=0}^{d} \frac{1}{200d^3} \left(2^{\binom{k_1}{2}} \|v_{k_1}\|^2 + 2^{\binom{k_2}{2}} \|v_{k_2}\|^2 \right) \\
&\leqslant \sum_{k=0}^{d} \frac{2^{\binom{k}{2}}}{100d^2} \|v_k\|^2 \leqslant v^\top (D/8d) v
\end{aligned}$$

□

Coefficient Decay in the Factorization: Proof of Lemma 11.5.10

We turn to the proof of Lemma 11.5.10, for which we want the following characterization of the effect of the separating factorization on the underlying graph of a ribbon.

We require the following combinatorial quantities:

Definitions for Lemma 11.5.12

1. $I, J, S_\ell, S_r \subseteq [n]$ of size at most d.
2. Ribbons $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$ satisfying 1,3*,2 but not 4 for $S_\ell, S_r, I, J \subseteq [n]$. (Remember that $\mathcal{R}_m$ may be improper.)
3. Ribbons $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ which are the separating factorization of $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$, with separators S'_ℓ, S'_r. (Remember that $\mathcal{R}'_m$ may be improper.)
4. p, the number of vertex-disjoint paths from S_ℓ to S_r in $\mathcal{R}_m$.
5. p', the number of vertex-disjoint paths from S'_ℓ to S'_r in $\mathcal{R}'_m$.
6. $r = (|\mathcal{V}(\mathcal{R}_\ell)| + |\mathcal{V}(\mathcal{R}_m)| + |\mathcal{V}(\mathcal{R}_\ell)| - |S_\ell| - |S_r|) - (|\mathcal{V}(\mathcal{R}'_\ell)| + |\mathcal{V}(\mathcal{R}'_m)| + |\mathcal{V}(\mathcal{R}'_\ell)| - |S'_\ell| - |S'_r|)$, the number of intersections among R_ℓ, R_m, R_r.
7. $\mathfrak{D} = \mathcal{Z}(\mathcal{R}'_m) \setminus \mathcal{Z}(\mathcal{R}_m)$, the newly degree-0 (we write *isolated*) vertices in $\mathcal{R}'_m$.
8. $\mathfrak{U} \subseteq \mathcal{V}(\mathcal{R}_\ell) \cup \mathcal{V}(\mathcal{R}_m) \cup \mathcal{V}(\mathcal{R}_r)$, the set of vertices appearing in more than one of $\mathcal{V}(\mathcal{R}_\ell), \mathcal{V}(\mathcal{R}_m)$, and $\mathcal{V}(\mathcal{R}_r)$. Note that $\mathfrak{U} \subseteq \mathcal{V}(\mathcal{R}'_m)$.

Lemma 11.5.12.

$$\underbrace{|S'_\ell| + |S'_r| - (|S_\ell| + |S_r|)}_{\text{increase in separator size}} + \underbrace{p - p'}_{\text{lost paths between separators}} + \underbrace{|\mathfrak{D}|}_{\text{new isolated vertices}} \leqslant \underbrace{r}_{\text{number of intersections}} .$$

The following series of claims will help us in the proof of Lemma 11.5.12

Claim 11.5.13. $I \cap \mathcal{V}(\mathcal{R}'_m) \subseteq S'_\ell$ and $J \cap \mathcal{V}(\mathcal{R}'_m) \subseteq S'_r$.

Proof of claim. If $u \in I \cap \mathcal{V}(\mathcal{R}'_m)$ then since $I \subseteq \mathcal{V}(\mathcal{R}'_\ell)$, we have $u \in \mathcal{V}(\mathcal{R}'_\ell) \cap \mathcal{V}(\mathcal{R}'_m) = S'_\ell$, and similarly for the second part. □

Next we have a simple analysis of which vertices may possibly be newly isolated.

Claim 11.5.14. $\mathfrak{D} \subseteq \mathfrak{U}$.

Proof of claim. Let $u \in \mathfrak{D}$. If $u \in S_\ell$ or $u \in S_r$ we are done. Otherwise, if $u \in I$ or $u \in J$, then u appeared in more than one of $\mathcal{V}(\mathcal{R}_\ell), \mathcal{V}(\mathcal{R}_m), \mathcal{V}(\mathcal{R}_r)$ by the definition of the canonical factorization.

If neither of these cases hold, then u was incident to an edge in at least one of $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$. Since that edge does not exist in $\mathcal{R}'_m$, it must have appeared at least twice among the edge sets of $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$, and therefore u appeared at least twice among the vertex sets, thus proving the claim. □

Next we show that some vertices in $\mathfrak{U}$ cannot become isolated.

Claim 11.5.15. By Menger's theorem, there are $|S'_\ell|$ vertex-disjoint paths from $\mathfrak{U} \cap \mathcal{V}(\mathcal{R}_\ell)$ to I in $\mathcal{R}_\ell$. Let $u_\ell^{(1)}, \ldots, u_\ell^{(|S'_\ell|)}$ be distinct vertices so that $u^{(i)}$ is the last vertex in $\mathfrak{U}$ along the i-th vertex disjoint path. Let $u_r^{(1)}, \ldots, u_r^{(|S'_r|)}$ be similarly defined. None of the vertices u may be in $\mathfrak{D}$.

Proof of claim. Fix one of these vertices u, and consider its neighbor v one step farther along the path to I (or J). By definition, the vertex v does not appear in more than one of $\mathcal{V}(\mathcal{R}_\ell), \mathcal{V}(\mathcal{R}_m), \mathcal{V}(\mathcal{R}_r)$. If $v \in \mathcal{R}'_m$, then the edge (u, v) must be

in $\mathcal{R}'_m$, and so u is not isolated in $\mathcal{R}'_m$. If $v \notin \mathcal{R}'_m$, then u must be in $S'_\ell \cup S'_r$, in which case by definition $u \notin \mathfrak{D}$. □

We set up sets q of vertices to divide up the intersecting vertices among $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$ according to which ribbons witness the intersection.

Claim 11.5.16. Let

$$\begin{aligned} q_{\ell,m,r} &\stackrel{\text{def}}{=} (\mathcal{V}(\mathcal{R}_r) \cap \mathcal{V}(\mathcal{R}_m) \cap \mathcal{V}(\mathcal{R}_\ell)) \setminus (S_\ell \cup S_r) \\ q_{\ell,r} &\stackrel{\text{def}}{=} (\mathcal{V}(\mathcal{R}_\ell) \cap \mathcal{V}(\mathcal{R}_r)) \setminus \mathcal{V}(\mathcal{R}_m) \\ q_{\ell,m} &\stackrel{\text{def}}{=} (\mathcal{V}(\mathcal{R}_\ell) \cap \mathcal{V}(\mathcal{R}_m)) \setminus (S_\ell \cup \mathcal{V}(\mathcal{R}_r)) \\ q_{r,m} &\stackrel{\text{def}}{=} (\mathcal{V}(\mathcal{R}_r) \cap \mathcal{V}(\mathcal{R}_m)) \setminus (S_r \cup \mathcal{V}(\mathcal{R}_\ell)) \,. \end{aligned}$$

The sets q are pairwise disjoint, and

$$r = 2|q_{\ell,m,r}| + |q_{\ell,r}| + |q_{\ell,m}| + |q_{r,m}| + |S_\ell \cap (\mathcal{V}(\mathcal{R}_r) \setminus S_r)| + |S_r \cap (\mathcal{V}(\mathcal{R}_\ell) \setminus S_\ell)| \,.$$

Also, $\mathfrak{U} = q_{\ell,m,r} \cup q_{\ell,r} \cup q_{\ell,m} \cup q_{r,m} \cup S_\ell \cup S_r$.

Proof. By inspection. □

We are prepared to prove Lemma 11.5.12.

Proof of Lemma 11.5.12. We start by bounding the number of vertices in $\mathfrak{U} \setminus \mathfrak{D}$. By Claim 11.5.15, there are at least $|\{u_\ell^{(1)}, \ldots, u_\ell^{(|S'_\ell|)}, u_r^{(1)}, \ldots, u_r^{|S'_r|}\}|$ such vertices.

Let a be the number of pairs i, j so that $u_\ell^{(i)} = u_r^{(j)}$. Then there are vertex-disjoint paths $w_1, \ldots, w_a$ from S'_ℓ to S'_r. The path w corresponding to $u_\ell^{(i)} = u_r^{(j)}$ is given by following $u_\ell^{(i)}$'s path from I to $\mathcal{U}$, ending at $u_\ell^{(i)}$, then following $u_r^{(j)}$'s path from $\mathfrak{U}$ to J. This gives a path from I to J, which must have a subpath from S'_ℓ to S'_r.

Now consider the p vertex-disjoint paths from S_ℓ to S_r in $\mathcal{R}_m$. We claim that

$$p - |S_\ell \cap S_r| \leqslant |q_{\ell,m,r}| + |S_\ell \cap \mathcal{V}(\mathcal{R}_r) \setminus S_r| + |S_r \cap \mathcal{V}(\mathcal{R}_\ell) \setminus S_\ell| + |\mathfrak{U} \setminus (\{u_\ell^{(1)}, \ldots, u_\ell^{(|S'_\ell|)}, u_r^{(1)}, \ldots, u_r^{|S'_r|}\} \cup \mathfrak{D})| + (p' - a) \quad (11.5.1)$$

In words, every nontrivial path from S_ℓ to S_r contributes to at least one of:

- $|q_{\ell,m,r}|$, the number of 3-way intersections,
- intersections between S_ℓ and $\mathcal{V}(\mathcal{R}_r)$ (but not S_r), intersections between $\mathcal{V}(\mathcal{R}_\ell)$ and S_r (but not S_ℓ),
- vertices in $\mathfrak{U}$ which are guaranteed not to become isolated (and which we have not yet accounted for), or
- vertex-disjoint paths from S'_ℓ to S'_r (which we have not yet accounted for).

Fix one such path. If it intersects $q_{\ell,m,r}$, $S_l \cap \mathcal{V}(\mathcal{R}_r)$, or $S_r \cap \mathcal{V}(\mathcal{R}_l)$ we are done, so suppose otherwise. If it is contained entirely in $q_{\ell,m} \cup q_{r,m} \cup (S_\ell \setminus \mathcal{V}(\mathcal{R}_r)) \cup (S_r \setminus \mathcal{V}(\mathcal{R}_l))$, then there is some edge along the path connecting a vertex in $\mathcal{V}(\mathcal{R}_\ell) \cap \mathcal{V}(\mathcal{R}_m) \setminus \mathcal{V}(\mathcal{R}_r)$ with one in $\mathcal{V}(\mathcal{R}_r) \cap \mathcal{V}(\mathcal{R}_m) \setminus \mathcal{V}(\mathcal{R}_\ell)$. That edge can occur nowhere else among $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$, and so the incident vertices must not be in $\mathfrak{D}$. At the same time, if there is any vertex along the path which is outside $\mathfrak{U}$, then the nearest vertices along the path to either side which do lie in $\mathfrak{U}$ also must be outside $\mathfrak{D}$.

In either case, there are two vertices along the path in $\mathfrak{U} \setminus \mathfrak{D}$. If either of these is not among the u vertices, we are done. If both are, then by definition of the u vertices this creates a path from I to J, and so from S'_ℓ to S'_r. Furthermore, this path must be vertex disjoint from the paths $w_1, \ldots, w_a$ previously constructed,

since the u vertices involved in those paths were $\mathcal{V}(\mathcal{R}_\ell) \cap \mathcal{V}(\mathcal{R}_r)$. This proves (11.5.1).

It's time to put things together. By Claim 11.5.14, we can bound $|\mathfrak{D}|$ by

$$|\mathfrak{D}| \leqslant |\mathfrak{U}| - |\mathfrak{U} \setminus \mathfrak{D}|.$$

We have $|\mathfrak{U} \setminus \mathfrak{D}| \geqslant |S'_\ell| + |S'_r| - a + |\mathfrak{U} \setminus (\{u_\ell^{(1)}, \dots, u_\ell^{(|S'_\ell|)}, u_r^{(1)}, \dots, u_r^{|S'_r|}\} \cup \mathfrak{D})|$, and $|\mathfrak{U}| = |q_{\ell,m,r}| + |q_{\ell,r}| + |q_{\ell,m}| + |q_{r,m}| + |S_\ell \cup S_r|$. This gives us

$$\begin{aligned}|\mathfrak{D}| \leqslant |q_{\ell,m,r}| + |q_{\ell,r}| + |q_{\ell,m}| + |q_{r,m}| + |S_\ell \cup S_r| - |S'_\ell| - |S'_r| + a - \\ |\mathfrak{U} \setminus (\{u_\ell^{(1)}, \dots, u_\ell^{(|S'_\ell|)}, u_r^{(1)}, \dots, u_r^{|S'_r|}\} \cup \mathfrak{D})|\,.\end{aligned}$$

Adding (11.5.1) to both sides and rearranging, we get

$$\begin{aligned}p - p' + |\mathfrak{D}| \leqslant 2|q_{\ell,m,r}| + |S_\ell \cap (\mathcal{V}(\mathcal{R}_r) \setminus S_r)| + |S_r \cap (\mathcal{V}(\mathcal{R}_\ell) \setminus S_\ell)| + |q_{\ell,r}| + |q_{\ell,m}| + |q_{r,m}| \\ + |S_\ell \cup S_r| - |S'_\ell| - |S'_r| + |S_\ell \cap S_r|\,,\end{aligned}$$

and substituting $r = 2|q_{\ell,m,r}| + |S_\ell \cap (\mathcal{V}(\mathcal{R}_r) \setminus S_r)| + |S_r \cap (\mathcal{V}(\mathcal{R}_\ell) \setminus S_\ell)| + |q_{\ell,r}| + |q_{\ell,m}| + |q_{r,m}|$ gives

$$p - p' + |\mathfrak{D}| \leqslant r + |S_\ell \cup S_r| - |S'_\ell| - |S'_r| + |S_\ell \cap S_r|\,.$$

Notice that $|S_\ell \cup S_r| + |S_\ell \cap S_r| = |S_\ell| + |S_r|$, so we can rearrange to obtain the lemma. □

Now we can prove Lemma 11.5.10.

Proof of Lemma 11.5.10. First of all, we note that $c_i(\mathcal{R}_m)$ depends only on the shape of $\mathcal{R}_m$ by symmetry of our construction. We turn to the quantitative bound.

The proof is by induction. The coefficients $c_0(\mathcal{R}_m)$ are nonzero only for ribbons $\mathcal{R}_m$ which have $\mathcal{Z}(\mathcal{R}_m) = \emptyset$ and admitting $|S_\ell| = |S_r| = p$ paths from S_ℓ to S_r.

Thus in the case that $i = 0$, the statement reduces to $c_0(\mathcal{R}_m) \leqslant 1$, which is true by definition.

Suppose the lemma holds for c_i, and consider c_{i+1}. By definition, for an (improper) S'_ℓ, S'_r-ribbon $\mathcal{R}'_m$ and ribbons $\mathcal{R}'_\ell, \mathcal{R}'_r$ satisfying 1 and 2,

$$c_{i+1}(\mathcal{R}'_m) = \sum_{\substack{\mathcal{R}_\ell,\mathcal{R}_m,\mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \text{ for some } S_\ell,S_r \\ r \text{ intersections outside } S_\ell,S_r \\ \text{separating factorization } \mathcal{R}'_\ell,\mathcal{R}'_m,\mathcal{R}'_r,S'_\ell,S'_r}} c_i(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^r . \tag{11.5.2}$$

We introduce the shorthand $s' = \frac{|S'_\ell|+|S'_r|}{2}$. Consider first a particular term in the sum, $c_i(\mathcal{R}_m)(\omega/n)^r$, where $\mathcal{R}_m$ is an improper S_ℓ, S_r ribbon, and let $|\mathfrak{D}| = |\mathcal{Z}(\mathcal{R}'_m) \setminus \mathcal{Z}(\mathcal{R}_m)|$. By induction and Lemma 11.5.12,

$$\begin{aligned}
&\left(\frac{\omega}{n}\right)^r \cdot c_i(\mathcal{R}_m) \\
&\leqslant \left(\frac{\omega}{n}\right)^r \cdot \left(\frac{\omega}{n}\right)^s \cdot n^{\frac{p-|\mathcal{Z}(\mathcal{R}_m)|-i/2}{2}+\varepsilon s} && \text{by induction} \\
&= \left(\frac{\omega}{n}\right)^{s'} \cdot \left(\frac{\omega}{n}\right)^{r-s'+s} \cdot \cdot n^{\frac{p-|\mathcal{Z}(\mathcal{R}_m)|-i/2}{2}+\varepsilon s} \\
&= \left(\frac{\omega}{n}\right)^{s'} \cdot n^{-\varepsilon(r-s'+s)} \cdot n^{-\frac{1}{2}(r-s'+s)} \cdot n^{\frac{p-|\mathcal{Z}(\mathcal{R}_m)|-i/2}{2}+\varepsilon s} && \text{using } \omega = n^{1/2-\varepsilon} \\
&\leqslant \left(\frac{\omega}{n}\right)^{s'} \cdot n^{-\varepsilon(r-s'+s)} \cdot n^{-\frac{1}{2}(s'-s+p-p'+|\mathfrak{D}|)} \cdot n^{\frac{p-|\mathcal{Z}(\mathcal{R}_m)|-i/2}{2}+\varepsilon s} && \text{by Lemma 11.5.12} \\
&= \left(\frac{\omega}{n}\right)^{s'} \cdot n^{-\varepsilon(r-s'+s)} \cdot n^{\frac{p'-|\mathcal{Z}(\mathcal{R}'_m)|-i/2-s'+s}{2}+\varepsilon s} && \text{canceling terms, using } |\mathcal{Z}(\mathcal{R}'_m)| = |\mathfrak{D}| + |\mathcal{Z}(\mathcal{R}_m)| \\
&= n^{-\varepsilon r} \cdot \left(\frac{\omega}{n}\right)^{s'} \cdot n^{\frac{p'-|\mathcal{Z}(\mathcal{R}'_m)|-i/2-(s'-s)}{2}+\varepsilon s'} \\
&\leqslant n^{-\varepsilon r} \cdot \left(\frac{\omega}{n}\right)^{s'} \cdot n^{\frac{p'-|\mathcal{Z}(\mathcal{R}'_m)|-(i+1)/2}{2}+\varepsilon s'} && \text{using } s'-s \geqslant 1/2, \text{ by Lemma 11.4.13}
\end{aligned}$$

Next we assess how many nonzero terms are in the sum (11.5.2) for a fixed r and a fixed $\mathcal{R}'_m$. For each vertex of $\mathcal{R}'_m$, there are 7 possibilities for which ribbon(s) it came from in $\{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r\}$ so there are at most 7^τ choices overall (recall that

$\mathcal{R}'_m$ has at most τ vertices for the terms we are looking at). Once we have chosen which ribbon(s) each vertex of $\mathcal{R}'_m$ came from, everything is fixed except for possible edges of $\mathcal{R}'_m$ which appear at least twice in $\mathcal{R}_\ell$, $\mathcal{R}_m$, and $\mathcal{R}_r$. There are two possibilities for each possible edge of $\mathcal{R}'_m$ which appears twice in $\mathcal{R}_\ell$, $\mathcal{R}_m$, and $\mathcal{R}_r$ and four possibilities for each possible edge of $\mathcal{R}'_m$ which appers three times in $\mathcal{R}_\ell$, $\mathcal{R}_m$, and $\mathcal{R}_r$. However, note that any such edge must be between an intersected vertex and either another intersected vertex or a vertex in $S_\ell \cup S_r$. Thus, there are at most $r\tau$ possible edges of $\mathcal{R}'_m$ which appear at least twice in $\mathcal{R}_\ell$, $\mathcal{R}_m$, and $\mathcal{R}_r$ and the total number of possibilities for these edges is at most $4^{r\tau}$.

