Università Commerciale "Luigi Bocconi"
Ph.D. Program in Statistics and Computer Science

Algorithms Beyond the Union Bound

Polynomial Optimization and Discrepancy Theory

Lucas Pesenti

Abstract

The *union bound* is a classical tool in the probabilistic method for proving the existence of objects with extremal features by showing that a random object satisfies each feature with high probability. This approach has powered major results spanning theoretical computer science, combinatorics, random matrix theory, and statistical physics — such as the existence of graph sparsifiers, satisfying assignments to constraint satisfaction problems, and low-energy configurations in spin glass models. However, the union bound ignores correlations between different features, and thereby often leads to suboptimal results in these applications.

This thesis presents an alternative approach: proving existence through the design and analysis of algorithms that explicitly construct the desired objects. Each of the above problems can be reduced to minimizing either low-degree polynomials or linear-image norms in high dimensions. For these objectives, we analyze algorithms inspired by *continuous optimization*.

In the first part, we propose an orthogonal representation for first-order algorithms optimizing random quadratic polynomials using *Fourier analysis*. We show that in the high-dimensional limit, these algorithms can be analyzed by tracking their dynamics in a *tree basis*. This reframes random polynomial optimization as a combinatorial problem, which we explicitly solve in some cases. Our approach also yields the first direct justification, in this setting, of the heuristic *cavity method* from physics.

The second part extends the approach to general polynomial optimization. We introduce a multiscale union bound argument for random tensors, extending results of Friedman and Wigderson on the spectral gap of random hypergraphs. We further present a new rounding scheme for *semidefinite programming* relaxations, leading to improved approximation guarantees for homogeneous cubic optimization and the Max-3-Sat problem.

In the third part, we design an algorithm for minimizing norms of linear functions via Newton's method applied to a *regularized* objective function. By varying the regularizer, we recover and generalize foundational results in *discrepancy theory*. This framework yields an improved bound in Spencer's classical result from the 1980s, and establishes that the Beck–Fiala and Komlós conjectures hold for a new class of *pseudorandom* instances.

Together, these results show that algorithms can not only match, but also exceed the power of probabilistic arguments for finding extremal objects.

To L.T.

Table of Contents

CHAPTER 1.

Introduction

This thesis takes an algorithmic perspective on problems at the intersection of combinatorics, random matrix theory, statistical physics, and optimization. Such questions often reduce to finding rare objects satisfying a collection of desirable properties $\mathcal{P}$, or proving that none exist.

The classical approach to these problems is the *probabilistic method*: consider a random object and apply a *union bound* to argue that, with positive probability, it satisfies all desired properties simultaneously:

$$\sum_{P \in \mathcal{P}} \Pr_{\boldsymbol{x}} (\boldsymbol{x} \notin P) < 1 \implies \bigcap_{P \in \mathcal{P}} P \neq \varnothing \,. \tag{1.1}$$

While simple and effective, the argument (1.1) fails to account for correlations between properties. As a result, it leads to suboptimal bounds for many fundamental problems.

To address these limitations, this thesis explores an alternative route: designing efficient algorithms that directly target the desired objects. We show that analyzing such algorithms yields not just constructive bounds, but often the simplest, most conceptual, and even strongest proofs of results that were traditionally tackled with probabilistic arguments.

1.1. Motivating examples

We introduce four illustrative problems that will serve as recurring themes throughout this thesis. They are both foundational in their respective field and share a common feature: being amenable to a simple, though suboptimal, probabilistic argument. In contrast, the algorithmic and optimization perspective will lead to results that are not only stronger, but also more illuminating.

1.1.1. Satisfying Boolean formulas

The 3-Sat problem is a prototypical hard problem in theoretical computer science. It appears on Karp's original list of NP-complete problems [Kar72] and is very often used as a starting point for reductions to prove that other problems are NP-hard.

Problem 1.1 (Max-3-Sat)**.** A 3-Sat instance is a Boolean formula given as a list of distinct clauses. Each clause is a disjunction of three literals (a variable or its negation). The goal is to find a truth assignment maximizing the number of satisfied clauses. We denote by n the number of variables.

A central perspective in this thesis is that this problem can be viewed as maximizing a *cubic polynomial* over the *hypercube*, i.e., a function $f : \{-1, 1\}^n \to \mathbb{R}$ of the form

$$f(x_1, \ldots, x_n) = a + \sum_{i=1}^{n} b_i x_i + \sum_{i,j=1}^{n} c_{i,j} x_i x_j + \sum_{i,j,k=1}^{n} d_{i,j,k} x_i x_j x_k \tag{1.2}$$

for real coefficients determined by the formula. Each truth assignment corresponds to a vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{-1, 1\}^n$, and the maximum value of f over the hypercube equals the maximal number of clauses that can be satisfied simultaneously.

Unless P = NP, Problem 1.1 cannot be solved exactly in polynomial time. Instead, we study *approximation algorithms* for this problem, under the promise that the input formula is satisfiable (i.e., when there exists an assignment satisfying all clauses).

As a baseline, a *random* assignment satisfies any clause with probability $7/8$, so by linearity of expectation, it satisfies on average a $7/8$-fraction of the clauses. Håstad and Venkatesh [HV04] show that sampling independently a polynomial number of random assignments yields one that satisfies a $7/8 + \Omega(n^{-3/2})$ fraction of clauses.[1]

Conversely, Håstad [Hås01] shows that unless P = NP, no polynomial-time algorithm can find an assignment satisfying a $7/8 + \varepsilon$ fraction of clauses in a satisfiable 3-Sat formula, *for any constant* $\varepsilon > 0$ (independent of n). This leaves a polynomial gap regarding how much of a gain one can achieve compared to the guarantees of a random assignment.

Which of [Hås01] or [HV04] truly reflects the power of efficient algorithms?

In this thesis, we give the first improvement on [HV04] using tools from convex optimization. The formulation (1.2) will guide our approach, as with our next example, which is an average-case variant of polynomial optimization.

1.1.2. Finding ground states in spin glass models

Our second example is a random optimization problem originating in statistical mechanics: finding the configuration of minimal energy for atoms in disordered magnetic materials. In

[1] The analysis relies on a simple anti-concentration inequality and does not use the promise that the formula is satisfiable. Hence, it also gives a lower bound on "how unsatisfiable" an arbitrary 3-Sat formula can be. Improving it or proving a matching upper bound remains an open problem.

a statistical physics model, each atom has a spin x_i, and a spin configuration $\boldsymbol{x} = (x_1, \dots, x_n)$ has an associated energy $H(\boldsymbol{x})$. At zero temperature, configurations concentrate on *ground states* of the system, which are the minimizers of $\boldsymbol{x} \mapsto H(\boldsymbol{x})$.

The Sherrington–Kirkpatrick and the spherical model

The *Sherrington–Kirkpatrick model* [SK75] is a mean-field statistical physics model where the energy function H is given by a random quadratic polynomial.

Problem 1.2 (Sherrington–Kirkpatrick model)**.** Consider n atoms indexed by $\{1, \dots, n\}$. Each pair $\{i, j\}$ is independently assigned a random ± 1 coupling that favors alignment $(+1)$ or anti-alignment (-1) with equal probability. The goal is to partition the atoms into two groups to maximize the "negative energy" (i.e., minimize the energy), defined as the number of aligned pairs in the same group plus anti-aligned pairs in different groups.

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be defined by $A_{ij} = 1$ if atoms i and j tend to align, and $A_{ij} = -1$ otherwise. $\boldsymbol{A}$ is a matrix with i.i.d. random ± 1 entries (up to the symmetry). A partition into two groups can be represented by a vector $\boldsymbol{x} \in \{-1, 1\}^n$, where the sign of x_i indicates the group assignment of atom i. Then the objective of Problem 1.2 becomes

$$\max_{\boldsymbol{x} \in \{-1,1\}^n} \frac{1}{n^{1.5}} \sum_{i,j=1}^{n} A_{ij} x_i x_j \,, \tag{1.3}$$

where the $n^{1.5}$ normalization makes the optimum typically on the $\Theta(1)$ scale, as we will see below. The diagonal elements of $\boldsymbol{A}$ contribute negligibly as $n \to \infty$, so that (1.3) is equivalent to optimizing over $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_\infty \leqslant 1\}$.[2]

A natural, tractable upper bound on (1.3) is obtained by relaxing the Boolean constraint to a spherical one, namely with $\mathcal{S}^{n-1}(\sqrt{n}) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = \sqrt{n}\}$:

$$\max_{\boldsymbol{x} \in \mathcal{S}^{n-1}(\sqrt{n})} \frac{1}{n^{1.5}} \sum_{i,j=1}^{n} A_{ij} x_i x_j \,. \tag{1.4}$$

Unlike (1.3), (1.4) has a special spectral structure: it is equivalent to computing the largest eigenvalue of $\boldsymbol{A}$.

Though both are polynomial optimization problems, there are two key differences between Problem 1.2 and Problem 1.1: (1) the objective function of the former is a homogeneous quadratic polynomial instead of a non-homogeneous cubic polynomial, and (2) we now study the average-case setting where the coefficients of the polynomial are typical.[3] It is also possible to directly view (1.3) as encoding a *constraint satisfaction problem* analogous to 3-Sat, namely the Max-Cut problem on a not-too-sparse random graph [DMS17].

[2] If $\operatorname{diag}(\boldsymbol{A}) = 0$, this is a multilinear polynomial, whose maximum over the solid cube is attained at a vertex.
[3] Most results in this thesis are *universal*, i.e., robust to different notions of "typicality" [CH06].

Ground state energy

Numerical simulations suggest that both (1.3) and (1.4) converge to deterministic constants as $n \to \infty$, with high probability over the randomness in $\boldsymbol{A}$: These constants are respectively $P_* \approx 1.52$ for (1.3) and 2 for (1.4). Surprisingly, a rigorous explanation of this phenomenon proved to be a major challenge for both mathematicians and physicists.

A simple union bound shows that (1.3) is $O(1)$ with high probability: for fixed $\boldsymbol{x} \in \{-1,1\}^n$, $\sum_{i,j} A_{ij} x_i x_j$ is a sum of independent random variables, so is approximately Gaussian with standard deviation $O(n)$. It is thus exponentially unlikely to exceed $\Omega(n^{3/2})$. Taking a union bound over the 2^n possible choices of $\boldsymbol{x}$ completes the proof.[4] However, this union bound argument, as well as its powerful refinements [Tal21], fall short of giving sharp constants, and thus cannot explain the above observation.

The sharp constant in (1.4) follows from classical results in random matrix theory, using the *trace method* [FK81, BY88]. The key idea is that while the largest eigenvalue of $\boldsymbol{A}$ may not be directly accessible, the ℓ_p-norm of eigenvalues of $\boldsymbol{A}$ have a combinatorial interpretation as a sum over walks encoded by $\boldsymbol{A}$. Counting these walks for larger and larger p provides increasingly tight approximation to the spectral norm of $\boldsymbol{A}$, and a similar argument also yields a matching lower bound. However, this approach heavily relies on the spectral interpretation of the spherical constraint.

The case of (1.3) is far more subtle. In the 1980s, Parisi [Par80] predicted the existence of a limit P_* and expressed it as the solution of an optimization problem that is independent of n. His influential insights remained non-rigorous for decades, eventually culminating in a proof of existence of P_* [GT02] and a proof of the Parisi formula by Guerra and Talagrand [Gue03, Tal06]. A key idea underlying these works is *interpolation* between Gaussian processes.

While powerful, these tools are highly specialized and technically involved. For instance, the ground state energy of the *bipartite* Sherrington–Kirkpatrick model

$$\max_{\boldsymbol{x},\boldsymbol{y} \in \{-1,1\}^n} \frac{1}{n^{1.5}} \sum_{i,j=1}^{n} A_{ij} x_i y_j$$

remains rigorously unknown to this day [CM25]. This suggests that existing methods may be overly rigid.

Are there simple and robust methods for analyzing the value of random optimization problems?

In this thesis, we propose a new framework for proving *lower bounds* on quantities such as (1.3), using a basis for low-degree algorithms optimizing random quadratic polynomials.

[4] A similar approach using ε-nets applies to (1.4).

While mathematically rigorous, our approach parallels heuristic arguments from physics based on the *cavity method*.

1.1.3. Sparsifying graphs

Our final examples depart from polynomial optimization per se, but we will see that similar themes arise.

Graph sparsification is a fundamental compression primitive: given a dense graph, the goal is to construct a sparse subgraph that approximates its structure. The sparsifier can then be used to efficiently approximate queries on the original graph, leading to faster and more memory-efficient algorithms, both in theory and in practice [ST14, BSST13, HGN+24].

Cut sparsification

We focus on the notion of *cut sparsification* [BK96], which aims to preserve the weight of all cuts in the graph. For example, if an edge is critical in the sense that removing it would disconnect the graph, it should appear in the sparsifier.

Problem 1.3 (Cut-preserving sparsifiers)**.** Given an undirected graph $G = (V, E)$, a *cut sparsifier* of G with error $\varepsilon > 0$ is a weighted subgraph H of G such that

$$\forall S \subseteq V\,, \qquad (1-\varepsilon)\,\mathrm{cut}_G(S) \leqslant \mathrm{cut}_H(S) \leqslant (1+\varepsilon)\,\mathrm{cut}_G(S)\,, \tag{1.5}$$

where the *cut value*, $\mathrm{cut}_G(S)$, denotes the total weight of edges in G with one endpoint in S and one endpoint in $V \setminus S$ (and similarly for H).

The results of Benczúr and Karger [BK96] and Spielman, Teng, and Srivastava [ST11, SS11] imply that any graph on n vertices admits a cut sparsifier with $O(n \log n \cdot \varepsilon^{-2})$ edges and error $\varepsilon > 0$. These constructions all use *importance sampling*: assign an importance to each edge, sample edges independently proportional to these importances, and repeat until the sparsifier reaches the desired size. This procedure also plays a central role in many extensions of Problem 1.3 [Lee23, JLLS23, JLLS24].

As an illustration, consider the case where G is the complete graph.[5] By symmetry, all edges are equally important, so the procedure reduces to sampling independently $m = O(n \log n \cdot \varepsilon^{-2})$ edges and assigning uniform weights so as to preserve the total edge mass. A Chernoff bound implies that each subset of $k \ll n$ vertices has cut value in the sparsifier within $1 \pm \varepsilon$ of its value in G with probability at least $1 - e^{-\Omega(\varepsilon^2 mk/n)}$. A union bound over all $\binom{n}{k}$ subsets succeeds precisely once $m = \Omega(n \log n \cdot \varepsilon^{-2})$. The $\log n$ factor is tight: random subgraphs of the complete graph with $\ll n \log n$ edges are not connected.

[5] In this case, a cut sparsifier is equivalent to a combinatorial expander, except that we allow edge weights.

Yet independent sampling fails to find the best possible sparsifiers. Already in our toy example, while Erdős-Renyi random graphs require average degree $\Omega(\log n)$ to expand, there exist constant-degree expanders. In a breakthrough work, Batson, Spielman, and Srivastava [BSS14] generalized this observation by showing that any graph admits a sparsifier with $O(n \cdot \varepsilon^{-2})$ edges. Their algorithm is deterministic and based on convex optimization. We will explore variants in Part III of this thesis.

From sparsification to discrepancy

More broadly, sparsification is abstracted as follows: given a collection $(\boldsymbol{A}_t)_{t\in\mathcal{T}}$ of vectors or matrices, find a subset $\mathcal{S} \subseteq \mathcal{T}$ and weights $(w_s)_{\in\mathcal{S}}$ such that $|\mathcal{S}| \ll |\mathcal{T}|$ and $\sum_{s\in\mathcal{S}} w_s \boldsymbol{A}_s \approx \sum_{t\in\mathcal{T}} \boldsymbol{A}_t$. In graph sparsification, the $\boldsymbol{A}_t$ are the normalized Laplacian matrices of the edges, and $\approx$ means preserving quadratic forms over Boolean test vectors.

This formulation reveals a natural link to *discrepancy theory*. There, the goal is rather to find a *coloring* $\boldsymbol{x}\colon \mathcal{T} \to \{-1, 1\}$ with low *discrepancy* $\|\sum_{t\in\mathcal{T}} x_t \boldsymbol{A}_t\|$, measured under some norm $\|\cdot\|$. A natural sparsification strategy is to compute such a low-discrepancy coloring, and retain only the elements assigned the minority color. The discrepancy guarantee ensures the resulting sum approximates the original sum, but its support has shrunk by a factor of 2. Iterating this process yields a sparsifier of the desired size.[6] This reduction underlies several recent advances in sparsification [RR20, JRT24, RR22, BRR23, LWZ25].

Spencer's theorem

Erdős asked the following question about the discrepancy of bounded vectors:

Problem 1.4 (ℓ_∞-vector balancing)**.** Define the *discrepancy* of a matrix $\boldsymbol{A}$ as

$$\operatorname{disc}(\boldsymbol{A}) \coloneqq \min_{\boldsymbol{x}\in\{-1,1\}^n} \|\boldsymbol{A}\boldsymbol{x}\|_\infty \,. \tag{1.6}$$

What is the largest discrepancy of a $d \times n$ matrix with entries in $[-1, 1]$?

An illustrative special case of Problem 1.4 is $\boldsymbol{A} \in \{0, 1\}^{d\times n}$, interpreted as the incidence matrix of d subsets of $\{1, \ldots, n\}$. Every $\boldsymbol{x} \in \{-1, 1\}^n$ is a *coloring* of the elements with two colors. (1.6) asks for a coloring that minimizes the maximal color imbalance over all sets.

A standard upper bound comes from the probabilistic method: if $\boldsymbol{x} \sim \{-1, 1\}^n$ is uniformly random, each coordinate of $\boldsymbol{A}\boldsymbol{x}$ is approximately Gaussian with standard deviation $O(\sqrt{n})$. A Chernoff bound followed by a union bound over all d coordinates shows that $\boldsymbol{A}$ has a coloring of discrepancy $O(\sqrt{n \log d})$.

[6] When it succeeds, this reduction produces *unweighted* sparsifiers. In some settings, such sparsifiers may not even exist. However, this can be relaxed using partial colorings $\boldsymbol{x} \in [-1, 1]^{\mathcal{T}}$ which have a constant fraction of ± 1 coordinates.

Once again, this argument fails to give the sharp answer to Problem 1.4. This gap is especially striking in the case $d = n$: the best known lower bounds to Problem 1.4 are random (or random-like) instances. They have discrepancy $(1 + o(1))\sqrt{n}$ with high probability, yet the probabilistic method yields an extra $\sqrt{\log n}$ factor. In a seminal work, Spencer [Spe85] proved that every $\boldsymbol{A} \in [-1, 1]^{n \times n}$ has a coloring with discrepancy at most $6\sqrt{n}$. This result gave the paper its memorable title: *Six standard deviations suffice*.

The failure of the probabilistic method suggests that low-discrepancy colorings are rare and so perhaps also hard to find algorithmically. Bansal [Ban10] refuted this intuition by giving a polynomial-time algorithm achieving discrepancy $O(\sqrt{n})$ for $d = n$. This breakthrough opened a new algorithmic perspective on discrepancy theory, leading to many further works [LM15, ES18, Rot17, LRR17, BDG19, BDGL19, BM20, BLV22].

Outstanding questions

Despite this progress, many variants of Problem 1.3 and Problem 1.4 remain open:

1. *Matrix Spencer problem*: Given symmetric matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{R}^{n \times n}$ of spectral norm at most 1, does there always exist a coloring $\boldsymbol{x} \in \{-1, 1\}^n$ such that $\sum_i x_i \boldsymbol{A}_i$ has spectral norm $O(\sqrt{n})$? This is the non-commutative analog of Problem 1.4, with connections to quantum information theory [HRS22, BJM23]. Despite significant interest, only special cases have been resolved [LRR17, HRS22, DJR22, BJM23].
2. *Komlós (and Beck–Fiala) conjecture*: It states that if $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ has columns of ℓ_2-norm at most 1, then it has discrepancy $O(1)$. It is a sparse or multiscale analog of Problem 1.4. The best-known bound is $O(\sqrt{\log n})$ [Ban98, BDG19, BG17, BDGL19].
3. Tight constants in Problem 1.3 and Problem 1.4: A lower bound on the number of edges needed to sparsify any graph is provided by the minimal number of edges an expander must have (as given by Alon–Boppana bound). The best construction uses twice as many edges [BSS14]. Similarly, the constant 6 in Spencer's original result has only been slightly improved [Bel13].
4. *Kadison–Singer problem*: Can we find *unweighted* sparsifiers of unweighted graphs when all edges have equal importance? Such sparsifiers are known to exist [MSS15], but no efficient algorithm is known for finding them.
5. More broadly, extra $\sqrt{\log d}$ factors appear throughout computer science as artifacts of applying Chernoff or matrix Chernoff bounds. Removing these is a major bottleneck: for instance, through the analysis of *Kikuchi matrices*, this would have strong consequences in extremal combinatorics [HKM23] and coding theory [AGKM23].

In this thesis, we develop a new approach to Problem 1.4 that not only improves existing bounds, but also sheds light on several of these outstanding questions.

1.2. Algorithms beyond the union bound

In recent years, tools from *continuous optimization* have begun to challenge the classical union bound argument. A unifying principle underlies these advances: many extremal problems reduce to minimizing a low-degree polynomial function (Problem 1.1 and Problem 1.2) or the norm of a linear function (Problem 1.3 and Problem 1.4).

1.2.1. Polynomial optimization

We begin with a fundamental problem capturing Problem 1.1, Problem 1.2, and many more: *polynomial optimization*. Let p be a degree-3 polynomial in n variables:

$$p(\boldsymbol{x}) = a + \sum_{i=1}^{n} b_i x_i + \sum_{i,j=1}^{n} c_{i,j} x_i x_j + \sum_{i,j,k=1}^{n} d_{i,j,k} x_i x_j x_k \,.$$

More generally, we will consider polynomials of *constant* degree, while the number of variables n goes to infinity. The quadratic and cubic cases capture most phenomena of interest; higher (but constant) degrees typically follow similarly.

Our goal is to optimize p over a domain $\Omega \subseteq \mathbb{R}^n$:

$$\max_{\boldsymbol{x} \in \Omega} p(\boldsymbol{x}) \,. \tag{1.7}$$

(1.7) is a non-convex optimization problem. While Ω can encode a variety of constraints, we will focus on two canonical settings: the *unit sphere* $\Omega = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$ and the *hypercube* $\Omega = \{-1, 1\}^n$. The former corresponds to unconstrained optimization, while the latter is a proxy for constrained optimization. We note that the discrete nature of $\{-1, 1\}^n$ will not be a crucial feature for us (for multilinear polynomials, there is no difference between optimizing over $[-1, 1]^n$ or $\{-1, 1\}^n$).

The general formulation (1.7) is remarkably expressive. It captures a wide range of problems across computer science and physics:

- constraint satisfaction problems such as Problem 1.1, solving sparse linear equations over $\mathbb{F}_2$ (Max-k-Xor), or finding the maximum cut in a graph (Max-Cut);
- finding ground states in spin glass models, such as the Sherrington–Kirkpatrick model (Problem 1.2) for degree-2, or pure 3-spin models for degree-3;
- computing the injective norm of a tensor [BBH+12];
- finding a planted clique in a random graph [FK08];
- finding optimal Lyapunov certificates of stability in control theory [Par00];
- the best separable state problem [BKS17] and the QMA(2) vs EXP conjecture in quantum information [AIM14].

However, most prior work on polynomial optimization has focused (often exclusively) on degree-2 polynomials. This reflects a fundamental barrier: the techniques used to study quadratic polynomials often break down entirely when moving to higher degrees.

1.2.2. Linear algebra for tensors

A guiding intuition is that the coefficients of a homogeneous quadratic polynomial form a matrix, while those of a cubic polynomial form a tensor. This effectively limits the applicability of standard linear algebraic tools (spectrum, rank, trace powers, etc.) to higher-degree polynomials. While many attempts have been made to generalize these notions to tensors, the resulting theory remains limited.

To illustrate, consider an Erdős-Renyi random graph where each edge appears independently with probability p. Define the centered adjacency matrix $\boldsymbol{A}$ by $A_{ij} = 1 - p$ if there is an edge $\{i, j\}$ and $A_{ij} = -p$ otherwise. The spectral norm of $\boldsymbol{A}$, i.e., $\|\boldsymbol{A}\|_2 := \max_{\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=1} \sum_{i,j=1}^{n} A_{ij}x_iy_j$, is a widely used *spectral certificate*. For example, it can be used to upper bound the maximum cut or the maximum independent set of random graphs [Tre17a]. The analysis of this certificate relies on a textbook bound of the form:

Lemma 1.5 ([FK81, FKS89, FO05]). *If $np = \Omega(\log n)$, then $\|\boldsymbol{A}\|_2 = O(\sqrt{np})$.*

However, the typical approach to prove Lemma 1.5 uses the trace method, a technique with no natural analog in the tensor setting. This is a recurring theme for many problems involving random graphs. As a result, these basic questions remain poorly understood in the analogous random hypergraph models.

Can we extend Lemma 1.5 to tensors, and if so, how?

We answer this question in the affirmative by introducing a new approach that bypasses the trace method and yields stronger results than those obtainable via a simple union bound. In the long term, we hope that such techniques will contribute to the development of a spectral theory for random hypergraphs with power and versatility comparable to that of the well-established spectral theory of random graphs.

1.2.3. Approximation algorithms

The increase in complexity is also reflected algorithmically: while optimizing arbitrary degree-2 polynomials over the unit sphere is tractable, the degree-3 case is NP-hard. Here are hard instances for this problem, coming from the reduction of Nesterov [Nes03]:

Example 1.6 (Maximum clique as a degree-3 polynomial)**.** Given any graph $G = (V, E)$, define a cubic polynomial over variables $\boldsymbol{x}$ indexed by $V \cup E$ as follows:

$$f(\boldsymbol{x}) = \sum_{e=\{u,v\}\in E} x_u x_v x_e \,.$$

Maximizers of f over the unit sphere of $\mathbb{R}^{V\cup E}$ correspond to maximum cliques in G [MS65], which are NP-hard to find.

This reduction actually reveals a deeper phenomenon: when moving from degree-2 to degree-3, the spherical maximization problem becomes hard to solve even *approximately*.

Definition 1.7. An algorithm achieves approximation ratio $\alpha \geqslant 1$ if it always outputs a solution with value at least $1/\alpha$ times the optimum.

Assuming the exponential time hypothesis, optimizing n-variate cubic polynomials over the sphere within a $\exp(\log^{\frac{1}{2}-\varepsilon} n)$-factor takes exponential time for any $\varepsilon > 0$ [BBH+12]. It is plausible that the truth is even worse: the best-known polynomial-time algorithms achieve only $\widetilde{O}(\sqrt{n})$-approximation [HLZ10, HKPT24]. These algorithms rely on the same technique of rounding semidefinite programming relaxations of the problem. Stronger lower bounds are known for this approach: they cannot achieve approximation ratio better than $\widetilde{O}(n^{1/4})$, even in subexponential time [HSS15, PR22].[7] This barrier is qualitatively different from the NP-hardness in Example 1.6: it arises even for *random* instances.

1.2.4. Certifiable approximation

A subtle issue arises with approximation algorithms in the presence of randomness. In such cases, we typically ask that the approximation guarantee holds with high probability. Ideally, an algorithm should also not only perform well on average — it should *know* when it does. This leads to the notion of *certifiability*: an approximation algorithm has certifiable guarantees if it either returns a solution with a provable guarantee on its quality, or outputs "I don't know". For maximization problems such as (1.7), such a certificate often takes the form of an efficiently computable upper bound on the optimum — one that (1) always holds, and (2) is, with high probability, close to the true optimum.

Certifiability is desirable (both in theory and in practice) since it connects naturally to *robustness*. For instance, an algorithm that certifies an upper bound on the maximum of a random polynomial can be repurposed to recover a low-rank signal planted in its coefficients [HSS15]. However, this flexibility comes at a cost: certifiable approximation is sometimes much harder than the underlying search problem.

[7] These lower bounds implicitly only formally apply to the *canonical* relaxation. As we will see, some algorithms with slightly better guarantees use alternative formulations.

This gap is especially visible in Problem 1.2. On the one hand, evidence from [MRX20, KB21, GJJ+20] suggests that the best efficient certifiable approximation is given by computing the optimum of the spherical relaxation (1.4), followed by rounding the solution to a Boolean vector. On the other hand, algorithms that do not produce certificates can get arbitrarily close to the optimum of (1.3). This suggests a fundamental "certification–optimization" gap: a difference between what we can find, and what we can certify [Kun21].

Remarkably, a similar phenomenon appears even on deterministic input. For maximizing arbitrary cubic polynomials over the hypercube, the only known approximation algorithm, due to Khot and Naor [KN08], is randomized and does not provide any certificate.

Is there a certification–optimization gap for the guarantees achieved by [KN08]?

We show that the answer is no, by introducing an approximation algorithm with certifiable guarantees matching the result of Khot and Naor. Moreover, we show that certifiable guarantees can guide the design of new algorithms, including our result for Problem 1.1. Our approach is based on *convex relaxations* of polynomial optimization.

1.2.5. Roundings of convex relaxations

Rather than solving directly (1.7), we solve a relaxed convex problem, whose optimal dual solution provides a certifiable upper bound on the optimum. The main challenge lies in turning this certificate into a feasible solution for the original problem. However, rounding algorithms beyond the quadratic case remain poorly understood.

The sum-of-squares hierarchy

The *sum-of-squares* (SoS) hierarchy of convex relaxations is a general framework for designing approximation algorithms, consisting of two steps:

1. Optimize the objective over a *relaxed* set of solutions with the *ellipsoid method*.
2. *Round* the optimal relaxed solution to a feasible solution for the original problem.

Under the Unique Games Conjecture, this strategy achieves the best-possible approximation factor for any constraint satisfaction problem [Rag08].[8]

SoS relaxations operate on objects called *pseudo-expectations*. A degree-d pseudo-expectation is a linear map $\widetilde{\mathbb{E}}\colon \mathbb{R}[x_1, \dots, x_n]_{\leqslant d} \to \mathbb{R}$ that mimics the behavior of an expectation under some distribution, but only for polynomials of degree at most d. It satisfies self-consistency conditions such as $\widetilde{\mathbb{E}}\, p(\boldsymbol{x})^2 \geqslant 0$ for all polynomial p of degree at most $d/2$, which holds trivially for actual expectations.

[8] This result holds up to an arbitrary small constant additive error in the approximation guarantees. In particular, it is vacuous for the "advantage over a random assignment" regime we target for Max-3-Sat.

The set $\mathcal{P}_d(\Omega)$ of degree-d pseudo-expectations over domains Ω such as the sphere or the hypercube is convex and admits an efficient separation oracle. Thus, the degree-d SoS relaxation of (1.7):

$$\max_{\widetilde{\mathbb{E}} \in \mathcal{P}_d(\Omega)} \widetilde{\mathbb{E}}\, p(\boldsymbol{x})$$

is a convex program that can be solved (to high precision) in time $n^{O(d)}$. Note that the degree of the relaxation must be at least as large as the degree of p.

Roundings for quadratic optimization

As with linear programming, the main challenge lies in the *rounding* step: Given an SoS solution $\widetilde{\mathbb{E}} \in \mathcal{P}_d(\Omega)$, we seek an efficiently samplable distribution $\mathcal{D}$ supported on Ω such that $\mathbb{E}_{\overline{\boldsymbol{x}} \sim \mathcal{D}}\, p(\overline{\boldsymbol{x}})$ is not too small compared to $\widetilde{\mathbb{E}}\, p(\boldsymbol{x})$.

In the quadratic case, there is a generic rounding procedure. Since $\widetilde{\mathbb{E}}$ satisfies the constraint $\widetilde{\mathbb{E}} \langle \boldsymbol{x} - \widetilde{\mathbb{E}}\, \boldsymbol{x}, \boldsymbol{u} \rangle^2 \geqslant 0$ in any direction $\boldsymbol{u}$, we can sample a Gaussian vector

$$\boldsymbol{g} \sim \mathcal{N}\left(\widetilde{\mathbb{E}}\, \boldsymbol{x}, (\boldsymbol{x} - \widetilde{\mathbb{E}}\, \boldsymbol{x})(\boldsymbol{x} - \widetilde{\mathbb{E}}\, \boldsymbol{x})^\top\right),$$

whose first and second moments perfectly match those of the SoS solution. Rounding reduces to the task of mapping $\boldsymbol{g}$ to a feasible point $\overline{\boldsymbol{x}} \in \Omega$, e.g. by taking the sign [GW95] or truncating and rescaling [CW04]. Some examples of the many algorithmic applications of this simple rounding include approximation algorithms for 2-variable constraint satisfaction problems, cut norms of matrices [AN06], and correlation clustering [CW04]. See §8.1.2 for additional background.

Roundings for higher-degree optimization

In sharp contrast, there are very few known rounding techniques for higher-degree sum-of-squares relaxations. One notable exception is *global correlation rounding* [BRS11], which proceeds by repeatedly sampling variables from their marginal distribution under the pseudo-expectation, and conditioning on the outcome. It has been successfully applied to quadratic optimization problems such as Max-Cut and generalizations [BRS11, RT12], list decoding of error-correcting codes [JST23], expansion [GS11], and graph coloring [AG11]. All these applications rely on structural assumptions about the coefficients ensuring that each conditioning step reduces the variance of the SoS solution. Such assumptions do not hold for the general class of problems considered in this thesis.

In fact, even the simple problem of rounding a convex relaxation of cubic optimization is not well understood. The only general result is the rounding scheme of [BGG+17] of (a variant of) polynomial optimization achieving $O(\sqrt{n})$-approximation[9] over the unit

[9] [BGG+17] considers the variant where one wants to maximize the *absolute value* of a polynomial and this distinction makes a material difference to the difficulty of the problem.

sphere. Over the hypercube, no known rounding algorithm (for any convex relaxation) achieves non-trivial guarantees. The only approximation result for cubic optimization over the hypercube is [KN08], that circumvents convex relaxations and uses instead anti-concentration and techniques from convex geometry.

Understanding rounding for higher-degree polynomial optimization is tightly connected to major open problems. We already mentioned that Problem 1.1 is a cubic optimization problem, whose approximability remains open. Further, sufficiently strong (and yet far from what known hardness results and integrality gaps rule out) approximation algorithms for special cases of degree-3 and degree-4 polynomial optimization can refute the Small-Set Expansion [BBH+12] hypothesis, settle the Aaronson–Impagliazzo–Moshkovitz [AIM14] conjecture that relates to the power of quantum entanglement, and refute the celebrated planted clique hypothesis [FK08].

In short, the lack of rounding algorithms for high-degree SoS relaxations is a central roadblock in our understanding of polynomial optimization beyond the quadratic case. Developing new tools to overcome this barrier is one of the goals of this thesis.

1.3. Detailed overview of our contributions

This thesis offers new algorithmic perspectives on optimizing: random quadratic polynomials (§1.3.1), random sparse cubic polynomials (§1.3.2), arbitrary cubic polynomials (§1.3.3), and discrepancy objectives (§1.3.4).

1.3.1. Fourier analysis of random quadratic optimization

Our first contribution is a new framework for analyzing algorithms that optimize random quadratic polynomials. The approach is built on a basis of *Fourier diagrams* that captures the asymptotic behavior of symmetric, low-degree iterative algorithms.

Indeed, recent years have seen the emergence of new algorithms for random polynomial optimization problems. While the power method and its variants can solve (1.4), Montanari [Mon19] proposed a polynomial-time algorithm that, assuming a widely believed statistical physics conjecture, finds a maximizer of (1.3) to arbitrary constant accuracy as $n \to \infty$. On a high level, Montanari's algorithm is a nonlinear generalization of the power method, where the nonlinearities are used to enforce the Boolean constraints.

However, both the algorithm and its analysis are relatively difficult to motivate and describe at this point. To better understand the mechanisms behind this algorithm and its potential for proving lower bounds in random optimization, we introduce a new framework in Part I to analyze nonlinear iterations.

Nonlinear iterative algorithms on random matrices

We consider algorithms taking as input a quadratic polynomial $p(\boldsymbol{x}) = \sum_{i,j=1}^{n} A_{ij} x_i x_j$ whose coefficients A_{ij} are symmetric ($A_{ij} = A_{ji}$), but are otherwise independent mean-0, variance-$\frac{1}{n}$ random variables. These algorithms output a vector $\boldsymbol{x} \in \mathbb{R}^n$, which for example may be a feasible solution to (1.3) or (1.4). Our goal is to analyze them, e.g., by computing the objective value that they achieve.

The best-known algorithms for such tasks, in null models [Sub20, Mon19, EM20, EMS21, Sel24] as well as in models with a planted signal [RF12, MR16] belong to the class of *generalized first-order methods* [CMW20]. These algorithms start from an arbitrary initialization $\boldsymbol{x}_0 = (1, \dots, 1)$. At each step, they either (1) multiply the current state by $\boldsymbol{A}$ ($\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t$); or (2) apply a polynomial function coordinate-wise ($\boldsymbol{x}_{t+1} = f_t(\boldsymbol{x}_t)$, for some constant-degree polynomial $f_t : \mathbb{R} \to \mathbb{R}$). The algorithm outputs $\boldsymbol{x}_T$ after a constant number T of iterations.[10] These nonlinear generalizations of the power method have two key features:

1. They are $\mathfrak{S}_n$-symmetric: swapping the indices of two rows and columns of $\boldsymbol{A}$ results in a corresponding swap in the output vector.[11]
2. Each output is a low-degree polynomial function of the entries of $\boldsymbol{A}$. Low-degree polynomials are an expressive class of algorithms [Wei25].

Our toolkit provides several new insights into the behavior and algorithms with such features. First, it provides an explicit description of the joint distribution of their iterates as $n \to \infty$, leading to a simple analysis of why they succeed. Such compact descriptions were previously known only for a more restricted class of algorithms. Second, in contrast to all previous mathematically rigorous approaches, it closely mirrors the analysis of nonlinear iterative algorithms in physics (via the *cavity method*). In particular, it offers a simple combinatorial interpretation of concepts that physicists used to guide the design of their impressive algorithms, such as approximate message passing and Onsager correction terms. This combinatorial perspective suggests principles for designing new algorithms, and in turn lower bounds on the optima of random maximization problems.

The Fourier diagram basis

We introduce the *Fourier diagram basis*, an orthogonal basis of $\mathfrak{S}_n$-symmetric polynomials for expressing nonlinear iterative algorithms. Each Fourier diagram is represented by a

[10] This in fact is not sufficient for the planted models mentioned earlier, that require $O(\log n)$ iterations to obtain non-trivial correlation with the signal from random initialization; see §5.5.

[11] Why this is without loss of generality may be explained by the fact the random matrix distribution satisfies the same symmetry; see [KMW24, §3.4] for related results.

rooted graph $\alpha = (V(\alpha), E(\alpha))$ and expresses a vector-valued polynomial $\boldsymbol{Z}_\alpha(\boldsymbol{A}) \in \mathbb{R}^n$:

$$Z_\alpha(\boldsymbol{A})_i := \sum_{\substack{\varphi\colon V(\alpha)\hookrightarrow[n] \\ \varphi(\circledcirc)=i}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}\,, \qquad \text{for all } i \in [n]\,, \tag{1.8}$$

where $\circledcirc \in V(\alpha)$ is the root vertex. The first few basis elements are depicted on Figure 1.1.

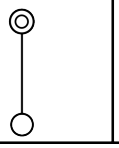
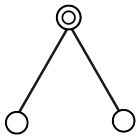
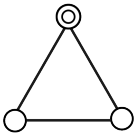
$Z_\alpha(\boldsymbol{A})_i$	1	$\sum_{\substack{j \\ j\neq i}} A_{ij}$	$\sum_{\substack{j,k \\ k\neq j\neq i}} A_{ij}A_{ik}$	$\sum_{\substack{j,k \\ k\neq j\neq i}} A_{ij}A_{ik}A_{jk}$
α	◎			

Figure 1.1. Some Fourier diagrams with the symmetric polynomial that they express.

We emphasize that we sum over *injective* maps $\varphi\colon V(\alpha) \hookrightarrow [n]$ in (1.8). This is a crucial point for all the following results and the key novelty from this work; see §2.5 for a description of the basis obtained by summing over arbitrary φ.

Chapter 2 shows that the iterates of nonlinear iterative algorithms can be expanded as linear combinations of Fourier diagrams (1.8). Thus, to understand the (asymptotic) behavior of such algorithms, it suffices to describe the (asymptotic) joint distribution of the basis elements. This is the purpose of our first main result, established in Chapter 3. The joint distribution of the Fourier diagrams can be read directly from their graph structure:

Theorem 1.8 (Informal version of Theorem 3.14)**.** *The joint distribution of* $\{Z_\alpha(\boldsymbol{A})\}$ *as* $n \to \infty$ *is given by:*

1. $\{\boldsymbol{Z}_\tau(\boldsymbol{A}) : \tau$ *tree whose root has degree 1*$\}$ *are asymptotically independent Gaussian vectors with independent coordinates. We refer to these diagrams as* Gaussian tree diagrams.
2. *For all tree* τ, $\boldsymbol{Z}_\tau(\boldsymbol{A})$ *is asymptotically a multivariate Hermite polynomial in the Gaussian tree diagrams.*
3. *If* α *is connected and has a cycle, then* $\boldsymbol{Z}_\alpha(\boldsymbol{A})$ *is asymptotically negligible (i.e., its entries are on a scale at least* $\sqrt{n}$ *smaller than the tree diagrams).*

For example, in Figure 1.1, asymptotically: the second diagram is a Gaussian vector, the third diagram is a Hermite polynomial in this vector (in fact, the degree-2 Hermite polynomial), and the fourth diagram is negligible.

The key takeaway from Theorem 1.8 is:

Only the restriction on tree diagrams matters.

This fact has a simple combinatorial intuition. Suppose that $\boldsymbol{A}$ is normalized to have entries of variance $\frac{1}{n}$. Then (1.8) is a sum over $n^{|V(\alpha)|-1}$ terms, each of which being of magnitude $n^{-|E(\alpha)|/2}$. If we heuristically treat these terms as independent, cancellations make (1.8) of order $n^{\frac{1}{2}(|V(\alpha)|-|E(\alpha)|-1)}$, which remains constant precisely when α is a tree. Every additional edge makes the diagram smaller by a factor $\sqrt{n}$.

What insight do we gain by focusing on tree diagrams? The key advantage is that algorithmic operations in generalized first-order methods have simple combinatorial interpretations when expressed in the tree diagram basis. First, as Theorem 1.8 suggests, applying a Hermite polynomial entrywise to a Gaussian tree diagram asymptotically produces another tree diagram. More generally, a nonlinear polynomial can be applied to a diagram by expanding it in the Hermite basis and acting term by term.

Nonlinearities cannot increase the depth of a tree diagram. Only applying one iteration of the power method can do so:

Theorem 1.9 (See §3.5)**.** *Let τ be a Fourier tree diagram. Then $\boldsymbol{A} \cdot \boldsymbol{Z}_\tau(\boldsymbol{A})$ is asymptotically the sum of $\boldsymbol{Z}_{\tau^+}(\boldsymbol{A})$ and $\boldsymbol{Z}_{\tau^-}(\boldsymbol{A})$, where τ^+ (resp. τ^-) is obtained by extending (resp. contracting) the root of τ by one edge.*

Figure 1.2 summarizes how Fourier tree diagrams evolve under algorithmic operations.

(a) Applying a Hermite polynomial entrywise to a Gaussian tree diagram grafts copies of the tree at the root.

(b) Multiplying by $\boldsymbol{A}$ creates a forward and a backward term.

Figure 1.2. The effect of algorithmic operations on the Fourier tree representation.

Theorem 1.9 sheds light on a special class of generalized first-order methods: *approximate message passing* (AMP) algorithms. For example, Montanari's algorithm for Problem 1.2 belongs to this class. Originally developed in physics, AMP algorithms are designed to produce *Gaussian* iterates by subtracting a carefully chosen correction term (the *Onsager correction*) at each iteration. While the mathematical expression of this correction is often viewed as intricate, we show in Chapter 4 that it exactly cancels the τ^- term in Theorem 1.9. The remaining τ^+ term is a Gaussian tree diagram by Theorem 1.8, which explains why AMP iterates remain Gaussian.

The cavity method and free probability

Taken together, Theorem 1.8 and Theorem 1.9 yield a simple two-step procedure for analyzing the performance of generalized first-order methods:

1. Construct the *tree approximation* $(\widehat{\boldsymbol{x}}_t)_{t\geqslant 0}$ of the original iterates $(\boldsymbol{x}_t)_{t\geqslant 0}$ by applying the algorithmic operations only to tree diagrams. We show that $\|\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_t\|_\infty \lesssim n^{-1/2}$ (Theorem 3.17).
2. Compute the objective value achieved by $\widehat{\boldsymbol{x}}_t$. For quadratic objective functions, it converges to $\langle \widehat{\boldsymbol{x}}_t, \widehat{\boldsymbol{x}}_t^+ + \widehat{\boldsymbol{x}}_t^- \rangle$ by Theorem 1.9.

A key insight is that the combinatorics of the tree approximation mirrors the independence assumptions of the non-rigorous *cavity method* from physics (see §4.1 for background on the cavity method). We demonstrate the connection in Chapter 4 by making multiple heuristic physics arguments analyzing message-passing algorithms fully rigorous:

- equivalence between *belief propagation* and AMP (Theorem 4.1);
- independence of messages incoming at a vertex (Theorem 4.9);
- the *state evolution* formula (Theorem 4.10).

While these results have appeared in previous work, existing proofs (such as those based on Bolthausen's conditioning method [Bol14]) are technically involved. In contrast, our approach follows exactly the physics derivation, and formalizes every heuristic equality by showing that the equality actually holds rigorously up to a sum of cyclic Fourier diagrams, which are negligible by Theorem 1.8.

Working directly in the tree basis also allows us to characterize the *optimal* symmetric, constant-degree algorithm for constrained random quadratic optimization, yielding the following lower bound on the optima of both (1.3) and (1.4):

Theorem 1.10 (see Theorem 6.9)**.** *Let $\mathcal{T}$ be a collection of one-dimensional random variables indexed by rooted trees, whose distributions match the asymptotic distribution of a single coordinate of the tree diagrams from Theorem 1.8. Then for any integer $p \geqslant 2$,*

$$\max_{\|\boldsymbol{x}\|_p \leqslant 1} \sum_{i,j=1}^{n} A_{ij} x_i x_j \geqslant (2 - o(1)) \cdot n^{1-\frac{2}{p}} \cdot \max_{\substack{Z \in \operatorname{span}(\mathcal{T}) \\ \mathbb{E}\, Z^p \leqslant 1}} \mathbb{E}\left[ZZ^+\right] . \tag{1.9}$$

We illustrate Theorem 1.10 by recovering a tight lower bound for the spherical model in Chapter 6. Although similar algorithmic lower bounds have been explored for planted models [MR16] and spin glass models [EMS21], the key novelty is that the optimization problem on the right-hand side of (1.9) has a simple combinatorial interpretation. On the one hand, the objective $\mathbb{E}\,[ZZ^+]$ encourages mass to be spread across many translated Gaussian diagrams; on the other hand, the constraint $\mathbb{E}\, Z^p \leqslant 1$ can only be satisfied if each Gaussian is offset by Hermite tree diagrams depending on it. Describing the optimizer in the limit $p \to \infty$ is an exciting open problem raised in this thesis.

Finally, our framework also connects naturally with *free probability*. Voiculescu [Voi91] pioneered the analysis of the spectrum of large random matrices by relating them to their idealized infinite-dimensional version living in the so-called *Fock space*. The path diagrams in the set $\mathcal{T}$ from Theorem 1.10 form an explicit basis for this Fock space.[12] These path diagrams suffice to encode spectral information because the power method involves no nonlinearities, and Theorem 1.9 preserves path diagrams. The remaining tree diagrams in $\mathcal{T}$ enrich the space to encode constrained optimization, playing a role analogous to that of free probability in the unconstrained (ℓ_2) case. This connection underlies a recent generalization of the Fourier diagram method to deterministic delocalized matrix ensembles, via the theory of traffic probability [GJKP25].

The long-time behavior

The analysis so far applies to algorithms that run for a constant number of steps. This suffices for Problem 1.2, where constant-factor approximation is conjectured to be achieved after constantly many iterations. By contrast, in statistical estimation problems (e.g. the spiked Wigner model; see Example 2.3), $\Omega(\log n)$ iterations are required to extract a planted signal starting from a random initialization. This presents a major challenge for existing analyses of iterative algorithms, which either make unrealistic assumptions such as assuming access to a warm start, or rely on complex and problem-specific arguments [RV18, MV21, MV22, LW22, LFW23].

Theorem 1.8 does not extend either to this long-time regime in general. The idea is that nonlinear algorithms can exploit high moments of the input, breaking universality (see §5.4). However, in Chapter 5, we show that the tree approximation remains valid for an iterative algorithm approximating the top eigenvector of a random matrix. These results suggest a path toward a mathematical analysis of a ubiquitous but in many cases non-rigorous statistical physics algorithm: message-passing from random or spectral initialization.

1.3.2. Multiscale union bound for random hypergraphs

We next switch from dense to *sparse* random polynomials, and describe our extensions of the results of §1.2.2 to cubic polynomials. We develop a multiscale union bound argument that generalizes the guarantees of the trace method for Lemma 1.5 to tensors.

Friedman and Wigderson [FW95] defined the second eigenvalue of hypergraphs or tensors, generalizing the second eigenvalue of the adjacency matrix of a graph:

Definition 1.11 (Friedman–Wigderson second eigenvalue)**.** Let $\boldsymbol{T}$ be adjacency tensor of an Erdős-Renyi 3-uniform hypergraph, i.e., $T_{ijk} = 1$ with probability p and $T_{ijk} = 0$ with

[12]The author thanks Tim Kunisky and Robert Wang for discussions leading to this observation.

probability $1-p$, independently for all triplets of distinct vertices $\{i,j,k\}$. Let

$$\|\boldsymbol{T}\| := \max_{\|\boldsymbol{x}\|_2=\|\boldsymbol{y}\|_2=\|\boldsymbol{z}\|_2=1} \sum_{i,j,k=1}^{n} T_{ijk}x_iy_jz_k\,, \quad \lambda_{\mathrm{FW}}(\boldsymbol{T}) := \|\boldsymbol{T}-\mathbb{E}\,\boldsymbol{T}\|\,.$$

[FW95] asked when random hypergraphs exhibit a spectral gap, in the sense that $\lambda_{\mathrm{FW}}(\boldsymbol{T})$ is bounded away from $\|\boldsymbol{T}\|$. We refer to such hypergraphs as *quasi-random*, in analogy with the converse of the expander mixing lemma [BL06]. [FW95] showed that *random regular* 3-uniform hypergraphs become quasirandom once the number of hyperedges is $\Omega(n^2)$.

For Erdős-Renyi hypergraphs, the threshold as a function of the average number of edges $m := p\binom{n}{3}$, appears much smaller. This can be shown by *flattening* the tensor into an $n^2 \times n^2$ matrix, and bounding its norm with the spectral norm of that matrix. This technique is widely used for designing spectral certificates, e.g. in the context of refutation of constraint satisfaction problems [AOW15]. In our setting, this argument implies that Erdős-Renyi hypergraphs are quasi-random once $m \geqslant n^{1.5}\,\mathrm{polylog}(n)$. Moreover, these extra $\mathrm{polylog}(n)$ factors are necessary when using this certificate.

Chapter 7 establishes:

Theorem 1.12 (Informal version of Theorem 7.3)**.** $\|\boldsymbol{T}\| - \lambda_{\mathrm{FW}}(\boldsymbol{T}) = \Theta(pn^{1.5})$

As a corollary, Erdős-Renyi hypergraphs are quasi-random as soon as $m = \Omega(n^{1.5})$.

How to prove Theorem 1.12? As discussed earlier, the flattening bound incurs superfluous logarithmic factors, and the trace method from Lemma 1.5 does not generalize to tensors. A natural candidate for a sharp answer is (generic) *chaining*, a sophisticated union bound technique known non-constructively to give tight bounds up to constant factors [Tal21]. Prior work has constructed chaining bounds for related problems, but these still incur logarithmic losses [BR16, BGJ+25]. A tight generic chaining construction here is expected to be technical, due to the intricate geometry of the underlying metric space [LvHY18]. This mirrors the long-standing open question of proving general matrix concentration bounds using chaining [Rud99].

Instead, our proof of Theorem 1.12 uses a direct multiscale union bound argument, extending an approach initiated by Kahn and Szemerédi [FKS89]. We believe that this perspective may guide tighter chaining constructions for tensors, with potential applications to problems such as hypergraph sparsification [Lee23].

1.3.3. Roundings beyond quadratic polynomials

In this section, we move from random to worst-case cubic polynomial optimization. We develop new rounding algorithms for this problem over the sphere and the hypercube, improving the best known approximation guarantees.

Our first main result is an improvement over the coin flipping argument of Håstad and Venkatesh [HV04] for Problem 1.1:

Theorem 1.13 (see Theorem 8.29)**.** *There is a polynomial-time algorithm that, given a satisfiable 3-Sat formula, outputs an assignment satisfying a* $\frac{7}{8} + \widetilde{\Omega}(n^{-3/4})$ *fraction of the clauses.*

Known results are summarized on Figure 1.3.

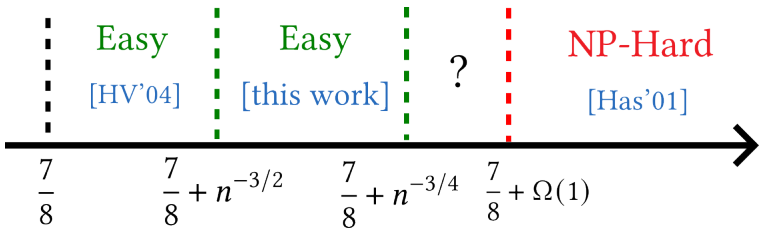


Figure 1.3. Known and new approximation guarantees for satisfiable 3-Sat instances.

The proof of Theorem 1.13 relies on the polynomial formulation from (1.2). Depending on whether the degree-1, 2, or 3 part of the polynomial dominates at an optimal solution, we apply different strategies. When the degree-1 or 2 parts dominate, standard approximation algorithms for quadratic polynomials can be applied. When the degree-3 part dominates, this requires a new rounding algorithm that we present next.

Our second main result is a rounding algorithm for homogeneous cubic optimization:

Theorem 1.14 (Informal; see Theorems 8.14 and 8.22)**.** *For every* $k \geqslant 6$*, there is an* $n^{O(k)}$*-time algorithm that rounds the canonical degree-*k *sum-of-squares relaxation for a homogeneous multilinear cubic maximization problem (over the unit sphere or the hypercube) and achieves an* $O(\sqrt{n/k})$ *approximation.*

This improves on the prior result of [BGG+17], which gave a similar guarantee over the sphere only. Their approach uses "weak decoupling" inequalities and involve reasoning about eigenvectors of the SoS solution. In particular, such techniques seem to have no natural analogs over the Boolean hypercube.

To prove Theorem 1.14, we introduce a new rounding algorithm based on *polynomial reweighting*, an operation on pseudo-expectations analogous to reweighting a probability distribution by a low-degree polynomial.

Our third result shows that there is a different, *compressed*, SoS relaxation of the problem of size $2^{O(k)} \cdot \mathrm{poly}(n)$ (instead of $n^{O(k)}$), preserving the guarantees of Theorem 1.14:

Theorem 1.15 (Informal; see Theorems 8.17 and 8.23)**.** *There is a* $2^{O(k)} n^{O(1)}$*-time algorithm that takes input a homogeneous multilinear cubic polynomial* $f(\boldsymbol{x})$ *in* n *variables and outputs an assignment that achieves an* $O(\sqrt{n/k})$*-approximation to the optimum of* f *over the hy-*

percube or the unit sphere. Moreover, our algorithm is obtained by rounding a semidefinite programming relaxation of the cubic optimization problem.

Over the sphere, Theorem 1.15 improves the approximation guarantees of [BGG+17]. In particular, we obtain an $O(\sqrt{n/\log n})$ approximation algorithm in polynomial time, as opposed to quasi-polynomial time in [BGG+17]. Over the hypercube, our algorithm for Theorem 1.15 is *deterministic* and our guarantees in the polynomial-time regime match the ones of [KN08]. This answers affirmatively the question asked in §1.2.4 about whether the guarantees of [KN08] had any certifiable analog.

The proof of Theorem 1.15 relies on derandomizing a crucial set of inequalities that arise in our rounding algorithm via polynomial reweightings. Unlike prior pruning approaches [GS12, BRS11], which reduce the number of constraints or variables in the relaxation, our construction *adds* a small number of carefully chosen auxiliary variables and constraints. This approach is reminiscent of the degree-reduction method of Steurer and Tiegel [ST21] in robust statistics, though their result exploits problem-specific structure.

1.3.4. Discrepancy minimization via regularization

The iterative algorithms for spin glass models described in §1.3.1 can be interpreted as second-order methods finding stationary points of a regularized version of the polynomial objective, known as the TAP free energy [Sub20, Mon19, JSS25]. In Part III, we reinterpret this idea to design new algorithms for worst-case discrepancy minimization.

To illustrate this approach, consider Problem 1.4. Up to duplicating rows to account for both $\boldsymbol{A}$ and $-\boldsymbol{A}$ (see Remark 9.2), the goal is to minimize the discrepancy objective

$$\boldsymbol{x} \in \{-1,1\}^n \mapsto \max_{i\in[n]} (\boldsymbol{A}\boldsymbol{x})_i = \max_{\boldsymbol{r}\in\Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x}\rangle \ , \tag{1.10}$$

where $\Delta_d \coloneqq \{\boldsymbol{r} \in \mathbb{R}^d_{\geqslant 0} : \sum_{i=1}^d r_i = 1\}$.

Early algorithmic approaches to Spencer's theorem use a *sticky walk* in $[-1,1]^n$ [Ban10, LM15], which maintains a partial coloring $\boldsymbol{x} \in [-1,1]^n$ and freezes each coordinate x_i once it reaches ± 1. However, these algorithms operate in multiple stages to balance discrepancy and freezing constraints, and do not naturally fit into a continuous optimization framework.

This mismatch arises because (1.10) is not well suited to standard continuous optimization. For example, sticky gradient descent on $[-1,1]^n$ reduces to selecting the row $\boldsymbol{a}_j$ of $\boldsymbol{A}$ that maximizes $\left|\langle \boldsymbol{a}_j, \boldsymbol{x}\rangle\right|$ for the current partial coloring $\boldsymbol{x}$, and stepping in the $\pm\boldsymbol{a}_j$ direction. Yet this strategy fails to make progress: starting from the origin, the first gradient step is immediately pulled back, as the gradient points back to zero. Second-order methods appear to offer no help either, as the objective is piecewise linear.

We revisit the continuous optimization perspective and show that, with an appropriate use of *regularization*, it can be made effective for discrepancy objectives. Specifically, we

introduce a regularized version of (1.10) by adding a concave function $\omega\colon \mathbb{R} \to \mathbb{R}$ that encourages the mass of $\boldsymbol{r} \in \Delta_d$ to be spread across coordinates:

$$\Phi_{\boldsymbol{A}}(\boldsymbol{x}) := \max_{\boldsymbol{r}\in\Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x}\rangle + \sum_{j=1}^{d} \omega(r_j)\,. \tag{1.11}$$

Our algorithm is a second-order method: a sticky variant of Newton method applied to $\Phi_{\boldsymbol{A}}$. To prevent the oscillatory behavior observed with gradient descent, we constrain updates to lie in the subspace orthogonal to the current partial coloring $\boldsymbol{x}$.[13]

The behavior of (1.11) depends on the regularizer ω, and we leverage this flexibility to adapt our framework to different settings. For Problem 1.4, we show in Chapter 9 that choosing $\omega(r) = r^p$ for $p \in (0, 1)$ yields an algorithm matching Spencer's guarantees. The special case $p = \frac{1}{2}$ recovers the potential function used in [BSS14] to construct linear-size graph sparsifiers.

One key advantage of the optimization viewpoint is that it enables modular proofs of "best of both worlds" results by summing multiple objectives of the form (1.11). This perspective has been central in algorithmic discrepancy theory, and has been repurposed for sparsification [JRT24] and for rounding linear programming relaxations in combinatorial optimization [Ban19]. It also underlies the applications developed throughout this thesis.

Tighter constant for Spencer's theorem

In Chapter 11, we show — mirroring Spencer's original paper — that in fact, 4.1 *standard deviations suffice.* Moreover, such a coloring can be found efficiently with our second-order optimization algorithm.

Theorem 1.16 (Informal version of Theorem 11.1)**.** *Any $\boldsymbol{A} \in [-1, 1]^{n\times n}$ has discrepancy at most $4.1\sqrt{n}$. Moreover, such a coloring can be found in polynomial time.*

The best known asymptotic lower bound for this constant is 1, attained when $\boldsymbol{A}$ is a random matrix, or a structured matrix with similar properties such as a Hadamard matrix. For small n, numerical simulations suggest the existence of structured matrices whose discrepancy is close to 2.

Motivated by Theorem 1.16, we also attempted to improve the constant in the size–approximation tradeoff for graph sparsifiers. However, we provide evidence in §11.4 that our discrepancy-theoretic framework may not improve this constant. This is because it cannot easily distinguish normalized edge Laplacian matrices from general normalized sums of rank-one matrices, for which we conjecture that such improvements are impossible.

[13] This constraint also appears in algorithms for random polynomial optimization [Sub20].

Random-like instances of Komlós conjecture

In Chapter 10, we prove that Komlós conjecture holds for pseudorandom matrices:

Theorem 1.17 (Informal; see Theorem 10.5 and Theorem 10.6)**.** *Suppose that* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *has columns of* ℓ_2*-norm at most 1. Define*

$$\lambda(\boldsymbol{A}) := \left\|\boldsymbol{A}^{\odot 2}\boldsymbol{\Pi}\right\|_2 = \sup_{\substack{\|\boldsymbol{v}\|_2=1 \\ \sum_{i=1}^n v_i=0}} \|\boldsymbol{A}^{\odot 2}\boldsymbol{v}\|_2\,,$$

where $\boldsymbol{A}^{\odot 2} := (A_{ij}^2)_{i,j\in[n]}$ *and* $\boldsymbol{\Pi}$ *is the projection orthogonal to* $(1,\dots,1)$*.*

Then, $\boldsymbol{A}$ *has a coloring of discrepancy* $O(\sqrt{\lambda(\boldsymbol{A})\cdot\log n})$ *that can be found efficiently.*

To compare this result with prior work, consider the special case where $\boldsymbol{A}$ is a rotation matrix. This case remains open and we believe it is a key instance of Komlós conjecture. In this setting, the top eigenvector of $\boldsymbol{A}^{\odot 2}$ is the all-ones vector with eigenvalue 1, so that $\lambda(\boldsymbol{A})$ coincides with the second eigenvalue of this entrywise squared matrix.

After applying a rotation to the hypercube $\{-1,1\}^n$, every corner lies at ℓ_2-distance $\sqrt{n}$ from the origin. Komlós conjecture asserts that there always exists a corner whose ℓ_∞-distance to the origin is $O(1)$. Theorem 1.17 establishes this conclusion under the assumption that the second largest eigenvalue of $\boldsymbol{A}^{\odot 2}$ is $O(1/\log n)$. We show in §10.3 that this quantity is $O(1/\sqrt{n})$ for some models of random orthogonal matrices. Moreover, it is always at most 1 when $\boldsymbol{A}$ is a rotation matrix, so our bound subsumes the best-known general result for arbitrary instances to Komlós conjecture: Banaszczyk's discrepancy bound of $O(\sqrt{\log n})$ [Ban98, BDG19].[14]

1.4. Roadmap

Part I: The Fourier Diagram Basis

Chapter 2. We introduce the Fourier diagram basis, illustrate it with examples, and analyze how algorithmic operations act on the full Fourier representation. We motivate the term *Fourier*, and conclude by contrasting this basis with the monomial basis used in prior work.

Chapter 3. We state and prove our main theorem characterizing the joint distribution of Fourier diagrams. We then derive the effect of algorithmic operations on the asymptotic Fourier tree basis. Finally, we establish a general state evolution result applicable to nonlinear iterative algorithms.

[14] Shortly before the publication of this thesis, this bound was improved to $\widetilde{O}(\log^{1/4} n)$ [BJ25].

Chapter 4. We connect our approach with the cavity method from statistical physics. We provide background on the cavity method and derive several of its key algorithmic predictions rigorously by working in the space of tree diagrams.

Chapter 5. We study generalizations where the number of iterations grows with the input dimension. We discuss obstacles to extending our framework, and prove that the tree approximation remains valid for approximating the top eigenvector.

Part II: Polynomial Optimization

Chapter 6. We show how the theory from Part I applies to random quadratic polynomial optimization. We give a tight analysis of power iteration, its shifted variant, and the optimal low-degree polynomial for approximating the top eigenvector. We revisit and reinterpret Montanari's algorithm for hypercube optimization. Finally, we study the analogous problem for worst-case polynomials.

Chapter 7. We prove our main theorem bounding the second eigenvalue of sparse random tensors in the sense of Friedman and Wigderson. We present two multiscale union bound approaches that significantly improve over the naive approach.

Chapter 8. We present new rounding algorithms for higher-degree polynomial optimization and prove our main theorems on certifiable approximation guarantees for cubic optimization over both the sphere and the hypercube. We discuss extensions to higher degrees and establish our approximation result for Max-3-Sat.

Part III: Discrepancy Theory

Chapter 9. We present our new framework for discrepancy minimization based on regularization. We provide several interpretations of our method: as a continuous-time process, as a barrier argument, and as an online optimization game. We give a complete proof of Spencer's theorem using this framework. Finally, we extend our framework to matrix discrepancy.

Chapter 10. We state and prove our main results for pseudorandom instances of the Komlós and Beck–Fiala conjectures. We also explore connections between our framework and alternative arguments based on duality and compression.

Chapter 11. We show that Spencer's theorem holds with an improved constant. We discuss potential directions to further tighten this constant, including links to the optimal regret bound in learning with expert advice. We also argue that our framework cannot improve the constant for spectral sparsification.

1.5. Bibliographic notes

Part I is based on joint work with Chris Jones. A preliminary version of these results appeared in the proceedings of ICALP 2025 [JP25]. Chapter 8 is joint work with Tim Hsieh, Pravesh Kothari, and Luca Trevisan, and appeared in the proceedings of SODA 2024 [HKPT24].

Chapter 9, Chapter 10, and §11.1 are based on joint work with Adrian Vladu. With the exception of the extension to the matrix case in §9.5, these results appeared in the proceedings of SODA 2023 [PV23].

The remaining chapters, namely Chapter 6, Chapter 7, and Chapter 11, are original and unpublished contributions by the author. The author thanks Tim Hsieh for collaboration on the open problem mentioned in §6.4; Chris Jones and Tim Kunisky for discussions related to Theorem 6.9; and Luca Trevisan for his guidance, which led to the results in Chapter 7.

Notation

If $n \geqslant 0$ is an integer, we define $[n] \coloneqq \{1, \dots, n\}$. If $n, k \geqslant 0$ are integers, then $\binom{[n]}{k}$ denotes the set of subsets of $[n]$ of size k.

Vectors and matrices

We use boldface to denote vectors and matrices. $\mathbf{0}$ is the all-zeros vector, and $\mathbf{1}$ is the all-ones vector. The tuple $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$ is the standard basis of $\mathbb{R}^n$, where $e_{i,j} = 1$ if $i = j$ and $e_{i,j} = 0$ otherwise.

If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then $\boldsymbol{x} \odot \boldsymbol{y}$ denotes the componentwise product: $(\boldsymbol{x} \odot \boldsymbol{y})_i \coloneqq x_i \cdot y_i$ for all $i \in [n]$. For any integer $k \geqslant 1$, we define $\boldsymbol{x}^{\odot k} \coloneqq \boldsymbol{x} \odot \cdots \odot \boldsymbol{x}$ (k terms).

Let $\boldsymbol{A} \in \mathbb{R}^{d \times n}$. For $i \in [d]$, we denote by $\boldsymbol{A}_i$ the i-th row, and for $j \in [n]$, we denote by $\boldsymbol{A}^j$ the j-th column of $\boldsymbol{A}$.

If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, define $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \coloneqq \sum_{i=1}^n x_i y_i$. If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d \times n}$, define $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \coloneqq \sum_{i=1}^d \sum_{j=1}^n A_{ij} B_{ij}$.

If $\boldsymbol{x} \in \mathbb{R}^n$ and $p \in (0, \infty)$, define $\|\boldsymbol{x}\|_p \coloneqq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. We also define $\|\boldsymbol{x}\|_\infty \coloneqq \max_{i \in [n]} |x_i|$ and $\|\boldsymbol{x}\|_0 \coloneqq |\{i \in [n] : x_i \neq 0\}|$.

We denote by $\mathcal{S}^{n-1} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}$ the unit ℓ_2-sphere, and by $\Delta_n \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n_{\geqslant 0} : \sum_{i=1}^n x_i = 1\}$ the unit simplex.

If $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, then $\|\boldsymbol{A}\|_2 \coloneqq \max_{\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1} \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle$ is the spectral norm, and $\|\boldsymbol{A}\|_{\mathrm{F}} \coloneqq \langle \boldsymbol{A}, \boldsymbol{A} \rangle^{1/2}$ is the Frobenius norm.

If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we define $\operatorname{tr} \boldsymbol{A} \coloneqq \sum_{i=1}^n A_{ii}$ to be the (unnormalized) trace of $\boldsymbol{A}$. For symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$, we write $\boldsymbol{A} \succeq 0$ if $\boldsymbol{A}$ if positive semidefinite (i.e., all eigenvalues of $\boldsymbol{A}$ are non-negative), and $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A} \succeq 0$.

We denote by $\mathfrak{S}_n$ the symmetric group on n elements. If $\sigma \in \mathfrak{S}_n$ and $\boldsymbol{u} \in \mathbb{R}^n$, then $\sigma(\boldsymbol{u})_i \coloneqq u_{\sigma(i)}$. If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, then $\sigma(\boldsymbol{A})_{ij} \coloneqq A_{\sigma(i), \sigma(j)}$.

Tensors

For an integer $t \geqslant 1$, we say $\boldsymbol{T}$ is a t-uniform tensor over $\mathbb{R}^n$ when $\boldsymbol{T} = (T_{\boldsymbol{i}})_{\boldsymbol{i} \in [n]^t}$. The tensor $\boldsymbol{T}$ is symmetric if $T_{\boldsymbol{i}} = T_{\sigma(\boldsymbol{i})}$ for all $\sigma \in \mathfrak{S}_t$.

If $i \in [n]$, the i-th slice of $\boldsymbol{T}$ is $\boldsymbol{T}_i \in \mathbb{R}^{n^{t-1}}$, defined by $(\boldsymbol{T}_i)_{\boldsymbol{j}} \coloneqq T_{i,\boldsymbol{j}}$ for all $\boldsymbol{j} \in [n]^{t-1}$.

Given $\boldsymbol{u}_1, \dots, \boldsymbol{u}_t \in \mathbb{R}^n$, their tensor product is $\boldsymbol{u}_1 \otimes \cdots \otimes \boldsymbol{u}_t \in \mathbb{R}^{n^t}$ defined by

$$(\boldsymbol{u}_1 \otimes \cdots \otimes \boldsymbol{u}_t)_{\boldsymbol{i}} \coloneqq \prod_{j=1}^t u_j(i_j) \quad \text{for } \boldsymbol{i} \in [n]^t.$$

Probability theory

All asymptotics are with respect to $n \to \infty$ unless otherwise stated.

Let $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ and $\boldsymbol{Z}$ be random vectors. We write $\boldsymbol{X}_n \xrightarrow{\text{a.s.}} \boldsymbol{Z}$ if $\boldsymbol{X}_n$ converges almost surely to $\boldsymbol{Z}$, i.e., $\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{Z}$ with probability one.

We write $\boldsymbol{X}_n \xrightarrow{\text{d}} \boldsymbol{Z}$ if $\boldsymbol{X}_n$ converges in distribution to $\boldsymbol{Z}$, meaning that for every bounded continuous function f, we have $\mathbb{E}\, f(\boldsymbol{X}_n) \to \mathbb{E}\, f(\boldsymbol{Z})$.

We say that a sequence of events indexed by n holds with high probability if their probability tends to 1 as $n \to \infty$.

We will refer to the generalized (probabilist's) Hermite polynomials as $h_k(\,\cdot\,; \sigma^2)$, where h_k is the degree-k monic orthogonal polynomial for $\mathcal{N}(0, \sigma^2)$. If Z_i is an independent $\mathcal{N}(0, \sigma_i^2)$ random variable for all $i \in \mathcal{I}$, then $\left(\prod_{i\in\mathcal{I}} h_{k_i}(Z_i; \sigma_i^2)\right)_{k\in\mathbb{N}^{\mathcal{I}}}$ is an orthogonal basis for polynomials in $(Z_i)_{i\in\mathcal{I}}$ with respect to the expectation over $(Z_i)_{i\in\mathcal{I}}$.

Asymptotics

We use standard asymptotic notations O, Ω, o, ω, and Θ, as well as their equivalent forms $\lesssim$, $\gtrsim$, $\ll$, $\gg$, and $\asymp$. Unless otherwise specified, all parameters in such notations are universal constants.

Part I.

The Fourier Diagram Basis

CHAPTER 2.

The Fourier Diagram Basis of Wigner Matrices

This chapter introduces the *Fourier diagram basis*, an orthogonal basis of permutation-symmetric polynomials in the entries of a random matrix. We keep in this chapter a random matrix theory perspective. In later chapters, this technology will be instrumental to understand the dynamics of non-linear iterative algorithms generalizing the matrix power method, for solving problems such as random quadratic polynomial optimization.

While the Fourier diagram basis can be defined in different settings, we focus in this chapter on the simple case of *Wigner matrices*, i.e., random symmetric matrices with independent entries above the diagonal. The specific distribution of the entries is not important; in particular, we will obtain several *universality* results that hold across a wide class of models.

Table of contents

This chapter is based on [JP25].

2.1. Introduction

Throughout Part I, we will work with the following Wigner random matrix model.

Assumption 2.1 (Wigner model)**.** Let μ and μ_0 be two subgaussian[1] distributions on $\mathbb{R}$ such that $\mathbb{E}_{X\sim\mu}[X] = 0$ and $\mathbb{E}_{X\sim\mu}[X^2] = 1$.

Let $\boldsymbol{A}$ be a random $n\times n$ symmetric matrix with independent entries (up to the symmetry) which are either $\sqrt{n}A_{ii} \sim \mu_0$ on the diagonal or $\sqrt{n}A_{ij} \sim \mu$ off the diagonal.

We next define the model of iterative algorithm used throughout Part I. We adopt the framework of generalized first-order methods, introduced by Celentano, Montanari, and Wu [CMW20].

Definition 2.2 (General first-order method)**.** The input is a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. The state of the algorithm at time t is a vector $\boldsymbol{x}_t \in \mathbb{R}^n$. Initially, $\boldsymbol{x}_0 = \mathbf{1}$. At each time t, we can execute one of the following two operations:

1. Multiply by $\boldsymbol{A}$, i.e.,
$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t\,.$$
2. Apply coordinatewise a polynomial[2] function independent of n, $f_t\colon \mathbb{R}^{t+1} \to \mathbb{R}$ to $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_0)$, i.e., for all $i \in [n]$,
$$x_{t+1,i} = f_t(x_{t,i}, \dots, x_{0,i})\,.$$

This definition captures a wide class of iterative algorithms, including power iteration and the best-known algorithm for optimizing a random degree-2 polynomial over the hypercube by Montanari [Mon19]. We will focus mostly on the case where $\boldsymbol{A}$ is sampled from a null model, without any planted signal. However, similar iterative methods are also ubiquitous in statistical estimation, as the following example demonstrates.

Example 2.3 (Spiked Wigner model)**.** Let $\boldsymbol{W}$ be a matrix satisfying Assumption 2.1, $\boldsymbol{x}^* \in \mathbb{R}^n$ be a signal vector, and $\lambda > 0$ be a fixed parameter (the *signal-to-noise ratio*). We observe

$$\boldsymbol{A} \coloneqq \frac{\lambda}{n}\cdot \boldsymbol{x}^*(\boldsymbol{x}^*)^\top + \boldsymbol{W}\,.$$

In the Bayesian setting, we further assume that $\boldsymbol{x}^*$ is drawn from some product prior:

$$\boldsymbol{x}^*_i \overset{\text{i.i.d.}}{\sim} P_0\,, \quad i \in [n]\,,$$

for some fixed prior distribution P_0 on $\mathbb{R}$. The goal is to recover $\boldsymbol{x}^*$ by observing only $\boldsymbol{A}$.

[1] A distribution μ on $\mathbb{R}$ is *subgaussian* if there exists a constant $C > 0$ such that for all $q \in \mathbb{N}$, $\mathbb{E}_{X\sim\mu}[|X|^q] \leqslant C^q q^{q/2}$.

[2] Restriction to polynomial functions is a technical assumption which is not present in the original definition.

The quality of a solution $\boldsymbol{x}$ output by an algorithm is typically measured by the overlap $\langle \boldsymbol{x}, \boldsymbol{x}^* \rangle$. When P_0 is given a priori, the algorithm achieving the best-known overlap is an iterative algorithm in the sense of Definition 2.2 [RF12]. For example, when $P_0 = \mathcal{N}(0, 1)$, there is no structure in the planted vector to leverage, so that algorithm is simply the power method ($\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t$). When P_0 is the uniform distribution on $\{-1, 1\}$ (the $\mathbb{Z}_2$-*synchronization problem*), that algorithm uses non-linearities based on tanh: $\mathbb{R} \to [-1, 1]$ to map the coordinates of the iterates to $[-1, 1]$.

2.1.1. An example of Fourier diagram computation

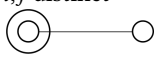
We introduce the Fourier diagram basis through an example: computing the vector $\boldsymbol{A}(\boldsymbol{A}\mathbf{1})^{\odot 2}$, which is the iterate of a simple nonlinear iterative algorithm. In general, calculation with diagrams is a bit like a symbolic version of the trace method from random matrix theory [Bor19].

For simplicity, we assume in this section that A satisfies Assumption 2.1 with $A_{ii} = 0$ for all $i \in [n]$.

We will use rooted multigraphs to represent vectors.[3] Multigraphs may include multiedges and self-loops. In our figures, the root will be drawn as a circled vertex ◎. The vector $\mathbf{1}$ will correspond to the singleton graph with one vertex (the root): ◎. Edges will correspond to A_{ij} terms.

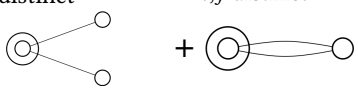
The vector $\boldsymbol{A}\mathbf{1}$ will be represented by the graph consisting of a single edge, with one of the endpoints being the root:

$$(\boldsymbol{A}\mathbf{1})_i = \sum_{j=1}^{n} A_{ij} = \sum_{\substack{j=1 \\ i,j \text{ distinct}}}^{n} A_{ij}$$

$\equiv$ 

where the second equality uses the assumption that A has zero diagonal. Now to apply the square function componentwise, we can decompose:

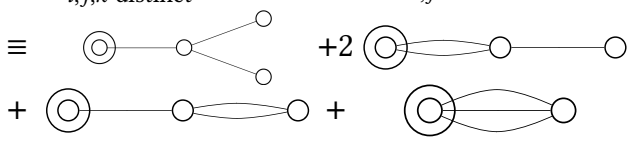
$$(\boldsymbol{A}\mathbf{1})_i^2 = \sum_{\substack{j,k=1 \\ i,j,k \text{ distinct}}}^{n} A_{ij}A_{ik} + \sum_{\substack{j=1 \\ i,j \text{ distinct}}}^{n} A_{ij}^2$$

$\equiv$ 

[3] Graphs with multiple roots can be used to represent matrices and tensors, although we will not need those here.

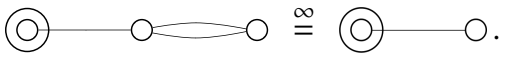
Moving on, we apply $\boldsymbol{A}$ to this representation by casing on whether the new index i matches one of the previous indices. We group terms together using the symmetry of $\boldsymbol{A}$ and the fact that $A_{ii} = 0$.

$$(\boldsymbol{A}(\boldsymbol{A}\mathbf{1})^2)_i = \sum_{\substack{j,k,\ell=1 \\ i,j,k,\ell \text{ distinct}}}^{n} A_{ij}A_{jk}A_{j\ell} + 2 \sum_{\substack{j,k=1 \\ i,j,k \text{ distinct}}}^{n} A_{ij}^2 A_{jk}$$
$$+ \sum_{\substack{j,k=1 \\ i,j,k \text{ distinct}}}^{n} A_{ij}A_{jk}^2 + \sum_{\substack{j=1 \\ i,j \text{ distinct}}}^{n} A_{ij}^3$$

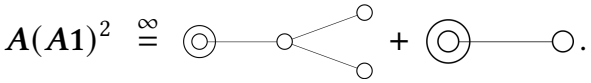


This is the non-asymptotic Fourier diagram representation of $\boldsymbol{A}(\boldsymbol{A}\mathbf{1})^2$.

We will see in Chapter 3 that in the limit $n \to \infty$, only some of these terms contribute to the *asymptotic* Fourier diagram basis representation. Asymptotically, *hanging* double edges can be removed from a diagram[4], so that the third diagram in the representation above satisfies, as $n \to \infty$,



The second and fourth diagrams in the representation of $\boldsymbol{A}(\boldsymbol{A}\mathbf{1})^2$ have entries on the scale $O(n^{-1/2})$ and so they will be dropped from the asymptotic diagram representation. In total,



We will show that as $n \to \infty$, the left diagram becomes a Gaussian vector with independent entries of variance 2, and the right diagram becomes a Gaussian vector with independent entries of variance 1. In fact, these $2n$ entries are asymptotically jointly independent. It can be verified numerically that approximately for large n, $\boldsymbol{A}(\boldsymbol{A}\mathbf{1})^2$ matches the sum of these two random vectors, the histogram of each vector's entries is Gaussian, and the vectors are approximately orthogonal.

2.2. Definition of the Fourier diagram basis

Definition 2.4. A *Fourier diagram* is an unlabeled undirected multigraph $\alpha = (V(\alpha), E(\alpha))$ with a special vertex labeled ◎ which we call the *root*. No vertices may be isolated except for the root. We let $\mathcal{A}$ be the set of all Fourier diagrams.

[4] To be convinced of this, the reader can think of the case where the entries of A are uniform $\pm\frac{1}{\sqrt{n}}$.

Definition 2.5. For a Fourier diagram $\alpha \in \mathcal{A}$ with root $\circledcirc$, define the vector $\boldsymbol{Z}_\alpha \in \mathbb{R}^n$ by

$$Z_{\alpha,i} = \sum_{\substack{\varphi\colon V(\alpha)\hookrightarrow[n] \\ \varphi(\circledcirc)=i}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}\,, \qquad \text{for all } i \in [n]\,,$$

where $\varphi\colon V(\alpha) \hookrightarrow [n]$ means that we sum over all injective maps from $V(\alpha)$ to $[n]$.

Among all Fourier diagrams, the ones corresponding to trees play a special role. They will constitute the *asymptotic Fourier diagram basis*.

Definition 2.6 ($\mathcal{S}$ and $\mathcal{T}$)**.** Let $\mathcal{S}$ be the set of unlabeled rooted trees such that the root has exactly one subtree (i.e. the root has degree 1). Let $\mathcal{T}$ be the set of all unlabeled rooted trees (non-empty, but allowing the singleton).

Definition 2.7 (Proper Fourier diagram)**.** A proper Fourier diagram is a Fourier diagram with no multiedges or self-loops (i.e. a rooted simple graph).

For *proper* Fourier diagrams $\alpha \in \mathcal{A}$, the following properties of $\boldsymbol{Z}_\alpha$ hold in a non-asymptotic sense, i.e., for arbitrary n:

1. $\boldsymbol{Z}_\alpha$ is a multilinear polynomial in the entries of A with degree $|E(\alpha)|$ (or $Z_\alpha = 0$ when $|V(\alpha)| > n$).
2. $\boldsymbol{Z}_\alpha$ has the symmetry that $Z_{\alpha,i}(\boldsymbol{A}) = Z_{\alpha,\pi(i)}(\pi(\boldsymbol{A}))$ for all permutations $\pi \in \mathfrak{S}_n$, where π acts on $\boldsymbol{A}$ by permuting the rows and columns simultaneously.
3. For each $i \in [n]$, the set $\{Z_{\alpha,i} : \text{proper Fourier diagram } \alpha \in \mathcal{A}\}$ is orthogonal with respect to the expectation over $\boldsymbol{A}$.
4. In fact, $\boldsymbol{Z}_\alpha$ is a symmetrized multilinear Fourier character (see §2.3). This implies the previous properties and it shows that the proper diagrams are an orthogonal basis for a class of symmetric functions of $\boldsymbol{A}$.

We may represent the algorithmic state $\boldsymbol{x}_t$ of a GFOM in the diagram basis,

$$\boldsymbol{x}_t = \sum_{\alpha\in\mathcal{A}} c_\alpha \boldsymbol{Z}_\alpha\,.$$

To multiply together or apply algorithmic operations on a diagram expression, we case on which indices repeat, like in the example in §2.1. See Lemmas 2.11 and 2.14 for a formal derivation of these rules.

2.3. The Fourier analysis viewpoint

The Fourier diagrams form an orthogonal basis that can be derived in a mechanical way using *symmetrization*.

We can use Fourier analysis to express a function or algorithm with respect to a natural basis. The unsymmetrized underlying analytical space consists of functions of the n^2 entries of $\boldsymbol{A}$. Since the entries of $\boldsymbol{A}$ are independent, the associated Fourier basis is the product basis for the different entries. When $\boldsymbol{A} \in \{-1,1\}^{n\times n}$ is a Rademacher random matrix, the Fourier characters are the multilinear monomials in $\boldsymbol{A}$. An arbitrary function $f : \{-1,1\}^{n\times n} \to \mathbb{R}$ is then expressed as

$$f(\boldsymbol{A}) = \sum_{\alpha\subseteq[n]\times[n]} c_\alpha \prod_{(i,j)\in\alpha} A_{ij}\,,$$

where c_α are the Fourier coefficients of f. When $\boldsymbol{A}$ is a symmetric matrix with zero diagonal, we only need Fourier characters for the top half of $\boldsymbol{A}$, and the basis simplifies to $\alpha \subseteq \binom{[n]}{2}$. That is, the possible α can be interpreted combinatorially as graphs on the vertex set $[n]$.

An observation that allows us to significantly simplify the representation is that many of the Fourier coefficients are equal for $\mathfrak{S}_n$-*equivariant* algorithms. A function $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is $\mathfrak{S}_n$-equivariant if it satisfies $f(\pi(\boldsymbol{A})) = f(\boldsymbol{A})$ or if $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}^n$ satisfies $f(\pi(\boldsymbol{A})) = \pi(f(\boldsymbol{A}))$ where π acts on $\boldsymbol{A}$ by permuting the rows and columns simultaneously. For scalar-valued functions, considering the action of $\mathfrak{S}_n$ on the vertex set of the Fourier characters $[n]$, any two Fourier characters α, β which are in the same orbit will have the same Fourier coefficient. Equivalently, if α and β are isomorphic as graphs, then their Fourier coefficients are the same. By grouping together all isomorphic Fourier characters, we obtain the symmetry-reduced representation defining the Fourier diagram basis,

$$f(\boldsymbol{A}) = \sum_{\text{nonisomorphic } \alpha\subseteq\binom{[n]}{2}} c_\alpha \left(\sum_{\varphi\colon V(\alpha)\hookrightarrow[n]} \prod_{\{u,v\}\in\alpha} A_{\varphi(u)\varphi(v)} \right).$$

Thus by construction, the diagrams are an orthogonal basis for symmetric low-degree polynomials of $\boldsymbol{A}$.

The above discussion was for Boolean matrices with $A_{ij} \sim \{\pm 1\}$. The natural generalization expresses polynomials in the basis of orthogonal polynomials for the entries A_{ij} (e.g. the Hermite polynomials when the $A_{ij} \sim \mathcal{N}(0, 1/n)$ [MW25, §3.2]).

Our results show that for the first-order algorithms we consider, only the multilinear part of the basis matters (i.e. the orthogonal polynomials which are degree 0 or 1 in each variable): up to negligible error, we can approximate $A_{ij}^2 \approx \frac{1}{n}$ and $A_{ij}^k \approx 0$ for $k \geqslant 3$. We use the monomial basis[5] to represent higher-degree polynomials instead of higher-degree orthogonal polynomials in order to simplify the presentation (except for the degree-2 orthogonal polynomial $A_{ij}^2 - \frac{1}{n}$ which expresses some error terms).

[5] The monomial "basis" is a misnomer in the cases when A_{ij} satisfies a polynomial identity such as $A_{ij}^2 = \frac{1}{n}$. In these cases, representation as a sum of diagrams is not unique. Our expressions should be interpreted as giving explicit sums of diagrams.

2.3.1. Consequences

In Definition 2.5, for a proper $\alpha \in \mathcal{A}$ (a graph instead of a multigraph), Z_α has entries which are homogeneous multilinear polynomials in the entries of the matrix $\boldsymbol{A}$. The next lemma shows that the proper diagrams with size at most n form an orthogonal basis of symmetric polynomials in $\boldsymbol{A}$ with respect to the expectation over $\boldsymbol{A}$.

Lemma 2.8. *For all $i, j \in [n]$ and distinct proper diagrams $\alpha, \beta \in \mathcal{A}$, $\mathbb{E}\left[Z_{\alpha,i} Z_{\beta,j}\right] = 0$.*

Proof. For each distinct $S, T \subseteq \binom{[n]}{2}$, the independence and centeredness of the off-diagonal entries of A proves that

$$\mathbb{E}\left[\prod_{\{i,j\}\in S} A_{ij} \prod_{\{k,\ell\}\in T} A_{k\ell}\right] = 0\,.$$

Two distinct diagrams sum over distinct sets of multilinear monomials, so this orthogonality extends to diagrams. □

The diagrams are not normalized for that inner product, but their variance can be estimated as follows:

Lemma 2.9. *For all $i \in [n]$ and proper $\alpha \in \mathcal{A} \setminus \{\circledcirc\}$ we have $\mathbb{E}\left[Z_{\alpha,i}\right] = 0$ and*

$$\begin{aligned}\mathbb{E}\left[Z_{\alpha,i}^2\right] &= |\mathrm{Aut}(\alpha)| \cdot \frac{(n-1)(n-2)\cdots(n-|V(\alpha)|+1)}{n^{|E(\alpha)|}} \\ &\underset{n\to\infty}{=} |\mathrm{Aut}(\alpha)| \cdot n^{|V(\alpha)|-1-|E(\alpha)|}(1+o(1))\,,\end{aligned}$$

where the last estimate holds when $|V(\alpha)| = o(\sqrt{n})$.

Proof. When α is proper, $Z_{\alpha,i}$ is a multilinear polynomial with zero constant coefficient, and so it has expectation 0. For the second moment, we have

$$\mathbb{E}\left[Z_{\alpha,i}^2\right] = \sum_{\substack{\varphi_1 \colon V(\alpha)\hookrightarrow[n] \\ \varphi_1(\circledcirc)=i}} \sum_{\substack{\varphi_2 \colon V(\alpha)\hookrightarrow[n] \\ \varphi_2(\circledcirc)=i}} \mathbb{E}\left[\prod_{\{u,v\}\in E(\alpha)} A_{\varphi_1(u)\varphi_1(v)} A_{\varphi_2(u)\varphi_2(v)}\right].$$

Since $\mathbb{E}\left[A_{jk}\right] = 0$ for $j \neq k$, the only terms with nonzero expectation have each A_{jk} occurring at least twice. As φ_1 are φ_2 are injective, each A_{jk} can only occur at most twice. Therefore, if we fix φ_1 the embeddings φ_2 that contribute a nonzero value are exactly graph isomorphisms onto $\mathrm{img}(\varphi_1)$. The total number of choices for φ_1 and φ_2 is $(n-1)\cdots(n-|V(\alpha)|+1)\cdot|\mathrm{Aut}(\alpha)|$ and the expectation of a nonzero term is

$$\prod_{\{j,k\}\in E(\alpha)} \mathbb{E}\left[A_{jk}^2\right] = \frac{1}{n^{|E(\alpha)|}}\,.$$

This completes the proof of the first part of the statement. Under the further assumption $|V(\alpha)| = o(\sqrt{n})$, the falling factorial can then be estimated as

$$\left|\log\left(\frac{(n-1)\dots(n-|V(\alpha)|+1)}{n^{|V(\alpha)|-1}}\right)\right| \leqslant \sum_{i=1}^{|V(\alpha)|-1}\left|\log\left(1-\frac{i}{n}\right)\right| \leqslant \sum_{i=1}^{|V(\alpha)|-1}\frac{i}{n} \underset{n\to\infty}{\longrightarrow} 0\,.$$

This implies that $(n-1)\dots(n-|V(\alpha)|+1) = (1+o(1))n^{|V(\alpha)|-1}$, as desired. □

We can already see from the previous lemma that if $\alpha \in \mathcal{T}$ is a tree, then the variance of $Z_{\alpha,i}$ is $\Theta(1)$, whereas if α is a connected graph with a cycle, then the variance of $Z_{\alpha,i}$ is $o(1)$.

We will use orthogonality repeatedly in the sequel through the following direct consequence of Lemma 2.8 and Lemma 2.9:

Corollary 2.10. *Let $\boldsymbol{x} = \sum_{\text{proper } \alpha\in\mathcal{A}} c_\alpha \boldsymbol{Z}_\alpha$. Then for any $\tau \in \mathcal{T}$,*

$$\mathbb{E}\left[x_i Z_{\tau,i}\right] = c_\tau\,\mathbb{E}\left[Z_{\tau,i}^2\right] \underset{n\to\infty}{=} c_\tau\,|\mathrm{Aut}(\tau)| + o(1)\,,$$

where the second estimate holds for $|V(\tau)| = o(\sqrt{n})$.

In particular, $\mathbb{E}[\boldsymbol{x}] = c_{\circledcirc}\mathbf{1}$ where $c_{\circledcirc}$ is the coefficient of the singleton diagram.

2.4. Operations on the diagram representation

We compute the diagrammatic effect of multiplying by $\boldsymbol{A}$ in the Fourier diagram basis. For any diagram $\alpha \in \mathcal{A}$, we use the notation α^+ to denote the diagram obtained from α by extending the root by one edge.

Lemma 2.11. *For all diagrams $\alpha \in \mathcal{A}$,*

$$\boldsymbol{A}\boldsymbol{Z}_\alpha = \boldsymbol{Z}_{\alpha^+} + \sum_{v\in V(\alpha)} \boldsymbol{Z}_{\text{contract } v \text{ and } \circledcirc \text{ in } \alpha^+}\,.$$

Proof.

$$(\boldsymbol{A}\boldsymbol{Z}_\alpha)_i = \sum_{j=1}^{n} A_{ij} \sum_{\substack{\varphi\colon V(\alpha)\hookrightarrow[n]\\ \varphi(\circledcirc)=j}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)} = \sum_{\varphi\colon V(\alpha)\hookrightarrow[n]} A_{i,\varphi(\circledcirc)} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}\,.$$

The sum over φ can be partitioned based on whether $i \in \mathrm{img}(\varphi)$. The terms with $i \notin \mathrm{img}(\varphi)$ sum to Z_{α^+}. The terms with $i \in \mathrm{img}(\varphi)$ sum to the different contractions of α^+ based on which vertex of α is labeled i. □

Switching to componentwise operations, the combinatorics is captured by the concepts of intersection patterns and intersection diagrams.

Definition 2.12 (Intersection pattern, $P \in \mathcal{P}(\alpha_1, \ldots, \alpha_k)$)**.** Let $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$. Let α be the diagram obtained by putting all α_i at the same root. An intersection pattern P is a partition of $V(\alpha) \setminus \{◎\}$ such that for all $i \in [k]$ and $v, w \in V(\alpha_i) \setminus \{◎\}$, v and w are not in the same block of the partition.

Let $\mathcal{P}(\alpha_1, \ldots, \alpha_k)$ be the set of intersection patterns between $\alpha_1, \ldots, \alpha_k$.

Definition 2.13 (Intersection diagram, α_P)**.** Let $\alpha \in \mathcal{A}$. Given a partition P of $V(\alpha)$, let α_P be the diagram obtained by contracting each block of P into a single vertex. Keep all edges (hence there may be new multiedges or self-loops).

By casing on which vertices are equal among the embeddings of $\alpha_1, \ldots, \alpha_k$ as in the proof of Lemma 2.11, we have:

Lemma 2.14. *For $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$, the componentwise product of $Z_{\alpha_1}, \ldots, Z_{\alpha_k}$ is*

$$Z_{\alpha_1} \odot \cdots \odot Z_{\alpha_k} = \sum_{P \in \mathcal{P}(\alpha_1, \ldots, \alpha_k)} Z_{\alpha_P}.$$

2.5. Repeated-label diagram basis

An alternative basis for the diagram space consists of diagrams in which labels are allowed to repeat. This representation has been defined by Ivkov and Schramm [IS24, §3.5]. We will not use this basis in this work, for reasons that will become apparent in the next chapter. One key observation at this point is that this alternate basis is *not* orthogonal.

Definition 2.15 ($\widetilde{Z}_\alpha$)**.** For a diagram α with root ◎, define $\widetilde{Z}_\alpha \in \mathbb{R}^n$ by

$$\widetilde{Z}_{\alpha,i} = \sum_{\substack{\varphi : V(\alpha) \to [n] \\ \varphi(◎) = i}} \prod_{\{u,v\} \in E(\alpha)} A_{\varphi(u)\varphi(v)}.$$

The only difference between $\widetilde{Z}_\alpha$ and Z_α is that the embedding φ must be injective in Z_α. To perform the change of basis in one direction is as easy as replacing $\widetilde{Z}_\alpha$ by a sum of Z_α based on which labels are repeated.

Lemma 2.16. *For $\alpha \in \mathcal{A}$,*

$$\widetilde{Z}_\alpha = \sum_{P \in \mathcal{P}(\alpha)} Z_{\alpha_P}$$

where $\mathcal{P}(\alpha)$ is the set of partitions of $V(\alpha)$ and α_P contracts the blocks of P (definition 2.13).

Proof. We have

$$\widetilde{Z}_{\alpha,i} = \sum_{\substack{\varphi: V(\alpha) \to [n] \\ \varphi(\circledcirc) = i}} \prod_{\{u,v\} \in E(\alpha)} A_{\varphi(u)\varphi(v)} \,.$$

The sum over φ can be divided based on which vertices are assigned the same label. The terms with a given partition P of $V(\alpha)$ are exactly $Z_{\alpha_P,i}$. □

The algorithmic operations are simpler to compute in this basis, although the asymptotic tree approximation does not seem to be easily visible in this basis (the tree diagrams do not span the same space, and a diagram which is an even cycle has entries with magnitude $\Theta(1)$ in $\widetilde{Z}_\alpha$ but negligible entries in Z_α).

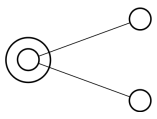
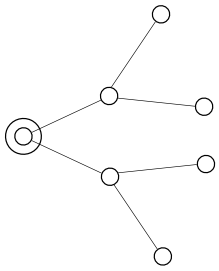
Given the current representation $\boldsymbol{x}_t = \sum_{\tau \in \mathcal{T}} c_\tau \widetilde{Z}_\tau$ the operations have the following effects on the $\widetilde{Z}_\tau$ (non-asymptotically i.e. without taking the limit $n \to \infty$).

1. *Multiplying by $\boldsymbol{A}$ extends the root.*
   We have $\boldsymbol{A}\widetilde{Z}_\alpha = \widetilde{Z}_{\alpha^+}$ where α^+ is obtained by extending the root by one edge.
2. *Componentwise products graft trees together.*
   To componentwise multiply $\widetilde{Z}_\alpha$ and $\widetilde{Z}_\beta$, we "graft" α and β by merging their roots.

Example 2.17. Consider the example,

$$\boldsymbol{x}_{t+1} = (\boldsymbol{A}\boldsymbol{x}_t)^{\odot 2} \qquad \boldsymbol{x}_0 = \boldsymbol{1}$$

where $\boldsymbol{1} \in \mathbb{R}^n$ is the all-ones vector and the square function is applied componentwise. The first few iterations are,

$\boldsymbol{x}_0 = \boldsymbol{1}$	$\boldsymbol{x}_1 = (\boldsymbol{A}\boldsymbol{1})^2$	$\boldsymbol{x}_2 = (\boldsymbol{A}(\boldsymbol{A}\boldsymbol{1})^{\odot 2})^{\odot 2}$
$x_{0,i} = 1$	$x_{1,i} = \sum_{j_1,j_2=1}^{n} A_{ij_1}A_{ij_2}$	$x_{2,i} = \sum_{j_1,j_2=1}^{n} \sum_{k_1,k_2=1}^{n} \sum_{\ell_1,\ell_2=1}^{n} A_{ij_1}A_{ij_2}A_{j_1k_1}A_{j_1\ell_1}A_{j_2k_2}A_{j_2\ell_2}$
		

The Fourier diagram basis and the repeated-label diagram basis may appear similar at this point. The key difference is that the first one has nicer properties in the limit $n \to \infty$, as we will see in the next chapter.

2.6. Summary

We introduced our diagrammatic framework to represent the iterates of algorithms applied to a Wigner matrix. Each vector iterate is an $\mathfrak{S}_n$-symmetric polynomial in the matrix entries and can be expanded in the *Fourier diagram basis*. Unlike the classical *repeated-label diagram basis*, the Fourier basis is orthogonal, which will be key for the asymptotic results developed in the next chapter.

CHAPTER 3.

The Asymptotic Tree Approximation

Building on Chapter 2, this chapter studies the behavior of the Fourier diagram basis in the high-dimensional limit $n \to \infty$. We begin by proving that the scaling of a diagram is determined by its number of excess edges, and that *tree diagrams* are precisely the dominant contributions. This is the essence of the *asymptotic tree approximation*. Next, we show that a subset of these trees forms a basis of asymptotically independent Gaussian variables, while the remaining diagrams are asymptotically Hermite polynomials in them. In particular, the joint distribution of a set of Fourier diagrams can be directly understood from their graph-theoretic properties. Finally, we apply these results to explicitly describe the asymptotic behavior of non-linear iterative algorithms.

Table of contents

This chapter is based on [JP25].

3.1. Asymptotic properties of the Fourier diagram basis

We still assume that $\boldsymbol{A}$ satisfies Assumption 2.1. Recall the set of tree diagrams $\mathcal{T}$ defined in §2.2. The constant-size tree diagrams $(\boldsymbol{Z}_\tau)_{\tau\in\mathcal{T}}$ exhibit the following key properties in the limit $n \to \infty$ and with respect to the randomness of $\boldsymbol{A}$.

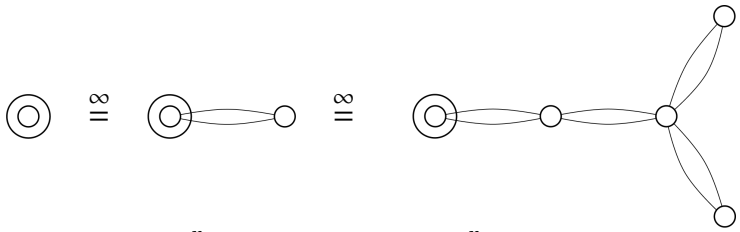
1. The coordinates of $\boldsymbol{Z}_\tau \in \mathbb{R}^n$ for any $\tau \in \mathcal{T}$ are asymptotically independent and identically distributed.
2. The random variables $Z_{\sigma,1}$ for $\sigma \in \mathcal{S}$ (the tree diagrams with one subtree) are asymptotically independent Gaussians with variance $|\mathrm{Aut}(\sigma)|$, where $\mathrm{Aut}(\sigma)$ are the graph automorphisms of σ which fix the root.
3. The random variable $Z_{\tau,1}$ for $\tau \in \mathcal{T}$ (the tree diagrams with multiple subtrees) is asymptotically equal to the multivariate Hermite polynomial

$$\prod_{\sigma\in\mathcal{S}} h_{d_\sigma}(Z_{\sigma,1}; |\mathrm{Aut}(\sigma)|)$$

   where d_σ is the number of children of the root whose subtree (including the root) equals $\sigma \in \mathcal{S}$.

The remaining Fourier diagrams not in $\mathcal{T}$ can be understood using the further asymptotic properties:

(iv) For any diagram $\alpha \in \mathcal{A}$, if α has a *hanging double edge* i.e. a double edge with one non-root endpoint of degree exactly 2, letting α_0 be the diagram with the hanging double edge and hanging vertex removed, then $\boldsymbol{Z}_\alpha$ is asymptotically equal to $\boldsymbol{Z}_{\alpha_0}$. For example, the following diagrams are asymptotically equal:



$$1 \quad \approx \quad \sum_{\substack{j=1\\ i\neq j}}^{n} A_{ij}^2 \quad \approx \quad \sum_{\substack{j,k,\ell,m=1\\ i,j,k,\ell,m \text{ distinct}}}^{n} A_{ij}^2 A_{jk}^2 A_{k\ell}^2 A_{km}^2$$

(v) For any *connected* $\alpha \in \mathcal{A}$, if removing the hanging trees of double edges from α creates a diagram in $\mathcal{T}$, then by the previous property, $\boldsymbol{Z}_\alpha$ is asymptotically equal to that diagram. If the result is not in $\mathcal{T}$, then $\boldsymbol{Z}_\alpha$ is asymptotically negligible.

(vi) The disconnected diagrams have only a minor and negligible role in the algorithms that we consider. See §3.4 for the description of these random variables.

To summarize the properties, given a sum $\boldsymbol{x}$ of connected diagrams, by removing the hanging double trees, and then removing all diagrams not in $\mathcal{T}$, the expression admits an

asymptotic Fourier diagram basis representation of the form

$$\boldsymbol{x} \overset{\infty}{=} \sum_{\tau \in \mathcal{T}} c_\tau \boldsymbol{Z}_\tau \,, \tag{3.1}$$

for some coefficients $c_\tau \in \mathbb{R}$ independent of n and $\boldsymbol{A}$. We call this the *tree approximation* to $\boldsymbol{x}$. Note that all tree diagrams have order 1 variance regardless of their size, which can be counter-intuitive.

3.2. The idealized Gaussian dynamic

The main appeal of the tree approximation in (3.1) is that when restricted to the tree-shaped diagrams, the GFOM operations have a very simple interpretation: they implement an idealized *Gaussian dynamics* which we describe now.

The idealized GFOM moves through an "asymptotic Gaussian probability space" which is naturally the one corresponding to the $n \to \infty$ limit of the diagrams. Based on the diagram classification, this consists of an infinite family of independent Gaussian vectors $(\boldsymbol{Z}_\sigma)_{\sigma \in \mathcal{S}}$. However, due to symmetry, all of the coordinates follow the same dynamic, so we can compress the representation of the dynamic down to a one-dimensional random variable X_t which is the asymptotic distribution of each coordinate $x_{t,i}$. We call X_t the *asymptotic state* of $\boldsymbol{x}_t$.

For example, Approximate Message Passing (AMP) is a special type of GFOM whose iterates are asymptotically Gaussian i.e. X_t is a Gaussian random variable for all t (in general GFOMs, although X_t is defined in terms of Gaussians, it is not necessarily Gaussian).

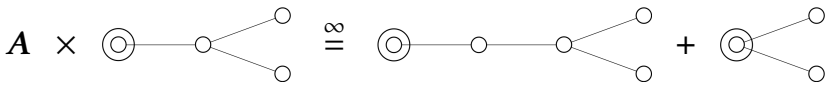
The algorithmic operations restricted to the trees and the corresponding evolution of the asymptotic state X_t are as follows. Two important operations on a tree-shaped diagram are extending/contracting the root by one edge.

Definition 3.1 (+ and − operators)**.** We define $+ : \mathcal{T} \to \mathcal{S}$ and $- : \mathcal{S} \to \mathcal{T}$ by:

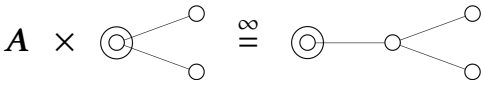
- If $\tau \in \mathcal{T}$, let τ^+ be the diagram obtained by extending the root by one edge (i.e. adding one new vertex and one edge connecting it to the root of τ, and re-rooting τ^+ at this new vertex).
- If $\tau \in \mathcal{S}$, let τ^- be the diagram obtained by contracting the root by one edge (i.e. removing the root vertex and the unique edge from it, and re-rooting τ^- at the endpoint of that edge).

Recall that the possible operations of a GFOM are either multiplying the state by $\boldsymbol{A}$ or applying a function componentwise. The effect of these two operations on the tree-shaped diagrams are:

- If $\sigma \in \mathcal{S}$, then $\boldsymbol{A}\boldsymbol{Z}_\sigma$ is asymptotically the sum of the diagrams σ^+ and σ^- obtained by respectively extending and contracting the root by one edge. For example,



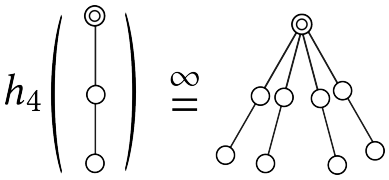
  If $\tau \in \mathcal{T} \setminus \mathcal{S}$, then $\boldsymbol{A}\boldsymbol{Z}_\tau$ is asymptotically only the τ^+ term. For example,



- From the classification of diagrams, if $\tau \in \mathcal{T}$ consists of d_σ copies of $\sigma \in \mathcal{S}$, then

$$\prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(\boldsymbol{Z}_\sigma; |\mathrm{Aut}(\sigma)|) \overset{\infty}{=} \boldsymbol{Z}_\tau \,. \tag{3.2}$$

  Therefore, to compute $f(\boldsymbol{Z}_\sigma : \sigma \in \mathcal{S})$, we expand f in the Hermite polynomial basis associated to $\mathcal{S}$, and apply the rule (3.2) to all the terms. For example,



These operations correspond to the following Gaussian dynamic.

Definition 3.2 (Asymptotic Gaussian space, Ω)**.** Let $(Z^\infty_\sigma)_{\sigma \in \mathcal{S}}$ be a set of independent centered (one-dimensional) Gaussian random variables with variances $\mathrm{Var}(Z^\infty_\sigma) = |\mathrm{Aut}(\sigma)|$.

If $\tau \in \mathcal{T}$ can be decomposed as d_σ copies of each $\sigma \in \mathcal{S}$ branching from the root, we define

$$Z^\infty_\tau = \prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(Z^\infty_\sigma; |\mathrm{Aut}(\sigma)|) \,.$$

We call *asymptotic states* the elements in the linear span of $(Z^\infty_\tau)_{\tau \in \mathcal{T}}$. We can view them both as polynomials in the formal variables $(Z^\infty_\sigma)_{\sigma \in \mathcal{S}}$ and as real-valued random variables. The set of asymptotic states is denoted Ω.

Definition 3.3 (Asymptotic state)**.** If $\boldsymbol{x} \in \mathbb{R}^n$ is such that $\boldsymbol{x} \overset{\infty}{=} \sum_{\tau \in \mathcal{T}} c_\tau \boldsymbol{Z}_\tau$, we define the *asymptotic state* of $\boldsymbol{x}$ by

$$X = \sum_{\tau \in \mathcal{T}} c_\tau Z^\infty_\tau \,.$$

The state evolution of the algorithm can now be described concisely as:

- If $\boldsymbol{x}_t$ has asymptotic state X_t, then the asymptotic state of $\boldsymbol{A}\boldsymbol{x}_t$ is $X_t^+ + X_t^-$. Here we extend the + and − operators linearly to sums of $\boldsymbol{Z}_\tau$ or Z^∞_τ (let $\boldsymbol{Z}^-_\tau = (Z^\infty_\tau)^- = 0$ if $\tau \in \mathcal{T} \setminus \mathcal{S}$).
- If $\boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_0$ have asymptotic states $X_{t-1}, \dots, X_0$ and f is any polynomial, then the asymptotic state of $f(\boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_0)$ is $f(X_{t-1}, \dots, X_0)$.

3.3. Equality up to combinatorially negligible diagrams

The idea behind $\overset{\infty}{=}$ is to make a purely combinatorial definition so that we can use combinatorial arguments on the diagrams. First, we have the following key moment bound which estimates the magnitude in n of a diagram $\boldsymbol{Z}_\alpha$ based on combinatorial properties of α.

Definition 3.4 ($I(\alpha)$)**.** For a diagram $\alpha \in \mathcal{A}$, let $I(\alpha)$ be the subset of non-root vertices such that every edge incident to that vertex has multiplicity $\geqslant 2$ or is a self-loop.

Lemma 3.5. *Let $q \in \mathbb{N}$ be a constant independent of n and $\alpha \in \mathcal{A}$ be a constant-size diagram. Then for $i \in [n]$,*

$$\left|\mathbb{E}\left[Z_{\alpha,i}^q\right]\right| \leqslant O\left(n^{\frac{q}{2}(|V(\alpha)|-1-|E(\alpha)|+|I(\alpha)|)}\right).$$

A similar norm bound for matrices is a crucial ingredient in Fourier analysis of matrix-valued functions [AMP20]. The proof of Lemma 3.5 is in Appendix A.2.2.

Based on this computation, we define a *combinatorially negligible diagram* to be one whose moments decay with n. Since we will be working with diagram expressions that are linear combinations of different diagrams, the following definition also handles diagrams rescaled by some coefficient depending on n.

Definition 3.6 (Combinatorially negligible and order 1)**.** Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real-valued coefficients such that $a_n = \Theta(n^{-k})$ for some $k \geqslant 0$ with $2k \in \mathbb{Z}$. Let $\alpha \in \mathcal{A}$ be a constant-size diagram.

1. We say that $a_n\boldsymbol{Z}_\alpha$ is *combinatorially negligible* if

$$|V(\alpha)| - 1 - |E(\alpha)| + |I(\alpha)| \leqslant 2k - 1\,.$$

   For $a_n = 0$, we also say that $a_n\boldsymbol{Z}_\alpha$ is combinatorially negligible.
2. We say that $a_n\boldsymbol{Z}_\alpha$ has *combinatorial order 1* if

$$|V(\alpha)| - 1 - |E(\alpha)| + |I(\alpha)| = 2k\,.$$

We will only consider settings where the coefficients are small enough so that all diagram expressions have combinatorial order at most 1 (that is, negligible or order 1).

Definition 3.7 ($\overset{\infty}{=}$)**.** We say that $\boldsymbol{x} \overset{\infty}{=} \boldsymbol{y}$ if there exists real coefficients $(c_\alpha)_{\alpha\in\mathcal{A}}$ depending on n and supported on diagrams of constant size such that

$$\boldsymbol{x} - \boldsymbol{y} = \sum_{\alpha\in\mathcal{A}} c_\alpha \boldsymbol{Z}_\alpha\,,$$

where $c_\alpha\boldsymbol{Z}_\alpha$ is combinatorially negligible for all $\alpha \in \mathcal{A}$.

Later, we will prove results of the form $\boldsymbol{x} \overset{\infty}{=} \widehat{\boldsymbol{x}}$ where $\boldsymbol{x}$ is the state of an algorithm and $\widehat{\boldsymbol{x}}$ is some asymptotic approximation of $\boldsymbol{x}$. In order to interpret these results, we note that $\overset{\infty}{=}$ implies very strong forms of convergence of the error to 0. The proof of the following lemma can be found in Appendix A.2.2.

Lemma 3.8. *Suppose that $\boldsymbol{A} = \boldsymbol{A}(n)$ is a sequence of random matrices satisfying Assumption 2.1. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are diagram expressions such that $\boldsymbol{x} \overset{\infty}{=} \boldsymbol{y}$, then $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \widetilde{O}(n^{-1/2})$ with high probability.*

Next, we prove a very important property of $\overset{\infty}{=}$. The combinatorially negligible diagrams remain combinatorially negligible after applying additional algorithmic operations.

Lemma 3.9. *If $\boldsymbol{x}, \boldsymbol{y}$ are diagram expressions with $\boldsymbol{x} \overset{\infty}{=} \boldsymbol{y}$, then*

$$\boldsymbol{A}\boldsymbol{x} \overset{\infty}{=} \boldsymbol{A}\boldsymbol{y}\,.$$

Moreover, if $\boldsymbol{x}_1, \dots, \boldsymbol{x}_t, \boldsymbol{y}_1, \dots, \boldsymbol{y}_t$ are diagram expressions with $\boldsymbol{x}_i \overset{\infty}{=} \boldsymbol{y}_i$ for all $i \in [t]$, then

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t) \overset{\infty}{=} f(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\,,$$

for any polynomial function $f : \mathbb{R}^t \to \mathbb{R}$ applied componentwise.

The proof of Lemma 3.9 is in Appendix A.2.2. The intuitive view of this lemma is that a diagram with a cycle still has the cycle after the algorithmic operations and thus remains negligible. The proof in Appendix A.2.2 is a syntactic version.

We can also show combinatorially that the error of removing a hanging double edge from any diagram is negligible. The proof proceeds by extending the definition of diagrams to allow new types of residual edges that are only used in the analysis (see Appendix A.2.1).

Lemma 3.10. *Let $a_n \boldsymbol{Z}_\alpha$ be a term of combinatorial order at most 1 such that α has a hanging double edge. Let α_0 be α with the hanging double edge and hanging vertex removed. Then*

$$a_n \boldsymbol{Z}_\alpha \overset{\infty}{=} a_n \boldsymbol{Z}_{\alpha_0}\,.$$

3.4. Classification of constant-size diagrams

We classify the asymptotic limits of constant-size diagrams and prove that all of their constant-order joint moments are within $O(n^{-1/2})$ of the asymptotic limit. In addition to the vector Fourier diagrams from Definition 2.4, we will classify *scalar Fourier diagrams*, which are simply unlabeled undirected multigraphs (the only difference with vector diagrams being that they do not have a root). The notation for scalar diagrams is analogous.

Definition 3.11 (Scalar Fourier diagrams)**.** Let $\mathcal{A}_{\text{scalar}}$ be the set of all unlabeled undirected multigraphs with no isolated vertices. Let $\mathcal{T}_{\text{scalar}}$ be the set of non-empty unlabeled trees.

Given a scalar Fourier diagram $\alpha \in \mathcal{A}_{\text{scalar}}$, we define $Z_\alpha \in \mathbb{R}$ by

$$Z_\alpha = \sum_{\varphi\colon V(\alpha)\hookrightarrow[n]} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}.$$

We allow the empty scalar Fourier diagram which represents the constant 1.

Definition 3.12 ($\mathcal{F}_{\text{scalar}}$ and $\mathcal{F}$)**.** Let $\mathcal{F}_{\text{scalar}}$ be the set of unlabeled forests with no isolated vertices. Let $\mathcal{F}$ be the set of unlabeled forests such that one vertex is the special root vertex ◎. No vertices may be isolated except for the root.

The scalar diagrams are not normalized "correctly" by default. Z_ρ for $\rho \in \mathcal{F}_{\text{scalar}}$ has order $n^{c/2}$ where c is the number of connected components in ρ. The proper normalization divides by $n^{c/2}$ to put all the diagrams on the same scale. The notion of $\overset{\infty}{=}$ and combinatorial negligibility also extend in a natural way to scalar diagrams. See Appendix A.3 for these definitions.

We classify the diagrams in $\mathcal{A}$ and $\mathcal{A}_{\text{scalar}}$. First, the next lemma identifies which of the diagrams are non-negligible. This lemma is for *connected* vector diagrams; scalar diagrams and disconnected vector diagrams have a similar characterization in Lemma A.17.

Lemma 3.13. *Let $\alpha \in \mathcal{A}$ be a connected Fourier diagram. Then Z_α is either combinatorially negligible or combinatorially order 1. Moreover, it is combinatorially order 1 if and only if the following four conditions hold simultaneously:*

1. *Every multiedge has multiplicity 1 or 2.*
2. *There are no cycles.*
3. *The subgraph of multiplicity 1 edges is connected and contains the root if it is nonempty (i.e. the multiplicity 2 edges consist of hanging trees).*
4. *There are no self-loops or 2-labeled edges (Appendix A.2.1).*

Proof. By assumption, every vertex is connected to the root. With the exception of the root, we can assign injectively one edge to every vertex in $V \setminus I(\alpha)$ and two edges to every vertex in $I(\alpha)$ as follows. Run a breadth-first search from the root and assign to each vertex the multiedge that was used to discover it. This encoding argument implies

$$(|V(\alpha)| - |I(\alpha)| - 1) + 2|I(\alpha)| \leqslant |E(\alpha)|\,.$$

Hence Z_α is combinatorially negligible or combinatorially order 1, and it is combinatorially order 1 if and only if this inequality is an equality. This holds if and only if there are no cycles, multiplicity $\geqslant 2$ edges, self-loops, or 2-labeled edges in α, and the edges incident to $V(\alpha) \setminus I(\alpha)$ in the direction of the root all have multiplicity 1. □

As a result, the non-negligible connected diagrams in $\mathcal{A}$ are asymptotically equal to trees in $\mathcal{T}$ after using Lemma 3.10 to remove the hanging double edges (disconnected diagrams $\alpha \in \mathcal{A}$ and scalar diagrams $\alpha \in \mathcal{A}_{\text{scalar}}$ are likewise asymptotically equal to a forest in $\mathcal{F}$ or $\mathcal{F}_{\text{scalar}}$).

The next Theorem 3.14 completes the classification by showing that the non-negligible diagrams in $\mathcal{T}, \mathcal{F}$, and $\mathcal{F}_{\text{scalar}}$ are asymptotically Gaussians and Hermite polynomials. The proof is in Appendix A.4. Also see Theorem A.23 for a version of the theorem in terms of moments.

Theorem 3.14 (Classification). *Suppose that $\boldsymbol{A} = \boldsymbol{A}(n)$ is a sequence of random matrices satisfying Assumption 2.1.*

The non-negligible scalar diagrams can be classified as follows:

- *If $\tau \in \mathcal{T}_{\text{scalar}}$, then $n^{-\frac{1}{2}} Z_\tau \xrightarrow{d} \mathcal{N}(0, |\text{Aut}(\tau)|)$.*
- *If $\rho \in \mathcal{F}_{\text{scalar}}$ has c connected components, then*

$$n^{-\frac{c}{2}} Z_\rho \overset{\infty}{=} \prod_{\tau \in \mathcal{T}_{\text{scalar}}} h_{d_\tau}(n^{-\frac{1}{2}} Z_\tau; |\text{Aut}(\tau)|)\,,$$

  where d_τ is the number of copies of τ in ρ.

The non-negligible vector diagrams can be classified as follows:

- *If $\sigma \in \mathcal{S}$ and $i \in [n]$, then $Z_{\sigma,i} \xrightarrow{d} \mathcal{N}(0, |\text{Aut}(\sigma)|)$.*
- *If $\tau \in \mathcal{T}$, then $\boldsymbol{Z}_\tau \overset{\infty}{=} \prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(\boldsymbol{Z}_\sigma; |\text{Aut}(\sigma)|)$ where d_σ is the number of isomorphic copies of σ starting from the root of τ, and the Hermite polynomial is applied componentwise.*
- *If $\alpha \in \mathcal{F}$ has c floating components (connected components which are not the component of the root), letting $\alpha_{\circledcirc}$ be the component of the root (a vector diagram) and α_{float} be the floating part (a scalar diagram), then $n^{-\frac{c}{2}} \boldsymbol{Z}_\alpha \overset{\infty}{=} n^{-\frac{c}{2}} Z_{\alpha_{\text{float}}} \boldsymbol{Z}_{\alpha_{\circledcirc}}$.*

Moreover, the random variables

$$\left\{Z_{\sigma,i} : \sigma \in \mathcal{S}, i \in [n]\right\} \cup \left\{n^{-\frac{1}{2}} Z_\tau : \tau \in \mathcal{T}_{\text{scalar}}\right\}$$

are asymptotically independent (Definition 3.15).

Finally, we formalize what we mean by *asymptotic independence* of vectors whose dimension can grow with n.

Definition 3.15 (Asymptotic independence). A family of real-valued random variables $(X_{n,i})_{n \in \mathbb{N}, i \in \mathcal{I}_n}$ is *asymptotically independent* if:

$$\forall q \in \mathbb{N}.\ \exists \varepsilon = \varepsilon(q) \underset{n \to \infty}{\longrightarrow} 0.\ \forall k \in \mathbb{N}^{\mathcal{I}_n} : \sum_{i \in \mathcal{I}_n} k_i = q.\ \left| \mathbb{E}\left[\prod_{i \in \mathcal{I}_n} X_{n,i}^{k_i}\right] - \prod_{i \in \mathcal{I}_n} \mathbb{E}\left[X_{n,i}^{k_i}\right] \right| \leqslant \varepsilon(q)\,.$$

Note that $\mathcal{I}_n$ may be infinite.

3.5. Tree approximation of GFOMs

Inductively following the rules given explicitly in §2.4, we may represent the algorithmic state $\boldsymbol{x}_t$ of a GFOM in the diagram basis. We define the *tree approximation* $\widehat{\boldsymbol{x}}_t$ to be the analogous diagram expression obtained by performing the algorithmic operations on only the tree diagrams, removing hanging double edges and removing the cyclic diagrams that appear.

Next, we derive how these operations behave explicitly. Suppose we start from $\tau \in \mathcal{T}$ and compute $\boldsymbol{A}\boldsymbol{Z}_\tau$. Which diagrams appearing in Lemma 2.11 are non-negligible? Following the asymptotic classification of non-negligible diagrams (§3.4), it is only τ^+ and τ^- (the latter only appears if the root of τ has degree 1, in which case τ^- is the result of intersecting ◎ and the child of the root then removing a double edge). Hence we conclude

$$\boldsymbol{A}\boldsymbol{Z}_\tau \overset{\infty}{=} \begin{cases} \boldsymbol{Z}_{\tau^+} + \boldsymbol{Z}_{\tau^-} & \text{if } \tau \in \mathcal{S} \\ \boldsymbol{Z}_{\tau^+} & \text{if } \tau \in \mathcal{T} \setminus \mathcal{S}\,. \end{cases}$$

Given tree diagrams $\tau_1, \dots, \tau_k \in \mathcal{T}$, the asymptotically non-negligible terms in the product in Lemma 2.14 are identified as follows. Let $\widetilde{\tau}$ be a non-negligible diagram appearing in the result, i.e. $\widetilde{\tau}$ is a tree with hanging trees of double edges. Since $\tau_1, \dots, \tau_k$ are connected, the hanging double trees must hang off the root vertex of $\widetilde{\tau}$ in order to avoid cycles. Additionally, they must arise as the overlap of two complete copies of the tree. Thus the asymptotically non-negligible terms are the partial matchings between isomorphic branches of the roots of the τ_i. Two copies of a branch $\sigma \in \mathcal{S}$ can be matched up into a tree of double edges in $|\mathrm{Aut}(\sigma)|$ ways.

Based on these observations, the *tree approximation* is formally defined to be the result of applying the algorithmic operations and removing the non-trees at each step.

Definition 3.16 (Tree approximation of a GFOM, $\widehat{\boldsymbol{x}}_t$)**.** Let $\boldsymbol{x}_t \in \mathbb{R}^n$ be the state of a GFOM. We recursively define the tree approximation of $\boldsymbol{x}_t$, denoted by $\widehat{\boldsymbol{x}}_t$, to be a diagram expression in the span of $(\boldsymbol{Z}_\tau)_{\tau \in \mathcal{T}}$.

1. Initially, $\widehat{\boldsymbol{x}}_0 = \boldsymbol{Z}_{◎}$.
2. If $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t$, define $\widehat{\boldsymbol{x}}_{t+1} = (\widehat{\boldsymbol{x}}_t)^+ + (\widehat{\boldsymbol{x}}_t)^-$.
3. If $\boldsymbol{x}_{t+1} = f_t(\boldsymbol{x}_t, \dots, \boldsymbol{x}_0)$ coordinatewise for some polynomial $f_t : \mathbb{R}^t \to \mathbb{R}$, define $\widehat{\boldsymbol{x}}_{t+1}$ by applying each monomial of f_t to $\widehat{\boldsymbol{x}}_t, \dots, \widehat{\boldsymbol{x}}_0$ separately and summing the results. To apply a monomial on $\widehat{\boldsymbol{x}}_t, \dots, \widehat{\boldsymbol{x}}_0$, expand each $\widehat{\boldsymbol{x}}_s$ in the diagram basis and sum all the cross product terms. The result of multiplying q tree diagrams $\tau_1, \dots, \tau_q \in \mathcal{T}$ is

$$\sum_{M \in \mathcal{M}(\tau_1, \dots, \tau_q)} c_M \boldsymbol{Z}_{\tau_M}\,,$$

where:

a) $\mathcal{M}(\tau_1, \dots, \tau_q)$ is the set of (partial) matchings of isomorphic branches of $\tau_1, \dots, \tau_q$ such that no two branches from the same τ_i are matched.
b) τ_M is the tree obtained by merging the roots of $\tau_1, \dots, \tau_q$ and removing all subtrees matched in M.
c) $c_M = \prod_{\{\sigma,\sigma'\}\in M} |\mathrm{Aut}(\sigma)|$.

Theorem 3.17 (Tree approximation of GFOMs)**.** *Let $t \geqslant 0$ be a constant independent of n and $\boldsymbol{A} = \boldsymbol{A}(n)$ be a sequence of random matrices satisfying Assumption 2.1. Let $\boldsymbol{x}_t \in \mathbb{R}^n$ be the state of a GFOM and let $\widehat{\boldsymbol{x}}_t$ be its tree approximation. Then $\boldsymbol{x}_t \overset{\infty}{=} \widehat{\boldsymbol{x}}_t$. In particular,*

$$\|\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_t\|_\infty = \widetilde{O}(n^{-1/2}) \text{ with high probability}. \tag{3.3}$$

Proof. We can prove $\boldsymbol{x}_t \overset{\infty}{=} \widehat{\boldsymbol{x}}_t$ inductively. By Lemma 3.9, each of the combinatorially negligible diagrams in $\boldsymbol{x}_t$ remains combinatorially negligible at time $t+1$. Meanwhile, the combinatorially non-negligible tree diagrams in $\widehat{\boldsymbol{x}}_t$ get updated to $\widehat{\boldsymbol{x}}_{t+1}$. The error bound (3.3) follows from Lemma 3.8. □

Remark 3.18. The leading order guarantee of Theorem 3.17 is best possible in general (up to logarithmic factors). Similar but more complicated equations can be given for the lower-order error terms in (3.3). For example, since the other connected diagrams with E edges and V vertices have magnitude $n^{(V-1-E)/2}$, the first lower-order term of order $n^{-1/2}$ consists of connected diagrams with exactly one cycle. The GFOM operations on this set of diagrams describe how the error evolves at this order.

Remark 3.19. One technical caveat of our analysis is that many nonlinearities used in applications are not polynomial functions (e.g. ReLU, tanh). We note that existing polynomial approximation arguments in the literature (see for example [MW25, IS24]) should apply here to prove that the tree approximation holds for GFOMs with Lipschitz denoisers f_t up to arbitrarily small $\frac{1}{\sqrt{n}}\|\cdot\|_2$ error. This is however strictly weaker than the guarantees of Theorem 3.17.

3.6. General state evolution

From the ideas established so far, we directly deduce *state evolution* for GFOM algorithms, capturing that the coordinates of $\boldsymbol{x}_t$ are asymptotically independent trajectories of an explicit random variable X_t. Recall the definition of the asymptotic state X_t from Definition 3.3.

Theorem 3.20 (General state evolution)**.** *Let t be a constant and $\boldsymbol{A} = \boldsymbol{A}(n)$ be a sequence of random matrices satisfying Assumption 2.1. Let $\boldsymbol{x}_t \in \mathbb{R}^n$ be the state of a GFOM and let X_t be the asymptotic state of $\boldsymbol{x}_t$. Then:*

1. $\frac{1}{n}\sum_{i=1}^{n} x_{t,i} \overset{\infty}{=} \mathbb{E}[X_t]$ *and therefore,*

$$\frac{1}{n}\sum_{i=1}^{n} x_{t,i} = \mathbb{E}[X_t] + \widetilde{O}(n^{-\frac{1}{2}}) \text{ with high probability}.$$

2. X_t *satisfies the explicit recurrence defined at the end of §3.2.*

Proof. This will be proven in Appendix A.5 as the following lemma.

Lemma 3.21. *Let $\boldsymbol{x}$ be a vector diagram expression with asymptotic state $X \in \Omega$. Then as scalar diagrams, $\frac{1}{n}\sum_{i=1}^{n} x_i \overset{\infty}{=} \mathbb{E}[X]$.*

For the second item, the tree approximation $\boldsymbol{x}_t = \widehat{\boldsymbol{x}}_t$ holds by Theorem 3.17. The asymptotic state X_t corresponding to $\widehat{\boldsymbol{x}}_t$ then satisfies the explicit recursion on trees presented in §3.2. □

We conclude this section by working out a few lemmas which help compute asymptotic states. We will use them in §4.4 to compute the state evolution of approximate message passing.

The set of asymptotic states Ω has an inner product coming from the expectation over the Gaussians $(Z_\sigma^\infty)_{\sigma\in\mathcal{S}}$. Since these random variables are independent Gaussians, the multivariate Hermite polynomials $(Z_\tau^\infty)_{\tau\in\mathcal{T}}$ form an orthogonal basis of Ω with respect to this inner product. Recall the + and − operators from Definition 3.1.

Fact 3.22. *+ and − are bijections between $\mathcal{T}$ and $\mathcal{S}$ which are inverses of each other and preserve $|\mathrm{Aut}(\tau)|$.*

A key observation is that X^+ is always a centered Gaussian random variable for any $X \in \Omega$, since every resulting tree is in $\mathcal{S}$.

Fact 3.23. *For all $X \in \Omega$, $(X^+)^- = X$ and $(X^-)^+$ is the orthogonal projection of X to the subspace spanned by $\mathcal{S}$.*

We deduce that + and − are adjoint operators on Ω:

Lemma 3.24. *For all $X, Y \in \Omega$, $\mathbb{E}[X^+Y] = \mathbb{E}[XY^-]$.*

Proof. Since $(Z_\tau^\infty)_{\tau\in\mathcal{T}}$ is a basis of the vector space Ω, it suffices to check this for each pair of basis elements $\tau, \rho \in \mathcal{T}$. By orthogonality, $\mathbb{E}\left[Z_{\tau^+}^\infty Z_\rho^\infty\right]$ is nonzero if and only if $\tau^+ = \rho$ and in this case it takes value $|\mathrm{Aut}(\tau^+)|$. By Fact 3.22, this occurs if and only if $\rho \in \mathcal{S}$ and $\tau = \rho^-$. Moreover, in this case the value is also $|\mathrm{Aut}(\tau^+)| = |\mathrm{Aut}(\tau)|$, as needed. □

Lemma 3.25. *For all $X, Y \in \Omega$, $\mathbb{E}[XY] = \mathbb{E}[X^+Y^+]$ and $\mathbb{E}\left[(X^-)^2\right] \leqslant \mathbb{E}\left[X^2\right]$.*

Proof. For the first statement, apply Lemma 3.24 on X and Y^+, then use Fact 3.23. For the second statement, apply Lemma 3.24 on X^- and X to get $\mathbb{E}\left[(X^-)^+ X\right] = \mathbb{E}\left[(X^-)^2\right]$. Since $(X^-)^+$ projects away some terms from X by Fact 3.23, the left-hand side is upper bounded by $\mathbb{E}\left[X^2\right]$. □

3.7. Summary

This key chapter established the *tree approximation* of $\mathfrak{S}_n$-symmetric polynomials in a Wigner matrix. Every iteration can be mapped to its idealized version by removing all its projections on cyclic diagrams. The idealized version follows a simplified Gaussian dynamic, which we connect with a infamous statistical physics method in the next chapter.

CHAPTER 4.

Rigorous Implementation of the Cavity Method

This chapter connects the tree approximation framework developed in Chapter 3 with the *cavity method*, a powerful but traditionally non-rigorous technique from statistical physics. The cavity method is widely used to analyze message-passing algorithms, and to predict quantities such as the free energy of statistical physics models.

We demonstrate that our tree approximation can be used to rigorously justify a central application of the cavity method: the *state evolution* formula for approximate message-passing (AMP) algorithms. Unlike previous rigorous approaches, which often depart significantly from the original physical intuition, our proof closely mirrors the folklore arguments from physics by justifying them line by line.

Table of contents

This chapter is based on [JP25].

4.1. Background on the cavity method

Belief Propagation (BP) and Approximate Message Passing (AMP) are the main class of nonlinear iterative algorithms that are studied using physical techniques. BP is a general tool for statistical inference on graphical models which performs exact inference when the underlying graph is a tree. The behavior of "loopy BP" on interaction graphs with cycles is more subtle; the *cavity method* can be used to predict the asymptotic dynamics of loopy BP on mean-field models (i.e. when the underlying graphical model is dense and random).

We first explain the idea behind the cavity method on the example of the replica-symmetric belief propagation iteration for the Sherrington–Kirkpatrick (SK) model, which is the original setting in which the method was conceived by Mézard, Parisi, and Virasoro [MPV87, Chapter V]. The goal here is to estimate the marginals of the following Gibbs distribution on $\boldsymbol{x} \in \{-1, 1\}^n$:

$$p(\boldsymbol{x}) \propto \exp\left(\beta \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle + h \sum_{i=1}^{n} x_i\right),$$

where $\boldsymbol{A}$ is a random symmetric matrix with i.i.d. $\mathcal{N}(0, 1/n)$ entries and $\beta, h > 0$ are fixed parameters. We will focus on a particular regime of (β, h) known as the replica-symmetric or high temperature region of the SK model.

Let $m_i = \mathbb{E}_{\boldsymbol{x}\sim p}[x_i]$. By isolating a single coordinate $i \in [n]$ and looking at the influence of other coordinates on it, Mézard, Parisi, and Virasoro derive the *cavity equations*, which are fixed-point equations approximately satisfied by m_i,

$$m_{i\to j} = f\left(\sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} m_{k\to i}\right), \quad m_i \approx f\left(\sum_{k=1}^{n} A_{ik} m_{k\to i}\right), \tag{4.1}$$

where $f(x) = \tanh(\beta x + h)$ and $m_{i\to j}$ are new variables. Algorithmically, we can think of an iterative *belief propagation* algorithm that tries to compute a solution to these equations,

$$m_{i\to j}^{t+1} = f\left(\sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} m_{k\to i}^{t}\right), \quad m_i^{t+1} = f\left(\sum_{k=1}^{n} A_{ik} m_{k\to i}^{t}\right), \tag{4.2}$$

initialized at say $m_{i\to j}^{0} = 1$. This iteration occurs on a set of *cavity messages* $m_{i\to j}$ for $i, j \in [n]$ which conceptually are "the belief of vertex i about its own value, disregarding j".