All together there are at most $2^{O(r\tau)}$ nonzero terms for fixed r. This means that the total contribution from such terms is at most

$$2^{O(r\tau)} \cdot n^{-\varepsilon r} \cdot \left(\frac{\omega}{n}\right)^{s'} \cdot n^{\frac{p'-|\mathcal{Z}(\mathcal{R}'_m)|-(i+1)/2}{2}+\varepsilon s'}$$

As long as $\tau \leqslant (\varepsilon/C)\log n$ for some universal constant C, we have $2^{O(r\tau)} \cdot n^{-\varepsilon r} \ll 1/\tau$ for all $r \geqslant 1$. All in all, we obtain

$$c_{i+1}(\mathcal{R}'_m) \leqslant \left(\frac{\omega}{n}\right)^{s'} \cdot n^{\frac{p'-|\mathcal{Z}(\mathcal{R}'_m)|-(i+1)/2}{2}+\varepsilon s'}$$

which completes the induction. □

11.5.4 $\mathcal{L}\,\mathcal{L}^\top$ is Well-Conditioned — Proof of Lemma 11.5.3

In this section we prove Lemma 11.5.3, restated here.

Lemma (Restatement of Theorem 11.5.3)**.** *With high probability,* $\Pi\,\mathcal{L}\,\Pi\,\mathcal{L}^\top\,\Pi \geq \Omega(\omega/n)^{d+1} \cdot \Pi$*, where as usual* Π *is the projector to* $\mathrm{Span}\{e_C \,:\, C \in C_{\leqslant d}\}$.

Proof of Lemma 11.5.3. We recall the definition of $\mathcal{L}$.

$$\mathcal{L}(I,S) = \left(\frac{\omega}{n}\right)^{-\frac{|S|}{2}} \sum_{\substack{\mathcal{R} \text{ having } 1 \\ |\mathcal{V}(\mathcal{R}_\ell)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|} \chi_{\mathcal{R}_\ell}\,.$$

Consider a diagonal entry $\mathcal{L}(S,S)$. Since every ribbon $\mathcal{R}$ appearing in its expansion must have 1, in particular it has no edges inside S. Thus, by the same argument as in Lemma 11.3.3, with probability at least $1 - O(n^{-10\log n})$,

$$\mathcal{L}(S,S) = \left(\frac{\omega}{n}\right)^{\frac{|S|}{2}} (1 \pm n^{-\Omega(\varepsilon)})\,.$$

Let $\mathcal{L}^{\text{off-diag}}$ be given by

$$\mathcal{L}^{\text{off-diag}}(I,S) = \begin{cases} \mathcal{L}(I,S) \text{ if } I \neq S \\ 0 \text{ otherwise} \end{cases}.$$

We will consider the block of $\mathcal{L}^{\text{off-diag}}$ with rows indexed by sets of size s_ℓ and columns indexed by sets of size s_r for some $s_\ell, s_r \leqslant d$. For a fixed $t \leqslant \tau$, let $U_1^{(s_\ell,s_r,t)}, \ldots, U_q^{(s_\ell,s_r,t)}$ be all the graphs on vertex set $[t]$ with distinguished subsets of vertices A, B of size s_ℓ, s_r respectively, and where

- $A \neq B$,
- there are no edges inside B,
- every vertex in U outside $A \cup B$ is reachable from A without passing through B, and
- B is the unique minimum-size vertex separator in U separating A from B.

Then let $M_i^{(s_\ell,s_r,t)}$ be given by

$$M_i^{(s_\ell,s_r,t)}(I,S) = \left(\frac{\omega}{n}\right)^{t-\frac{s_r}{2}} \cdot \sum_{\mathcal{R} \text{ an } (I,S)\text{-ribbon with shape } U_i^{(s_\ell,s_r,t)}} \chi_{\mathcal{R}}\,.$$

By assumption on $U_i^{(s_\ell,s_r,t)}$, there are s_r vertex-disjoint paths from A to B. Let $r = |A \cap B|$. By Lemma 11.5.8, with probability at least $1 - O(n^{-100\log n})$,

$$\begin{aligned}\left\|M_i^{(s_\ell,s_r,t)}\right\| &\leqslant \left(\frac{\omega}{n}\right)^{\frac{s_r}{2}} \cdot \left(\frac{\omega}{n}\right)^{t-s_r} \cdot n^{\frac{t-s_r}{2}} \cdot 2^{O(t)} \cdot (\log n)^{O(t-r+(s_r-r))} \\ &= \left(\frac{\omega}{n}\right)^{\frac{s_r}{2}} \cdot n^{-\varepsilon(t-s_r)} \cdot 2^{O(t)} \cdot (\log n)^{O(t-s_r)},\end{aligned}$$

where in the last step we have used that $t \geqslant s_\ell + s_r - r$ and $s_r \leqslant s_\ell$, which holds by the vertex-separator requirement on B. There are at most $2^{\binom{t}{2}-\binom{s_r}{2}+O(t)}$ choices for $U_i^{(s_\ell,s_r,t)}$ when s_ℓ, s_r, t are fixed, by the requirement that U have no edges inside B. Summing over all q for a fixed t, we get by triangle inequality

$$\left\|\sum_{i\leqslant q} M_i^{(s_\ell,s_r,t)}\right\| \leqslant \left(\frac{\omega}{n}\right)^{\frac{s_r}{2}} \cdot 2^{\binom{t}{2}-\binom{s_r}{2}+O(t)} \cdot n^{-\varepsilon(t-s_r)} \cdot (\log n)^{O(t-s_r)}$$

with probability $1 - O(n^{-99\log n})$. By our assumptions on d, τ, and ε, this is at most $(\omega/n)^{s_r/2} \cdot 1/d^4$.

The following standard manipulations now prove the lemma. Let $D' \in \mathbb{R}^{\binom{[n]}{\leqslant d}}$ be the diagonal matrix with $D'(S,S) = (\omega/n)^{|S|/2}$ if S is a clique in G and 0 otherwise. Then we can decompose $\mathcal{L} = D + E + \mathcal{L}^{\text{off-diag}}$, where E is a diagonal matrix with $|E(S,S)| \leqslant n^{-\Omega(\varepsilon)} \cdot (\omega/n)^{|S|/2}$. Then we have

$$\begin{aligned}\Pi\,\mathcal{L}\,\Pi\,\mathcal{L}^\top\,\Pi = D^2 \\ + \Pi(D\Pi\,\mathcal{L}^{\text{off-diag}} + D\Pi E + E\Pi D + E\Pi\,\mathcal{L}^{\text{off-diag}} + \mathcal{L}^{\text{off-diag}}\,\Pi D + \mathcal{L}^{\text{off-diag}}\,\Pi E \\ + E\Pi E + \mathcal{L}^{\text{off-diag}}\,\Pi\,\mathcal{L}^{\text{off-diag}})\Pi\end{aligned}$$

Each of the above matrices aside from D^2 is a $d \times d$ block matrix, where the (s_ℓ, s_r) block is $\binom{[n]}{s_\ell} \times \binom{[n]}{s_r}$ dimensional and has norm at most $(\omega/n)^{(s_\ell+s_r)/2} \cdot d^{-4}$. By the same argument as in the proof of Lemma 11.5.2, using Cauchy-Schwarz to combine the d^2 blocks, we obtain the lemma. □

11.5.5 High-Degree Matrices Have Small Norms

In this section we prove Lemma 11.5.4, restated here:

Lemma (Restatement of Theorem 11.5.4)**.** *With high probability,* $\|\xi_0 - \ldots + \xi_{2d}\| \leqslant n^{-16d}$.

We recall the definition of ξ_i. For a coefficient function on ribbons $c_{i-1}(\mathcal{R}_m)$, we have a matrix $\mathcal{E}$ given by

$$\mathcal{E}(I,J) = \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell|, |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \\ |\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau \\ \text{separating factorization} \\ \mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, S'_\ell, S'_r}} c_{i-1}(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r},$$

and another one, $\mathcal{E}'$, given by

$$\mathcal{E}'(I,J) = \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell|, |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \text{ satisfying} \\ 1,3^*,2 \text{ and not } 4 \\ \text{separating factorization} \\ \mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r, S'_\ell, S'_r \\ |\mathcal{V}(\mathcal{R}'_\ell)|, |\mathcal{V}(\mathcal{R}'_m)|, |\mathcal{V}(\mathcal{R}'_r)| \leqslant \tau}} c_{i-1}(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \cdot \chi_{\mathcal{R}'_\ell} \cdot \chi_{\mathcal{R}'_m} \cdot \chi_{\mathcal{R}'_r} \cdot$$

Then the matrix ξ_i is given by $\mathcal{E} - \mathcal{E}'$.

We will actually prove a bound on the Frobenious norm of each matrix ξ_i. The following will allow us to control the magnitude of the entries. It follows immediately from our concentration bound Lemma 11.6.1, which is proved via the moment method. (Under the slightly stronger assumption $\tau \ll \varepsilon \log n / \log\log n$, it would also follow from standard hypercontractivity.)

Lemma 11.5.17. *Suppose c_T are a collection of coefficients, one for each $T \subseteq \binom{[n]}{2}$, and there is a constant C such that*

1. *If $|T| > C\tau$ then $c_T = 0$.*
2. *Otherwise, $|c_T| \leqslant (\omega/n)^{|T|/C - Cd}$.*

Then with probability at least $1 - O(n^{-100 \log n})$ it occurs that $\left|\sum_{T \subseteq \binom{[n]}{2}} c_T \cdot \chi_T\right| \leqslant n^{-20d}$.

We will also need several facts about the coefficients of ribbons in the expansion of each matrix ξ_i.

Lemma 11.5.18. *Every triple $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$ appearing with nonzero coefficient in ξ_c satisfies $|\mathcal{V}(\mathcal{R}_\ell)| + |\mathcal{V}(\mathcal{R}_m)| + |\mathcal{V}(\mathcal{R}_r)| = \Theta(\tau)$.*

Proof. To appear with nonzero coefficient, the triple $\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r$ with separating factorization $\mathcal{R}'_\ell, \mathcal{R}'_m, \mathcal{R}'_r$ must either have

$$|\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, |\mathcal{V}(\mathcal{R}_r)| \leqslant \tau \quad \text{but} \quad |\mathcal{V}(\mathcal{R}'_\ell)| > \tau \text{ or } |\mathcal{V}(\mathcal{R}'_m)| > \tau \text{ or } |\mathcal{V}(\mathcal{R}'_\ell)| > \tau\,,$$

or

$$|\mathcal{V}(\mathcal{R}'_\ell)|, |\mathcal{V}(\mathcal{R}'_m)|, |\mathcal{V}(\mathcal{R}'_r)| \leqslant \tau \quad \text{but} \quad |\mathcal{V}(\mathcal{R}_\ell)| > \tau \text{ or } |\mathcal{V}(\mathcal{R}_m)| > \tau \text{ or } |\mathcal{V}(\mathcal{R}_\ell)| > \tau\,.$$

In the first case, we must have one of $|\mathcal{V}(\mathcal{R}_\ell)| \geqslant \tau/3$ or $|\mathcal{V}(\mathcal{R}_m)| \geqslant \tau/3$ or $|\mathcal{V}(\mathcal{R}_r)| \geqslant \tau/3$. In the second, we must have $|\mathcal{V}(\mathcal{R}_\ell)|, |\mathcal{V}(\mathcal{R}_m)|, \mathcal{V}(\mathcal{R}_r)| \leqslant 3\tau$. □

We are prepared to prove Lemma 11.5.4.

Proof of Lemma 11.5.4. We will apply Lemma 11.5.17 to $\xi_i(I, J)$ for each $i \leqslant 2d$ and $I, J \subseteq [n]$ with $|I|, |J| \leqslant d$. So consider the Fourier expansion of $\xi_i(I, J)$, given

by

$$\xi_i(I,J) = \sum_{T \subseteq \binom{[n]}{2}} c_T \cdot \chi_T \,.$$

From Lemma 11.5.18, we obtain that if $|T| > C\tau$ then $c_T = 0$, for some absolute constant C. For smaller T we need a bound on the magnitude $|c_T|$. The coefficient c_T is bounded by

$$|c_T| \leqslant \sum_{\substack{S_\ell, S_r \subseteq [n] \\ |S_\ell|, |S_r| \leqslant d}} \left(\frac{\omega}{n}\right)^{-\frac{|S_\ell|+|S_r|}{2}} \sum_{\substack{\mathcal{R}_\ell, \mathcal{R}_m, \mathcal{R}_r \\ \text{nonzero in } \xi_i(I,J) \text{ as in } 11.5.18 \\ \chi_{\mathcal{R}_\ell} \cdot \chi_{\mathcal{R}_m} \cdot \chi_{\mathcal{R}_r} = \chi_T}} c_{i-1}(\mathcal{R}_m) \left(\frac{\omega}{n}\right)^{|\mathcal{V}(\mathcal{R}_\ell)|+|\mathcal{V}(\mathcal{R}_r)|+|\mathcal{V}(\mathcal{R}_m)|-\frac{|S_\ell|+|S_r|}{2}} \tag{11.5.3}$$

By Lemma 11.5.10, we have $c_{i-1}(\mathcal{R}_m) \leqslant n^d \leqslant (\omega/n)^{-2d}$. At the same time, there are at most $2^{O(\tau^2)}$ nonzero terms in the sum (11.5.3). Thus by Lemma 11.5.18 and our assumptions on d, τ, and ε, the coefficient c_T is at most $(\omega/n)^{\tau/C - Cd}$ for some absolute constant C.

Applying Lemma 11.5.17, we obtain $|\xi_i(I,J)| \leqslant n^{-20d}$ with probability $1 - O(n^{-100\log n})$. Taking a union bound over all $n^{2d} \leqslant n^{2\log n}$ entries of ξ_i, and over all $i \leqslant 2d$, we obtain that $\|\xi_0 - \ldots + \xi_{2d}\| \leqslant \|\xi_0 - \ldots + \xi_{2d}\|_F \leqslant n^{-16d}$ with probability $1 - O(n^{-96\log n})$. □

11.6 Omitted Proofs

11.6.1 Concentration for Linear Constraints

In this section we prove Lemma 11.3.3. We will use the following elementary concentration bound repeatedly. (It is the scalar version of the matrix concentra-

tion bound Lemma 11.5.8; we state and prove a scalar version here because it is a good warmup for Lemma 11.5.8.)

Lemma 11.6.1. *Let $\mathcal{T}$ be a family of subsets of $\binom{[n]}{2}$ so that for every $T, T' \in \mathcal{T}$ there exists $\sigma : [n] \to [n]$ a permutation of vertices so that $\sigma(T) = T'$. Let t be the number of vertices incident to edges in any $T \in \mathcal{T}$. For every $s \geqslant 0$ and every even ℓ,*

$$\mathop{\mathbb{P}}_{G\sim G(n,1/2)}\left\{\left|\sum_{T\in\mathcal{T}} \chi_T(G)\right| \leqslant s\right\} \geqslant 1 - \frac{n^{t\ell/2}\cdot(t\ell)^{t\ell}}{s^\ell}.$$

Proof. Let $\ell \in \mathbb{N}$ be a parameter to be chosen later. We will estimate $\mathbb{E}_{G\sim G(n,1/2)}[(\sum_{T\in\mathcal{T}} \chi_T)^\ell]$.

$$\begin{aligned}\mathop{\mathbb{E}}_{G\sim G(n,1/2)}\left[\left(\sum_{T\in\mathcal{T}}\chi_T\right)^\ell\right] &= \sum_{T_1,\dots,T_\ell\in\mathcal{T}} \mathop{\mathbb{E}}_{G\sim G(n,1/2)} \prod_{j\leqslant\ell}\chi_{T_j}\\ &= \left|\{(T_1,\dots,T_\ell) : \mathbb{E}\prod_{j\leqslant\ell}\chi_{T_j} = 1\}\right|.\end{aligned}$$

In order to have $\mathbb{E}\prod_{j\leqslant\ell}\chi_{T_j} = 1$, every edge in the multiset $\bigcup_{j\leqslant\ell} T_j$ must appear at least twice, so every vertex in the multiset $\bigcup_{j\leqslant\ell}\mathcal{V}(T_j)$ also appears at least twice. Thus, this multiset contains at most $t\ell/2$ distinct vertices. Since each $T_j \in \mathcal{T}$, each is uniquely determined by an ordered tuple of t elements of $[n]$. Thus, there are at most $n^{t\ell/2}\cdot(t\ell)^{t\ell}$ distinct choices for $(T_1,\dots,T_\ell)$, so

$$\mathop{\mathbb{E}}_{G\sim G(n,1/2)}\left[\left(\sum_{T\in\mathcal{T}}\chi_T\right)^\ell\right] \leqslant n^{t\ell/2}\cdot(t\ell)^{t\ell}.$$

For even ℓ, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{T\in\mathcal{T}}\chi_T\right| > s\right\} &= \mathbb{P}\left\{\left|\sum_{T\in\mathcal{T}}\chi_T\right)\right|^\ell > s^\ell\right\}\\ &\leqslant \frac{n^{t\ell/2}\cdot(t\ell)^{t\ell}}{s^\ell}.\end{aligned}$$

□

Lemma (Restatement of Theorem 11.3.3)**.** *With high probability,* $\tilde{\mathbb{E}}[1] = 1 \pm n^{-\Omega(\varepsilon)}$ *and* $\tilde{\mathbb{E}}[\sum_{i\in[n]} x_i] = \omega \cdot (1 \pm n^{-\Omega(\varepsilon)})$.

Proof. We will prove the statement regarding $\tilde{\mathbb{E}}[1]$; the bound for $\tilde{\mathbb{E}}[\sum_{i\in[n]} x_i]$ is almost identical.

Recall the Fourier expansion

$$\tilde{\mathbb{E}}[1] - 1 = \sum_{\substack{T \subseteq \binom{[n]}{2} \\ 2 \leqslant |\mathcal{V}(T)| \leqslant \tau}} \left(\frac{\omega}{n}\right)^{|\mathcal{V}(T)|} \cdot \chi_T \,.$$

Considering each $T \subseteq \binom{[n]}{2}$ as a graph, we partition $\{T \subseteq \binom{[n]}{2} : |\mathcal{V}(T)| = t\}$ into p_t families $\{\mathcal{T}_i^t\}_{i=1}^{p}$ by placing T and T' in the same family iff there exists a permutation $\sigma : [n] \to [n]$ of vertices so that $\sigma(T) = T'$. Thus,

$$\tilde{\mathbb{E}}[1] - 1 = \sum_{t=2}^{\tau} \left(\frac{\omega}{n}\right)^t \sum_{i=1}^{p_t} \sum_{T\in\mathcal{T}_i^t} \chi_T \leqslant \sum_{t=2}^{\tau} \left(\frac{\omega}{n}\right)^t \sum_{i=1}^{p_t} \left| \sum_{T\in\mathcal{T}_i^t} \chi_T \right| .$$

By Lemma 11.6.1 (taking $\ell = (\log n)^2$), and since $t \leqslant \tau \leqslant \log n$, each $\mathcal{T}_i^t$ satisfies

$$\mathbb{P}\left\{ \left| \sum_{T\in\mathcal{T}_i^t} \chi_T \right| < O(n^{t/2} \cdot (\log n)^{3t}) \right\} \geqslant 1 - (\tau \cdot 2^{t^2} \cdot n^{\log n})^{-1} .$$

By a union bound over all $p_t \leqslant 2^{t^2}$ families $\mathcal{T}_i^t$, we get that with high probability,

$$|\tilde{\mathbb{E}}[1] - 1| \leqslant \tau \cdot \max_{t\leqslant\tau} \left(2^{t^2} \cdot \left(\frac{\omega}{\sqrt{n}}\right)^t \right) .$$

For $\tau \leqslant (\varepsilon/2)\log n$ and $\omega = n^{1/2-\varepsilon}$, this is at most $n^{-\Omega(\varepsilon)}$. □

.

11.6.2 Combinatorial Proofs about Ribbons

In this section we prove Lemma 11.4.3, restated here:

Lemma (Restatement of Theorem 11.4.3)**.** *Let $\mathcal{R}$ be an (I, J)-ribbon. There is a unique minimum vertex separator S of $\mathcal{R}$ such that S separates I and Q for any vertex separator Q of $\mathcal{R}$. We call S the* leftmost *separator in $\mathcal{R}$. We define the* rightmost *separator analogously and we denote them by $S_L(\mathcal{R})$ and $S_R(\mathcal{R})$ respectively.*

We start by defining a natural partial order on the set of vertex separators in a ribbon $\mathcal{R}$.

Definition 11.6.2. We write $Q_1 \preccurlyeq Q_2$ for two vertex separators Q_1 and Q_2 of an (I, J)-ribbon $\mathcal{R}$ if Q_1 separates I and Q_2.