The physical techniques predict the asymptotic trajectory of the messages $m_{i\to j}^{t}$ and the outputs m_i^t in (4.2) with respect to the randomness of the matrix $\boldsymbol{A}$. They say that $\boldsymbol{m}^t$ will

have approximately independent and identically distributed entries,

$$m_i^t \sim f(Z_t)\,, \qquad \text{where } Z_t \sim \mathcal{N}(0, \sigma_t^2)\,, \quad \sigma_1^2 = 1\,, \quad \sigma_{t+1}^2 = \mathbb{E}\, f(Z_t)^2\,. \tag{4.3}$$

A heuristic replica symmetric cavity approach for proving (4.3) would go as follows. We make an *independence assumption* that the incoming terms $m_{k\to i}^t$ in the non-backtracking summation $\sum_{k=1,k\neq j}^n A_{ik} m_{k\to i}^t$ of (4.2) are independent, as if the messages were coming up from disjoint branches of a tree. By symmetry, the messages are identically distributed. Then, we appeal to the central limit theorem to deduce

$$\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m_{k\to i}^t \sim \mathcal{N}\left(0, \mathbb{E}\left[(m_{k\to i}^t)^2\right]\right)\,.$$

From here, we get that the outgoing message satisfies $m_{i\to j}^t \sim f(Z_t)$ for $Z_t \sim \mathcal{N}(0, \sigma_t^2)$ with σ_t^2 defined by the recurrence in (4.3). Using a similar argument, we get $m_i^t \sim f(Z_t)$.

[MPV87] also derived from (4.1) a simpler form of self-consistent equations involving only the marginals themselves, known as the Thouless–Anderson–Palmer equations [TAP77],

$$m_i \approx f\left(\sum_{\substack{k=1\\k\neq i}}^n A_{ik} m_k - \beta\left(1 - \frac{1}{n}\sum_{k=1}^n m_k^2\right) m_i\right)\,. \tag{4.4}$$

The subtracted term on the right-hand side in which m_i re-occurs is the *Onsager reaction term*. In the same way that belief propagation (4.2) tries to compute solutions to the cavity equations (4.1), an *approximate message passing* algorithm can be iterated to compute approximate solutions to (4.4),

$$m_i^{t+1} = f\left(\sum_{\substack{k=1\\k\neq i}}^n A_{ik} m_k^t - \beta\left(1 - \frac{1}{n}\sum_{k=1}^n (m_k^t)^2\right) m_i^{t-1}\right)\,. \tag{4.5}$$

The approximate equivalence between the BP iteration (4.2) and the AMP iteration (4.5) is a folklore cavity method argument which we elaborate next.

4.2. Equivalence between message-passing iterations

Belief propagation. We consider BP iterations on $\boldsymbol{A}$ of the form

$$m^0_{i\to j} = 1\,, \qquad m^t_{i\to j} = f_t\left(\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m^{t-1}_{k\to i}, \dots, \sum_{\substack{k=1\\k\neq j}}^n A_{ik} m^0_{k\to i},\ m^0_{i\to j}\right), \tag{4.6}$$
$$m^t_i = \widetilde{f}_t\left(\sum_{k=1}^n A_{ik} m^{t-1}_{k\to i}, \dots, \sum_{k=1}^n A_{ik} m^0_{k\to i},\ m^0_{i\to j}\right),$$

for a sequence of functions $f_t, \widetilde{f}_t\colon \mathbb{R}^{t+1} \to \mathbb{R}$. (4.6) is a generalization of (4.2) to iterations "with memory" i.e. that can use all the previous messages. At any timestep t, the $(m^t_{i\to j})_{1\leqslant i,j\leqslant n}$ are *cavity messages* that try to compute some information about the i-th variable by ignoring the edge between i and j, while the $(m^t_i)_{1\leqslant i\leqslant n}$ are the output of the algorithm.

Approximate message passing. On the other side, we have an *approximate message passing* (AMP) algorithm of the form

$$\boldsymbol{w}^0 = \mathbf{1}\,, \qquad \boldsymbol{w}^{t+1} = \boldsymbol{A} f_t(\boldsymbol{w}^t, \dots, \boldsymbol{w}^0) - \sum_{s=1}^t b_{s,t} f_{s-1}(\boldsymbol{w}^{s-1}, \dots, \boldsymbol{w}^0)\,, \tag{4.7}$$
$$\boldsymbol{m}^t = \widetilde{f}_t(\boldsymbol{w}^t, \dots, \boldsymbol{w}^0)\,, \tag{4.8}$$

where $b_{s,t}$ is defined to be the scalar quantity

$$b_{s,t} = \frac{1}{n}\sum_{i=1}^n \frac{\partial f_t}{\partial w^s}(w^t_i, \dots, w^0_i)\,.$$

One practical advantage of AMP compared to BP is that is has a smaller number of messages to track, $O(n)$ vs $O(n^2)$.

Theorem 4.1 (Equivalence of BP and AMP)**.** *Let $T \geqslant 1$ be a constant independent of n, $f_t, \widetilde{f}_t\colon \mathbb{R}^{t+1} \to \mathbb{R}$ for $t \leqslant T$ be a sequence of polynomials independent of n, and $\boldsymbol{A} = \boldsymbol{A}(n)$ be a sequence of random matrices satisfying Assumption 2.1. Generate $\boldsymbol{m}^{t,\mathrm{BP}}$ according to (4.6) and $\boldsymbol{m}^{t,\mathrm{AMP}}$ according to (4.8). Then*

$$\boldsymbol{m}^{t,\mathrm{AMP}} \overset{\infty}{=} \boldsymbol{m}^{\mathrm{BP}}\,,$$

so in particular, with high probability,

$$\|\boldsymbol{m}^{t,\mathrm{AMP}} - \boldsymbol{m}^{t,\mathrm{BP}}\|_\infty = \widetilde{O}(n^{-1/2})\,.$$

4.2.1. Heuristic derivation of Theorem 4.1

The equivalence between BP and AMP is folklore in the statistical physics community, thanks to the following simple cavity-based reasoning. It can be found for example in the seminal paper [DMM09, §A] or the survey [ZK16, §IV.E].

We start by rewriting the BP iteration, letting $\boldsymbol{w}^0 = \mathbf{1}$ and $w_i^{t+1} = \sum_{k=1}^n A_{ik} m_{k\to i}^t$. The output of BP is computed as

$$m_i^{t+1} = \widetilde{f}_{t+1}\left(w_i^{t+1}, \dots, w_i^0\right) .$$

Hence it suffices to show that $\boldsymbol{w}^t$ asymptotically follows the AMP iteration (4.7). First, (4.6) can be rewritten

$$m_{i\to j}^{t+1} = f_{t+1}\left(w_i^{t+1} - A_{ij} m_{j\to i}^t, \dots, w_i^1 - A_{ij} m_{j\to i}^0, w_i^0\right) .$$

Given that the entries A_{ij} are on the scale of $1/\sqrt{n}$, which we expect to be much smaller than the magnitude of the messages, we perform a first-order Taylor approximation (the partial derivatives are with respect to the coordinates of f_{t+1} and the last coordinate is ignored because w_i^0 is constant):

$$m_{i\to j}^{t+1} \approx f_{t+1}\left(w_i^{t+1}, \dots, w_i^1, w_i^0\right) - A_{ij} \sum_{s=1}^{t+1} m_{j\to i}^{s-1} \frac{\partial f_{t+1}}{\partial w^s}\left(w_i^{t+1}, \dots, w_i^1, w_i^0\right) . \tag{$*$}$$

Plugging this approximation in the definition of w_i^{t+1},

$$\begin{aligned} w_i^{t+1} &\approx \sum_{k=1}^n A_{ik} f_t(w_k^t, \dots, w_k^0) - \sum_{k=1}^n A_{ik}^2 \sum_{s=1}^t m_{i\to k}^{s-1} \frac{\partial f_t}{\partial w^s}(w_k^t, \dots, w_k^0) \\ &\approx \sum_{k=1}^n A_{ik} f_t(w_k^t, \dots, w_k^0) - \sum_{k=1}^n \frac{1}{n} \sum_{s=1}^t f_{s-1}(w_i^{s-1}, \dots, w_i^0) \frac{\partial f_t}{\partial w^s}(w_k^t, \dots, w_k^0) \qquad (**) \\ &= \sum_{k=1}^n A_{ik} f_t(w_k^t, \dots, w_k^0) - \sum_{s=1}^t b_{s,t} f_{s-1}(w_i^{s-1}, \dots, w_i^0) . \end{aligned}$$

This shows that w_i^{t+1} approximately satisfies the AMP recursion (4.7), as desired.

The intuition behind $(**)$ is that because we are summing over k, we may expand A_{ik}^2 and $m_{i\to k}^{s-1}$ on the first order and replace them by averages which do not depend on k:

$$\begin{aligned} A_{ik}^2 &\approx \mathbb{E}\left[A_{ik}^2\right] = \frac{1}{n} , \\ m_{i\to k}^{s-1} &= f_{s-1}\left(w_i^{s-1} - A_{ik} m_{k\to i}^t, \dots, w_i^1 - A_{ik} m_{k\to i}^0, w_i^0\right) \\ &\approx f_{s-1}\left(w_i^{s-1}, \dots, w_i^0\right) . \end{aligned}$$

4.2.2. Diagram proof of Theorem 4.1

In fact, the previous heuristic argument can be made directly rigorous by working with the tree approximation. It suffices to justify (∗) and (∗∗) in order to prove Theorem 4.1.

The BP iteration takes place on $\boldsymbol{m}^t \in \mathbb{R}^{n^2}$ instead of $\mathbb{R}^n$ which is not captured by our previous definitions. Most of the work below is setting up definitions to fit this iteration into our framework. We define diagrams for vectors $\boldsymbol{x} \in \mathbb{R}^{n(n-1)}$ whose (i, j) entry is written $x_{i\to j}$ (for simplicity, we assume $A_{ii} = 0$ so that the messages $m^t_{i\to i}$ can be ignored).

Definition 4.2 (Cavity diagrams)**.** A cavity diagram is an unlabeled undirected graph $\alpha = (V(\alpha), E(\alpha))$ with two distinct, ordered root vertices ⓞⓞ. No vertices may be isolated except for the roots.

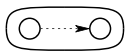
For any cavity diagram α, we define $\boldsymbol{Z}_\alpha \in \mathbb{R}^{n(n-1)}$ by

$$Z_{\alpha,i\to j} = \sum_{\substack{\varphi\colon V(\alpha)\to[n] \\ \varphi \text{ injective} \\ \varphi(ⓞⓞ)=(i,j)}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u),\varphi(v)}\,,$$

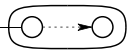
for any distinct $i, j \in [n]$.

Below is the representation of the first iterate of (4.6) with cavity diagrams. In the pictures, we draw an arrow from the first root to the second root to indicate the order. If a (multi)edge exists in the graph between the roots, then the arrow is on the edge; otherwise we use a dashed line to indicate that there is no edge.

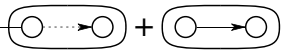
$$m^0_{i\to j} =$$



$$\sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} m^0_{k\to i} =$$



$$\sum_{k=1}^{n} A_{ik} m^0_{k\to i} =$$



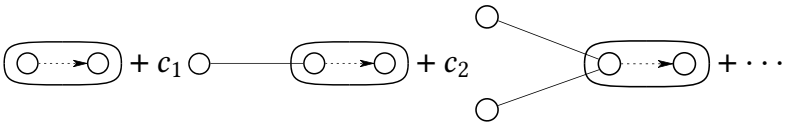
Multiplying $A_{ik} m^t_{k\to i}$ creates a new edge between k and i in $m^t_{k\to i}$. Summing over k "unroots" the first root. A case distinction needs to be made in the summation depending on if $k = i$ or $k = j$ or $k \notin \{i, j\}$. The case $k = i$ is ignored assuming that $A_{ii} = 0$. The case $k = j$ yields the "backward step" while the remaining case $k \neq j$ is the "forward step".

To apply f_1, we need to multiply $i \to j$ diagrams componentwise, which is achieved by fixing/merging the roots i, j and summing over the part outside the roots. For some

coefficients $c_0, c_1, c_2, \ldots$ we have[1]

$$m^1_{i\to j} = f_1\left(\sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} m^0_{k\to i}\right)$$

$$= c_0 \quad + c_1 \quad + c_2 \quad + \cdots$$



The output m_i^{t+1} uses the non-cavity quantities $\sum_{k=1}^n A_{ik} m^t_{k\to i}$. The cavity diagrams are converted back to the usual diagram basis as follows.

Claim 4.3 (Conversion of cavity diagrams)**.** *For any cavity diagram α and $i \in [n]$,*

$$\sum_{j=1}^{n} A_{ij} Z_{\alpha, j\to i} = Z_{\alpha', i}\,,$$

where α' is the diagram (in the sense of Definition 2.4) obtained from α by adding an edge between the two roots of α and unrooting the first root.

Since the final output is computed by converting all cavity diagrams back to regular diagrams using the previous claim, the definition of combinatorial negligibility and the $\overset{\infty}{=}$ notation can be extended to cavity diagrams. We make the following definitions.

Definition 4.4. A cavity diagram α is combinatorially negligible if the diagram α' obtained in Claim 4.3 is combinatorially negligible. We naturally extend the $\overset{\infty}{=}$ notation to cavity diagrams as in Definition 3.7.

Claim 4.5. *Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be in the span of the cavity diagrams such that $\boldsymbol{x} \overset{\infty}{=} \boldsymbol{x}'$. If we let*

$$y_{i\to j} = \sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} x_{k\to i}\,, \qquad y'_{i\to j} = \sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} x'_{k\to i}\,,$$

then $\boldsymbol{y} \overset{\infty}{=} \boldsymbol{y}'$.

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_t$ are in the span of cavity diagrams, $\boldsymbol{x}_i \overset{\infty}{=} \boldsymbol{x}'_i$ for all $i \in [n]$, and $f\colon \mathbb{R}^t \to \mathbb{R}$ is a polynomial function applied componentwise, then

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t) \overset{\infty}{=} f(\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_t)\,.$$

Claim 4.5 follows directly from Lemma 3.9.

This completes the diagrammatic description of the belief propagation algorithm. We are now ready to rigorously justify the approximations made during the heuristic argument.

[1] The exact values of the coefficients c_i are not necessary to compute.

Lemma 4.6 ((∗)).

$$m^t_{i\to j} \overset{\infty}{=} f_t\left(w^t_i,\dots,w^0_i\right) - A_{ij}\sum_{s=1}^{t} m^{s-1}_{j\to i}\frac{\partial f_t}{\partial w^s}\left(w^t_i,\dots,w^0_i\right) .$$

Proof. Since f_t is a polynomial, it has an exact Taylor expansion. The terms of degree higher than 1 in the Taylor expansion create at least 2 edges between the roots i and j. All cavity diagrams with 2 edges between the roots are combinatorially negligible because the unrooting operation of Claim 4.3 adds one more edge between i and j, and diagrams with multiedges of multiplicity > 2 are combinatorially negligible (Lemma 3.13). □

Lemma 4.7 ((∗∗)).

$$\sum_{k=1}^{n} A^2_{ik} m^{s-1}_{i\to k}\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k) \overset{\infty}{=} \frac{1}{n} f_{s-1}(w^{s-1}_i,\dots,w^0_i)\sum_{k=1}^{n}\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k) .$$

Proof. First, we argue about the replacement of $m^{s-1}_{i\to k}$. We have

$$m^{s-1}_{i\to k} = f_{s-1}\left(\sum_{\substack{\ell=1\\ \ell\neq k}}^{n} A_{i\ell} m^{s-2}_{\ell\to i},\ \dots,\ \sum_{\substack{\ell=1\\ \ell\neq k}}^{n} A_{i\ell} m^{0}_{\ell\to i},\ m^0_{i\to k}\right) .$$

The difference between this and $f_{s-1}(w^{s-1}_i,\dots,w^0_i)$ are the backtracking terms $A_{ik}m^r_{k\to i}$. All terms in the entire Taylor expansion of the polynomial on the right-hand side around $w^{s-1}_i,\dots,w^0_i$ will introduce at least one additional factor of A_{ik}, which combines with the A^2_{ik} present in the summation over k to become a negligible multiplicity > 2 edge (Lemma 3.13). This shows that

$$\sum_{k=1}^{n} A^2_{ik} m^{s-1}_{i\to k}\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k) \overset{\infty}{=} f_{s-1}(w^{s-1}_i,\dots,w^0_i)\sum_{k=1}^{n} A^2_{ik}\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k) . \tag{4.9}$$

Second, we argue about the replacement of A^2_{ik}. This double edge is only non-negligible if it is hanging (Lemma 3.13). Among the diagrams in $\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^t_0)$ the only one which does not attach something to k is the singleton diagram ◎. The coefficient of this diagram is the expected value (Corollary 2.10),

$$\mathbb{E}\left[\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k)\right] .$$

The expected value is equal to the empirical expectation up to negligible terms (Lemma 3.21),

$$\mathbb{E}\left[\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k)\right] \overset{\infty}{=} \frac{1}{n}\sum_{k=1}^{n}\frac{\partial f_t}{\partial w^s}(w^t_k,\dots,w^0_k) .$$

This implies

$$\sum_{k=1}^{n} A_{ik}^2 \frac{\partial f_t}{\partial w^s}(w_k^t, \dots, w_k^0) \overset{\infty}{=} \frac{1}{n} \sum_{k=1}^{n} \frac{\partial f_t}{\partial w^s}(w_k^t, \dots, w_k^0)\,. \tag{4.10}$$

The desired statement follows from combining (4.9) and (4.10). □

Proof of Theorem 4.1. Replace the $\approx$ signs in the heuristic argument from §4.2.1 by $\overset{\infty}{=}$ and use Claim 4.5 repeatedly. □

4.3. Proving the cavity assumptions

We examine the belief propagation iteration (4.6) more closely. The BP iterates have the following asymptotic structure.

Lemma 4.8. *$m_{i\to j}^t$ is asymptotically equivalent to a linear combination of cavity diagrams which have a tree hanging off of i, no edges between the roots, and nothing attached to j.*

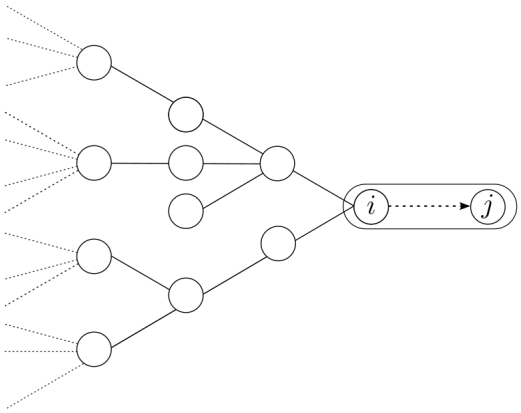


Figure 4.1. Diagram representation of the cavity messages $m_{i\to j}^t$. Each cavity diagram in the asymptotic cavity diagram representation of $m_{i\to j}^t$ is a tree rooted at i.

Proof. Let α be a cavity diagram with the stated form. The vector whose (i,j)-th entry is $\sum_{k=1}^n A_{ik} Z_{\alpha,k\to i}$ is the sum of the diagrams which add an edge between the roots of α, that can be of 3 types: (1) the "forward step" diagram which puts the j root as a new vertex, (2) the "backtracking step" diagram which interchanges the first and second roots of α, and (3) other diagrams where j intersects with a vertex from $V(\alpha) \setminus \{i\}$.

All diagrams of type (3) are negligible (and stay so when applying further operations to them), because they create a cycle of length > 2. The backtracking step in (2) is canceled by

summing over $k \neq j$ in the belief propagation iteration. What asymptotically remains is the forward step (1) which again has the stated form. Additionally, componentwise functions preserve the stated form. □

Theorem 4.9. *For any $j \in [n]$, the incoming messages at j, $\{m^t_{i\to j} : i \in [n] \setminus \{j\}\}$, are asymptotically independent (Definition 3.15).*

Proof. When j is ignored, the cavity diagrams in the asymptotic representation of $m^t_{i\to j}$ in Lemma 4.8 are equivalent to non-cavity diagrams (replacing n by $n-1$). From the classification theorem (Theorem 3.14), these are asymptotically independent. □

4.4. State evolution formula for BP/AMP

We show how to simplify the asymptotic state appearing in Theorem 3.20 for the special case of approximate message passing. Recall the + and − operators from §3.2.

Theorem 4.10 (Asymptotic state for AMP)**.** *Under the same assumptions as Theorem 4.1, the asymptotic state of $(\boldsymbol{w}_t)_{t\leqslant T}$ satisfies the recursion*

$$W_0 = 1\,, \qquad W_{t+1} = f_t(W_t, \dots, W_0)^+\,. \tag{4.11}$$

In particular, W_t is a centered Gaussian and for all $s, t \leqslant T$, the covariances are

$$\mathbb{E}\left[W_{s+1}W_{t+1}\right] = \mathbb{E}\left[f_s(W_s, \dots, W_0)f_t(W_t, \dots, W_0)\right]\,.$$

Combining Theorem 4.10 and part (ii) of Theorem 3.20 recovers the typical formulation of state evolution for AMP algorithms. We note that while the formula for computing iterates of AMP (4.8) might look mysterious at first sight, the AMP recursion in the asymptotic space (4.11) is much easier to interpret.

We now prove Theorem 4.10. Note that (4.7) is not directly captured by the definition of a GFOM because $b_{s,t}$ requires computing an average over coordinates. This is only a technical issue: by Lemma 3.21, empirical expectations are concentrated up to combinatorially negligible terms. Hence, the following inductive definition of a GFOM for $\boldsymbol{w}_t \in \mathbb{R}^n$ and its corresponding asymptotic state W_t is asymptotically equivalent to (4.7):

$$\boldsymbol{w}_0 = \mathbf{1}\,, \quad \boldsymbol{w}_{t+1} = \boldsymbol{A}f_t(\boldsymbol{w}_t, \dots, \boldsymbol{w}_0) - \sum_{s=1}^{t} \mathbb{E}\left[\frac{\partial f_t}{\partial w_t}(W_t, \dots, W_0)\right] f_{s-1}(\boldsymbol{w}_{s-1}, \dots, \boldsymbol{w}_0)\,. \tag{4.12}$$

The Onsager correction term in (4.12) will be rigorously interpreted as a backtracking term using diagrams.

Lemma 4.11. *Let $W_1, \dots, W_t \in \Omega$ be Gaussian (i.e. each W_s is in the span of $(Z_\sigma^\infty)_{\sigma \in \mathcal{S}}$). Then for any polynomial function $f : \mathbb{R}^t \to \mathbb{R}$,*

$$f(W_1, \dots, W_t)^- = \sum_{s=1}^{t} \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] W_s^- .$$

Proof. Expand $f(W_1, \dots, W_t)$ as

$$\begin{aligned} f(W_1, \dots, W_t) &= \sum_{\sigma \in \mathcal{S}} c_\sigma Z_\sigma^\infty + \sum_{\tau \in \mathcal{T} \setminus \mathcal{S}} c_\tau Z_\tau^\infty , \\ f(W_1, \dots, W_t)^- &= \sum_{\sigma \in \mathcal{S}} c_\sigma Z_{\sigma^-}^\infty , \end{aligned}$$

for some coefficients $c_\tau \in \mathbb{R}$. When $\sigma \in \mathcal{S}$, we have

$$\begin{aligned} c_\sigma \, |\mathrm{Aut}(\sigma)| &= \mathbb{E}\left[Z_\sigma^\infty f(W_1, \dots, W_t)\right] && \text{(orthogonality)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_\sigma^\infty W_s\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] && \text{(Lemma A.5)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_{\sigma^-}^\infty W_s^-\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] . && \text{(Lemma 3.24)} \end{aligned}$$

The second expectation does not depend on σ. Summing the first expectation over σ produces W_s^- as desired. □

Now we complete the proof of Theorem 4.10.

Proof of Theorem 4.10. We prove by induction on t that $W_{t+1} = f_t(W_t, \dots, W_0)^+$. For $t = 0$, we have $\boldsymbol{w}_1 = \boldsymbol{A} f_0(\mathbf{1})$ so $W_1 = f_0(W_0)^+$ and the statement holds.

Now suppose that the statement holds for $W_1, \dots, W_t$ for some $t < T$. The asymptotic state of $\boldsymbol{A} f_t(\boldsymbol{w}_t, \dots, \boldsymbol{w}_0)$ is $f_t(W_t, \dots, W_0)^+ + f_t(W_t, \dots, W_0)^-$. By the induction hypothesis and Fact 3.23, for any $s \leqslant t$,

$$W_s^- = f_{s-1}(W_{s-1}, \dots, W_0) .$$

Combining this with Lemma 4.11, we see that the asymptotic state of the Onsager correction term equals $f_t(W_t, \dots, W_0)^-$. This concludes the induction.

In particular, $W_{t+1} = f_t(W_t, \dots, W_0)^+$ has no constant term and is in the span of $\mathcal{S}$, so it has a centered Gaussian distribution. The covariances are, for all $s, t \leqslant T$,

$$\begin{aligned} \mathbb{E}\left[W_{s+1} W_{t+1}\right] &= \mathbb{E}\left[f_s(W_s, \dots, W_0)^+ f_t(W_t, \dots, W_0)^+\right] \\ &= \mathbb{E}\left[f_s(W_s, \dots, W_0) f_t(W_t, \dots, W_0)\right] , \end{aligned}$$

where the last equality follows from Lemma 3.25. This completes the proof. □

4.5. Summary

This chapter demonstrated that the cavity method from physics — at least, its algorithmic applications — is correct simply because every heuristic part of the argument holds "up to cyclic diagrams".

CHAPTER 5.

Beyond a Constant Number of Iterations

In summary, we have so far described the trajectory of first-order algorithms for a *constant* number of iterations. We now turn to the question of how these algorithms behave when the number of iterations scales with the dimension n of the matrix.

A primary motivation is to understand the convergence of iterative methods: whether they approach a fixed point, or continue searching indefinitely without success.

A second motivation is to study algorithms with a warm start, such as spectral initialization [MV21, MV22, LW22]. If the initialization step itself can be implemented via a first-order method (e.g., power iteration), we may hope to analyze the composite algorithm using our diagrammatic framework. This approach is demonstrated for constant-depth composition in Figure 5.1.

Table of contents

This chapter (except §5.5) is based on [JP25].

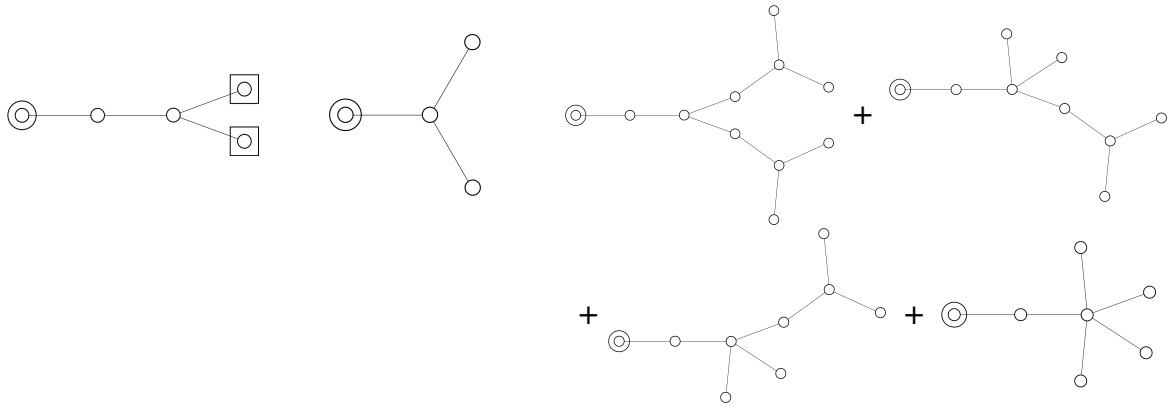


Figure 5.1. Composing diagram representations. The leaves of a diagram access the entries of the input, so we draw a box around each one to indicate that the input's entries are not yet fixed. When another diagram is used as input, it is duplicated at each leaf. The treelike diagrams in the result are a sum over i of contracting i path edges from both sides of the merged root/leaf. Note that here the second and third output diagram are the same.

5.1. Combinatorial phase transitions

In order to show that that the long-time behavior is a delicate question, we will compute in §5.4 that some diagrams of $\omega(1)$ size are no longer asymptotically Gaussian, breaking the classification Theorem 3.14. Larger-degree vertices in a diagram can access high moments of the entries of other diagrams, which will detect that these quantities are not exactly Gaussian.

However, in typical first-order algorithms, high-degree diagrams only appear in a controlled way. Thus we expect that for a class of "nice" GFOMs, the Gaussian tree approximation continues to hold for many more iterations. To demonstrate this, we examine *debiased power iteration*, which is the iterative algorithm

$$\boldsymbol{x}_0 = \mathbf{1}\,, \quad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\,. \tag{5.1}$$

(5.1) has a very simple tree approximation (the t-path diagram). Due to its basic representation in the tree space, this algorithm will re-appear in §5.5 and §6.2.2.

Note that by Theorem 4.1, for constantly many iterations this algorithm is asymptotically equivalent to power iteration on the non-backtracking walk matrix, which is the algorithm

$$\boldsymbol{m}_0 = \mathbf{1}\,, \qquad \boldsymbol{m}_{t+1} = \boldsymbol{B}\boldsymbol{m}_t\,,$$

$$x_{t+1,i} = \sum_{k=1}^{n} A_{ik} m_{t,k\to i}\,,$$

where $\boldsymbol{B} \in \mathbb{R}^{n^2\times n^2}$ is the weighted non-backtracking walk matrix defined by $B_{i\to j,k\to \ell} = A_{k\ell}$ if $j = k$ and $i \neq \ell$, and $B_{i\to j,k\to \ell} = 0$ otherwise.

We distinguish several regimes of $T = T(n)$ depending on the obstacles that arise when trying to generalize the tree approximation for (5.1) to a larger number of iterations.

- When $T \ll \frac{\log n}{\log\log n}$, we expect the proofs of Theorem 3.14 and Theorem 3.17 to generalize with minimal changes. The total number of diagrams that arise can be bounded by $T^{O(T)}$ which is $n^{o(1)}$ in this regime.
- When $T \approx \frac{\log n}{\log\log n}$, there are $T^{O(T)} = \mathrm{poly}(n)$ many diagrams to keep track of. This could overpower the magnitude of some cyclic diagrams, and make the naive union bound argument fail. This barrier is also the one of previous non-asymptotic analyses of AMP [RV18, CR24].
- When $T \ll n^{\delta}$ for some small constant $\delta > 0$, we show in the next sections that the tree approximation of debiased power iteration still holds by a more careful accounting of the error terms. We predict that this can be extended up to $T \ll \sqrt{n}$.
- When $T \approx \sqrt{n}$, the tree diagrams with T vertices are exponentially small in magnitude (see Lemma 2.9) and the number of non-tree diagrams starts to become overwhelmingly large. At the conceptual level, random walks of length $\geqslant \sqrt{n}$ in an n-vertex graph are likely to collide. Therefore, it is unclear whether or not the tree diagrams of size $\geqslant \sqrt{n}$ are significantly different from diagrams with cycles. This threshold also appears in recent analyses of AMP [LFW23], although it is not a barrier for their result.

5.2. Analyzing power iteration via combinatorial walks

For constantly many iterations of debiased power iteration, by Theorem 3.17, we know that $\boldsymbol{x}_t$ is well-approximated by the t-path diagram, denoted $\boldsymbol{Z}_{t\text{-path}}$. Here we prove that this approximation holds much longer. To simplify the calculation, we assume:

Assumption 5.1. Let $\boldsymbol{A}$ be a random $n \times n$ symmetric matrix with $A_{ii} = 0$ and A_{ij} drawn independently from the uniform distribution over $\left\{-\frac{1}{\sqrt{n-1}}, \frac{1}{\sqrt{n-1}}\right\}$ for all $i < j$.

We prove that for this iterative algorithm we can extend Theorem 3.17 to a polynomial number of iterations, hence overcoming some obstructions mentioned in §5.1. A similar argument can also show that $\boldsymbol{Z}_{t\text{-path}}$ remain approximately independent Gaussians for t in the same regime. Taken together, we see that the "usual" state evolution formula for constantly many iterations continues to hold much longer, up to conjecturally $\sqrt{n}$ iterations.

Theorem 5.2. *Suppose that $\boldsymbol{A} = \boldsymbol{A}(n)$ satisfies Assumption 5.1 and generate $\boldsymbol{x}_t$ according to (5.1). Then there exist universal constants $c, \delta > 0$ such that for all $t \leqslant cn^{\delta}$,*

$$\left\|\boldsymbol{x}_t - \boldsymbol{Z}_{t\text{-path}}\right\|_{\infty} \xrightarrow{\text{a.s.}} 0\,.$$

To obtain the tree approximation of algorithms with poly(n) many iterations, we need to very carefully count combinatorial factors that were neglected in §3.2. The total number of diagrams in the unapproximated diagram expansion is very large, and furthermore, each diagram can arise in many different ways if it has high-degree vertices. To perform the analysis, we decompose $\boldsymbol{x}_t$ in terms of walks of length t; we need to track walks instead of diagrams so that we do not throw away additional information about high-degree vertices.

Our goal is to show that the walk without any back edge (the t-path) dominates asymptotically. We will proceed as in the proof of Theorem 3.14 by bounding the q-th moment of $\boldsymbol{x}_t - \boldsymbol{Z}_{t\text{-path}}$. This moment can be represented diagrammatically using q-tuples of non-backtracking walks with at least one back edge.

Definition 5.3. A (q,t)-traversal $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)$ is an ordered sequence of q walks, each of length t and starting from the same vertex:

$$\boldsymbol{\gamma}_i = (\{u_{i,1} = \circledcirc, u_{i,2}\}, \{u_{i,2}, u_{i,3}\}, \ldots, \{u_{i,t}, u_{i,t+1}\})\,, \quad \text{for all } i \in [q].$$

Each traversal $\boldsymbol{\gamma}$ is naturally associated to an improper diagram $(V(\boldsymbol{\gamma}), E(\boldsymbol{\gamma}))$ with $V(\boldsymbol{\gamma}) = \{u_{i,j} : i \in [q], j \in [t]\}$ and $E(\boldsymbol{\gamma}) = \{(u_{i,j}, u_{i,j+1}) : i \in [q], j \in [t-1]\}$ (viewed as a multiset). We use the notation $Z_{\boldsymbol{\gamma}} = Z_{(V(\boldsymbol{\gamma}),E(\boldsymbol{\gamma}))}$ following Definition 2.5.

- A traversal is even if each edge appears an even number of times in $\bigcup_{i\in[q]} \boldsymbol{\gamma}_i$.
- A traversal is non-backtracking if every walk of the traversal is non-backtracking, i.e. $u_{i,j+1} \neq u_{i,j-1}$ for all $i \in [q]$ and $j \in \{2, \ldots, t-1\}$.
- A traversal is non-full-forward if every walk of the traversal has a back edge, namely for all $i \in [q]$, there exist $j_1 \neq j_2$ such that $u_{i,j_1} = u_{i,j_2}$.

Let $\mathcal{W}_t^q$ be the set of (q,t)-traversals that are simultaneously even, non-backtracking, non-full-forward, and have no self-loops.

Definition 5.3 is motivated by the following decomposition:

Claim 5.4. *Suppose that $\boldsymbol{x}_t$ is generated according to (5.1) and $\boldsymbol{A}$ satisfies Assumption 5.1. Then,*

$$\mathbb{E}\left[(\boldsymbol{x}_t - \boldsymbol{Z}_{t\text{-path}})^q\right] = \sum_{\boldsymbol{\gamma}\in\mathcal{W}_t^q} \mathbb{E}\left[Z_{\boldsymbol{\gamma}}\right]\,.$$

We now proceed to proving Theorem 5.2. We will bound the magnitude of $\mathbb{E}\left[Z_{\boldsymbol{\gamma},i}\right]$ for $\boldsymbol{\gamma} \in \mathcal{W}_t^q$, then count the number of traversals in $\mathcal{W}_t^q$. Both bounds will depend on $\frac{E}{2} - V + 1$

(where V is the number of vertices of the traversal and E the number of edges), which quantifies how close the traversal is to a tree of double edges.

Our first insight is that the traversals contributing to $(\boldsymbol{x}_t - \boldsymbol{Z}_{t\text{-path}})^q$ become further from trees as q increases because each walk must have a back edge.

Lemma 5.5. *For any $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ with V vertices and E edges, $\frac{E}{2} - V + 1 \geqslant \frac{q}{2}$.*

Proof. Assign to each vertex all the edges going into it in $\boldsymbol{\gamma}$. Each non-root vertex must have at least 2 incoming edges: the edge which explores it, and since $\boldsymbol{\gamma}$ is even and non-backtracking, an edge which revisits it a second time. Since $\boldsymbol{\gamma}$ is non-full-forward, each $\boldsymbol{\gamma}_i$ has a back edge; the first back edge in each $\boldsymbol{\gamma}_i$ yields an additional incoming edge for each i (either it points to the root, which has not yet been counted, or by assumption that it is the *first* back edge in $\boldsymbol{\gamma}_i$, it cannot cover both incident edges from the first visit). We have

$$E \geqslant 2(V-1) + q\,,$$

as needed. □

Lemma 5.6. *For any $i \in [n]$ and $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ with V vertices and E edges,*

$$\left|\mathbb{E}\left[Z_{\boldsymbol{\gamma},i}\right]\right| \leqslant O\left(n^{-\left(\frac{E}{2}-V+1\right)}\right).$$

Proof. Using Assumption 5.1, we can directly count

$$\begin{aligned}\left|\mathbb{E}\left[Z_{\boldsymbol{\gamma},i}\right]\right| &\leqslant O(1) \cdot \frac{(n-1)(n-2)\cdots(n-V+1)}{n^{\frac{E}{2}}} \\ &= O\left(n^{V-1-\frac{E}{2}}\right).\end{aligned}$$

□

Finally, the following lemma captures the counting of traversals. Its proof is deferred to the next section.

Lemma 5.7. *The number of $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ with V vertices and E edges is at most*

$$O_q(t)^{6\left(\frac{E}{2}-V+1\right)+2q},$$

where $O_q(\cdot)$ hides a constant depending only on q.

Proof of Theorem 5.2. We decompose the sum over $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ according to the value of $r = \frac{E}{2} - V + 1$ using Lemma 5.6 and Lemma 5.7:

$$\mathbb{E}\left[(x_{t,i} - Z_{t\text{-path},i})^q\right] \leqslant O_q(t)^{2q} \sum_{r \geqslant \frac{q}{2}} O_q(t)^{6r} n^{-r}\,.$$

If t satisfies $t \leqslant cn^{\delta}$ with $0 < \delta < 1/6$, the sum is a geometrically decreasing series and therefore it is bounded by the first term which is $O_q(t^{5q} n^{-\frac{q}{2}})$. Under the condition $\delta < 1/10$, for q being a large enough integer we obtain for some $\varepsilon > 0$,

$$\mathbb{E}\left[(x_{t,i} - Z_{t\text{-path},i})^q\right] \leqslant O(1/n^{2+\varepsilon}) .$$

This is enough to imply that $\|\boldsymbol{x}_t - \boldsymbol{Z}_{t\text{-path}}\|_\infty \xrightarrow{\text{a.s.}} 0$ by a union bound over the n coordinates, then Markov's inequality and the Borel-Cantelli lemma. □

5.3. Counting combinatorial walks

Our goal here is to prove Lemma 5.7.

In the extreme case $V \approx \frac{E}{2}$ where the moment bound Lemma 5.6 is the weakest, typical traversals $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ look like trees of double edges with a constant number of back edges. In this regime, most vertices will have degree exactly 4. Following this intuition, our encoding will proceed by compressing the long paths of degree-4 vertices connected by double edges.

Definition 5.8. For $\boldsymbol{\gamma} \in \mathcal{W}_t^q$, let $\boldsymbol{\gamma}_c$ be the traversal obtained by replacing all maximally long paths of degree-4 vertices in $\boldsymbol{\gamma}$ by a single special marked edge between the endpoints of the paths, and removing the internal vertices of the path. (The paths should be broken at the root so that it is not removed.)

Note that these operations can create self-loops in $\boldsymbol{\gamma}_c$.

Lemma 5.9. *For any $\boldsymbol{\gamma} \in \mathcal{W}_t^q$,*

$$|E(\boldsymbol{\gamma}_c)| \leqslant 3|E(\boldsymbol{\gamma})| - 6(|V(\boldsymbol{\gamma})| - 1) + 2q .$$

Proof. For $k \in \mathbb{N}$, let $V_k(\boldsymbol{\gamma})$ be the set of non-root vertices of $\boldsymbol{\gamma}$ of degree exactly k. Since $\boldsymbol{\gamma}$ is an even traversal, we get by double counting the number of edges in $\boldsymbol{\gamma}$

$$2|V_2(\boldsymbol{\gamma})| + 4|V_4(\boldsymbol{\gamma})| + 6\left(|V(\boldsymbol{\gamma})| - |V_2(\boldsymbol{\gamma})| - |V_4(\boldsymbol{\gamma})| - 1\right) \leqslant 2|E(\boldsymbol{\gamma})| .$$

Moreover, the number of edges removed during the compression is $2|V_4(\boldsymbol{\gamma})|$. This means that

$$|E(\boldsymbol{\gamma})| - |E(\boldsymbol{\gamma}_c)| = 2|V_4(\boldsymbol{\gamma})| \geqslant 6(|V(\boldsymbol{\gamma})| - 1) - 4|V_2(\boldsymbol{\gamma})| - 2|E(\boldsymbol{\gamma})| .$$

Finally, since $\boldsymbol{\gamma}$ is non-backtracking, non-root degree-2 vertices can only be created in $\boldsymbol{\gamma}$ by pairing endpoints of the walks, so that $|V_2(\boldsymbol{\gamma})| \leqslant q/2$. The desired inequality immediately follows. □

We are now ready to prove Lemma 5.7.

Proof of Lemma 5.7. We encode a traversal $\boldsymbol{\gamma} \in \mathcal{W}_t^q$ as follows:

1. We first encode $\boldsymbol{\gamma}_c$. We write down the sequence of vertices of each walk and indicate whether each step should be the first step of a marked edge (Definition 5.8). Every time we traverse a marked edge for the second time, instead of recording the next vertex of the walk, we record the identifier of the marked edge. We also add a single bit of information to each edge to indicate whether it is the last edge of its walk. The target space of the encoding has size $O(|E(\boldsymbol{\gamma}_c)|)^{|E(\boldsymbol{\gamma}_c)|}$.
2. We then expand the marked edges in $\boldsymbol{\gamma}_c$ of which there are at most $|E(\boldsymbol{\gamma}_c)|/2$. For each marked edge, we write down the length of the path that it replaced. This can be encoded using "stars and bars". Initially allocating 2 edges to each marked edge, there are at most $\binom{E}{|E(\boldsymbol{\gamma}_c)|/2}$ such objects.

We claim that this encoding allows to reconstruct $\boldsymbol{\gamma}$, and its length can be bounded by

$$\begin{aligned} O(|E(\boldsymbol{\gamma}_c)|)^{|E(\boldsymbol{\gamma}_c)|}\binom{E}{|E(\boldsymbol{\gamma}_c)|/2} &\leqslant O(|E(\boldsymbol{\gamma}_c)|)^{|E(\boldsymbol{\gamma}_c)|} O\left(\frac{E}{|E(\boldsymbol{\gamma}_c)|}\right)^{|E(\boldsymbol{\gamma}_c)|/2} \\ &= O_q(t)^{|E(\boldsymbol{\gamma}_c)|}\,. \end{aligned}$$

The proof follows after plugging in the bound of Lemma 5.9. □

5.4. High-degree tree diagrams are not Gaussian

Care must be taken when studying the diagrams of superconstant size. In this section we compute that the star-shaped diagram with $\log n$ leaves and the root at a leaf is not Gaussian (its fourth moment is significantly larger than the square of its second moment).[1] This diagram appears after only $T = O(\log \log n)$ iterations in the recursion

$$\boldsymbol{x}_1 = \boldsymbol{A}\mathbf{1} \qquad \boldsymbol{x}_{t+1} = (\boldsymbol{x}_t)^2 \qquad \boldsymbol{x}_{T+1} = \boldsymbol{A}\boldsymbol{x}_T\,.$$

However, we expect that this diagram does not contribute significantly to nicer GFOMs that strictly alternate between multiplication by $\boldsymbol{A}$ and constant-degree componentwise operations.

Fixing d, let γ denote (d-star graph)$^+$. We compute that $\mathbb{E}\left[Z_{\gamma,1}^4\right] \gg \mathbb{E}\left[Z_{\gamma,1}^2\right]^2$ when $d \approx \log n$. By Lemma 2.9, the variance is

$$\mathbb{E}\left[Z_{\gamma,1}^2\right] = (1 + o(1))\,|\mathrm{Aut}(\gamma)| = (1 + o(1))d!\,.$$

[1] Similarly, adding an edge between two of the leaves creates a cyclic diagram with negligible variance but non-negligible fourth moment.

When computing the fourth moment $\mathbb{E}\left[Z_{\gamma,1}^4\right]$ for constant d, the terms that are dominant consist of (1) a perfect matching between the four edges incident to the root, (2) perfect matchings between their d children. There are $3(d!)^2$ such terms, recovering the fourth moment of a Gaussian with variance $d!$.

For $d = \log n$, another type of term becomes dominant. These are the terms where all four edges incident to the root are equal, then we have a perfect matching on $4d$ objects divided into four groups of size d such that no two objects from the same group are matched. Denote the latter set of matchings by $\mathcal{M}(d, d, d, d)$.

Lemma 5.10. *Up to a multiplicative* $\mathrm{poly}(d)$ *factor,* $|\mathcal{M}(d, d, d, d)| \gtrsim 3^d (d!)^2$.

These terms come with a $\frac{1}{n}$ factor due to the multiplicity 4 edge. When $d = \Omega(\log n)$, the extra factor of 3^d overpowers the $\frac{1}{n}$ and makes the fourth moment much larger than the the squared variance $(d!)^2$.

Proof of Lemma 5.10. We establish a recursion. There are $(3d)(3d-1)\cdots(2d+1)$ ways to match up the objects in the first group, which can be partitioned in $O(d^2)$ ways depending on how many objects in each other group are matched. We will recurse on the "maximum-entropy" case in which the first group matches $d/3$ elements from each other group, using the following claim.

Claim 5.11. *Let* $d, k \in \mathbb{N}$ *such that* $\frac{d}{k-1}$ *is an integer. Counting the matchings between* d *objects and a subset of* $(k-1)d$ *objects in* $k-1$ *groups, as a function of the number of objects matched in each group, the number of matchings is maximized when there are* $\frac{d}{k-1}$ *matched elements per group.*

Proof of Claim 5.11. Letting $n_1, \ldots, n_{k-1}$ be the number of matched elements per group, we may directly compute this number as

$$\prod_{i=1}^{k-1} (d)_{n_i}$$

where $(d)_k = d(d-1)\cdots(d-k+1)$ is the falling factorial. When n_i and n_j are replaced by $n_i - 1$ and $n_j + 1$, the ratio of new to old values is

$$\frac{d - n_j}{d - n_i + 1}$$

which is at least 1 if $n_i \geqslant n_j + 1$. Hence the n_i are equal at the maximum. □

Using Claim 5.11, up to a factor of $O(d^2)$,

$$|\mathcal{M}(d, d, d, d)| \gtrsim (3d)(3d-1)\cdots(2d+1)|\mathcal{M}(2d/3, 2d/3, 2d/3)|$$

$$\asymp \left(\frac{3d}{e}\right)^{3d} \left(\frac{e}{2d}\right)^{2d} |\mathcal{M}(2d/3, 2d/3, 2d/3)|$$

where the second equality holds up to a $\mathrm{poly}(d)$ factor by Stirling's approximation:

Fact 5.12 (Stirling's approximation). *Up to a multiplicative* $\mathrm{poly}(d)$ *factor,* $d! \asymp \left(\frac{d}{e}\right)^d$.

Recursing via the same principle,

$$\begin{aligned}
|\mathcal{M}(2d/3, 2d/3, 2d/3)| &\gtrsim (4d/3)(4d/3-1)\cdots(2d/3+1)|\mathcal{M}(d/3, d/3)| \\
&= (4d/3)(4d/3-1)\cdots(2d/3+1)(d/3)! \\
&\asymp \left(\frac{4d}{3e}\right)^{4d/3} \left(\frac{3e}{2d}\right)^{2d/3} \left(\frac{d}{3e}\right)^{d/3}
\end{aligned}$$

where the last equation follows from Fact 5.12. In total,

$$|\mathcal{M}(d, d, d, d)| \gtrsim 3^d \left(\frac{d}{e}\right)^{2d}. \qquad \square$$

5.5. The BBP transition

In this final section, we speculate on how Theorem 5.2 could be used to recover the lower bound in the *BBP transition* of the spiked Wigner model.

The goal in the spiked Wigner model (Example 2.3) is to understand under what conditions on λ it is possible to recover u from an observation of $\boldsymbol{A}$ alone. Since the spectrum of $\boldsymbol{W}$ is contained in $[-2 - o(1), 2 + o(1)]$ with high probability, one may predict that $\boldsymbol{A}$ will exhibit an outlier eigenvalue when $\lambda \geqslant 2$. Perhaps surprisingly, this phenomenon already occurs when $\lambda \geqslant 1$. This marks the onset of the BBP transition, named after Baik, Ben Arous, and Péché's work [BBP05] on the analogous transition in the Wishart setting.

Theorem 5.13 (BBP transition for Wigner matrices). *Let* $\boldsymbol{A}$ *be drawn from the spiked Wigner model, i.e.,*

$$\boldsymbol{A} = \lambda \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{W},$$

where $\boldsymbol{W}$ *satisfies Assumption 5.1 and* $\boldsymbol{u}$ *is uniformly random on the sphere* S^{n-1}. *Let* $\lambda_{\max}$ *and* $\boldsymbol{u}_{\max}$ *denote the largest eigenvalue of* $\boldsymbol{A}$ *and its corresponding (unit) eigenvector.*

1. *If* $\lambda < 1$, *then* $\langle \boldsymbol{u}, \boldsymbol{u}_{\max}\rangle^2 \to 0$ *and* $\lambda_{\max} = 2 + o(1)$.
2. *If* $\lambda \geqslant 1$, *then as* $n \to \infty$, *with high probability,*

$$\begin{aligned}
\langle \boldsymbol{u}, \boldsymbol{u}_{\max}\rangle^2 &= 1 - \frac{1}{\lambda^2} + o(1), \\
\lambda_{\max} &= \lambda + \frac{1}{\lambda} + o(1).
\end{aligned}$$

The eigenvalue BBP transition for the spiked Wigner model was proved by Féral and Péché via the trace method [FP07]. An alternative, algebraic proof appears in [BN11, §4.1].

We now sketch a constructive approach to Theorem 5.13 by analyzing the following nonbacktracking power iteration on $\boldsymbol{A}$:

$$\boldsymbol{x}_0 = \mathbf{1}\,, \quad \boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{x}_0\,, \quad \forall t \geqslant 2,\ \boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\,. \tag{5.2}$$

While we expect similar arguments to work for related iterative schemes, (5.2) has a particularly simple representation in the Fourier diagram basis: asymptotically, it corresponds to the length-t path diagram. Unlike in the null model, where certain algorithms converge to an approximate top eigenvector in a constant number of iterations (see §6.1.1), in the spiked model, $\Theta(\log n)$ iterations are required.

Our first lemma gives an exact decomposition of the iterates of (5.2) in terms of iterates under the null model. The proof is a straightforward induction, which we omit.

Claim 5.14. *Let*

$$\begin{aligned} \boldsymbol{Q}_{-1} = \mathbf{0}\,, \quad \boldsymbol{Q}_0 = \mathbf{1}\,, \quad \boldsymbol{Q}_{t+1} = \boldsymbol{W}\boldsymbol{Q}_t - \boldsymbol{Q}_{t-1}\,, \\ \boldsymbol{P}_{-1} = \mathbf{0}\,, \quad \boldsymbol{P}_0 = \boldsymbol{u}\,, \quad \boldsymbol{P}_{t+1} = \boldsymbol{W}\boldsymbol{P}_t - \boldsymbol{P}_{t-1}\,. \end{aligned}$$

Then for every $t \geqslant 1$,

$$\boldsymbol{x}_t = \lambda \langle \boldsymbol{x}_{t-1}, \boldsymbol{u} \rangle \boldsymbol{u} + \lambda \sum_{s=1}^{t-1} \langle \boldsymbol{x}_{t-s-1}, \boldsymbol{u} \rangle \boldsymbol{P}_s + \boldsymbol{Q}_t\,.$$

Claim 5.14 shows that to understand the behavior of $\boldsymbol{x}_t$, it suffices to understand the sequences $\boldsymbol{P}_t$ and $\boldsymbol{Q}_t$, which are iterations in the null model $\boldsymbol{W}$. By Theorem 5.2, the vectors $\boldsymbol{Q}_t$ are well-approximated by their tree approximation (the length-t path diagram) even when t is polynomial in n.

We conjecture that the same holds for $\boldsymbol{P}_t$ as well.

Definition 5.15. Define the tree approximations of $\boldsymbol{Q}_t$ and $\boldsymbol{P}_t$ as

$$\widehat{\boldsymbol{Q}}_t := \sum_{\boldsymbol{i}\colon [t+1] \hookrightarrow [n]} \prod_{j=1}^{t} A_{i(j),i(j+1)}\,, \qquad \widehat{\boldsymbol{P}}_t := \sqrt{n} \cdot \sum_{\boldsymbol{i}\colon [t+1] \hookrightarrow [n]} u_{i(t+1)} \prod_{j=1}^{t} A_{i(j),i(j+1)}\,.$$

Conjecture 5.16. *Suppose $\boldsymbol{A}$ satisfies Assumption 2.1. For any $t = t(n)$ with*

$$\limsup_{n\to\infty} t(n)/\log n < \infty$$

we have, with high probability,

$$\max_{s \leqslant t} \left\| \boldsymbol{Q}_s - \widehat{\boldsymbol{Q}}_s \right\|_\infty = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right), \ \max_{s \leqslant t} \left\| \sqrt{n} \cdot \boldsymbol{P}_s - \widehat{\boldsymbol{P}}_s \right\|_\infty = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right).$$

and

$$\max_{\substack{s,s' \leqslant t \\ s \neq s'}} \left\langle \widehat{P}_s, \widehat{P}_{s'} \right\rangle = \widetilde{O}\left(\sqrt{n}\right) , \max_{\substack{s,s' \leqslant t \\ s \neq s'}} \left\langle \widehat{Q}_s, \widehat{Q}_{s'} \right\rangle = \widetilde{O}\left(\sqrt{n}\right) , \max_{s,s' \leqslant t} \left\langle \widehat{P}_s, \widehat{Q}_{s'} \right\rangle = \widetilde{O}\left(\sqrt{n}\right) .$$

Note that these conditions are natural generalizations to the regime $t = \omega(1)$ of the fact that distinct Fourier diagrams are approximately orthonormal with high probability.

Let us now explain how Theorem 5.13 is related to Conjecture 5.16. The key idea is that Conjecture 5.16 implies that the decomposition in Claim 5.14 is approximately orthogonal, allowing us to explicitly compute the norm and projections of $\boldsymbol{x}_t$. For example, this reasoning yields

$$\begin{aligned} \langle \boldsymbol{x}_t, \boldsymbol{u} \rangle^2 &\approx \lambda^2 \langle \boldsymbol{x}_{t-1}, \boldsymbol{u} \rangle^2 , \\ \|\boldsymbol{x}_t\|_2^2 &\approx \lambda^2 \sum_{s=1}^{t-1} \langle \boldsymbol{x}_s, \boldsymbol{u} \rangle^2 + n , \end{aligned}$$

where the $\approx$ sign hides multiplicative factors of the form $1 + \widetilde{O}(1/\sqrt{n})$. Since with a random initialization, we have $\langle \boldsymbol{x}_0, \boldsymbol{u} \rangle^2 \approx 1/n$, if $\lambda > 1$, then after $t = C \log n$ steps for a large constant C, we obtain

$$\frac{\langle \boldsymbol{x}_t, \boldsymbol{u} \rangle^2}{\|\boldsymbol{x}_t\|_2^2} \approx \frac{\lambda^{2t}}{\lambda^2 \cdot \frac{\lambda^{2t}}{\lambda^2 - 1}} = 1 - \frac{1}{\lambda^2} ,$$

which matches the eigenvector overlap in Theorem 5.13. Establishing Conjecture 5.16 is an interesting direction for future work.

5.6. Summary

In this chapter, we showed that the tree approximation remains valid for $\omega(1)$ iterations when approximating the top eigenvector of a matrix. We also discussed a potential application of these ideas to recovering a phase transition in random matrix theory. Similarly, in the next chapter, we will be able to implement our idea for optimization in the null model, which only requires a constant number of iterations.

Part II.

Polynomial Optimization

CHAPTER 6.

On Optima of Polynomials

This transition chapter illustrates how the concept of *tree approximation*, developed in Part I, can be leveraged to analyze the optima of random polynomials. As a testbed, we study the problem of maximizing a random quadratic polynomial over the ℓ_p-ball:

$$\max_{\|\boldsymbol{x}\|_p \leqslant 1} p(\boldsymbol{x}), \qquad p(\boldsymbol{x}) := \sum_{i,j=1}^{n} c_{ij} x_i x_j, \tag{6.1}$$

whose coefficients are i.i.d. normalized Gaussians.

For $p = 2$, (6.1) converges to 2 as $n \to \infty$. We first show that the basic power method converges to this value in $\Theta(n^{2/3})$ steps. We then prove the main consequence of Part I, Theorem 6.9, which reduces lower bounds on quadratic optimization to a combinatorial problem in the tree basis. We explicitly solve it for $p = 2$, and re-interpret Montanari's algorithm [Mon19] for $p = \infty$. Finally, we initiate our study of worst-case counterparts by characterizing the optimal value of arbitrary quadratic polynomial optimization.

Table of contents

The results of this chapter are unpublished.

6.1. Tight analysis of power iteration

We aim here at comparing iterative algorithms for the simplest random optimization problem one can think of,

$$\begin{aligned} \max \qquad & |\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle| \\ \text{s.t.} \qquad & \boldsymbol{x} \in \mathbb{R}^n, \quad \|\boldsymbol{x}\|_2 = 1 \end{aligned} \tag{6.2}$$

where $\boldsymbol{A}$ is an $n \times n$ Wigner matrix with entries of variance $1/n$. This is the spectral norm of $\boldsymbol{A}$, and a standard trace method argument shows that the optimum of (6.2) is $2 + o(1)$ with high probability, as $n \to \infty$.

The textbook algorithm for computing the spectral norm of a matrix is the *power method*, which is nothing else than gradient descent with infinite step size applied to (6.2), i.e.,

$$\boldsymbol{x}_0 = \mathbf{1}\,, \qquad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t\,. \tag{6.3}$$

A folklore argument shows that this algorithm achieves $(1 - O(\varepsilon))$-approximation to (6.2) in $O(\log n/\varepsilon)$ iterations when $\boldsymbol{A}$ is a positive semidefinite matrix [Tre17b, Theorem 9.6]. However, this is not true anymore when $\boldsymbol{A}$ has both positive and negative eigenvalues, because of cancellations happening in the spectrum.

Our main result in this section is that when $\boldsymbol{A}$ is a Wigner matrix, power iteration typically converges in $\approx n^{2/3}$ iterations.

We first give an elementary argument showing the much weaker bound that power iteration does not converge within a constant (independent of n) number of iterations. While the proof of the following proposition is easy and does not need our entire theory, it will serve as an introduction for the methodology we will use for other examples. It also points out at the key symmetry of power iteration that is responsible for its slow behavior. As we will see later, algorithms that break this symmetry converge in constantly many iterations.

Proposition 6.1. *Suppose that $\boldsymbol{A}$ satisfies Assumption 2.1. Then for any constant $T \geqslant 0$ independent of n, the T-th iterate of (6.3) satisfies*

$$\frac{|\langle \boldsymbol{x}_T, \boldsymbol{A}\boldsymbol{x}_T\rangle|}{\|\boldsymbol{x}_T\|_2^2} = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

with high probability over $\boldsymbol{A}$.

Proof. By Definition 3.3, $\boldsymbol{x}_t$ has asymptotic state X_t defined by induction by

$$X_0 = 1\,, \qquad X_t = X_{t-1}^{+} + X_{t-1}^{-}\,.$$

First, using Theorem 3.17,

$$\frac{1}{n}\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t\rangle = 2\,\mathbb{E}\left[X_t X_t^+\right] + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

with high probability. Similarly,

$$\frac{1}{n}\langle \boldsymbol{x}_t, \boldsymbol{x}_t\rangle = \mathbb{E}\,X_t^2 + \widetilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

with high probability.

Moreover, for any $T \geqslant 0$, we have $X_T = \sum_{t=0}^{T} c_t P_t$, where P_t is the length-t path diagram, and c_t counts the number of Dyck paths[1] starting from $(0,0)$ and ending at (T,t).

Since $(P_1, P_2, \ldots, P_T)$ are orthonormal for $\mathbb{E}$ (they are simply different diagrams), we have:

1. $\mathbb{E}\,X_T^2 = \sum_{t=0}^{T} c_t^2$ is counting the number of Dyck paths from $(0,0)$ to $(2T,0)$, which is exactly the Catalan number C_T.
2. $\mathbb{E}\left[X_T X_T^+\right] = \sum_{t=0}^{T-1} c_t c_{t-1} = 0$ because $c_t = 0$ when t and T have different parity.

Putting everything together, we get

$$\frac{|\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t\rangle|}{\|\boldsymbol{x}_t\|_2^2} = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right),$$

with high probability over $\boldsymbol{A}$, as desired. □

Remark 6.2. Note that in the previous proof, we can make the coefficients c_t of power iteration in the Fourier diagram basis explicit using the reflection principle. More precisely, by reflecting a non-Dyck path at the first hitting time of $y = -1$, we get a bijection to arbitrary paths starting from $(0,0)$ and ending at $(T, -(t+2))$. Therefore,

$$c_t = \begin{cases} \binom{T}{\frac{T+t}{2}} - \binom{T}{1+\frac{T+t}{2}} & \text{if } t \text{ and } T \text{ have the same parity} \\ 0 & \text{otherwise} \end{cases}$$

Although we took a more explicit route, the proof can be summarized by the fact that $\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t\rangle$ is the average entry of the vector $\boldsymbol{x}_t \odot \boldsymbol{A}\boldsymbol{x}_t$, which like any $\mathfrak{S}_n$-symmetric polynomial in $\boldsymbol{A}$, has asympotically independent entries. Therefore, it is very concentrated around its expectation, and $|\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t\rangle|$ has to concentrate around the same value. The fact that $\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t\rangle$ has expectation 0 can be seen more directly: since power iteration does not

[1] A Dyck path is a path in the (x,y)-plane with allowed steps $(x,y) \to (x+1, y+1)$ and $(x,y) \to (x+1, y-1)$, and whose y-coordinates are all non-negative.

break the symmetry between the positive and negative eigenspace of $\boldsymbol{A}$, the distribution of $\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t \rangle$ has to be symmetric around the origin.

We conclude that the special symmetry of power iteration is responsible for its dimension-dependent complexity. This phenomenon is well-known in statistical physics, and suggests a simple modification of power iteration.

Our main theorem in this subsection is the following. It generalizes Proposition 6.1 from $t = O(1)$ to $t \ll n^{2/3}$. For technical reasons, we assume now that $\boldsymbol{A}$ is a GOE matrix, but we expect that small modifications to our approach would apply to Wigner matrices as well.

Theorem 6.3. *Let A be a GOE matrix (i.e., $A_{ij} = A_{ji} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$ for all $i < j$, and $A_{ii} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/n)$). Then for any $\varepsilon > 0$ and $t \leqslant O(n^{\frac{2}{3}-\varepsilon})$, the t-th iterate of (6.3) satisfies*

$$\frac{|\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t \rangle|}{\|\boldsymbol{x}_t\|_2^2} \leqslant \widetilde{O}\left(\frac{t^{3/4}}{\sqrt{n}}\right) = o(1)\,,$$

with probability $1 - o(1)$.

Our proof of Theorem 6.3 is based on the eigenvalue rigidity of Wigner matrices, first established in [EYY12]. This is a uniform concentration result on the location of the eigenvalues.

Definition 6.4. For all $i \in [n]$, let $\lambda_i^* \in [-2, 2]$ be the classical location of the i-th eigenvalue of a Wigner matrix as predicted by the semicircle law, i.e., the unique solution to

$$\frac{1}{2\pi} \int_{\lambda_i^*}^{2} \sqrt{4 - x^2}\, \mathrm{d}x = \frac{i}{n}\,.$$

The following theorem is a corollary of [EYY12, Theorem 2.2]. Note that the uniform concentration of the *bulk* eigenvalues can be improved to $\widetilde{O}(1/n)$, but we will not need this stronger fact here.

Theorem 6.5. *Let $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ be the eigenvalues of a Wigner matrix. Then for any $\varepsilon > 0$, we have*

$$\max_{i \in [n]} \left|\lambda_i - \lambda_i^*\right| \leqslant n^{-\frac{2}{3}+\varepsilon}\,,$$

with probability $1 - o(1)$.

We will also use the following simple estimate on the classical locations of the semicircle eigenvalues.

Lemma 6.6. *If $t \ll n^{2/3}$, then*

$$\frac{1}{4^t} \sum_{i=1}^{n} \left(\lambda_i^*\right)^{2t} \asymp \frac{n}{t^{3/2}}\,. \tag{6.4}$$

Proof of Lemma 6.6. The asymptotics trivially hold when t is a constant, so we can assume without loss of generality that t is a large enough constant. As $\varepsilon \to 0$, we have

$$\frac{1}{\sqrt{2\pi}} \int_{2-\varepsilon}^{2} \sqrt{4-x^2}\,\mathrm{d}x = \frac{1}{\sqrt{\pi}} \int_0^{\varepsilon} \sqrt{u} \cdot \sqrt{1-\frac{u}{2}}\,\mathrm{d}u = \Theta\left(\varepsilon^{3/2}\right).$$

Equivalently, there exist constants $C_1, C_2 > 0$ such that for $\varepsilon > 0$ small enough,

$$C_1\varepsilon^{3/2} \leqslant \frac{1}{\sqrt{2\pi}} \int_{2-\varepsilon}^{2} \sqrt{4-x^2}\,\mathrm{d}x \leqslant C_2\varepsilon^{3/2}.$$

Therefore, for small enough $\varepsilon > 0$, we have that $i \in [n]$ satisfies $\frac{i}{n} \leqslant C_1\varepsilon^{3/2}$, then $\lambda_i^* \geqslant 2-\varepsilon$, and if $\frac{i}{n} \geqslant C_2\varepsilon^{3/2}$, then $\lambda_i^* \leqslant 2-\varepsilon$. This means than the number of $i \in [n]$ such that $\lambda_i^* \geqslant 2-\varepsilon$ is at least $C_1 n\varepsilon^{3/2}$ and at most $C_2 n\varepsilon^{3/2}$. Up to a factor 2, the same holds for the number of $i \in [n]$ such that $\left|\lambda_i^*\right| \geqslant 2-\varepsilon$.