Next, we check that the definition above indeed is a partial order.

Lemma 11.6.3. *For any set of minimum vertex separators Q_1, Q_2, Q_3 an (I, J)-ribbon, we have:*

1. $Q_1 \preccurlyeq Q_1$.
2. *If $Q_1 \preccurlyeq Q_2$ and $Q_2 \preccurlyeq Q_3$, then, $Q_1 \preccurlyeq Q_3$.*
3. *If $Q_1 \preccurlyeq Q_2$ and $Q_2 \preccurlyeq Q_1$, then, $Q_1 = Q_2$.*

Proof. The first statement is immediate from the definition. For the second, consider a path P from I to Q_3 in $\mathcal{R}$. Since $Q_2 \preccurlyeq Q_3$, P passes through a vertex in Q_2. Thus, P contains a subpath that connects I and Q_2. But since $Q_1 \preccurlyeq Q_2$, this subpath must pass through Q_1. Thus, any such P must pass through Q_1 and thus, $Q_1 \preccurlyeq Q_3$.

Finally, for the third statement, let $k = |Q_1| = |Q_2|$. Then, using Menger's theorem (Fact 11.2.2, there is a set of k vertex disjoint paths $P_1, P_2, \ldots, P_k$ between I and J. By virtue of Q_1, Q_2 being *minimum* vertex separators of $\mathcal{R}$, Q_1 and Q_2 must intersect each P_i in exactly one vertex. It is then immediate that the only way $Q_1 \preccurlyeq Q_2$ and $Q_2 \preccurlyeq Q_1$ if every P_i intersects Q_1, Q_2 in the same vertex. □

Now we can prove Lemma 11.4.3.

Proof of Lemma 11.4.3. It is enough to show that for any two minimum separators Q_1, Q_2 of size k in R, there are separators Q_L, Q_R such that $Q_L \preccurlyeq Q_1 \preccurlyeq Q_R$ and $Q_L \preccurlyeq Q_2 \preccurlyeq Q_R$. We now construct Q_L and Q_R as required.

Let $U = Q_1 \cap Q_2$ and $V = Q_1 \Delta Q_2$. Let $W_L \subseteq V$ be the set of vertices w such that there is a path from I to w that doesn't pass through $Q_1 \cup Q_2$. Similarly, let $W_R \subseteq V$ be the set of vertices such that there is a path from w to some vertex in J that doesn't pass through any vertex in $Q_1 \cup Q_2$. Then we first observe:

Claim 11.6.4. $W_L \cap W_R = \emptyset$.

Proof of Claim. Assume otherwise and let $w \in W_L \cap W_R$. Then there is a path between I and J that doesn't go through any vertex in at least one of Q_1 or Q_2 contradicting that both are in fact vertex separators. □

Next, we have:

Claim 11.6.5. Let $Q_L = U \cup W_L$ and $Q_R = U \cup W_R$. Then Q_L, Q_R are both vertex separators in R.

Proof of Claim. We only give the argument for Q_L, the other case is similar. Assume there is a path P from I to J that does not pass through Q_L. P must

intersect $Q_1 \cup Q_2$. Then there is a vertex $v \in Q_1 \cup Q_2$ such that there is a path I to v which intersects no other vertices in $Q_1 \cup Q_2$. This implies that either $v \in U$ or $v \in W_L$. But by our construction of W_L this is a contradiction. □

Finally, we note that both Q_L, Q_R must in fact be *minimum* vertex separators.

Claim 11.6.6. $|Q_L| = |Q_R| = |Q_1| = |Q_2| = k$

Proof of Claim. Let $|Q_1| = |Q_2| = k$. Then $2k = |Q_1| + |Q_2| = 2|U| + |V| \geqslant 2|U| + |W_L| + |W_R| = |U \cup W_L| + |U \cup W_R| = |Q_L| + |Q_R|$. Since Q_L and Q_R are vertex separators, $|Q_L|, |Q_R| \geqslant k$. Thus, $|Q_L| = |Q_R| = k$. □

Finally, we have the ordering requirement on Q_L and Q_R.

Claim 11.6.7. $Q_L \leqslant Q_1$ and $Q_2 \leqslant Q_R$.

Proof of Claim. Let P be a path from I to Q_1, let v be the first vertex on this path which is in $Q_1 \cup Q_2$. Then, $v \in U$ or $v \in W_L$. Thus, $Q_L \leqslant Q_1$. The other case is similar. □

This concludes the proof of the lemma. □

11.6.3 Spectral Norms

The results in this section are in essence due to Medarametla and Potechin [127]. For completeness, we state and prove them here in the language and notation of the current paper, with minor modifications as needed.

Lemma (Restatement of Theorem 11.5.8)**.** *Let U be a graph on $t \leqslant O(\log n)$ vertices, with two distinguished subsets of vertices A and B, and suppose:*

- *U admits p vertex-disjoint paths from $A \setminus B$ to $B \setminus A$.*
- *$|A \cap B| = r$.*
- *Every vertex outside $A \cup B$ has degree at least 1.*

Let $M = M(G)$ be the graphical matrix with shape U. Then, whp, $\|M\| \leqslant n^{\frac{t-p-r}{2}} \cdot 2^{O(t)} \cdot (\log n)^{O(t-r+p)}$.

Proof of Lemma 11.5.8. We proceed by the trace power method, with a dependence-breaking step beforehand.

Breaking Dependence Let $q_1, \ldots, q_p$ be vertex-disjoint paths from $A \setminus B$ to $B \setminus A$ in U. Without loss of generality we can take each to intersect $A \setminus B$ and $B \setminus A$ only at its endpoints. We will partition the space of labelings σ into disjoint sets $S_1, \ldots, S_m$. For each S_k there will be a partition V_1^k, V_2^k of $[n]$ so that $\sigma(\bigcup_{j \leqslant p} q_j) \subseteq V_1^k$ and $\sigma(U \setminus (\bigcup_{j \leqslant p} q_j)) \subseteq V_2^k$ for every $\sigma \in S_k$. Let $(V_1^1, V_2^1), \ldots, (V_1^m, V_2^m)$ be a sequence of independent uniformly random partitions of $[n]$. Call a labeling σ *good* at k if the preceeding conditions apply to σ for the partition V_1^k, V_2^k and not for any $V_1^{k'}, V_2^{k'}$ for some $k' < k$. Let $S_k = \{\sigma : \sigma \text{ is good at } k\}$.

Claim 11.6.8. There is $m = O(2^t \cdot t \cdot \log n)$ so that $\bigcup_{k=1}^m S_k$ contains every labeling $\sigma : U \to G$.

Proof. For a fixed σ,

$$\mathbb{P}\{\sigma \text{ not good for some } k \leqslant m\} \leqslant (1 - 2^{-t})^m$$

since every vertex $u \in U$ is in V_i with probability $1/2$. If $m \geqslant 10t2^t \log n$, then by a union bound over all $\sigma : U \to G$ (of which there are at most n^t), we get $\mathbb{P}\{$ all σ good for some $k \leqslant m \} > 0$. □

Henceforth, let $S_1, \ldots, S_m$ be the partition guaranteed by the preceeding claim. For $k \leqslant m$, let $M_k(I, J) = \sum_{\sigma \in S_k : \sigma(A)=I, \sigma(B)=J} \mathrm{val}(\sigma)$. Then $M = \sum_{k=1}^{m} M_k$.

Moment Calculation Let $\ell = \ell(n)$ be a parameter to be chosen later. By the triangle inequality, $\|M\| \leqslant \sum_{k=1}^{m} \|M_k\|$. Fix k. We expand $\mathbb{E}_G \operatorname{Tr}(M_k^\top M_k)^\ell$ as

$$\mathbb{E} \operatorname{Tr}(M_k^\top M_k)^\ell = \mathbb{E} \sum_{\substack{\sigma_1, \ldots, \sigma_{2\ell} \in S_k \\ \sigma_{2i}(A) = \sigma_{2i-1}(A) \\ \sigma_{2i}(B) = \sigma_{2i+1}(B)}} \prod_{j=1}^{2\ell} \mathrm{val}(\sigma_j) \,.$$

(Here arithmetic with indices i is modulo 2ℓ, so for example we take $2i + 1 = 1$.) For any σ,

$$\mathrm{val}(\sigma) = \prod_{(i,j) \in U} G_{\sigma(i), \sigma(j)} \,.$$

Notice that for all $\sigma_1, \ldots, \sigma_{2\ell}$, the expectation $\mathbb{E} \prod_{j=1}^{2\ell} \mathrm{val}(\sigma_j)$ is either 0 or 1. We will bound the number of $\sigma_1, \ldots, \sigma_{2\ell}$ for which $\mathbb{E} \prod_{j=1}^{2\ell} \mathrm{val}(\sigma_j) = 1$ by bounding the number of distinct labels such a family of labelings may assign to vertices in U.

Fix $\sigma_1, \ldots, \sigma_{2\ell} \in S_k$. Consider the family $q_1, \ldots, q_p$ of vertex-disjoint paths. Every edge in every q_j receives one pair of labels from each σ_i. Consider these labels arranged on 2ℓ adjoined copies of each q_j, one for each σ (giving p paths with $2\ell \sum_{j \leqslant p} |q_j|$ edges in total, where $|q_j|$ is the number of edges in q_j). Every pair of labels $\{\sigma_i(v), \sigma_i(w)\}$ appearing on an edge (v, w) in this graph must also appear on some distinct edge (v', w') in order to have $\mathbb{E} \prod_{i=1}^{2\ell} \mathrm{val}(\sigma_i) = 1$; otherwise the disjointness of V_1^k, V_2^k would be violated. Merging edges which

received the same pair of labels, we arrive at a graph with at most p connected components and at most $\ell \sum_{j \leqslant p} |q_j|$ edges, and so at most $\ell \sum_{j \leqslant p} |q_j| + p$ vertices. Thus, the vertices in $q_1, \ldots, q_p$ together receive at most $\ell \sum_{j \leqslant p} |q_j| + p$ distinct labels among all $\sigma_1, \ldots, \sigma_{2\ell}$.

Next we account for labels of $v \notin (\bigcup_{j \leqslant p} q_j \cup A \cup B)$. If $\mathbb{E}_G \prod_{i=1}^{2\ell} \mathrm{val}(\sigma_i) = 1$ then the 2ℓ-size multiset $\{\sigma_i(v)\}_{i \leqslant 2\ell}$ of labels for such v contains at most ℓ distinct labels, since by assumption v has degree at least 1 in U.

Next we account for labels of vertices in $A \setminus (B \cup \bigcup_{j \leqslant p} q_j)$ and $B \setminus (A \cup \bigcup_{j \leqslant p} q_j)$. Every such vertex receives a label from every σ_i, but σ_{2i} and σ_{2i-1} must agree on A-labels and σ_{2i} and σ_{2i+1} must agree on B-labels. So in total there are at most $\ell(|A| + |B| - 2p - 2r)$ distinct labels for such vertices.

This means that among the labels $\sigma_i(j)$ for all $j \notin A \cap B$, there are at most

$$\underbrace{\ell \sum_{j \leqslant p} |q_j| + p}_{\text{labels from paths}} \quad + \quad \underbrace{\ell(|A| + |B| - 2p - 2r)}_{\text{additional vertices in } A \cup B \setminus (A \cap B)}$$

$$\underbrace{+ \ell(t - (|A| + |B| - r) - (\sum_j |q_j| - p)) = \ell(t - p - r) + p}_{\text{vertices in } U \setminus (\bigcup_j q_j \cup A \cup B)}$$

unique labels.

Finally, consider the labels of the r vertices $j_1, \ldots, j_r$ in $A \cap B$. The first labelling σ_1 assigns these vertices some $\sigma_1(j_1), \ldots, \sigma_1(j_r)$ labels in G. Since σ_2 agrees with σ_1 on A-vertices, we must have $\sigma_2(j_1) = \sigma_1(j_1), \ldots, \sigma_1(j_r) = \sigma_2(j_r)$. Since σ_3 agrees with σ_2 on B-vertices, we must have $\sigma_3(j_1) = \sigma_2(j_1), \ldots, \sigma_3(j_r) = \sigma_2(j_r)$, and so on. So there are at most r unique labels for such vertices.

Now we can assess how many choices there are for $\sigma_1, \ldots, \sigma_{2\ell} \in S_k$ so that $\mathbb{E} \prod_{i \leqslant 2\ell} \mathrm{val}(\sigma_i) = 1$. To choose such a collection $\sigma_1, \ldots, \sigma_{2\ell}$, we proceed in stages.

Stage 1. Choose the labels $\sigma_i(j_1), \ldots, \sigma_i(j_r)$ of all the vertices in $A \cap B$. Here there are at most n^r options.

Stage 2. For each pair (i, j), where $j \notin A \cap B$, choose whether $\sigma_i(j)$ it will be the first appearance of the index $\sigma_i(j) \in [n]$ or if there is some $i' < i$ and j' so that $\sigma_{i'}(j') = \sigma_i(j)$. Here there are $2^{2\ell t}$ options.

Stage 3. Choose the labels $\sigma_i(j) \in [n]$ for all $j \notin A \cap B$ and pairs (i, j) which in Stage 2 we chose to be the first appearance of a label. If there are x such vertices, there are at most n^x options.

Stage 4. Choose the labels $\sigma_i(j) \in [n]$ for all the pairs (i, j), with $j \notin A \cap B$, which in Stage 2 we chose not to be the first apperance of a label. Here there are at most $x^{2\ell t - 2\ell r - x}$ options.

All together, there are at most $n^r \cdot 2^{2\ell t} \cdot n^x \cdot x^{2\ell(t-r)-x} \leqslant n^r \cdot 2^{2\ell t} \cdot n^x \cdot (2\ell t)^{2\ell(t-r)-x}$ choices for a given x. Since $4lt \ll n$, summing up over all $x \leqslant \ell(t-p-r)+p$ the total number of choices is at most $2n^r \cdot 2^{2\ell t} \cdot n^{\ell(t-p-r)+p} \cdot (2\ell t)^{\ell(t-r+p)-p}$. Putting it together,

$$\mathbb{E}\,\mathrm{Tr}(M_k^\top M_k)^\ell \leqslant 2n^r \cdot n^{\ell(t-p-r)+p} \cdot (2\ell t)^{\ell(t-r+p)-p}\,.$$

Now using Markov's inequality and standard manipulations, for any s,

$$\begin{aligned}
\mathbb{P}\{\|M_k\| \geqslant s\} &= \mathbb{P}\{\|M_k^\top M_k\|^\ell \geqslant s^{2\ell}\} \\
&\leqslant \frac{\mathbb{E}\,\|(M_k^\top M_k)^\ell\|}{s^{2\ell}} \quad \text{by Markov's} \\
&\leqslant \frac{\mathbb{E}\,\mathrm{Tr}(M_k^\top M_k)^\ell}{s^{2\ell}} \quad \text{since } \|(M_k^\top M_k)^\ell\| \leqslant \mathrm{Tr}(M_k^\top M_k)^\ell
\end{aligned}$$

$$\leqslant \frac{2n^r \cdot 2^{2\ell t} \cdot n^{\ell(t-p-r)+p} \cdot (2\ell t)^{\ell(t-r+p)-p}}{s^{2\ell}}$$

Taking $\ell = (\log n)^3$ and using $p \leqslant t \leqslant O(\log n)$, there is $s = 2^t \cdot n^{(t-p-r)/2}(\log n)^{O(t-r+p)}$ so that $\mathbb{P}\{\|M_k\| \geqslant s\} \leqslant n^{-100\log n} m^{-1}$. By a union bound, $\mathbb{P}\{\|M_k\| \leqslant s \text{ for all } k\} \geqslant 1 - n^{-100\log n}$, so $\|M\| \leqslant sm$ with probability $1 - n^{-100\log n}$. Since $m \leqslant 2^{O(t)} \cdot \log(n)^{O(1)}$, this completes the proof. □

11.7 Extension to Sparse Principal Component Analysis

Before we overview the proof of Theorem 11.1.3, we offer some further context for the theorem statement.

Remark 11.7.1 (Relation to the spiked-Wigner model of sparse principal component analysis). Theorem 11.1.3 proves an SoS lower bound for a refutation problem. As usual, the refutation problem is closely related to a hypothesis testing problem and an estimation problem, both of which come from the following alternative distribution: the spiked-Wigner model of sparse principal component analysis. Let W be a symmetric matrix with iid entries from $\mathcal{N}(0,1)$, and let v be a random k-sparse unit vector with entries $\{\pm 1/\sqrt{k}, 0\}$. Let $B = W + \lambda vv^\mathsf{T}$. The hypothesis testing problem is to distinguish between a single sample from B and a sample from W.

There are two main algorithms for this problem, both captured by the SoS hierarchy. The first, applicable when $\lambda \gg \sqrt{n}$, is vanilla principal component analysis: the top eigenvalue of B will be larger than the top eigenvalue of W. The second, applicable when $\lambda \gg k$, is diagonal thresholding: the diagonal entries of B which corresponds to nonzero coordinates will be noticeably large [62]. Interpreting Theorem 11.1.3 in the hypothesis testing setting suggests that once λ

is well outside these parameter regimes, i.e. when $\lambda < n^{1/2-\varepsilon}, k^{1-\varepsilon}$ for arbitrarily small $\varepsilon > 0$, even degree $n^{\Omega(\varepsilon)}$, natural SoS programs do not distinguish between B and W.

Remark 11.7.2 (Interpretation as an integrality gap). A second interpretation of Theorem 11.1.3, independent of any planted problem, is as a strong integrality gap for random instances for the problem of maximizing a quadratic form over k-sparse vectors. Consider the actual maximum of $\langle x, Ax \rangle$ for random ($\{\pm 1\}$ or Gaussian) A over k-sparse unit vectors x. There are roughly $2^{k \log n}$ points in a $\frac{1}{2}$-net for such vectors, meaning that by standard arguments,

$$\max_{\|x\|=1, x \text{ is } k\text{-sparse}} \langle x, Ax \rangle \leqslant O(\sqrt{k} \log n) .$$

With the parameters of the theorem, this means that the integrality gap of the degree $n^{\Omega(\varepsilon)}$ SoS relaxation is at least $\min(n^{\rho/2-\varepsilon}, n^{1/2-\rho/2-\varepsilon})$ when $k = n^{\rho}$.

Remark 11.7.3 (Relation to spiked-Wishart model). Theorem 11.1.3 most closely concerns the spiked-Wigner model of sparse PCA. (Here "Wigner" refers to independence of the entries of the matrix A.) Often, sparse PCA is instead studied in the (perhaps more realistic) *spiked-Wishart model*, where the input is m samples $x_1, \ldots, x_m$ from an n-dimensional Gaussian vector $\mathcal{N}(0, \mathrm{Id} + \lambda \cdot vv^{\top})$, where v is a unit-norm k-sparse vector. Here the question is: as a function of the sparsity k, the ambient dimension n, and the signal strength λ, how many samples m are needed to recover the vector v? The natural approach to recovering v in this setting is to solve a convex relaxation of the problem of maximizing he quadratic form of the empirical covariance $M = \sum_{i \leqslant m} x_i x_i^{\top}$ over k-sparse unit vectors (the maximization problem itself is NP-hard even to approximate in the worst case [48]).

Theoretically, one may apply pseudocalibration and our proof techniques from

Theorem 11.1.1 and Theorem 11.1.3 directly to the spiked-Wishart model, but this carries the expense of substantial technical complication. We may however make intelligent guesses about the behavior of SoS relaxations for the spiked-Wishart model on the basis of Theorem 11.1.3 alone. As in the spiked Wigner model, there are essentially two known algorithms to recover a planted sparse vector v in the spiked Wishart model: vanilla PCA and diagonal thresholding [62]. We conjecture that, as in the spiked Wigner model, the SoS hierarchy requires $n^{\Omega(1)}$ degree to improve the number of samples required by these algorithms by any polynomial factor.

Concretely, considering the case $\lambda = 1$ for simplicity, we conjecture that there are constants c, ε^* such that for every $\varepsilon \in (0, \varepsilon^*)$ if $m \leqslant \min(k^{2-\varepsilon}, n^{1-\varepsilon})$ and $x_1, \ldots, x_m \sim \mathcal{N}(0, \mathrm{Id})$ are iid, then with high probability for every $\rho \in (0, 1)$ if $k = n^\rho$,

$$SoS_{d,k}\left(\sum_{i \leqslant m} x_i x_i^\mathsf{T}\right) \geqslant \min(n^{1-\varepsilon}k, k^{2-\varepsilon})$$

for all $d \leqslant n^{c \cdot \varepsilon}$.

This conjecture is supported both by Theorem 11.1.3 and by the lack of successful n^ε-simple statistics in the spiked Wishart model.

11.7.1 Proof Overview for Theorem 11.1.3

Our proof of Theorem 11.1.3 is very similar to the analogous proof for planted clique, Theorem 11.1.1.

We state here just the main PSDness lemma, which describes the pseudocalibration construction in the sparse PCA setting.

Lemma 11.7.4. *Let $d \in \mathbb{N}$ and let $N_d = \sum_{s \leqslant d} n(n-1)\cdots(n-(s-1))$ be the number of $\leqslant d$-tuples with unique entries from $[n]$. Let $\mu(A)$ be the density of the following distribution on $n \times n$ matrices A with respect to the uniform distribution on $\{\pm 1\}^{\binom{n}{2}}$.*

Planted distribution: *Let $k = k(n) \in \mathbb{N}$ and $\lambda = \lambda(n) \in \mathbb{R}$, and $\gamma > 0$, and assume $\lambda \leqslant k$. Sample a uniformly random k-sparse vector $v \in \mathbb{R}^n$ with entries $\pm 1, 0$. Form the matrix $B = vv^\top$. For each nonzero entry of B independently, replace it with a uniform draw from $\{\pm 1\}$ with probability $1 - \lambda/k$ (maintaining the symmetry $B = B^\top$). For each zero entry of B, replace it with a uniform draw from $\{\pm 1\}$ (maintaining the same symmetry). Finally, choose every $i \in [n]$ with probability $n^{-\gamma}$ independently; for those indices that were not chosen, replace every entry in the corresponding row and column of B with random ± 1 entries.[1] Output the resulting matrix A. (We remark that this matrix is a Boolean version of the more standard spiked-Wigner model $B + \lambda vv^\top$ where B has iid standard normal entries and v is a random k-sparse unit vector with entries from $\{\pm 1/\sqrt{k}, 0\}$.)*

Let $\Lambda : \{\pm 1\}^{\binom{n}{2}} \to \mathbb{R}^{N_d \times N_d}$ be the following function

$$\Lambda(A) = \mu(A) \cdot \mathop{\mathbb{E}}_{v|A} v^{\otimes \leqslant 2d}$$

where the expectation is with respect to the planted distribution above. For $D = D(n) \in \mathbb{N}$, let $\Lambda^{\leqslant D}$ be the entrywise projection of Λ into the Boolean functions of degree at most D.

There are constants $C, \varepsilon^* > 0$ such that for every $\gamma > 0$ and $\rho \in (0,1)$ and every $\varepsilon \in (0, \varepsilon^*)$ (all independent of n), if $k = n^\rho$ and $\lambda \leqslant \min\{n^{\rho-\varepsilon}, n^{1/2-\varepsilon}\}$, and if $Cd/\varepsilon < D < n^{\varepsilon/C}$, then for large enough n

$$\mathop{\mathbb{P}}_{A \sim \{\pm 1\}^{\binom{n}{2}}} \{\Lambda^{\leqslant D}(A) \succeq 0\} \geqslant 1 - o(1)\,.$$

[1] This additional $n^{-\gamma}$ noising step is a technical convenience which has the effect of somewhat decreasing the number of nonzero entries of v and decreasing the signal-strength λ.

At this point it is useful to consider a more familiar planted model, which the lemma above mimics. Let W be a $n \times n$ symmetric matrix with iid entries from $\mathcal{N}(0,1)$. Let $v \in \mathbb{R}^n$ be a k-sparse unit vector, with entries in $\{\pm 1/\sqrt{k}, 0\}$. Let $A = W + \lambda vv^\mathsf{T}$. Notice that if $\lambda \gg k$, then diagonal thresholding on the matrix W identifies the nonzero coordinates of v. (This is the analogue of the covariance-thresholding algorithm in the spiked-Wishart version of sparse PCA.) On the other hand, if $\lambda \gg \sqrt{n}$ then (since typically $\|W\| \approx \sqrt{n}$), ordinary PCA identifies v. The lemma captures computational hardness for the problem of distinguishing a single sample from A from a sample from the null model W both diagonal thresholding and ordinary PCA fail.

The differences between the planted distribution in Lemma 11.7.4 and the spiked Wigner distribution for sparse PCA are mainly technical conveniences which help in applying the techniques from the proof of Theorem 11.1.1. In particular we note the trick of rerandomizing each *row* of the matrix B with probability $n^{-\gamma}$. The result of this trick is that moments $\mathbb{E}_B \chi_\alpha(B)$ of B (where $\alpha \subseteq \binom{n}{2}$ and χ_α is a Fourier character) decay like $n^{-\gamma \cdot (\#\text{ of vertices in } \alpha)}$.

11.8 Chapter Notes

Theorem 11.1.1 appeared originally in [29], joint work with Boaz Barak, Jon Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. Theorem 11.1.3 appeared in [88], joint work with Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, David Steurer, and Tselil Schramm.

Prior SoS Lower Bounds for Planted Clique Theorem 11.1.1 was preceded by several less-tight SoS lower bounds for planted clique. Using a dual witness originally constructed by Feige and Krauthgamer to prove lower bounds against the weaker Lovasz-Schrijver hierarchy, Meka, Potechin and Wigderson showed that degree-d SoS cannot refute the existence of $n^{\Omega(1/d)}$-cliques in $G(n, 1/2)$ [71, 128]. Deshpande and Montanari improved the analysis of this certificate for $d = 4$, showing a lower bound against $n^{1/3}$-refutation where Meka, Potechin, and Wigderson showed a lower bound against only $n^{1/4}$-refutation (ignoring subpolynomial factors) [63].

Hopkins, Kothari, Potechin, Raghavendra, and Schramm then showed the first tight SoS lower bounds for degree-4 SoS, proving a lower bound against $\sqrt{n}$-refutation (again ignoring logarithmic factors) [89]. Their lower bound required a new pseudodistribution construction, which can be viewed in retrospect as a special case of the pseudocalibration construction where each entry of the moment matrix is projected to a small number of carefully chosen degree-$O(1)$ graph functions.

Prior Lower Bounds for Sparse PCA The study of computational lower bounds for sparse PCA began with work of Berthet and Rigollet showing a reduction from planted clique to sparse PCA in the regime where the planted spike has about $\sqrt{n}$ nonzero entries [38].

Krauthgamer, Nadler, and Vilenchik showed that degree-2 SoS does not solve sparse PCA up to the information-theoretic limit [109]. Ma and Wigderson proved a similarly-strong lower bound against degree-4 SoS [122]. Our lower bound Theorem 11.1.3 is the first to go beyond *quasipolynomial* hardness for

sparse PCA, suggesting that subexponential time may be required to improve by polynomial factors on guarantees of existing algorithms.

History and Importance of Planted Clique Arising from the 1976 work of Karp [102], the planted clique problem was formally defined by Jerrum [98] and Kucera [112]. The strongest guarantees known to be achievable in polynomial time are originally due to Alon, Krivelevich, and Sudakov, who describe for every constant $\varepsilon > 0$ a polynomial-time algorithm to find planted cliques of size $\omega = \varepsilon\sqrt{n}$ [9].

Over the years planted clique and related problems have found applications to important questions in a variety of areas including community detection [82], finding signals in molecular biology [148], discovering motifs in biological networks [130, 97], computing Nash equilibrium [86, 20], property testing [8], sparse principal component analysis [38], compressed sensing [106], cryptography [100, 15] and even mathematical finance [16].

Thus, the question of whether the currently known algorithms can be improved is of great interest. It is unlikely that lower bounds for planted clique can be derived from conjectured complexity class separations such as $\mathbf{P} \neq \mathbf{NP}$, precisely because it is an average-case problem [72, 41]. Thus, our best evidence for its difficulty comes from showing limitations on powerful *classes* of algorithms. In particular, since many of the algorithmic approaches for this and related problems involve spectral techniques and convex programs, limitations for these types of algorithm are of significant interest.

Our interest in showing SoS lower bounds for planted clique stems from its substantial power in solving similar planted problems (for example, all of the

algorithms presented in the first part of this thesis). In particular, in some key instances SoS appears more powerful than weaker convex hierarchies (e.g. the Lovasz-Schrijver hierarchy) for which planted clique lower bounds were known prior to our work. Thus, prior to the proof of Theorem 11.1.1 there appeared to be a real possibility that SoS could detect slightly sub-polynomial-size planted cliques in polynomial time, or at least beat brute-force search.

CHAPTER 12

EQUIVALENCE OF SOS AND SIMPLE MATRIX STATISTICS

In this chapter we state and prove Theorem 12.1.5, the formal version of Theorem 1.1.2, on equivalence of SoS hypothesis testing algorithms and hypothesis tests based on simple *matrix* statistics. In the language of Chapter 2, a D-simple matrix statistic is the maximum eigenvalue of a matrix whose entries are D-simple statistics. Theorem 12.1.5 can be viewed in (at least) two ways:

Algorithmists's view: For nice-enough hypothesis testing problems, if an efficient SoS-based testing algorithm exists then so does similarly-efficient spectral one, based on the maximum eigenvalue of a matrix whose entries are low-degree polynomials of problem instances. We have already seen examples of this phenomenon: in Chapters 6 and 7 we presented spectral algorithms based on matrix-valued polynomials designed from SoS proofs. Theorem 12.1.5 says that this was no accident: similar spectral algorithms exist for *any* nice-enough hypothesis testing problem.

Complexity theorist's view: Conjecture 2.2.4, which says that superlogarithmic almost-independence fools P, is a statement about all polynomial time algorithms, so we are unlikely to prove it outright any time soon. Theorem 12.1.5 (or rather its converse, as we state it below) specializes Conjecture 2.2.4, replacing all polynomial-time hypothesis tests with polynomial-size SoS-based tests, and replacing D-simple statistics with D-simple matrix statistics (in the definition of almost-independence), and therefore represents substantial evidence in favor of Conjecture 2.2.4. (In fact, perhaps the strongest SoS-based evidence we could hope for would result if Theorem 12.1.5 were strengthened to replace simple matrix statistics with simple statistics

outright – improving the theorem in this way is a fascinating open problem.)

Our proof technique for Theorem 12.1.5 employs pseudocalibration. In our pseudocalibration-based arguments in Chapter 11, the main technical difficulty was in proving that a pseudocalibrated moment matrix is PSD with high probability. Our proof here avoids this difficulty with a new idea: instead of arguing directly about the truncated/low-degree moments of a planted distribution, we find a nearby PSD matrix-valued function to those moments and argue via convex duality that it is still suitable to rule out SoS algorithms.

Chapter Overview To state the main result Theorem 12.1.5 formally requires some definitions, which we tackle in the next section. They formalize (for the first time, as far as the author is aware) a notion of successful SoS hypothesis test, and define a set of noise-robust hypothesis testing problem to which the main theorem applies.

After we prove the main result in Section 12.2 and Section 12.3, we show in Section 12.4 that the conditions of the main theorem hold for two important high-dimensional inference problems: planted clique and the spiked tensor model/tensor principal component analysis. (The conditions of the main theorem actually hold for a wide range of inference problems, including the k-community stochastic block model, sparse PCA, random constraint satisfaction, and more. Verifying this is a series of routine calculations, most of which we omit: they can be found in [88].)

12.1 Main Result

Because our main result covers many hypothesis testing problems simultaneously, stating it with the right level of generality requires some care. Before we state Theorem 12.1.5, we get set up by defining the *robust inference problems* it applies to.

Hypothesis Testing Against Product Distributions

We begin by describing a class of hypothesis testing problems. For $\mathcal{A}$ a set of real numbers, we will use $\mathcal{I} = \mathcal{A}^N$ denote a space of instances indexed by N variables—for the sake of concreteness, it will be useful to think of $\mathcal{I}$ as $\{0,1\}^N$; for example, we could have $N = \binom{n}{2}$ and $\mathcal{I}$ as the set of all graphs on n vertices.

However, the results that we will show here continue to hold in other contexts, where the space of all instances is $\mathbb{R}^N$ or $[q]^N$.

We will study binary hypothesis testing with a pair of distributions on $\mathcal{I} = \mathcal{A}^N$: a null distribution ν, which will always be a product distribution, and an alternative distribution μ. As usual, for us the goal in this setting is to correctly decide with probability better than $1/2$ whether an instance $\mathcal{I} \in \mathcal{I}$ has been sampled from ν or from μ, when $\mathcal{I}$ is chosen with probability $1/2$ from μ and otherwise from ν.

In the hypothesis testing problems we consider, μ is usually a distribution over instances $\mathcal{I}$ with some hidden, or *planted* structure. So, we often call μ the *planted distribution*.

Polynomial Systems

Our goal is to define a broad notion of successful SoS hypothesis testing algorithms. To that effect, we define next a set of hypothesis testing problems which would be solvable given an oracle to solve systems of polynomial equations.

Program 12.1.1 (Polynomial System). Let $\mathcal{A}, \mathcal{B}$ be sets of real numbers, let $n, N \in \mathbb{N}$, and let $\mathcal{I} = \mathcal{A}^N$ be a space of instances and $\mathcal{X} \subseteq \mathcal{B}^n$ be a space of solutions. A *polynomial system* is a set of polynomial equalities

$$g_j(x, \mathcal{I}) = 0 \quad \forall j \in [m],$$

where $\{g_j\}_{j=1}^m$ are polynomials in the *program variables* $\{x_i\}_{i\in[n]}$, representing $x \in \mathcal{X}$, and in the *instance variables* $\{\mathcal{I}_j\}_{j\in[N]}$, representing $\mathcal{I} \in \mathcal{I}$. We define $\deg_{\text{prog}}(g_j)$ to be the degree of g_j in the program variables, and $\deg_{\text{inst}}(g_j)$ to be the degree of g_j in the instance variables.

Remark 12.1.2. For the sake of simplicity, the polynomial system 12.1.1 has no inequalities. Inequalities can be incorporated in to the program by converting each inequality in to an equality with an additional slack variable. Our main theorem still holds, but for some minor modifications of the proof, as outlined in Section 12.3.

Example 12.1.3 (k-clique)**.** Consider a quadratic program which checks if a graph on n vertices contains a clique of size k. We can express this with the polynomial system over program variables $x \in \mathbb{R}^n$ and instance variables $\mathcal{I} \in \{0,1\}^{\binom{n}{2}}$, where $\mathcal{I}_{ij} = 1$ iff there is an edge from i to j, as follows:

$$\left\{\sum\nolimits_{i\in[n]} x_i - k = 0\right\} \cup \{x_i(x_i - 1) = 0\}_{i\in[n]} \cup \{(1 - \mathcal{I}_{ij})x_i x_j = 0\}_{i,j\in\binom{[n]}{2}}.$$

Planted Distributions

We will be concerned with planted distributions of a particular form. For us, a planted distribution is specified by a set $\mathcal{X} \subseteq \mathcal{B}^n$ of *solutions*, where $\mathcal{B}$ is a subset of $\mathbb{R}$, and for each $x \in \mathcal{X}$ a distribution μ_x on $\mathcal{I} = \mathcal{A}^N$. These specify the following joint distribution μ on $\mathcal{X} \times \mathcal{I}$: first draw $x \sim \mathcal{X}$ uniformly, then draw $\mathcal{I}$ from μ_x.

We say that a planted distribution μ is ε-satisfying for a polynomial system $\{g_j(x, \mathcal{I})\}$ if with $(x, \mathcal{I}) \sim \mu$ satisfies $g_j(x, \mathcal{I}) = 0$ with probability at least $1 - \varepsilon$.

We have in mind the following kind of SoS-based hypothesis test: given a polynomial system $\mathcal{S}$, the goal is to test an $\varepsilon(n, d)$-satisfing distribution μ for $\mathcal{S}$ against a product distribution ν. The algorithm is: given an instance $\mathcal{I}$, using semidefinite programming, check whether there is a degree-d SoS refutation of the system $\{g_j(x, \mathcal{I})\}_{g_j \in \mathcal{S}}$, with respect to the variables x. If there is, output NULL. Otherwise, output PLANTED.

Sub-instances

Suppose that $\mathcal{I} = \mathcal{A}^N$ is a family of instances; then given an instance $\mathcal{I} \in \mathcal{I}$ and a subset $S \subseteq [N]$, let $\mathcal{I}_S$ denote the sub-instance consisting of coordinates within S. Further, for a distribution Θ over subsets of $[N]$, let $\mathcal{I}_S \sim_\Theta \mathcal{I}$ denote a subinstance generated by sampling $S \sim \Theta$. Let $\mathcal{I}_\downarrow$ denote the set of all sub-instances of an instance $\mathcal{I}$, and let $\mathcal{I}_\downarrow$ denote the set of all sub-instances of all instances.

Robust Inference

Our result will pertain to polynomial systems that define planted distributions which have the property that feasible solutions to sub-instances generalize to feasible solutions over the entire instance. We call this property "robust inference."

Definition 12.1.4. Let $\mathcal{I} = \mathcal{A}^N$ be a family of instances, let Θ be a distribution over subsets of $[N]$, let $\mathcal{S}$ be a polynomial system as in 12.1.1, with $\deg_{\text{prog}} \mathcal{S} \leqslant 2D$, and let μ be a planted distribution. Then the polynomial system $\mathcal{S}$ is said to satisfy the *robust inference property for probability distribution μ on $\mathcal{I}$ and subsampling distribution* Θ, if given a subsampling $\mathcal{I}_S$ of an instance $\mathcal{I}$ from μ, one can infer a setting of the program variables x^* that remains feasible to $\mathcal{S}$ for most settings of $\mathcal{I}_{\overline{S}}$.

Formally, μ is $\varepsilon(n,d)$ satisfying for $\mathcal{S}$, and there exists a map $x : \mathcal{I}_{\downarrow} \to \mathbb{R}^n$ such that

$$\mathop{\mathbb{P}}_{\mathcal{I}\sim\mu, S\sim\Theta, \tilde{\mathcal{I}}\sim\nu_{|\mathcal{I}_S}} [x(\mathcal{I}_S) \text{ is a feasible for } \mathcal{S} \text{ on } \mathcal{I}_S \circ \tilde{\mathcal{I}}] \geqslant 1 - \varepsilon(n,d)$$

for some negligible function $\varepsilon(n,d)$. To specify the error probability, we will say that polynomial system is $\varepsilon(n,d)$-*robustly inferable*.

Statement of Theorem 12.1.5

We are now ready to state our main theorem.

Theorem 12.1.5 (Formal version of Theorem 1.1.2)**.** *Suppose that $\mathcal{S}$ is a polynomial system as defined in 12.1.1, of degree at most $2d$ in the program variables and degree at most k in the instance variables. Let $B > d \cdot k \in \mathbb{N}$ such that*

1. *The polynomial system $\mathcal{S}$ is $\frac{1}{n^{8B}}$-robustly inferable with respect to the planted distribution μ and the sub-sampling distribution Θ.*
2. *For $\mathcal{I} \sim \nu$, the polynomial system $\mathcal{S}$ admits a degree-d SoS refutation with numbers bounded by n^B with probability at least $1 - \frac{1}{n^{8B}}$.*

Let $D \in \mathbb{N}$ be such that for any subset $\alpha \subseteq [N]$ with $|\alpha| \geqslant D - 2dk$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \subseteq S] \leqslant \frac{1}{n^{8B}}$$

There exists a degree $2D$ matrix polynomial $Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ such that,

$$\frac{\mathbb{E}_{\mathcal{I} \sim \mu}[\lambda^+_{max}(Q(\mathcal{I}))]}{\mathbb{E}_{\mathcal{I} \sim \nu}[\lambda^+_{max}(Q(\mathcal{I}))]} \geqslant n^{B/2}$$

Remark 12.1.6. Our argument implies a stronger result that can be stated in terms of the eigenspaces of the subsampling operator.