We deduce that when t is a large enough constant,

$$\sum_{i=1}^{n} \left(\lambda_i^*\right)^{2t} \gtrsim \frac{n}{t^{3/2}} \cdot \left(2-\frac{1}{t}\right)^{2t} \gtrsim \frac{4^t n}{t^{3/2}}.$$

For the other direction, we decompose the sum as

$$\sum_{i=1}^{n} \left(\lambda_i^*\right)^{2t} = \sum_{j\geqslant 0} \sum_{\substack{i\in[n] \\ 2-\left|\lambda_i^*\right| \in [2^{-(j+1)}, 2^{-j})}} \left(\lambda_i^*\right)^{2t}.$$

Applying the above estimate on the number of eigenvalues ε-close to 2 for small enough ε, there exists a universal constant $c > 0$ such that

$$\begin{aligned}
\sum_{i=1}^{n} \left(\lambda_i^*\right)^{2t} &\lesssim \sum_{j\geqslant 0} \left(2-2^{-j}\right)^{2t} n 2^{-3(j+1)/2} + n(2-c)^t \\
&\leqslant 4^t n \sum_{j\geqslant 0} \left(1-2^{-j+1}\right)^{2t} 2^{-3(j+1)/2} + n(2-c)^t \\
&\lesssim \frac{4^t n}{t^{3/2}}.
\end{aligned}$$

This finishes the proof. □

Theorem 6.7 (Bernstein inequality on the sphere)**.** *Let x be uniformly random vector on the sphere of radius $\sqrt{n}$. Then for any $\boldsymbol{a} \in \mathbb{R}^n$ and $t \geqslant 0$,*

$$\Pr\left(\left|\sum_{i=1}^{n} a_i\left(x_i^2-1\right)\right| \geqslant t\right) \leqslant \exp\left(-C\min\left(n, \frac{t^2}{\|\boldsymbol{a}\|_2^2}, \frac{t}{\|\boldsymbol{a}\|_\infty}\right)\right).$$

Proof of Theorem 6.7. Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. By rotational invariance, the random variable $\sqrt{n} \cdot \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$ is uniformly distributed on the sphere of radius $\sqrt{n}$. With probability $1 - e^{-\Omega(n)}$, we have $\|\boldsymbol{g}\|_2 \leqslant O(\sqrt{n})$, and the result follows from Bernstein inequality applied to $\sum_i a_i g_i^2$. □

Proof of Theorem 6.3. Let $\delta \in (0,1)$. The expression $\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t \rangle = \left\langle \mathbf{1}, \boldsymbol{A}^{2t+1}\mathbf{1} \right\rangle$ is odd in $\boldsymbol{A}$, so up to replacing $\boldsymbol{A}$ by $-\boldsymbol{A}$ and using a union bound, it suffices to bound

$$p(t) \coloneqq \Pr\left(\langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t \rangle \geqslant \delta \|\boldsymbol{x}_t\|_2^2\right),$$

We start by rewriting the expression in an orthogonal eigenbasis $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ of $\boldsymbol{A}$, corresponding to eigenvalues $(\lambda_1, \dots, \lambda_n)$. Let $\alpha_i \coloneqq \langle \boldsymbol{x}_0, \boldsymbol{u}_i \rangle$. When $\boldsymbol{A}$ is a GOE matrix, by rotational invariance, the eigenbasis can be chosen so that $\boldsymbol{U} = [\boldsymbol{u}_1 \mid \dots \mid \boldsymbol{u}_n]$ is has a random orthogonal matrix (Haar measure in $\mathcal{O}(n)$). In this case, $\boldsymbol{\alpha} = \boldsymbol{U}^\top \mathbf{1}$ is uniformly distributed on the sphere of radius $\sqrt{n}$.[2]

Next, we have

$$\boldsymbol{x}_t = \boldsymbol{A}^t \mathbf{1} = \sum_{i=1}^n \lambda_i^t \alpha_i \boldsymbol{u}_i,$$

so we can rewrite equivalently

$$p(t) = \Pr\left(\sum_{i=1}^n \lambda_i^{2t+1} \alpha_i^2 \geqslant \delta \sum_{i=1}^n \lambda_i^{2t} \alpha_i^2\right) = \Pr\left(\sum_{i=1}^n \left(\lambda_i^{2t+1} - \delta \lambda_i^{2t}\right) \alpha_i^2 \geqslant 0\right). \tag{6.5}$$

We next show that we can replace λ_i by λ_i^* in this expression as long as $t \ll n^{2/3}$. Applying Theorem 6.5, we know that for all $i \in [n]$ such that $\left|\lambda_i^*\right| \geqslant 0.1$ and for any $\varepsilon > 0$, we have

$$\lambda_i^{2t} = \left(\lambda_i^*\right)^{2t} \left(1 \pm O\left(t n^{-\frac{2}{3}+\varepsilon}\right)\right), \quad \lambda_i^{2t+1} = \left(\lambda_i^*\right)^{2t+1} \left(1 \pm O\left(t n^{-\frac{2}{3}+\varepsilon}\right)\right).$$

On the other hand, the indices $i \in [n]$ such that $\left|\lambda_i^*\right| \leqslant 0.1$ have a negligible contribution to the sum in (6.5),

$$\sum_{\substack{i \in [n] \\ \left|\lambda_i^*\right| \leqslant 0.1}} \left|\lambda_i^{2t+1} - \delta \lambda_i^{2t}\right| \alpha_i^2 \leqslant \max_{\substack{i \in [n] \\ \left|\lambda_i^*\right| \leqslant 0.1}} \left|\lambda_i^{2t+1} - \delta \lambda_i^{2t}\right| \leqslant \eta \coloneqq O(1) \cdot 0.1^{2t+1}.$$

After reparametrizing $\tilde{\delta} \coloneqq \delta - O\left(t n^{-\frac{2}{3}-\varepsilon}\right)$, we have

$$p(t) \leqslant \Pr\left(\sum_{i=1}^n \left(\left(\lambda_i^*\right)^{2t+1} - \tilde{\delta}\left(\lambda_i^*\right)^{2t}\right) \alpha_i^2 \geqslant -\eta\right).$$

[2] This is the step that crucially uses the assumption that $\boldsymbol{A}$ is a GOE matrix.

Let $c_i := \left(\lambda_i^*\right)^{2t+1} - \tilde{\delta}\left(\lambda_i^*\right)^{2t}$ and $f(\alpha) := \sum_{i=1}^n c_i \alpha_i^2$. Assume that n is odd for simplicity. The sequence $(\lambda_i^*)_{i\in[n-1]}$ is symmetric around 0, and since $\mathbb{E}\,\alpha_i^2 = 1$,

$$\mathbb{E}\, f = -\tilde{\delta}\sum_{i=1}^{n-1}\left(\lambda_i^*\right)^{2t} - 2^{2t}(2-\tilde{\delta}) \leqslant -\tilde{\delta}\sum_{i=1}^{n}\left(\lambda_i^*\right)^{2t}. \tag{6.6}$$

We have $\eta + \mathbb{E}\, f < 0$ when t is a large enough constant, so by Theorem 6.7,

$$\begin{aligned} p(t) &\leqslant \Pr\left(f(\alpha) - \mathbb{E}\, f \geqslant -\mathbb{E}\, f - \eta\right) \\ &\leqslant \exp\left(-C\min\left(n, \frac{(\mathbb{E}\, f+\eta)^2}{\|\boldsymbol{c}\|_2^2}, \frac{-\mathbb{E}\, f-\eta}{\|\boldsymbol{c}\|_\infty}\right)\right). \end{aligned}$$

Moreover, by (6.6), as long as $\tilde{\delta} \geqslant 0.1^{-t}$, we have

$$\frac{(\mathbb{E}\, f+\eta)^2}{\|\boldsymbol{c}\|_2^2} \gtrsim \frac{\left(\tilde{\delta}\sum_{i=1}^n\left(\lambda_i^*\right)^{2t} - \eta\right)^2}{\sum_{i=1}^n\left(\lambda_i^*\right)^{4t}} \gtrsim \tilde{\delta}^2 \cdot \frac{n}{t^{3/2}},$$

where the last step follows from applying Lemma 6.6 to both the numerator and the denominator. Applying similarly Lemma 6.6 to the other term, if $\tilde{\delta} \geqslant 0.1^{-t}$,

$$\frac{-\mathbb{E}\, f-\eta}{\|\boldsymbol{c}\|_\infty} \gtrsim \frac{\tilde{\delta}\sum_{i=1}^n\left(\lambda_i^*\right)^{2t} - \eta}{4^t} \gtrsim \tilde{\delta}\cdot\frac{n}{t^{3/2}}.$$

Putting everything together, we get that as long as $\tilde{\delta} \geqslant 0.1^{-t}$ and $t \ll n^{2/3}$, we have $p(t) \leqslant e^{-\Omega(\tilde{\delta}^2 n/t^{3/2})}$. This goes to 0 if $\delta \gg t^{3/4}/\sqrt{n}$, as desired. □

Remark 6.8. The analysis of Theorem 6.3 is essentially tight. This is because the gap between the largest eigenvalue in magnitude and the second largest eigenvalue in magnitude of a Wigner matrix scales like $n^{2/3}$. As witnessed in our proof of Theorem 6.3, the Rayleigh quotient achieved by power iteration is driven by the ratio

$$\frac{\left|\sum_{i=1}^n \lambda_i^{2t+1}\right|}{\sum_{i=1}^n \lambda_i^{2t}}$$

which gets close to 2 whenever t is large compared to the aforementioned eigenvalue gap.

6.1.1. Symmetry-breaking power iteration

One may initially wonder whether the dependency of the runtime on the spectral gap is necessary for estimating the maximal eigenvalue. We show that this is not the case, and

that a very minor variant of the basic power method dramatically improves the speed of convergence.

Consider

$$\boldsymbol{x}_0 = \mathbf{1}\,, \qquad \boldsymbol{x}_{t+1} = (\boldsymbol{I} + \boldsymbol{A})\boldsymbol{x}_t$$

While $\boldsymbol{A}$ and $\boldsymbol{I} + \boldsymbol{A}$ have the same eigenvectors, the spectrum of $\boldsymbol{I} + \boldsymbol{A}$ is not symmetric around $\mathbf{0}$.

We can reproduce the proof of Proposition 6.1 in this setting. $\boldsymbol{x}_t$ has asymptotic state X_t, where

$$X_0 = 1\,, \qquad X_{t+1} = X_t^+ + X_t^- + X_t\,.$$

Therefore, we can decompose $X_t = \sum_{t=0}^{T} c_t P_t$ in the Fourier diagram basis, where P_t is the length-t path diagram, and c_t counts the number of Motzkin paths[3] starting at $(0,0)$ and ending at (T,t). Then

1. $\mathbb{E}\, X_T^2 = \sum_{t=0}^{T} c_t^2$ counts the number of Motzkin paths from $(0,0)$ to $(2T,0)$. This is exactly the Motzkin number M_{2T} (OEIS sequence A026945).
2. $\mathbb{E}\left[X_T X_T^+\right] = \sum_{t=0}^{T-1} c_t c_{t+1}$ counts the number of Motzkin paths from $(0,0)$ to $(2T+1,0)$ whose T-th increment is $(+1,+1)$. Call N_T this number. By reflecting a path around $x = T/2$, we get that this also counts the number of Motzkin paths whose T-th increment is $(+1,-1)$. Therefore, $2N_T = M_{2T+1} - M_{2T}$, since M_{2T} is the number of Motzkin paths of length $2T+1$ where the T-th increment is $(+1,0)$.

In summary, the value achieved by this algorithm is

$$2 \cdot \frac{M_{2T+1} - M_{2T}}{2M_{2T}} = \frac{M_{2T+1}}{M_{2T}} - 1\,.$$

Standard asymptotics of Motzkin numbers [FS09, Example VI.3] ensure that

$$M_n = C \cdot \frac{3^{n+1}}{n^{3/2}} \left(1 + O\left(\frac{1}{n}\right)\right),$$

for some constant $C > 0$, so that $M_{2T+1}/M_{2T} = 3 + O(1/T)$, and so the algorithm achieves value $2 - O(1/T)$ as $T \to \infty$.

6.2. Optimization in the tree basis

The algorithm from the previous section achieves the optimal value as $T \to \infty$. We now present an alternative route to design the algorithm which achieves the *best possible* value as $n \to \infty$, among all iterative algorithms running for any fixed number of T iterations.

[3] A Motzkin path is a path in the (x,y)-plane with allowed steps $(x,y) \to (x+1,y+1)$, $(x,y) \to (x+1,y)$, and $(x,y) \to (x+1,y-1)$, and whose y-coordinates are all non-negative.

6.2.1. The main theorem

The following theorem is the key connection between the Fourier diagram basis and the value of random quadratic polynomial optimization:

Theorem 6.9 (Optimization in the tree basis)**.** *Assume that $\boldsymbol{A}$ satisfies Assumption 2.1. Let $\mathcal{T}$ be a collection of one-dimensional random variables indexed by rooted trees, whose distributions match the asymptotic distribution of a single coordinate of the tree diagrams from Theorem 3.14. Then for any integer $p \geqslant 2$,*

$$\max_{\|\boldsymbol{x}\|_p \leqslant 1} \sum_{i,j=1}^{n} A_{ij} x_i x_j \geqslant (2 - o(1)) \cdot n^{1-\frac{2}{p}} \cdot \sup_{\substack{Z \in \operatorname{span}(\mathcal{T}) \\ \mathbb{E} Z^p \leqslant 1}} \mathbb{E}\left[ZZ^+\right].$$

The proof of Theorem 6.9 is an easy consequence of the tools we developed in Part I.

Proof. First, we claim that for any $Z \in \mathcal{T}$, there is an iterative algorithm whose asymptotic state is equal to Z (viewed as a tree). This follows from an inductive argument on the number of vertices in the tree.

1. Suppose that the root of Z has degree 1. Let $k \geqslant 1$ be such that Z can be decomposed from the root into a path of length k, followed by a tree Z', which is either a singleton or has a root of degree larger than 1. By induction, let $\boldsymbol{z}'$ be the iterate of an algorithm with asymptotic state Z'. Consider the iteration $\boldsymbol{z}_{-1} = 0$, $\boldsymbol{z}_0 = \boldsymbol{z}'$, and $\boldsymbol{z}_{i+1} = \boldsymbol{A}\boldsymbol{z}_i - \boldsymbol{z}_{i-1}$ for all $i \geqslant 0$. Then $\boldsymbol{z}_k$ has asymptotic state Z.
2. Otherwise, Z can be obtained by grafting several branches at the root. An iterate with asymptotic state Z can be obtained by applying the appropriate multivarite Hermite polynomial to the iterates computing the different branches.

More generally, an algorithm whose asymptotic state is an arbitrary element in the span of $\mathcal{T}$ can be obtained by taking linear combinations of the previous construction.

Let $\boldsymbol{z}$ be the output of an algorithm with asymptotic state Z. Then $\boldsymbol{z}^{\odot p}$ has asymptotic state Z^p, so

$$\frac{1}{n}\|\boldsymbol{z}\|_p^p = \frac{1}{n}\sum_{i=1}^{n} z_i^p \to \mathbb{E} Z^p \leqslant 1 \tag{6.7}$$

by Theorem 3.17. Moreover, $\frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z}\rangle \to 2\,\mathbb{E}\left[ZZ^+\right]$ by the same result. Combining with (6.7), we get:

$$\frac{\langle \boldsymbol{z}, \boldsymbol{A}\boldsymbol{z}\rangle}{\|\boldsymbol{z}\|_p^2} \geqslant (2 - o(1)) \cdot \frac{n\,\mathbb{E}\left[ZZ^+\right]}{n^{2/p}},$$

which completes the proof after rearranging the inequality. □

We now show a direct application of Theorem 6.9 in the case $p = 2$. We design an optimal low-degree algorithm by solving the right-hand side of Theorem 6.9:

$$\sup_{\substack{Z \in \mathrm{span}(\mathcal{T}) \\ \mathbb{E} Z^2 \leqslant 1}} 2\,\mathbb{E}\left[ZZ^+\right] .$$

Since $\mathcal{T}$ forms an orthogonal basis under the inner product defined by $\mathbb{E}$, this reduces to selecting a linear combination of trees whose squared coefficients sum to 1, in order to maximize the correlation between Z and Z^+.

Remark 6.10. For $p = 2$, our method is an asymptotic variant of the *Randomized Krylov Method* [Tro20, §2.4], in the sense that we will select an optimal iterate in the span of $\{\mathbf{1}, \boldsymbol{A}\mathbf{1}, \ldots, \boldsymbol{A}^k\mathbf{1}, \ldots\}$ (i.e., the span of all path diagrams). In particular, one can think of Theorem 6.9 as a generalization of Krylov subspace methods beyond the unconstrained setting.

6.2.2. AMP power iteration

A natural candidate for optimizing this objective is illustrated in Figure 6.1:

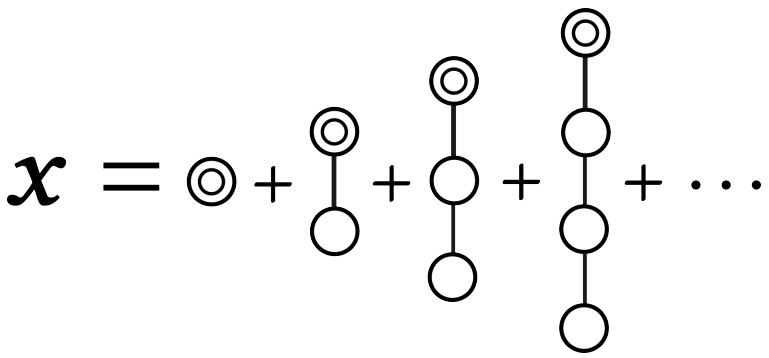


Figure 6.1. The infinite sequence of path diagrams.

This corresponds to the asymptotic state of the AMP power iteration algorithm introduced in Chapter 4:

$$z_0 = \mathbf{1}\,, \quad z_{t+1} = \boldsymbol{A}z_t - z_{t-1}\,, \quad x_t = \sum_{s=0}^{t-1} z_s\,.$$

More formally, for any $T \geqslant 0$, define $X_T = \sum_{t=0}^{T-1} P_t$, where P_t denotes the length-t path diagram. Then $X_T^+ = \sum_{t=1}^{T} P_t$, so that

$$\mathbb{E}\left[X_T X_T^+\right] = T - 1\,, \quad \mathbb{E} X_T^2 = T\,.$$

Thus, the resulting value of the objective is

$$\frac{2\,\mathbb{E}\left[X_T X_T^+\right]}{\mathbb{E}\,X_T^2} = 2 - \frac{1}{T}\,.$$

Note that this matches the guarantees of the algorithm from §6.1.1, although with a better constant factor in front of $1/T$.

Remark 6.11. The Chebychev polynomials of the second kind $(U_t)_{t\geqslant 0}$ satisfy the similar recurrence $U_{t+1}(x) = 2xU_t(x) - U_{t-1}(x)$. Hence, we have $\boldsymbol{z}_t = U_t(\boldsymbol{A}/2)\mathbf{1}$. This iteration also appears in the numerical linear algebra literature as a Krylov subspace method for the more general problem of computing the top eigenvalue of a general positive semidefinite matrix in the absence of spectral gap [KW92, Tro20].

6.2.3. The optimal algorithm for spherical maximization

There is an even better algorithm achieving value $2 - O(T^{-2})$, as we show now.

Theorem 6.12. *For any $Z \in \operatorname{span}(\mathcal{T})$ supported on trees of depth at most T,*

$$\frac{2\,\mathbb{E}\left[ZZ^+\right]}{\mathbb{E}\,Z^2} \leqslant 2\cos\left(\frac{\pi}{T+2}\right)\,.$$

Moreover, equality is achieved for

$$Z^* = \sum_{t=0}^{T} \sin\left(\frac{t\pi}{T+2}\right) P_t\,,$$

where P_t denotes the length-t path diagram.

Note that a nonlinear iteration with asymptotic state Z^* can be obtained by simply changing the coefficients when recombining the iterates in AMP power iteration:

$$\boldsymbol{z}_0 = \mathbf{1}\,, \quad \boldsymbol{z}_{t+1} = \boldsymbol{A}\boldsymbol{z}_t - \boldsymbol{z}_{t-1}\,, \quad \boldsymbol{x}_t = \sum_{s=0}^{t-1} \sin\left(\frac{s\pi}{t+1}\right)\boldsymbol{z}_s\,.$$

Proof. We define a *piece* to be a sequence of trees $(\tau_1, \dots, \tau_k)$ such that $\tau_{i+1} = \tau_i^+$ for all $1 \leqslant i < k$. Decompose Z into a union of disjoint maximal pieces. Then observe that both $\mathbb{E}\left[ZZ^+\right]$ and $\mathbb{E}\,Z^2$ decompose across pieces, so by an averaging argument, it suffices to prove the upper bound for a single piece.

Now, suppose that $Z = \sum_{i=1}^k c_i\tau_i$ for some coefficients $c_i \in \mathbb{R}$, where $\tau_{i+1} = \tau_i^+$ for all $i < k$, and $k \leqslant T+1$. Then,

$$\frac{2\,\mathbb{E}\left[ZZ^+\right]}{\mathbb{E}\,Z^2} = \frac{2\sum_{i=1}^{k-1} c_i c_{i+1}}{\sum_{i=1}^k |\operatorname{Aut}(\tau_i)|\,c_i^2} \leqslant \frac{\langle \boldsymbol{c}, \boldsymbol{B}\boldsymbol{c}\rangle}{\|\boldsymbol{c}\|_2^2}\,, \tag{6.8}$$

where $\boldsymbol{B}$ is the adjacency matrix of the length-k path graph ($B_{i,i+1} = B_{i+1,i} = 1$ for all $i < k$, and zeros elsewhere). The largest eigenvalue of $\boldsymbol{B}$ is $2\cos(\pi/(k+1))$, attained with the eigenvector $c_i = \sin(i\pi/(k+1))$.

Finally, the inequality in (6.8) is tight when each τ_i is the length-i path diagram, and the objective is maximized when $k = T+1$, which completes the proof. □

6.3. Random optimization over the hypercube

While we are not able to describe the optimal asymptotic state in a similar way when $p = \infty$ in Theorem 6.9, we give some interpretation of Montanari's algorithm from [Mon19] in the tree basis.

Montanari's algorithm is a special type of AMP algorithm called *iterative AMP* (or martingale AMP), which uses (4.7) with the functions

$$f_t(\boldsymbol{w}_t, \dots, \boldsymbol{w}_0) = \boldsymbol{w}_t \odot u_t(\boldsymbol{w}_{t-1}, \dots, \boldsymbol{w}_0) \tag{6.9}$$

for chosen functions $u_t : \mathbb{R}^t \to \mathbb{R}$ applied componentwise, where $\odot$ denotes componentwise multiplication. The candidate output of the algorithm is $\boldsymbol{x}_T = \sum_{t=1}^T \boldsymbol{w}_t \odot u_t(\boldsymbol{w}_{t-1}, \dots, \boldsymbol{w}_0) = \sum_{t=1}^T f_t(\boldsymbol{w}_t, \dots, \boldsymbol{w}_0)$.

The special property of iterative AMP is that it sums up *independent* Gaussian vectors $\boldsymbol{w}_t$ scaled componentwise by the functions u_t. The independence of the Gaussian vectors $\boldsymbol{w}_t$ is contained in the state evolution for AMP as follows.[4] By Theorem 4.10, the asymptotic states W_t, U_t, X_t of $\boldsymbol{w}_t, \boldsymbol{u}_t, \boldsymbol{x}_t$ satisfy $U_0 = W_0 = 1$,

$$U_t = u_t(W_{t-1}, \dots, W_0), \quad W_{t+1} = (U_t W_t)^+, \quad X_t = \sum_{s=1}^t U_t W_t .$$

Claim 6.13. *U_t is in the span of trees in $\mathcal{T}$ with depth at most $t-1$ and W_t is in the span of trees in $\mathcal{S}$ with depth exactly t.*

Proof of Claim 6.13. Arguing inductively, as componentwise functions do not increase the depth, U_t is in the span of trees from $\mathcal{T}$ of depth at most $t-1$. In the product $U_t W_t$, the trees of depth t in W_t cannot be cancelled by any trees of lower depth from U_t. Therefore all trees in $U_t W_t$ and $W_{t+1} = (U_t W_t)^+$ have depth exactly t and $t+1$ respectively, as needed. □

Claim 6.13 provides a very clear explanation of where the independent Gaussians are coming from: the W_t have different depths, and Gaussian diagrams of different depths are asymptotically independent Gaussian vectors.

[4] The algorithm of [Mon19] uses a non-polynomial f_t which is not directly covered by our state evolution proof. However, Ivkov and Schramm [IS24] prove that this AMP can be approximated by polynomial f_t.

Optimality via state evolution. The objective value achieved by the iteration can also be computed using state evolution. For the Sherrington–Kirkpatrick model, the objective is $\max_{\boldsymbol{x}\in\{-1,1\}^n} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$. The value achieved by the iteration is:

$$\begin{aligned}
\frac{1}{n}\boldsymbol{x}_T^\top \boldsymbol{A}\boldsymbol{x}_T &\overset{\infty}{=} \mathbb{E}\left[X_T(X_T^+ + X_T^-)\right] && \text{(Lemma 3.21)}\\
&= 2\,\mathbb{E}\left[X_T X_T^+\right] && \text{(Lemma 3.24)}\\
&= 2\sum_{s,t=1}^{T} \mathbb{E}\left[U_s W_s (U_t W_t)^+\right]\\
&= 2\sum_{s,t=1}^{T} \mathbb{E}\left[U_s W_s W_{t+1}\right]\\
&= 2\sum_{t=2}^{T} \mathbb{E}\left[U_t W_t^2\right] && \text{(Independence of the } W_t\text{)}\\
&= 2\sum_{t=2}^{T} \mathbb{E}\left[U_t\right]\mathbb{E}\left[W_t^2\right] && \text{(Claim 6.13 and independence of the } W_t\text{)}
\end{aligned}$$

This gives an asymptotic description of both the iterates and the objective value achieved by the algorithm. The remaining key step used by [Mon19] is to observe that when the number of steps T is taken large, the point $\boldsymbol{x}_T$ heuristically approaches a martingale process $\mathrm{d}X_t = U_t\,\mathrm{d}B_t$ with the steps $\boldsymbol{w}_t$ converging to the Brownian motion. This limit only holds if we choose u_t to satisfy $\mathbb{E}\left[U_t^2\right] = 1$ so that $\mathbb{E}\left[W_t^2\right] = \mathbb{E}\left[U_t^2\right]\mathbb{E}\left[W_{t-1}^2\right] = \mathbb{E}\left[W_{t-1}^2\right]$ for all t. In this limit, the function $u_t(\boldsymbol{w}_{t-1},\dots,\boldsymbol{w}_0)$ is chosen in the best possible way so as to maximize the objective value:

$$\begin{aligned}
\max\ & 2\int_0^1 \mathbb{E}\left[U_t\right]\mathrm{d}t\\
\text{s.t.}\ & (U_t)_{t\in[0,1]} \text{ is progressively measurable w.r.t. a Brownian motion } (B_t)_{t\in[0,1]}\\
& \mathbb{E}\left[U_t^2\right] = 1 \text{ for all } t\in[0,1]\\
& \int_0^1 U_t\,\mathrm{d}B_t \in [-1,+1] \text{ a.s.}
\end{aligned}$$

This optimization problem is convex and dual to an "extended Parisi formula" for the optimal value of the SK model [EMS21, §4]. The remaining important technical step is to show that this program is well-posed, and that the maximizer of this program, which can be written in terms of the solution to the Parisi PDE, is smooth enough that it can be discretely approximated by the limit $T\to\infty$.

6.4. Beyond random polynomials

Due to concentration of measure, the maxima of random polynomials typically concentrate around deterministic values (their expectations) as $n \to \infty$. But what about arbitrary (multilinear) polynomials? How small can their maximum value be, relative to the size of their coefficients? In particular, how "unsatisfiable" can a 3-Sat formula be?

Motivated by this last question, we focus on optimization over the hypercube, and measure the size of the coefficients of a multilinear polynomial p using their ℓ_1-norm:

Definition 6.14. Let $p(\boldsymbol{x}) = \sum_{S \subseteq [n]} c_S \prod_{i \in S} x_i$ be a multilinear polynomial. We define:

$$\|p\|_1 := \sum_{S \subseteq [n]} |c_S| \,, \quad \|p\|_2 := \left(\sum_{S \subseteq [n]} c_S^2 \right)^{1/2} .$$

6.4.1. Quadratic polynomials

For quadratic polynomials, the following simple example provides an upper bound:

Example 6.15 (Max-Cut on the complete graph)**.** Let $p(\boldsymbol{x}) = -\frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} x_i x_j$. Then

$$\max_{\boldsymbol{x} \in \{-1,1\}^n} p(\boldsymbol{x}) \lesssim \frac{1}{n} = \frac{\|p\|_1}{n} .$$

This upper bound is tight in general, as the following one-sided anti-concentration shows:

Fact 6.16 (Lemma 3.2 of [AGK04])**.** *Let $p \colon \{-1,1\}^n \to \mathbb{R}$ be a degree-k polynomial such that $\mathbb{E}\, p = 0$. Then,*

$$\Pr_{\boldsymbol{x} \sim \{-1,1\}^n} \left(p(\boldsymbol{x}) \geqslant \frac{\|p\|_2}{4 \cdot 3^k} \right) \geqslant \Omega(9^{-k}) .$$

In particular, for constant k, there always exists an assignment with value $\Omega(\|p\|_1 \cdot n^{-k/2})$.

6.4.2. Cubic polynomials

For cubic polynomials, the best known upper bound is achieved by a random construction:

Example 6.17 (Pure 3-spin Ising model)**.** Let $c_{ijk} \overset{\text{i.i.d.}}{\sim} \{-1, 1\}$ and define

$$p(\boldsymbol{x}) = \sum_{1 \leqslant i,j,k \leqslant n} c_{ijk} x_i x_j x_k .$$

Then $\|p\|_1 = n^3$. Moreover, for any fixed $\boldsymbol{x} \in \{-1,1\}^n$, the standard deviation of $p(\boldsymbol{x})$ is $\Theta(n^{3/2})$, so a union bound implies that the maximum of p over $\{-1,1\}^n$ is $O(n^2)$, which is $O(\|p\|_1 / n)$.

For general cubic polynomials, the best-known lower bound in *absolute value* is given by the following theorem, which is implicit in [DFKO06].

Theorem 6.18. *Let $p\colon \{-1,1\}^n \to \mathbb{R}$ be a degree-k polynomial. There is an algorithm that outputs $\boldsymbol{x} \in \{-1,1\}^n$ satisfying*

$$|p(\boldsymbol{x})| \geqslant 2^{-O(k)} \sum_{i=1}^{n} \sqrt{\mathrm{Inf}_i[p]}\,,$$

where $\mathrm{Inf}_i[p] = \sum_{S \ni i} c_S^2$, when $p(\boldsymbol{x}) = \sum_S c_S \boldsymbol{x}^S$.

By Cauchy–Schwarz, this implies a lower bound of $\Omega(\|p\|_1 / n)$ for the maximum in absolute value of degree-3 polynomials. Note that this implies a similar lower bound when maximizing homogeneous cubic polynomials, using the symmetry $p(\boldsymbol{x}) = -p(-\boldsymbol{x})$.

However, this leaves open the question of the best possible lower for non-homogeneous cubic polynomials:

Problem 6.19 (Extremal value of degree-3 polynomials)**.** Determine the scaling of the minimal possible maximum of multilinear degree-3 polynomials over the Boolean hypercube:

$$\max_{\boldsymbol{x} \in \{-1,1\}^n} \sum_{\substack{S \subseteq [n] \\ |S| \leqslant 3}} c_S \prod_{i \in S} x_i\,,$$

where the coefficients satisfy $\sum |c_S| \leqslant 1$.

The best known lower bound is $\Omega(n^{-3/2})$ by Fact 6.16, and the best known upper bound is $O(1/n)$ by Example 6.17.

A potential approach to Problem 6.19 would be to analyze constructions that adaptively modify Example 6.17 by planting structured instances such as Example 6.15. However, the author was unable to analyze these constructions rigorously.

6.4.3. A generic cubic optimization algorithm

We now describe a generic algorithm for maximizing homogeneous cubic polynomials over the hypercube that simultaneously achieves: (1) nontrivial advantage over a random assignment (in the sense of the previous subsection), and (2) a good approximation to the maximum value of the polynomial.

For reasons explained in §8.1.3, we assume without loss of generality that the input is a *decoupled* homogeneous cubic polynomial:

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := \sum_{i,j,k=1}^{n} T_{ijk} x_i y_j z_k\,,$$

where the 3-tensor $\boldsymbol{T}$ satisfies $T_{ijk} = 0$ whenever i, j, k are not all distinct.

The algorithm is presented in Algorithm 1.

Algorithm 1 Generic homogeneous cubic optimization algorithm

1. Sample a random $\overline{\boldsymbol{x}} \sim \{-1, 1\}^n$.
2. Solve the SDP relaxation of $\boldsymbol{y}^\top \boldsymbol{M}(\overline{\boldsymbol{x}})\boldsymbol{z}$, and apply Grothendieck rounding (Theorem 8.9) to the matrix $\boldsymbol{M}(\overline{\boldsymbol{x}})$ defined as

$$M(\overline{\boldsymbol{x}})_{ij} := \sum_{k=1}^{n} T_{kij}\overline{x}_k \,.$$

   Let $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{z}}$ be the resulting assignments.
3. Repeat steps 1–2 $\mathrm{poly}(n)$ times, and return the best solution $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}})$.

While the original algorithm from [KN08] uses a reduction to computing the ℓ_1-diameter of a convex body, we show that it is equivalent to Algorithm 1. We provide a self-contained analysis that matches the guarantees of both [KN08] and [BMO+15]. This analysis will serve as a starting point for our algorithms with certifiable guarantees in Chapter 8.

Proposition 6.20. *The output $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}})$ of Algorithm 1 satisfies with high probability:*

$$\langle \boldsymbol{T}, \overline{\boldsymbol{x}} \otimes \overline{\boldsymbol{y}} \otimes \overline{\boldsymbol{z}} \rangle \geqslant \Omega\left(\sqrt{\frac{\log n}{n}}\right) \cdot \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \{-1,1\}^n} \langle \boldsymbol{T}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \rangle$$

The key lemma to prove Proposition 6.20 is the following anti-concentration inequality.

Lemma 6.21 (Lemma 3.2 of [KN08])**.** *For any $\delta \in (0, 1/2)$, there is a constant $c(\delta) > 0$ such that for any $\boldsymbol{a} \in \mathbb{R}^n$, if $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\{\pm 1\}$ random variables, then*

$$\Pr\left[\sum_{i=1}^{n} a_i \varepsilon_i \geqslant \sqrt{\frac{\delta \log n}{n}} \cdot \|\boldsymbol{a}\|_1\right] \geqslant \frac{c(\delta)}{n^\delta} \,.$$

A simple proof of this lemma can be derived from our derandomization argument in §8.5.

Proof of Proposition 6.20. Let $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^* \in \{\pm 1\}^n$ be an optimal solution. First observe that it is always optimal to set $x_i^* := \mathrm{sgn}(\langle \boldsymbol{T}_i, \boldsymbol{y}^* \otimes \boldsymbol{z}^* \rangle)$, so that $\mathrm{OPT} := f(\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^*) = \sum_{i=1}^n |\langle \boldsymbol{T}_i, \boldsymbol{y}^* \otimes \boldsymbol{z}^* \rangle|$. Then, for any $\overline{\boldsymbol{x}} \in \{\pm 1\}^n$, the algorithm outputs $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}})$ such that

$$f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \Omega(1) \cdot \max_{\boldsymbol{y},\boldsymbol{z} \in \{\pm 1\}^n} f(\overline{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}) \geqslant \Omega(1) \cdot f(\overline{\boldsymbol{x}}, \boldsymbol{y}^*, \boldsymbol{z}^*) \,.$$

However, by Lemma 6.21, with at least inverse polynomial probability, a random $\overline{x}$ satisfies

$$\begin{aligned} f(\overline{x}, y^*, z^*) &= \sum_{i=1}^{n} \overline{x}_i \langle T_i, y^* \otimes z^* \rangle \\ &\geqslant \Omega(1) \cdot \sqrt{\frac{\log n}{n}} \cdot \sum_{i=1}^{n} |\langle T_i, y^* \otimes z^* \rangle| \\ &= \Omega(1) \cdot \sqrt{\frac{\log n}{n}} \cdot \mathrm{OPT}\,. \end{aligned}$$

Thus, by repeating $\mathrm{poly}(n)$ times, with high probability the algorithm outputs an assignment that has value $\Omega\left(\sqrt{\frac{\log n}{n}}\right) \cdot \mathrm{OPT}$. □

Next, we show that this algorithm also always finds a ± 1-assignment with value matching the guarantees we deduced from [DFKO06] in §6.4.2.

Proposition 6.22. *The output $(\overline{x}, \overline{y}, \overline{z})$ of Algorithm 1 satisfies with high probability:*

$$\langle T, \overline{x} \otimes \overline{y} \otimes \overline{z} \rangle \gtrsim \frac{\|T\|_1}{n}\,.$$

Proof. Let $y^*(\overline{x}), z^*(\overline{x})$ be the maximizer of $y^\top M(\overline{x}) z$. Further, denote by

$$z'_k = \mathrm{sgn}\left(\sum_{i,j=1}^{n} T_{ijk} \overline{x}_i \overline{x}_j\right)$$

the optimal assignment given $\overline{x}$. The analysis of Grothendieck rounding implies that

$$\begin{aligned} \langle T, \overline{x} \otimes \overline{y} \otimes \overline{z} \rangle &\geqslant \Omega(1) \cdot \langle T, \overline{x} \otimes y^*(\overline{x}) \otimes z^*(\overline{x}) \rangle \\ &\geqslant \Omega(1) \cdot \langle T, \overline{x} \otimes \overline{x} \otimes z' \rangle\,. \end{aligned}$$

Now, for every $k \in [n]$, by Fact 6.16 (applied to the degree-2 polynomial in $\overline{x}$),

$$\mathop{\mathbb{E}}_{\overline{x}} \left| \sum_{i,j=1}^{n} T_{ijk} \overline{x}_i \overline{x}_j \right| \gtrsim \left(\mathop{\mathbb{E}}_{\overline{x}} \left(\sum_{i,j=1}^{n} T_{ijk} \overline{x}_i \overline{x}_j \right)^2 \right)^{\frac{1}{2}} = \left(\sum_{i,j=1}^{n} T_{ijk}^2 \right)^{\frac{1}{2}}\,.$$

This implies that with the optimal choice z',

$$\mathop{\mathbb{E}}_{\overline{x}} \langle T, \overline{x} \otimes \overline{x} \otimes z' \rangle \gtrsim \sum_{k=1}^{n} \left(\sum_{i,j=1}^{n} T_{ijk}^2 \right)^{\frac{1}{2}} \gtrsim \frac{\|T\|_1}{n}\,.$$

Finally, $\langle T, \overline{x} \otimes \overline{x} \otimes z' \rangle$ is always bounded by $\|T\|_1$, so by Markov's inequality, it exceeds $\Omega(\|T\|_1 / n)$ with probability $\Omega(1/n)$. In particular, we get such an assignment with high probability after repeating the experiment $\mathrm{poly}(n)$ times. This completes the proof. □

In fact, the same proof shows that this algorithm always finds an assignment with value $\frac{1}{2} + \Omega(\frac{1}{\sqrt{D}})$, in any 3-Xor formula in which every variable appears in at most D clauses, matching the main result of [BMO+15].

6.5. Summary

In this chapter, we analyzed several algorithms for maximizing random quadratic polynomials over the sphere, ranging from basic power iteration to optimally tuned AMP power iteration. We started discussing the worst-case analog of the problem, and noted that it remains open in the case of non-homogeneous cubic polynomials.

CHAPTER 7.

The Second Eigenvalue of Random Hypergraphs

In this chapter, we estimate the maximum of *random* and *sparse* polynomials of the form

$$f(\boldsymbol{x}) = \sum_{i,j,k=1}^{n} T_{ijk} x_i x_j x_k \,, \tag{7.1}$$

whose coefficients are independent mean-0 and variance-1, but their higher moments grow with n. Such polynomials arise in the study of spectral certificates for random hypergraphs.

Our main contributions in this chapter (Theorems 7.3 and 7.4) are tight upper bounds on (7.1) in the "ultra-sparse" regime where the average number of non-zero entries per row of the tensor is $o(1)$. Our proof is based on a discretization scheme capturing the multiscale sparsity of test vectors, reminiscent of ideas appearing in the study of Rademacher processes.

Table of contents

The results in this chapter are unpublished. The author thanks Luca Trevisan for suggesting to study this problem, and Kevin Lucca for discussions related to tensor concentration.

7.1. Introduction

We study the following generalization of the Erdős-Renyi random graph model to hypergraphs:

Definition 7.1. Fix $t \geqslant 2$ and let $p = p(n) \in [0,1]$. The distribution $\mathcal{H}_t(n,p)$ over t-uniform tensors $\boldsymbol{T}$ is defined as follows: for each $\boldsymbol{i} \in \binom{[n]}{t}$ independently, let $T_{\boldsymbol{i}} = 1$ with probability p and $T_{\boldsymbol{i}} = 0$ otherwise; and set: $T_{\sigma(\boldsymbol{i})} = T_{\boldsymbol{i}}$ for every permutation $\sigma \in \mathfrak{S}_t$. Finally, let $T_{\boldsymbol{i}} = 0$ for every $\boldsymbol{i}$ whose coordinates are not all distinct.

We also consider the following notion of tensor norm:

Definition 7.2. For any t-uniform tensor $\boldsymbol{T}$, define

$$\|\boldsymbol{T}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \left| \sum_{\boldsymbol{i} \in [n]^t} T_{\boldsymbol{i}} \prod_{u=1}^{t} x_{i(u)} \right| = \max_{\|\boldsymbol{x}\|_2=1} \left| \left\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes t} \right\rangle \right| . \tag{7.2}$$

When $\boldsymbol{T}$ is a symmetric tensor (i.e., $T_{\boldsymbol{i}} = T_{\sigma(\boldsymbol{i})}$ for all $\boldsymbol{i} \in [n]^t$ and $\sigma \in \mathfrak{S}_t$), $\|\boldsymbol{T}\|_2$ is equivalent to the *decoupled* maximization problem $\langle \boldsymbol{T}, \boldsymbol{x}_1 \otimes \ldots \otimes \boldsymbol{x}_t \rangle$ over unit vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$.[1] Such an equivalence holds for matrices, and can then be argued by induction on t; see e.g. [FW95].

Our goal, anticipated in §1.3.2, is to understand for which scalings of $p = p(n)$ does the quantity $\|\boldsymbol{T}\|_2$ become noticeably larger than $\|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|_2$. When $t = 2$, this threshold can be predicted by estimating the magnitude of the quadratic form of very dense test vectors only [Tre17a]. Following this insight, for general t, we expect $|\langle \boldsymbol{T}, \mathbf{1}^{\otimes t} \rangle|$ to start growing once $p = \Omega(n^{-t/2})$, while $|\langle \boldsymbol{T} - \mathbb{E}\boldsymbol{T}, \mathbf{1}^{\otimes t} \rangle|$ remains bounded until $p \sim 1/n$. This suggests a natural threshold at $p \sim n^{-t/2}$ for the Erdős-Renyi model.

The main result of this chapter, established in §7.3, confirms this phase transition for $t = 3$:

Theorem 7.3. *Suppose $p(n) = O(n^{-1-\varepsilon})$ for some constant $\varepsilon > 0$. Let $\boldsymbol{T} \sim \mathcal{H}_3(n,p)$. Then,*

$$\|\boldsymbol{T}\|_2 = \Omega(pn^{1.5}) ,$$
$$\|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|_2 = O(1) .$$

As argued in §1.3.2, this shows that 3-uniform Erdős-Renyi hypergraphs are quasirandom once they have $\Omega(n^{1.5})$ hyperedges.

For more general t, we prove the following weaker bound in §7.2:

[1] This alternative definition is usually called the *injective tensor norm*.

Theorem 7.4. *Let $t \in \{3, 4\}$. Suppose that $\boldsymbol{T} \sim \mathcal{H}_t(n, p)$, where $p \leqslant O(n^{-1-\varepsilon})$ for some constant $\varepsilon > 0$. Then, with high probability,*

$$\|\boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}\|_2 \leqslant K \log^2 \alpha_t\,,$$

where $K = K(\varepsilon)$ is a constant depending only on ε, and

$$\alpha_3 := \max_{i,j \in [n]} \left|\{(a, k, \ell) \in [n]^3 : T_{aik} = 1 \text{ and } T_{aj\ell} = 1\}\right|,$$
$$\alpha_4 := \max_{i,j \in [n]} \left|\{(k, \ell) \in [n]^2 : T_{ijk\ell} = 1\}\right|.$$

7.1.1. From tensor norms to refutation

Before proceeding, we give a concrete motivation behind Theorem 7.3.

Refuting random Boolean formulas. Random t-SAT is the average-case analog of Problem 1.1. An instance on n variables and m clauses is generated by sampling m random t-tuples of variables and adding them to the formula with uniformly random literals. With high probability, such formulas are unsatisfiable once m/n exceeds a threshold depending on t. On the algorithmic side, one can ask about the existence of efficient *certificates* of unsatisfiability. A conjecture of Feige states that this problem exhibits a fundamental statistical-computational gap, in that such certificates cannot be found in polynomial time unless m is much larger than n [Fei02].

The state of the art is that for even t, efficient certificates of unsatisfiability are known when $m = \Omega(n^{t/2})$, while for odd t, prior work until recently required $m = \Omega(n^{t/2} \cdot \mathrm{polylog}(n))$ [AOW15]. These polylog terms are believed to be technical artifacts due to difficulties arising with odd-order tensors. The intuition is that for even t, one can fit the entries the tensor in a square matrix of dimension $\frac{t}{2} \times \frac{t}{2}$, but this idea does not generalize naturally when t is odd. Theorem 7.3 goes beyond these flattening-based methods.

Bounding large independent sets. Given a random t-SAT instance on n variables, consider the following t-uniform hypergraph:

1. Create $2n$ vertices, representing each variable and its negation.
2. For each clause, add a hyperedge corresponding to the assignment forbidden by that clause.

A simple observation is that any satisfying assignment corresponds to an *independent set* of size n in this hypergraph. However, random hypergraphs with more than $n^{t/2}$ hyperedges do not typically have such large independent sets. Therefore, the size of the maximum

independent set of this graph can certify unsatisfiability of the formula. While it is not efficiently computable, we proceed to further upper bound it by a *spectral* certificate.

Although there are small correlations between the hyperedges, we are going to assume that the adjacency tensor of the constructed hypergraph is sufficiently close in distribution to $\mathcal{H}_t(2n, p)$, so that Theorem 7.3 continues to hold. Similarly to the graph case, $\lambda_{\mathrm{FW}}(\boldsymbol{T}) = \|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|_2$ is a spectral certificate for the maximum independent set in $\boldsymbol{T}$. Indeed, if $S \subseteq [n]$ is an independent set (i.e., $T_{\boldsymbol{i}} = 0$ for all $\boldsymbol{i} \in S^t$) and $\boldsymbol{x} = \mathbf{1}_S$, then for $\boldsymbol{T} \sim \mathcal{H}_t(n, p)$,

$$\frac{\left|\left\langle \boldsymbol{T} - \mathbb{E}\boldsymbol{T}, \boldsymbol{x}^{\otimes t}\right\rangle\right|}{\|\boldsymbol{x}\|_2^t} = \frac{p\,|S|^3}{|S|^{\frac{3}{2}}} = p\,|S|^{\frac{3}{2}}\ .$$

Thus, the maximum independent set is bounded by $\|\boldsymbol{T}\|_2^{2/3} p^{-2/3}$. By Theorem 7.3, this certificate rules out independent sets of size $\Omega(n)$ when the number of hyperedges is $\Omega(n^{3/2})$, thereby certifying unsatisfiability of random formulas when $m = \Omega(n^{3/2})$. Notably, this avoids any extra polylog losses compared to the flattening argument.

Theorem 7.3 does not directly resolve the algorithmic question of refuting random t-Sat for odd t, because we do not know how well can $\|\boldsymbol{T}\|_2$ be approximated on sparse random inputs (for dense inputs, the conjectured approximation ratio is polynomial in n). However, we believe that the proof of Theorem 7.3 motivates the construction of tensor norms that are (1) efficiently approximable, and (2) amenable to a multiscale union bound argument. We note that the question of refuting random t-Sat instances with $\Omega(n^{t/2})$ clauses for odd t was recently resolved in [dT23] using a completely different spectral certificate.

7.1.2. Chaining and Rademacher processes

Both Theorems 7.3 and 7.4 are beyond the reach of standard techniques such as flattening, trace methods, or simple union bounds. Prior work on tensor concentration, such as [BGJ+25, Boe24], focuses on dense or moderately sparse tensors and yields bounds that are too weak in our setting. Many previous results in random matrix theory rely on discretization arguments, and we do not attempt to survey them here.

Our proof method is close in spirit to the method of *chaining* [Tal21]. A classical way to approach the quantity $\mathbb{E}\,\|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|$ for a random tensor $\boldsymbol{T} \sim \mathcal{H}_3(n, p)$ is to start from its symmetrized version

$$\mathop{\mathbb{E}}_{\boldsymbol{T}\sim\mathcal{H}_3(n,p)} \|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\| \asymp \mathop{\mathbb{E}}_{\boldsymbol{T}\sim\mathcal{H}_3(n,p)} \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{-1,1\}^{\otimes n^3}} \|\boldsymbol{\varepsilon} \odot \boldsymbol{T}\|. \tag{7.3}$$

(See, e.g., [Ver18, Lemma 6.4.2] for a proof of this equivalence, which holds for random vectors in general normed spaces.)

Conditioned on $\boldsymbol{T}$, the expectation $\mathbb{E}\,\|\boldsymbol{\varepsilon} \odot \boldsymbol{T}\|$ is the supremum of a Rademacher process. In this perspective, $\boldsymbol{T}$ acts as a fixed $\{0,1\}$-sparsity pattern for the Rademacher tensor $\boldsymbol{\varepsilon} \odot \boldsymbol{T}$. One may then try to further upper bound this supremum by the supremum of an associated Gaussian process via Talagrand's comparison inequality,

$$\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon} \sim \{-1,1\}^{\otimes n^3}} \|\boldsymbol{\varepsilon} \odot \boldsymbol{T}\| \lesssim \mathop{\mathbb{E}}_{\boldsymbol{g} \sim \mathcal{N}(0,1)^{\otimes n^3}} \|\boldsymbol{g} \odot \boldsymbol{T}\|. \tag{7.4}$$

The hope would be to control the right-hand side of (7.4) using the chaining machinery of [Tal21] for Gaussian processes, obtaining an upper bound that depends only on combinatorial properties of $\boldsymbol{T}$ (that would be negligible with high probability when $\boldsymbol{T}$ is the adjacency tensor of a sparse random Erdős-Renyi hypergraph). However, this strategy fails: at the expected phase transition threshold, $\boldsymbol{T}$ has about $n^{1.5}$ nonzero entries, and with high probability at least one of the corresponding Gaussians is of order $\Omega(\sqrt{\log n})$, which already spoils the desired $O(1)$ bound.

Returning to (7.3), a more successful approach is to exploit that the supremum of a Rademacher process can be entirely understood by combining Gaussian-process tools with the trivial inequality $\left|\sum_i \varepsilon_i x_i\right| \leqslant \|\boldsymbol{x}\|_1$ when $\boldsymbol{\varepsilon}$ has ± 1 entries [Tal21, Chapter 5]. Recently, Latała [Lat24] used this idea to prove bounds for the matrix analog of (7.3). The intermediate, weaker estimate in [Lat24, §4] is based on an argument that is very similar to the one we give in §7.2. Whether the construction of §7.3 can be used to extend the results of [Lat24] from matrices to higher-order tensors remains a tantalizing open problem.

7.2. Lifting discrete to continuous test vectors

In this section, we prove Theorem 7.4. While not explicitly relying on chaining, our argument is similarly based on discretizing and grouping together vectors with similar properties. This section is inspired by the strategy of Bilu and Linial for the proof of their converse of the expander mixing lemma [BL06]. We will see in §7.3 that a more careful grouping and union bound argument yields the optimal answer for 3-uniform tensors.

7.2.1. Preliminaries

We first prove that it suffices to bound the multilinear form over vectors whose coordinates are discretized:

Lemma 7.5. *Let $t \geqslant 2$ and $\boldsymbol{T}$ be a t-uniform tensor such that $T_{\boldsymbol{i}} = 0$ for all t-tuple $\boldsymbol{i}$ that contains some element multiple times. Then:*

$$\max_{\boldsymbol{x} \in \mathbb{R}^n} \frac{\left|\left\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes t} \right\rangle\right|}{\|\boldsymbol{x}\|_2^t} \leqslant 2^t \cdot \max_{\boldsymbol{x} \in \mathcal{P}} \frac{\left|\left\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes t} \right\rangle\right|}{\|\boldsymbol{x}\|_2^t}$$

where $\mathcal{P} := (\{\pm 2^{-i} : i \geqslant 1\} \cup \{0\})^n$ *is the set of* n*-dimensional vectors whose nonzero entries are negative powers of two.*

Proof. The proof is a simple extension of an argument of [BL06]. Let $\boldsymbol{x} \in \mathbb{R}^n$ be normalized so that $\|\boldsymbol{x}\|_\infty \leqslant \frac{1}{2}$. Then, for any $i \in [n]$, there exists $t_i \in [1, \infty]$, $\varepsilon_i \in \{-1, 1\}$, and $\delta_i \in [0, 1)$ such that $x_i = \varepsilon_i(1 + \delta_i)2^{-t_i}$. Now, set, independently for every i,

$$\widetilde{x}_i := \begin{cases} \varepsilon_i 2^{-t_i+1} & \text{with probability } \delta_i \\ \varepsilon_i 2^{-t_i} & \text{with probability } 1 - \delta_i \end{cases}.$$

Clearly, we always have $\widetilde{x}_i^2 \leqslant 4x_i^2$, and $\mathbb{E}\,\widetilde{x}_i = x_i$ by construction. Since $\boldsymbol{x} \mapsto \langle \boldsymbol{T}, \boldsymbol{x}^{\otimes t}\rangle$ is multilinear by assumption, we get $\mathbb{E}\langle \boldsymbol{T}, \widetilde{\boldsymbol{x}}^{\otimes t}\rangle = \langle \boldsymbol{T}, \boldsymbol{x}^{\otimes t}\rangle$, and in particular there exists a test vector in $\mathcal{P}$ whose normalized multilinear form is at least a $4^{-t/2} = 2^{-t}$ fraction of the unconstrained optimum. □

Next, we prove that a boun/ over $\{0, 1\}$-valued test vectors can be lifted to a bound over $\{-1, 0, 1\}$-valued test vectors up to a constant depending only on t:

Lemma 7.6. *Let* $t \geqslant 2$ *and* $\boldsymbol{T}$ *be a* t*-uniform tensor. Then,*

$$\max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t \in \{-1,0,1\}^n} \frac{|\langle \boldsymbol{T}, \boldsymbol{x}_1 \otimes \ldots \otimes \boldsymbol{x}_t\rangle|}{\|\boldsymbol{x}_1\|_2 \ldots \|\boldsymbol{x}_t\|_2} \leqslant 2^t \cdot \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t \in \{0,1\}^n} \frac{|\langle \boldsymbol{T}, \boldsymbol{x}_1 \otimes \ldots \otimes \boldsymbol{x}_t\rangle|}{\|\boldsymbol{x}_1\|_2 \ldots \|\boldsymbol{x}_t\|_2}.$$

Proof. Write any $\boldsymbol{x} \in \{-1, 0, 1\}^n$ as $\boldsymbol{x} = \boldsymbol{x}^+ - \boldsymbol{x}^-$ for $\boldsymbol{x}^+, \boldsymbol{x}^- \in \{0, 1\}^n$. By the triangle inequality,

$$\begin{aligned} |\langle \boldsymbol{T}, \boldsymbol{x}_1 \otimes \ldots \otimes \boldsymbol{x}_t\rangle| &= \left|\left\langle \boldsymbol{T}, (\boldsymbol{x}_1^+ - \boldsymbol{x}_1^-) \otimes \ldots \otimes (\boldsymbol{x}_t^+ - \boldsymbol{x}_t^-)\right\rangle\right| \\ &\leqslant \sum_{S \in \{-,+\}^k} \left|\left\langle \boldsymbol{T}, \boldsymbol{x}_1^{S_1} \otimes \ldots \otimes \boldsymbol{x}_t^{S_t}\right\rangle\right| \\ &\leqslant 2^t \cdot \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t \in \{0,1\}^n} \frac{|\langle \boldsymbol{T}, \boldsymbol{x}_1 \otimes \ldots \otimes \boldsymbol{x}_t\rangle|}{\|\boldsymbol{x}_1\|_2 \ldots \|\boldsymbol{x}_t\|_2}. \end{aligned}$$

This concludes the proof. □

7.2.2. Tensors of even arity

In this subsection, we prove:

Theorem 7.7. *Let* $\varepsilon > 0$*. Suppose* $\boldsymbol{T} \sim \mathcal{H}_4(n, p)$ *with* $p \leqslant O(n^{-1-\varepsilon})$*. Then, with high probability,*

$$\|\boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}\|_2 \leqslant K \log^2 \alpha,$$

where K *is a constant depending only on* ε*, and* $\alpha := \max_{1 \leqslant i,j \leqslant n} \left|\{(k, \ell) \in [n]^2 : T_{ijk\ell} = 1\}\right|$.

This shows that random Erdős-Renyi hypergraphs exhibit a spectral gap whenever $\|\boldsymbol{T}\|_2 \gg \log^2 \beta$, which happens when $p \gg (\log\log n)^2 \cdot n^{-2}$. This is a polyloglog-factor away from the predicted phase transition threshold.

To prove Theorem 7.7, we start by bounding the *decoupled* form over all $\{0,1\}$-valued test vectors.

Lemma 7.8. *Let $\boldsymbol{T}$ and $\varepsilon > 0$ be as Theorem 7.7. Then with high probability,*

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}\in\{0,1\}^n} \frac{|\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\otimes\boldsymbol{u}\rangle|}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2\|\boldsymbol{z}\|_2\|\boldsymbol{u}\|_2} \leqslant K\,,$$

for some constant K depending only on ε.

Proof. Let E_{abcd} ($0 \leqslant a \leqslant b \leqslant c \leqslant d \leqslant n$) be the event that there exist $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u} \in \{0,1\}^n$ such that $\boldsymbol{x}$ (resp. $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{u}$) has exactly a nonzero entries (resp. b, c, d) and

$$|\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\otimes\boldsymbol{u}\rangle| > K\sqrt{abcd} = K\cdot P\,.$$

where, to simplify notations, we define $P \coloneqq P(a,b,c,d) = \sqrt{abcd}$. Our goal is to show that $\bigcup_{a,b,c,d} E_{abcd}$ occurs only with negligible probability. We proceed by grouping together E_{abcd} based on the values of $a \leqslant b \leqslant c \leqslant d$.

Large P. First, suppose that $P \geqslant n$. Then, by Bernstein inequality applied with variance proxy bounded by $pabcd = pP^2$ and uniform bound 1 on the summands,

$$-\log\Pr(E_{abcd}) \geqslant \frac{K^2}{2}\cdot\frac{P^2}{pP^2 + \frac{1}{3}KP} \geqslant \frac{K^2}{4}\min\left(\frac{1}{p}, \frac{3P}{K}\right).$$

Since by assumption we have $1/p = \Omega(n^{1+\varepsilon})$ and $P \geqslant n$, whenever K is large enough, this overcounts the number of choices for the support of $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}$, which is clearly at most 2^{4n}. Therefore, we can take a union bound to show that none of the E_{abcd} satisfying $P \geqslant n$ occurs, with probability at least $1 - n^{-10}$.

Therefore, from now on we assume $P < n$. In this regime, the contribution from the expectation of $\boldsymbol{T}$ is negligible:

Claim 7.9. *For any $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u} \in \{0,1\}^n$ with respectively a,b,c,d nonzero entries,*

$$\langle \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\otimes\boldsymbol{u}\rangle \leqslant pP^2\,.$$

When $P < n$, we have $pP^2 \leqslant npP \leqslant P$, so it suffices to upper bound

$$\Pr\left(\langle \boldsymbol{T}, \boldsymbol{x}\otimes\boldsymbol{y}\otimes\boldsymbol{z}\otimes\boldsymbol{u}\rangle > \frac{K}{2}\cdot P\right).$$

Large d. Now, suppose that $a+b+c+d > \sqrt{abcd} = P$. In particular, we have $4d > P$.

By Claim 7.9, if E_{abcd} occurs, then we must have $\langle \boldsymbol{T}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \otimes \boldsymbol{u} \rangle > \frac{K}{2} \cdot P$ for some $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}$ with a, b, c, d nonzero entries. By averaging, it must be that for some $i, j, k \in [n]$,

$$\sum_{\ell=1}^{n} T_{ijk\ell} \geqslant \sum_{\ell=1}^{n} T_{ijkl} x_i y_j z_k u_\ell \geqslant \frac{1}{abc} \cdot \frac{K}{2} \cdot P = \frac{Kd}{2P} > \frac{K}{8},$$

using the identity $abc = \frac{p^2}{d}$ and the assumption $4d > P$. Finally, we use:

Claim 7.10. *For any $C > 0$, there exists $K(C) > 0$ such that $\max_{i,j,k \in [n]} \sum_{\ell=1}^{n} T_{ijk\ell} \leqslant K(C)$ with probability at least $1 - n^{-C}$.*

Therefore, the probability that any E_{abcd} satisfying $a+b+c+d > P$ occurs is at most n^{-10}, uniformly over all such a, b, c, d.

Remaining case. Finally, suppose that $a+b+c+d \leqslant P \leqslant n$. In this regime, the number of choices for $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}$ is at most

$$\binom{n}{a}\binom{n}{b}\binom{n}{c}\binom{n}{d} \leqslant n^{a+b+c+d}.$$

On the other hand, for any fixed $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}$, $\langle \boldsymbol{T}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \otimes \boldsymbol{u} \rangle$ is the sum of independent Bernoulli random variables with mean pP^2, so by the multiplicative version of the Chernoff bound (see, e.g., Claim 7.16 below),

$$\Pr(E_{abcd}) \leqslant n^{a+b+c+d} \cdot \exp\left(-\frac{1}{3} KP \log\left(\frac{K}{pP}\right)\right). \tag{7.5}$$

(This expansion is valid since $\frac{K}{pP} \geqslant 4$ in our regime of parameters.)

The quantity in the exponential in (7.5) exceeds $10P \log n \geqslant 10(a+b+c+d) \log n$ when K is a large enough constant depending on ε, so we conclude that no E_{abcd} satisfying the assumption can occur, with probability $1 - n^{-10}$.

Conclusion. Our three cases cover all regimes for a, b, c, d. It remains to take a union bound over the n^4 possible values for a, b, c, d. Therefore, the statement holds with high probability, which concludes the proof. □

Our next lemma lifts the bound from $\{0,1\}$-valued vectors to any vectors whose entries are negative powers of 2:

Lemma 7.11. *Suppose that $\boldsymbol{T} \sim \mathcal{H}_4(n,p)$, where $p \leqslant O(n^{-1-\varepsilon})$ for some constant $\varepsilon > 0$. Then with high probability,*

$$\max_{\boldsymbol{x}\in\mathcal{P}} \frac{\left|\left\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{\otimes 4}\right\rangle\right|}{\|\boldsymbol{x}\|_2^4} \lesssim K \log^2 \alpha\,,$$

where $\mathcal{P}$ is defined in Lemma 7.5, $K = K(\varepsilon)$ is the constant from Lemma 7.8, and

$$\alpha := \max_{i,j\in[n]} \left|\{(k,\ell) \in [n]^2 : T_{ijk\ell} = 1\}\right|\,.$$

Proof. Let $\boldsymbol{x} \in \mathcal{P}$, decomposed so that $\boldsymbol{x} = \sum_{i\geqslant 1} 2^{-i}\boldsymbol{x}^{(i)}$ where $x_j^{(i)} = \pm 1$ if $x_j = \pm 2^{-i}$ and 0 otherwise. If we denote by s_i the number of nonzero entries in $\boldsymbol{x}^{(i)}$, it follows that $\|\boldsymbol{x}\|_2^2 = \sum_i 2^{-2i}s_i$. We can assume without loss of generality that $\|\boldsymbol{x}\|_2 = 1$. First,

$$|\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{\otimes 4}\rangle| \leqslant \sum_{i,j,k,\ell=1}^{n} 2^{-(i+j+k+\ell)} |\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{(i)}\otimes\boldsymbol{x}^{(j)}\otimes\boldsymbol{x}^{(k)}\otimes\boldsymbol{x}^{(\ell)}\rangle|\,.$$

Fix some integer parameter $\gamma \geqslant 1$ to be fixed later.

- Let $\mathcal{I} := \{i \leqslant j \leqslant k \leqslant \ell : k+\ell \leqslant i+j+\gamma\}$. By Lemmas 7.6 and 7.8, it holds, with high probability,

$$\begin{aligned} 2^{-(i+j+k+\ell)} |\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{(i)}\otimes\boldsymbol{x}^{(j)}\otimes\boldsymbol{x}^{(k)}\otimes\boldsymbol{x}^{(\ell)}\rangle| &\leqslant K\sqrt{2^{-2i}s_i 2^{-2j}s_j 2^{-2k}s_k 2^{-2\ell}s_\ell} \\ &\leqslant 2K\left(2^{-2i}s_i 2^{-2j}s_j + 2^{-2k}s_k 2^{-2\ell}s_\ell\right)\,. \end{aligned}$$

  Plugging into our original left-hand side, we obtain

$$\begin{aligned} \sum_{(i,j,k,\ell)\in\mathcal{I}} 2^{-(i+j+k+\ell)} |\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{(i)}\otimes\boldsymbol{x}^{(j)}\otimes\boldsymbol{x}^{(k)}\otimes\boldsymbol{x}^{(\ell)}\rangle| &\leqslant 4K \sum_{(i,j,k,l)\in\mathcal{I}}^{n} 2^{-2i-2j} s_i s_j \\ &\leqslant 4K \sum_{i,j=1}^{n} 2^{-2i-2j} s_i s_j \sum_{i+j\leqslant k+\ell\leqslant i+j+\gamma} 1 \\ &\leqslant 4K\gamma^2\,, \end{aligned}$$

  using the normalization $\sum_i 2^{-2i}s_i = 1$.
- Now, consider the indices $\mathcal{J} := \{i \leqslant j \leqslant k \leqslant \ell : k+\ell > i+j+\gamma\}$. For all $i, j \in [n]$, by the triangle inequality we have

$$\sum_{k,\ell=1}^{n} |\langle \boldsymbol{T} - \mathbb{E}\,\boldsymbol{T}, \boldsymbol{x}^{(i)}\otimes\boldsymbol{x}^{(j)}\otimes\boldsymbol{x}^{(k)}\otimes\boldsymbol{x}^{(\ell)}\rangle| \leqslant \langle \boldsymbol{T}, |\boldsymbol{x}^{(i)}|\otimes|\boldsymbol{x}^{(j)}|\otimes\sum_{k,\ell=1}^{n}|\boldsymbol{x}^{(k)}|\otimes|\boldsymbol{x}^{(\ell)}|\rangle$$

$$\leqslant s_i s_j \max_{i,j\in[n]} \sum_{k,\ell=1}^{n} T_{ijk\ell} = s_i s_j \alpha\,,$$

where we are using that the nonzero entries of $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^{(\ell)}$ are disjoint. Plugging back in,

$$\begin{aligned}
&\sum_{(i,j,k,\ell)\in\mathcal{J}} 2^{-(i+j+k+\ell)} |\langle \boldsymbol{T} - \mathbb{E}\boldsymbol{T}, \boldsymbol{x}^{(i)} \otimes \boldsymbol{x}^{(j)} \otimes \boldsymbol{x}^{(k)} \otimes \boldsymbol{x}^{(\ell)} \rangle| \\
\leqslant &\sum_{(i,j,k,\ell)\in\mathcal{J}} 2^{-(2i+2j+\gamma)} |\langle \boldsymbol{T} - \mathbb{E}\boldsymbol{T}, \boldsymbol{x}^{(i)} \otimes \boldsymbol{x}^{(j)} \otimes \boldsymbol{x}^{(k)} \otimes \boldsymbol{x}^{(\ell)} \rangle| \\
\leqslant &\alpha 2^{-\gamma} \sum_{i,j=1}^{n} 2^{-2(i+j)} s_i s_j = \alpha 2^{-\gamma}\,.
\end{aligned}$$

Combining both cases and picking $\gamma = \log \alpha$, we obtain

$$|\langle \boldsymbol{T} - \mathbb{E}\boldsymbol{T}, \boldsymbol{x}^{\otimes 4} \rangle| \leqslant 4K\gamma^2 + \alpha 2^{-\gamma} \lesssim K \log^2 \alpha\,,$$

as desired. □

Proof of Theorem 7.7. The proof directly follows from combining Lemmas 7.5 and 7.11. □

7.2.3. Generalization to odd-arity tensors

We briefly explain how to generalize the strategy of §7.2.2 to odd-arity tensors, focusing on $t = 3$. We will prove a much better bound in the next section.

Theorem 7.12. *Suppose $\boldsymbol{T} \sim \mathcal{H}_3(n,p)$ where $p \leqslant O(n^{-1-\varepsilon})$ for some constant $\varepsilon > 0$. Then with high probability,*

$$\|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|_2 \leqslant K \log^2 \alpha$$

where $K = K(\varepsilon)$ is a constant depending only on ε, and

$$\alpha := \max_{1\leqslant i,j\leqslant n} \left|(a,k,\ell) : T_{aik} = 1 \text{ and } T_{aj\ell} = 1\}\right|\,.$$

We only sketch the proof of Theorem 7.12, as it is very similar to §7.2.2. The main difference is that we start by applying the "Cauchy-Schwarz trick" (which is a classical idea for refuting random k-Xor instances for odd k). Let $\widetilde{\boldsymbol{T}} = \boldsymbol{T} - \mathbb{E}\boldsymbol{T}$ be the centered tensor. For any $x \in \mathcal{S}^{n-1}$,

$$\left(\sum_{1\leqslant i,j,k\leqslant n} \widetilde{T}_{ijk} x_i x_j x_k \right)^2 \leqslant \sum_{1\leqslant i,j_1,j_2,k_1,k_2\leqslant n} \widetilde{T}_{i,j_1,k_1} \widetilde{T}_{i,j_2,k_2} x_{j_1} x_{j_2} x_{k_1} x_{k_2}\,.$$

Let $M_{ij,k\ell} := \sum_{r=1}^{n} \widetilde{T}_{ikr}\widetilde{T}_{j\ell r}$ (note that the other way to flatten the tensor would give a very poor upper bound). We bound $(\boldsymbol{x}^{\otimes 2})^\top \boldsymbol{M}\boldsymbol{x}^{\otimes 2}$ for any $\boldsymbol{x} \in \mathcal{S}^{n-1}$. First, observe that the ℓ_1-norm of the row of $\boldsymbol{M}$ corresponding to the tuple (i, j) is

$$\sum_{1\leqslant k,\ell,r\leqslant n} \widetilde{T}_{ikr}\widetilde{T}_{j\ell r}.$$

For $p = n^{-1.5}$, the distribution of this random variable is close to Poisson(1). Hence the maximum ℓ_1-norm of a row of $\boldsymbol{M}$ is about $\log n/\log\log n$. This quantity will play the role of α in the odd-arity case.

The proof of Theorem 7.12 proceeds by first establishing that whenever $p \ll 1/n$,

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}\in\{0,1\}^n} \frac{\sum_{i,j,k,\ell=1}^{n} M_{ij,k\ell}x_i y_j z_k u_\ell}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2\|\boldsymbol{z}\|_2\|\boldsymbol{u}\|_2} \leqslant O(1)\,,$$

using a case analysis similar to the previous subsection. Then, for any $\boldsymbol{x} = \sum_i \boldsymbol{x}^{(i)}2^{-i}$ with $\boldsymbol{x}^{(i)} \in \{-1, 0, 1\}^n$, we write

$$(\boldsymbol{x}^{\otimes 2})^\top \boldsymbol{M}\boldsymbol{x}^{\otimes 2} = \sum_{i,j,k,\ell\geqslant 0} 2^{-(i+j+k+\ell)}(\boldsymbol{x}^{(i)} \otimes \boldsymbol{x}^{(j)})^\top \boldsymbol{M}(\boldsymbol{x}^{(k)} \otimes \boldsymbol{x}^{(\ell)})\,.$$

A similar argument to the previous subsection shows that the contribution of indices $i \leqslant j \leqslant k \leqslant \ell$ with $k + \ell \leqslant i + j + \gamma$ to the quadratic form is at most $O(\gamma^2)$, and the contribution of remaining indices is at most $O(\alpha)$, where α is the maximal ℓ_1-norm of a row of $\boldsymbol{M}$. Setting $\gamma = \log\alpha$, we get a $\log^2\alpha$ upper bound on the second eigenvalue of $\boldsymbol{T}$, which is $(\log\log n)^2$ for a random 3-uniform hypergraph with $n^{1.5}$ hyperedges.

7.3. A direct multiscale union bound

In this final section, we prove Theorem 7.3. The lower bound comes from the all ones test vector that achieves cubic form $pn^{1.5}$. We now focus on establishing the upper bound. We fix $\boldsymbol{T}$ to be a random 3-tensor satisfying the assumptions of Theorem 7.3. Our proof strategy borrows ideas from Kahn and Szemerédi, and Feige and Ofek [FKS89, FO05].

7.3.1. Preliminaries

Notation 7.13. Given $A, B, C \subseteq [n]$ and a tensor $\boldsymbol{T} \in \mathbb{R}^{n^3}$, we let

$$\boldsymbol{T}(A, B, C) := \sum_{a\in A}\sum_{b\in B}\sum_{c\in C} T_{abc}\,.$$

To bound the norm of $\boldsymbol{T} - \mathbb{E}\boldsymbol{T}$, we start by discretizing test vectors into negative powers of 2. By Lemma 7.5, this can be done at the cost of a constant factor only. That is, we focus from now on on bounding the right-hand side of:

Lemma 7.14.

$$\|\boldsymbol{T} - \mathbb{E}\boldsymbol{T}\|_2 \lesssim \max_{\mathcal{A},\mathcal{B},\mathcal{C}} \sum_{i,j,k\geqslant 0} 2^{-(i+j+k)} (\boldsymbol{T} - \mathbb{E}\boldsymbol{T})(A_i, B_j, C_k)\,,$$

where the maximum runs over all 3-tuples of partitions $\mathcal{A} \coloneqq (A_i)_{i\geqslant 0}$, $\mathcal{B} \coloneqq (B_j)_{j\geqslant 0}$, $\mathcal{C} \coloneqq (C_k)_{k\geqslant 0}$ *of* $[n]$ *satisfying* $\sum_{i\geqslant 0} |A_i| 2^{-2i} = \sum_{j\geqslant 0} |B_j| 2^{-2j} = \sum_{k\geqslant 0} |C_k| 2^{-2k} = 1$.

Definition 7.15. We say that a partition $\mathcal{A} = (A_i)_{i\geqslant 0}$ is *normalized* if $\sum_{i\geqslant 0} 2^{-2i}|A_i| = 1$.

We will use repeatedly the following version of Chernoff bounds.

Claim 7.16. *Let* $X_1, \dots, X_n$ *be independent Bernoulli random variables and let*

$$\mu \coloneqq \sum_{i=1}^{n} \Pr(X_i = 1)\,.$$

Then for all $t \geqslant 4$,

$$\Pr\left(\sum_{i=1}^{n} X_i \geqslant t\mu\right) \leqslant e^{-\frac{1}{3}t\mu \log t}\,.$$

7.3.2. Technical lemmas

Let $\delta > 0$ to be fixed later.

Lemma 7.17 (Dense test vectors)**.** *Let* $S_1 = \{(i,j,k) : 2^{-(i+j+k)} \leqslant \frac{1}{n}\}$. *Then with probability* $1 - e^{-\Omega(n)}$, *it holds for all normalized partitions* $\mathcal{A}, \mathcal{B}, \mathcal{C}$*:*

$$\sum_{(i,j,k)\in S_1} 2^{-(i+j+k)} (\boldsymbol{T} - \mathbb{E}\boldsymbol{T})(A_i, B_j, C_k) = O(1)\,.$$

Proof. First, we keep only indices i, j, k such that $2^{-i}, 2^{-j}, 2^{-k} \geqslant \frac{1}{n^4}$ (clearly, the sum over the other indices can contribute at most $O(1)$). In particular, we only need to take a union bound against $\lesssim (\log n)^3 \cdot 2^{3n} = e^{O(n)}$ elements.

Up to the symmetries, the left-hand side is a sum of independent random variables bounded by $\frac{1}{n}$, with variance proxy:

$$\sum_{i,j,k\geqslant 0} 2^{-(2i+2j+2k)} \sum_{a\in A_i} \sum_{b\in B_j} \sum_{c\in C_k} \mathbb{E}\left[(T_{abc} - p)^2\right] \lesssim p\,.$$

Therefore, by Bernstein inequality, the left-hand side in the statement of the lemma exceeds t with probability at most $\min(e^{-Cnt}, e^{-Ct^2/p})$ for some constant $C > 0$. Since $p \ll \frac{1}{n}$, using large enough constants, we get that it is $O(1)$ with probability at least $1 - e^{-\Omega(n)}$. We conclude by taking a union bound. □

Claim 7.18 (Target union bound). *Suppose that for any fixed sets $A, B, C \subseteq [n]$, we know that $\boldsymbol{T}(A,B,C) \leqslant f(|A|,|B|,|C|)$ with probability at least $1-n^{-\Omega(|A|+|B|+|C|)}$. Then with probability $1 - \frac{1}{\mathrm{poly}(n)}$, $\boldsymbol{T}(A,B,C) \leqslant f(|A|,|B|,|C|)$ holds simultaneously for every $A, B, C \subseteq [n]$.*

Proof. We first take a union bound over all A, B, C of fixed size $a, b, c \geqslant 0$ to get that $\boldsymbol{T}(A,B,C) \leqslant f(|A|,|B|,|C|)$ for all sets A, B, C of size a, b, c simultaneously with probability $1 - (\binom{n}{a}\binom{n}{b}\binom{n}{c})^{-\Omega(1)}$. By another union bound, this then holds for all sets $A, B, C \subseteq [n]$ with probability $1 - \frac{1}{\mathrm{poly}(n)}$. □

Lemma 7.19 (Case $|A_i| \cdot |B_j| \gg n$). *Let*

$$S_2 = S_2(A,B,C) := \left\{(i,j,k) : 2^{-(i+j+k)} \geqslant \frac{1}{n}, \frac{|A_i||B_j||C_k|}{\max(|A_i|,|B_j|,|C_k|)} \geqslant n \log n\right\}.$$

Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$, it holds for all normalized partitions $\mathcal{A}, \mathcal{B}, \mathcal{C}$ that

$$\sum_{(i,j,k)\in S_2} 2^{-(i+j+k)}(\boldsymbol{T} - \mathbb{E}\boldsymbol{T})(A_i, B_j, C_k) = O(1).$$

Proof. By Claim 7.16, $(\boldsymbol{T}-\mathbb{E}\boldsymbol{T})(A_i,B_j,C_k) \lesssim \frac{|A_i||B_j||C_k|}{n}$ holds with probability $1-e^{-\frac{C|A_i||B_j||C_k|}{n}}$ for some constant $C > 0$. Now if $(i,j,k) \in S_2$, we know that $\frac{|A_i||B_j||C_k|}{n}$ is at least $\log n \cdot \max(|A_i|,|B_j|,|C_k|)$. Hence, by Claim 7.18, we get that with probability $1 - \frac{1}{\mathrm{poly}(n)}$,

$$\begin{aligned}\sum_{(i,j,k)\in S_2} 2^{-(i+j+k)}\boldsymbol{T}(A_i,B_j,C_k) &\leqslant \sum_{(i,j,k)\in S_2} 2^{-(i+j+k)}\frac{|A_i||B_j||C_k|}{n}\\ &\leqslant \sum_{i,j,k\geqslant 0} 2^{-(2i+2j+2k)}|A_i||B_j||C_k|\\ &= 1.\end{aligned}$$

□

Lemma 7.20 (Bound on the expectation). *For any normalized partitions $\mathcal{A}, \mathcal{B}, \mathcal{C}$,*

$$\sum_{i,j,k\geqslant 0:|A_i||B_j||C_k|\leqslant n^{2+\varepsilon}} 2^{-(i+j+k)}(\mathbb{E}\boldsymbol{T})(A_i,B_j,C_k) = O(1).$$

Proof. This equals

$$n^{-(1+\varepsilon)}\sum_{i,j,k\geqslant 0} 2^{-(i+j+k)}|A_i||B_j||C_k| \leqslant n^{-\frac{\varepsilon}{2}}\sum_{i,j,k\geqslant 0} 2^{-(i+j+k)}\sqrt{|A_i||B_j||C_k|} = O(1),$$

where the last inequality follows from Cauchy-Schwarz and the normalization of $\mathcal{A}, \mathcal{B}, \mathcal{C}$. □

Lemma 7.21 (Irregular sizes). *Let*

$$S_3 = S_3(A, B, C) := \{(i, j, k) : \max(i, j) \leqslant k, \max(|A_i|, |B_j|) \geqslant |C_k|, \frac{|A_i||B_j||C_k|}{\max(|A_i|, |B_j|)} \leqslant \frac{n^{-\delta}}{p}\}.$$

Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$, *it holds for all normalized partitions* $\mathcal{A}, \mathcal{B}, \mathcal{C}$ *that*

$$\sum_{(i,j,k)\in S_3} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) = O(1).$$

Proof. By a Chernoff bound, we have $\boldsymbol{T}(A_i, B_j, C_k) = O(\max(|A_i|, |B_j|))$ with probability at least $1 - n^{-\Omega(1)\cdot\max(|A_i|,|B_j|)}$, because the second assumption ensures that

$$\log\left(\frac{\max(|A_i|, |B_j|)}{|A_i||B_j||C_k|p}\right) = \Omega(\log n).$$

Therefore, by Claim 7.18, we have with probability $1 - \frac{1}{\mathrm{poly}(n)}$ that

$$\sum_{(i,j,k)\in S_3} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) \lesssim \sum_{i,j\geqslant 0} 2^{-(i+j)} \max(|A_i|, |B_j|) \sum_{k\geqslant \max(i,j)} 2^{-k} \leqslant 1.$$

This concludes the proof. □

Lemma 7.22 (Case $|A_i| \ll n^{0.01}$). *Let*

$$S_6 = S_6(A, B, C) := \left\{i, j, k : k \geqslant \max(i, j), \min(|A_i|, |B_j|) \leqslant \frac{1}{pn^{1+\delta}}\right\}.$$

Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$, *it holds for all normalized partitions* $\mathcal{A}, \mathcal{B}, \mathcal{C}$ *that*

$$\sum_{(i,j,k)\in S_6} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) = O(1).$$

Proof. The argument is symmetric in i and j, so we assume without loss of generality that $|A_i| \leqslant |B_j|$. Since $|A_i||B_j|pn \ll |B_j|$, by Claim 7.16 we have $\boldsymbol{T}(A_i, B_j, [n]) = O(|B_j|)$ with probability $1 - n^{-\Omega(|B_j|)}$. Then, by Claim 7.18, with probability $1 - \frac{1}{\mathrm{poly}(n)}$,

$$\begin{aligned}\sum_{(i,j,k)\in S_6} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) &\lesssim \sum_{i,j\geqslant 0} 2^{-(i+j)} \sum_{k\geqslant j} 2^{-k} \boldsymbol{T}(A_i, B_j, C_k) \\ &\leqslant \sum_{i,j\geqslant 0} 2^{-i} 2^{-2j} \boldsymbol{T}(A_i, B_j, [n]) \\ &= O(1).\end{aligned}$$

This concludes the proof. □

Lemma 7.23 (Case $|C_k| \ll |A_i| \cdot |B_j|$)**.** *Let*

$$S_4 \coloneqq \left\{(i,j,k) : \max(|A_i|,|B_j|) \leqslant |C_k| \leqslant (|A_i||B_j|)^{1-\delta}, 4|A_i||B_j| \leqslant p^{-1}\right\} .$$

Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$*, it holds for all normalized partitions* $\mathcal{A}, \mathcal{B}, \mathcal{C}$ *that*

$$\sum_{(i,j,k)\in S_4} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) = O(1) .$$

Proof. By Claim 7.16, we have $\boldsymbol{T}(A_i, B_j, C_k) = O(|C_k|)$ with probability $1 - n^{-\Omega(|C_k|)}$ (this works provided that $4|A_i||B_j| \leqslant p^{-1}$). Moreover, if this holds, since $2^{2i} \geqslant |A_i|$, $2^{2j} \geqslant |B_j|$, $2^{2k} \geqslant |C_k|$:

$$\sum_{(i,j,k)\in S_4} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) \lesssim \sum_{i,j,k\geqslant 0} 2^{-(i+j+k)} (|A_i||B_j||C_k|)^{\frac{1-\delta}{2}} = O(1) ,$$

and we conclude by Claim 7.18. □

Lemma 7.24 (Case $|C_k| \gg |B_j|$)**.** *Let*

$$S_5 = S_5(A,B,C) \coloneqq \{(i,j,k) : j \leqslant k, 4\min(|A_i|,|B_j|) \leqslant n^{0.5}\log n, \\ |C_k| \geqslant \max(n^{\delta}, |A_i|^{1+\delta}, |B_j|^{1+\delta})\} .$$

Then with probability at least $1 - \frac{1}{\mathrm{poly}(n)}$*, it holds for all normalized partitions* $\mathcal{A}, \mathcal{B}, \mathcal{C}$ *that*

$$\sum_{(i,j,k)\in S_5} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) = O(1) .$$

Proof. Imagine that $|A_i| \leqslant |B_j|$ to simplify (the argument will be symmetric by switching the roles of i and j). By Claim 7.16, we have $\boldsymbol{T}(A_i, B_j, [n]) = O(\log n \cdot |B_j|)$ with probability $1 - n^{-|B_j|}$ (note that $4|A_i||B_j|n^{-1/2} \leqslant |B_j|\log n$ holds by assumption). If this holds, then using $2^{-k} \leqslant |C_k|^{-\frac{1}{2}} \leqslant |B_j|^{-\frac{1}{2}-\frac{\delta}{4}} \cdot n^{-\frac{\delta^2}{4}}$, we get

$$\begin{aligned}\sum_{(i,j,k)\in S_5} 2^{-(i+j+k)} \boldsymbol{T}(A_i, B_j, C_k) &\leqslant \sum_{i,j\geqslant 0} 2^{-(i+j)} \boldsymbol{T}(A_i, B_j, [n]) |B_j|^{-\frac{1}{2}-\frac{\delta}{4}} \cdot n^{-\frac{\delta^2}{4}} \\ &\leqslant \sum_{i,j\geqslant 0} 2^{-(i+j)} |B_j|^{-\frac{1}{2}-\frac{\delta}{4}} \\ &\leqslant 1 ,\end{aligned}$$

where the last inequality follows from $|B_j| \leqslant 2^{2j}$. The result then follows from Claim 7.18. □

7.3.3. Putting everything together: proof of Theorem 7.3

Without loss of generality we treat the case of i, j, k satisfying $0 \leqslant i \leqslant j \leqslant k$. We prove that (i, j, k) necessarily belongs to $S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6$.

Suppose that the assumptions of Lemma 7.17 do not apply. Then $2^{-(i+j+k)} > \frac{1}{n}$. Now suppose that the assumptions of Lemma 7.19 do not apply. Then,

$$\frac{|A_i||B_j||C_k|}{\max(|A_i|, |B_j|, |C_k|)} \leqslant n \log n \,. \tag{7.6}$$

Now, suppose that the assumptions of Lemma 7.21 do not apply. Then by (7.6),

$$|C_k| \geqslant \max(|A_i|, |B_j|) \,. \tag{7.7}$$

In particular, by Lemma 7.20, the contribution of the expectation is negligible in this regime. So we reduce to bounding $\boldsymbol{T}$ instead of $T - \mathbb{E}\boldsymbol{T}$. Next, suppose that the assumptions of Lemma 7.23 do not apply. Then by (7.6) and (7.7),

$$|C_k| > (|A_i||B_j|)^{1-\delta} \,. \tag{7.8}$$

Next, suppose that the assumptions of Lemma 7.24 do not apply. Then,

$$|C_k| \leqslant n^{\delta} \quad \text{or} \quad |C_k| \leqslant \max(|A_i|, |B_j|)^{1+\delta} \tag{7.9}$$

In the first case, if we pick $\delta \leqslant \varepsilon/2$, it must be by (7.7) that $|A_i| \leqslant \frac{1}{pn^{1+\delta}}$, so Lemma 7.22 applies. In the second case, after combining with (7.8), we get

$$\max(|A_i|, |B_j|) \leqslant (|A_i||B_j|)^{\frac{1-\delta}{1+\delta}} \,,$$

so if we pick δ to be a small enough constant, then both $|A_i|$ and $|B_j|$ have to be $O(n^{\delta})$, and Lemma 7.22 applies again with the choice $\delta := \varepsilon/2$.

7.4. Summary

We proved a tight bound on the Friedman–Wigderson second eigenvalue of sparse Erdős-Renyi 3-uniform hypergraphs. Our argument avoids a generic chaining construction in favor of a carefully designed union bound.

CHAPTER 8.

Certifiable Approximation for Polynomial Optimization

In this chapter, we introduce new approximation algorithms for the problem of maximizing an arbitrary homogeneous, multilinear, cubic polynomial

$$f(\boldsymbol{x}) = \sum_{\substack{i,j,k=1 \\ i,j,k \text{ distinct}}}^{n} c_{ijk} x_i x_j x_k \,,$$

over $\boldsymbol{x} \in \mathcal{S}^{n-1}$ or $\boldsymbol{x} \in \{-1,1\}^n$. Our algorithms have several features:

1. They have certifiable approximation guarantees, in the sense of §1.2.4.
2. They come with an entire tradeoff between time and approximation.
3. Already in the polynomial-time regime, they improve over the approximation guarantees of the prior work [BGG+17].
4. They are based on a new technique for rounding higher-degree sum-of-squares relaxations that we introduce.

We give an application of our results to the design of improved approximation algorithms for Max-3-Sat (Problem 1.1).

Table of contents

These results appeared in [HKPT24].

8.1. Preliminaries

8.1.1. Sum-of-squares relaxations

We refer the reader to the monograph [FKP19] and the lecture notes [BS16] for a detailed exposition of the sum-of-squares method and its usage in algorithm design.

A *degree-ℓ pseudo-distribution* μ over variables $x_1, x_2, \ldots, x_n$ corresponds to a linear operator $\widetilde{\mathbb{E}}_\mu$ that maps polynomials of degree $\leqslant \ell$ to real numbers and satisfies $\widetilde{\mathbb{E}}_\mu\ 1 = 1$ and $\widetilde{\mathbb{E}}_\mu[p^2] \geqslant 0$ for every polynomial $p(x_1, x_2, \ldots, x_n)$ of degree $\leqslant \ell/2$. We say that such a pseudo-distribution satisfies the *hypercube* constraints if $\widetilde{\mathbb{E}}_\mu[px_i^2] = \widetilde{\mathbb{E}}_\mu[p]$ for every polynomial p of degree $\leqslant \ell - 2$ and every $i \in [n]$. We say that such a pseudo-distribution satisfies the *unit sphere* constraints if $\widetilde{\mathbb{E}}_\mu[\|\boldsymbol{x}\|_2^2 p] = \widetilde{\mathbb{E}}_\mu[p]$ for every p of degree $\leqslant \ell - 2$.

Given a polynomial p (with the ℓ_1-norm of the coefficients being $\|p\|_1$) over $x_1, x_2, \ldots, x_n$, a pseudo-distribution of degree ℓ over the unit sphere or the hypercube that maximizes p within an additive $\varepsilon\|p\|_1$ error can be found in time $n^{O(\ell)}\operatorname{polylog}(n/\varepsilon)$ via the ellipsoid method.

Reweighting. Given a pseudo-distribution μ over the unit sphere or the hypercube, a *reweighting* of μ by a sum-of-squares polynomial q satisfying $\widetilde{\mathbb{E}}_\mu[q] > 0$ is a pseudo-distribution μ' that maps any polynomial p to $\widetilde{\mathbb{E}}_{\mu'}[p] = \widetilde{\mathbb{E}}_\mu[pq]/\widetilde{\mathbb{E}}_\mu[q]$. For any μ of degree ℓ and q of degree $r < \ell$, μ' is a pseudo-distribution of degree at least $\ell - r$. Furthermore, if μ satisfies the unit sphere (or the hypercube) constraints then so does μ' as long as $r \leqslant \ell - 2$.

Sum-of-squares proofs. Let $f_1, f_2, \ldots, f_m$ and g be multivariate polynomials in x. A *sum-of-squares* proof that the constraints $\{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ imply $g \geqslant 0$ consists of sum-of-squares polynomials $(p_S)_{S\subseteq[m]}$ such that $g = \sum_{S\subseteq[m]} p_S \prod_{i\in S} f_i$. The *degree* of such a sum-of-squares proof equals the maximum of the degree of $p_S \prod_{i\in S} f_i$ over all S appearing in the sum above. We write $\{f_i \geqslant 0,\ \forall i \in [m]\} \left|\frac{x}{t}\right. \{g \geqslant 0\}$ where t is the degree of the sum-of-squares proof.

We recall the following connection between SoS proofs and pseudo-distributions:

Fact 8.1. *Suppose $\{f_i \geqslant 0, \forall i \in [m]\} \left|\frac{}{t}\right. \{g \geqslant 0\}$ for some polynomials f_i and g. Let μ be a pseudo-distribution of degree $\geqslant t$ satisfying $\{f_i \geqslant 0\}_{i\in[m]}$. Then, $\widetilde{\mathbb{E}}_\mu[g] \geqslant 0$.*

We next state some standard facts (see [FKP19] for references).

Fact 8.2 (SoS generalized triangle inequality)**.** *Let $k \in \mathbb{N}$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$ be indeterminates.*

$$\{x_i \geqslant 0,\ \forall i \in [n]\} \left|\frac{x_1,\ldots,x_n}{k}\right. \left\{\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^k \leqslant \frac{1}{n}\sum_{i=1}^n x_i^k\right\}.$$

Moreover, if k is even, then

$$\vdash_{k}^{x_1,\dots,x_n} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^k \leqslant \frac{1}{n} \sum_{i=1}^{n} x_i^k \right\} .$$

Note that the $k = 2$ case is the SoS version of the Cauchy-Schwarz inequality for vectors.

Fact 8.3 (Cauchy-Schwarz for pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $\boldsymbol{x}$. Then, for any degree-$2d$ pseudo-distribution μ over $\boldsymbol{x}$, we have*

$$\widetilde{\mathbb{E}}_{\mu}[fg] \leqslant \sqrt{\widetilde{\mathbb{E}}_{\mu}[f^2]} \sqrt{\widetilde{\mathbb{E}}_{\mu}[g^2]} .$$

Fact 8.4 (Hölder's inequality for pseudo-distributions)**.** *Let f, g be polynomials of degree at most d in indeterminate $\boldsymbol{x}$, and fix any even $t \in \mathbb{N}$. Then, for any degree-td pseudo-distribution μ over $\boldsymbol{x}$, we have*

$$\widetilde{\mathbb{E}}_{\mu}[f^{t-1}g] \leqslant \left(\widetilde{\mathbb{E}}_{\mu}[f^t] \right)^{\frac{t-1}{t}} \left(\widetilde{\mathbb{E}}_{\mu}[g^t] \right)^{\frac{1}{t}} .$$

Furthermore, for any $t \in \mathbb{N}$ and any degree-$2td$ pseudo-distribution μ, we have $\widetilde{\mathbb{E}}_{\mu}\left[f^{2t-2}\right] \leqslant \widetilde{\mathbb{E}}_{\mu}\left[f^{2t}\right]^{\frac{t-1}{t}}$.

8.1.2. Roundings for quadratic polynomial optimization

The problem of worst-case quadratic polynomial optimization with certifiable guarantees was conjecturally settled 20 years ago in a series of work that used semidefinite programming to design approximation algorithms. The starting point was the 0.878-approximation algorithm of Goemans and Williamson [GW95] for Max-Cut. The algorithm of Goemans and Williamson is based on rounding the degree-2 sum-of-squares relaxation of Max-Cut using *random hyperplane rounding*.

Example 8.5 (Random hyperplane rounding)**.** We identify vertex cuts with vectors in $\{-1, 1\}^n$ in the natural way. The fraction of edges cut by $\boldsymbol{x} \in \{-1, 1\}^n$ in a graph $G = (V, E)$ can be expressed as

$$f_G(\boldsymbol{x}) = \frac{1}{4} \mathop{\mathbb{E}}_{\{u,v\}\sim E} (x_u - x_v)^2 . \tag{8.1}$$

Let μ be any degree-2 pseudo-distribution on $\boldsymbol{x} \in \{-1, 1\}^n$. Since $\widetilde{\mathbb{E}}_{\mu} f_G(\boldsymbol{x}) = \widetilde{\mathbb{E}}_{\mu} f_G(-\boldsymbol{x})$, up to symmetrizing the pseudo-distribution, we assume without loss of generality that $\widetilde{\mathbb{E}}_{\mu} \boldsymbol{x} = 0$. Random hyperplane rounding proceeds by:

1. First, drawing $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \widetilde{\mathbb{E}}_\mu \boldsymbol{x}\boldsymbol{x}^\top)$ matching the first two moments of μ.
2. Then, letting $\overline{x}_i := \mathrm{sign}(g_i)$ for all $i \in [n]$.

The analysis is based on the following inequality on two-dimensional Gaussians: for any jointly Gaussian random variables (g, h),

$$\mathbb{E}\left[(\mathrm{sign}(g) - \mathrm{sign}(h))^2\right] \geqslant K_{\mathrm{GW}} \cdot \mathbb{E}\left[(g-h)^2\right],$$

where $K_{\mathrm{GW}} \approx 0.878$.

Random hyperplane rounding and its generalization RPR^2, i.e., randomized projection followed by randomized rounding [FL06], are the main rounding technique for quadratic polynomial optimization. We proceed to describe the best-known rounding algorithms for maximizing quadratic polynomials with *arbitrary* coefficients.

Spherical optimization. Maximizing a degree-2 polynomial over the sphere is equivalent to computing a maximal eigenvalue, which can be done in polynomial time without loss in the approximation ratio. This can of course be captured by a rounding algorithm of degree-2 sum-of-squares.

Lemma 8.6 (Lossless rounding on the unit sphere)**.** *Given any degree-2 pseudo-distribution μ over $\boldsymbol{x} \in \mathcal{S}^{n-1}$ and $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, there is an algorithm that outputs $\overline{\boldsymbol{x}} \in \mathcal{S}^{n-1}$ such that*

$$\langle \overline{\boldsymbol{x}}, \boldsymbol{M}\overline{\boldsymbol{x}} \rangle \geqslant \widetilde{\mathop{\mathbb{E}}_{\mu}} \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle \ .$$

Proof. Let $\boldsymbol{X} := \widetilde{\mathbb{E}}_\mu [\boldsymbol{x}\boldsymbol{x}^\top]$ and write the eigendecomposition $\boldsymbol{X} = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$, where $\lambda_i \geqslant 0$ and $\boldsymbol{v}_i \in \mathcal{S}^{n-1}$ for all $i \in [n]$. Since μ is a pseudo-distribution over the unit sphere, we have

$$\mathrm{tr}(\boldsymbol{X}) = \sum_{i=1}^n \lambda_i = 1\,,$$

so $\{\lambda_i\}_{i \in [n]}$ defines a valid probability distribution. Now sample $\overline{\boldsymbol{x}} = \boldsymbol{v}_i$ with probability λ_i for all $i \in [n]$. Then,

$$\mathbb{E}\,\langle \overline{\boldsymbol{x}}, \boldsymbol{M}\overline{\boldsymbol{x}} \rangle = \sum_{i=1}^n \lambda_i \langle \boldsymbol{v}_i, \boldsymbol{M}\boldsymbol{v}_i \rangle = \langle \boldsymbol{X}, \boldsymbol{M} \rangle = \widetilde{\mathop{\mathbb{E}}_{\mu}} \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle \ .$$

In particular, one of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ must have quadratic form at least $\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle$. □

Hypercube optimization. The analog rounding algorithm for maximizing degree-2 polynomial on the hypercube is due to Charikar and Wirth [CW04]. Its approximation ratio is $O(\log n)$, and there is evidence that this is best possible in polynomial time [ABH+05].

Theorem 8.7 (Charikar-Wirth rounding [CW04]). *For any $T > 0$, given any degree-2 pseudo-distribution μ over $\boldsymbol{x} \in \{\pm 1\}^n$, there is a polynomial-time sampleable distribution $\mathcal{D}$ supported on $\{\pm 1\}^n$ such that for any matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ with zero diagonal entries,*

$$\mathop{\mathbb{E}}_{\overline{\boldsymbol{x}} \sim \mathcal{D}} \langle \overline{\boldsymbol{x}}, \boldsymbol{M}\overline{\boldsymbol{x}} \rangle \geqslant \frac{1}{T^2} \cdot \widetilde{\mathbb{E}}_{\mu} \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle - 8e^{-\frac{T^2}{2}} \sum_{i,j=1}^{n} |M_{ij}| \,.$$

In particular, if μ is an optimal pseudo-distribution for the degree-2 SoS relaxation of

$$\max_{\boldsymbol{x} \in \{\pm 1\}^n} \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle \,,$$

then by picking $T = \Theta(\sqrt{\log n})$,

$$\mathop{\mathbb{E}}_{\overline{\boldsymbol{x}} \sim \mathcal{D}} \langle \overline{\boldsymbol{x}}, \boldsymbol{M}\overline{\boldsymbol{x}} \rangle \geqslant \Omega\left(\frac{1}{\log n}\right) \cdot \max_{\boldsymbol{x} \in \{\pm 1\}^n} \langle \boldsymbol{x}, \boldsymbol{M}\boldsymbol{x} \rangle \,.$$

The rounding in the proof of Theorem 8.7 is a truncated hyperplane rounding. First, sample a Gaussian $\boldsymbol{g}$ with first and second moments matching the corresponding pseudo-moments of μ, and set

$$\overline{x}_i := \begin{cases} \frac{g_i}{T} & \text{if } |g_i| \leqslant T \\ 1 & \text{if } g_i > T \\ -1 & \text{if } g_i < T \end{cases}$$

Note that $\overline{\boldsymbol{x}}$ has entries in $[-1, 1]$ instead of $\{-1, 1\}$, but applying a simple independent rounding in every coordinate preserves the value of any multilinear polynomial in expectation. The idea of the analysis is that when $T \sim \sqrt{\log n}$, all contribution comes from the untruncated variables, and the degree-2 moments in untruncated variables are preserved up to a $1/T^2$ scaling factor.

Grothendieck rounding. While the $O(\log n)$-loss in Theorem 8.7 is believed to be tight, there is one important special case in which it is possible to get a constant-factor approximation: for *decoupled* polynomials.

Definition 8.8 (Decoupled polynomial)**.** A homogeneous polynomial $p(\boldsymbol{x})$ of degree d in n variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ is *decoupled* if there exists a partition $S_1, \ldots, S_d$ of $[n]$ such that

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{i} \in S_1 \times \ldots \times S_d} c_{\boldsymbol{i}} \prod_{j=1}^{d} x_{i_j} \,.$$

for some coefficients $\{c_{\boldsymbol{i}}\}_{\boldsymbol{i} \in S_1 \times \ldots \times S_d}$.

For example, decoupled polynomials of degree-2 have the bipartite form

$$p(x_1, \dots, x_n, y_1, \dots, y_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j} x_i y_j \,. \tag{8.2}$$

For maximizing decoupled quadratic polynomials over the hypercube, a rounding based on the celebrated *Grothendieck inequality* achieves constant-factor approximation.

Theorem 8.9 (Grothendieck rounding [AN06]). *There is a rounding algorithm that, given any degree-2 pseudo-distribution μ over $\boldsymbol{x}, \boldsymbol{y} \in \{\pm 1\}^n$ and $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, outputs $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}} \in \{\pm 1\}^n$ such that*

$$\langle \overline{\boldsymbol{x}}, \boldsymbol{M} \overline{\boldsymbol{y}} \rangle \geqslant \frac{1}{K_{\mathrm{G}}} \cdot \widetilde{\mathbb{E}}_{\mu} \langle \boldsymbol{x}, \boldsymbol{M} \boldsymbol{y} \rangle \,,$$

where $K_{\mathrm{G}} < 1.783$ is the Grothendieck constant.

8.1.3. Decoupling

The reason why the decoupling structure helps for polynomial optimization is not too mysterious. Suppose we want to maximize a polynomial of the form (8.2) over $\{-1, 1\}^n$. A natural strategy would be to separately optimize over the $\boldsymbol{x}$- and $\boldsymbol{y}$-variables. In fact, given an assignment of the $\boldsymbol{x}$ variables, the optimization problem over $\boldsymbol{y}$ can be solved in closed form, namely

$$y_j^* = \operatorname{sign}\left(\sum_{i=1}^{n} c_{i,j} x_i \right)$$

achieving value

$$p(x, y^*) = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} c_{i,j} x_i \right| \,.$$

This kind of idea generalizes to higher-degree decoupled polynomials as well.

Following this intuition, a standard technique for polynomial optimization problems is *decoupling*, which relates the optimum of $\langle \boldsymbol{T}, \boldsymbol{x}^{\otimes 3} \rangle$ (the "coupled" polynomial) to the optimum of $\langle \boldsymbol{T}, \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} \rangle$ (the "decoupled" polynomial). Something quite magical happens for maximizing *degree-3* polynomials over the n-dimensional hypercube or the unit sphere: the coupled and decoupled maximization problems are equivalent up to a constant factor.

Lemma 8.10 (Decoupling [KN08, HLZ10]). *Let Ω be either $\{-1, 1\}^n$ or $\mathbb{S}^{n-1}$. Let f be a multilinear homogeneous degree-3 polynomial in n variables,*

$$f(\boldsymbol{x}) = \sum_{i,j,k=1}^{n} T_{ijk} x_i x_j x_k$$

(where $\boldsymbol{T}$ is a symmetric 3-tensor). Consider also the decoupled version of f:

$$\widetilde{f}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \sum_{i,j,k=1}^{n} T_{ijk}x_iy_jz_k\,.$$

Then,

$$\max_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \geqslant \frac{2}{9}\cdot \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\Omega} \widetilde{f}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\,.$$

In §8.6.2, we prove a generalization of Lemma 8.10 to all homogeneous polynomials of *odd-degree*.

We emphasize that Lemma 8.10 fails dramatically for homogeneous polynomials of even degree. The canonical counter-example is the Max-Cut polynomial of the complete graph (Example 6.15):

$$p(\boldsymbol{x}) = -\sum_{i,j=1}^{n} x_ix_j\,, \quad \widetilde{p}(\boldsymbol{x},\boldsymbol{y}) = -\sum_{i,j=1}^{n} x_iy_j\,.$$

Clearly, $\max_{\boldsymbol{x},\boldsymbol{y}\in\{-1,1\}^n}\widetilde{p}(\boldsymbol{x},\boldsymbol{y}) = n^2$, but $\max_{\boldsymbol{x}\in\{-1,1\}^n} p(\boldsymbol{x}) = O(n)$. This explains why there is a gap between Theorem 8.7 and Theorem 8.9.

8.1.4. Anti-concentration

We will need the following anti-concentration result that can be deduced from standard hypercontractivity (see e.g. [AGK04, Lemma 3.2]).

Lemma 8.11. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^n$ satisfying the following: there exists a constant $B > 0$ such that $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[p(\boldsymbol{x})^4] \leqslant B^d \cdot \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[p(\boldsymbol{x})^2\right]^2$ for every degree-d polynomial p. Then, for any degree-d polynomial p,*

$$\Pr_{\boldsymbol{x}\sim\mathcal{D}}\left[p(\boldsymbol{x}) > \mathop{\mathbb{E}}_{\mathcal{D}} p\right] \geqslant 2^{-\frac{4}{3}}B^{-d}\,.$$

Relevant special cases for what follows are when $\mathcal{D}$ is the uniform distribution over $\{\pm1\}^n$ or the standard Gaussian distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)$, which both satisfy the assumption with $B = 9$ (see e.g. for reference [O'D14, Theorem 9.21] and [Bog98, Theorem 1.6.2]).

8.2. A simple $O(\sqrt{n})$-certifiable upper bound

We start by describing a simple argument proving that the canonical SoS relaxation of maximizing homogeneous polynomials of degree-3 over the hypercube has integrality gap $O(\sqrt{n})$. This gives the first certifiable upper bound for this maximization problem.

Let $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \sum_{1\leqslant i,j,k\leqslant n} T_{ijk}x_iy_jz_k$ where $(T_{i,j,k})_{1\leqslant i,j,k\leqslant n}$ is a symmetric 3-tensor. We want to approximate

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\{\pm1\}^n} f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\,. \tag{8.3}$$

Let $\mathcal{D}$ be a pairwise independent distribution over $\{\pm1\}^n$. We can use a construction in which $|\mathrm{supp}(\mathcal{D})| = O(n)$. We will assume without loss of generality that if $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$ then also $-\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$.

We consider the following approximation algorithm: for each $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$, find a constant factor approximation of

$$\max_{\boldsymbol{y},\boldsymbol{z}\in\{\pm1\}^n} f(\widehat{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{z}) = \max_{\boldsymbol{y},\boldsymbol{z}\in\{\pm1\}^n} \sum_{1\leqslant i,j,k\leqslant n} T_{i,j,k}\widehat{x}_iy_jz_k\,, \tag{8.4}$$

using Theorem 8.9. Output the best solution over all choices of $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$.

Call r the maximum of the Grothendieck relaxation of (8.4) over all $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$. The algorithm outputs a solution of value $\Omega(r)$. We want to prove that the standard degree-4 SoS relaxation of (8.3) has optimum at most $r\cdot\sqrt{n}$, which establishes an $O(\sqrt{n})$ integrality gap.

Let μ be the optimal pseudo-distribution for the degree-4 SoS relaxation of (8.3) and SOS be its value. Let $q_i(\boldsymbol{y},\boldsymbol{z}) = \boldsymbol{y}^\top \boldsymbol{T}_i\boldsymbol{z} = \sum_{1\leqslant j,k\leqslant n} T_{i,j,k}y_jz_k$, and write $\boldsymbol{q} = (q_1,\ldots,q_n)$ such that $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \langle \boldsymbol{x},\boldsymbol{q}\rangle$. We have that

$$\mathrm{SOS}^2 = \left(\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x},\boldsymbol{q}\rangle\right)^2 \leqslant \widetilde{\mathbb{E}}_\mu[\langle \boldsymbol{x},\boldsymbol{q}\rangle^2] \leqslant n\cdot \widetilde{\mathbb{E}}_\mu \|\boldsymbol{q}\|_2^2$$

by Cauchy-Schwarz (Fact 8.3).

On the other hand, by definition of r and our assumption that $\mathrm{supp}(\mathcal{D})$ is symmetric, for every $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$ there is a degree-2 SoS proof (over variables $\boldsymbol{y}$, $\boldsymbol{z}$) that

$$f(\widehat{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{z}) \leqslant r \text{ and } f(\widehat{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{z}) \geqslant -r\,.$$

Hence, for every $\widehat{\boldsymbol{x}} \in \mathrm{supp}(\mathcal{D})$ there is a degree-4 SoS proof that

$$f(\widehat{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{z})^2 = \langle \widehat{\boldsymbol{x}},\boldsymbol{q}\rangle^2 \leqslant r^2\,,$$

which means that

$$r^2 \geqslant \widetilde{\mathbb{E}}_\mu \mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}},\boldsymbol{q}\rangle^2 = \widetilde{\mathbb{E}}_\mu \|\boldsymbol{q}\|_2^2 \geqslant \frac{1}{n}\cdot \mathrm{SOS}^2\,,$$

where we use the fact that the pairwise independence of $\mathcal{D}$ implies that

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}},\boldsymbol{v}\rangle^2 = \|\boldsymbol{v}\|_2^2$$

for all $\boldsymbol{v} \in \mathbb{R}^n$. Putting things together we get $\mathrm{SOS}^2 \leqslant n\cdot r^2$, which completes the argument.

8.3. An $O(\sqrt{n})$-factor approximation with rounding

We now give a simple polynomial-time certification and rounding algorithm via constant-degree SoS that achieves approximation $O(\sqrt{n})$ for cubic optimization over the hypercube. Our key technical ingredient is a new use of polynomial reweightings of pseudo-distributions (see Lemma 8.13).

Theorem 8.12. *For any decoupled homogeneous degree-3 polynomial*

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \sum_{i,j,k=1}^{n} T_{ijk} x_i y_j z_k \,,$$

the degree-6 SoS relaxation of $\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\{\pm1\}^n} f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ *has integrality gap at most* $O(\sqrt{n})$.

Furthermore, there is a polynomial-time rounding algorithm that, given a degree-6 pseudo-distribution μ *with* $\mathrm{SOS} := \widetilde{\mathbb{E}}_\mu f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) > 0$, *outputs a solution* $\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}} \in \{\pm1\}^n$ *with value* $f(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}}) \geqslant \Omega\left(\frac{\mathrm{SOS}}{\sqrt{n}}\right)$.

We describe another algorithm outputting a certificate with similar guarantees in §8.2. In comparison, the proof in this section will come together with a rounding and will allow us to build up towards a more general tradeoff between time and approximation in §8.4.

Recall from §6.4.3 that the strategy of [KN08] is to first sample $\overline{\boldsymbol{x}} \sim \{\pm1\}^n$ and then solve for $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{z}}$ using Grothendieck rounding. One might expect that a similar strategy works to round an optimal SoS solution. However, given an optimal pseudo-distribution μ, it is not clear how $\widetilde{\mathbb{E}}_\mu \sum_{i=1}^n \overline{x}_i(\boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z})$ relates to the SoS value $\widetilde{\mathbb{E}}_\mu \sum_{i=1}^n x_i(\boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z})$. In fact, the former can be much smaller than the latter or even zero.

Denote $q_i(\boldsymbol{y},\boldsymbol{z}) := \boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z}$ for convenience. Our key idea is that even though $\widetilde{\mathbb{E}}_\mu \langle \overline{\boldsymbol{x}}, \boldsymbol{q}\rangle$ may be small, we can reweight the pseudo-distribution μ and get another pseudo-distribution μ' such that $\widetilde{\mathbb{E}}_{\mu'} \langle \overline{\boldsymbol{x}}, \boldsymbol{q}\rangle \gtrsim (\widetilde{\mathbb{E}}_\mu \langle \overline{\boldsymbol{x}}, \boldsymbol{q}\rangle^2)^{1/2}$. Furthermore, the quantity on the right-hand side can be related to the SoS value (for a typical $\overline{\boldsymbol{x}}$). One may view this procedure as raising the (pseudo-) expectation of a random variable to be close to its (pseudo-) standard deviation, which is reminiscent of the scalar fixing lemma of [BKS17].

We capture this idea in the following lemma:

Lemma 8.13. *Let* $p(x_1,\dots,x_n)$ *be a degree-*t *polynomial and let* μ *be a degree-*$3t$ *pseudo-distribution over* $(x_1,\dots,x_n)$. *There is reweighting of* μ *by a degree-*$2t$ *polynomial such that the resulting degree-*t *pseudo-distribution* μ' *satisfies*

$$\left|\widetilde{\mathbb{E}}_{\mu'}[p]\right| \geqslant \frac{1}{3} \cdot \sqrt{\widetilde{\mathbb{E}}_{\mu}[p^2]} \,.$$

Proof. Let $m := \sqrt{\widetilde{\mathbb{E}}_\mu [p^2]} > 0$. We can assume that $\left|\widetilde{\mathbb{E}}_\mu[p]\right| < \frac{m}{3}$, otherwise we are done without any reweighting.

First, suppose that $\left|\widetilde{\mathbb{E}}_\mu\left[p^3\right]\right| \geqslant \frac{m^3}{3}$. Reweight μ by the degree-$2t$ SoS polynomial p^2 and let μ' be the resulting pseudo-distribution. Then we have $\left|\widetilde{\mathbb{E}}_{\mu'}[p]\right| = \left|\frac{\widetilde{\mathbb{E}}_\mu[p^3]}{\widetilde{\mathbb{E}}_\mu[p^2]}\right| \geqslant \frac{1}{3}\cdot\sqrt{\widetilde{\mathbb{E}}_\mu\left[p^2\right]}$.

Now suppose that $\left|\widetilde{\mathbb{E}}_\mu\left[p^3\right]\right| < \frac{m^3}{3}$. Reweight μ by the degree-$2t$ SoS polynomial $(p+m)^2$ and let μ' be the resulting pseudo-distribution. Note that:

$$\widetilde{\mathop{\mathbb{E}}_\mu}\left[(p+m)^2\right] = 2m^2 + 2m\cdot\widetilde{\mathop{\mathbb{E}}_\mu}\left[p\right] \in \left[\frac{4m^2}{3}, \frac{8m^2}{3}\right],$$

where we used $\left|\widetilde{\mathbb{E}}_\mu[p]\right| < \frac{m}{3}$. In particular, we are reweighting by a polynomial with non-zero pseudo-expectation, so this is a well-defined operation. Moreover,

$$\widetilde{\mathop{\mathbb{E}}_\mu}\left[(p+m)^2 p\right] \geqslant 2m\cdot\widetilde{\mathop{\mathbb{E}}_\mu}\left[p^2\right] - \left|\widetilde{\mathop{\mathbb{E}}_\mu}\left[p^3\right]\right| - m^2\cdot\left|\widetilde{\mathop{\mathbb{E}}_\mu}[p]\right| \geqslant \frac{4m^3}{3}.$$

Putting everything together, we obtain $\left|\widetilde{\mathbb{E}}_{\mu'}[p]\right| \geqslant \frac{1}{2}\cdot\sqrt{\widetilde{\mathbb{E}}_\mu\left[p^2\right]}$.

Thus, we get the desired result in both cases. □

We are now ready to prove Theorem 8.12.

Proof of Theorem 8.12. Let $q_i(\boldsymbol{y},\boldsymbol{z}) \coloneqq \boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z}$ for each $i \in [n]$. For simplicity of notation, we will drop the dependence on $\boldsymbol{y},\boldsymbol{z}$ and denote $\boldsymbol{q} = (q_1,\dots,q_n)$. Then, we have

$$\mathrm{SOS} = \sum_{i=1}^n \widetilde{\mathop{\mathbb{E}}_\mu}[x_i q_i] \leqslant \sum_{i=1}^n \sqrt{\widetilde{\mathop{\mathbb{E}}_\mu}[q_i^2]} \leqslant \sqrt{n\cdot\sum_{i=1}^n \widetilde{\mathop{\mathbb{E}}_\mu}[q_i^2]} = \sqrt{n\cdot\widetilde{\mathop{\mathbb{E}}_\mu}\|\boldsymbol{q}\|_2^2},$$

by Cauchy-Schwarz and its pseudo-expectation version (Fact 8.3). Next, since

$$\mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm1\}^n}\langle \boldsymbol{v},\boldsymbol{h}\rangle^2 = \|\boldsymbol{v}\|_2^2$$

is a polynomial identity,

$$\mathrm{SOS}^2 \leqslant n\cdot\mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm1\}^n}\widetilde{\mathop{\mathbb{E}}_\mu}\langle \boldsymbol{q},\boldsymbol{h}\rangle^2, \tag{8.5}$$

where we recall that $\boldsymbol{q} = (q_1,\dots,q_n)$ are degree-2 polynomials in $\boldsymbol{y},\boldsymbol{z}$. We now describe the rounding algorithm.

1. Sample $\boldsymbol{h} \sim \{\pm1\}^n$, and set $\overline{\boldsymbol{x}} \coloneqq \boldsymbol{h}$.
2. Reweight the pseudo-distribution μ via Lemma 8.13 to obtain a degree-2 pseudo-distribution μ' such that $\left|\widetilde{\mathbb{E}}_{\mu'}\langle \boldsymbol{q},\boldsymbol{h}\rangle\right| \geqslant \frac{1}{3}\sqrt{\widetilde{\mathbb{E}}_\mu\langle \boldsymbol{q},\boldsymbol{h}\rangle^2}$.

3. Use Grothendieck rounding (Theorem 8.9) on μ' to obtain solutions $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ satisfying $\overline{\boldsymbol{y}}^\top \left(\sum_{i=1}^n h_i \boldsymbol{T}_i\right) \overline{\boldsymbol{z}} \geqslant \frac{1}{K_G} \cdot \left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right|$ (we can get the guarantees with the absolute value by flipping the sign of h).

First, note that $\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^2$ is a degree-2 polynomial in $\boldsymbol{h}$, so by Paley-Zygmund inequality,

$$\Pr_{\boldsymbol{h}\sim\{\pm 1\}^n}\left[\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^2 \geqslant \frac{\mathrm{SOS}^2}{n}\right] \geqslant \Omega(1)\,,$$

meaning that we get a "good" $\boldsymbol{h}$ with constant probability. For a good h, it also holds that

$$f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \frac{1}{K_G}\left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right| \geqslant \frac{1}{3K_G}\sqrt{\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^2} \geqslant \Omega\left(\frac{\mathrm{SOS}}{\sqrt{n}}\right).$$

Thus, repeating the above $\mathrm{poly}(n)$ times, we can obtain a solution with value $\Omega\left(\frac{\mathrm{SOS}}{\sqrt{n}}\right)$ with high probability. This completes the proof. □

8.4. Going beyond $O(\sqrt{n})$-approximation via higher-degree SoS

We now switch to a general time/approximation tradeoff for the problem by rounding higher levels of the SoS hierarchy.

Theorem 8.14. *Let k, n be integers such that $2 \leqslant k \leqslant n$. For any decoupled homogeneous degree-3 polynomial $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{i,j,k=1}^n T_{ijk} x_i y_j z_k$, the canonical degree-$(6k)$ SoS relaxation of $\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\{\pm 1\}^n} f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ has integrality gap at most $O\left(\sqrt{\frac{n}{k}}\right)$.*

Furthermore, there is an $n^{O(k)}$-time rounding algorithm that, given a degree-$(6k)$ pseudo-distribution with $\mathrm{SOS} := \widetilde{\mathbb{E}}_\mu f > 0$, outputs a solution $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ with value $f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \Omega\left(\sqrt{\frac{k}{n}}\right) \cdot \mathrm{SOS}$.

Recall that the $\sqrt{n}$ approximation factor in the previous section was coming from relating the SoS value $\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x}, \boldsymbol{q}\rangle$ to a quantity of the form $\mathbb{E}_{\boldsymbol{h}\sim\{\pm 1\}^n} \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{h}, \boldsymbol{q}\rangle^2$. To make use of higher levels of the SoS hierarchy, we will now connect the SoS value to higher moments of the form $\mathbb{E}_{\boldsymbol{h}\sim\{\pm 1\}^n} \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{h}, \boldsymbol{q}\rangle^{2k}$. The proof of Theorem 8.14 will then follow from a high-degree version of the polynomial reweighting from Lemma 8.13.

One can interpret the inequality from the previous section

$$\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x}, \boldsymbol{q}\rangle^2 \leqslant n \cdot \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm 1\}^n} \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{h}, \boldsymbol{q}\rangle^2$$

as the SoS analog of the inequality $\|\boldsymbol{q}\|_1^2 \leqslant n \cdot \|\boldsymbol{q}\|_2^2 = n \cdot \mathbb{E}_{\boldsymbol{h}\sim\{\pm1\}^n} \langle \boldsymbol{h}, \boldsymbol{q}\rangle^2$ that holds for any $\boldsymbol{q} \in \mathbb{R}^n$ by Cauchy-Schwarz and an explicit variance equality. Our higher-level proof also has a classical analog, namely:

$$\|\boldsymbol{q}\|_1 \leqslant O(1) \cdot \sqrt{\frac{n}{k}} \cdot \left(\mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm1\}^n} \langle \boldsymbol{h}, \boldsymbol{q}\rangle^{2k} \right)^{\frac{1}{2k}} . \tag{8.6}$$

Such an inequality holds for any $\boldsymbol{q} \in \mathbb{R}^n$ [Mon90]. To see that, decompose $\boldsymbol{q}$ and $\boldsymbol{h}$ into k (arbitrary) blocks $\boldsymbol{q}^{(1)}, \dots, \boldsymbol{q}^{(k)}$ and $\boldsymbol{h}^{(1)}, \dots, \boldsymbol{h}^{(k)}$ of size roughly $\frac{n}{k}$. By Paley-Zygmund inequality, we get that $\left|\left\langle \boldsymbol{q}^{(i)}, \boldsymbol{h}^{(i)} \right\rangle\right| \geqslant \Omega(1) \cdot \left\|\boldsymbol{q}^{(i)}\right\|_2 \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \left\|\boldsymbol{q}^{(i)}\right\|_1$ holds with at least constant probability for any fixed $i \in [k]$. So with probability at least $2^{-O(k)}$ we have $|\langle \boldsymbol{q}, \boldsymbol{h}\rangle| \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \|\boldsymbol{q}\|_1$, which in turn implies (8.6).

Although this proof is streamlined, the part using Paley-Zygmund and independence across the k blocks does not directly translate into a sum-of-squares proof. We now give a different and degree-$O(k)$ sum-of-squares proof of the inequality.

Lemma 8.15. *Let $k < n \in \mathbb{N}$, and let $\boldsymbol{x} = (x_1, \dots, x_n)$ be indeterminates and $\boldsymbol{v} = (v_1, \dots, v_n)$ be such that each v_i is a polynomial of degree $\leqslant t$. Then,*

$$\left\{x_i^2 = 1,\ \forall i \in [n]\right\} \vdash^{\boldsymbol{x},\boldsymbol{v}}_{2(t+1)k} \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm1\}^n} \left[\langle \boldsymbol{h}, \boldsymbol{v}\rangle^{2k}\right] \geqslant \left(\frac{k}{4n}\right)^k \langle \boldsymbol{x}, \boldsymbol{v}\rangle^{2k} .$$

Proof. We divide $[n]$ into k blocks, each of size at most $\lceil \frac{n}{k} \rceil$. For $t \in [k]$, let $\boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)}$ be the vectors $\boldsymbol{x}, \boldsymbol{v}$ restricted to the t-th block. Then,

$$\vdash^{\boldsymbol{x},\boldsymbol{v}}_{2(t+1)k} \langle \boldsymbol{x}, \boldsymbol{v}\rangle^{2k} = \left(\sum_{t=1}^k \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle\right)^{2k} \leqslant 2^k \cdot \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^k} \left(\sum_{t=1}^k \varepsilon_t \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle\right)^{2k} , \tag{8.7}$$

since $\left(\sum_{t=1}^k \varepsilon_t \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle\right)^{2k}$ is a square for each $\boldsymbol{\varepsilon} \in \{\pm1\}^k$. Expanding the above and using the fact that all odd moments of $\boldsymbol{\varepsilon} \sim \{\pm1\}^k$ vanish, we get

$$\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^k} \left(\sum_{t=1}^k \varepsilon_t \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle\right)^{2k} = \sum_{\boldsymbol{\gamma}\in\mathbb{N}^k : |\boldsymbol{\gamma}|=k} c_{\boldsymbol{\gamma}} \prod_{t=1}^k \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^{2\gamma_t} , \tag{8.8}$$

where $c_{\boldsymbol{\gamma}} := \frac{(2k)!}{\prod_{t=1}^k (2\gamma_t)!}$. Here $|\boldsymbol{\gamma}| = \sum_{t=1}^k \gamma_t$ and $\boldsymbol{\gamma}$ represents a multiset of $[k]$ of size $|\boldsymbol{\gamma}|$. Next, by SoS Cauchy-Schwarz (Fact 8.2), we have that

$$\left\{x_i^2 = 1,\ \forall i \in [n]\right\} \vdash^{\boldsymbol{x},\boldsymbol{v}}_{2(t+1)} \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^2 \leqslant \|\boldsymbol{x}^{(t)}\|_2^2 \cdot \|\boldsymbol{v}^{(t)}\|_2^2 \leqslant \left(\frac{2n}{k}\right) \cdot \|\boldsymbol{v}^{(t)}\|_2^2 ,$$

since $\boldsymbol{x}^{(t)}$ has dimension at most $\lceil \frac{n}{k} \rceil \leqslant \frac{2n}{k}$. Next, using the identity

$$\|\boldsymbol{v}^{(t)}\|_2^2 = \mathop{\mathbb{E}}_{\boldsymbol{h}^{(t)}} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^2$$

where $\boldsymbol{h}^{(t)} \sim \{\pm 1\}^{\dim(\boldsymbol{x}^{(t)})}$,

$$\begin{aligned}
\left\{x_i^2 = 1,\ \forall i \in [n]\right\} \underset{2(t+1)k}{\vdash_{\boldsymbol{x},\boldsymbol{v}}} \prod_{t=1}^{k} \left\langle \boldsymbol{x}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^{2\gamma_t} &\leqslant \prod_{t=1}^{k} \left(\frac{2n}{k}\right)^{\gamma_t} \|\boldsymbol{v}^{(t)}\|_2^{2\gamma_t} \\
&= \left(\frac{2n}{k}\right)^{k} \prod_{t=1}^{k} \left(\mathop{\mathbb{E}}_{\boldsymbol{h}^{(t)}} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^2 \right)^{\gamma_t} \\
&\leqslant \left(\frac{2n}{k}\right)^{k} \prod_{t=1}^{k} \mathop{\mathbb{E}}_{\boldsymbol{h}^{(t)}} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^{2\gamma_t},
\end{aligned}$$

where the last inequality uses Fact 8.2. Combining the above with (8.7) and (8.8), we have

$$\left\{x_i^2 = 1,\ \forall i \in [n]\right\} \underset{2(t+1)k}{\vdash_{\boldsymbol{x},\boldsymbol{v}}} \langle \boldsymbol{x}, \boldsymbol{v} \rangle^{2k} \leqslant \left(\frac{4n}{k}\right)^{k} \cdot \mathop{\mathbb{E}}_{\boldsymbol{h} \sim \{\pm 1\}^n} \sum_{\boldsymbol{\gamma} \in \mathbb{N}^k : |\boldsymbol{\gamma}| = k} c_{\boldsymbol{\gamma}} \prod_{t=1}^{k} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^{2\gamma_t}.$$

Finally, since $\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(k)}$ are uniformly random Boolean vectors, multiplying $\boldsymbol{h}^{(t)}$ by $\varepsilon_t \sim \{\pm 1\}$ does not change the distribution. Thus, applying (8.8) again, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{h}} \sum_{\boldsymbol{\gamma} \in \mathbb{N}^k : |\boldsymbol{\gamma}| = k} c_{\boldsymbol{\gamma}} \prod_{t=1}^{k} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle^{2\gamma_t} &= \mathop{\mathbb{E}}_{\boldsymbol{h}} \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}} \left(\sum_{t=1}^{k} \varepsilon_t \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle \right)^{2k} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{h}} \left(\sum_{t=1}^{k} \left\langle \boldsymbol{h}^{(t)}, \boldsymbol{v}^{(t)} \right\rangle \right)^{2k} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{h}} \langle \boldsymbol{h}, \boldsymbol{v} \rangle^{2k}.
\end{aligned}$$

This completes the proof. □

Our second key ingredient is the analog of Lemma 8.13 for high moments.

Lemma 8.16. *Let $k \in \mathbb{N}$. Let p be a degree-t polynomial in variables $\boldsymbol{x} \in \mathbb{R}^n$, and let μ be a degree-$(2k+2)t$ pseudo-distribution. There is a degree-$2kt$ reweighting of μ such that the resulting pseudo-distribution μ' satisfies*

$$\left| \widetilde{\mathop{\mathbb{E}}}_{\mu'} p \right| \geqslant \frac{1}{3} \cdot \left(\widetilde{\mathop{\mathbb{E}}}_{\mu} \left[p^{2k} \right] \right)^{\frac{1}{2k}}.$$

Proof. Let $m := \left(\widetilde{\mathbb{E}}_\mu \left[p^{2k}\right]\right)^{\frac{1}{2k}} > 0$.

First, consider reweighting μ by the degree-$2kt$ sum-of-squares polynomial p^{2k} and denote by μ_1 the resulting pseudo-distribution. Then, we have $\left|\widetilde{\mathbb{E}}_{\mu_1}[p]\right| = \frac{\left|\widetilde{\mathbb{E}}_\mu\left[p^{2k+1}\right]\right|}{m^{2k}}$. We are done if this is larger than $\frac{m}{3}$, hence it remains to handle the case $\left|\widetilde{\mathbb{E}}_\mu\left[p^{2k+1}\right]\right| \leqslant \frac{m^{2k+1}}{3}$.

Now, reweight μ by p^{2k-2} and denote by μ_2 the resulting pseudo-distribution. Note that by the pseudo-distribution version of Cauchy-Schwarz (Fact 8.3), as long as μ is a degree-$(2k+2)t$ pseudo-distribution,

$$0 < \widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k}\right]^2 \leqslant \widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k-2}\right] \cdot \widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k+2}\right],$$

so that $\widetilde{\mathbb{E}}_\mu\left[p^{2k-2}\right] > 0$ and the reweighting is well-defined. Furthermore, we have $\left|\widetilde{\mathbb{E}}_{\mu_2}[p]\right| = \frac{\left|\widetilde{\mathbb{E}}_\mu\left[p^{2k-1}\right]\right|}{\widetilde{\mathbb{E}}_\mu\left[p^{2k-2}\right]}$. Once again, we are done if this is larger than $\frac{m}{3}$, so we assume from now on that $\left|\widetilde{\mathbb{E}}_\mu\left[p^{2k-1}\right]\right| \leqslant \frac{m}{3} \cdot \widetilde{\mathbb{E}}_\mu\left[p^{2k-2}\right]$.

Finally, we consider the reweighting of μ by the SoS polynomial $(p+m)^2 p^{2k-2}$ and call μ_3 the resulting pseudo-distribution. We have:

$$\widetilde{\mathop{\mathbb{E}}_\mu}\left[(p+m)^2 p^{2k-2}\right] = m^{2k} + 2m \cdot \widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k-1}\right] + m^2 \cdot \widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k-2}\right] \in \left(0, \frac{8m^{2k}}{3}\right],$$

where we also use $\widetilde{\mathbb{E}}_\mu\left[p^{2k-2}\right] \leqslant m^{2k-2}$ (which follows from Fact 8.4). In particular, the reweighting for μ_3 is well-defined. Similarly, we have

$$\widetilde{\mathop{\mathbb{E}}_\mu}\left[(p+m)^2 p^{2k-1}\right] \geqslant 2m^{2k+1} - \left|\widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k+1}\right]\right| - m^2 \cdot \left|\widetilde{\mathop{\mathbb{E}}_\mu}\left[p^{2k-1}\right]\right| \geqslant \frac{4m^{2k+1}}{3}.$$

Thus, $\widetilde{\mathbb{E}}_{\mu_3}[p] \geqslant \frac{m}{2}$ holds in this case, which concludes the proof. □

We are now ready to prove Theorem 8.14.

Proof of Theorem 8.14. Similarly to the proof of Theorem 8.12, we start by defining $q_i = q_i(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z}$. By Fact 8.4 and Lemma 8.15,

$$\mathrm{SOS}^{2k} = \left(\widetilde{\mathop{\mathbb{E}}_\mu}\langle \boldsymbol{x}, \boldsymbol{q}\rangle\right)^{2k} \leqslant \widetilde{\mathop{\mathbb{E}}_\mu}\langle \boldsymbol{x}, \boldsymbol{q}\rangle^{2k} \leqslant O\left(\frac{n}{k}\right)^k \widetilde{\mathop{\mathbb{E}}_\mu} \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\{\pm 1\}^n} \langle \boldsymbol{h}, \boldsymbol{q}\rangle^{2k}.$$

Here we require μ to be a degree-$6k$ pseudo-distribution.

Since $\boldsymbol{h} \mapsto \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}$ is a degree-$2k$ polynomial, by anti-concentration of low-degree polynomials (Lemma 8.11), we can sample $\boldsymbol{h} \in \{\pm 1\}^n$ such that

$$\left(\widetilde{\mathop{\mathbb{E}}_\mu}\langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}\right)^{1/2k} \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \cdot \mathrm{SOS}$$

with probability at least $2^{-O(k)}$.

The rounding algorithm is as follows,

1. Sample $\boldsymbol{h} \sim \{\pm 1\}^n$ and set $\overline{\boldsymbol{x}} = \boldsymbol{h}$.
2. Reweight the pseudo-distribution via Lemma 8.16 such that
$$\left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right| \geqslant \frac{1}{3}\left(\widetilde{\mathbb{E}}_{\mu} \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}\right)^{1/2k}.$$
The SoS degree required for the reweighting is $2(2k+2) \leqslant 6k$.
3. Use Grothendieck rounding (Theorem 8.9) on μ' to obtain solutions $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ for the quadratic polynomial $\boldsymbol{y}^\top \left(\sum_{i=1}^n h_i \boldsymbol{T}_i\right) \boldsymbol{z}$.

The Grothendieck rounding gives us solutions $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ with value
$$\Omega(1) \cdot \left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right|.$$
Thus, with probability at least $2^{-O(k)}$, we get assignments $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ such that
$$f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \Omega\left(\sqrt{\frac{k}{n}}\right) \cdot \mathrm{SOS}.$$
This completes the proof. □

8.5. Polynomial-size SDPs via compressed SoS relaxations

This section is dedicated to the proof of the following theorem.

Theorem 8.17. *Let k, n be integers such that $1 \leqslant k \leqslant n$. There is a $2^{O(k)} n^{O(1)}$-time certification algorithm that, given a decoupled homogeneous degree-3 polynomial $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{1 \leqslant i,j,k \leqslant n} T_{ijk} x_i y_i z_k$ achieves $O(\sqrt{n/k})$-approximation to $\mathrm{OPT} := \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \{\pm 1\}^n} f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$. Moreover, there is a corresponding rounding algorithm running in $2^{O(k)} n^{O(1)}$ time that outputs a solution $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ with value $f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \Omega\left(\sqrt{\frac{k}{n}}\right) \cdot \mathrm{OPT}$.*

Roughly, we will proceed by "compressing" the SDP relaxations analyzed in the previous sections. We will use some explicit hitting set of size $2^k n^{O(1)}$ and use it to define some constant-degree SoS relaxations with $2^k n^{O(1)}$ variables and one additional axiom.

8.5.1. The blockwise construction of the hitting set

Before explaining how to write down the relaxations, we describe the construction of our hitting set over a small sample space that "fools" high moments in every direction. We will

mimic the anti-concentration proof from §8.4 by decomposing the n-dimensional vectors into k blocks.

Definition 8.18. Let $n, k \in \mathbb{N}$ such that k divides n. Define the distribution $\mathcal{D}$ over $\widehat{\boldsymbol{x}} \in \{\pm 1\}^n$ as follows.

1. Sample $\widehat{\boldsymbol{b}}$ from a 4-wise independent distribution over $\{\pm 1\}^{\frac{n}{k}}$, that is, let $\widehat{\boldsymbol{b}} = f(\boldsymbol{s})$, where $f : \{\pm 1\}^r \to \{\pm 1\}^{\frac{n}{k}}$ is a 4-wise independent pseudorandom generator with seed $\boldsymbol{s} \sim \{\pm 1\}^r$ and $r = O(\log n)$.
2. Sample $\widehat{\boldsymbol{c}} \sim \{\pm 1\}^k$ independently of $\widehat{\boldsymbol{b}}$.
3. Let $\widehat{\boldsymbol{x}} := \widehat{\boldsymbol{c}} \otimes \widehat{\boldsymbol{b}}$. In other words, decompose $\widehat{\boldsymbol{x}}$ into k blocks $\widehat{\boldsymbol{x}}^{(1)}, \dots, \widehat{\boldsymbol{x}}^{(k)}$ of size $\frac{n}{k}$ and set $\widehat{\boldsymbol{x}}^{(i)} := \widehat{\boldsymbol{c}}_i \cdot \widehat{\boldsymbol{b}}$ for all $i \in [k]$.