Specifically, suppose we define

$$\mathcal{S}_\varepsilon \stackrel{\text{def}}{=} \left\{ \alpha \mid \mathop{\mathbb{P}}_{S \sim \Theta}\{\alpha \subseteq S\} \leqslant \varepsilon \right\}$$

Then, the distinguishing polynomial exhibited by Theorem 12.1.5 satisfies $Q \in$ Span$\{$ monomials $\mathcal{I}_\alpha | \alpha \in \mathcal{S}_\varepsilon\}$. This refinement can yield tighter bounds in cases where all monomials of a certain degree are not equivalent to each other.

For example, in the Planted Clique problem, each monomial corresponds to a subset $\alpha \subseteq \binom{n}{2}$, and the right measure of the degree of a monomial is the number of vertices among $[n]$ incident to α, thought of as a graph. This is by contrast to the usual notion of degree of a monomial, which corresponds just to $|\alpha|$.

Next, in Section 12.2, we describe our main tool to prove Theorem 12.1.5, which is a novel combination of convex duality and pseudocalibration. Then in Section 12.3 we prove Theorem 12.1.5.

12.2 Moment-Matching Pseudodistributions

We assume the setup from Section 12.1: we have a family of instances $\mathscr{I} = \mathcal{A}^N$, a polynomial system $\mathcal{S} = \{g_j(x, \mathcal{I})\}_{j\in[m]}$ with a family of solutions $\mathcal{X} = \mathcal{B}^n$, a null distribution ν which is a product distribution over $\mathscr{I}$, and a planted distribution μ over $\mathscr{I}$.

The contrapositive of Theorem 12.1.5 is that if $\mathcal{S}$ is robustly inferable with respect to μ and a distribution over sub-instances Θ, and if there is no spectral algorithm for distinguishing μ and ν, then with high probability over ν there is no degree-d SoS refutation for the polynomial system $\mathcal{S}$.

We will prove this converse by using convex duality to show that if no spectral algorithm exists, there is a *witness* to this fact, in the form of an *approximate, pseudocalibrated pseudodistribution*. That is, the object we obtain will not be a pseudodistribution, strictly speaking, but it will be close enough to allow us to prove Theorem 12.1.5. In this section, we carry out this convex duality argument.

Since most $\mathcal{I}$ in the support of μ come with $x \in \mathcal{X}$ such that $(x, \mathcal{I})$ is feasible for $\mathcal{S}$ (if μ is $\varepsilon(n, d)$-satisfying), we will start our setup by using these feasible solutions to $\mathcal{S}$. In particular, we use that they are also feasible solutions to the dual of the SoS SDP which searches for refutations of $\mathcal{S}$.

With this in mind, let $\Lambda : \mathscr{I} \to (\mathbb{R}^{[n]^{\leqslant d}\times[n]^{\leqslant d}})_+$ be any function from the support

of μ over $\mathcal{I}$ to PSD matrices with the following form:

$$\Lambda(\mathcal{I}) = \widehat{\mu}(\mathcal{I}) \cdot M(\mathcal{I})$$

where $\widehat{\mu}$ is the relative density of μ with respect to ν, so that $\widehat{\mu}(\mathcal{I}) = \mu(\mathcal{I})/\nu(\mathcal{I})$, and M is some matrix valued function such that $M(\mathcal{I}) \succeq 0$ and $\|M(\mathcal{I})\| \leqslant n^B$ for all $\mathcal{I} \in \mathcal{I}$. Our goal is to find a PSD matrix-valued function P that matches the low-degree moments of Λ in the variables $\mathcal{I}$, while being supported over most of $\mathcal{I}$ (rather than just over the support of μ). This insistence on moment matching is the start of our use of the pseudocalibration idea.

Let the function $P : \mathcal{I} \to (\mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}})_+$ be given by minimizer of the following exponentially large convex program over matrix-valued functions,

Program 12.2.1 (Pseudodistribution Program).

$$\min \quad \|P\|_{Fr,\nu}^2 \tag{12.2.1}$$

$$s.t. \quad \langle Q, P\rangle_\nu = \langle Q, \Lambda'\rangle_\nu \quad \forall Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}},\ \deg_{\text{inst}}(Q) \leqslant D \tag{12.2.2}$$

$$P \succeq 0$$

$$\Lambda' = \Lambda + \eta \cdot \mathrm{Id}, \quad 2^{-2^{2^n}} > \eta > 0 \tag{12.2.3}$$

In our typical applications, the matrix-valued function M will have $\operatorname{Tr} M(\mathcal{I}) = n^{O(d)}$ for all $\mathcal{I}$. Since $\operatorname{Tr} M(\mathcal{I}) = \langle \mathrm{Id}, M(\mathcal{I})\rangle$, the constraint (12.2.2) fixes $\mathbb{E} \operatorname{Tr}(P)$, and so the objective function (12.2.1) can be viewied as minimizing $\mathbb{E} \operatorname{Tr}(P^2)$, a proxy for the collision probability of the distribution, which is a measure of entropy.

Eventually, we will see that any solution P to Program 12.2.1 with $\|P\|_{Fr,\nu}^2 \ll n^B$ will allow us to rule out the existence of SoS refutations of $\mathcal{S}$ for typical $\mathcal{I} \sim \nu$.

Remark 12.2.2. We have perturbed Λ in (12.2.3) so that we can easily show that strong duality holds in the proof of 12.2.4. For the remainder of the paper we

ignore this perturbation, as we can accumulate the resulting error terms and set η to be small enough so that they can be neglected.

The dual of the above program will allow us to relate the existence of an SoS refutation to the existence of a spectral algorithm.

Program 12.2.3 (Low-Degree Distinguisher).

$$\begin{aligned} \max \quad & \langle \Lambda, Q \rangle_\nu \\ s.t. \quad & Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}, \ \deg_{\text{inst}}(Q) \leqslant D \\ & \|Q_+\|^2_{Fr,\nu} \leqslant 1, \end{aligned}$$

where Q_+ is the projection of Q to the PSD cone.

Claim 12.2.4. 12.2.3 is a manipulation of the dual of 12.2.1, so that if 12.2.1 has optimum $c > 1$, 12.2.3 has optimum at least $\Omega(\sqrt{c})$.

Before we present the proof of the claim, we summarize its central consequence in the following theorem: if 12.2.1 has a large objective value (and therefore does not provide a feasible SoS solution), then there is a spectral algorithm.

Theorem 12.2.5. *Fix a function $M : \mathcal{I} \to \mathbb{R}_+^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ be such that $\mathrm{Id} \succeq M \succeq 0$. Let $\lambda^+_{\max}(\cdot)$ be the function that gives the largest non-negative eigenvalue of a matrix. Suppose $\Lambda = \mu \cdot M$ then the optimum of 12.2.1 is equal to $\mathrm{opt} > 1$ only if there exists a low-degree matrix polynomial Q such that,*

$$\mathop{\mathbb{E}}_{\mathcal{I} \sim \mu} [\lambda^+_{max}(Q(\mathcal{I}))] \geqslant \Omega(\sqrt{\mathrm{opt}}/n^d)$$

while,

$$\mathop{\mathbb{E}}_{\mathcal{I} \sim \nu} [\lambda^+_{max}(Q(\mathcal{I}))] \leqslant 1 \,.$$

Proof. By Claim 12.2.4, if the value of 12.2.1 is opt > 1, then there is a polynomial Q achieves a value of $\Omega(\sqrt{\text{opt}})$ for the dual. It follows that

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu}[\lambda^{+}_{max}(Q(\mathcal{I}))] \geqslant \frac{1}{n^d}\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu}[\langle \mathrm{Id}, Q_{+}(\mathcal{I}))\rangle] \geqslant \frac{1}{n^d}\langle \Lambda, Q_{+}\rangle_{\nu} = \Omega(\sqrt{\text{opt}}/n^d),$$

where we have used that $0 \preceq M \preceq \mathrm{Id}$. (Here Q_+ denotes the PSD projection of Q.) On the other hand,

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\lambda^{+}_{max}(Q(\mathcal{I}))] \leqslant \sqrt{\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\lambda^{+}_{\max}(Q(\mathcal{I}))^2]} \leqslant \sqrt{\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\|Q_{+}(\mathcal{I})\|^2_{Fr}} \leqslant 1.$$

□

It is interesting to note that the specific structure of the PSD matrix valued function M plays no role in the above argument—since M serves as a proxy for monomials in the solution as represented by the program variables $x^{\otimes d}$, it follows that the choice of how to represent the planted solution is not critical. Although seemingly counterintuitive, this is natural because the property of being distinguishable by low-degree distinguishers or by SoS proofs is a property of ν and μ.

We wrap up the section by presenting a proof of Claim 12.2.4.

Proof of 12.2.4. We take the Lagrangian dual of 12.2.1. Our dual variables will be some combination of low-degree matrix polynomials, Q, and a PSD matrix A:

$$\mathcal{L}(P, Q, A) = \|P\|^2_{Fr,\nu} - \langle Q, P - \Lambda'\rangle_{\nu} - \langle A, P\rangle_{\nu} \qquad s.t. \quad A \succeq 0.$$

It is easy to verify that if P is not PSD, then A can be chosen so that the value of $\mathcal{L}$ is ∞. Similarly if there exists a low-degree polynomial upon which P and Λ differ in expectation, Q can be chosen as a multiple of that polynomial so that the value of $\mathcal{L}$ is ∞.

Now, we argue that Slater's conditions are met for 12.2.1, as $P = \Lambda'$ is strictly feasible. Thus strong duality holds, and therefore

$$\min_{P} \max_{A \succeq 0, Q} \mathcal{L}(P, Q, A) \leqslant \max_{A \succeq 0, Q} \min_{P} \mathcal{L}(P, Q, A).$$

Taking the partial derivative of $\mathcal{L}(P, Q, A)$ with respect to P, we have

$$\frac{\partial}{\partial P} \mathcal{L}(P, Q, A) = 2 \cdot P - Q - A.$$

where the first derivative is in the space of functions from $\mathscr{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$. By the convexity of $\mathcal{L}$ as a function of P, it follows that if we set $\frac{\partial}{\partial P} \mathcal{L} = 0$, we will have the minimizer. Substituting, it follows that

$$\begin{aligned} \min_{P} \max_{A \succeq 0, Q} \mathcal{L}(P, Q, A) &\leqslant \max_{A \succeq 0, Q} \frac{1}{4} \|A + Q\|_{Fr,\nu}^2 - \frac{1}{2} \langle Q, A + Q - \Lambda' \rangle_\nu - \frac{1}{2} \langle A, A + Q \rangle_\nu \\ &= \max_{A \succeq 0, Q} \langle Q, \Lambda' \rangle_\nu - \frac{1}{4} \|A + Q\|_{Fr,\nu}^2 \end{aligned} \tag{12.2.4}$$

Now it is clear that the maximizing choice of A is to set $A = -Q_-$, the negation of the negative-semi-definite projection of Q. Thus (12.2.4) simplifies to

$$\begin{aligned} \min_{P} \max_{A \succeq 0, Q} \mathcal{L}(P, Q, A) &\leqslant \max_{Q} \langle Q, \Lambda' \rangle_\nu - \frac{1}{4} \|Q_+\|_{Fr,\nu}^2 \\ &\leqslant \max_{Q} \langle Q, \Lambda \rangle_\nu + \eta \operatorname{Tr}_\nu(Q_+) - \frac{1}{4} \|Q_+\|_{Fr,\nu}^2, \end{aligned} \tag{12.2.5}$$

where we have used the shorthand $\operatorname{Tr}_\nu(Q_+) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{I} \sim \nu} \operatorname{Tr}(Q(\mathcal{I})_+)$. Now suppose that the low-degree matrix polynomial Q^* achieves a right-hand-side value of

$$\langle Q^*, \Lambda \rangle_\nu + \eta \cdot \operatorname{Tr}_\nu(Q_+^*) - \frac{1}{4} \|Q_+^*\|_{Fr,\nu}^2 \geqslant c.$$

Consider $Q' = Q^* / \|Q_+^*\|_{Fr,\nu}$. Clearly $\|Q'_+\|_{Fr,\nu} = 1$. Now, multiplying the above inequality through by the scalar $1/\|Q_+^*\|_{Fr,\nu}$, we have that

$$\begin{aligned} \langle Q', \Lambda \rangle_\nu &\geqslant \frac{c}{\|Q_+^*\|_{Fr,\nu}} - \eta \cdot \frac{\operatorname{Tr}_\nu(Q_+^*)}{\|Q_+^*\|_{Fr,\nu}} + \frac{1}{4} \|Q_+^*\|_{Fr,\nu} \\ &\geqslant \frac{c}{\|Q_+^*\|_{Fr,\nu}} - \eta \cdot n^d + \frac{1}{4} \|Q_+^*\|_{Fr,\nu}. \end{aligned}$$

Therefore $\langle Q', \Lambda \rangle_\nu$ is at least $\Omega(c^{1/2})$, as if $\|Q'_+\|_{Fr,\nu} \geqslant \sqrt{c}$ then the third term gives the lower bound, and otherwise the first term gives the lower bound.

Thus by substituting Q', the square root of the maximum of (12.2.5) within an additive ηn^d lower-bounds the maximum of the program

$$\begin{aligned} \max \quad & \langle Q, \Lambda \rangle_\nu \\ s.t. \quad & Q : \mathscr{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}, \quad \deg_{\text{inst}}(Q) \leqslant D \\ & \|Q_+\|^2_{Fr,\nu} \leqslant 1. \end{aligned}$$

This concludes the proof. □

12.3 Proof of Theorem 12.1.5

We will prove Theorem 12.1.5 by contradiction. Let us assume that there exists no degree-$2D$ matrix polynomial that distinguishes ν from μ. First, the lack of distinguishers implies the following fact about scalar polynomials.

Lemma 12.3.1. *Under the assumption that there are no degree-$2D$ distinguishers, for every degree-D scalar polynomial Q,*

$$\|Q\|^2_{Fr,\mu} \leqslant n^B \|Q\|^2_{Fr,\nu}$$

Proof. Suppose not, then the degree-$2D$ 1×1 matrix polynomial $\mathrm{Tr}(Q(\mathcal{I})^2)$ will be a distinguisher between μ and ν. □

Constructing Λ First, we will use the robust inference property of μ to construct a pseudo-distribution Λ. Recall again that we have defined $\widehat{\mu}$ to be the relative

density of μ with respect to ν, so that $\widehat{\mu}(\mathcal{I}) = \mu(\mathcal{I})/\nu(\mathcal{I})$. For each subset $S \subseteq [N]$, define a PSD matrix-valued function $\Lambda_S : \mathscr{I} \to (\mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}})_+$ as,

$$\Lambda_S(\mathcal{I}) = \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}}[\widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] \cdot x(\mathcal{I}_S)^{\leqslant d}(x(\mathcal{I}_S)^{\leqslant d})^T$$

where we use $\mathcal{I}_S$ to denote the restriction of $\mathcal{I}$ to $S \subset [N]$, and $\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}$ to denote the instance given by completing the sub-instance $\mathcal{I}_S$ with the setting $\mathcal{I}'_{\overline{S}}$. Notice that Λ_S is a function depending only on $\mathcal{I}_S$—this fact will be important to us. Define $\Lambda \stackrel{\text{def}}{=} \mathbb{E}_{S\sim\Theta} \Lambda_S$. Observe that Λ is a PSD matrix-valued function that satisfies

$$\langle \Lambda_{\emptyset,\emptyset}, 1\rangle_\nu = \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} [\widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] = \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}}} \mathop{\mathbb{E}}_{\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}\sim\nu} [\widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] = 1 \tag{12.3.1}$$

Since $\Lambda(\mathcal{I})$ is an average over $\Lambda_S(\mathcal{I})$, each of which is a feasible solution with high probability, $\Lambda(\mathcal{I})$ is close to a feasible solution to the SDP relaxation for $\mathcal{I}$. The following Lemma formalizes this intuition.

Define $\mathcal{G} \stackrel{\text{def}}{=} \text{Span}\{\chi_S \cdot G_j \mid j \in [m], S \subseteq [N]\}$, and use $\Pi_{\mathcal{G}}$ to denote the orthogonal projection into $\mathcal{G}$.

Lemma 12.3.2. *Suppose 12.1.1 satisfies the ε-robust inference property with respect to planted distribution μ and subsampling distribution Θ and $\|x(\mathcal{I}_S)^{\leqslant d}\|_2^2 \leqslant K$ for all $\mathcal{I}_S$. Then for every $G \in \mathcal{G}$, we have*

$$\langle \Lambda, G\rangle_\nu \leqslant \sqrt{\varepsilon} \cdot K \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}\sim\mu} \|G(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}})\|_2^2 \right)^{\frac{1}{2}}$$

Proof. We begin by expanding the left-hand side by substituting the definition of Λ. We have

$$\begin{aligned}
\langle \Lambda, G\rangle_\nu &= \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \langle \Lambda_S(\mathcal{I}_S), G(\mathcal{I})\rangle \\
&= \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \langle x(\mathcal{I}_S)^{\leqslant d}(x(\mathcal{I}_S)^{\leqslant d})^T, G(\mathcal{I})\rangle
\end{aligned}$$

And because the inner product is zero if $x(\mathcal{I}_S)$ is a feasible solution,

$$\leqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \mathbf{1}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I})] \cdot \left\|x(\mathcal{I}_S)^{\leqslant d}\right\|_2^2 \cdot \|G(\mathcal{I})\|_{Fr}$$

$$\leqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \mathbf{1}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I})] \cdot K \cdot \|G(\mathcal{I})\|_{Fr}$$

And now letting $\tilde{\mathcal{I}}_{\overline{S}}$ denote the completion of $\mathcal{I}_S$ to $\mathcal{I}$, so that $\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}} = \mathcal{I}$, we note that the above is like sampling $\mathcal{I}'_{\overline{S}}, \tilde{\mathcal{I}}_{\overline{S}}$ independently from ν and then reweighting by $\widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})$, or equivalently taking the expectation over $\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}} = \mathcal{I}' \sim \mu$ and $\tilde{\mathcal{I}}_{\overline{S}} \sim \nu$:

$$= \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu}\mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \cdot \mathbf{1}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}})] \cdot K \cdot \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}$$

and by Cauchy-Schwarz,

$$\leqslant K \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu}\mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \cdot \mathbf{1}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}})]\right)^{\frac{1}{2}}$$
$$\cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu}\mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2\right)^{\frac{1}{2}}$$

The lemma follows by observing that the first term in the product above is exactly the non-robustness of inference probability ε. □

Corollary 12.3.3. *If $G \in \mathcal{G}$ is a degree-D polynomial in $\mathcal{I}$, then under the assumption that there are no degree-$2D$ distinguishers for ν, μ,*

$$\langle \Lambda, G \rangle_\nu \leqslant \sqrt{\varepsilon} \cdot K \cdot n^B \cdot \|G\|_{Fr,\nu}$$

Proof. For each fixing of $\tilde{\mathcal{I}}_S$, $\|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_2^2$ is a degree-$2D$-scalar polynomial in $\mathcal{I}$. Therefore by Lemma 12.3.1 we have that,

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2 \leqslant n^B \cdot \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2 .$$

Substituting back in the bound in Lemma 12.3.2 the corollary follows. □

Now, since there are no degree-D matrix distinguishers Q, for each S in the support of Θ we can apply reasoning similar to Theorem 12.2.5 to conclude that there is a high-entropy PSD matrix-valued function P_S that matches the degree-D moments of Λ_S.

Lemma 12.3.4. *If there are no degree-D matrix distinguishers Q for μ, ν, then for each $S \sim \Theta$, there exists a solution P_S to 12.2.1 (with the variable $\Lambda := \Lambda_S$) and*

$$\|P_S\|_{Fr,\nu} \leqslant n^{\frac{(B+d)}{4}} \leqslant n^{B/2} \tag{12.3.2}$$

This does not follow directly from Theorem 12.2.5, because a priori a distinguisher for some specific S may only apply to a small fraction of the support of μ. However, we can show that 12.2.1 has large value for Λ_S only if there is a distinguisher for μ, ν.

Proof. By 12.2.4, it suffices for us to argue that there is no degree-D matrix polynomial Q which has large inner product with Λ_S relative to its Frobenius norm. So, suppose by way of contradiction that Q is a degree-D matrix that distinguishes Λ_S, so that $\langle Q, \Lambda_S \rangle_\nu \geqslant n^{B+d}$ but $\|Q\|_{Fr,\nu} \leqslant 1$.