The following observation can be deduced for example from the classical construction of k-wise independent sets of random variables [Jof74].

Claim 8.19. *The distribution $\mathcal{D}$ can be obtained as the uniform distribution over a sample space of size $2^k n^{O(1)}$. In particular, for any $\boldsymbol{x} \in \operatorname{supp}(\mathcal{D})$, $\Pr_{\widehat{\boldsymbol{x}} \sim \mathcal{D}}[\widehat{\boldsymbol{x}} = \boldsymbol{x}] \geqslant 2^{-k} n^{-O(1)}$.*

We will also need the following result, which is a direct consequence of the Paley-Zygmund inequality and the 4-wise independence of $\widehat{\boldsymbol{b}}$.

Claim 8.20. *For all $\boldsymbol{w} \in \mathbb{R}^{\frac{n}{k}}$, $\Pr_{\widehat{\boldsymbol{b}}}\left[\left|\left\langle \widehat{\boldsymbol{b}}, \boldsymbol{w} \right\rangle\right| \geqslant \frac{1}{2}\|\boldsymbol{w}\|_2\right] \geqslant \Omega(1)$.*

Finally, the following lower bound on the moments of $\widehat{\boldsymbol{x}}$ will be the key ingredient to prove that our relaxation provides a correct certificate to the optimum.

Lemma 8.21 (Large moments in every direction)**.** *For all $\boldsymbol{w} \in \mathbb{R}^n$,*

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}} \sim \mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{w} \rangle^{2k} \geqslant \Omega\left(\frac{k}{n}\right)^k n^{-O(1)} \|\boldsymbol{w}\|_1^{2k}.$$

Proof. We first decompose $\boldsymbol{w}$ into k blocks $\boldsymbol{w}^{(1)}, \dots, \boldsymbol{w}^{(k)}$ of size $\frac{n}{k}$, in such a way that $\langle \widehat{\boldsymbol{x}}, \boldsymbol{w} \rangle = \sum_{i=1}^k \widehat{\boldsymbol{c}}_i \left\langle \widehat{\boldsymbol{b}}, \boldsymbol{w}^{(i)} \right\rangle$. Now for any fixed block $i \in [k]$, we know from Claim 8.20 that with at least constant probability over $\widehat{\boldsymbol{b}}$, it holds that $\left|\left\langle \widehat{\boldsymbol{b}}, \boldsymbol{w}^{(i)} \right\rangle\right| \geqslant \frac{1}{2}\|\boldsymbol{w}^{(i)}\|_2 \geqslant \frac{1}{2}\sqrt{\frac{k}{n}}\|\boldsymbol{w}^{(i)}\|_1$, where the last inequality follows from Cauchy-Schwarz. In turn, by linearity of expectation,

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{b}}}\left[\sum_{i=1}^k \left|\left\langle \widehat{\boldsymbol{b}}, \boldsymbol{w}^{(i)} \right\rangle\right|\right] \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}}\|\boldsymbol{w}\|_1.$$

In particular, there exists some $\boldsymbol{x} \in \operatorname{supp}(\mathcal{D})$ satisfying $|\langle \boldsymbol{x}, \boldsymbol{w} \rangle| \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}}\|\boldsymbol{w}\|_1$. By Claim 8.19, this $\boldsymbol{x}$ must be drawn with probability at least $2^{-k} n^{-O(1)}$ from $\mathcal{D}$. Finally, we

apply Markov's inequality to get

$$\begin{aligned}\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{w}\rangle^{2k} &\geqslant \Omega\left(\frac{k}{n}\right)^k \|\boldsymbol{w}\|_1^{2k} \cdot \Pr_{\widehat{\boldsymbol{x}}\sim\mathcal{D}}\left(|\langle \widehat{\boldsymbol{x}}, \boldsymbol{w}\rangle| \geqslant \Omega(1)\cdot\sqrt{\frac{k}{n}}\|\boldsymbol{w}\|_1\right)\\ &\geqslant \Omega\left(\frac{k}{n}\right)^k n^{-O(1)}\|\boldsymbol{w}\|_1^{2k}.\end{aligned}$$

This concludes the proof. □

8.5.2. Proof of Theorem 8.17

We are now ready to state and analyze the SDP relaxation. The high-level intuition is the following: write $q_i := \sum_{j,k} T_{ijk} y_j z_k$ for all $i \in [n]$, so that our goal is now to maximize $\langle \boldsymbol{x}, \boldsymbol{q}\rangle$, which by symmetry is equivalent to maximizing $\langle \boldsymbol{x}, \boldsymbol{q}\rangle^2$. Instead of maximizing over $\boldsymbol{x} \in \{\pm 1\}^n$ we essentially pick a random $\widehat{\boldsymbol{x}}$ from a distribution $\mathcal{D}$ that has large $2k$-th moments in every direction. Then we replace the objective function $\mathbb{E}_{\mathcal{D}} \max_{\boldsymbol{w}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}\rangle^2$ by the following proxy:

$$\max_{\mu \text{ pseudo-distribution on } \boldsymbol{w}} \frac{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \widetilde{\mathbb{E}}_{\mu} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}\rangle^{2(k+1)}}{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \widetilde{\mathbb{E}}_{\mu} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}\rangle^{2k}}.$$

As k grows, this yields a sequence of increasingly better approximations leveraging higher moments of the variables. Since expanding the $2k$-th powers would require solving an SDP of size $n^{\Omega(k)}$, we introduce auxiliary variables $\{M_{\widehat{\boldsymbol{x}}}\}$ corresponding to $\langle \widehat{\boldsymbol{x}}, \boldsymbol{q}\rangle^k$ in combinatorial solutions.

Proof of Theorem 8.17. Assume without loss of generality that k divides n. Let $\mathcal{D}$ be the pseudorandom distribution from Definition 8.18. Furthermore, we fix a guess $\alpha \geqslant 0$ for the value of the optimum of the cubic optimization problem (the final certification and rounding algorithms will be obtained by binary searching for the best possible value of α).

The relaxation. We solve for *feasibility* the degree-12 SoS program over the following variables:

- variables y_j and z_k for all $j, k \in [n]$. To lighten notations we let $q_i := q_i(\boldsymbol{y}, \boldsymbol{z}) = \sum_{1\leqslant j,k\leqslant n} T_{ijk} y_j z_k$ for all $i \in [n]$ (each q_i is a degree-2 polynomial) and write $\boldsymbol{q} = (q_1, \ldots, q_n)$.
- variables $M_{\boldsymbol{x}}$ for each $\boldsymbol{x} \in \mathrm{supp}(\mathcal{D})$.

and under the following additional polynomial constraints:

$$y_j^2 = 1 \quad \text{for all } j \in [n],$$

$$z_k^2 = 1 \quad \text{for all } k \in [n]\,,$$
$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \left[M_{\widehat{\boldsymbol{x}}}^2 \left(\langle \widehat{\boldsymbol{x}}, \boldsymbol{q} \rangle^2 - \alpha^2 \right) \right] \geqslant 0\,. \tag{8.9}$$

By construction of $\mathcal{D}$, the relaxation has $2^k n^{O(1)}$ variables and constraints.

The rounding algorithm. First, we check that any feasible solution to the SoS program can be rounded into an integral solution $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ achieving value $\Omega(\alpha)$. Suppose that there exists some degree-12 pseudo-distribution μ (over $\boldsymbol{y}$, $\boldsymbol{z}$ and $M_{\boldsymbol{x}}$) satisfying all the constraints. Then by (8.9), there exists $\overline{\boldsymbol{x}} \in \operatorname{supp}(\mathcal{D})$ satisfying

$$\alpha^2 \leqslant \frac{\widetilde{\mathbb{E}}_\mu M_{\overline{\boldsymbol{x}}}^2 \langle \overline{\boldsymbol{x}}, \boldsymbol{q} \rangle^2}{\widetilde{\mathbb{E}}_\mu M_{\overline{\boldsymbol{x}}}^2} = \mathop{\widetilde{\mathbb{E}}}_{\mu'} \langle \overline{\boldsymbol{x}}, \boldsymbol{q} \rangle^2\,,$$

where μ' is the degree-6 pseudo-distribution obtained by reweighting μ by the SoS polynomial $M_{\overline{\boldsymbol{x}}}^2$. We now use Lemma 8.13 to construct from μ' a degree-2 pseudo-distribution μ'' that satisfies

$$\mathop{\widetilde{\mathbb{E}}}_{\mu'} \langle \overline{\boldsymbol{x}}, \boldsymbol{q} \rangle^2 \leqslant 9 \left(\mathop{\widetilde{\mathbb{E}}}_{\mu''} \langle \overline{\boldsymbol{x}}, \boldsymbol{q} \rangle \right)^2\,.$$

Finally we use Grothendieck rounding (Theorem 8.9) on μ'' to find $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$ such that

$$f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \frac{1}{K_G} \mathop{\widetilde{\mathbb{E}}}_{\mu''} f(\overline{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{K_G} \mathop{\widetilde{\mathbb{E}}}_{\mu''} \langle \overline{\boldsymbol{x}}, \boldsymbol{q} \rangle \geqslant \Omega(1) \cdot \alpha\,.$$

Approximation factor. Our final algorithm consists of a binary search to get the largest value of $\alpha \geqslant 0$ that makes the SoS program above feasible. Then, some explicit multiple of α coming from the analysis of our rounding provides a correct upper bound certificate on OPT.

We now check that this achieves approximation $O(\sqrt{n/k})$. Fix any triplet $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^* \in \{\pm 1\}^n$ and let $q_i^* = \sum_{1 \leqslant j,k \leqslant n} T_{ijk} y_j^* z_k^*$ for all $i \in [n]$. Suppose that $(\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^*)$ achieves the optimum of the original problem, so that $\mathrm{OPT} = \langle \boldsymbol{x}^*, \boldsymbol{q}^* \rangle = \|\boldsymbol{q}^*\|_1$. We set $(\boldsymbol{y}, \boldsymbol{z}) = (\boldsymbol{y}^*, \boldsymbol{z}^*)$ and $M_{\boldsymbol{x}} = \langle \boldsymbol{x}, \boldsymbol{q}^* \rangle^k$ for all $\boldsymbol{x} \in \operatorname{supp}(\mathcal{D})$, and we prove that this defines a feasible solution. By Hölder's inequality,

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k+2} \geqslant \left(\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k} \right)^{\frac{2k+2}{2k}}\,,$$

and Lemma 8.21 then yields

$$\frac{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} M_{\widehat{\boldsymbol{x}}}^2 \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^2}{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} M_{\widehat{\boldsymbol{x}}}^2} = \frac{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k+2}}{\mathbb{E}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k}}$$

$$\geqslant \left(\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}} \sim \mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k} \right)^{1/k}$$
$$\geqslant \Omega(1) \cdot \frac{k}{n} \cdot n^{-O\left(\frac{1}{k}\right)} \mathrm{OPT}^2 .$$

Assume without loss of generality that $k = \Omega(\log n)$ (since otherwise, one can always increase k to $\Theta(\log n)$ without affecting the target runtime). Then, as long as $\alpha \leqslant O(1) \cdot \sqrt{\frac{k}{n}} \cdot \mathrm{OPT}$, (8.9) is satisfied. This completes the proof. □

8.6. Extensions

8.6.1. Optimization over the unit sphere

In this section, we prove the approximation results for cubic optimization over the unit sphere matching our results over the hypercube:

- We show that the canonical degree-$6k$ SoS relaxation has integrality gap $O(\sqrt{n/k})$ by giving a corresponding rounding algorithm.
- We then prove that a pruned SDP can achieve approximation $O(\sqrt{n/k})$ in time $2^{O(k)}\mathrm{poly}(n)$.

Analysis of the canonical degree-k SoS relaxation

We prove that the canonical degree-k SoS relaxation for optimizing over the unit sphere has integrality gap at most $O\left(\sqrt{\frac{n}{k}}\right)$. The proof mirrors the hypercube case (Theorem 8.14), although the analysis is much simpler here since we can directly relate the SoS value to the moments of the Gaussian distribution.

Theorem 8.22. *Fix $1 \leqslant k \leqslant n$. Given any decoupled homogeneous degree-3 polynomial $f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{1 \leqslant i,j,k \leqslant n} T_{ijk} x_i y_j z_k$, the canonical degree-$6k$ SoS relaxation of*

$$\max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathcal{S}^{n-1}} f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$$

has integrality gap $O\left(\sqrt{\frac{n}{k}}\right)$.

Furthermore, given any degree-$6k$ pseudo-distribution μ over $\left(\mathcal{S}^{n-1}\right)^3$ such that $\mathrm{SOS} := \widetilde{\mathbb{E}}_\mu f > 0$, there is an $n^{O(k)}$-time randomized rounding algorithm that outputs with high probability $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \mathcal{S}^{n-1}$ such that $f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) \geqslant \Omega\left(\sqrt{\frac{k}{n}}\right) \cdot \mathrm{SOS}$.

Proof. Similarly to the proof of Theorem 8.12, we start by defining $q_i = q_i(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{y}^\top \boldsymbol{T}_i \boldsymbol{z}$ and consider

$$\mathrm{SOS}^{2k} = \left(\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x}, \boldsymbol{q}\rangle\right)^{2k} \leqslant \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{x}, \boldsymbol{q}\rangle^{2k} \leqslant \widetilde{\mathbb{E}}_\mu \left[\|\boldsymbol{x}\|_2^{2k} \cdot \|\boldsymbol{q}\|_2^{2k}\right] = \widetilde{\mathbb{E}}_\mu \|\boldsymbol{q}\|_2^{2k},$$

using Cauchy-Schwarz and the pseudo-expectation version of Hölder's inequality (Fact 8.4).

Then, let $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. We know from standard estimates on the moments of the Gaussian distribution that for any vector $\boldsymbol{v} \in \mathbb{R}^n$, $\mathbb{E}_{\boldsymbol{h}} \langle \boldsymbol{h}, \boldsymbol{v}\rangle^{2k} = c_k \|\boldsymbol{v}\|_2^{2k}$, where $c_k = (2k-1)!! \geqslant (k/2)^k$. Thus, we have

$$\mathrm{SOS}^{2k} \leqslant \widetilde{\mathbb{E}}_\mu \|\boldsymbol{q}\|_2^{2k} \leqslant \left(\frac{2}{k}\right)^k \cdot \mathop{\mathbb{E}}_{\boldsymbol{h}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)} \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}.$$

Since $\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}$ is a degree-$2k$ polynomial in $\boldsymbol{h}$, by Lemma 8.11, $\boldsymbol{h}$ satisfies $\left(\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}\right)^{\frac{1}{2k}} \geqslant \Omega(\sqrt{k}) \cdot \mathrm{SOS}$ with probability at least $2^{-O(k)}$. Moreover, with probability at least $1 - 2^{-\Omega(n)}$ we have $\|\boldsymbol{h}\|_2 \leqslant O(\sqrt{n})$.

Hence, our rounding algorithm goes as follows,

1. Sample $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\overline{\boldsymbol{x}} = \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}$.
2. Reweight the pseudo-distribution μ via Lemma 8.16 to obtain a degree-2 pseudo-distribution μ' such that $\left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right| \geqslant \frac{1}{3}\left(\widetilde{\mathbb{E}}_\mu \langle \boldsymbol{q}, \boldsymbol{h}\rangle^{2k}\right)^{\frac{1}{2k}}$.
3. Use the lossless rounding for quadratic forms over the sphere (Lemma 8.6) on μ' to obtain solutions $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \mathcal{S}^{n-1}$ such that $\overline{\boldsymbol{y}}^\top (\sum_{i=1}^n h_i \boldsymbol{T}_i) \overline{\boldsymbol{z}} \geqslant \left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right|$ (we can flip the sign of $\boldsymbol{h}$ to get the guarantee with the absolute value).

Putting everything together, it holds with probability at least $2^{-O(k)}$ that

$$f(\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}) = \frac{1}{\|\boldsymbol{h}\|_2} \cdot \overline{\boldsymbol{y}}^\top \left(\sum_{i=1}^n h_i \boldsymbol{T}_i\right) \overline{\boldsymbol{z}} \geqslant \frac{1}{\|\boldsymbol{h}\|_2} \left|\widetilde{\mathbb{E}}_{\mu'} \langle \boldsymbol{q}, \boldsymbol{h}\rangle\right| \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \cdot \mathrm{SOS}.$$

Repeating this $\mathrm{poly}(n, 2^k)$ times, we obtain a rounding algorithm that satisfies the desired guarantees with high probability. □

Analysis of compressed SoS relaxations over the unit sphere

In this section, we prove the analogous statement of Theorem 8.17 over the sphere.

Theorem 8.23. *Fix $1 \leqslant k \leqslant n$. There is a $2^{O(k)} n^{O(1)}$-time certification algorithm that given a decoupled homogeneous degree-3 polynomial*

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{1\leqslant i,j,k\leqslant n} T_{ijk} x_i y_i z_k$$

achieves $O(\sqrt{n/k})$-approximation to

$$\mathrm{OPT} := \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathcal{S}^{n-1}} f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\,.$$

Moreover, there is a corresponding rounding algorithm running in $2^{O(k)}n^{O(1)}$ time that outputs a solution $\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}} \in \mathcal{S}^{n-1}$ with value $f(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}}) \geqslant \Omega\left(\sqrt{\frac{k}{n}}\right)\cdot \mathrm{OPT}$.

Our proof relies on a hitting set construction analogous to Lemma 8.21.

Lemma 8.24. *For any $1 \leqslant k \leqslant n$ with $k = \Omega(\log n)$, there exists a distribution $\mathcal{D}$ over $\mathcal{S}^{n-1}$ supported on at most $2^{O(k)}n^{O(1)}$ vectors such that for all $\boldsymbol{w} \in \mathbb{R}^n$,*

$$\left(\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}}\sim\mathcal{D}} \langle\widehat{\boldsymbol{x}},\boldsymbol{w}\rangle^{2k}\right)^{\frac{1}{2k}} \geqslant \Omega(1)\cdot\sqrt{\frac{k}{n}}\|\boldsymbol{w}\|_2\,.$$

Proof. Assume without loss of generality that k divides n. We define the distribution $\mathcal{D}$ over $\widehat{\boldsymbol{x}} \in \mathcal{S}^{n-1}$ as follows.

1. Sample $\widehat{\boldsymbol{b}} \in \{\pm1\}^{\frac{n}{k}}$ from a 4-wise independent distribution. Similarly to Claim 8.19, this can be achieved by taking the uniform distribution over a subset of $\{\pm1\}^{\frac{n}{k}}$ of size $n^{O(1)}$.
2. Sample independently $\widehat{\boldsymbol{c}}$ uniformly over an ε-net of $\mathcal{S}^{k-1}$ of size $O(1/\varepsilon)^k$ for $\varepsilon = \frac{1}{10}$.
3. Output $\widehat{\boldsymbol{x}} = \sqrt{\frac{k}{n}}\cdot\widehat{\boldsymbol{c}}\otimes\widehat{\boldsymbol{b}}$. Note in particular that $\widehat{\boldsymbol{x}}$ is a unit vector.

Decompose $\boldsymbol{w}$ into k blocks $\boldsymbol{w}^{(1)},\dots,\boldsymbol{w}^{(k)}$ of size $\frac{n}{k}$ such that

$$\langle\widehat{\boldsymbol{x}},\boldsymbol{w}\rangle = \sqrt{\frac{k}{n}}\cdot\sum_i \widehat{c}_i\left\langle\widehat{b},\boldsymbol{w}^{(i)}\right\rangle\,.$$

Claim 8.20 states that for each $1 \leqslant i \leqslant \frac{n}{k}$,

$$\Pr_{\widehat{\boldsymbol{b}}}\left[\left|\left\langle\widehat{\boldsymbol{b}},\boldsymbol{w}^{(i)}\right\rangle\right| \geqslant \frac{1}{2}\|\boldsymbol{w}^{(i)}\|_2\right] \geqslant \Omega(1)\,.$$

This implies that

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{b}}}\sum_{i=1}^{k}\left\langle\widehat{\boldsymbol{b}},\boldsymbol{w}^{(i)}\right\rangle^2 \geqslant \Omega(1)\cdot\|\boldsymbol{w}\|_2^2\,.$$

On the other hand, for any $\boldsymbol{b} \in \{\pm1\}^{\frac{n}{k}}$, we can find $\boldsymbol{c} = \boldsymbol{c}(\boldsymbol{b})$ in the ε-net such that

$$\sum_{i=1}^{k} c_i\left\langle\boldsymbol{b},\boldsymbol{w}^{(i)}\right\rangle \geqslant \Omega(1)\cdot\left(\sum_{i=1}^{k}\left\langle\boldsymbol{b},\boldsymbol{w}^{(i)}\right\rangle^2\right)^{\frac{1}{2}}\,.$$

Therefore, there must exist $\boldsymbol{x} \in \operatorname{supp}(\mathcal{D})$ such that $\langle \boldsymbol{x}, \boldsymbol{w} \rangle \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \cdot \|\boldsymbol{w}\|_2$, and this $\boldsymbol{x}$ is drawn with probability at least $2^{-O(k)} n^{-O(1)}$. Finally, by Markov inequality,

$$\begin{aligned}\mathop{\mathbb{E}}_{\widehat{\boldsymbol{x}} \sim \mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{w} \rangle^{2k} &\geqslant \Pr_{\widehat{\boldsymbol{x}} \sim \mathcal{D}} \left(|\langle \widehat{\boldsymbol{x}}, \boldsymbol{w} \rangle| \geqslant \Omega(1) \cdot \sqrt{\frac{k}{n}} \cdot \|\boldsymbol{w}\|_2 \right) \cdot \Omega\left(\frac{k}{n}\right)^k \|\boldsymbol{w}\|_2^{2k} \\ &= 2^{-O(k)} n^{-O(1)} \cdot \Omega\left(\frac{k}{n}\right)^k \|\boldsymbol{w}\|_2^{2k} .\end{aligned}$$

This completes the proof assuming $k = \Omega(\log n)$. □

We are now ready to prove Theorem 8.23.

Proof of Theorem 8.23. The proof is identical to the proof of Theorem 8.17, with the following exceptions:

- The Boolean constraints $y_j^2 = 1$ and $z_k^2 = 1$ for all $j, k \in [n]$ become spherical constraints: $\|\boldsymbol{y}\|_2^2 = 1$ and $\|\boldsymbol{z}\|_2^2 = 1$.
- Instead of Grothendieck rounding, we use the (lossless) rounding for quadratic forms over the sphere from Lemma 8.6.
- If $\boldsymbol{x}^*, \boldsymbol{y}^*, \boldsymbol{z}^*$ achieve the optimum for the cubic maximization problem, then $\mathsf{OPT} = \|\boldsymbol{q}^*\|_2$, where $q_i^* = \sum_{1 \leqslant j,k \leqslant n} T_{ijk} y_j^* z_k^*$ for all $i \in [n]$.
- The last sequence of inequalities in the proof of Theorem 8.17 becomes as follows after using Lemma 8.24 instead of Lemma 8.21:

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{D}} \left[M_{\widehat{\boldsymbol{x}}}^2 \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^2 \right]}{\mathbb{E}_{\mathcal{D}} \left[M_{\widehat{\boldsymbol{x}}}^2 \right]} &= \frac{\mathbb{E}_{\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k+2}}{\mathbb{E}_{\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k}} \\ &\geqslant \left(\mathop{\mathbb{E}}_{\mathcal{D}} \langle \widehat{\boldsymbol{x}}, \boldsymbol{q}^* \rangle^{2k} \right)^{1/k} \\ &\geqslant \Omega(1) \cdot \frac{k}{n} \cdot n^{-O(\frac{1}{k})} \mathsf{OPT}^2 .\end{aligned}$$

We conclude in the same way as in the proof of Theorem 8.17. □

8.6.2. Optimizing higher-degree polynomials

Decoupling inequalities

We first prove a decoupling lemma for all odd-degree polynomials.

Lemma 8.25 (High-degree version of Lemma 8.10)**.** *Let d be an odd integer. Let $f(\boldsymbol{x}) = \langle \boldsymbol{T}, \boldsymbol{x}^{\otimes d} \rangle$ be a multilinear homogeneous degree-d polynomial in n variables (where $\boldsymbol{T}$ is a symmetric d-tensor), and let $\widetilde{f}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}) = \langle \boldsymbol{T}, \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(d)} \rangle$ be the decoupled polynomial of f. Then, given any $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)} \in \{\pm 1\}^n$, there exists a sampleable distribution $\mathcal{D}$ over $\{\pm 1\}^n$ such that*

$$\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{D}} [f(\boldsymbol{y})] = \frac{d!}{d^d} \cdot \widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right).$$

As a consequence,

$$\max_{\boldsymbol{y} \in \{\pm 1\}^n} f(\boldsymbol{y}) \geqslant \frac{d!}{d^d} \cdot \max_{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)} \in \{\pm 1\}^n} \widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right).$$

Proof. The distribution $\mathcal{D}$ can be sampled as follows,

- Let $b_1, \dots, b_{d-1}$ be i.i.d. uniform ± 1 random variables and let $b_d := b_1 \cdots b_{d-1}$. Note that the distribution of $\boldsymbol{b} = (b_1, \dots, b_d)$ is $(d-1)$-wise independent and $b_1 b_2 \cdots b_d = 1$.
- Independently for each $i \in [n]$, sample y_i uniformly from $\{b_j x_i^{(j)}\}_{j \in [d]}$.

Since each y_i is sampled independently conditioned on $\boldsymbol{b}$, we have that for any pairwise distinct indices $i_1, \dots, i_d \in [n]$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{D}} \left[y_{i_1} \cdots y_{i_d}\right] &= \mathop{\mathbb{E}}_{\boldsymbol{b}} \left[\prod_{k=1}^{d} \left(\frac{1}{d} \sum_{j=1}^{d} b_j x_{i_k}^{(j)}\right)\right] \\
&= d^{-d} \sum_{j_1, \dots, j_d \in [d]} \mathop{\mathbb{E}}_{\boldsymbol{b}} \left[b_{j_1} \cdots b_{j_d}\right] \cdot x_{i_1}^{(j_1)} \cdots x_{i_d}^{(j_d)}.
\end{aligned}$$

Since $\boldsymbol{b}$ follows a $(d-1)$-wise independent distribution, $b_1 b_2 \cdots b_d = 1$ and d is odd, $\mathbb{E}_{\boldsymbol{b}}[b_{j_1} \cdots b_{j_d}]$ is nonzero (equals 1) if and only if $j_1, \dots, j_d$ are all distinct, i.e., $\{j_1, \dots, j_d\} = [d]$.

Thus, the summation above reduces to summing over permutations of $[d]$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{D}} \left[y_{i_1} \cdots y_{i_d}\right] &= \frac{d!}{d^d} \cdot \mathop{\mathbb{E}}_{\pi \sim \mathbb{S}_d} \left[x_{i_1}^{(\pi(1))} \cdots x_{i_d}^{(\pi(d))}\right] \\
&= \frac{d!}{d^d} \cdot \mathop{\mathbb{E}}_{\pi \sim \mathbb{S}_d} \left[x_{i_{\pi(1)}}^{(1)} \cdots x_{i_{\pi(d)}}^{(d)}\right],
\end{aligned}$$

where $\pi \sim \mathfrak{S}_d$ denotes a random permutation of $[d]$. Finally, as $\boldsymbol{T}$ is a symmetric tensor, we deduce

$$\mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mathcal{D}} [f(\boldsymbol{y})] = \frac{d!}{d^d} \cdot \widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right).$$

This proves the first statement of the lemma. The second follows immediately. □

For even-degree polynomials, Lemma 8.25 simply cannot hold:

Example 8.26 (Impossibility of decoupling for quadratics)**.** Consider the matrix $\boldsymbol{Q} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^\top$, and define the multilinear quadratic polynomial $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$ and the decoupled polynomial $\widetilde{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{y}$. Then, it is easy to verify that $\max_{\boldsymbol{x}\in\{\pm1\}^n} f(\boldsymbol{x}) = n$ but $\max_{\boldsymbol{x},\boldsymbol{y}\in\{\pm1\}^n} \widetilde{f}(\boldsymbol{x}, \boldsymbol{y}) = n^2 - n$, which is a $\mathrm{poly}(n)$ gap.

On the other hand, if we only consider $\max_{\boldsymbol{y}} |f(\boldsymbol{y})|$ like in the setting of [BGG+17] (as opposed to $\max_{\boldsymbol{y}} f(\boldsymbol{y})$), then decoupling inequalities with the same guarantees as Lemma 8.25 hold for *any* degree:

Lemma 8.27 (Decoupling for absolute values, any degree)**.** *Let $d \in \mathbb{N}$. Let $f(\boldsymbol{x}) = \langle \boldsymbol{T}, \boldsymbol{x}^{\otimes d} \rangle$ be a multilinear homogeneous degree-d polynomial in n variables (where T is a symmetric d-tensor), and let $\widetilde{f}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}) = \langle \boldsymbol{T}, \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(d)} \rangle$ be the decoupled polynomial of f. Then,*

$$\max_{\boldsymbol{y}\in\{\pm1\}^n} |f(\boldsymbol{y})| \geqslant \frac{d!}{d^d} \cdot \max_{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(d)}\in\{\pm1\}^n} \widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right) .$$

Proof. We use the trick that $\max_{\boldsymbol{y}\sim\{\pm1\}^n} |f(\boldsymbol{y})| = \max_{\boldsymbol{y}\in[-1,1]^n} |f(\boldsymbol{y})|$. Thus, given assignments $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)} \in \{\pm1\}^n$, it suffices to round to a $\boldsymbol{y} \in [-1, 1]^n$.

We next state the well-known polarization identity for degree-d homogeneous polynomials:

$$\widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right) = \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \left[\frac{\varepsilon_1 \cdots \varepsilon_d}{d!} f\left(\varepsilon_1 \boldsymbol{x}^{(1)} + \cdots + \varepsilon_d \boldsymbol{x}^{(d)}\right)\right] .$$

Define $\boldsymbol{y}_{\boldsymbol{\varepsilon}} := \frac{1}{d}(\varepsilon_1 \boldsymbol{x}^{(1)} + \cdots + \varepsilon_d \boldsymbol{x}^{(d)}) \in [-1, 1]^n$. Then, rewriting the above and using the triangle inequality,

$$\widetilde{f}\left(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}\right) = \frac{d^d}{d!} \cdot \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \left[\varepsilon_1 \cdots \varepsilon_d \cdot f(\boldsymbol{y}_{\boldsymbol{\varepsilon}})\right] \leqslant \frac{d^d}{d!} \cdot \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}\sim\{\pm1\}^n} \left[|f(\boldsymbol{y}_{\boldsymbol{\varepsilon}})|\right] .$$

Thus, there exists some $\boldsymbol{y}_{\boldsymbol{\varepsilon}} \in [-1, 1]^n$ such that

$$|f(\boldsymbol{y}_{\boldsymbol{\varepsilon}})| \geqslant \frac{d!}{d^d} \cdot \widetilde{f}(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}) . \qquad \square$$

Rounding SoS relaxations for high-degree polynomials

We now give a simple polynomial-time certification and rounding algorithm using the canonical SoS relaxation that achieves approximation $O(n^{\frac{d}{2}-1})$ for optimizing *decoupled* homogeneous degree-d polynomials over the hypercube. The proof is almost identical to the cubic case (Theorem 8.12).

Theorem 8.28. *Let $d \geqslant 3$ and $n \geqslant 1$ be integers. Given any decoupled homogeneous degree-d polynomial $f(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)}) = \langle \boldsymbol{T}, \boldsymbol{x}^{(1)} \otimes \cdots \otimes \boldsymbol{x}^{(d)} \rangle$, the degree-$2d$ SoS relaxation of*

$$\max_{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(d)}\in\{\pm1\}^n} f(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(d)})$$

has integrality gap at most $O(n^{\frac{d}{2}-1})$. *Furthermore, given a degree-2d pseudo-distribution* μ *such that* $\mathrm{SOS} := \widetilde{\mathbb{E}}_\mu f > 0$, *there is a randomized* $n^{O(d)}$*-time rounding algorithm that outputs* $\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(d)} \in \{\pm 1\}^n$ *such that with high probability,*

$$f(\overline{\boldsymbol{x}}^{(1)}, \dots, \overline{\boldsymbol{x}}^{(d)}) \geqslant \Omega(n^{-\frac{d}{2}+1}) \cdot \mathrm{SOS}\,.$$

Proof. For $i_3, \dots, i_d \in [n]$, let $q_{i_3,\dots,i_d}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) := \left\langle \boldsymbol{T}_{i_3,\dots,i_d}, \boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \right\rangle$, a degree-2 polynomial in $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, where $\boldsymbol{T}_{i_3,\dots,i_d}$ is an $n \times n$ matrix corresponding to a slice of the tensor $\boldsymbol{T}$.

For simplicity of notation, we drop the dependence on $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ and write

$$\boldsymbol{Q} := \boldsymbol{Q}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = (q_{i_3,\dots,i_d})_{i_3,\dots,i_d \in [n]}$$

as an order-$(d-2)$ tensor whose entries are degree-2 polynomials in $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$.

Then, we have

$$\begin{aligned}
\mathrm{SOS} &= \sum_{i_3,\dots,i_d \in [n]} \widetilde{\mathbb{E}}_\mu \left[q_{i_3,\dots,i_d} \cdot x^{(3)}_{i_3} \cdots x^{(d)}_{i_d} \right] \\
&\leqslant \sum_{i_3,\dots,i_d \in [n]} \sqrt{\widetilde{\mathbb{E}}_\mu \left[q^2_{i_3,\dots,i_d} \right]} \\
&\leqslant \sqrt{n^{d-2} \sum_{i_3,\dots,i_d \in [n]} \widetilde{\mathbb{E}}_\mu \left[q^2_{i_3,\dots,i_d} \right]} \\
&= n^{\frac{d}{2}-1} \sqrt{\widetilde{\mathbb{E}}_\mu \|\boldsymbol{Q}\|_{\mathrm{F}}^2},
\end{aligned}$$

using Cauchy-Schwarz and its pseudo-expectation version. Here, $\|\boldsymbol{Q}\|_{\mathrm{F}}^2$ denotes the sum of squared coefficients of $\boldsymbol{Q}$.

Let $\boldsymbol{h}^{(3)}, \dots, \boldsymbol{h}^{(d)}$ be i.i.d. uniform random vectors from $\{\pm 1\}^n$. Then,

$$\mathop{\mathbb{E}}_{\boldsymbol{h}^{(3)},\dots,\boldsymbol{h}^{(d)}} \left\langle \boldsymbol{Q}, \boldsymbol{h}^{(3)} \otimes \cdots \otimes \boldsymbol{h}^{(d)} \right\rangle^2 = \|\boldsymbol{Q}\|_{\mathrm{F}}^2\,.$$

Denote $\boldsymbol{H} := \boldsymbol{h}^{(3)} \otimes \cdots \otimes \boldsymbol{h}^{(d)}$. Then,

$$\mathrm{SOS}^2 \leqslant n^{d-2} \cdot \mathop{\mathbb{E}}_{\boldsymbol{H}} \widetilde{\mathbb{E}}_\mu \langle \boldsymbol{Q}, \boldsymbol{H} \rangle^2\,,$$

where the coefficients of $\boldsymbol{Q}$ are degree-2 polynomials in $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$.

We now describe the rounding algorithm.

1. Sample $\boldsymbol{h}^{(3)}, \dots, \boldsymbol{h}^{(d)} \sim \{\pm 1\}^n$, and set $\overline{\boldsymbol{x}}^{(j)} := \boldsymbol{h}^{(j)}$ for $j \geqslant 3$. Denote $\boldsymbol{H} := \boldsymbol{h}^{(3)} \otimes \cdots \otimes \boldsymbol{h}^{(d)}$.

2. Apply Lemma 8.13 to obtain from μ a degree-2 pseudo-distribution μ' satisfying
$$\left|\widetilde{\mathbb{E}}_{\mu'} \langle Q, H\rangle\right| \geqslant \frac{1}{3}\sqrt{\widetilde{\mathbb{E}}_{\mu} \langle Q, H\rangle^2}\,.$$
3. Use Grothendieck rounding (Theorem 8.9) on μ' to obtain $\overline{x}^{(1)}, \overline{x}^{(2)} \in \{\pm 1\}^n$ such that
$$\left\langle Q(x^{(1)}, x^{(2)}), H\right\rangle \geqslant \frac{1}{K_G} \cdot \left|\widetilde{\mathbb{E}}_{\mu'} \langle Q, H\rangle\right|\,.$$

Note that $\widetilde{\mathbb{E}}_{\mu} \langle Q, H\rangle^2$ is a degree-$2(d-2)$ polynomial in $h^{(3)}, \dots, h^{(d)}$. By Lemma 8.11 and hypercontractivity over the hypercube,
$$\Pr_{H}\left[\widetilde{\mathbb{E}}_{\mu} \langle Q, H\rangle^2 \geqslant \mathop{\mathbb{E}}_{H} \widetilde{\mathbb{E}}_{\mu} \langle Q, H\rangle^2\right] \geqslant 2^{-O(d)}\,.$$
Hence, with probability at least $2^{-O(d)}$, we get a "good" H such that $\widetilde{\mathbb{E}}_{\mu} \langle Q, H\rangle^2 \geqslant n^{-(d-2)} \cdot \mathrm{SOS}^2$.

Putting everything together, with probability at least $2^{-O(d)}$, we obtain $\overline{x}^{(1)}, \dots, \overline{x}^{(d)}$ such that
$$f(\overline{x}^{(1)}, \dots, \overline{x}^{(d)}) \geqslant \Omega(n^{-\frac{d}{2}+1}) \cdot \mathrm{SOS}\,.$$
Repeating the above $\mathrm{poly}(n, 2^d)$ times, we can obtain a solution with value $\Omega(n^{-\frac{d}{2}+1}) \cdot \mathrm{SOS}$ with high probability. This completes the proof. □

Combining Theorem 8.28 with our decoupling inequalities Lemma 8.25 and Lemma 8.27, we deduce that the same approximation guarantees hold in the general, non-decoupled case, for maximizing an odd-degree homogeneous polynomial, or maximizing the absolute value of an homogeneous polynomial of any degree. While we stated our results on the hypercube, the same holds for maximizing over the unit sphere, using the Gaussian rounding from Theorem 8.22 instead of Grothendieck's inequality.

8.7. Improved approximation algorithms for Max-3-SAT

In this section, we consider 3-Sat formulas where each 3-tuple of variables appears at most once, i.e. there are no two clauses with the same set of variables. Håstad and Venkatesh [HV04] used an anti-concentration result of [AGK04] to prove that any 3-Sat formula with m clauses has value at least $\frac{7}{8} + \Omega(\frac{1}{\sqrt{m}})$ (which is achieved by a random assignment with probability $\Omega(\frac{1}{m})$).

We prove the following improvement over this result:

Theorem 8.29. *There is a polynomial-time randomized algorithm that, given a satisfiable 3-Sat formula with n variables, finds with high probability an assignment satisfying a $(\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{3}{4}}))$-fraction of the clauses.*

Notations. A 3-Sat clause C with variables x_1, x_2, x_3 and literals $(\sigma_1, \sigma_2, \sigma_3) \in \{\pm 1\}^3$ can be written as

$$\psi_C(x_1, x_2, x_3) = \frac{7}{8} - \frac{1}{8}(\sigma_1 x_1 + \sigma_2 x_2 + \sigma_3 x_3 + \sigma_1\sigma_2 x_1 x_2 + \sigma_2\sigma_3 x_2 x_3 + \sigma_1\sigma_3 x_1 x_3 + \sigma_1\sigma_2\sigma_3 x_1 x_2 x_3) .$$

Here we adopt the convention that -1 is True and $+1$ is False, and we can see that $\psi_C(x) = 0$ if $\sigma_1 x_1 = \sigma_2 x_2 = \sigma_3 x_3 = +1$, and $\psi_C(x) = 1$ otherwise. Thus, a 3-Sat formula can be represented as a function $\psi : \{\pm 1\}^n \to [0, 1]$,

$$\psi(x) = \frac{7}{8} + f_1(x) + f_2(x) + f_3(x) ,$$

where f_1, f_2, f_3 are homogeneous polynomials of degree 1, 2 and 3 respectively.

Observe that

$$\max_{x\in\{\pm1\}^n} |f_1(\boldsymbol{x})| , \quad \max_{x\in\{\pm1\}^n} |f_2(\boldsymbol{x})| \leqslant \frac{3}{8} , \quad \max_{x\in\{\pm1\}^n} |f_3(\boldsymbol{x})| \leqslant \frac{1}{8} .$$

In particular, this last statement implies the following crucial observation:

Observation 8.30. If $\boldsymbol{x}^*$ is a satisfying assignment, then $f_1(\boldsymbol{x}^*) + f_2(\boldsymbol{x}^*) \geqslant 0$.

Before proceeding to describing the algorithm, we show the following version of degree-3 decoupling that augments the guarantees of Lemma 8.10 by controlling also the degree-1 and degree-2 parts:

Lemma 8.31 (Recoupling). *Given $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{\pm 1\}^n$, there exists a polynomial-time sampleable distribution $\mathcal{D}$ over $\{\pm 1\}^n$ such that*

1. *For any degree-3 homogeneous multilinear polynomial*

$$f(\boldsymbol{x}) = \sum_{i,j,k\in[n]} T_{ijk} x_i x_j x_k$$

(where $\boldsymbol{T}$ is a symmetric 3-tensor), let the corresponding decoupled polynomial be

$$\widetilde{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \sum_{i,j,k\in[n]} T_{ijk} x_i y_j z_k .$$

Then,

$$\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{D}} [f(\boldsymbol{x}')] = \frac{2}{9} \cdot \widetilde{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) .$$

2. *For any degree-2 homogeneous multilinear polynomial $g(\boldsymbol{x}) = \sum_{i,j\in[n]} M_{ij} x_i x_j$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{x}'\sim\mathcal{D}} [g(\boldsymbol{x}')] = \frac{1}{9} \cdot (g(\boldsymbol{x}) + g(\boldsymbol{y}) + g(\boldsymbol{z})) .$$

3. $\mathbb{E}_{x' \sim \mathcal{D}} [x'] = 0$.

Proof. The distribution $\mathcal{D}$ can be sampled as follows:

- Sample b_1 and b_2 independently and uniformly in $\{\pm 1\}$, and let $b_3 = b_1 b_2$. Then (b_1, b_2, b_3) has a pairwise independent distribution and $b_1 b_2 b_3 = 1$.
- Independently for each $i \in [n]$, sample x'_i uniformly in the multiset $\{b_1 x_i, b_2 y_i, b_3 z_i\}$.

Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [f(x')] &= \sum_{i,j,k} T_{ijk} \cdot \mathbb{E} \left[\frac{b_1 x_i + b_2 y_i + b_3 z_i}{3} \cdot \frac{b_1 x_j + b_2 y_j + b_3 z_j}{3} \cdot \frac{b_1 x_k + b_2 y_k + b_3 z_k}{3} \right] \\
&= \frac{1}{27} \sum_{i,j,k} T_{ijk} \cdot (x_i y_j z_k + x_i z_j y_k + x_j y_i z_k + x_j y_k z_i + x_k y_j z_i + x_k y_i z_j) \\
&= \frac{2}{9} \cdot \widetilde{f}(x, y, z) \,.
\end{aligned}$$

Similarly, for degree-2 multilinear polynomials,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x' \sim \mathcal{D}} [g(x')] &= \sum_{i,j} M_{ij} \cdot \mathbb{E} \left[\frac{b_1 x_i + b_2 y_i + b_3 z_i}{3} \cdot \frac{b_1 x_j + b_2 y_j + b_3 z_j}{3} \right] \\
&= \frac{1}{9} \cdot (g(x) + g(y) + g(z)) \,.
\end{aligned}$$

The third part follows similarly. □

The algorithm. We now describe the algorithm that achieves the guarantees of Theorem 8.29. We define $\delta := \frac{c}{\sqrt{n} \log n}$ (for some small constant $c > 0$ to be picked at the end of Lemma 8.35).

The first three steps in the algorithm correspond to each of the degree-1, 2, or 3 part being large. On a high level, our strategy is quite natural. If the degree-2 part is large, we use the classical roundings for degree-2 polynomials. If the degree-3 part is large, then we use our rounding for decoupled degree-3 polynomials from §8.3. In those two cases, we get an assignment with value $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{1}{2}})$. If the degree-1 part is large, we introduce a different algorithm based on a degree-2 SoS relaxation with additional axioms. Our rounding in this case is inspired by the proof of Theorem 8.7 and uses an additional idea to make the degree-3 part negligible. In this last case, we get an assignment with value $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{3}{4}})$.

We analyze separately these three cases now.

Lemma 8.32 (Large degree-1 part)**.** *Suppose that μ is a degree-2 pseudo-distribution such that:*

$$\left| \widetilde{\mathop{\mathbb{E}}_{\mu}} [f_1(x)] \right| \geqslant \delta \,, \quad \widetilde{\mathop{\mathbb{E}}_{\mu}} [f_1(x) + f_2(x)] \geqslant 0 \,.$$

Algorithm 2 Approximation algorithm for 3-Sat

Input: A 3-Sat instance in n variables: $\psi(\boldsymbol{x}) = \frac{7}{8} + f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) + f_3(\boldsymbol{x})$, where f_1, f_2, f_3 are homogeneous polynomials of degree 1, 2, and 3 respectively.
Output: An assignment $\boldsymbol{x} \in \{\pm 1\}^n$ with value $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{3}{4}})$ with high probability.

1. Solve feasibility for the following two degree-2 SoS programs over variables $\boldsymbol{x} = (x_1, \ldots, x_n)$:
   a) With axioms $x_i^2 = 1$ for all $i \in [n]$, $f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \geqslant 0$, and $f_1(\boldsymbol{x}) > \delta$;
   b) With axioms $x_i^2 = 1$ for all $i \in [n]$, $f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) \geqslant 0$, and $f_1(\boldsymbol{x}) < -\delta$.
2. If neither program is feasible, move to the next case.
3. Otherwise, let μ be either a feasible degree-2 pseudo-distribution for (a), or the negation of a feasible degree-2 pseudo-distribution for (b).
   - Sample $\boldsymbol{g} \sim \mathcal{N}(\widetilde{\mathbb{E}}_\mu \boldsymbol{x}, \widetilde{\mathbb{E}}_\mu (\boldsymbol{x} - \widetilde{\mathbb{E}}_\mu \boldsymbol{x})(\boldsymbol{x} - \widetilde{\mathbb{E}}_\mu \boldsymbol{x})^\top)$.
   - For all $i \in [n]$, set
   $$\overline{x}_i = \begin{cases} \frac{g_i}{T} & \text{if } |g_i| \leqslant T \\ \operatorname{sign}(g_i) & \text{otherwise} \end{cases}$$
   for some parameter $T = T(n) > 0$ (to be chosen in Lemma 8.32).
   - Sample
   $$x_i^{(1)} = \begin{cases} 1 & \text{with probability } \frac{1+p\overline{x}_i}{2} \\ -1 & \text{with probability } \frac{1-p\overline{x}_i}{2} \end{cases}$$
   independently for all $i \in [n]$, where $p = p(n)$ (to be chosen in Lemma 8.32).
4. Get $\boldsymbol{x}'$ by optimizing $f_2(\boldsymbol{x})$ using Charikar–Wirth rounding (Theorem 8.7). Set
   $$\boldsymbol{x}^{(2)} = \operatorname*{arg\,max}_{\boldsymbol{y} \in \{\boldsymbol{x}', -\boldsymbol{x}'\}} f_2(\boldsymbol{y}).$$
5. Let $\widetilde{f_3}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ be the decoupled polynomial of $f_3(\boldsymbol{x})$.
   - Run the degree-6 SoS relaxation of cubic optimization of $\widetilde{f_3}$ over variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{\pm 1\}^n$ with additional axioms $f_2(\boldsymbol{y}) \geqslant -\delta, \quad f_2(\boldsymbol{z}) \geqslant -\delta$.
   - Sample $\overline{\boldsymbol{x}} \sim \{\pm 1\}^n$ and reweight the pseudo-distribution μ as in the algorithm of Theorem 8.12.
   - Apply Charikar–Wirth rounding (Theorem 8.7) on the reweighted μ' to get $\overline{\boldsymbol{y}}, \overline{\boldsymbol{z}} \in \{\pm 1\}^n$.
   - Obtain $\boldsymbol{x}^{(3)}$ by recoupling $\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}, \overline{\boldsymbol{z}}$ via Lemma 8.31.
6. Pick the best assignment among $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}$.
7. Repeat steps 1–6 $\operatorname{poly}(n)$ times and output the best assignment obtained.

Then $\boldsymbol{x}^{(1)}$ *has value* $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{3}{4}})$.

Proof. First, up to negating the pseudo-distribution (which does not affect the degree-2 part), we can assume without loss of generality that $\widetilde{\mathbb{E}}_\mu [f_1(\boldsymbol{x})] \geqslant \delta$.

The rounding proceeds in two steps. We introduce some parameters $p = p(n) \in [0,1]$ and $T = T(n) > 0$ to be fixed later.

1. First, sample a Gaussian with mean and covariance matching what the degree-2 pseudo-distribution indicates, that is: $\boldsymbol{g} \sim \mathcal{N}(\widetilde{\mathbb{E}}_\mu \boldsymbol{x}, \widetilde{\mathbb{E}}_\mu (\boldsymbol{x} - \widetilde{\mathbb{E}}_\mu \boldsymbol{x})(\boldsymbol{x} - \widetilde{\mathbb{E}}_\mu \boldsymbol{x})^\top)$. Then, let $\overline{\boldsymbol{x}} \in [-1,1]^n$ be defined as follows: for each $i \in [n]$, let $\overline{x}_i = \frac{g_i}{T}$ if $|g_i| \leqslant T$ and $\overline{x}_i = \mathrm{sign}(g_i)$ otherwise.
2. Use $\overline{\boldsymbol{x}}$ as a bias for sampling $\boldsymbol{x}^{(1)}$: sample for each $i \in [n]$ independently: $x_i^{(1)} = 1$ with probability $\frac{1+p\overline{x}_i}{2}$ and $x_i^{(1)} = -1$ with probability $\frac{1-p\overline{x}_i}{2}$.

We now analyze this rounding. Define

$$\Delta_i \coloneqq \mathbb{E}\, \overline{x}_i - \frac{1}{T}\mathbb{E}\, g_i, \quad \Delta_{ij} \coloneqq \mathbb{E}\left[\overline{x}_i \overline{x}_j\right] - \frac{1}{T^2}\mathbb{E}\left[g_i g_j\right] .$$

Then, at the end of the first step, we have (here, $\|f_1\|_1$ and $\|f_2\|_1$ denote the sum of the absolute value of the coefficients of f_1 and f_2, respectively):

$$\mathbb{E}\,[f_1(\overline{\boldsymbol{x}})] \geqslant \frac{1}{T}\widetilde{\mathbb{E}}\,[f_1(\boldsymbol{x})] - \|f_1\|_1 \cdot \max_{1 \leqslant i \leqslant n} |\Delta_i| , \tag{8.10}$$

$$\mathbb{E}\,[f_2(\overline{\boldsymbol{x}})] \geqslant \frac{1}{T^2}\widetilde{\mathbb{E}}\,[f_2(\boldsymbol{x})] - \|f_2\|_1 \cdot \max_{1 \leqslant i,j \leqslant n} |\Delta_{ij}| . \tag{8.11}$$

Claim 8.33. $\max_{1 \leqslant i \leqslant n} |\Delta_i|$ *and* $\max_{1 \leqslant i,j \leqslant n} |\Delta_{ij}|$ *are both at most* $O(1) \cdot e^{-\frac{T^2}{8}}$.

Proof. Fix $i \in [n]$. First,

$$\Delta_i = \frac{1}{T}\mathbb{E}\left[g_i(\mathbf{1}_{|g_i| \leqslant T} - 1)\right] + \mathbb{E}\left[\mathrm{sign}(g_i)\mathbf{1}_{|g_i| > T}\right] .$$

Since μ is a degree-2 pseudo-distribution over the hypercube, g_i is a Gaussian with mean $\widetilde{\mathbb{E}}_\mu x_i \in [-1,1]$ and variance 1. Define $p_i \coloneqq \Pr(|g_i| > T) \leqslant e^{-T^2/4}$ (this holds provided T is a large enough constant). By the triangle inequality and Cauchy-Schwarz, we have

$$|\Delta_i| \leqslant \frac{1}{T}\left|\mathbb{E}\left[g_i \mathbf{1}_{|g_i| > T}\right]\right| + p_i \leqslant \frac{1}{T}\sqrt{2p_i} + p_i \leqslant O(1) \cdot e^{-\frac{T^2}{8}} .$$

Fix now $i, j \in [n]$. Similarly,

$$\begin{aligned}\Delta_{ij} = {} & \frac{1}{T^2}\mathbb{E}\left[g_i g_j \left(\mathbf{1}_{|g_i| \leqslant T, |g_j| \leqslant T} - 1\right)\right] + \frac{1}{T}\mathbb{E}\left[\mathrm{sign}(g_i) g_j \mathbf{1}_{|g_i| > T, |g_j| \leqslant T}\right] \\ & + \frac{1}{T}\mathbb{E}\left[g_i \mathrm{sign}(g_j)\mathbf{1}_{|g_i| \leqslant T, |g_j| > T}\right] + \mathbb{E}\left[\mathrm{sign}(g_i)\mathrm{sign}(g_j)\mathbf{1}_{|g_i| > T, |g_j| > T}\right] .\end{aligned}$$

Note that (g_i, g_j) is a 2-dimensional Gaussian with marginal means bounded by 1 in absolute value and $\left|\mathbb{E}[g_i g_j]\right| = \left|\widetilde{\mathbb{E}}_\mu[x_i x_j]\right| \leqslant 1$ by Fact 8.3. Hence, using once again the triangle inequality and Cauchy-Schwarz,

$$\begin{aligned}
|\Delta_{ij}| &\leqslant \frac{1}{T^2}\left|\mathbb{E}[g_i g_j \mathbf{1}_{|g_i|>T \text{ or } |g_j|>T}]\right| + \frac{1}{T}\left(\mathbb{E}\left[|g_j|\mathbf{1}_{|g_i|>T}\right] + \mathbb{E}\left[|g_i|\mathbf{1}_{|g_j|>T}\right]\right) + p_i \\
&\leqslant \frac{1}{T^2}\sqrt{\mathbb{E}\left[g_i^2 g_j^2\right](p_i + p_j)} + \frac{\sqrt{2}}{T}\left(\sqrt{p_i} + \sqrt{p_j}\right) + p_i \\
&\leqslant O(1)\cdot e^{-\frac{T^2}{8}}\,.
\end{aligned}$$

□

Since f_1 and f_2 come from a 3-Sat instance with $m \leqslant n^3$ clauses, we have $\|f_1\|_1, \|f_2\|_1 \leqslant O(n^3)$. Thus, if we pick $T = \sqrt{48\log n}$, we get

$$\begin{aligned}
\mathbb{E}\left[f_1(\overline{x}) + f_2(\overline{x})\right] &\geqslant \frac{1}{\sqrt{48\log n}}\widetilde{\mathbb{E}}\left[f_1(x)\right] + \frac{1}{48\log n}\widetilde{\mathbb{E}}\left[f_2(x)\right] - O\left(\frac{1}{n^2}\right) \\
&\geqslant -O\left(\frac{1}{n^2}\right),
\end{aligned}$$

where we used our assumption $\widetilde{\mathbb{E}}_\mu\left[f_1(x) + f_2(x)\right] \geqslant 0$, together with the fact that

$$\widetilde{\mathbb{E}}\left[f_1(x)\right] \geqslant 0\,.$$

Next, at the end of second step, it holds that:

$$\mathbb{E}\left[f_1(x^{(1)})\right] = p\,\mathbb{E}\left[f_1(\overline{x})\right], \quad \mathbb{E}\left[f_2(x^{(1)})\right] = p^2\,\mathbb{E}\left[f_2(\overline{x})\right], \tag{8.12}$$

$$\mathbb{E}\left[f_3(x^{(1)})\right] = p^3\,\mathbb{E}\left[f_3(\overline{x})\right] \geqslant -p^3\,. \tag{8.13}$$

Hence, by (8.12) and (8.10):

$$\begin{aligned}
\mathbb{E}\left[f_1(x^{(1)}) + f_2(x^{(1)})\right] &\geqslant (p - p^2)\,\mathbb{E}\left[f_1(\overline{x})\right] + p^2\,\mathbb{E}\left[f_1(\overline{x}) + f_2(\overline{x})\right] \\
&\geqslant \Omega(1)\cdot\frac{\delta(p-p^2)}{\sqrt{\log n}} - O\left(\frac{p^2}{n^2}\right) - O\left(\frac{p}{n^2}\right).
\end{aligned}$$

Setting $p = \frac{n^{-\frac{1}{4}}}{\log n}$, we get

$$\mathbb{E}\left[f_1(x^{(1)}) + f_2(x^{(1)})\right] \geqslant \Omega(1)\cdot\frac{n^{-\frac{3}{4}}}{\log^{2.5} n},$$

and we conclude by combining with (8.13). □

Lemma 8.34 (Large degree-2 part)**.** *Suppose that* $\max f_2(\boldsymbol{x}) \geqslant \delta$*. Then* $\boldsymbol{x}^{(2)}$ *has value* $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{1}{2}})$.

Proof. Using Charikar-Wirth rounding (Theorem 8.7 with $T = \Theta(\sqrt{\log n})$), we get that

$$f_2(\boldsymbol{x}') \geqslant \Omega\left(\frac{1}{\log n}\right) \max_{\boldsymbol{x}\in\{\pm 1\}^n} f_2(\boldsymbol{x})\,.$$

Moreover, f_2 is invariant by the choice of the signing $\pm\boldsymbol{x}'$, while $f_1 + f_3$ changes sign. Thus, $\boldsymbol{x}^{(2)}$ has value $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{1}{2}})$ (recall that $\delta = \frac{c}{\sqrt{n}\log n}$). □

Lemma 8.35 (Large degree-3 part)**.** *Suppose that* $\max_{\boldsymbol{x}} f_2(\boldsymbol{x}) \leqslant \delta$ *and* $|f_1(\boldsymbol{x}^*)| \leqslant \delta$ *for some assignment* $\boldsymbol{x}^*$ *satisfying the 3-SAT formula. Then* $\boldsymbol{x}^{(3)}$ *has value* $\frac{7}{8} + \widetilde{\Omega}(n^{-\frac{1}{2}})$.

Proof. We run the canonical degree-6 SoS relaxation on the decoupled polynomial $\widetilde{f_3}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ associated to f_3 with variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \{\pm 1\}^n$, and the additional axioms $f_2(\boldsymbol{y}) \geqslant -\delta$ and $f_2(\boldsymbol{z}) \geqslant -\delta$. Let μ denote the resulting pseudo-distribution.

Since $\max_{\boldsymbol{x}} f_2(\boldsymbol{x}) \leqslant \delta$, $|f_1(\boldsymbol{x}^*)| \leqslant \delta$ and $\delta = o(1)$, we must have $f_3(\boldsymbol{x}^*) \geqslant \frac{1}{8} - o(1)$. Using Observation 8.30, we get that the delta pseudo-distribution centered at $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{z} = \boldsymbol{x}^*$ is a feasible solution to the SoS relaxation, so that μ satisfies

$$\widetilde{\mathbb{E}}_{\mu} \widetilde{f_3}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \geqslant \frac{1}{8} - o(1), \quad \widetilde{\mathbb{E}}_{\mu} f_2(\boldsymbol{y}) \geqslant -\delta, \quad \widetilde{\mathbb{E}}_{\mu} f_2(\boldsymbol{z}) \geqslant -\delta\,.$$

Next, we sample $\overline{\boldsymbol{x}} \sim \{\pm 1\}^n$ and reweight μ using Lemma 8.13 to get a degree-2 pseudo-distribution μ'. The same analysis shows that with at least constant probability, $\overline{\boldsymbol{x}}$ satisfies

$$\widetilde{\mathbb{E}}_{\mu'} \widetilde{f_3}(\overline{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}) \geqslant \Omega\left(\frac{1}{\sqrt{n}}\right)\,. \tag{8.14}$$

Furthermore, we claim that for $C > 1$,

$$\Pr_{\overline{\boldsymbol{x}}\sim\{\pm 1\}^n}[f_2(\overline{\boldsymbol{x}}) \geqslant -C\delta] \geqslant 1 - \frac{1}{C}\,.$$

To see this, note that $\mathbb{E}_{\overline{\boldsymbol{x}}} f_2(\overline{\boldsymbol{x}}) = 0$ because f_2 is multilinear. The above bound then follows from the assumption $\max_{\boldsymbol{x}} f_2(\boldsymbol{x}) \leqslant \delta$. Therefore, by a union bound, with at least constant probability we get a good $\overline{\boldsymbol{x}}$ that satisfies simultaneously (8.14) and $f_2(\overline{\boldsymbol{x}}) \geqslant -C\delta$. By repeating the sampling $\mathrm{poly}(n)$ times, we can find such an $\overline{\boldsymbol{x}} \in \{\pm 1\}^n$ with high probability.

Now, fix $\overline{\boldsymbol{x}} \in \{\pm 1\}^n$ satisfying the previous conditions. We apply Theorem 8.7 to the degree-2 pseudo-distribution μ' over $(\boldsymbol{y}, \boldsymbol{z}) \in \{\pm 1\}^{2n}$ for some $T > 0$ to be fixed later.

Denoting by $\|f_2\|_1$ (resp. $\|f_3\|_1$) the sum of the absolute value of the coefficients of f_2 (resp. f_3), we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\overline{\boldsymbol{y}}}[f_2(\overline{\boldsymbol{y}})] &\geqslant \frac{1}{T^2}\mathop{\widetilde{\mathbb{E}}}_{\mu'}[f_2(\boldsymbol{y})] - 8e^{-\frac{T^2}{2}}\|f_2\|_1\,,\\
\mathop{\mathbb{E}}_{\overline{\boldsymbol{z}}}[f_2(\overline{\boldsymbol{z}})] &\geqslant \frac{1}{T^2}\mathop{\widetilde{\mathbb{E}}}_{\mu'}[f_2(\boldsymbol{z})] - 8e^{-\frac{T^2}{2}}\|f_2\|_1\,,\\
\mathop{\mathbb{E}}_{\overline{\boldsymbol{y}},\overline{\boldsymbol{z}}}\left[\widetilde{f_3}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}})\right] &\geqslant \frac{1}{T^2}\mathop{\widetilde{\mathbb{E}}}_{\mu'}\left[\widetilde{f_3}(\overline{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{z})\right] - 8e^{-\frac{T^2}{2}}\|f_3\|_1\,.
\end{aligned}$$

The constraints $f_2(\boldsymbol{y}) \geqslant -\delta$ and $f_2(\boldsymbol{z}) \geqslant -\delta$ still hold for the reweighted pseudo-distribution μ'. Moreover, $\|f_2\|_1$ and $\|f_3\|_1$ are both $O(n^3)$, so by picking $T = \sqrt{8\log n}$, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{\overline{\boldsymbol{y}}}[f_2(\overline{\boldsymbol{y}})] &\geqslant -O\left(\frac{\delta}{\log n}\right) - O\left(\frac{1}{n}\right),\\
\mathop{\mathbb{E}}_{\overline{\boldsymbol{z}}}[f_2(\overline{\boldsymbol{z}})] &\geqslant -O\left(\frac{\delta}{\log n}\right) - O\left(\frac{1}{n}\right),\\
\mathop{\mathbb{E}}_{\overline{\boldsymbol{y}},\overline{\boldsymbol{z}}}\left[\widetilde{f_3}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}})\right] &\geqslant \Omega\left(\frac{1}{\sqrt{n}\log n}\right) - O\left(\frac{1}{n}\right).
\end{aligned}$$

Finally, once we have $\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}}$, we recouple them to get $\boldsymbol{x}^{(3)}$ by Lemma 8.31, obtaining

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}^{(3)}} f_3(\boldsymbol{x}^{(3)}) &= \frac{2}{9}\cdot\widetilde{f_3}(\overline{\boldsymbol{x}},\overline{\boldsymbol{y}},\overline{\boldsymbol{z}})\,,\\
\mathop{\mathbb{E}}_{\boldsymbol{x}^{(3)}} f_2(\boldsymbol{x}^{(3)}) &= \frac{1}{9}\cdot(f_2(\overline{\boldsymbol{x}}) + f_2(\overline{\boldsymbol{y}}) + f_2(\overline{\boldsymbol{z}}))\,,\\
\mathop{\mathbb{E}}_{\boldsymbol{x}^{(3)}} f_1(\boldsymbol{x}^{(3)}) &= 0\,.
\end{aligned}$$

Thus, we have

$$\mathbb{E}\,\psi(\boldsymbol{x}^{(3)}) \geqslant \frac{7}{8} + \Omega\left(\frac{1}{\sqrt{n}\log n}\right) - O(\delta) = \frac{7}{8} + \widetilde{\Omega}\left(\frac{1}{\sqrt{n}}\right),$$

where the last equality holds provided that we pick the constant $c > 0$ in the definition of δ to be small enough. This concludes the proof. □

Proof of Theorem 8.29. We prove that one of the assumptions of Lemma 8.32, Lemma 8.34 or Lemma 8.35 must hold. Fix some satisfying assignment $\boldsymbol{x}^*$ to the 3-Sat formula.

1. If $|f_1(\boldsymbol{x}^*)| > \delta$, then one of the two SoS programs from Step 1 of Algorithm 2 is feasible, and the assumptions of Lemma 8.32 hold.
2. If $\max_{\boldsymbol{x}} f_2(\boldsymbol{x}) > \delta$, then the assumptions of Lemma 8.34 hold.
3. If $|f_1(\boldsymbol{x}^*)| \leqslant \delta$ and $\max_{\boldsymbol{x}} f_2(\boldsymbol{x}) \leqslant \delta$, then the assumptions of Lemma 8.35 hold.

Hence, in all cases we get a random assignment $\widetilde{\boldsymbol{x}} \in \{\pm 1\}^n$ satisfying $\mathbb{E}\,\psi(\widetilde{\boldsymbol{x}}) \geqslant \frac{7}{8} + \widetilde{\Omega}(n^{-\frac{3}{4}})$. By repeating the rounding $\mathrm{poly}(n)$ times, we can get such an assignment with high probability. □

8.8. Summary

This core chapter introduced our new rounding algorithm for higher-degree sum-of-squares relaxations, our compression of SDP relaxations which yields certification guarantees matching the result of Khot and Naor, and our improved approximation algorithm for Max-3-Sat.

Part III.

Discrepancy Theory

CHAPTER 9.

Spencer's Theorem via Regularization

In this third and final part of the thesis, we shift focus from polynomial optimization to the problem of *discrepancy minimization*. In discrepancy theory, the objective is the norm of a linear function:

$$f(\boldsymbol{x}) := \left\| \sum_{i=1}^{n} x_i \boldsymbol{u}_i \right\|, \tag{9.1}$$

where $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ are input vectors and $\|\cdot\|$ is a fixed norm. Unlike the problems considered in Part II, this function is usually not well-approximated by low-degree polynomials. However, a parallel story will appear: as before, (9.1) is a non-convex function, and as before, we study both the cases where $\boldsymbol{x}$ lies on the unit sphere $\mathcal{S}^{n-1}$ or the hypercube $\{-1, 1\}^n$ — the latter being the classical and more challenging setting in discrepancy theory.

In this chapter, we introduce a new framework for discrepancy minimization based on two ideas: (1) *regularizing* the objective (9.1), and (2) applying a second-order optimization method to the resulting problem. We apply this framework by to a new proof of a seminal result of Spencer from the 1980s [Spe85], which corresponds to the case where the vectors $\boldsymbol{u}_i$ have uniformly bounded entries and the norm $\|\cdot\|$ is ℓ_∞. Our proof is simple and has several new features compared to existing arguments.

Table of contents

A preliminary version of the results in this chapter appeared in [PV23].

9.1. Introduction

Our main result in this chapter is a new proof of the following result of Spencer [Spe85]:

Theorem 9.1. *For any $\boldsymbol{A} \in [-1,1]^{d \times n}$, there exists $\boldsymbol{x} \in \{-1,1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty \lesssim \sqrt{n \log\left(\frac{2d}{n}\right)}. \tag{9.2}$$

Note that the left-hand side of (9.2) matches (9.1), where the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are the columns of $\boldsymbol{A}$.

Several different proofs of Theorem 9.1 are known. The original argument of [Spe85] is based on the pigeonhole principle. A different method pioneered by Gluskin in the 1980s uses insights from convex geometry [Glu89, Gia97, Rot17, ES18]. Bansal [Ban10] analyzed for the first time a polynomial time algorithm finding a coloring matching the guarantees of (9.2). Bansal's algorithm is a random walk in the hypercube whose increments are chosen by solving a semidefinite program. This approach was simplified by Lovett and Meka [LM15]. More recently, Levy, Ramadas, and Rothvoss [LRR17] and Bansal, Laddha, and Vempala [BLV22] introduced new deterministic algorithms for the problem which are closer to our framework. We will compare our approach with theirs.

Random coloring. As emphasized in Chapter 1, Spencer's guarantees go much beyond what the probabilistic method shows. When $n = d$, Theorem 9.1 shows that any sequence of n bounded vectors in $\mathbb{R}^n$ have a coloring $\boldsymbol{x}$ of discrepancy $O\left(\sqrt{n}\right)$. However, if $\boldsymbol{x} \sim \{-1,1\}^n$ is a uniformly random coloring, then the discrepancy of a fixed constraint exceeds t with probability $e^{-t^2/2n}$. To take a union bound over all n constraints, we would need to pick $t = O(\sqrt{n \log n})$, which gives a discrepancy bound which is away by a $\sqrt{\log n}$-factor to the bound in Theorem 9.1. This is not an artifact of the analysis: in some instances, the set of colorings matching Spencer's guarantees has exponentially small measure.

9.2. A new approach to Spencer's theorem

In this section, we give an overview of our approach to prove Theorem 9.1. Mirroring Part II, our proof will be algorithmic: we design an algorithm that takes as input $\boldsymbol{A}$ and outputs a coloring $\boldsymbol{x}$ satisfying (9.2). This section is informal; proofs are delayed to §9.3 and §9.4.

9.2.1. Newton's method on a regularized objective

We introduce our algorithm in the case $d = n$ for simplicity.

Our algorithm is a sticky walk in the hypercube. We build a deterministic sequence $\boldsymbol{x}(t) \coloneqq (x_1(t), \dots, x_n(t))$ for times $t \in [0, T]$. At any time step t, $\boldsymbol{x}(t)$ will be an element of the solid hypercube $[-1, 1]^n$ that represents a partial coloring. We start from $\boldsymbol{x}(0) \coloneqq (0, \dots, 0)$ and the dynamic ends when $\boldsymbol{x}(t)$ hits a corner of the hypercube. We define the set of active coordinates of a fractional coloring $\boldsymbol{x} \in [-1, 1]^n$ as

$$F \coloneqq F(\boldsymbol{x}) = \{j \in [n] : x_j \notin \{-1, 1\}\}\,.$$

The final algorithm described in §9.3 and §9.4 will essentially be a discretization of this continuous dynamic.

Regularized objective. In order to control the quantity $\|\boldsymbol{A}\boldsymbol{x}\|_\infty$ over the duration of the walk, we now define a smooth proxy for the ℓ_∞-norm. The following standard observation allows to replace the ℓ_∞-norm by a one-sided version:

Remark 9.2. Given $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, the matrix $\boldsymbol{A}' \coloneqq \begin{bmatrix} \boldsymbol{A} \\ -\boldsymbol{A} \end{bmatrix}$ has dimension $2d \times n$, and satisfies for any $\boldsymbol{x} \in \{-1, 1\}^n$:

$$\max_{j \in [d]} (\boldsymbol{A}'\boldsymbol{x})_j = \|\boldsymbol{A}\boldsymbol{x}\|_\infty\,.$$

Until Chapter 11, we will not make any distinction between d and $2d$, so up to changing $\boldsymbol{A}$ to $\boldsymbol{A}'$, this observation allows to bound only the maximal entry of $\boldsymbol{A}\boldsymbol{x}$ without loss of generality.

Naturally, for any $\boldsymbol{y} \in \mathbb{R}^n$, we can equivalently write:

$$\max_{i \in [n]} y_i = \max_{\boldsymbol{r} \in \Delta_n} \langle \boldsymbol{r}, \boldsymbol{y} \rangle\,, \quad \text{where } \Delta_n \coloneqq \left\{ \boldsymbol{r} \in \mathbb{R}^n_{\geqslant 0} : \sum_{i \leqslant n} r_i = 1 \right\}.$$

Instead, we consider the following *regularized* version of the right-hand side, which is the maximization problem where we added an $\ell_{1/2}$-type penalty for each element of the simplex:[1]

$$\omega^*(\boldsymbol{y}) \coloneqq \max_{\boldsymbol{r} \in \Delta_n} \langle \boldsymbol{r}, \boldsymbol{y} \rangle + 2 \sum_{i \leqslant n} r_i^{\frac{1}{2}}\,.$$

In what follows, $\omega^*(\boldsymbol{A}\boldsymbol{x})$ will play a role of proxy for $\|\boldsymbol{A}\boldsymbol{x}\|_\infty$. It is not hard to see (Lemma 9.11) that we only lose a $2\sqrt{n}$ additive factor through this approximation. Therefore, for proving Theorem 9.1, it suffices to bound the total increase of the regularized maximum. We will discuss in §9.2.3 and §9.4 the choice of this particular $\ell_{1/2}$-regularizer.

[1] The choice of this specific penalty may appear arbitrary at this point; we justify it in §9.2.2 and Remark 9.13.

Algorithm 3 Continuous dynamic for discrepancy minimization

1: **while** $F \neq \emptyset$ **do**
2: $\quad \boldsymbol{\delta} \leftarrow \underset{\substack{\boldsymbol{\delta}:\langle \boldsymbol{\delta}, \boldsymbol{x}\rangle = 0 \\ \text{and } \operatorname{supp}(\boldsymbol{\delta}) \subseteq F}}{\arg\min} \langle \boldsymbol{A}^\top \nabla \omega^*(\boldsymbol{A}\boldsymbol{x}), \boldsymbol{\delta}\rangle + \frac{1}{2} \cdot \boldsymbol{\delta}^\top \boldsymbol{A}^\top \nabla^2 \omega^*(\boldsymbol{A}\boldsymbol{x}) \boldsymbol{A}\boldsymbol{\delta}$
3: $\quad \boldsymbol{x} \leftarrow \boldsymbol{x} + \varepsilon \boldsymbol{\delta}$
4: $\quad F \leftarrow \{i \in [n] : x_i \notin \{-1, 1\}\}$
5: **end while**
6: **return** x

Continuous dynamic. We sketch our dynamic in pseudo-code in Algorithm 3. Essentially, we impose two conditions on the update direction $\boldsymbol{\delta}$: $\operatorname{supp}(\boldsymbol{\delta}) \subseteq F$ ensures that the walk stays in the solid hypercube by fixing the coordinates of $\boldsymbol{x}$ when they reach ± 1, while $\langle \boldsymbol{\delta}, \boldsymbol{x}\rangle = 0$ ensures that the dynamic will eventually converge to a corner of the hypercube. Under these constraints, we select the direction that minimizes the best quadratic approximation of our potential function $\omega^*(\boldsymbol{A}\boldsymbol{x})$. In this sense, this is essentially a Newton step.

Local analysis. We would like to bound the increase in potential,

$$\frac{\mathrm{d}\omega^*(\boldsymbol{A}\boldsymbol{x})}{\mathrm{d}\|\boldsymbol{x}\|_2^2} \approx \langle \nabla \omega^*(\boldsymbol{A}\boldsymbol{x}), \boldsymbol{A}\boldsymbol{\delta}\rangle + \frac{1}{2}\langle \boldsymbol{A}\boldsymbol{\delta}, \nabla^2 \omega^*(\boldsymbol{A}\boldsymbol{x}) \boldsymbol{A}\boldsymbol{\delta}\rangle\,,$$

where $\boldsymbol{\delta}$ is the minimizer on line 2 of Algorithm 3.

Since the quadratic term is invariant by sign changes $\pm\boldsymbol{\delta}$, we can always upper bound the term that is linear in $\boldsymbol{\delta}$ by 0. Thus, it suffices to prove that the matrix $\boldsymbol{A}^\top \nabla^2 \omega^*(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}$ has a small eigenvalue on the subspace $S := \{\boldsymbol{\delta} \in \mathbb{R}^n : \langle \boldsymbol{\delta}, \boldsymbol{x}\rangle = 0 \text{ and } \operatorname{supp}(\boldsymbol{\delta}) \subseteq F\}$. For this, we need to understand better the regularization construction — as we will see in Lemma 9.12, it follows from standard convex analysis arguments that

$$\nabla^2 \omega^*(\boldsymbol{A}\boldsymbol{x}) \preccurlyeq \operatorname{diag}(\boldsymbol{\nabla})^{\frac{3}{2}} \qquad \text{for some vector } \boldsymbol{\nabla} \in \Delta_n\,.$$

By further use of the orthogonality trick, we can select a slightly smaller subspace than S whose elements "do not see" the rows i for which $\nabla_i \gtrsim 1/|F|$. A random element $\boldsymbol{\delta}$ in this subspace achieves quadratic form at most $\|\boldsymbol{\delta}\|_2^2/\sqrt{|F|}$ in expectation. The details can be found in Lemma 9.15. This ultimately implies

$$\mathrm{d}\omega^*(\boldsymbol{A}\boldsymbol{x}) \lesssim \frac{\mathrm{d}\|\boldsymbol{x}\|_2^2}{\sqrt{|F|}}\,. \tag{9.3}$$

Analysis of the whole dynamic. For $k \in [n]$, denote by $t_k := \min\{t \geqslant 0 : |F(t)| \leqslant k\}$ the first time for which the number of active coordinates reaches k. From the constraint that the update direction is always orthogonal to the current partial coloring, we get after integrating (9.3) over $t \in [0, T]$ that

$$\omega^*(\boldsymbol{A}\boldsymbol{x}(T)) - \omega^*(\boldsymbol{0}) \lesssim \sum_{k=1}^{n} \frac{\|\boldsymbol{x}(t_{k-1})\|_2^2 - \|\boldsymbol{x}(t_k)\|_2^2}{\sqrt{k}}.$$

Finally, we apply summation by parts and use the fact that $\sum_{i \leqslant k} \|\boldsymbol{x}(t_{i-1})\|_2^2 - \|\boldsymbol{x}(t_i)\|_2^2 \leqslant k$:

$$\omega^*(\boldsymbol{A}\boldsymbol{x}(T)) - \omega^*(\boldsymbol{0}) \lesssim \sqrt{n} + \sum_{k=1}^{n-1} \frac{1}{k^{\frac{3}{2}}} \sum_{i \leqslant k} \|\boldsymbol{x}(t_{i-1})\|_2^2 - \|\boldsymbol{x}(t_i)\|_2^2 \lesssim \sqrt{n}\,.$$

Combining this with our previous observation that $\omega^*(\boldsymbol{0}) \lesssim \sqrt{n}$ concludes our proof sketch of Theorem 9.1 in the case $n = d$.

9.2.2. Barrier function interpretation

We highlight here one motivation behind the specific choice of the $\ell_{1/2}$-regularizer. We defined ω^* as the solution to a convex optimization problem which is trying to smooth out the ℓ_∞-norm. This optimization problem has an equivalent dual formulation which shows that it is also an "auto-adjusted" barrier function:

Claim 9.3. *For any $\boldsymbol{y} \in \mathbb{R}^n$,*

$$\omega^*(y) := \max_{\boldsymbol{r} \in \Delta_n} \langle \boldsymbol{r}, \boldsymbol{y} \rangle + 2 \sum_{i=1}^{n} r_i^{\frac{1}{2}} = \min_{\lambda > \max_{i \in [n]} x_i} \lambda + \sum_{i=1}^{n} \frac{1}{\lambda - y_i}\,.$$

The proof of Claim 9.3 is a standard dual derivation.

The barrier function $\boldsymbol{y} \mapsto \sum_{i=1}^{n} \frac{1}{\lambda - y_i}$ appears in multiple fields; in free probability, it is known as the *Cauchy* or *Stieltjes transform* $G_{\boldsymbol{y}}(\lambda)$. The inverse of that function, the *K-transform* $K_{\boldsymbol{y}}$, maps $(0, \infty)$ to $(\max_i y_i, \infty)$, so that

$$\omega^*(\boldsymbol{y}) = \min_{\lambda > \max_i y_i} \lambda + G_{\boldsymbol{y}}(\lambda) = \min_{v > 0} v + K_{\boldsymbol{y}}(v)\,.$$

This barrier function was first introduced by Batson, Spielman, and Srivastava [BSS14] in the context of graph sparsification, and later developed by Marcus, Spielman, and Srivastava [MSS22, MSS18] as part of their theory of finite free probability. While in previous works [BSS14, BLV22], a parameter $\lambda = \lambda(t)$ has to be carefully chosen and tracked during the analysis, here we make the natural choice of λ which makes λ and $G_{\boldsymbol{y}}(\lambda)$ on the same scale. As we will see, this turns out to significantly simplify the analysis of this kind of methods.

9.2.3. Regret minimization interpretation

We now give an alternative derivation of our algorithm as a 2-players game between a builder and an inspector. We refer to the series of blog posts [Tre19] for other applications of this perspective. We note that connections between regret minimization, barrier potential functions, and second-order differential equations appear in various fields; see [KS05].

Using the observation from Remark 9.2, the discrepancy of a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ can be written in the following min-max form:

$$\operatorname{disc}(\boldsymbol{A}) = \min_{\boldsymbol{x} \in \{-1,1\}^n} \max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x} \rangle \ ,$$

where we rewrote the inner optimization problem to make it obviously convex. With this formulation in mind, it is natural to view discrepancy minimization as a 2-players game between a builder who constructs a coloring $\boldsymbol{x} \in \{-1, 1\}^n$, and an inspector who picks test vectors $\boldsymbol{r} \in \Delta_d$.

We consider a version of this game indexed by time $t \in [T]$. At every time step,

1. The inspector picks a test vector $\boldsymbol{r}(t)$.
2. The builder picks an update vector $\boldsymbol{\delta}(t)$.
3. The inspector gets payoff $\langle \boldsymbol{r}(t), \boldsymbol{A}\boldsymbol{\delta}(t) \rangle$.

Let $\boldsymbol{x}(T) = \boldsymbol{\delta}(1) + \ldots + \boldsymbol{\delta}(T)$ be the vector constructed by the builder. At the end of the T rounds, the best payoff in hindsight for the inspector is

$$\max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x}(T) \rangle \ , \tag{9.4}$$

which is precisely the discrepancy objective. Following the standard online optimization terminology, we call regret of the inspector the quantity

$$\operatorname{Regret}(T) := \max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x}(T) \rangle - \sum_{t=1}^{T} \langle \boldsymbol{r}(t), \boldsymbol{A}\boldsymbol{\delta}(t) \rangle \ . \tag{9.5}$$

Strategy for the inspector. Following this analogy, the inspector will play a low-regret strategy to make sure that her total payoff is not much smaller than the discrepancy objective (9.4). A standard class of low-regret strategy for online convex optimization are Follow the Regularized Leader (FTRL) strategies. For a given regularizer $\omega : \Delta_d \to \mathbb{R}_{\geqslant 0}$, the inspector picks the test vector

$$\boldsymbol{r}(t+1) = \arg\max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x}(t) \rangle + \omega(\boldsymbol{r}) \, .$$

Note that without the $\omega(\boldsymbol{r})$ term, no non-trivial upper bound on the regret of such a strategy would be possible. This is because the builder could pick two discrepancy constraints and alternate update vectors that incur a large cost on these two constraints. In this case, the regret (9.5) would be as large as $T/2$.

Regret upper bound. This regret minimization framework can be equivalently seen as a way to come up with good smooth potential functions. Let

$$\Phi(\boldsymbol{x}) \coloneqq \max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x} \rangle + \omega(\boldsymbol{r}) \,.$$

The standard way to upper bound the regret of a strategy in online learning consists in using Φ as a potential function and analyze how much it increases during the game. Indeed, since ω takes non-negative values, we have

$$\Phi(x(T)) \geqslant \max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{A}\boldsymbol{x} \rangle \;, \quad \Phi(\boldsymbol{x}(0)) = \max_{\boldsymbol{r} \in \Delta_d} \omega(\boldsymbol{r}) \,.$$

Moveover, by construction, we have $\nabla\Phi(\boldsymbol{x}(t)) = \boldsymbol{A}^\top \boldsymbol{r}(t)$. Hence, if we work in the regime where the update vectors are small enough (we impose this as a constraint to the builder),

$$\Phi(\boldsymbol{x}(t)) - \Phi(\boldsymbol{x}(t-1)) \approx \langle \nabla\Phi(\boldsymbol{x}(t)), \boldsymbol{A}\boldsymbol{\delta}(t) \rangle + \frac{1}{2} \left\langle \boldsymbol{A}\boldsymbol{\delta}(t), \nabla^2\Phi(\boldsymbol{x}(t))\boldsymbol{A}\boldsymbol{\delta}(t) \right\rangle \,.$$

So summing the first term on the right hand side over $t \leqslant T$ gives exactly the total payoff of the inspector. In conclusion,

$$\mathrm{Regret}(T) \leqslant \Phi(\boldsymbol{x}(0)) + \frac{1}{2} \sum_{t=1}^{T} \left\langle \boldsymbol{\delta}(t), \boldsymbol{A}^\top \nabla^2\Phi(\boldsymbol{x}(t))\boldsymbol{A}\boldsymbol{\delta}(t) \right\rangle \,.$$

Strategy for the builder. First, the builder picks the update vectors $\boldsymbol{\delta}(t)$ to make the payoff of the inspector negative. This implies that the regret (9.5) is an upper bound on the discrepancy objective (9.4). Then, the builder picks update vectors that makes the upper bound of the regret as small as possible. In this way, we recover Algorithm 3, and its analysis starts from the above regret bound.

We will come back to this interpretation in §11.2.

9.3. The regularization framework

In this section, we define formally our regularization framework.

9.3.1. An iterative meta-algorithm

We start by describing a generic iterative algorithm for discrepancy minimization that will serve as a basis for incorporating the potential functions based on regularization. Following §9.2, we will construct a sequence of partial colorings $\boldsymbol{x}(t) \in [-1, 1]^n$ for integer times $t = 0, 1, \ldots$. Each step consists in picking an update vector $\boldsymbol{\delta}$ and adding it to $\boldsymbol{x}(t)$. Whenever some coordinate of $\boldsymbol{x}(t)$ becomes ± 1, we say that the coordinate is *frozen*. We will also say of an unfrozen coordinate that it is *active*.

The oracle

Suppose that we are given some blackbox algorithm oracle that encapsulates all the possible choices of directions of the update vector. In the sequel, oracle$(\boldsymbol{A}, \boldsymbol{x})$ will correspond to a subset of vectors that do not increase too much the value of the regularized potential function when $\boldsymbol{x}$ is the current partial coloring.

Assumption 9.4. Let $C > 0$ be some universal constant. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ and a partial coloring $\boldsymbol{x} \in [-1, 1]^n$, oracle$(\boldsymbol{A}, \boldsymbol{x})$ satisfies (with $F \coloneqq \{j \in [n] : x_j \notin \{-1, 1\}\}$):

- If $|F| \geqslant C$, oracle$(\boldsymbol{A}, \boldsymbol{x})$ is a subset of $\mathbb{R}^F$ such that the intersection of oracle$(\boldsymbol{A}, \boldsymbol{x})$ with any halfspace of $\mathbb{R}^F$ contains a half-line.
- If $|F| < C$, it returns the value undefined.

The meta-algorithm

With a given oracle, the meta-algorithm for discrepancy minimization is Algorithm 4.

Algorithm 4 Generic iterative algorithm for discrepancy minimization

Input: $\boldsymbol{A} \in \mathbb{R}^{d \times n}$, $L \in (0, 1)$
Output: $\boldsymbol{x} \in \{\pm 1\}^n$ (a low-discrepancy coloring of $\boldsymbol{A}$)
Let $\boldsymbol{x}(0) \leftarrow (0, \ldots, 0)$ and $t \leftarrow 0$.
while oracle$(\boldsymbol{A}, \boldsymbol{x}(t))$ is not undefined **do**
    Choose any unit vector $\boldsymbol{\delta}$ in

$$\text{oracle}(\boldsymbol{A}, \boldsymbol{x}(t)) \cap \{\boldsymbol{\delta} \in \mathbb{R}^n : \langle \boldsymbol{\delta}, \boldsymbol{x}(t) \rangle = 0\}.$$

    Let

$$\varepsilon(t) \leftarrow \min\left\{\varepsilon > 0 : \exists j \in [n], x_j(t) \notin \{-1, 1\} \text{ and } x_j(t) + \varepsilon \delta_j \in \{-1, 1\}\right\}.$$

    Set

$$\boldsymbol{x}(t+1) \leftarrow \boldsymbol{x}(t) + \min(L, \varepsilon(t))\boldsymbol{\delta}.$$

    Update $t \leftarrow t + 1$.
end while
Let $T \leftarrow t$ and $x_j^* \leftarrow \text{sign}(x_j(T))$ for all $j \in [n]$.
return x^*.

The following three immediate observations on Algorithm 4 will be central to our framework:

Observation 9.5. For any $t = 0, \ldots, T-1$, $\|\boldsymbol{x}(t+1) - \boldsymbol{x}(t)\|_\infty \leqslant L$.

Observation 9.6. The final step on line 10 adds at most $C \max_{i\in[d],j\in[n]} |A_{ij}|$ to the discrepancy of the coloring, where C is the constant from Assumption 9.4.

Observation 9.7. There can be at most n/L^2 iterations of the main loop of Algorithm 4. Therefore, Algorithm 4 runs in polynomial time as long as oracle runs in polynomial time and $L \geqslant n^{-O(1)}$.

Proof. When $\varepsilon(t) \leqslant L$, at least one additional coordinate will reach ± 1 and will be frozen at the end of the iteration. This can happen at most n times. When $\varepsilon(t) > L$, since we pick our update vector orthogonal to $\boldsymbol{x}$, we have $\|\boldsymbol{x}(t+1)\|_2^2 = \|\boldsymbol{x}(t)\|_2^2 + L^2$. This can happen at most n/L^2 times. □

For our purposes, we will always set $L = n^{-O(1)}$ and computing oracle will only require elementary linear algebraic operations in $\mathbb{R}^n$ (intersection, orthogonal complements, direct sums, computation of eigenspaces, etc.).

9.3.2. Regularized maximum

Our main tool for building proxies for discrepancy is the following regularized version of the maximal entry of a vector.

Definition 9.8. For any convex function $\phi : \Delta_d \to \mathbb{R}$, we define $\phi^* : \mathbb{R}^d \to \mathbb{R}$ by

$$\phi^*(\boldsymbol{y}) \coloneqq \max_{\boldsymbol{r}\in\Delta_d} \langle \boldsymbol{r}, \boldsymbol{y}\rangle - \phi(\boldsymbol{r})\,.$$

We will call ϕ the *regularizer*. It maps elements of the simplex to some penalty in a convex way. By symmetry, it makes sense to focus on regularizers of the form $\phi(r) = \sum_{i\in[d]} \varphi(r_i)$ for some one-dimensional convex $\varphi\colon \mathbb{R} \to \mathbb{R}$. The following two special cases will play an important role in our theory.

Definition 9.9. For any $0 < q < 1$, the *ℓ_q-regularization* of the maximum, parametrized by $\eta > 0$, is the function $\omega^*_{q,\eta}\colon \mathbb{R}^d \to \mathbb{R}$ such that

$$\omega^*_{q,\eta}(\boldsymbol{y}) \coloneqq \max_{\boldsymbol{r}\in\Delta_d} \langle \boldsymbol{r}, \boldsymbol{y}\rangle + \frac{1}{\eta q}\sum_{i=1}^{d} r_i^q, \qquad \text{for any } y \in \mathbb{R}^m\,.$$

Definition 9.10. The *(negative) entropy regularization* of the maximum, parametrized by $\eta > 0$, is the function $\operatorname{smax}\colon \mathbb{R}^d \to \mathbb{R}$ such that

$$\operatorname{smax}_\eta(\boldsymbol{y}) \coloneqq \max_{\boldsymbol{r}\in\Delta_d} \langle \boldsymbol{r}, \boldsymbol{y}\rangle - \frac{1}{\eta}\sum_{i=1}^{d} r_i \log r_i, \qquad \text{for any } \boldsymbol{y} \in \mathbb{R}^m\,.$$

It is not hard to see that in this case, the solution of the maximization problem can be written in closed form: $\mathrm{smax}_\eta(y) = \frac{1}{\eta}\log\left(\sum_{i=1}^d \exp(\eta y_i)\right)$, thereby recovering the usual formulation of the softmax function.

9.3.3. Regularization bounds

We now present our two main technical lemmas that give an analytic justification for the ℓ_q and negative entropy regularization. The first one (Lemma 9.11) estimates the additive error incurred when tracking the regularized version of the maximum instead of the true maximum. For constant η and q, the approximation is worse for ℓ_q-regularization than for negative entropy regularization (polynomial vs logarithmic in the size of the vector).

Lemma 9.11. *Let $y \in \mathbb{R}^d$ and $q \in (0,1)$. If $M(\boldsymbol{y}) \coloneqq \max_{1\leqslant i\leqslant d} y_i$,*

$$M(\boldsymbol{y}) \leqslant \omega^*_{q,\eta}(\boldsymbol{y}) \leqslant M(\boldsymbol{y}) + \frac{d^{1-q}}{\eta q} \text{ and } M(\boldsymbol{y}) \leqslant \mathrm{smax}_\eta(\boldsymbol{y}) \leqslant M(\boldsymbol{y}) + \frac{\log d}{\eta}\,.$$

Proof. The lower bounds follow from picking $\boldsymbol{r}$ to be the Dirac mass function centered on the maximum coordinate. For the upper bounds, note that on the one hand, for all $\boldsymbol{r} \in \Delta_d$, $\langle \boldsymbol{r}, \boldsymbol{y}\rangle \leqslant M(\boldsymbol{y})$, and on the other hand, $\sum_i r_i^q \leqslant d^{1-q}$ (resp. $-\sum_i r_i \log r_i \leqslant \log d$) by Jensen's inequality. □

The second one (Lemma 9.12) bounds the first two terms in the Taylor expansion of the potential function. In the sequel, this will allow us to control the increase in ℓ_∞-norm when making a small update in our iterative algorithm. As we demonstrated in §9.2, what matters in this expansion is the second-order term. Indeed, in applications to discrepancy, we will always trivially upper bound the first-order term by simply picking an update that is positively correlated with the gradient (which will be an easy additional condition to impose).

Lemma 9.12. *Fix $\boldsymbol{y} \in \mathbb{R}^d$ and $q \in (0,1)$. Let $\boldsymbol{\nabla} \coloneqq \nabla\omega^*_{q,\eta}(\boldsymbol{y})$. Then $\boldsymbol{\nabla} \in \Delta_d$ and for all $\boldsymbol{\delta} \in \mathbb{R}^d$ with $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1-q}{8\eta}$,*

$$\omega^*_{q,\eta}(\boldsymbol{y}+\boldsymbol{\delta}) \leqslant \omega^*_{q,\eta}(\boldsymbol{y}) + \langle \boldsymbol{\nabla}, \boldsymbol{\delta}\rangle + \frac{\eta}{1-q}\sum_{i=1}^d \nabla_i^{2-q}\delta_i^2\,.$$

Similarly, if $\boldsymbol{\nabla} \coloneqq \nabla\mathrm{smax}_\eta(y)$, then $\boldsymbol{\nabla} \in \Delta_d$ and for all $\boldsymbol{\delta} \in \mathbb{R}^d$ with $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1}{3\eta}$,

$$\mathrm{smax}_\eta(\boldsymbol{y}+\boldsymbol{\delta}) \leqslant \mathrm{smax}_\eta(\boldsymbol{y}) + \langle \boldsymbol{\nabla}, \boldsymbol{\delta}\rangle + \eta\sum_{i=1}^d \nabla_i\delta_i^2\,.$$

Proof. Consider first the ℓ_q-regularizer with $\eta = 1$. To lighten notation, we write ω^* for $\omega^*_{q,\eta}$. Recall that

$$\omega^*(\boldsymbol{y}) = \max_{\boldsymbol{r} \in \Delta_d} \langle \boldsymbol{r}, \boldsymbol{y} \rangle + \frac{1}{q} \sum_{i=1}^{d} r_i^q \,. \tag{9.6}$$

By Danskin's theorem (see e.g. [Ber99, Proposition B.25]), we have $\nabla \omega^*(\boldsymbol{y}) = \boldsymbol{r}^* \in \Delta_d$, where $\boldsymbol{r}^*$ is the optimum in (9.6). For the KKT conditions to hold, we must have for some $\lambda \colon \mathbb{R}^d \to \mathbb{R}$ (the Lagrange multiplier associated to the equality constraint of the simplex):

$$y_i + (r_i^*)^{q-1} = \lambda(\boldsymbol{y}) \quad \text{for all } i \in [d] \,.$$

The Lagrange multipliers associated to the inequality constraints disappear by complementary slackness since necessarily $r_i^* \neq 0$. Also we must have $\lambda(\boldsymbol{y}) > \max_{i \in [d]} y_i$ by the previous equality. In fact, $\lambda(\boldsymbol{y})$ is the unique solution to

$$\sum_{i \in [d]} (\lambda(\boldsymbol{y}) - y_i)^{1/(q-1)} = 1 \,.$$

In summary, $\nabla \omega^*(\boldsymbol{y}) = (\lambda(\boldsymbol{y}) \mathbf{1} - \boldsymbol{y})^{\odot \frac{1}{q-1}} \in \Delta_d$. Differentiating once more, we see that

$$\nabla^2 \omega^*(\boldsymbol{y}) = \frac{1}{1-q} \left(\operatorname{diag}(\nabla \omega^*(\boldsymbol{y})^{\odot 2-q}) - (\nabla \lambda(\boldsymbol{y}))(\nabla \omega^*(\boldsymbol{y})^{\odot 2-q})^\top \right) .$$

Let $\boldsymbol{M} \coloneqq (\nabla \lambda(\boldsymbol{y}))(\nabla \omega^*(\boldsymbol{y})^{\odot 2-q})^\top$. Observe that $\boldsymbol{M}$ has rank 1 and must be symmetric as the Hessian itself is symmetric. Further, $\lambda(\boldsymbol{y})$ is a nondecreasing function of y_i for all $i \in [d]$, so that every entry of $\boldsymbol{M}$ is nonnegative. It follows that $\boldsymbol{M}$ is positive semidefinite, and thus

$$\nabla^2 \omega^*(\boldsymbol{y}) \preccurlyeq \frac{1}{1-q} \operatorname{diag}(\nabla \omega^*(\boldsymbol{y})^{\odot 2-q}) \,. \tag{9.7}$$

Now fix $\boldsymbol{\delta} \in \mathbb{R}^d$. The function $s \mapsto \sum_i (s - y_i)^{\frac{1}{q-1}}$ defined for $s > \max_i y_i$ is nonincreasing, so for all i,

$$|\lambda(\boldsymbol{y} + \boldsymbol{\delta}) - \lambda(\boldsymbol{y})| \leqslant \|\boldsymbol{\delta}\|_\infty \quad \text{and} \quad \lambda(\boldsymbol{y}) \geqslant 1 + y_i \,. \tag{9.8}$$

Now fix $i \in [d]$ and suppose that $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1-q}{8}$. We write

$$\begin{aligned} (\nabla \omega^*(\boldsymbol{y} + \boldsymbol{\delta}))_i^{2-q} &= (\nabla \omega^*(\boldsymbol{y}))_i^{2-q} \left(1 + \frac{\lambda(\boldsymbol{y} + \boldsymbol{\delta}) - \lambda(\boldsymbol{y}) - \delta_i}{\lambda(\boldsymbol{y}) - y_i} \right)^{\frac{2-q}{q-1}} \\ &\leqslant (\nabla \omega^*(\boldsymbol{y}))_i^{2-q} \exp\left(\frac{2-q}{1-q} \cdot \frac{\lambda(\boldsymbol{y}) + \delta_i - \lambda(\boldsymbol{y} + \boldsymbol{\delta})}{\lambda(\boldsymbol{y} + \boldsymbol{\delta}) - y_i - \delta_i} \right) , \end{aligned}$$

where we used the inequality $\log(1+x) \geqslant \frac{x}{1+x}$. Now plug in the bounds from (9.8):

$$\begin{aligned}(\nabla\omega^*(\boldsymbol{y}+\boldsymbol{\delta}))_i^{2-q} &\leqslant (\nabla\omega^*(\boldsymbol{y}))_i^{2-q} \exp\left(\frac{2-q}{1-q}\cdot\frac{\lambda(\boldsymbol{y})+\delta_i-\lambda(\boldsymbol{y}+\boldsymbol{\delta})}{1+\lambda(\boldsymbol{y}+\boldsymbol{\delta})-\lambda(\boldsymbol{y})-\delta_i}\right)\\ &\leqslant (\nabla\omega^*(\boldsymbol{y}))_i^{2-q} \exp\left(\frac{2}{1-q}\cdot\frac{2\left\|\boldsymbol{\delta}\right\|_\infty}{1-2\left\|\boldsymbol{\delta}\right\|_\infty}\right)\\ &\leqslant 2(\nabla\omega^*(\boldsymbol{y}))_i^{2-q}\,. \end{aligned} \tag{9.9}$$

Finally, from Taylor's inequality, under the same assumption $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1-q}{8}$,

$$\begin{aligned}|\omega^*(\boldsymbol{y}+\boldsymbol{\delta}) - \omega^*(\boldsymbol{y}) - \langle\nabla\omega^*(\boldsymbol{y}),\boldsymbol{\delta}\rangle| &\leqslant \frac{1}{2}\sup_{u\in[0,1]}\left|\boldsymbol{\delta}^\top\nabla^2\omega^*(\boldsymbol{y}+u\boldsymbol{\delta})\boldsymbol{\delta}\right|\\ &\leqslant \frac{1}{1-q}\sum_{i=1}^d(\nabla\omega^*(\boldsymbol{y}))_i^{2-q}\delta_i^2\,,\end{aligned}$$

where the last inequality follows from (9.7) and (9.9).

For the entropy regularizer and $\eta = 1$, it holds that

$$\nabla\mathrm{smax}(\boldsymbol{y}) = \frac{\exp(\boldsymbol{y})}{\sum_{i=1}^d\exp(y_i)} \quad\text{and}\quad \nabla^2\mathrm{smax}(\boldsymbol{y}) \preccurlyeq \mathrm{diag}(\nabla\mathrm{smax}(\boldsymbol{y}))\,.$$

Therefore for all $i \in [d]$,

$$(\nabla\mathrm{smax}(\boldsymbol{y}+\boldsymbol{\delta}))_i \leqslant (\nabla\mathrm{smax}(\boldsymbol{y}))_i\exp(2\left\|\boldsymbol{\delta}\right\|_\infty)\,,$$

and we conclude in the same way as for the ℓ_q-regularizers.

For general η, observe that $\nabla\omega^*_{q,\eta}(\boldsymbol{y}) = \nabla\omega^*_{q,1}(\eta\boldsymbol{y})$ and $\nabla^2\omega^*_{q,\eta}(\boldsymbol{y}) = \eta\nabla^2\omega^*_{q,1}(\eta\boldsymbol{y})$ (and similarly for smax_η). Therefore, the same argument based on Taylor's inequality gives the desired result as long as $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1-q}{8\eta}$ for $\omega^*_{q,\eta}$ and $\|\boldsymbol{\delta}\|_\infty \leqslant \frac{1}{3\eta}$ for smax_η. □

Remark 9.13. This proof gives an analytic explanation for why we might prefer ℓ_q-regularization to negative entropy regularization in certain situations, although the approximation error from Lemma 9.11 is worse (for the same value of η). Observe that a typical entry ∇_i of the gradient is much smaller that 1. Hence, ℓ_q-regularization can be advantageous whenever we can leverage the fact that ∇_i^{2-q} is typically much smaller than ∇_i. This type of tradeoff has been applied both in the bandit literature [AB10] and for graph sparsification [ALO15]. As we will next see, this is also the case in Spencer's setting.

9.4. Full proof of Spencer's theorem

We now give a complete proof of Spencer's theorem in the general case where the matrix has d rows and n columns. Our choice of $q \in (0,1)$ in the ℓ_q-regularization is going to depend on the ratio d/n.

Theorem 9.14. *Let $n \leqslant d$. There is a deterministic algorithm running in polynomial time that for each $\boldsymbol{A} \in [-1,1]^{d\times n}$, finds $\boldsymbol{x} \in \{\pm 1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty \lesssim \left(\sqrt{n \log\left(\frac{2d}{n}\right)}\right).$$

We start by proving the following lemma, which will allow us to find an update vector that does not increase too much the ℓ_q-regularization of the maximal coordinate when there are k active coordinates remaining.

Lemma 9.15. *Let $4 \leqslant k \leqslant 2d-2$, $\boldsymbol{u}_1, \dots, \boldsymbol{u}_d$ be unit vectors in $\mathbb{R}^k$ and $\boldsymbol{\nabla} \in \Delta_d$. Consider*

$$\boldsymbol{M} := \sum_{i=1}^{d} \nabla_i^{2-q} \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

There is a subspace S of dimension at least 2 such that for all $\boldsymbol{v} \in S$,

$$\langle \boldsymbol{v}, \boldsymbol{M}\boldsymbol{v}\rangle \leqslant 8k^{q-2}\|\boldsymbol{v}\|_2^2.$$

Moreover, this subspace can be found in polynomial time.

Proof. Without loss of generality, suppose that $\nabla_1 \geqslant \ldots \geqslant \nabla_d$. Let

$$S_1 := \left\{\boldsymbol{v} \in \mathbb{R}^k : \langle \boldsymbol{v}, \boldsymbol{u}_i\rangle = 0 \text{ for all } i = 1, \dots, \left\lceil \frac{k}{2}\right\rceil - 1\right\}.$$

Observe that $\nabla_{\lceil \frac{k}{2}\rceil} \leqslant \frac{2}{k}$, so for all $\boldsymbol{v} \in S_1$,

$$\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v} \leqslant \left(\frac{2}{k}\right)^{1-q} \sum_{i=1}^{d} \nabla_i \langle \boldsymbol{u}_i, \boldsymbol{v}\rangle^2. \tag{9.10}$$

Let $\boldsymbol{R} := \sum_{i\leqslant d} \nabla_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ and consider an orthonormal basis $\boldsymbol{w}_1, \dots, \boldsymbol{w}_\ell$ of S_1 such that $\boldsymbol{w}_1^\top \boldsymbol{R}\boldsymbol{w}_1 \leqslant \ldots \leqslant \boldsymbol{w}_\ell^\top \boldsymbol{R}\boldsymbol{w}_\ell$. Select S to be the span of $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$. Observe that

$$\sum_{j=1}^{\ell} \boldsymbol{w}_j^\top \boldsymbol{R}\boldsymbol{w}_j = \sum_{i=1}^{d} \nabla_i \sum_{j=1}^{\ell} \langle \boldsymbol{w}_j, \boldsymbol{u}_i\rangle^2 \leqslant 1.$$

Thus, by an averaging argument, it holds that

$$\boldsymbol{v}^\top \boldsymbol{R}\boldsymbol{v} \leqslant \frac{1}{\ell - 1}\|\boldsymbol{v}\|_2^2 \leqslant \frac{2}{k-2}\|\boldsymbol{v}\|_2^2$$

for all $\boldsymbol{v} \in S$. We conclude by combining with (9.10) and using the assumption on k. □

Proof of Theorem 9.14. Up to doubling the number of rows, we apply Remark 9.2 to reduce to the case where $\boldsymbol{A}$ satisfies for all $\boldsymbol{x} \in \{-1,1\}^n$, $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{i\in[d]} (\boldsymbol{A}\boldsymbol{x})_i$.

We set the parameter L of Algorithm 4 to be $L := \frac{1-q}{8\eta n}$, where q and η are the parameters of the ℓ_q-regularizer to be fixed later.

We now describe our construction of oracle$(\boldsymbol{A}, \boldsymbol{x}(t))$ with $F(t)$ being the set of active coordinates of $\boldsymbol{x}(t)$ and $k = k(t) := |F(t)|$. To simplify notations, we write $\boldsymbol{x} = \boldsymbol{x}(t)$ and $F = F(t)$. Observe that $\|\boldsymbol{A}\boldsymbol{\delta}\|_\infty \leqslant n\|\boldsymbol{\delta}\|_\infty$. Hence, by Lemma 9.12, there exists $\boldsymbol{\nabla} \in \Delta_n$ such that for all update $\boldsymbol{\delta} \in \mathbb{R}^d$ with $\|\boldsymbol{\delta}\|_\infty \leqslant L$,

$$\omega^*_{q,\eta}(\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{\delta})) - \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}) \leqslant \langle \boldsymbol{\nabla}, \boldsymbol{A}\boldsymbol{\delta}\rangle + \frac{\eta}{1-q}\sum_{i=1}^{d} \nabla_i^{2-q}\langle \boldsymbol{A}_i, \boldsymbol{\delta}\rangle^2 .$$

By assumption, $\sum_{j\in F} A_{i,j}^2 \leqslant k$, so we can apply Lemma 9.15 to get a 2-dimensional subspace S such that for all $\boldsymbol{\delta} \in S$,

$$\sum_{i=1}^{n} \nabla_i^{2-q}\langle \boldsymbol{A}_i, \boldsymbol{\delta}\rangle^2 \lesssim \frac{\|\boldsymbol{\delta}\|_2^2}{k^{1-q}} . \tag{9.11}$$

The second-order term is invariant if we change $\boldsymbol{\delta}$ to $-\boldsymbol{\delta}$, but the first-order term changes sign. We return from oracle$(\boldsymbol{A}, \boldsymbol{x}(t))$ the subspace S intersected with the halfspace $\{\boldsymbol{\delta} \in \mathbb{R}^n : \langle \boldsymbol{A}^\top\boldsymbol{\nabla}, \boldsymbol{\delta}\rangle \leqslant 0\}$.

Now we switch to the global analysis of Algorithm 4 and estimate what is the total discrepancy incurred over the whole walk. Assumption 9.4 is here satisfied for $C = 3$, so since the entries of $\boldsymbol{A}$ are bounded, the last step of Algorithm 4 only changes the discrepancy of the final coloring by an additive constant.

Denote by β_k the sum of the ℓ_2-squared norm of the update vectors starting from the point where there are at most k unfrozen coordinates remaining. Recall that we always choose our update vector orthogonal to the current position, so that $\beta_k \leqslant k$. We now sum by parts the main term of the increases (9.11) over the execution of the algorithm,

$$\sum_{k=4}^{n} \frac{\beta_k - \beta_{k-1}}{k^{1-q}} = \frac{\beta_n}{n^{1-q}} + \sum_{k=4}^{n-1} \beta_k \left(\frac{1}{k^{1-q}} - \frac{1}{(k+1)^{1-q}}\right) \leqslant n^q + \sum_{k=4}^{n-1} \frac{1}{k^{1-q}} \leqslant \frac{2n^q}{q} . \tag{9.12}$$

Thus, by (9.12) and Lemma 9.11, the final coloring $\boldsymbol{x}(T)$ satisfies

$$\|\boldsymbol{A}\boldsymbol{x}(T)\|_\infty \leqslant \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(T)) \lesssim \frac{d^{1-q}}{\eta q} + \frac{\eta}{q(1-q)} n^q .$$

The result follows by setting

$$\eta = \sqrt{\frac{(1-q)d^{1-q}}{n^q}} \quad \text{and} \quad q = 1 - \frac{1}{\log_2(\frac{2d}{n})} . \qquad \square$$

At this point, it is worth looking at what happens in this proof if we replace the ℓ_q-regularizer with the entropic regularizer. For simplicity, consider the case where $d = O(n)$. While the constant cost is only $\log n/\eta$, we are not able to win anything in the local update as in Lemma 9.15 and we would get an ηn loss in the potential during the walk. Optimizing over η would give discrepancy $\sqrt{n \log n}$. Negative entropy regularization in this context corresponds merely to a derandomization of the Chernoff and union bound argument.

In fact, one could repeat the same analysis by replacing the regularizer by a general function of the form $\phi(r) = \sum_{i \leqslant d} \varphi(r_i)$ for some convex, non-positive function $\varphi : \mathbb{R} \to \mathbb{R}$. Under additional conditions on φ (for example the fact that $x \mapsto x\varphi''(x)$ is non-increasing) one would obtain a discrepancy of

$$O\left(\sqrt{-n\varphi\left(\frac{1}{n}\right)\sum_{k \leqslant n} \frac{k}{\varphi''\left(\frac{1}{k}\right)}}\right). \tag{9.13}$$

With this bound established, we can quickly verify that setting φ to be the negative entropy, we obtain $\varphi(1/n) = -\log n/n$ and $\varphi''(1/k) = k$, which immediately recovers a discrepancy of $O(\sqrt{n \log n})$.

Given this expression in (9.13), it appears that we can derive the best possible regularizer by solving a differential equation. Since there is no silver bullet for such problems, one can simply test various elementary functions. Setting $\varphi(x) = -x^q$ for $0 < q < 1$ we verify the required condition and obtain $\varphi(1/n) = -1/n^q$, and $\varphi''(1/k) = q(1-q)/k^{2-q}$, which removes the logarithmic factor for constant q.

Remark 9.16 (Spherical discrepancy)**.** A slight variation of the same algorithm, which does not freeze variables, automatically achieves optimal bounds for *spherical discrepancy*. This setting is a relaxation of the Komlós problem (see Chapter 10), where the columns of the input matrix are vectors with at most unit ℓ_2-norm, but the sought coloring only has an ℓ_2-norm constraint i.e. $\|\boldsymbol{x}\|_2 = \sqrt{n}$, rather than $\boldsymbol{x} \in \{-1, 1\}^n$ [JM20]. The key difference between this setting and that of Komlós is that we are not forced to lose degrees of freedom by freezing variables, so throughout the entire execution of the algorithm we have $\Theta(n)$ degrees of freedom to update the partial coloring.

To show this, we simply observe that at all times there is an update that does not increase the discrepancy of rows with large *global* ℓ_2 norm, which represent only at most a constant fraction of the entire set of rows, by Markov's inequality. The rate of increase in discrepancy entirely depends on the ℓ_2-norm of the rows of the underlying matrix (restricted to the unfrozen variables, which in this case are all the variables). Following through with the same argument we used for Spencer, we obtain discrepancy $O(1)$.

9.5. Matrix discrepancy

In this section, we extend our discrepancy minimization framework to the matrix setting. Specifically, we consider objectives of the form

$$f(x) := \left\| \sum_{i=1}^{n} x_i A_i \right\|,$$

where $A_1, \dots, A_n \in \mathbb{R}^{d \times d}$ are symmetric matrices and $\|\cdot\|$ is a matrix norm. We focus on the spectral norm, which generalizes the ℓ_∞-norm for vectors. Our regularization-based approach extends naturally to this setting.

We study a matrix generalization of Spencer's theorem, in which the A_i are arbitrary symmetric matrices with bounded spectral norm. When the A_i are diagonal, we recover the classical Spencer setting. The *matrix Spencer conjecture* posits that an $O(\sqrt{n})$ discrepancy bound continues to hold even when the matrices do not commute.

Another key application of matrix discrepancy bounds is the construction of graph sparsifiers, as previewed in §1.1.3. In this case, the matrices A_i are normalized Laplacian matrices of individual edges. The construction of [BSS14] has been recovered in [LWZ25] using our framework.

9.5.1. Background on matrix discrepancy

We denote by $\|\cdot\|_2$ the spectral norm of a matrix, i.e.,

$$\|A\|_2 := \max_{\|x\|_2 = \|y\|_2 = 1} \langle x, Ay \rangle \ .$$

Conjecture 9.17 (Matrix Spencer conjecture [Zou12])**.** *Let $A_1, \dots, A_n \in \mathbb{R}^{n \times n}$ be symmetric matrices such that $\|A_i\|_2 \leqslant 1$ for all $i \in [n]$. Then there exists $x_1, \dots, x_n \in \{-1, 1\}$ such that*

$$\left\| \sum_{i=1}^{n} x_i A_i \right\|_2 = O(\sqrt{n}) \ . \tag{9.14}$$

One obvious difference between the matrix Spencer conjecture and Spencer's theorem is that the left-hand side of (9.14) corresponds to an infinite (or, after discretization, exponentially large) number of constraints. When the matrices $A_1, \dots, A_n$ have a common eigenbasis, then the problem effectively reduces to bounding only n quadratic forms (one for each of the common basis elements), and so we recover Spencer's theorem as a special case of this conjecture.

There are other ways in which matrix discrepancy behaves differently to the vector discrepancy, as the following example shows. A similar construction independently appeared in [DJR22, §6.2].

Example 9.18 (The "matrix Komlós conjecture" is false.)**.** Let G be the star graph on $n+1$ vertices and let $\boldsymbol{A}_i = (\boldsymbol{e}_i - \boldsymbol{e}_{n+1})(\boldsymbol{e}_i - \boldsymbol{e}_{n+1})^\top$ be the Laplacian matrices of the n edges of G (here $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{n+1})$ denotes the canonical basis of $\mathbb{R}^{n+1}$). Note that the $\boldsymbol{A}_i$'s satisfy $\|\boldsymbol{A}_i\|_{\mathrm{F}} = 4$. By analogy with the Komlós conjecture, one might expect that the bound on the *Frobenius norm* implies the existence of much better colorings of the $\boldsymbol{A}_i$'s than those guaranteed by Conjecture 9.17. However, this intuition is false, and in fact, Conjecture 9.17 is tight on this example.

This is because for any $\boldsymbol{x} \in \{-1,1\}^n$, the first n coordinates of $\sum_i x_i \boldsymbol{A}_i \boldsymbol{e}_{n+1}$ are $x_1, \dots, x_n$, which certifies that $\left\|\sum_{i=1}^n x_i \boldsymbol{A}_i\right\|_2 \geqslant \|\boldsymbol{x}\|_2 = \sqrt{n}$.

Random coloring

Picking $x_1, \dots, x_n$ uniformly at random comes within a $\sqrt{\log}$ factor away from the conjecture:

Proposition 9.19. *Let $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{R}^{n \times n}$ be symmetric matrices such that $\|\boldsymbol{A}_i\|_2 \leqslant 1$ and $\operatorname{rank}(\boldsymbol{A}_i) \leqslant r$. Sample a random coloring $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \{-1,1\}$. Then,*

$$\mathbb{E}\left\|\sum_{i=1}^n x_i \boldsymbol{A}_i\right\|_2 = O(\sqrt{n \log r})\,.$$

Proof. We give a self-contained proof of this fact using the trace method. For any integer p,

$$\begin{aligned}\left(\mathbb{E}\left\|\sum_{i=1}^n x_i \boldsymbol{A}_i\right\|_2\right)^{2p} &\leqslant \mathbb{E}\operatorname{tr}\left(\sum_{i=1}^n x_i \boldsymbol{A}_i\right)^{2p} \\ &= \sum_{i:[2p]\to[n]} \mathbb{E}\left[x_{i(1)} \dots x_{i(2p)}\right] \operatorname{tr}\left(\boldsymbol{A}_{i(1)} \dots \boldsymbol{A}_{i(2p)}\right)\,.\end{aligned}$$

For a $i : [2p] \to [n]$ to contribute a non-zero value to the sum on the right-hand side, every distinct element among $i(1), \dots, i(2p)$ have to occur an even number of times. There are at most $n^p \cdot (2p-1)!! \leqslant (2np)^p$ such maps. Moreover, by Hölder inequality and the assumptions $\|\boldsymbol{A}_i\|_2 \leqslant 1$ and $\operatorname{rank}(\boldsymbol{A}_i) \leqslant r$,

$$\operatorname{tr}\left(\boldsymbol{A}_{i(1)} \dots \boldsymbol{A}_{i(2p)}\right) \leqslant \|\boldsymbol{A}_{i(1)}\|_{2p} \dots \|\boldsymbol{A}_{i(2p)}\|_{2p} \leqslant r\,.$$

Putting everything together, we get

$$\mathbb{E}\left\|\sum_{i=1}^n x_i \boldsymbol{A}_i\right\|_2 \leqslant \sqrt{2np} \cdot r^{\frac{1}{2p}}\,,$$

which is $O(\sqrt{n \log r})$ when setting $p = \Theta(\log r)$, as desired. □

The Kadison–Singer problem

We close this background by mentioning another setting in matrix discrepancy, introduced in Weaver's work on discrepancy-theoretic versions of the Kadison–Singer problem [Wea04].

Marcus, Spielman, and Srivastava [MSS15] famously showed that there exist universal constants $K \geqslant 2$ and $\varepsilon > 0$ such that for any vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n \in \mathbb{R}^d$ such that $\|\boldsymbol{u}_i\|_2 \leqslant 1$ and $\left\|\sum_i \boldsymbol{u}_i \boldsymbol{u}_i^\top\right\|_2 \leqslant K$, there exists a coloring $\boldsymbol{x} \in \{-1, 1\}^n$ such that $\left\|\sum_i x_i \boldsymbol{u}_i \boldsymbol{u}_i^\top\right\|_2 \leqslant K - \varepsilon$.

However, designing an efficient algorithm for finding such a coloring remains an outstanding open question. Through the discrepancy–sparsification connection, this result is related to the existence of *unweighted* graph sparsifiers.

9.5.2. Regularization for matrix discrepancy

We generalize in the natural way the regularization framework from the ℓ_∞-norm of a vector to the spectral norm of a matrix. The ℓ_q-regularized spectral norm first appeared in [BSS14] in its barrier (dual) form for $q = \frac{1}{2}$, and in [ALO15] in the following primal form:

Definition 9.20. For any symmetric matrix $\boldsymbol{X}$, let

$$\omega_q^*(\boldsymbol{X}) := \max_{\boldsymbol{R} \geq 0, \operatorname{tr} \boldsymbol{R} = 1} \langle \boldsymbol{R}, \boldsymbol{X} \rangle + \frac{1}{q} \operatorname{tr} \boldsymbol{R}^q \,.$$

Remark 9.21 (Dual barrier form)**.** Similarly to the vector case, we can compute the dual of the optimization problem as

$$\begin{aligned} \omega_{\frac{1}{2}}^*(\boldsymbol{X}) &= \max_{\boldsymbol{R} \geq 0} \min_{\lambda \in \mathbb{R}} \langle \boldsymbol{R}, \boldsymbol{X} \rangle + 2 \operatorname{tr} \boldsymbol{R}^{1/2} + \lambda(1 - \operatorname{tr} \boldsymbol{R}) \\ &= \min_{\lambda \in \mathbb{R}} \lambda + \max_{\boldsymbol{R} \geq 0} \langle \boldsymbol{R}, \boldsymbol{X} - \lambda \boldsymbol{I} \rangle + 2 \operatorname{tr} \boldsymbol{R}^{1/2} \\ &= \min_{\lambda > \lambda_{\max}(\boldsymbol{X})} \lambda + \operatorname{tr}((\lambda \boldsymbol{I} - \boldsymbol{X})^{-1}) \,. \end{aligned}$$

So, in comparison to the barrier approach in [BSS14, BLV22], where a parameter λ has to be carefully tracked during the analysis, here the barrier is auto-adjusting by balancing the choice of λ with the additive λ term.

The dual derivation reveals in particular:

Lemma 9.22.

$$\nabla \omega_q^*(\boldsymbol{X}) = (\lambda \boldsymbol{I} - \boldsymbol{X})^{\frac{1}{q-1}} \,,$$

where $\lambda > \lambda_{\max}(\boldsymbol{X})$ *is the unique solution to* $\operatorname{tr}((\lambda \boldsymbol{I} - \boldsymbol{X})^{\frac{1}{q-1}}) = 1$.

Next, we extend the regularization bounds from §9.3.3 to the matrix case.

Lemma 9.23. *Let X be a symmetric $d \times d$ matrix. Then,*

$$\lambda_{\max}(X) \leqslant \omega_q^*(X) \leqslant \lambda_{\max}(X) + \frac{d^{1-q}}{q}.$$

Proof. Identical to the vector case. □

Lemma 9.24. *There exists a universal constant $c > 0$ such that for any $\varepsilon > 0$, the following holds. Suppose that $q = 1 - \frac{1}{\gamma}$ for some integer γ. Let X and Δ be a symmetric $d \times d$ matrix such that $\|\Delta\|_2 \leqslant c\varepsilon$. Denote $G(X) := \nabla\omega_q^*(X)$. Then,*

$$\omega_q^*(X+\Delta) \leqslant \omega_q^*(X) + \langle G(X), \Delta\rangle + \frac{1+\varepsilon}{2}\sum_{k=1}^{\gamma} \operatorname{tr}\left(G(X)^{\frac{k}{\gamma}}\Delta G(X)^{\frac{\gamma-k+1}{\gamma}}\Delta\right).$$

Proof. Denoting by ∂ the derivative in direction Δ with respect to X, using the rule for matrix differentiation of $Z \mapsto Z^{\frac{1}{q-1}} = Z^{-\gamma}$,

$$\partial G(X) = -\sum_{k=1}^{\gamma} G(X)^{\frac{k}{\gamma}}(\partial\lambda(X)\cdot I - \Delta)G(X)^{\frac{\gamma-k+1}{\gamma}}.$$

Moreover, by the constraint on $\lambda(X)$ that enforces that $G(X)$ stays in the simplex, we have $0 = \partial \operatorname{tr} G(X) = \operatorname{tr}(\partial G(X))$, so that

$$\partial\lambda(X) = \frac{\operatorname{tr}\left(G(X)^{2-q}\Delta\right)}{\operatorname{tr}\left(G(X)^{2-q}\right)}. \tag{9.15}$$

Hence, the quadratic form of the Hessian of ω^* in direction Δ is

$$\begin{aligned}\operatorname{tr}\left(\partial G(X)\cdot\Delta\right) &= \sum_{k=1}^{\gamma}\operatorname{tr}\left(G(X)^{\frac{k}{\gamma}}\Delta G(X)^{\frac{\gamma-k+1}{\gamma}}\Delta\right) - \frac{\gamma\operatorname{tr}\left(G(X)^{2-q}\Delta\right)^2}{\operatorname{tr} G(X)^{2-q}}\\ &\leqslant \sum_{k=1}^{\gamma}\operatorname{tr}\left(G(X)^{\frac{k}{\gamma}}\Delta G(X)^{\frac{\gamma-k+1}{\gamma}}\Delta\right).\end{aligned}$$

To make sure that there is no contribution of higher-order terms in the Taylor expansion, it remains to bound:

Lemma 9.25.

$$\sup_{u\in[0,1]}\operatorname{tr}\left(G(X+u\Delta)^{\frac{k}{\gamma}}\Delta G(X+u\Delta)^{\frac{\gamma-k+1}{\gamma}}\Delta\right) \leqslant (1+\varepsilon)\operatorname{tr}\left(G(X)^{\frac{k}{\gamma}}\Delta G(X)^{\frac{\gamma-k+1}{\gamma}}\Delta\right)$$

Lemma 9.24 follows from Lemma 9.25 using Taylor inequality. □

Proof of Lemma 9.25. First applying to Cauchy-Schwarz to (9.15), we have $|\partial\lambda(X+u\Delta)| \leqslant \|\Delta\|_2$, so λ changes by at most $\|\Delta\|_2$ on the line from X to $X+\Delta$. From there, we get that $G^{-1/\gamma}$ has Lipschitz constant (for the spectral norm) $2\|\Delta\|_2$ on that interval. Whenever this is a small enough constant, we can expand

$$G(X+u\Delta)^{\frac{k}{\gamma}} = G(X)^{\frac{k}{\gamma}}\left(I + \left(G(X+u\Delta)^{-\frac{1}{\gamma}} - G(X)^{-\frac{1}{\gamma}}\right)G(X)^{\frac{1}{\gamma}}\right)^{-k}.$$

Under the assumption on $\|\Delta\|_2$ being a small enough multiple of ε, we get $G(X+u\Delta)^{\frac{k}{\gamma}} \preceq G(X)^{\frac{k}{\gamma}}(1+\frac{\varepsilon}{2})$ and similarly $G(X+u\Delta)^{\frac{\gamma-k+1}{\gamma}} \preceq G(X)^{\frac{\gamma-k+1}{\gamma}}(1+\frac{\varepsilon}{2})$. Finally, the result folllows by a successive application of both PSD inequalities to the trace term. □

9.5.3. The matrix Spencer problem

A series of recent works on low-rank instances of the matrix Spencer problem [HRS22, DJR22] culminated in a proof [BJM23] that Conjecture 9.17 holds under the additional assumption $\operatorname{rank}(A_i) \leqslant n/\operatorname{polylog}(n)$ for all $i \in [n]$. This proof applies a deep result from random matrix theory [BBvH23]. In this section, we highlight connections between that approach and our regularization framework, which may enable a similar implementation.

The strategy in [BJM23]

[BJM23] applies as a black-box the following result to find a good partial coloring:

Theorem 9.26 (Refined Non-Commutative Khintchine [BBvH23]). *Let $A_1, \dots, A_n$ be symmetric $d \times d$ matrices and $g \sim \mathcal{N}(0, I_n)$. Then,*

$$\mathbb{E}\left\|\sum_{i=1}^n g_i A_i\right\|_2 \lesssim \left\|\sum_{i=1}^n A_i^2\right\|_2^{\frac{1}{2}} + \operatorname{polylog}(d)\cdot\left\|\operatorname{Cov}\left(\sum_{i=1}^n g_i A_i\right)\right\|_2^{\frac{1}{2}}.$$

A weaker interpretation of Theorem 9.26 yields an improved bound on the discrepancy for a random ± 1 coloring when the matrices $A_1, \dots, A_n$ have small covariance parameter. The key observation in [BJM23] is that when the A_i have low rank, the covariance matrix of $\sum_i g_i A_i$ has small spectral norm after projecting out a few adversarial directions.

It is plausible that matching the guarantee of Theorem 9.26 for a random coloring within our regularization framework, combined with orthogonality constraints during the sticky walk, could recover the result from [BJM23].

The proof of Theorem 9.26

The proof of Theorem 9.26 constructs an interpolation path between $\sum_i g_i \boldsymbol{A}_i$ and an idealized model whose spectral norm matches the standard deviation proxy from matrix Chernoff bounds — crucially, without incurring an extra $\sqrt{\log d}$ factor.

We now sketch how one might try to analyze the performance of a random coloring matching Theorem 9.26. The ℓ_q-potential that matches matrix Chernoff is $q = 1 - \frac{1}{\log d}$. Consider the potential

$$\Phi(\boldsymbol{x}) = \omega^* \left(\sum_{i=1}^n x_i \boldsymbol{A}_i \right).$$

Then by Lemma 9.24, the second-order contribution from perturbing $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{\delta}$ is

$$\left\langle \boldsymbol{\delta}, \nabla^2 \Phi(\boldsymbol{x}) \boldsymbol{\delta} \right\rangle = \sum_{k=1}^{\gamma} \operatorname{tr}\left(\boldsymbol{G}^{\frac{k}{\gamma}} \Delta \boldsymbol{G}^{\frac{\gamma-k+1}{\gamma}} \Delta \right), \tag{9.16}$$

where $\gamma = \log d$, $\Delta = \sum_{i=1}^n \delta_i \boldsymbol{A}_i$, and $\boldsymbol{G} = \nabla \Phi(\boldsymbol{x})$.

The goal is to find a direction $\boldsymbol{\delta}$ such that (9.16) is small. A naive approach, ignoring non-commutativity, uses the following:

Lemma 9.27. *Let Δ be symmetric and $\boldsymbol{G}$ be positive semidefinite. For all $\alpha, \beta > 0$, we have*

$$\operatorname{tr}(\boldsymbol{G}^\alpha \Delta \boldsymbol{G}^\beta \Delta) \leqslant \operatorname{tr}(\boldsymbol{G}^{\alpha+\beta} \Delta^2).$$

Remark 9.28. The gap in the inequality,

$$I_\alpha(\boldsymbol{G}, \Delta) = \operatorname{tr}(\boldsymbol{G}\Delta^2) - \operatorname{tr}(\boldsymbol{G}^\alpha \Delta \boldsymbol{G}^{1-\alpha} \Delta),$$

is known in quantum information theory as the *Wigner–Yanase–Dyson (WYD) skew information*. Lieb [Lie73] showed that I_α is concave as a function of $\boldsymbol{G}$ when $0 < \alpha < 1$. It would be interesting to better understand the maximizers of $I_\alpha(\cdot, \Delta)$ when $\boldsymbol{G}$ is viewed as a vector in the simplex. For example, when $\alpha = 1/2$, the method of Lagrange multipliers implies that any maximizer satisfies $G_i \propto (\sum_{j \neq i} G_j^{1/2} A_{ij}^2)^2$ for all $i \in [n]$.

Proof. Without loss of generality, assume $\boldsymbol{G}$ is diagonal. Then,

$$\begin{aligned} \operatorname{tr}(\boldsymbol{G}^\alpha \Delta \boldsymbol{G}^\beta \Delta) &= \sum_{i,j=1}^n G_{ii}^\alpha G_{jj}^\beta \Delta_{ij}^2 \\ &= \sum_{i,j=1}^n G_{ii}^{\alpha+\beta} \Delta_{ij}^2 - \frac{1}{2} \sum_{i,j=1}^n (G_{ii}^\alpha - G_{jj}^\alpha)(G_{ii}^\beta - G_{jj}^\beta) \Delta_{ij}^2. \end{aligned}$$

The first term is $\operatorname{tr}(\boldsymbol{G}^{\alpha+\beta} \Delta^2)$, and the subtracted term is always nonnegative. □

Applying this lemma bounds (9.16) by

$$\left\langle \boldsymbol{\delta}, \nabla^2 \Phi(\boldsymbol{x}) \boldsymbol{\delta} \right\rangle \leqslant \gamma \operatorname{tr}\left(G^{2-q} \Delta^2\right) .$$

This makes the expression easier to bound, but the prefactor $\gamma = \log d$ ultimately causes a $\sqrt{\log d}$ loss. The insight from [BBvH23] is that such trace inequalities are too loose: the idealized model satisfies them strictly, with a dimension-dependent improvement. Developing a regularization-based proof that tightly bounds (9.16) remains an intriguing open problem.

High-rank instances

Finally, we note that the low-rank assumption, while technically convenient, can seem artificial. Many instances relevant to applications have full rank. For example, Pravesh Kothari communicated the following high-rank instances:

Problem 9.29. Let $\boldsymbol{P}_1, \dots, \boldsymbol{P}_n$ be $n \times n$ permutation matrices. Show that there exists $\boldsymbol{x} \in \{-1, 1\}^n$ such that

$$\left\| \sum_{i=1}^{n} x_i \boldsymbol{P}_i \right\|_2 \lesssim \sqrt{n} .$$

It remains unclear whether our algorithm from §9.4 (that is, Newton's method with an $\ell_{\frac{1}{2}}$-regularizer) can solve Problem 9.29.

9.6. Summary

We introduced a new framework to prove discrepancy bounds using Newton's method on a regularized discrepancy objective function. We showed how to recover Spencer's theorem and highlighted possible avenues for the matrix Spencer problem. In the next chapter, we will see another application of this framework.

CHAPTER 10.

Discrepancy of Sparse and Pseudorandom Vectors

Spencer's bound characterizes the worst-case discrepancy of ℓ_∞-bounded vectors. Its extremal instances such as Hadamard matrices have dense entries. What if the input has instead *sparse* or *multiscale* entries? Do such constraints make vector balancing fundamentally easier? The *Beck–Fiala* and *Komlós* conjectures posit that this is the case.

In this chapter, we adapt our framework to prove new bounds on *pseudorandom* instances of the Beck–Fiala and Komlós conjectures. For rotation matrices, our result improves on the classical result of Banaszczyk [Ban98]. Theorems 10.5 and 10.6 are the main results of this chapter. We also discuss alternative approaches based on duality and compression arguments, and relate them to our framework.

Table of contents

The results in this chapter appeared in [PV23].

10.1. Introduction

The main conjecture of interest for this chapter is the following.

Conjecture 10.1 (Komlós conjecture)**.** *There exists a universal constant $K > 0$ such that for any $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n \in \mathbb{R}^d$ with $\|\boldsymbol{u}_i\|_2 \leqslant 1$ for each $i \in [n]$,*

$$\min_{\boldsymbol{x} \in \{-1,1\}^n} \left\| \sum_{i=1}^n x_i \boldsymbol{u}_i \right\|_\infty \leqslant K \,.$$

Methods based on iteratively finding good partial colorings achieve a bound of $K = O(\log n)$ [Spe85]. The state-of-the-art is a $O(\sqrt{\log n})$-bound due to Banaszczyk [Ban98]. It is also known that this bound can be achieved constructively, using a random walk guided by SDPs [BDG19, BG17], the "Gram-Schmidt walk" algorithm [BDGL19], and a derandomization of it based on the multiplicative weights method [LRR17]. On the other side, there has been only limited attempt to try to refute this conjecture, and the best known lower bound is $K \geqslant 1 + \sqrt{2}$ [Kun23].

In the case where all the entries of all the vectors are on the same scale, we essentially fall back to Spencer's theorem. Hence, the key difficulty with Conjecture 10.1 lies in handling entries of different magnitudes. In fact, the special case where the vectors $\boldsymbol{u}_i$ have sparse $\{0, 1\}$-entries was also conjectured by Beck and Fiala [BF81].

Conjecture 10.2 (Beck–Fiala conjecture)**.** *For any $s = s(n) \in \mathbb{N}$, given any collection of vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n \in \{0, 1\}^d$ where $\boldsymbol{u}_i$ is s-sparse for each $i \in [n]$,*

$$\min_{\boldsymbol{x} \in \{-1,1\}^n} \left\| \sum_{i=1}^n x_i \boldsymbol{u}_i \right\|_\infty \leqslant O(\sqrt{s}) \,.$$

There, the state-of-the art is $O(\min(s, \sqrt{s \log n}))$ [BF81, Ban98], where the second bound follows from the result of Banaszczyk for the Komlós conjecture mentioned earlier.

Our contributions. Recently, Potukuchi [Pot20] gave a new bound for the Beck–Fiala conjecture depending on a pseudo-randomness parameter.