It follows by definition of Λ_S that

$$\begin{aligned} n^{B+d} \leqslant \langle Q, \Lambda_S \rangle_\nu &= \mathop{\mathbb{E}}_{\mathcal{I} \sim \nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}} \sim \nu} \widehat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \langle Q(\mathcal{I}), x(\mathcal{I}_S)^{\leqslant d} (x(\mathcal{I}_S)^{\leqslant d})^\top \rangle \\ &= \mathop{\mathbb{E}}_{\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}} \sim \mu} \left\langle \mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}} \sim \nu} Q(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}}), x(\mathcal{I}_S)^{\leqslant d} (x(\mathcal{I}_S)^{\leqslant d})^\top \right\rangle \\ &\leqslant \mathop{\mathbb{E}}_{\mu} \left[\lambda^+_{\max} \left(\mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}} \sim \nu} Q(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}}) \right) \right] \cdot \left\| x(\mathcal{I}_S)^{\leqslant d} \right\|_2^2 . \end{aligned}$$

So, we will show that $Q_S(\mathcal{I}) = \mathbb{E}_{\mathcal{I}'_{\overline{S}} \sim \nu} Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})$ is a degree-D distinguisher for μ. The degree of Q_S is at most D, since averaging over settings of the variables cannot increase the degree. Applying our assumption that $\|x(\mathcal{I}_S)^{\leqslant d}\|_2^2 \leqslant K \leqslant n^d$,

we already have $\mathbb{E}_\mu \lambda^+_{\max}(Q_S) > n^B$. It remains to show that $\mathbb{E}_\nu \lambda^+_{\max}(Q_S)$ is bounded. For this, we use the following fact about the trace.

Fact 12.3.5 (See e.g. Theorem 2.10 in [47]). *For a function $f : \mathbb{R} \to \mathbb{R}$ and a symmetric matrix A with eigendecomposition $\sum \lambda \cdot vv^\top$, define $f(A) = \sum f(\lambda) \cdot vv^\top$. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and convex, then the map $A \to \mathrm{Tr}(f(A))$ is convex for symmetric A.*

The function $f(t) = (\max\{0, t\})^2$ is continuous and convex over $\mathbb{R}$, so the fact above implies that the map $A \to \|A_+\|^2_{Fr}$ is convex for symmetric A. We can take Q_S to be symmetric without loss of generality, as in the argument above we only consider the inner product of Q_S with symmetric matrices. Now we have that

$$\|(Q_S(I))_+\|^2_{Fr} = \left\| \left(\mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}} \left[Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \right] \right)_+ \right\|^2_{Fr} \leqslant \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}} \left\| \left(Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \right)_+ \right\|^2_{Fr},$$

where the inequality is the definition of convexity. Taking the expectation over $\mathcal{I} \sim \nu$ gives us that $\|(Q_S)_+\|^2_{Fr,\nu} \leqslant \|Q_+\|^2_{Fr,\nu} \leqslant 1$, which gives us our contradiciton. □

Now, analogous to Λ, set $P \stackrel{\text{def}}{=} \mathbb{E}_{S \sim \Theta} P_S$.

Random Restriction We will exploit the crucial property that Λ and P are averages over functions that depend on subsets of variables. This has the same effect as a random restriction, in that $\langle P, R \rangle_\nu$ essentially depends on the low-degree part of R. Formally, we will show the following lemma.

Lemma 12.3.6. *(Random Restriction) Fix $D, \ell \in \mathbb{N}$. For matrix-valued functions $R : \mathcal{I} \to \mathbb{R}^{\ell \times \ell}$ and a family of functions $\{P_S : \mathcal{I}_S \to \mathbb{R}^{\ell \times \ell}\}_{S \subseteq [N]}$, and a distribution Θ*

over subsets of $[N]$,

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S^{<D}(\mathcal{I}_S)\rangle - \rho(D,\Theta)^{\frac{1}{2}}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{\frac{1}{2}}\|R\|_{Fr,\nu}$$

where

$$\rho(D,\Theta) = \max_{\alpha,|\alpha|\geqslant D}\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha\subseteq S].$$

Proof. We first re-express the left-hand side as

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle = \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S(\mathcal{I}_S)\rangle$$

where $R_S(\mathcal{I}_S) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{I}_{\overline{S}}}[R(\mathcal{I})]$ obtained by averaging out all coordinates outside S. Splitting the function R_S into its low-degree and high-degree parts, $R_S = R_S^{\leqslant D} + R_S^{>D}$, then applying a Cauchy-Schwartz inequality we get

$$\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S(\mathcal{I}_S)\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S^{<D}(\mathcal{I}_S)\rangle - \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{\frac{1}{2}}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|R_S^{\geqslant D}\|_{Fr,\nu}^2\right)^{\frac{1}{2}}.$$

Expressing $R^{\geqslant D}(\mathcal{I})$ in the Fourier basis, we have that over a random choice of $S\sim\Theta$,

$$\mathop{\mathbb{E}}_{S\sim\Theta}\|R_S^{\geqslant D}\|_{Fr,\nu}^2 = \sum_{\alpha,|\alpha|\geqslant D}\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha\subseteq S]\cdot\widehat{R}_\alpha^2 \leqslant \rho(D,\Theta)\cdot\|R\|_{Fr}^2$$

Substituting into the above inequality, the conclusion follows. □

Equality Constraints Since Λ is close to satisfying all the equality constraints $\mathcal{G}$ of the SDP, the function P approximately satisfies the low-degree part of $\mathcal{G}$. Specifically, we can prove the following.

Lemma 12.3.7. *Let $k \geqslant \deg_{\text{inst}}(G_j)$ for all $G_j \in \mathcal{S}$. With P defined as above and under the conditions of Theorem 12.1.5 for any function $G \in \mathcal{G}$,*

$$\left|\langle P, G^{\leqslant D}\rangle_\nu\right| \leqslant \frac{2}{n^{2B}}\|G\|_{Fr,\nu}$$

Proof. Recall that $\mathcal{G} = \text{Span}\{\chi_S \cdot G_j \mid j \in [m], S \subseteq [N]\}$ and let $\Pi_{\mathcal{G}}$ be the orthogonal projection into $\mathcal{G}$. Now, since $G \in \mathcal{G}$,

$$G^{\leqslant D} = (\Pi_{\mathcal{G}} G)^{\leqslant D} = (\Pi_{\mathcal{G}} G^{\leqslant D-2k})^{\leqslant D} + (\Pi_{\mathcal{G}} G^{>D-2k})^{\leqslant D}. \tag{12.3.3}$$

Now we make the following claim regarding the effect of projection on to the ideal $\mathcal{G}$, on the degree of a polynomial.

Claim 12.3.8. For every polynomial Q, $\deg(\Pi_{\mathcal{G}} Q) \leqslant \deg(Q) + 2k$. Furthermore for all α, $\Pi_{\mathcal{G}} Q^{>\alpha}$ has no monomials of degree $\leqslant \alpha - k$

Proof. To establish the first part of the claim it suffices to show that $\Pi_{\mathcal{G}} Q \in \text{Span}\{\chi_S \cdot G_j \mid |S| \leqslant \deg(Q) + k\}$, since $\deg(G_j) \leqslant k$ for all $j \in [m]$. To see this, observe that $\Pi_{\mathcal{G}} Q \in \text{Span}\{\chi_S \cdot G_j \mid |S| \leqslant \deg(Q) + k\}$ and is orthogonal to every $\chi_S \cdot G_j$ with $|S| > \deg(Q) + k$:

$$\langle \Pi_{\mathcal{G}} Q, \chi_S \cdot G_j\rangle_\nu = \langle Q, \Pi_{\mathcal{G}} \chi_S \cdot G_j\rangle_\nu = \langle Q, \chi_S \cdot G_j\rangle_\nu = \langle Q G_j, \chi_S\rangle_\nu = 0,$$

where the final equality is because $\deg(\chi_S) > \deg(G_j) + \deg(Q)$. On the other hand, for every subset S with $\deg(\chi_S) \leqslant \alpha - k$,

$$\langle \Pi_{\mathcal{G}} Q^{>\alpha}, \chi_S \cdot G_j\rangle = \langle Q^{>\alpha}, \Pi_{\mathcal{G}} \chi_S \cdot G_j\rangle = \langle Q^{>\alpha}, \chi_S \cdot G_j\rangle = 0,$$

since $\alpha > \deg(G_j) + \deg(\chi_S)$ This implies that $\Pi_{\mathcal{G}} Q^{>\alpha} \in \text{Span}\{\chi_S \cdot G_j \mid |S| > \alpha - k\}$ which implies that $\Pi_{\mathcal{G}} Q^{>\alpha}$ has no monomials of degree $\leqslant \alpha - k$. □

Incorporating the above claim into (12.3.3), we have that

$$G^{\leqslant D} = \Pi_{\mathcal{G}} G^{\leqslant D-2k} + (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]},$$

where the superscript $[D-3k, D]$ denotes the degree range. Now,

$$\langle P, G^{\leqslant D} \rangle_\nu = \langle P, \Pi_{\mathcal{G}} G^{\leqslant D-2k} \rangle_\nu + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]} \rangle_\nu$$

And since $\Pi_{\mathcal{G}} G^{\leqslant D-2k}$ is of degree at most D we can replace P by Λ,

$$= \langle \Lambda, \Pi_{\mathcal{G}} G^{\leqslant D-2k} \rangle_\nu + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]} \rangle_\nu$$

Now bounding the first term using 12.3.3 with a n^B bound on K,

$$\leqslant \left(\frac{1}{n^{8B}}\right)^{\frac{1}{2}} \cdot n^B \cdot (n^B \cdot \|\Pi_{\mathcal{G}} G^{\leqslant D-2k}_{\emptyset,\emptyset}\|_{Fr,\nu}) + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]} \rangle$$

And for the latter term we use Lemma 12.3.6,

$$\leqslant \frac{1}{n^{2B}} \|\Pi_{\mathcal{G}} G^{\leqslant D-2k}_{\emptyset,\emptyset}\|_{Fr,\nu} + \frac{1}{n^{4B}} \left(\mathop{\mathbb{E}}_S \|P_S\|^2_{Fr,\nu}\right)^{\frac{1}{2}} \|G\|_{Fr,\nu},$$

where we have used the fact that $(\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]}$ is high degree. By property of orthogonal projections, $\|\Pi_{\mathcal{G}} G^{\geqslant D-2k}\|_{Fr,\nu} \leqslant \|G^{\geqslant D-2k}\|_{Fr,\nu} \leqslant \|G\|_{Fr,\nu}$. Along with the bound on $\|P_S\|_{Fr,\nu}$ from (12.3.2), this implies the claim of the lemma. □

Finally, we have all the ingredients to complete the proof of Theorem 12.1.5.

Proof of Theorem 12.1.5. Suppose we sample an instance $\mathcal{I} \sim \nu$, and suppose by way of contradiction this implies that with high probability the SoS SDP relaxation is infeasible. In particular, this implies that there is a degree-d sum-of-squares refutation of the form,

$$-1 = a^{\mathcal{I}}(x) + \sum_{j \in [m]} g^{\mathcal{I}}_j(x) \cdot q^{\mathcal{I}}_j(x),$$

where $a^{\mathcal{I}}$ is a sum-of-squares of polynomials of degree at most $2d$ in x, and $\deg(q_j^{\mathcal{I}}) + \deg(g_j^{\mathcal{I}}) \leqslant 2d$. Let $A^{\mathcal{I}} \in \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ be the matrix of coefficients for $a^{\mathcal{I}}(c)$ on input $\mathcal{I}$, and let $G^{\mathcal{I}}$ be defined similarly for $\sum_{j \in [m]} g_j(x) \cdot q_j(x)$. We can rewrite the sum-of-squares refutation as a matrix equality,

$$-1 = \langle X^{\leqslant d}, A^{\mathcal{I}} \rangle + \langle X^{\leqslant d}, G^{\mathcal{I}} \rangle,$$

where $G^{\mathcal{I}} \in \mathcal{G}$, the span of the equality constraints of the SDP.

Define $s : \mathcal{I} \to \{0, 1\}$ as

$$s(\mathcal{I}) \stackrel{\text{def}}{=} \mathbf{1}[\exists \text{ a degree-}2d \text{ sos-refutation for } \mathcal{S}(\mathcal{I})]$$

By assumption, $\mathbb{E}_{\mathcal{I} \sim \nu}[s(\mathcal{I})] = 1 - \frac{1}{n^{8B}}$. Define matrix valued functions $A, G : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ by setting,

$$A(\mathcal{I}) \stackrel{\text{def}}{=} s(\mathcal{I}) \cdot A^{\mathcal{I}}$$

$$G(\mathcal{I}) \stackrel{\text{def}}{=} s(\mathcal{I}) \cdot G^{\mathcal{I}}$$

With this notation, we can rewrite the sos-refutation identity as a polynomial identity in X and $\mathcal{I}$,

$$-s(\mathcal{I}) = \langle X^{\leqslant d}, A(\mathcal{I}) \rangle + \langle X^{\leqslant d}, G(\mathcal{I}) \rangle .$$

Let $\mathbf{e}_{\emptyset,\emptyset}$ denote the $[n]^{\leqslant d} \times [n]^{\leqslant d}$ matrix with the entry corresponding to $(\emptyset, \emptyset)$ equal to 1, while the remaining entries are zero. We can rewrite the above equality as,

$$-\langle X^{\leqslant d}, s(\mathcal{I}) \cdot \mathbf{e}_{\emptyset,\emptyset} \rangle = \langle X^{\leqslant d}, A(\mathcal{I}) \rangle + \langle X^{\leqslant d}, G(\mathcal{I}) \rangle .$$

for all $\mathcal{I}$ and formal variables X.

Now, let $P = \mathbb{E}_{S \sim \Theta} P_S$ where each P_S is obtained by from the 12.2.1 with Λ_S. Substituting $X^{\leqslant d}$ with $P(\mathcal{I})$ and taking an expectation over $\mathcal{I}$,

$$\langle P, s(\mathcal{I}) \cdot \mathbf{e}_{\emptyset,\emptyset} \rangle_\nu = \langle P, A \rangle_\nu + \langle P, G \rangle_\nu \tag{12.3.4}$$

$$\geqslant \langle P, G\rangle_\nu \tag{12.3.5}$$

where the inequality follows because $A, P \succeq 0$. We will show that the above equation is a contradiction by proving that LHS is less than -0.9, while the right hand side is at least -0.5. First, the right hand side of (12.3.4) can be bounded by Lemma 12.3.7

$$\begin{aligned}
\langle P, G\rangle_\nu &= \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), G(\mathcal{I})\rangle \\
&\geqslant \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), G^{\leqslant D}(\mathcal{I})\rangle - \frac{1}{n^{4B}}\cdot\left(\mathop{\mathbb{E}}_{S}\|P_S\|_{Fr,\nu}^2\right)^{1/2}\cdot\|G\|_{Fr,\nu} \text{ (Lemma 12.3.6)} \\
&\geqslant -\frac{2}{n^{2B}}\cdot\|G\|_{Fr,\nu} - \frac{1}{n^{4B}}\left(\mathop{\mathbb{E}}_{S}\|P_S\|_{Fr,\nu}^2\right)^{\frac{1}{2}}\|G\|_{Fr,\nu} \qquad \text{(using Lemma 12.3.7)} \\
&\geqslant -\frac{1}{2}
\end{aligned}$$

where the last step used the bounds on $\|P_S\|_{Fr,\nu}$ from (12.3.2) and on $\|G\|_{Fr,\nu}$ from the n^B bound assumed on the SoS proofs in Theorem 12.1.5.

Now the negation of the left hand side of (12.3.4) is

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P(\mathcal{I}), s(\mathcal{I})\cdot\mathbf{e}_{\emptyset,\emptyset}\rangle \geqslant \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[P_{\emptyset,\emptyset}(\mathcal{I})\cdot 1] - \mathbb{E}[(s-1)^2]^{1/2}\cdot\|P\|_{Fr,\nu}$$

The latter term can be simplified by noticing that the expectation of the square of a 0,1 indicator is equal to the expectation of the indicator, which is in this case $\frac{1}{n^{8B}}$ by assumption. Also, since 1 is a constant, $P_{\emptyset,\emptyset}$ and $\Lambda_{\emptyset,\emptyset}$ are equivalent:

$$\begin{aligned}
&= \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\Lambda_{\emptyset,\emptyset}(\mathcal{I})\cdot 1] - \frac{1}{n^{4B}}\cdot\|P\|_{Fr,\nu} \\
&= 1 - \frac{1}{n^{4B}}\cdot\|P\|_{Fr,\nu} \qquad \text{(using (12.3.1))} \\
&= 1 - \frac{1}{n^{3B}} \qquad \text{(using (12.3.2))}
\end{aligned}$$

We have the desired contradiction in (12.3.4). □

12.3.1 Handling Inequalities

Suppose the polynomial system 12.1.1 includes inequalities of the form $h(\mathcal{I}, x) \geqslant 0$, then a natural approach would be to introduce a slack variable z and set $h(\mathcal{I}, x) - z^2 = 0$. Now, we can view the vector (x, z) consisting of the original variables along with the slack variables as the hidden planted solution. The proof of Theorem 12.1.5 can be carried out as described earlier in this section, with this setup. However, in many cases of interest, the inclusion of slack variables invalidates the robust inference property. This is because, although a feasible solution x can be recovered from a subinstance $\mathcal{I}_S$, the value of the corresponding slack variables could potentially depend on $\mathcal{I}_{\overline{S}}$. For instance, in a random CSP, the value of the objective function on the assignment x generated from $\mathcal{I}_S$ depends on all the constraints outside of S too.

The proof we described is to be modified as follows.

- As earlier, construct Λ_S using only the robust inference property of original variables x, and the corresponding matrix functions P_S.
- Convert each inequality of the form $h_i(\mathcal{I}, x) \geqslant 0$, in to an equality by setting $h_i(\mathcal{I}, x) = z_i^2$.
- Now we define a pseudo-distribution $\tilde{\Lambda}_S(\mathcal{I}_S)$ over original variables x and slack variables z as follows. It is convenient to describe the pseudo-distribution in terms of the corresponding pseudo-expectation operator. Specifically, if $x(\mathcal{I}_S)$ is a feasible solution for 12.1.1 then define

$$\tilde{E}[z_\sigma x_\alpha] \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \sigma_i \text{ odd for some } i \\ \prod_{i \in \sigma} (h_i(\mathcal{I}, x(\mathcal{I}_S)))^{\sigma_i/2} \cdot x(\mathcal{I}_S)_\alpha & \text{otherwise} \end{cases}$$

Intuitively, the pseudo-distribution picks the sign for each z_i uniformly at random, independent of all other variables. Therefore, all moments involving an odd power of z_i are zero. On the other hand, the moments of even powers of z_i are picked so that the equalities $h_i(\mathcal{I}, x) = z_i$ are satisfied.

It is easy to check that $\tilde{\Lambda}$ is psd matrix valued, satisfies (12.3.1) and all the equalities.

- While Λ_S in the original proof was a function of $\mathcal{I}_S$, $\tilde{\Lambda}_S$ is not. However, the key observation is that, $\tilde{\Lambda}_S$ is degree at most $k \cdot d$ in the variables outside of S. Each function $h_i(\mathcal{I}, x(\mathcal{I}_S))$ is degree at most k in $\mathcal{I}_{\overline{S}}$, and the entries of $\tilde{\Lambda}_S(\mathcal{I}_S)$ are a product of at most d of these polynomials.
- The main ingredient of the proof that is different from the case of equalities is the random restriction lemma which we outline below. The error in the random restriction is multiplied by $D^{dk/2} \leqslant n^{B/2}$; however this does not substantially change our results, since Theorem 12.1.5 requires $\rho(D, \Theta) < n^{-8B}$, which leaves us enough slack to absorb this factor (and in every application $\rho(D, \Theta) = p^{O(D)}$for some $p < 1$ sufficiently small that we meet the requirement that $D^{dk}\rho(D - dk, \Theta)$ is monotone non-increasing in D).

Lemma 12.3.9 (Random Restriction for Inequalities)**.** *Fix $D, \ell \in \mathbb{N}$. Consider a matrix-valued function $R : \mathcal{J} \to \mathbb{R}^{\ell \times \ell}$ and a family of functions $\{P_S : \mathcal{J} \to \mathbb{R}^{\ell \times \ell}\}_{S \subseteq [N]}$ such that each P_S has degree at most dk in $\mathcal{I}_{\overline{S}}$. If Θ is a distribution over subsets of $[N]$ with*

$$\rho(D, \Theta) = \max_{\alpha, |\alpha| \geqslant D} \mathbb{P}_{S \sim \Theta}[\alpha \subseteq S],$$

and the additional requirement that $D^{dk} \cdot \rho(D - dk, \Theta)$ is monotone non-increasing in

D, then

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle\\ &\quad\geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S^{<D}(\mathcal{I}_S)\rangle - D^{dk/2}\cdot\rho(D-dk,\Theta)^{\frac{1}{2}}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{2,\nu}^2\right)^{\frac{1}{2}}\|R\|_{Fr,\nu}\end{aligned}$$

Proof.

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle = \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S(\mathcal{I})\rangle$$

where $\tilde{R}_S(\mathcal{I})$ is now obtained by averaging out the values for all monomials whose degree in $\overline{S}$ is $> dk$. Writing $\tilde{R}_S = \tilde{R}_S^{\leqslant D} + \tilde{R}_S^{>D}$ and applying a Cauchy-Schwartz inequality we get,

$$\begin{aligned}&\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S(\mathcal{I})\rangle\\ &\quad\geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S^{<D}(\mathcal{I})\rangle - \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{\frac{1}{2}}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|\tilde{R}_S^{\geqslant D}\|_{Fr,\nu}\right)^{\frac{1}{2}}\end{aligned}$$

Over a random choice of S,

$$\mathop{\mathbb{E}}_{S\sim\Theta}\|\tilde{R}_S^{\geqslant D}\|_{Fr,\nu}^2 = \sum_{\alpha,|\alpha|\geqslant D}\mathop{\mathbb{P}}_{S\sim\Theta}[|\alpha\cap\overline{S}|\leqslant dk]\cdot\widehat{R}_\alpha^2 \leqslant D^{dk}\cdot\rho(D-dk,\Theta)\cdot\|R\|_{Fr}^2,$$

where we have used that $D^{dk}\rho(D-dk,\Theta)$ is a monotone non-increasing function of D. Substituting this in the earlier inequality the Lemma follows. □

12.4 Applications

In this section we provide two example inference problems for which the conditions of Theorem 12.1.5 hold: planted clique and the spiked tensor model. We will rely upon the (simple) proofs in Section 12.5, which show that some well-conditioned-ness facts about SoS proofs. Verifying that the conditions

hold for other inference problems, like densest-k-subgraph, random constraint satisfaction, sparse PCA and more, involves similar and routine calculations – we refer the reader to [88].

Problem 12.4.1 (Planted clique with clique of size n^δ)**.** Given a graph $G = (V, E)$ on n vertices, determine whether it comes from:

- **Uniform Distribution**: the uniform distribution over graphs on n vertices ($G(n, \frac{1}{2})$).
- **Planted Distribution**: the uniform distribution over n-vertex graphs with a clique of size at least n^δ

The usual polynomial program for *planted clique* in variables $x_1, \ldots, x_n$ is:

$$\begin{aligned} \text{obj} &\leqslant \sum_i x_i \\ x_i^2 &= x_i \quad \forall i \in [n] \\ x_i x_j &= 0 \quad \forall (i,j) \in E \end{aligned}$$

Lemma 12.4.2. *Theorem 12.1.5 applies to the above planted clique program, so long as* $\text{obj} \leqslant n^{\delta-\varepsilon}$ *for any* $\varepsilon \geqslant \frac{c \cdot d}{D-6d}$ *for a fixed constant* c.

Proof. Note that the definition of "degree" is a little subtle for planted clique;; see Remark 12.1.6 following the statement of Theorem 12.1.5.

In this case, the instance degree of the SoS relaxation is $k = 2$. We have from Corollary 12.5.3 that the degree-d SoS refutation is well-conditioned, with numbers bounded by $n^{c_1 \cdot d}$ for some constant $c_1/2$. Define $B = c_1 d \geqslant dk$.

Our subsampling distribution Θ is the distribution given by including every vertex with probability ρ, producing an induced subgraph of $\approx \rho n$ vertices. For

any set of edges α of instance degree at most $D - 6d$,

$$\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha \subseteq S] \leqslant \rho^{D-6d},$$

since the instance degree corresponds to the number of vertices incident on α.

This subsampling operation satisfies the subsample inference condition for the clique constraints with probability 1, since a clique in any subgraph of G is also a clique in G. Also, if there is a clique of size n^δ in G, then by a Chernoff bound

$$\mathop{\mathbb{P}}_{S\sim\Theta}[\exists \text{ clique of size } \geqslant (1-\beta)\rho n^\delta \in S] \geqslant 1 - \exp(-\frac{\beta^2 \rho n^\delta}{2}).$$

Choosing $\beta = \sqrt{\frac{10B\log n}{\rho n^\delta}}$, this gives us that Θ gives n^{-10B}-robust inference for the planted clique problem, so long as obj $\leqslant \rho n/2$. Choosing $\rho = n^{-\varepsilon}$ for ε so that

$$\rho^{D-6d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{c_2 d}{D-6d},$$

for some constant c_2, all of the conditions required by Theorem 12.1.5 now hold. □

Problem 12.4.3 (Spiked tensor model/tensor PCA)**.** Given an order-k tensor in $(\mathbb{R}^n)^{\otimes k}$, determine whether it comes from:

- **Uniform Distribution**: each entry of the tensor sampled independently from $\mathcal{N}(0,1)$.
- **Planted Distribution**: a spiked tensor, $\mathbf{T} = \lambda \cdot v^{\otimes k} + G$ where v is sampled uniformly from $\{\pm\frac{1}{\sqrt{n}}\}^n$, and where G is a random tensor with each entry sampled independently from $\mathcal{N}(0,1)$.

Given the tensor $\mathbf{T}$, the canonical program for the tensor PCA problem in variables $x_1, \ldots, x_n$ is:

$$\text{obj} \leqslant \langle x^{\otimes k}, \mathbf{T}\rangle$$

$$\|x\|_2^2 = 1$$

Lemma 12.4.4. *For $\lambda n^{-\varepsilon} \gg \log n$, Theorem 12.1.5 applies to the* tensor PCA *problem with* obj $\leqslant \lambda n^{-\varepsilon}$ *for any $\varepsilon \geqslant \frac{c \cdot d}{D-3d}$ for a fixed constant c.*

Proof. The degree of the SoS relaxation in the instance is $k = 1$. Since the entries of the noise component of the tensor are standard normal variables, with exponentially good probability over the input tensor $\mathbf{T}$ we will have no entry of magnitude greater than n^d. This, together with Corollary 12.5.3, gives us that except with exponentially small probability the SoS proof will have no values exceeding $n^{c_1 d}$ for a fixed constant c_1.

Our subsampling operation is to set to zero every entry of $\mathbf{T}$ independently with probability $1 - \rho$, obtaining a sub-instance $\mathbf{T}'$ on the nonzero entries. Also, for any $\alpha \in \binom{[n]^k}{D-3d}$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \in S] \leqslant \rho^{D-3d}.$$

This subsampling operation clearly preserves the planted solution unit sphere constraint. Additionally, let $\mathcal{R}$ be the operator that restricts a tensor to the nonzero entries. We have that $\langle \mathcal{R}(\lambda \cdot v^{\otimes k}), v^{\otimes k} \rangle$ has expectation $\lambda \cdot \rho$, since every entry of $v^{\otimes k}$ has magnitude $n^{-k/2}$. Applying a Chernoff bound, we have that this quantity will be at least $(1-\beta)\lambda\rho$ with probability at least n^{-10B} if we choose $\beta = \sqrt{\frac{10B \log n}{\lambda \rho}}$.

It remains to address the noise introduced by $G_{\mathbf{T}'}$ and resampling all the entries outside of the subinstance $\mathbf{T}'$. Each of these entries is a standard normal entry. The quantity $\langle (\mathrm{Id} - \mathcal{R})(N), v^{\otimes k} \rangle$ is a sum over at most n^k i.i.d. Gaussian entries each with standard deviation $n^{-k/2}$ (since that is the magnitude of $(v^{\otimes k})_\alpha$. The entire quantity is thus a Gaussian random variable with mean 0 and variance 1, and therefore with probability at least n^{-10B} this quantity will not exceed

$\sqrt{10B \log n}$. So long as $\sqrt{10B \log n} \ll \lambda\rho$, the signal term will dominate, and the solution will have value at least $\lambda\rho/2$.

Now, we set $\rho = n^{-\varepsilon}$ so that

$$\rho^{D-3d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{2c_1 d}{D-3d},$$

which concludes the proof (after making appropriate adjustments to the constant c_1). □

Remark 12.4.5. For problems where the null model is Gaussian (such as in the spiked tensor model), applying Theorem 12.1.5 yields the existence of distinguishers that are *low-degree* in a non-standard sense. Specifically, the degree of a monomial will be the number of distinct variables in it, irrespective of the powers to which they are raised. To obtain conclusions with respect to the usual notion of degree, Theorem 12.1.5 can be adapted to allow the Ornstein-Uhlenbeck noise operator in place of the subsampling distribution Θ.

12.5 Bounding the sum-of-squares proof ideal term

We give conditions under which sum-of-squares proofs are well-conditioned, using techniques similar to those that appear in [156] for bounding the bit complexity of SoS proofs. We begin with some definitions.

Definition 12.5.1. Let $\mathcal{P}$ be a polynomial optimization problem and let $\mathcal{D}$ be the uniform distribution over the set of feasible solutions S for $\mathcal{P}$. Define the degree-$2d$ moment matrix of $\mathcal{D}$ to be $X_{\mathcal{D}} = \mathbb{E}_{s\sim\mathcal{D}}[\widehat{s}^{\otimes 2d}]$, where $\widehat{s} = [1\ s]^\top$.

- We say that $\mathcal{P}$ *is k-complete on up to degree* $2d$ if every zero eigenvector of $X_{\mathcal{D}}$ has a degree-k derivation from the ideal constraints of $\mathcal{P}$.

Theorem 12.5.2. *Let $\mathcal{P}$ be a polynomial optimization problem over variables $x \in \mathbb{R}^n$ of degree at most $2d$, with objective function $f(x)$ and ideal constraints $\{g_j(x) = 0\}_{j\in[m]}$. Suppose also that $\mathcal{P}$ is $2d$-complete up to degree $2d$. Let G be the matrix of ideal constraints in the degree-$2d$ SoS proof for $\mathcal{P}$. Then if*

- *the SDP optimum value is bounded by $n^{O(d)}$*
- *the coefficients of the objective function are bounded by $n^{O(d)}$,*
- *there is a set of feasible solutions $\mathcal{S} \subseteq \mathbb{R}^n$ with the property that for each $\alpha \subseteq [n]^d$, $|\alpha| \leqslant d$ for which χ_α is not identically zero over the solution space, there exists some $s \in \mathcal{S}$ such that the square monomial $\chi_\alpha(s)^2 \geqslant n^{-O(d)}$,*

it follows that the SoS certificate for the problem is well-conditioned, with no value larger than $n^{O(d)}$.

To prove this, we essentially reproduce the proof of the main theorem of [156], up to the very end of the proof at which point we slightly deviate to draw a different conclusion.

Proof. Following our previous convention, the degree-$2d$ sum-of-squares proof for $\mathcal{P}$ is of the form

$$\text{sdpOpt} - f(x) = a(x) + g(x),$$

where the $g(x)$ is a polynomial in the span of the ideal constraints, and A is a sum of squares of polynomials. Alternatively, we have the matrix characterization,

$$\text{sdpOpt} - \langle F, \widehat{x}^{\otimes 2d}\rangle = \langle A, \widehat{x}^{\otimes 2d}\rangle + \langle G, \widehat{x}^{\otimes 2d}\rangle,$$

where $\widehat{x} = [1\ x]^\top$, F, A, and G are matrix polynomials corresponding to f, a, and g respectively, and with $A \succeq 0$.

Now let $s \in \mathcal{S}$ be a feasible solution. Then we have that

$$\text{sdpOpt} - \langle F, s^{\otimes 2d} \rangle = \langle A, s^{\otimes 2d} \rangle + \langle G, s^{\otimes 2d} \rangle = \langle A, s^{\otimes 2d} \rangle,$$

where the second equality follows because each $s \in S$ is feasible. By assumption the left-hand-side is bounded by $n^{O(d)}$.

We will now argue that the diagonal entries of A cannot be too large. Our first step is to argue that A cannot have nonzero diagonal entries unless there is a solution element in the solution Let $X_{\mathcal{D}} = \mathbb{E}[x^{\otimes 2d}]$ be the $2d$-moment matrix of the uniform distribution of feasible solutions to $\mathcal{P}$. Define Π to be the orthogonal projection into the zero eigenspace of $X_{\mathcal{D}}$. By linearity and orthonormality, we have that

$$\begin{aligned}\langle X_{\mathcal{D}}, A \rangle &= \left\langle X_{\mathcal{D}}, (\Pi + \Pi^{\perp}) A (\Pi + \Pi^{\perp}) \right\rangle \\ &= \left\langle X_{\mathcal{D}}, \Pi^{\perp} A \Pi^{\perp} \right\rangle + \left\langle X_{\mathcal{D}}, \Pi A \Pi^{\perp} \right\rangle + \left\langle X_{\mathcal{D}}, \Pi^{\perp} A \Pi \right\rangle + \langle X_{\mathcal{D}}, \Pi A \Pi \rangle .\end{aligned}$$

By assumption $\mathcal{P}$ is $2d$-complete on $\mathcal{D}$ up to degree $2d$, and therefore Π is derivable in degree $2d$ from the ideal constraints $\{g_j\}_{j \in [m]}$. Therefore, the latter three terms may be absorbed into G, or more formally, we can set $A' = \Pi^{\perp} A \Pi^{\perp}$, $G' = G + (\Pi + \Pi^{\perp}) A (\Pi + \Pi^{\perp}) - \Pi^{\perp} A \Pi^{\perp}$, and re-write the original proof

$$\text{sdpOpt} - \langle F, \widehat{x}^{\otimes 2d} \rangle = \langle A', \widehat{x}^{\otimes 2d} \rangle + \langle G', \widehat{x}^{\otimes 2d} \rangle. \tag{12.5.1}$$

The left-hand-side remains unchanged, so we still have that it is bounded by $n^{O(d)}$ for any feasible solution $s \in \mathcal{S}$. Furthermore, the nonzero eigenspaces of $X_{\mathcal{D}}$ and A' are identical, and so A' cannot be nonzero on any diagonal entry which is orthogonal to the space of feasible solutions.

Now, we argue that every diagonal entry of A' is at most $n^{O(d)}$. To see this, for each diagonal term χ_α^2, we choose the solution $s \in \mathcal{S}$ for which $\chi_\alpha(s)^2 \geqslant n^{-O(d)}$.

We then have by the PSDness of A' that

$$A'_{\alpha,\alpha} \cdot \chi_\alpha(s)^2 \leqslant \langle s^{\otimes 2d}, A' \rangle \leqslant n^{O(d)},$$

which then implies that $A'_{\alpha,\alpha} \leqslant n^{O(d)}$. It follows that $\mathrm{Tr}(A') \leqslant n^{O(d)}$, and again since A' is PSD,

$$\|A'\|_F \leqslant \sqrt{\mathrm{Tr}(A')} \leqslant n^{O(d)}. \tag{12.5.2}$$

Putting things together, we have from our original matrix identity (12.5.1) that

$$\begin{aligned}
\|G'\|_F &= \|\operatorname{sdpOpt} - A' - F\|_F \\
&\leqslant \|\operatorname{sdpOpt}\|_F + \|A'\|_F + \|F\|_F \quad \text{(triangle inequality)} \\
&\leqslant \|\operatorname{sdpOpt}\|_F + n^{O(d)} + \|F\|_F \quad \text{(from (12.5.2))}.
\end{aligned}$$

Therefore by our assumptions that $\|\operatorname{sdpOpt}\|, \|F\|_F = n^{O(d)}$, the conclusion follows. □

We now argue that the conditions of this theorem are met by several general families of problems. (See [88] for an expanded version of Corollary 12.5.3.)

Corollary 12.5.3. *The following problems have degree-2d SoS proofs with all coefficients bounded by $n^{O(d)}$:*

1. *The unit sphere: Any polynomial optimization problem with the only constraints being $\{\sum_{i\in[n]} x_i^2 = 1\}$ and objective value at most $n^{O(d)}$ over the set of feasible solutions. (Including* tensor PCA*).*
2. *The* max clique *problem.*

We prove this corollary below. For each of the above problems, it is clear that the objective value is bounded and the objective function has no large coefficients.

To prove this corollary, we need to verify the completeness of the constraint sets, and then demonstrate a set of feasible solutions so that each square term receives non-negligible mass from some solution.

A large family of completeness conditions were already verified by [156] and others (see the references therein):

Proposition 12.5.4 (Completeness of canonical polynomial optimization problems (from Corollary 3.5 of [156]))**.** *The following pairs of polynomial optimization problems $\mathcal{P}$ and distributions over solutions $\mathcal{D}$ are complete:*

1. *If the feasible set is $x \in \mathbb{R}^n$ with $\sum_{i\in[n]} x_i^2 = \alpha$, then $\mathcal{P}$ is d-complete on $\mathcal{D}$ up to degree d (e.g. if $\mathcal{P}$ is the tensor PCA problem).*
2. *If $\mathcal{P}$ is the* MAX CLIQUE *problem with feasible set $x \in \mathbb{R}^n$ with $\{x_i^2 = x_i\}_{i\in[n]} \cup \{x_i x_j = 0\}_{(i,j)\in E}$, then $\mathcal{P}$ is d-complete on $\mathcal{D}$ up to degree d.*

Proof of 12.5.3. We verify the conditions of Theorem 12.5.2 separately for each case.

1. The unit sphere: the completeness conditions are satisfied by 12.5.4. We choose the set of feasible solutions to contain a single point, $s = \frac{1}{\sqrt{n}} \cdot \vec{1}$, for which $\chi_\alpha^2(s) \geqslant n^{-d}$ as long as $|\alpha| \leqslant d$, which meets the conditions of Theorem 12.5.2.
2. The MAX CLIQUE problem: the completeness conditions are satisfied by 12.5.4. We choose the solution set $\mathcal{S}$ to be the set of $0,1$ indicators for cliques in the graph. Any α that corresponds to a non-clique in the graph has χ_α identically zero in the solution space. Otherwise, $\chi_\alpha(s)^2 = 1$ when $s \in \mathcal{S}$ is the indicator vector for the clique on α.

This concludes the proof. □

12.6 Chapter Notes

The content in this chapter originally appeared in [88], joint work with Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer.

BIBLIOGRAPHY

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017. 12, 140

[2] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *FOCS*, pages 670–688. IEEE Computer Society, 2015. 14, 131, 151

[3] Emmanuel Abbe and Colin Sandon. Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation. In *NIPS*, pages 1334–1342, 2016. 128, 129, 134, 135, 145, 173, 174

[4] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005. 18, 291

[5] Dimitris Achlioptas and Assaf Naor. The two possible values of the chromatic number of a random graph. *Ann. of Math. (2)*, 162(3):1335–1351, 2005. 418

[6] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. In *NIPS*, pages 33–40. Curran Associates, Inc., 2008. 15, 130, 140

[7] Sarah R. Allen, Ryan O'Donnell, and David Witmer. How to refute a random CSP. In *FOCS*, pages 689–708. IEEE Computer Society, 2015. 49, 104, 105

[8] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *STOC*, pages 496–505. ACM, 2007. 364

[9] Noga Alon, Michael Krivelevich, and Benny Sudakov. Finding a large hidden clique in a random graph. *Random Struct. Algorithms*, 13(3-4):457–466, 1998. 9, 49, 364

[10] Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *STOC*, pages 72–80. ACM, 2004. 169

[11] Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *J. ACM*, 42(4):844–856, 1995. 131, 161

[12] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham Kakade. A tensor spectral approach to learning mixed membership community models. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 867–881. JMLR.org, 2013. 139, 147, 148, 157

[13] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014. 148

[14] Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 1135–1164. JMLR.org, 2014. 18, 293

[15] Benny Applebaum, Boaz Barak, and Avi Wigderson. Public-key cryptography from different assumptions. In *STOC*, pages 171–180. ACM, 2010. 7, 364

[16] Sanjeev Arora, Boaz Barak, Markus Brunnermeier, and Rong Ge. Computational complexity and information asymmetry in financial products (extended abstract). In *ICS*, pages 49–65. Tsinghua University Press, 2010. 7, 364

[17] Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Probab.*, 15(1A):69–92, 2005. 18, 235, 291

[18] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *arXiv preprint arXiv:1711.05424*, 2017. 105

[19] Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013. 81

[20] Per Austrin, Mark Braverman, and Eden Chlamtac. Inapproximability of np-complete variants of nash equilibrium. *Theory of Computing*, 9:117–142, 2013. 7, 364

[21] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for

efficiently learning linear separators with noise. In *STOC*, pages 449–458, 2014. 294

[22] Jinho Baik, Gérard Ben Arous, Sandrine Péché, et al. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005. 128, 163

[23] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *CoRR*, abs/1408.2156, 2014. 293

[24] Jess Banks, Robert Kleinberg, and Cristopher Moore. The lovász theta function for random regular graphs and community detection in the hard regime. *arXiv preprint arXiv:1705.01194*, 2017. 418

[25] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 383–416. JMLR.org, 2016. 13

[26] Boaz Barak. Introduction to theoretical computer science, 2018. 80

[27] Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *STOC*, pages 307–326. ACM, 2012. 65, 113, 295

[28] Boaz Barak, Siu On Chan, and Pravesh K. Kothari. Sum of squares lower bounds from pairwise independence. In *STOC*, pages 97–106. ACM, 2015. 64, 105

[29] Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *FOCS*, pages 428–437. IEEE Computer Society, 2016. 68, 71, 96, 105, 362

[30] Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *STOC*, pages 31–40. ACM, 2014. 65, 107, 108, 111, 112, 113, 126, 259, 295

[31] Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning

and tensor decomposition via the sum-of-squares method. In *STOC*, pages 143–151. ACM, 2015. 65, 112, 148, 160

[32] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 417–445. JMLR.org, 2016. 105

[33] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *FOCS*, pages 472–481. IEEE Computer Society, 2011. 51, 112

[34] Boaz Barak and David Steurer. The sos algorithm over general domains. http://www.sumofsquares.org/public/lec-definitions-general.html, 2017. [Online; accessed 11-1-2017]. 50, 73, 74, 249, 284

[35] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112. IEEE Computer Society, 2010. 18, 292

[36] T. Bernholt. Robust estimators are hard to compute. Technical report, University of Dortmund, Germany, 2006. 238, 294

[37] Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941. 283, 284

[38] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 1046–1066. JMLR.org, 2013. 303, 363, 364

[39] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *STOC*, pages 594–603. ACM, 2014. 18, 293

[40] Vijay V. S. P. Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee. Certifying random polynomials over the unit sphere via sum of squares hierarchy. *CoRR*, abs/1605.00903, 2016. 84

[41] Andrej Bogdanov and Luca Trevisan. On worst-case to average-case reductions for NP problems. *SIAM J. Comput.*, 36(4):1119–1159, 2006. 364

[42] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs. In *FOCS*, pages 1347–1357. IEEE Computer Society, 2015. 14, 173

[43] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013. 76

[44] Jonah Brown-Cohen, Prasad Raghavendra, and Henry Yuen. personal communication. 60

[45] S. C. Brubaker. Robust PCA and clustering in noisy mixtures. In *SODA 2009*, pages 1078–1087, 2009. 294

[46] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011. 294

[47] Eric Carlen. Trace inequalities and quantum entropy: An introductory course. 2009. 382

[48] Siu On Chan, Dimitris Papailliopoulos, and Aviad Rubinstein. On the approximability of sparse PCA. In *COLT*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 623–646. JMLR.org, 2016. 359

[49] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. *CoRR*, abs/1611.02315, 2016. 294

[50] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006. 12

[51] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010. 148

[52] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of computer science, 1999. 40th annual symposium on*, pages 634–644. IEEE, 1999. 235, 291

[53] Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007. 18, 235, 291, 293

[54] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, 2014. 18, 292

[55] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. *Conference on Learning Theory*, 2017. 18, 293

[56] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *NIPS*, pages 41–48, 2004. 126

[57] Victor H. de la Peña and S. J. Montgomery-Smith. Bounds on the tail probability of U-statistics and quadratic forms. *Bull. Amer. Math. Soc. (N.S.)*, 31(2):223–227, 1994. 226

[58] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *CoRR*, abs/1109.3041, 2011. 24, 129, 142

[59] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Phase transition in the detection of modules in sparse networks. *CoRR*, abs/1102.1182, 2011. 13

[60] Laurent Demanet and Paul Hand. Scaling law for recovering the sparsest element in a subspace. *Inf. Inference*, 3(4):295–309, 2014. 126

[61] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. 130

[62] Yash Deshpande and Andrea Montanari. Sparse PCA via covariance thresholding. In *NIPS*, pages 334–342, 2014. 49, 107, 303, 358, 360

[63] Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 523–562. JMLR.org, 2015. 66, 232, 331, 363

[64] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without

the computational intractability. In *FOCS*, pages 655–664. IEEE Computer Society, 2016. 238, 245, 266, 294

[65] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *ICML*, 2017. 238, 294

[66] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. *arXiv preprint arXiv:1611.03473*, 2016. 294

[67] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073. ACM, 2018. 294

[68] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, 2018. 290

[69] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large k. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 59–68. ACM, 2015. 15

[70] Jian Ding, Allan Sly, and Nike Sun. Satisfiability threshold for random regular nae-sat. *Communications in Mathematical Physics*, 341(2):435–489, 2016. 15

[71] Uriel Feige and Robert Krauthgamer. The probable value of the lovász–schrijver relaxations for maximum independent set. *SIAM J. Comput.*, 32(2):345–370, 2003. 10, 46, 66, 302, 363

[72] Joan Feigenbaum and Lance Fortnow. Random-self-reducibility of complete sets. *SIAM J. Comput.*, 22(5):994–1005, 1993. 364

[73] Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. PAC learning axis-aligned mixtures of gaussians with no separation assumption. In *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 20–34. Springer, 2006. 18, 292

[74] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of*

statistical learning, volume 1. Springer series in statistics New York, NY, USA:, 2001. 107

[75] Robert Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28, 1962. 130

[76] Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In *STOC*, pages 761–770. ACM, 2015. 18, 148, 293

[77] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. In *APPROX-RANDOM*, volume 40 of *LIPIcs*, pages 829–849. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015. 49, 105

[78] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995. 51

[79] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *STOC*, pages 584–593. ACM, 2014. 148

[80] Dima Grigoriev. Complexity of positivstellensatz proofs for the knapsack. *Computational Complexity*, 10(2):139–154, 2001. 60

[81] Dima Grigoriev. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theor. Comput. Sci.*, 259(1-2):613–622, 2001. 35, 60, 64, 105

[82] Bruce E. Hajek, Yihong Wu, and Jiaming Xu. Computational lower bounds for community detection on random graphs. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 899–928. JMLR.org, 2015. 7, 364

[83] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics. The approach based on influence functions*. Wiley New York, 1986. 294

[84] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *STOC*, pages 753–760. ACM, 2015. 18, 292

[85] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis. 1970. 157

[86] Elad Hazan and Robert Krauthgamer. How hard is it to approximate the best nash equilibrium? *SIAM J. Comput.*, 40(1):79–91, 2011. 7, 364

[87] Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *J. ACM*, 60(6):45:1–45:39, 2013. 4

[88] Samuel B Hopkins, Pravesh Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. *Symposium on Foundations of Computer Science*, 2017. 96, 103, 105, 303, 362, 367, 391, 397, 399

[89] Samuel B. Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm. On the integrality gap of degree-4 sum of squares for planted clique. In *SODA*, pages 1079–1095. SIAM, 2016. 66, 331, 363

[90] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034. ACM, 2018. 251, 290

[91] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *STOC*, pages 178–191. ACM, 2016. 103, 105, 110, 120, 126

[92] Samuel B. Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 956–1006. JMLR.org, 2015. 4, 84, 87, 88, 103, 105, 107

[93] Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 379–390. IEEE, 2017. 233

[94] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS'13—Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science*, pages 11–19. ACM, New York, 2013. 18, 293

[95] Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. 237, 294

[96] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. 28

[97] Hamid Javadi and Andrea Montanari. The hidden subgraph problem. *arXiv preprint arXiv:1511.05254*, 2015. 7, 364

[98] Mark Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 3(4):347–360, 1992. 364

[99] D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978. 238, 294

[100] Ari Juels and Marcus Peinado. Hiding cliques for cryptographic security. *Des. Codes Cryptography*, 20(3):269–280, 2000. 7, 364

[101] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, pages 553–562. ACM, 2010. 18, 292

[102] Richard M. Karp. Probabilistic analysis of some combinatorial search problems. *Algorithms and Complexity: New Directions and Recent Results*, 1976. 364

[103] M. J. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993. 294

[104] Jonathan Kelner. personal communication, via Boaz Barak. 67

[105] A. Klivans, P. Long, and R. Servedio. Learning halfspaces with malicious noise. 2009. 294

[106] Pascal Koiran and Anastasios Zouzias. Hidden cliques and the certification of the restricted isometry property. *IEEE Trans. Information Theory*, 60(8):4999–5006, 2014. 364

[107] Pravesh K Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer. Sum of squares lower bounds for refuting any csp. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 132–145. ACM, 2017. 35, 64, 105

[108] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of*

the 50th Annual ACM SIGACT Symposium on Theory of Computing, pages 1035–1046. ACM, 2018. 290

[109] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al. Do semidefinite relaxations solve sparse pca up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 2015. 363

[110] Jean-Louis Krivine. Anneaux préordonnés. *Journal dâĂŹanalyse mathématique*, 12(1):307–326, 1964. 45

[111] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption: clustering sparse networks. *CoRR*, abs/1306.5550, 2013. 14

[112] Ludek Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2-3):193–212, 1995. 364

[113] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *FOCS*, pages 299–308. IEEE Computer Society, 2010. 18, 291

[114] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *FOCS*, pages 665–674. IEEE Computer Society, 2016. 238, 294

[115] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. 122

[116] G. Lerman, M. B. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *CoRR*, abs/1202.4044, 2012. 294

[117] Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. *arXiv preprint arXiv:1701.08010*, 2017. 105

[118] S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM J. Matrix Anal. Appl.*, 14(4):1064–1083, 1993. 157

[119] Elaine Levey and Thomas Rothvoss. A lasserre-based $(1+\varepsilon)$-approximation for $Pm \mid p_j = 1, \text{prec} \mid C_{\max}$. *CoRR*, abs/1509.07808, 2015. 112

[120] Jerry Li and Ludwig Schmidt. Robust and proper learning for mixtures of gaussians via systems of polynomial inequalities. In *Conference on Learning Theory*, 2017. 18, 292

[121] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *FOCS*, pages 438–446. IEEE Computer Society, 2016. 75, 78, 105, 148, 150, 160, 220, 224, 225, 226

[122] Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. In *Advances in Neural Information Processing Systems*, pages 1612–1620, 2015. 363

[123] Edward M Marcotte, Matteo Pellegrini, Ho-Leung Ng, Danny W Rice, Todd O Yeates, and David Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999. 12

[124] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. *CoRR*, abs/1311.3085, 2013. 165

[125] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *STOC*, pages 694–703. ACM, 2014. 14, 128, 129, 134, 135, 165

[126] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004. 234

[127] Dhruv Medarametla and Aaron Potechin. Bounds on the norms of uniform low degree graph matrices. In Klaus Jansen, Claire Mathieu, José D. P. Rolim, and Chris Umans, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France*, volume 60 of *LIPIcs*, pages 40:1–40:26. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2016. 331, 353

[128] Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *STOC*, pages 87–96. ACM, 2015. 66, 363

[129] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009. 105

[130] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon.

Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. 7, 364

[131] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, page iax001, 2017. 291

[132] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102. IEEE Computer Society, 2010. 18, 292

[133] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *STOC*, pages 814–827. ACM, 2016. 417

[134] Andrea Montanari and Nike Sun. Spectral algorithms for tensor completion. *CoRR*, abs/1612.07866, 2016. 105

[135] Cristopher Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *CoRR*, abs/1702.00467, 2017. 144

[136] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *CoRR*, abs/1311.4115, 2013. 14, 165, 167

[137] Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *COLT*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 356–370. JMLR.org, 2014. 14

[138] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *STOC*, pages 69–75. ACM, 2015. 128, 129, 134, 135, 153

[139] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields*, 162(3-4):431–461, 2015. 165

[140] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002. 12

[141] Jerzy Neyman and Egon S Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706):289–337, 1933. 29

[142] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014. 37, 38, 95

[143] Ryan O'Donnell. SOS is not obviously automatizable, even approximately. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:141, 2016. 46, 75, 249

[144] Ryan O'Donnell and Yuan Zhou. Approximability and proof complexity. In *SODA*, pages 1537–1556. SIAM, 2013. 295

[145] Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982. 130

[146] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. 234

[147] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. *CoRR*, abs/1612.07728, 2016. 81

[148] Pavel A Pevzner, Sing-Hoi Sze, et al. Combinatorial approaches to finding subtle signals in dna sequences. In *ISMB*, volume 8, pages 269–278, 2000. 7, 364

[149] Aaron Potechin. Sum of squares lower bounds from symmetry and a good story. *arXiv preprint arXiv:1711.11469*, 2017. 60

[150] Aaron Potechin and David Steurer. Exact tensor completion with sum-of-squares. *CoRR*, abs/1702.06237, 2017. 78, 105

[151] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *NIPS*, pages 3401–3409, 2014. 126

[152] Prasad Raghavendra. Optimal algorithms and inapproximability results for every csp? In *STOC*, pages 245–254. ACM, 2008. 6

[153] Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random csps below the spectral threshold. *CoRR*, abs/1605.00058, 2016. 35, 49, 84, 104

[154] Prasad Raghavendra and David Steurer. How to round any CSP. In *FOCS*, pages 586–594. IEEE Computer Society, 2009. 51

[155] Prasad Raghavendra and Ning Tan. Approximating csps with global cardinality constraints using SDP hierarchies. In *SODA*, pages 373–387. SIAM, 2012. 112

[156] Prasad Raghavendra and Benjamin Weitz. On the bit complexity of sum-of-squares proofs. *CoRR*, abs/1702.05139, 2017. 46, 75, 249, 394, 395, 398

[157] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011. 49

[158] Oded Regev and Aravindan Vijayraghavan. On learning mixtures of well-separated gaussians. In *Symposium on Foundations of Computer Science*, 2017. 18, 236, 237, 292

[159] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *NIPS*, pages 2897–2905, 2014. 87, 103

[160] Grant Schoenebeck. Linear level lasserre lower bounds for certain k-csps. In *FOCS*, pages 593–602. IEEE Computer Society, 2008. 64

[161] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 1760–1793. PMLR, 2017. 148, 150, 151, 218

[162] R. Servedio. Smooth boosting and learning with malicious noise. *JMLR*, 4:633–648, 2003. 294

[163] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 12

[164] Allan Sly, Nike Sun, and Yumeng Zhang. The number of solutions for

random regular nae-sat. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 724–731. IEEE, 2016. 15

[165] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *COLT*, volume 23 of *JMLR Proceedings*, pages 37.1–37.18. JMLR.org, 2012. 111, 126

[166] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. 2017. 238, 294

[167] Gilbert Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974. 45

[168] Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014. 18, 292

[169] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985. 234

[170] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. 76, 102, 279

[171] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531, 1975. 237, 238, 294

[172] Madhur Tulsiani. CSP gaps and reductions in the lasserre hierarchy. In *STOC*, pages 303–312. ACM, 2009. 60, 64

[173] Leslie G. Valiant. Learning disjunction of conjunctions. In *IJCAI*, pages 560–566. Morgan Kaufmann, 1985. 294

[174] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, page 113. IEEE Computer Society, 2002. 18, 235, 291

[175] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010. 87, 101, 118, 122, 124, 289

[176] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, january 2011, 2011. 118

[177] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. 42

[178] Wikipedia. Dirichlet distribution — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Dirichlet%20distribution&oldid=762020989, 2017. [Online; accessed 30-March-2017]. 141

[179] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983. 18, 293

[180] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016. 18, 293

[181] T. Zhang and G. Lerman. A novel m-estimator for robust pca. *J. Mach. Learn. Res.*, 15(1):749–808, January 2014. 294

APPENDIX A

OPEN PROBLEMS

Convex Relaxations for the Sparse Stochastic Block Model. *Design a natural, convex-relaxation-based algorithm to correctly label a* $(1/2 + \delta)$*-fraction of vertices for some* $\delta > 0$ *in a graph from the symmetric* 2*-community block model* $SBM(d, \varepsilon)$*, for any* $d > 1/\varepsilon^2$*.*

As we saw in Chapter 8, algorithms based on simple statistics – nonbacktracking walks and more – are able to accurately estimate communities for any $d > 1/\varepsilon^2$. Ultimately these guarantees are captured by low-degree spectral methods (the nonbacktracking operator), so they can also be captured by ad-hoc semidefinite programs. Nonetheless, such SDPs are unusual – they would not, for example, look at all like the SoS program one would get from the hypothesis testing/refutation/estimation SoS program constructions in Chapter 3. The best guarantees achieved by a more natural SDP are due to Montanari and Sen [133], who study the SDP

$$\max \langle X, A \rangle \text{ such that } X_{ii} = 1, X \succeq 0$$

where A is the adjacency matrix of a graph. They show that this SDP can recover a $(1/2 + \delta)$-fraction of the vertices when $d > 1/\varepsilon^2 + 1/d^{\Omega(1)}$.

In may be that the additive $1/d^{\Omega(1)}$ is a weakness of the SDP itself – this SDP, which is a relaxation of the maximum-likelihood estimator, simply may not obtain the same recovery guarantees as algorithms based on simple statistics. Indeed, it is possible that the maximum-likelihood estimator does not achieve such guarantees. In this case, a possible avenue to designing a better convex relaxation is to relax the *variational characterization of the posterior distribution*,

recalling that for a graph G from the block model with hidden communities x, the conditional distribution on x is given by

$$\{x \mid G\} = \arg \max_{\mu \in \Delta_{\{-1,1\}^n}} C(\varepsilon, d) \mathop{\mathbb{E}}_{x \sim \mu} \langle x, Gx \rangle + H(\mu)$$

where $C(\varepsilon, d)$ is a function of the parameters of the block model. Of course, this problem is already convex, but it is 2^n-dimensional. By replacing distributions μ with low degree pseudodistributions one might hope to find a lower-dimensional versions. This approach seems to require finding a new notion of the entropy of a pseudodistribution, which would be of interest in its own right.

SoS and colorability of random graphs. *Prove an $\Omega(n)$-degree SoS lower bound for refuting k-colorability of typical $G \sim G(n, \frac{d}{n})$ for constant d and $k \gg \sqrt{d}$.* A random n-node graph with average degree d (or a random d-regular graph) with high probability has chromatic number $\Omega(d/\log d)$ [5]. The strongest known polynomial-time refutation algorithms are able to certify only that $\Omega(\sqrt{d})$ colors are required, and improving on this $\sqrt{d}$ appears computationally intractable. The simple statistics heuristic suggests that testing whether a graph is sampled from $G(n, d/n)$ or a random k-color graph model should be computationally intractable for algorithms with running times at least $2^{n^{\Omega(1)}}$, when $k \gg \sqrt{d}$. (This is a special case of the hypothesis testing task in the k-community stochastic block model.)

However, the best available SoS lower bounds for this problem are stuck at degree-2 SoS [24]. While the pseudocalibration method should apply to this problem, it appears to capture many of the features which foil known techniques for analyzing the pseudocalibration construction – it involves sparse random matrices, which have more complicated spectra than the dense random matrices

analyzed in Chapter 11, but it has the non-local flavor of planted clique, so techniques used previously for CSP lower bounds seem not to apply. It appears likely that progress on SoS lower bounds for k-coloring would significantly improve our understanding of pseudocalibration in general.