Definition 10.3. Let $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ be a matrix. Define

$$\lambda(\boldsymbol{A}) \coloneqq \sup_{\substack{\|\boldsymbol{v}\|_2 = 1 \\ \langle \boldsymbol{v}, \boldsymbol{1} \rangle = 0}} \left\| \boldsymbol{A}^{\odot 2} \boldsymbol{v} \right\|_2 \,,$$

where $(\boldsymbol{A}^{\odot 2})_{ij} \coloneqq A_{ij}^2$ for all $i \in [d], j \in [n]$.

In the special case where $\boldsymbol{A}$ is the adjacency matrix of a s-regular graph, $\lambda(\boldsymbol{A})$ is the second largest eigenvalue of $\boldsymbol{A}$ and is bounded by s. As observed in [Pot20], $\lambda(\boldsymbol{A})$ is typically much smaller than this worst-case bound when $\boldsymbol{A}$ is the incidence matrix of a *random* regular set system. We will also check in §10.3 that this still holds for natural random instances of Komlós conjecture.

Theorem 10.4 (Theorem 1.1 in [Pot20]). *Let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n \in \{0,1\}^d$ be s-sparse vectors and let $\boldsymbol{A}$ be the $d \times n$ matrix with columns $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$. There is a randomized, polynomial-time algorithm outputting $\boldsymbol{x} \in \{-1,1\}^n$ such that*

$$\left\| \sum_{i=1}^{n} x_i \boldsymbol{u}_i \right\|_\infty \leqslant O(\sqrt{s} + \lambda(\boldsymbol{A})) \,.$$

The proof of Theorem 10.4 is based on iteratively running the random walk algorithm of Lovett and Meka [LM15]. We say that a discrepancy constraint $j \in [d]$ is *pseudo-random* at a given time if the ℓ_2-mass of that constraint restricted to active coordinates has decreased proportionally to the rate of active coordinates at that time. Potukuchi shows that (1) we can make progress on the pseudo-random constraints matching the $O(\sqrt{s})$ Beck–Fiala bound, and (2) when $\lambda(\boldsymbol{A})$ is small, constraints with large enough active ℓ_2-mass are pseudo-random. Finally, the bound on the discrepancy of the rows after their active ℓ_2-mass has become small crucially uses the fact that the instance is $\{0,1\}$-valued.

Our contributions in this chapter are new discrepancy bounds that generalize both Theorem 10.4 and Banaszczyk's bound [Ban98, BDG19]. Moreover, they hold both in the Beck–Fiala setting *and* in the Komlós setting. We apply these results in §10.3 to deduce random versions of Komlós conjecture. Next, we state the two theorems that we will prove in the next sections. For simplicity, we focus on the case of square matrices ($d = n$), although we naturally expect the techniques to generalize to the $d > n$ case.

Theorem 10.5 (Bound for pseudorandom Komlós instances). *Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be such that each column has ℓ_2-norm at most 1. Then there is a deterministic, polynomial-time algorithm to find $\boldsymbol{x} \in \{\pm 1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty = O(1 + \sqrt{\lambda(\boldsymbol{A}) \log n}) \,.$$

Theorem 10.6 (Bound for pseudorandom Beck–Fiala instances). *Let $\boldsymbol{A} \in \{0, \pm 1\}^{n \times n}$ be such that each column has at most s nonzero entries. Then there is a deterministic, polynomial-time algorithm to find $\boldsymbol{x} \in \{\pm 1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty = O(\sqrt{s} + \min(\sqrt{\lambda(\boldsymbol{A}) \log n}, \lambda(\boldsymbol{A}))) \,.$$

As immediate corollaries, Theorem 10.5 implies Komlós conjecture when $\lambda(\boldsymbol{A}) \lesssim \frac{1}{\log n}$ and Theorem 10.6 implies Beck–Fiala conjecture when $\lambda(\boldsymbol{A}) \lesssim \sqrt{s} + \frac{s}{\log n}$. This strictly improves

Theorem 10.4 for Beck–Fiala instances in the regime $\lambda(\boldsymbol{A}) \in [O(\log n), O(s)]$. Furthermore, Theorem 10.5 is the first result for pseudorandom instances of Komlós conjecture.

Remark 10.7. Unfortunately, the mere column-sparsity assumption in Theorem 10.6 does not suffice to ensure that $\lambda(\boldsymbol{A}) \leqslant s$.[1] However, as we will see (Remark 10.14), we can essentially replace $\lambda(\boldsymbol{A})$ by $\min(\lambda(\boldsymbol{A}), s)$ in the analysis. In this sense, our algorithm also matches Banaszczyk's bound.

10.2. Proof of the discrepancy bound for pseudorandom instances

10.2.1. Proof strategy and notations

We now describe our strategy for Theorems 10.5 and 10.6.

We start by giving some idea of how we will use the fact that $\lambda(\boldsymbol{A})$ is small in our discrepancy framework. The main insight of [Pot20] is that at any point in time in a discrepancy walk, we can control the ℓ_2-mass restricted to active coordinates of all rows simultaneously.

Lemma 10.8 (Lemma 2.3 in [Pot20])**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $F \subseteq [n]$ be of size k. Then, for any constant $D > 0$, there exists a subset $S \subseteq [n]$ such that $|S| \leqslant k/D^2$ and for any $i \notin S$,*

$$\sum_{j\in F} A_{ij}^2 \leqslant \frac{k}{n}\sum_{j=1}^{n} A_{ij}^2 + D\lambda(\boldsymbol{A})\,.$$

Intuitively, the term $\frac{k}{n}\sum_{j\leqslant n} A_{ij}^2$ corresponds to the ℓ_2-squared-mass we would expect the row A_i to have if the set of active coordinates F were picked at random. The parameter $\lambda(\boldsymbol{A})$ gives a bound on the deviation from this random behavior. In particular, the ℓ_2-mass of a row essentially decreases as in the average case as long as it is $\Omega(\sqrt{\lambda(\boldsymbol{A})})$. We include a proof of Lemma 10.8 here for completeness.

Proof. Let us denote $B_{ij} \coloneqq A_{ij}^2$ for all $i, j \in [n]$. If $\boldsymbol{u} \in \mathbb{R}^n$ is a vector orthogonal to $\mathbf{1}$, then by definition we have $\sum_i \langle \boldsymbol{B}_i, \boldsymbol{u}\rangle^2 \leqslant \lambda(\boldsymbol{A})^2\|\boldsymbol{u}\|_2^2$. Consider $\boldsymbol{u} \in \mathbb{R}^n$ such that $u_j \coloneqq 1 - \frac{k}{n}$ if $j \in F$ and $u_j \coloneqq -\frac{k}{n}$ if $j \notin F$. Then $\boldsymbol{u}$ is orthogonal to $\mathbf{1}$ and $\|\boldsymbol{u}\|_2^2 \leqslant k$. Hence,

$$\sum_{i=1}^{n}\left(\sum_{j\in F} A_{ij}^2 - \frac{k}{n}\sum_{j=1}^{n} A_{ij}^2\right)^2 \leqslant \lambda(\boldsymbol{A})^2 k\,.$$

The result follows from a simple counting argument. □

[1] For example, consider a vector v with half $+1$ and half -1 entries, take the first row of $\boldsymbol{A}$ to be $\boldsymbol{v}$ and fill the other rows with zeros. $\boldsymbol{A}$ has one nonzero entry per column but $\|\boldsymbol{A}\boldsymbol{v}\|_2 = \sqrt{n}\|\boldsymbol{v}\|_2$.

Roadmap of the proof

In order to prove Theorem 10.5 and Theorem 10.6, we will track two different types of potential functions, depending on which of the main term or the error term in Lemma 10.8 dominates. §10.2.2 will be devoted to bounding the discrepancy incurred in the regime where the row mass decreases as if the input were random. The analysis here will mirror our proof of Spencer's theorem. In §10.2.3, we will consider the case where the error term dominates. There we leverage the fact that the row mass has become small. In this setting, we will use a potential function that was introduced in [LRR17, Appendix B] to recover Banaszczyk's bound with the multiplicative weights update method.

Before starting the proof, we introduce some useful concepts and notations. From now on, we fix a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with column ℓ_2-norm bounded by 1. With the context being clear, we will write $\lambda := \lambda(\boldsymbol{A})$. Our algorithm will follow the structure of the meta-algorithm Algorithm 4, therefore we will use our usual notations: $\boldsymbol{x}(t)$ for the coloring at time t, $F(t)$ for the active coordinates at time t, etc.

Definition 10.9. For each row $i \in [n]$, we define

$$t_i := \min \left\{ t \geqslant 0 : \sum_{j \in F(t)} A_{ij}^2 \leqslant 8\lambda \right\} .$$

In words, t_i represents the time at which the i-th row stops behaving as if it were random. We will write $P(t) := \{i \in [n] : t \leqslant t_i\}$ for the set of "pseudorandom" rows at time t. The following observation explains what we mean by "pseudorandom" – we can pretend as if the freezing process decreases linearly the ℓ_2-mass of the rows. It is an easy consequence of Lemma 10.8.

Claim 10.10. *Fix any time step $t \geqslant 0$. There exists a subset of rows $I = I(t) \subseteq [n]$ such that $|I| \leqslant \frac{|F(t)|}{16}$ and for any $i \in P(t)$, $i \notin I$:*

$$\sum_{j \in F(t)} A_{ij}^2 \leqslant \frac{2|F(t)|}{n} \sum_{j=1}^{n} A_{ij}^2 .$$

For any row $i \in [n]$, we will track separately the contributions to its discrepancy for $t \leqslant t_i$ and $t > t_i$.

Random regime

Similarly to [Pot20], we group together the rows that have similar *total* ℓ_2-mass. For any $r \in \{1, \dots, \lceil \log_2 n \rceil\}$, let

$$R_r := \left\{ i \in [n] : \sum_{j=1}^{n} A_{ij}^2 \in (2^{r-1}, 2^r] \right\} .$$

We also consider $R_0 := \{i \in [n] : \sum_{j=1}^n A_{ij}^2 \leqslant 1\}$. An easy double counting argument bounds the size of each R_r: $|R_r| \leqslant n2^{1-r}$ for any $r \leqslant \lceil \log_2 n \rceil$. Our strategy will be to play several Spencer's games in parallel (restricted to the rows in R_r, for each value of r) and carefully allocate some "effective dimension" to each of them at any step of the walk.

We now define $\pi_{r,t} : \mathbb{R}^n \to \mathbb{R}^n$ to be a projection to the coordinates of R_r of the discrepancy of the rows that behave pseudorandomly. Once $t > t_i$, we keep tracking in the i-th row the same value $\langle \boldsymbol{A}_i, \boldsymbol{x}(t_i) \rangle$. In short, for any $\boldsymbol{x} \in \mathbb{R}^n$ and $i \in [n]$,

$$(\pi_{r,t}(\boldsymbol{x}))_i = \begin{cases} 0 & \text{if } i \notin R_r \\ \langle \boldsymbol{A}_i, \boldsymbol{x}(t_i) \rangle & \text{if } i \in R_r \text{ but } i \notin P(t) \\ \langle \boldsymbol{A}_i, \boldsymbol{x} \rangle & \text{if } i \in R_r \cap P(t) \end{cases}$$

Now we are able to define our potential functions in the random regime. For any $r = 0, \dots, \lceil \log_2 n \rceil$, let

$$\Phi_{r,t}(\boldsymbol{x}) := \omega^*_{\frac{1}{2}, \sqrt{n2^{1-r}}}(\pi_{r,t}(\boldsymbol{x})) \,.$$

The choice of $\eta_r = \sqrt{n2^{1-r}}$ as regularization parameter can be justified by the fact that there are at most $n2^{1-r}$ rows in the r-th group – so this is essentially the smallest value of η_r that makes the additive approximation error of the regularized maximum $O(1)$ (which is our target discrepancy in this regime).

Our main lemma states that it is possible to design oracle with a bit of slack in the dimension requirements, in such a way that all the potential functions Φ_r only pay a constant amortized increase over the duration of the walk.

Lemma 10.11. *There is a construction of* $\mathrm{oracle}(\boldsymbol{A}, \boldsymbol{x}(t))$ *that always returns a subspace of codimension at most* $\frac{k}{4} + O(1)$ *(with* k *being the number of active coordinates of* $\boldsymbol{x}(t)$*), such that for any* $r = 0, \dots, \lceil \log_2 n \rceil$, $\Phi_{r,T}(\boldsymbol{x}(T)) \leqslant O(1)$.

Small row regime

Set $\eta := \sqrt{\frac{\log n}{\lambda}}$ (it will be the parameter of the regularizer in this regime). We define $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ to be the following thresholded version of $\boldsymbol{A}$: for any $i, j \in [n]$, let $B_{ij} := A_{ij}$ if $A_{ij}^2 \leqslant \frac{1}{16K^2\eta^2}$ and $B_{ij} := 0$ otherwise. We will later see that it is sufficient to monitor the discrepancy of $\boldsymbol{B}$ instead of $\boldsymbol{A}$.

Recall that at this point of the walk, each row will have effective ℓ_2-mass $O(\lambda)$. For algorithms based on orthogonality constraints, there is not much difference between having bounded rows and bounded columns, so this justifies patching an algorithm for Banaszczyk's setting at this point. We implement the potential function from [LRR17] in our regularization framework.

Let $\pi'_t \colon \mathbb{R}^n \to \mathbb{R}^n$ to be such that for any $\boldsymbol{x} \in \mathbb{R}^n$ and $i \in [n]$,

$$(\pi'_t(\boldsymbol{x}))_i = \begin{cases} \langle \boldsymbol{B}_i, \boldsymbol{x} - \boldsymbol{x}(t_i) \rangle - K\eta \sum_{j=1}^n B_{ij}^2 (x_j - x_j(t_i))^2 & \text{if } i \notin P(t) \\ 0 & \text{if } i \in P(t) \end{cases}$$

for some constant $K > 0$ that we will fix in the proof. Finally, we define our potential function for this regime:

$$\Psi_t(\boldsymbol{x}) := \operatorname{smax}_\eta(\pi'_t(\boldsymbol{x}))\,.$$

Our main lemma states we can make this potential non-increasing (and moreover the subspace dimension necessary for that allows some slack).

Lemma 10.12. *There is a construction of* $\operatorname{oracle}(\boldsymbol{A}, \boldsymbol{x}(t))$ *that always returns a subspace of codimension at most* $\frac{k}{4} + O(1)$ *(where* k *is the number of active coordinates of* $\boldsymbol{x}(t)$*), such that* $\Psi_T(\boldsymbol{x}(T)) \leqslant \Psi_0(\boldsymbol{x}(0))$.

Remark 10.13. Some intuition for π'_t comes from the fact that to get a coloring of discrepancy $\sqrt{\lambda \log n}$ for an input matrix with *rows* of ℓ_2-norm bounded by λ, the Chernoff and union bound argument suffices. If $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ are the rows of the matrix, it is essentially consisting in arguing that when $\boldsymbol{x} \sim \{\pm 1\}^n$, it holds for any $\eta \geqslant 0$ that:

$$\log \mathbb{E} \exp\left(\eta \max_{i \in [n]} |\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle|\right) \leqslant \log \mathbb{E} \sum_{i=1}^n \exp(\eta |\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle|) \leqslant \log \sum_{i=1}^n \exp\left(\frac{\eta^2 \|\boldsymbol{v}_i\|_2^2}{2}\right).$$

If we interpret $\|\boldsymbol{v}_i\|_2^2$ as $\sum_{j \in [n]} v_{i,j}^2 x_j^2$, this might motivate us to look at the softmax of $\{\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle - \eta \langle \boldsymbol{v}_i^{\odot 2}, \boldsymbol{x}^{\odot 2} \rangle / 2\}$.

We are now ready to prove Theorem 10.5 and Theorem 10.6.

Proof of Theorem 10.5 and Theorem 10.6. We assume without loss of generality that

$$\max_{i \in [n]} (\boldsymbol{A}\boldsymbol{x})_i = \|\boldsymbol{A}\boldsymbol{x}\|_\infty$$

by the usual trick of doubling the rows. We define $\operatorname{oracle}(\boldsymbol{A}, \boldsymbol{x}(t))$ to be the intersection of the halfspace from Lemma 10.15 and Lemma 10.16, and of the subspace from Lemma 10.17.

On the one hand, Lemma 10.11 implies that for any $r \leqslant \lceil \log_2 n \rceil$ and $i \in R_r$,

$$\langle \boldsymbol{A}_i, \boldsymbol{x}(t_i) \rangle = \pi_{r,T}(\boldsymbol{x}(T))_i \leqslant \Phi_{r,T}(\boldsymbol{x}(T)) \leqslant O(1)\,.$$

On the other hand, Lemma 10.12 implies that for any $i \in [n]$,

$$\langle \boldsymbol{B}_i, \boldsymbol{x}(T) - \boldsymbol{x}(t_i) \rangle - K\eta \sum_{j=1}^n B_{ij}^2 (x_j(T) - x_j(t_i))^2 \leqslant \Psi_T(\boldsymbol{x}(T)) \leqslant \Psi_0(\boldsymbol{x}(0)) \leqslant \sqrt{\lambda \log n}\,.$$

Observe that

$$\sum_{j=1}^{n} B_{ij}^2 (x_j(T) - x_j(t_i))^2 \leqslant 4 \sum_{j=1}^{n} B_{ij}^2 \lesssim \lambda \,.$$

Hence, $\langle \boldsymbol{B}_i, \boldsymbol{x}(T) - \boldsymbol{x}(t_i) \rangle \lesssim \sqrt{\lambda \log n}$, and by Lemma 10.19, $|\langle \boldsymbol{A}_i - \boldsymbol{B}_i, \boldsymbol{x}(T) - \boldsymbol{x}(t_i) \rangle| \lesssim \sqrt{\lambda \log n}$ holds as well. It remains to use the triangle inequality:

$$\|\boldsymbol{A}\boldsymbol{x}(T)\|_\infty = O(1 + \sqrt{\lambda \log n}) \,.$$

This implies Theorem 10.5 and the first part of Theorem 10.6. For the second part of Theorem 10.6, simply observe that if $\boldsymbol{A}$ is a rescaled Beck–Fiala instance, namely $A_{ij} \in \{0, 1/\sqrt{s}\}$ for all $i, j \in [n]$, then the condition $\sum_{j \in F(t_i)} A_{ij}^2 = O(\lambda)$ implies by that $\boldsymbol{A}_i$ has at most $O(s\lambda)$ nonzero entries in $F(t_i)$, and so $\sum_{j \in F(t_i)} |A_{ij}| = O(\lambda\sqrt{s})$. In particular, the time steps $t \in [t_i, T]$ affect the discrepancy of $\boldsymbol{A}_i$ by at most $O(\lambda\sqrt{s})$. This shows that the constructed coloring also has discrepancy $O(1 + \lambda\sqrt{s})$ in this case, which is equivalent to the second part of the bound in Theorem 10.6. □

Remark 10.14. One may also recover Banaszszyk's bound for Komlós (or Beck–Fiala) instances by repeating the argument from §10.2.3, replacing λ by some universal constant larger than 1 and adding an additional orthogonality constraint to large rows, so that the algorithm can pretend that the rows all have bounded ℓ_2-mass. It follows from our previous observations on the link between the multiplicative weights update method and negative entropy regularization that this would be equivalent to the approach for recovering Banaszczyk's bound in [LRR17].

The plan for the next few sections is as follows. First, we prove Lemma 10.11 in §10.2.2 and Lemma 10.12 in §10.2.3. Then, we study the consequences of Theorem 10.5 and Theorem 10.6 for random instances in §10.3.

10.2.2. Discrepancy in the random regime

Our main goal in this subsection is to prove Lemma 10.11. Throughout this discussion, we fix a small constant $\varepsilon \in (0, 1/5)$. Our first lemma describes a construction of oracle that bounds locally the increase of the potential. This part is very similar in spirit to our proof of Spencer's theorem.

Lemma 10.15. *Let $\boldsymbol{x} := \boldsymbol{x}(t)$ and $k = k(t) := F(t)$. Let $R_0 = \lceil \log_2(32n/k) \rceil$. There exists a subspace $S = S(t) \subseteq F(t)$ such that S has codimension at most $\frac{k}{4}$ and if $\boldsymbol{\delta} \in S$ satisfies $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$,*

$$\Phi_{r,t}(\boldsymbol{x} + \boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) \lesssim u_r(\boldsymbol{\delta}) + \frac{1}{k} \left(\frac{k2^r}{n} \right)^{\frac{1-3\varepsilon}{2}} \|\boldsymbol{\delta}\|_2^2 \quad \text{for any } r \leqslant R_0$$

and

$$\Phi_{r,t}(\boldsymbol{x}+\boldsymbol{\delta}) = \Phi_{r,t}(\boldsymbol{x}) \quad \text{for any } r > R_0$$

where $\{u_r : r \leqslant R_0\}$ *are linear forms.*

Proof. For any $r = 0, \dots, R_0$, let $k_r = k_r(t) := C_1 \left(\frac{k2^r}{n}\right)^{\varepsilon} k$ be the effective subspace dimension devoted to rows in the r-th group, where $C_1 = C_1(\varepsilon)$ is chosen so that $\sum_{r \leqslant R_0} k_r \leqslant k/8$.

Let S' be the orthogonal complement of the span of the large rows and the row in I, namely $(\bigcup_{r>R_0}\{\boldsymbol{A}_i : i \in R_r\})^{\perp}$. These rows all have total ℓ_2-squared mass larger than $2^{R_0-1} \geqslant 16n/k$, so there are at most $k/16$ of them and thereby S_1 has codimension at most $k/16$.

Applying Lemma 9.12 to $r \leqslant R_0$, for some $\boldsymbol{\nabla}_r \in \Delta_n$, it holds for any $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$ that

$$\Phi_{r,t}(\boldsymbol{x}+\boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) \leqslant u_r(\boldsymbol{\delta}) + 2\sqrt{n2^{1-r}} \sum_{i \in R_r \cap P(t)} \nabla_{r,i}^{\frac{3}{2}} \langle \boldsymbol{A}_i, \boldsymbol{\delta}\rangle^2,$$

for some linear form $u_r : \mathbb{R}^n \to \mathbb{R}$.

Let I_r be the set of coordinates that are in the top $k_r/2$ entries of the gradient $\boldsymbol{\nabla}_r$. We define S_r to be the intersection of the orthogonal complement of the span of the rows in $I \cup I_r$ (where I is the set of rows from Claim 10.10 that satisfies $|I| \leqslant k/16$), and of the top $k_r/2$-dimensional eigenspace of

$$\sum_{i \in R_r \cap P(t), i \notin I \cup I_r} \nabla_{r,i}^{\frac{3}{2}} \boldsymbol{A}_i \boldsymbol{A}_i^{\top}$$

over $\mathbb{R}^{F(t)}$. Then if $\boldsymbol{\delta} \in S_r$,

$$\begin{aligned}
&\Phi_{r,t}(\boldsymbol{x}+\boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) - u_r(\boldsymbol{\delta}) \\
&\lesssim \frac{\|\boldsymbol{\delta}\|_2^2 \sqrt{n}}{2^{\frac{r}{2}} k_r} \sum_{i \in R_r \cap P(t), i \notin I \cup I_r} \nabla_{r,i}^{\frac{3}{2}} \sum_{j \in F(t)} A_{ij}^2 && \text{(since } \delta \in S_r\text{)} \\
&\lesssim \frac{\|\boldsymbol{\delta}\|_2^2 k}{k_r \sqrt{n} 2^{\frac{r}{2}}} \sum_{i \in R_r \cap P(t), i \notin I \cup I_r} \nabla_{r,i}^{\frac{3}{2}} \sum_{j=1}^{n} A_{ij}^2 && \text{(by Claim 10.10 and } i \in P(t), i \notin I\text{)} \\
&\leqslant \frac{\|\delta\|_2^2 k 2^{\frac{r}{2}}}{k_r \sqrt{n}} \sum_{i \in R_r \cap P(t), i \notin I \cup I_r} \nabla_{r,i}^{\frac{3}{2}} && \text{(since } i \in R_r\text{)} \\
&\lesssim \frac{\|\boldsymbol{\delta}\|_2^2 k 2^{\frac{r}{2}}}{k_r^{\frac{3}{2}} \sqrt{n}} && \text{(since } i \notin I_r\text{)} \\
&\lesssim \frac{1}{k} \left(\frac{k2^r}{n}\right)^{\frac{1-3\varepsilon}{2}} \|\boldsymbol{\delta}\|_2^2 . && \text{(by definition of } k_r\text{)}
\end{aligned}$$

Finally we set $S := S' \cap \bigcap_{r \leqslant R_0} S_r$. One can check that S has codimension at most $\frac{k}{16} + \frac{k}{16} + \sum_{r \leqslant R_0} k_r \leqslant \frac{k}{4}$. □

The following step is a trick to handle the first-order terms. Indeed, a caveat is that unlike in Spencer's setting, we cannot afford to move perpendicularly to all the gradients simultaneously.

Lemma 10.16. *Fix $\boldsymbol{x} \in \mathbb{R}^n$. Let S be a subspace such that for any $\boldsymbol{\delta} \in S$ and $r \leqslant R_0$,*

$$\Phi_{r,t}(\boldsymbol{x} + \boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) \lesssim u_r(\boldsymbol{\delta}) + \frac{1}{k}\left(\frac{k2^r}{n}\right)^{\frac{1-3\varepsilon}{2}} \|\boldsymbol{\delta}\|_2^2,$$

for some linear forms $\{u_r : r \leqslant R_0\}$.

Then for any $\boldsymbol{\delta} \in S$, at least one of $+\boldsymbol{\delta}$ or $-\boldsymbol{\delta}$ satisfies that for any $r \leqslant R_0$,

$$\Phi_{r,t}(\boldsymbol{x} \pm \boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) \lesssim \frac{1}{k}\left(\frac{k2^r}{n}\right)^{\varepsilon} \|\boldsymbol{\delta}\|_2^2.$$

Proof. By picking $\varepsilon < 1/5$, we have for any $\boldsymbol{\delta} \in S$

$$\sum_{r \leqslant R_0} 2^{-\varepsilon r}(\Phi_{r,t}(\boldsymbol{x} + \boldsymbol{\delta}) - \Phi_{r,t}(x) - u_r(\boldsymbol{\delta})) \leqslant \frac{1}{k}\left(\frac{k}{n}\right)^{\varepsilon} \|\boldsymbol{\delta}\|_2^2.$$

By a trivial upper bound, this means that for any $\boldsymbol{\delta} \in S, r \leqslant R_0$,

$$2^{-\varepsilon r}(\Phi_{r,t}(\boldsymbol{x} + \boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x})) \leqslant \sum_{s \leqslant R_0} 2^{-\varepsilon s} u_s(\boldsymbol{\delta}) + \frac{1}{k}\left(\frac{k}{n}\right)^{\varepsilon} \|\boldsymbol{\delta}\|_2^2.$$

In particular, by picking the signing $\pm\boldsymbol{\delta}$ that satisfies $\sum_{s \leqslant R_0} 2^{-\varepsilon s} u_s(\pm\boldsymbol{\delta}) \leqslant 0$, we get that for any $r \leqslant R_0$,

$$\Phi_{r,t}(\boldsymbol{x} \pm \boldsymbol{\delta}) - \Phi_{r,t}(\boldsymbol{x}) \lesssim \frac{1}{k}\left(\frac{k2^r}{n}\right)^{\varepsilon} \|\boldsymbol{\delta}\|_2^2.$$ □

Proof of Lemma 10.11. By combining Lemma 10.15 and Lemma 10.16, we know that at any step where there are k active coordinates remaining, the potential $\Phi_{r,t}$ increases by at most $\frac{1}{k}\left(\frac{k2^r}{n}\right)^{\varepsilon} \|\boldsymbol{\delta}\|_2^2$ if $r \leqslant 1 + \log_2(32n/k)$ and is unchanged for $r \geqslant 1 + \log_2(32n/k)$.

Now fix any $r \leqslant \lceil \log_2(n) \rceil$. By a similar argument to the one in the proof of Theorem 9.1, after letting β_k be the ℓ_2-squared mass injected into x starting from the first time for which there are at most k active coordinates remaining, we can upper bound

$$\Phi_{r,T}(\boldsymbol{x}(T)) - \Phi_{r,0}(\boldsymbol{0}) \lesssim \left(\frac{2^r}{n}\right)^{\varepsilon} \sum_{k=1}^{64n2^{-r}} \frac{\beta_k - \beta_{k-1}}{k^{1-\varepsilon}} = O(1).$$

The claimed bound follows after noting that our choice of parameters for the regularizers also implies $\Phi_{r,0}(\boldsymbol{0}) = O(1)$. □

10.2.3. Discrepancy in the small row regime

Our next goal is to prove Lemma 10.12, which is a direct consequence of the following lemma.

Lemma 10.17. *Fix any time t and let $k := |F(t)|$. There is a subspace S of $\mathbb{R}^{F(t)}$ of codimension at most $\frac{k}{4} + O(1)$ such that*

$$\Psi_t(\boldsymbol{x} + \boldsymbol{\delta}) \leqslant \Psi_t(\boldsymbol{x})$$

holds for any $\boldsymbol{\delta} \in S$ with $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$.

Before proving it, we recall the following well-known result (see for example [BDG19, Theorem 8] or [LRR17, Lemma 21]):

Lemma 10.18. *For any $w_1, \dots, w_m \geqslant 0$, $\boldsymbol{B} \in \mathbb{R}^{m \times n}$, and $\alpha \in (0,1)$, there exists a subspace S of $\mathbb{R}^n$ of codimension at most αn such that for all $\boldsymbol{\delta} \in S$,*

$$\sum_{i \leqslant m} w_i \left(\sum_{j \leqslant n} B_{ij} \delta_j \right)^2 \leqslant \frac{1}{\alpha} \sum_{i \leqslant m} w_i \sum_{j \leqslant n} B_{ij}^2 \delta_j^2 \,. \tag{10.1}$$

Proof. Up to considering $\sqrt{w_i}\boldsymbol{B}_i$, assume without of generality that $w_i = 1$ for all $i \in [m]$. Moreover, the statement is invariant if we remove the zero columns from $\boldsymbol{B}$ and replace δ_j by $\frac{\delta_j}{\|\boldsymbol{B}^j\|_2}$ for all $j \in [n]$. Therefore we can also assume that all columns of $\boldsymbol{B}$ have unit Euclidean length.

Now the right-hand side is just $\frac{\|\boldsymbol{\delta}\|_2^2}{\alpha}$ and the left-hand side is $\boldsymbol{\delta}^\top \sum_{i \leqslant m} \boldsymbol{B}_i \boldsymbol{B}_i^\top \boldsymbol{\delta}$. We can simply choose S to be the subspace of vectors $\boldsymbol{\delta}$ orthogonal to the top αn eigenspace of the linear operator $\sum_{i \leqslant m} \boldsymbol{B}_i \boldsymbol{B}_i^\top$, which has trace n. The result follows a counting argument. □

Proof of Lemma 10.17. To avoid overcharging notations we drop in this proof the dependencies on t and let $\boldsymbol{x} := \boldsymbol{x}(t)$, $F := F(t)$ and $P := P(t)$. When $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$, we can apply Taylor expansion (Lemma 9.12) – for some $\boldsymbol{\nabla} \in \Delta_n$:

$$\begin{aligned}
\Psi_t(\boldsymbol{x} + \boldsymbol{\delta}) - \Psi_t(\boldsymbol{x}) \leqslant & \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij} \delta_j + \eta \sum_{i \notin P} \nabla_i \left(\sum_{j \in F} B_{ij} \delta_j \right)^2 \\
& - 2K\eta \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij}^2 x_j \delta_j + 2K^2 \eta^3 \sum_{i \notin P^t} \nabla_i \left(\sum_{j \in F} B_{ij}^2 x_j \delta_j \right)^2 \\
& - K\eta \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij}^2 \delta_j^2 + K^2 \eta^3 \sum_{i \notin P} \nabla_i \left(\sum_{j \in F} B_{ij}^2 \delta_j^2 \right)^2 .
\end{aligned}$$

First, note that the last term scales as δ_j^4 so we can make it negligible by picking $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$.

We consider the subspace S_1 of codimension at most 2 that is the orthogonal complement of the span of the 2 vectors from the linear terms in $\boldsymbol{\delta}$ in the previous right-hand side. Also, by applying Lemma 10.18 to two different matrices with $\alpha = \frac{1}{8}$, we can find S_2 of codimension $k/4$ such that any $\boldsymbol{\delta} \in S_3$ satisfies the following two conditions:

$$\eta \sum_{i \notin P} \nabla_i \left(\sum_{j \in F} B_{ij} \delta_j \right)^2 \leqslant 8\eta \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij}^2 \delta_j^2 . \tag{10.2}$$

$$2K^2\eta^3 \sum_{i \notin P} \nabla_i \left(\sum_{j \in F} B_{ij}^2 x_j \delta_j \right)^2 \leqslant 16K^2\eta^3 \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij}^4 x_j^2 \delta_j^2 . \tag{10.3}$$

Note that whenever (10.3) is satisfied, it also follows from $|x_i| \leqslant 1$ and the assumption $B_{ij}^2 \leqslant \frac{1}{16K^2\eta^2}$ that

$$2K^2\eta^3 \sum_{i \notin P} \nabla_i \left(\sum_{j \in F} B_{ij}^2 x_j \delta_j \right)^2 \leqslant \eta \sum_{i \notin P} \nabla_i \sum_{j \in F} B_{ij}^2 \delta_j^2 .$$

Let $S \coloneqq S_1 \cap S_2$. S has codimension at most $k/4 + O(1)$ by construction. Picking $K \coloneqq 9$, we get

$$\Psi_t(\boldsymbol{x} + \boldsymbol{\delta}) - \Psi_t(\boldsymbol{x}) \leqslant 0 ,$$

for any $\boldsymbol{\delta} \in S$ satisfying $\|\boldsymbol{\delta}\|_\infty \leqslant 1/\mathrm{poly}(n)$. □

Finally we bound the error of replacing $\boldsymbol{A}$ by $\boldsymbol{B}$.

Lemma 10.19. *For any $i \in [n]$,*

$$|\langle \boldsymbol{A}_i - \boldsymbol{B}_i, \boldsymbol{x}(T) - \boldsymbol{x}(t_i) \rangle| \lesssim \sqrt{\lambda \log n} .$$

Proof. Fix $i \in [n]$. Let $F \coloneqq F(t_i)$ be the set of active coordinates when the i-th row becomes small. Since $\sum_{j \in F} A_{ij}^2 = O(\lambda)$, it must be that

$$|\{ j \in F : A_{ij} - B_{ij} \neq 0 \}| \lesssim \log n .$$

Therefore, by Cauchy-Schwarz,

$$|\langle \boldsymbol{A}_i - \boldsymbol{B}_i, \boldsymbol{x}(T) - \boldsymbol{x}(t_i) \rangle| \leqslant \sum_{j \in F} |A_{ij} - B_{ij}| \lesssim \sqrt{\log n} \sqrt{\sum_{j \in F} (A_{ij} - B_{ij})^2} \lesssim \sqrt{\lambda \log n} . \qquad \square$$

10.3. Application to random instances

10.3.1. Random orthogonal matrices

The next consequence of our bound for pseudorandom instances is that Komlós conjecture is true for *random* rotation matrices. An equivalent geometric way to state our result is the following: there exists a universal constant $C > 0$ such that when we randomly rotate the n-dimensional hypercube around the origin, with high probability there exists a corner at ℓ_∞-distance at most C from the origin.

Rotation matrices appear to be hard instances for proving the Komlós conjecture, as present proof techniques merely manage to match the discrepancy bounds to those for the suprema of Rademacher processes involving the transpose matrix. Improving beyond $O(\sqrt{\log n})$ discrepancy for orthogonal matrices would therefore provide new techniques for treating Rademacher/Gaussian processes for structured matrices. A first step in making progress on this front would therefore be to consider *random* orthogonal matrices.

What we mean by random rotation is a random matrix distributed according to the Haar measure on the orthogonal group $\mathcal{O}(n)$. The Haar measure is a natural generalization of the uniform distribution. We can just think of the sampling as picking the matrix columns to be i.i.d. standard Gaussians in $\mathbb{R}^n$ (which will be linearly independent almost surely), and orthonormalizing them with the Gram-Schmidt process.

We explicitly computed small moments of the entries of such a random matrix with the help of the Maple package IntHaar [GK21].

Claim 10.20. *Suppose $\boldsymbol{A}$ is distributed according to the Haar measure on $\mathcal{O}(n)$. Then,*

$$\mathbb{E}\left[A_{11}^8\right] = \frac{105}{n(n+2)(n+4)(n+6)},$$
$$\mathbb{E}\left[A_{11}^4 A_{12}^4\right] = \frac{9}{n(n+2)(n+4)(n+6)},$$
$$\mathbb{E}\left[A_{11}^2 A_{12}^2 A_{21}^2 A_{22}^2\right] = \frac{n^2+4n+15}{n(n+2)(n-1)(n+1)(n+4)(n+6)}.$$

Corollary 10.21 (Komlós conjecture for random rotations)**.** *There is a deterministic algorithm that given a Haar-distributed random matrix $\boldsymbol{A}$ on $\mathcal{O}(n)$, finds with high probability $\boldsymbol{x} \in \{\pm 1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty = O(1).$$

Proof. Our proof is inspired by the observations in the proof of [Ber01, Theorem 1] for the Haar measure on the unitary group. Consider $\boldsymbol{B} \coloneqq (\boldsymbol{A}^{\odot 2})^\top \boldsymbol{A}^{\odot 2}$. Using Claim 10.20, we see that

$$\mathbb{E}\operatorname{tr}\boldsymbol{B}^2 = \sum_{1 \leqslant i,j,k,l \leqslant n} \mathbb{E}\left[A_{ij}^2 A_{il}^2 A_{kj}^2 A_{kl}^2\right]$$

$$= n^2 \mathbb{E}\left[A_{11}^8\right] + 2n^2(n-1)\mathbb{E}\left[A_{11}^4 A_{12}^4\right] + n^2(n-1)^2 \mathbb{E}\left[A_{11}^2 A_{12}^2 A_{21}^2 A_{22}^2\right]$$
$$= 1 + O\left(\frac{1}{n}\right).$$

Note that 1 and $\lambda(\boldsymbol{A})^2$ are eigenvalues of the positive semidefinite matrix $\boldsymbol{B}$, so $1+\lambda(\boldsymbol{A})^4 \leqslant \operatorname{tr} \boldsymbol{B}^2$, and $\mathbb{E}\,\lambda(\boldsymbol{A})^4 \leqslant O\left(\frac{1}{n}\right)$ by the previous estimate. By Markov's inequality, $\lambda(\boldsymbol{A}) \leqslant \frac{1}{\log n}$ with high probability, and we conclude by applying Theorem 10.5. □

10.3.2. Random Gaussian matrices

Next we show that for matrices with random Gaussian entries, the corresponding λ parameter is small. Without loss of generality, we assume that the input matrix is square, as otherwise (in the regime $m \geqslant n$) we can add extra columns while only worsening λ. We assume that each entry is sampled i.i.d. from $\mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{1}{\sqrt{n}}$, so that all column norms are tightly concentrated around 1, i.e.

$$\Pr(1-\varepsilon \leqslant \|\boldsymbol{A}^j\|_2^2 \leqslant 1+\varepsilon) \geqslant 1 - 2\exp\left(-\frac{n\varepsilon^2}{8}\right),$$

which follows from standard concentration bounds.

Claim 10.22. *Given a random Gaussian matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, where entries are i.i.d. Gaussians $\mathcal{N}(0,\sigma^2)$ with $\sigma = \frac{1}{\sqrt{n}}$, one has that*

$$\max_{\langle \boldsymbol{u}, \mathbf{1}\rangle = 0} \frac{\|\boldsymbol{A}^{\odot 2}\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leqslant \frac{1}{\sqrt{n}}$$

with high probability.

Proof. Let $\boldsymbol{B} \coloneqq \boldsymbol{A}^{\odot 2} - \frac{\mathbf{1}\mathbf{1}^T}{n}$. Observe that

$$\max_{\|\boldsymbol{u}\|_2 = 1, \langle \boldsymbol{u}, \mathbf{1}\rangle = 0} \|\boldsymbol{A}^{\odot 2}\boldsymbol{u}\|_2 \leqslant \max_{\|\boldsymbol{u}\|_2 = 1} \|\boldsymbol{B}\boldsymbol{u}\|_2.$$

Now, $\boldsymbol{B}$ is a matrix with i.i.d. entries such that $\mathbb{E}\,B_{11} = 0$, $\mathbb{E}\,B_{11}^2 = O(1/n^2)$ and $\mathbb{E}\,B_{11}^4 = O(1/n^4)$, so by a standard result from random matrix theory (see e.g. [Tao12, Theorem 2.3.8]), it holds that $n\|\boldsymbol{B}\|_{\mathrm{op}} = O(\sqrt{n})$ with high probability. It readily follows that $\lambda(\boldsymbol{A}) \leqslant 1/\sqrt{n}$ with high probability. □

This shows that by applying Theorem 10.5 random Gaussian matrices have discrepancy $O(1)$.

Corollary 10.23. *Given a random Gaussian matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, where entries are i.i.d. Gaussians $\mathcal{N}(0,\sigma^2)$, $\sigma = \frac{1}{\sqrt{n}}$, there exists a deterministic algorithm that finds a coloring $\boldsymbol{x} \in \{\pm 1\}^n$ such that $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = O(1)$.*

10.4. The compression approach

An interesting special case of the Beck–Fiala conjecture is when the matrix $\boldsymbol{A}$ is the adjacency matrix of some s-regular graph. It turns out that in this setting, a folklore argument based on Lovász Local Lemma implies that there exists a coloring with discrepancy $O(\sqrt{s \log s})$. Although there is an algorithm to construct such a coloring in polynomial time, it is not captured by the iterative framework we introduced in this paper. It is in our opinion a great open problem to unify those two lines of work.

10.4.1. The Lovász Local Lemma algorithm

To formulate the problem more precisely, we provide a new streamlined and self-contained analysis of the algorithm matching the bound based on Lovász Local Lemma. In particular, it highlights the differences with the sticky walk approach. Our inspiration is an argument of [AIS19] for finding a satisfying assignment of bounded degree k-SAT instances.

Theorem 10.24 (Folklore)**.** *There is a randomized algorithm that, given $\boldsymbol{A} \in \{0,1\}^{n \times n}$ with at most s nonzero entries per row and at most s nonzero entries per column, finds with high probability in polynomial time a coloring $\boldsymbol{x} \in \{\pm 1\}^n$ such that $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = O(\sqrt{s \log s})$.*

Proof. We will call a row *bad* (w.r.t. an implicit full coloring) when its discrepancy is larger than $4\sqrt{s \log s}$. We consider the following algorithm. First, we generate a uniformly random coloring. Then we repeat t times the operation of picking the bad row with smallest index (unless there is none, in which case we stop) and resampling all the variables appearing in it.

Since each constraint contains at most s variables and each variable appears in at most s constraints, any constraint has nonempty intersection with at most s^2 other constraints.

Define C_t to be the set of all ordered sequences of t constraints that have nonzero probability to be picked in that order by the algorithm. The execution of the algorithm can be described as a rooted forest of t vertices, each one of these corresponding to a constraint that is picked. When a constraint is picked, it can create at most s^2 children, each of which corresponding to a constraint of lower index that intersects it and became bad after the resampling.

Therefore, we can encode an element $\boldsymbol{c} \in C_t$ by giving $\{c_i : \forall j \in \{1, \dots, i-1\}, c_j < c_i\}$, and a rooted forest on t vertices, each (except the roots) with labels between 1 and s^2. It follows from standard combinatorics that

$$|C_t| \leqslant 2^n \binom{2t}{t} s^{2t} = 2^n (2s)^{2t}.$$

Fix a sequence of resampled constraints $\boldsymbol{c} \in C_t$ and a sequence of $t+1$ colorings $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{t+1} \in \{\pm 1\}^n$. $\boldsymbol{c}$ and $\boldsymbol{u}$ can correspond to a potential execution of the algorithm only if $\boldsymbol{u}_i$ is $\boldsymbol{u}_{i+1}$ where the c_i-th constraint of $\boldsymbol{u}_i$ is bad. Applying Chernoff bounds, we see that there can be at most $2^s/s^8$ such $\boldsymbol{u}_i$'s. It follows by induction that there are at most $2^{st}/s^{8t}$ possible sequences $\boldsymbol{u}_1, \dots, \boldsymbol{u}_t$. On the other hand, for any fixed $\boldsymbol{u}_1, \dots, \boldsymbol{u}_{t+1}, c_1, \dots, c_t$, the probability that the algorithm follows exactly this sequence of constraints and colorings is at most 2^{-st}. Hence, by a union bound, the probability that the final coloring that the sequence of constraints is $c_1, \dots, c_t$ is at most $2^n s^{-8t}$.

To conclude, we can apply another union bound to get that the probability that the final discrepancy is larger than $\sqrt{2s \log s}$ is at most $|C_t| 2^n s^{-8t}$, which is $n^{-\Omega(1)}$ for some $t = n^{O(1)}$. □

Instead of working with fractional colorings, here we walk directly in the space of full colorings. While the row-sparsity assumption is not really restrictive in the sticky walk framework (as we can always pick update vectors orthogonal to large rows), it seems crucial for arguments based on Lovász Local Lemma.

10.4.2. Duality and compression

We believe that developing a proof technique capable of recovering both Theorem 10.24 and Theorem 9.1 is a compelling research direction — perhaps even equivalent to resolving the Beck–Fiala conjecture itself. Here, we outline a promising approach based on compression arguments.

In the context of the matrix Spencer problem, [HRS22], building on [ES18], proposed a method that proceeds by refuting the existence of dual certificates for a linear programming relaxation of the partial coloring problem. The idea is as follows: fix a discrepancy target Δ, and define the polytope of good partial colorings,

$$\mathcal{K}(\Delta) := \{\boldsymbol{x} \in [-1, 1]^n : \|\boldsymbol{A}\boldsymbol{x}\|_\infty \leqslant \Delta\}.$$

Now, fix a candidate coloring $\widehat{\boldsymbol{x}} \in \{-1, 1\}^n$, and consider the linear program in variable $\boldsymbol{x}$:

$$\frac{1}{n} \max_{\boldsymbol{x} \in \mathcal{K}(\Delta)} \langle \boldsymbol{x}, \widehat{\boldsymbol{x}} \rangle . \tag{10.4}$$

If the optimum of (10.4) is large for some guess $\widehat{\boldsymbol{x}}$, then we have found a good partial coloring. Otherwise, if the optimum is small, then the dual of (10.4) yields a certificate that can be used in a communication protocol to compress information beyond what is information-theoretically possible.

This compression-based duality argument can be used to recover Spencer's theorem. Moreover, the structure of the dual program suggests that similar ideas may extend to recover the guarantees of the Lovász Local Lemma via the proof of Theorem 10.24.

10.4.3. The twisted hypercubes

Finally, we introduce a family of discrepancy instances for which the tools we use to analyze our iterative framework fail to provide interesting bounds. It is an interesting question whether a more refined analysis will yield an improved discrepancy bound.

We view these examples as useful benchmarks for evaluating progress on improving discrepancy bounds.

Definition 10.25 (Twisted Hypercubes)**.** The graph on one vertex is the only twisted hypercube of dimension 0. A *twisted hypercube* of dimension d is then obtained by taking two copies of the same twisted hypercube of dimension $d-1$, and adding a matching between both vertex sets.[2]

Twisted hypercubes of dimension d have $n = 2^d$ vertices, each of degree $O(\log n)$. Theorem 10.24 implies that they have colorings of discrepancy $O(\sqrt{\log n \log\log n})$, but Theorem 10.6 only gives a bound of $O(\log n)$ (note that this would also follow from [BF81]). It remains an interesting open problem to construct colorings of twisted hypercubes with discrepancy $o(\log n)$ via the sticky walk approach.[3]

Remark 10.26. Although smoothing the twisted hypercube is not sufficient to apply our bounds on pseudorandom Beck–Fiala instances, we can at least transform it into the adjacency matrix of a (multi-)graph with constant spectral expansion, while only losing an $O(1)$-additive factor on the discrepancy of any coloring. Therefore, our question on the discrepancy of symmetric instances could be reduced to that of the discrepancy of symmetric *expanding* instances.

10.5. Summary

We established that the Komlós and Beck–Fiala conjectures hold for instances whose second eigenvalue satisfies $\lambda \ll 1/\log n$, a condition met with room to spare in several natural random models. We also discussed alternative proof strategies based on compression and duality, and outlined the prospect of unifying these with the iterative frameworks developed in this work.

[2] A slightly different construction consists in adding a matching between two *potentially distinct* twisted hypercubes of dimension $d-1$. This is for example the convention chosen in one previous use of the term "twisted hypercube" in the litterature [DPP+18]. We believe it does not make much difference in the discrepancy setting.

[3] One could also consider randomly twisted hypercubes, obtained by taking recursively uniformly random matchings. They form a family of structured random instances that are not captured by our pseudorandom bounds.

CHAPTER 11.

Optimal Constants in Discrepancy and Sparsification

In this final chapter, we refine the techniques developed in Part III to improve the leading constants in discrepancy problems. Our main result (Theorem 11.1) is that *4.1 standard deviations suffice* for Spencer's theorem. This result follows from analyzing an algorithm that minimizes an ℓ_q-regularized objective for an optimized choice of constant q, together with several refinements of our previous arguments that allow us to control constant-factor losses throughout the analysis.

We then discuss how this analysis might be further improved by considering variants of Spencer's problem. First, in the learning-with-experts setting, we show that our framework recovers the optimal constant, but only if we account for a rank-one Hessian term that we ignored in Spencer's setting. Second, for the related (and easier) ellipsoid discrepancy problem, we show that recovering the tight bound requires an amortized analysis of second-order contributions along the walk. Finally, we argue that our framework is unlikely to improve the constants for the construction of optimal spectral graph sparsifiers.

Table of contents

§11.1 and §11.3 appeared in [PV23].

11.1. An improved constant for Spencer's theorem

In his original paper, Spencer proves that any matrix in $[-1,1]^{n\times n}$ has discrepancy at most $5.4\sqrt{n}$ [Spe85]. The question of the optimal constant in this statement remains open, despite the attention it has received both on the lower and upper bound side [Bel13, BKMR25]. In this section, we improve Spencer's bound to $4.1\sqrt{n}$. Prior to this work, [Bel13, §5] improves Spencer's bound to $5.2\sqrt{n}$ and sketches how to obtain $3.7\sqrt{n}$, but some computations rely on personal communication. Unlike all these previous results, our proof is algorithmic.

Theorem 11.1. *For every $\boldsymbol{A} \in [-1,1]^{n\times n}$, there exists $\boldsymbol{x} \in \{-1,1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty \leqslant 4.1\sqrt{n} + O(1)\,.$$

Moreover, $\boldsymbol{x}$ can be found by a randomized algorithm running in polynomial time.

To prove Theorem 11.1, we revisit the argument from §9.4 by tracking constants more carefully. We start by giving an analog of Lemma 9.15 with a tighter leading constant.

Lemma 11.2. *There exists $C > 0$ such that the following holds for any $q \in (0,1)$ and $k, n \geqslant 1$. Let $\boldsymbol{M} = \sum_{i=1}^n \nabla_i^{2-q} \boldsymbol{u}_i \boldsymbol{u}_i^\top$ for some vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ in the unit ball of $\mathbb{R}^k$, and $\boldsymbol{\nabla} \in \Delta_n$.*

Then there exists a 2-dimensional subspace S such that the projection $\boldsymbol{\Pi}_S$ onto S satisfies

$$\|\boldsymbol{\Pi}_S \boldsymbol{M} \boldsymbol{\Pi}_S\|_2 \leqslant \left(1 + \frac{C}{k}\right) \frac{1}{k^{2-q}}\,.$$

Moreover, for any constant $\varepsilon > 0$, there is a randomized polynomial-time algorithm outputting a 2-dimensional subspace S satisfying with high probability

$$\|\boldsymbol{\Pi}_S \boldsymbol{M} \boldsymbol{\Pi}_S\|_2 \leqslant (1+\varepsilon)\left(1 + \frac{C}{k}\right) \frac{1}{k^{2-q}}\,.$$

Proof. Assume that $\nabla_1 \geqslant \ldots \geqslant \nabla_m$ without loss of generality. We will prove that if α is sampled uniformly in the interval $(\frac{1}{2}, 1)$, then

$$f(\boldsymbol{\nabla}) := \mathop{\mathbb{E}}_{\alpha\sim(\frac{1}{2},1)} \left[\sum_{i \geqslant \lfloor \alpha k \rfloor} \nabla_i^{2-q}\right] \leqslant \frac{k^{q-1}}{4} + O(k^{q-2}) = \mathop{\mathbb{E}}_{\alpha\sim(\frac{1}{2},1)} [(1-\alpha)k^{q-1}] + O(k^{q-2})\,. \quad (11.1)$$

We first explain why (11.1) implies the desired bound. Let $\alpha \in (\frac{1}{2}, 1)$ be such that

$$\frac{1}{(1-\alpha)k} \sum_{i \geqslant \lfloor \alpha k \rfloor} \nabla_i^{2-q} \leqslant k^{q-2} + O(k^{q-3})\,.$$

Then we can repeat the proof of Lemma 9.15 to get a 2-dimensional subspace S such that for all unit $\boldsymbol{v} \in S$,

$$\langle \boldsymbol{v}, \boldsymbol{M}\boldsymbol{v}\rangle \leqslant k^{q-2} + O(k^{q-3})\,,$$

which is equivalent to the desired statement since $\boldsymbol{M} \succeq 0$. Furthermore, if $\varepsilon > 0$ is constant, the corresponding α can be found with high probability by repeating the experiment and using Markov's inequality.

It remains to prove (11.1). We compute explicitly

$$f(\boldsymbol{\nabla}) = 2\int_{\frac{1}{2}}^{1} \sum_{i \geqslant \lfloor \alpha k \rfloor} \nabla_i^{2-q} \, \mathrm{d}\alpha = \sum_{i \geqslant \lfloor \frac{k}{2} \rfloor} \left(\frac{2(i+1)}{k} - 1 \right) \nabla_i^{2-q} .$$

Let $\mathcal{P} = \{\boldsymbol{\nabla} \in \Delta_m : \nabla_1 \geqslant \ldots \geqslant \nabla_m\}$. Since $f \colon \mathcal{P} \to \mathbb{R}$ is a convex function, it attains its maximum at an extreme point of $\mathcal{P}$. Those are of the form $\boldsymbol{z}_\ell := \left(\frac{1}{\ell} \ldots \frac{1}{\ell} \, 0 \ldots 0\right)$ (with ℓ nonzero coordinates) for some $\ell \in [m]$. Moreover, the maximum of f has to be attained when $\ell = \gamma k$, with $\gamma \in [\frac{1}{2}, 1]$. However, in that case,

$$\begin{aligned} f(\boldsymbol{z}_\ell) &= \ell^{q-2} \left(\frac{\ell(\ell+1) - \frac{k}{2}(\frac{k}{2}+1)}{k} - \left(\ell - \frac{k}{2} + 1 \right) + O(1) \right) \\ &= k^{q-1} \left(\sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} + \frac{1}{4\gamma^{3/2}} \right) + O(k^{q-2}) . \end{aligned}$$

Finally, $\gamma \mapsto \sqrt{\gamma} - \frac{1}{\sqrt{\gamma}} + \frac{1}{4\gamma^{3/2}}$ is increasing on $[\frac{1}{2}, 1]$, with maximum equal to $\frac{1}{4}$ for $\gamma = 1$. This concludes the proof of (11.1). □

We also sharpen the constant in front of the second-order term in Lemma 9.12.

Lemma 11.3. *There exist universal constants $C_1, C_2 \in (0,1)$ such that if $\boldsymbol{\nabla} := \nabla \omega^*_{q,\eta}(\boldsymbol{y})$, then for all $\boldsymbol{\delta} \in \mathbb{R}^n$ with $\|\boldsymbol{\delta}\|_\infty \leqslant C_1 \frac{1-q}{n\eta}$,*

$$\omega^*_{q,\eta}(\boldsymbol{y} + \boldsymbol{\delta}) \leqslant \omega^*_{q,\eta}(\boldsymbol{y}) + \langle \boldsymbol{\nabla}, \boldsymbol{\delta} \rangle + \frac{\eta}{2(1-q)} \left(1 + \frac{C_2}{n} \right) \sum_{i=1}^{n} \nabla_i^{2-q} \delta_i^2 .$$

Proof. The proof is identical to the proof of Lemma 9.12. We simply replace (9.9) by the stronger inequality following from the stronger assumption on $\|\boldsymbol{\delta}\|_\infty$. □

Proof of Theorem 11.1. We follow the proof of Theorem 9.14. We set the update size in Algorithm 4 to be $L = \frac{C_1}{4n^2}$, where C_1 is the constant from Lemma 11.3. We use the doubling trick Remark 9.2 to replace $\boldsymbol{A}$ by a $2n \times n$ matrix such that $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{i \in [n]} \langle \boldsymbol{A}_i, \boldsymbol{x} \rangle$ for any vector $\boldsymbol{x}$. We use the potential function $\omega^*_{q,\eta}$ for some parameters $q \in (0,1)$ and $\eta > 0$ to be optimized at the end.

First, by Lemma 9.11, the initial loss is

$$\omega^*_{q,\eta}(\boldsymbol{0}) \leqslant \frac{(2n)^{1-q}}{\eta q} .$$

Then, at any time t, we apply Lemma 11.3 to get that for any $\|\boldsymbol{\delta}\|_\infty \leqslant L$,

$$\omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{A}\boldsymbol{\delta}) - \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t)) \leqslant \langle \boldsymbol{A}\boldsymbol{\delta}, \boldsymbol{\nabla}\rangle + \frac{\eta}{2(1-q)}\left(1 + \frac{C_2}{n}\right)\sum_{i=1}^{2n} \nabla_i^{2-q}\langle \boldsymbol{A}_i, \boldsymbol{\delta}\rangle^2 ,$$

where $\boldsymbol{\nabla} = \nabla\omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t)) \in \Delta_{2n}$. Next, we apply Lemma 11.2, where the $\boldsymbol{u}_i$ are the normalized rows of $\boldsymbol{A}$ restricted to the active coordinates. Note that these rows have ℓ_2-norm at most $\sqrt{k}$, where k is the number of active coordinates. In this way, we find a 2-dimensional subspace S such that if $\boldsymbol{\delta} \in S$ has small enough ℓ_∞-norm,

$$\frac{1}{\|\boldsymbol{\delta}\|_2^2}\sum_{i=1}^{2n} \nabla_i^{2-q}\left\langle \frac{1}{\sqrt{k}}\boldsymbol{A}_i, \boldsymbol{\delta}\right\rangle^2 \leqslant \left(1 + \frac{C}{k}\right)\frac{1}{k^{2-q}}$$

We then choose a direction for $\boldsymbol{\delta} \in S$ so that $\langle \boldsymbol{\delta}, \boldsymbol{x}(t)\rangle = 0$, and a signing $\pm\boldsymbol{\delta}$ that makes the first-order term $\langle \boldsymbol{A}(\pm\boldsymbol{\delta}), \boldsymbol{\nabla}\rangle < 0$. With this choice of $\boldsymbol{\delta}$, the increase in the potential at time t is at most

$$\omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{A}\boldsymbol{\delta}) - \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t)) \leqslant \|\boldsymbol{\delta}\|_2^2 \frac{\eta}{2(1-q)}\left(1 + \frac{C_2}{n}\right)\left(1 + \frac{C}{k}\right)\frac{1}{k^{1-q}} .$$

Then, the summation by parts argument of §9.4 yields

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{x}(T)\|_\infty &\leqslant \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(T)) \\
&\leqslant \omega^*_{q,\eta}(\boldsymbol{0}) + \sum_{t=1}^{T} \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t)) - \omega^*_{q,\eta}(\boldsymbol{A}\boldsymbol{x}(t-1)) \\
&\leqslant \frac{(2n)^{1-q}}{\eta q} + \frac{\eta}{2(1-q)} \cdot \frac{n^q}{q} + O(1) .
\end{aligned}$$

Optimizing over $\eta > 0$ and $q \in (0, 1)$, we obtain that up to a constant additive error, the discrepancy of $\boldsymbol{A}$ is at most

$$2\sqrt{n}\min_{q\in(0,1)}\sqrt{\frac{1}{2^q q^2(1-q)}} \leqslant 4.1\sqrt{n},$$

achieved for $q \approx 0.71$. Finally, the algorithmic statement follows from using the constructive version of Lemma 11.2. □

11.2. Cancellations in regularization bounds

The best known asymptotic lower bound on the constant in Spencer's problem, as $n \to \infty$, is 1, achieved by Hadamard matrices (and by random matrices). In general, we do not

expect the analysis in §11.1 to be tight. In particular, the use of $\ell_{0.71}$-regularization seems somewhat ad hoc, and the analysis disregards information contained in the first-order term. This section provides evidence that our algorithm may achieve a better constant than we can currently prove. We demonstrate this by revisiting a toy problem: *learning with expert advice*. We show that the analysis in §11.1 does not recover the optimal guarantees for this problem, and we explain how to remedy this.

Learning with expert advice

We provide a short background on learning with expert advice, and refer to [KP17, Chapter 18] for additional context.

Consider the following game played over rounds $t = 1, \dots, T$. There are d *experts* giving predictions, and one player using the predictions of the experts. At the start of every round t, the player picks a distribution $\boldsymbol{r}_t \in \Delta_d$ over the experts $\{1, \dots, d\}$. After that, the gain $\ell_t(i) \in [0, \varepsilon]$ incurred by the prediction of each expert $i \in [d]$ is revealed. The choice of the player yields a reward $\langle \boldsymbol{r}_t, \boldsymbol{\ell}_t \rangle$.

The goal is to find a strategy to pick sequentially $\boldsymbol{r}_1, \dots, \boldsymbol{r}_t$ that minimizes the *regret*:

$$R_T \coloneqq \max_{i \in [d]} \sum_{t=1}^{T} \ell_t(i) - \sum_{t=1}^{T} \langle \boldsymbol{r}_t, \boldsymbol{\ell}_t \rangle \ .$$

To mirror the regret minimization interpretation of our algorithm (§9.2.3), we will assume the loss magnitude ε to be very small. One can think of learning with expert advice as an adversarial setting where losses are arbitrary bounded vectors.

Regret analysis from the Hessian

The optimal strategy for the player turns out to be a regularized strategy, i.e., to choose in advance a regularizer $\omega \colon \mathbb{R} \to \mathbb{R}_{\geqslant 0}$ and a parameter $\eta > 0$, and pick

$$\boldsymbol{r}_t \coloneqq \operatorname*{arg\,max}_{\boldsymbol{r} \in \Delta_m} \langle \boldsymbol{r}, \boldsymbol{L}_{t-1} \rangle + \frac{1}{\eta} \sum_{i=1}^{m} \omega(r_i) \,,$$

where we denote by $\boldsymbol{L}_t \coloneqq \sum_{s=1}^{t} \boldsymbol{\ell}_s$ the cumulative gains of each expert up to time t. Similarly to the discrepancy setting, the analysis tracks the potential function

$$\Phi_t \coloneqq \omega^* \left(\boldsymbol{L}_t \right) , \qquad \text{where } \omega^*(\boldsymbol{L}) \coloneqq \max_{\boldsymbol{r} \in \Delta_m} \langle \boldsymbol{r}, \boldsymbol{L} \rangle + \frac{1}{\eta} \sum_{i=1}^{m} \omega(r_i) \,.$$

In particular, we have the exact relation $\boldsymbol{r}_{t+1} = \nabla \omega^* \left(\boldsymbol{L}_t \right)$. Also, by construction, $\Phi_0 = \omega^*(\boldsymbol{0})$ and $\Phi_T \geqslant \max_{i \in [m]} L_t(i)$ always upper bounds the gain of the best expert. Applying this

upper bound and re-expressing as a function of ω^*, we obtain

$$\begin{aligned} R_T &\leqslant \Phi_T - \sum_{t=1}^{T} \langle \boldsymbol{r}_t, \boldsymbol{\ell}_t \rangle \\ &= \Phi_0 + \sum_{t=1}^{T} (\Phi_t - \Phi_{t-1}) - \sum_{t=1}^{T} \langle \boldsymbol{r}_t, \boldsymbol{\ell}_t \rangle \\ &= \omega^*(\mathbf{0}) + \sum_{t=1}^{T} \left[\omega^*(\boldsymbol{L}_{t-1} + \boldsymbol{\ell}_t) - \omega^*(\boldsymbol{L}_{t-1}) - \langle \nabla \omega^*(\boldsymbol{L}_{t-1}), \boldsymbol{\ell}_t \rangle \right] . \end{aligned}$$

When the gains are bounded by $[0, \varepsilon]$, as $\varepsilon \to 0$ this reduces to bounding the contribution from the Hessian term:

$$R_T \leqslant \omega^*(\mathbf{0}) + \frac{1}{2} \sum_{t=1}^{T} \left\langle \boldsymbol{\ell}_t, \nabla^2 \omega^*(\boldsymbol{L}_{t-1}) \boldsymbol{\ell}_t \right\rangle .$$

Tight constant for negative entropy regularization

We now specialize the previous analysis to the case $\omega(r) = -r \log r$, which makes the algorithm equivalent to the multiplicative weights update method. As in Lemma 9.12, a direct computation yields

$$\nabla \omega^*(\boldsymbol{L}) = \frac{\exp(\eta \boldsymbol{L})}{\sum_{i=1}^{d} \exp(\eta L(i))} , \qquad \nabla^2 \omega^*(\boldsymbol{L}) = \operatorname{diag}(\nabla \omega^*(\boldsymbol{L})) - \nabla \omega^*(\boldsymbol{L}) \nabla \omega^*(\boldsymbol{L})^\top .$$

However, to get the conclusion of Lemma 9.12, we ignore the subtracted rank-one term and simply upper bound it by 0. Moreover, a similar rank-1 term is removed in our analysis of ℓ_q-regularization (this appears explicitly in the proof of Lemma 9.12). Applying this coarse upper bound in the learning with expert advice setting would give

$$\left\langle \boldsymbol{\ell}_t, \nabla^2 \omega^*(\boldsymbol{L}_{t-1}) \boldsymbol{\ell}_t \right\rangle \leqslant \frac{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i)) \ell_t(i)^2}{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i))} = \mathop{\mathbb{E}}_{i \sim \nabla \omega^*(\boldsymbol{L}_{t-1})} \ell_t(i)^2 . \tag{11.2}$$

On the other hand, the second-order term actually *equals*

$$\left\langle \boldsymbol{\ell}_t, \nabla^2 \omega^*(\boldsymbol{L}_{t-1}) \boldsymbol{\ell}_t \right\rangle = \frac{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i)) \ell_t(i)^2}{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i))} - \left(\frac{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i)) \ell_t(i)}{\sum_{i=1}^{d} \exp(\eta L_{t-1}(i))} \right)^2 \tag{11.3}$$

$$= \mathop{\operatorname{Var}}_{i \sim \nabla \omega^*(\boldsymbol{L}_{t-1})} \ell_t(i) . \tag{11.4}$$

This makes a key difference: while the second moment of a distribution supported on $[0, \varepsilon]$ can be as large as ε^2 (as in (11.2)), its variance is always bounded by $\varepsilon^2/4$ (as in (11.4)). As a

result, the first argument gives a regret of $\sqrt{2T \log d}$, while the second gives $\sqrt{T \log d/2}$. This last bound turns out to be optimal for this problem [KP17, Proposition 18.3.8].

Although we do not currently know how to exploit such cancellations for ℓ_q-regularization, this suggests that the bound in §11.1 may be overly pessimistic.

11.3. Amortized analysis for ellipsoid discrepancy

In this section, we show that improved discrepancy bounds can arise from performing an *amortized* analysis of the gradient updates along the walk. We analyze a simpler geometric variant of Komlós problem, which we call *ellipsoid discrepancy*. This problem was originally studied by Banaszczyk [Ban90].

Definition 11.4. Let $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ be a positive definite matrix, Define for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{B}} \coloneqq \langle \boldsymbol{x}, \boldsymbol{B}\boldsymbol{y} \rangle \ , \quad \|\boldsymbol{x}\|_{\boldsymbol{B}}^2 \coloneqq \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{B}} \ .$$

We are interested in the following "Euclidean" version of the Komlós problem: Given a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ with columns of ℓ_2-norm at most 1,

$$\min_{\boldsymbol{x} \in \{-1,1\}^n} \|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} \, .$$

Banaszczyk [Ban90] proves that there always exist colorings with ellipsoid discrepancy $\sqrt{\operatorname{tr} \boldsymbol{B}}$. This is tight for any $\boldsymbol{B}$, as can be seen by taking the columns of $\boldsymbol{A}$ to be an orthonormal basis of eigenvectors of $\boldsymbol{B}$. Moreover, it follows implicitly from prior work, including the Gram-Schmidt walk algorithm [BDGL19], that this bound can be matched algorithmically up to constant factors.[1]

We give a different algorithmic proof, based on our iterative meta-algorithm Algorithm 4, that highlights the necessity of performing an amortized analysis in some cases.

Theorem 11.5. *There is deterministic, polynomial-time algorithm that given $B \succ 0$ and $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ whose columns have ℓ_2-norm at most 1, outputs $\boldsymbol{x} \in \{-1, 1\}^n$ such that*

$$\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} \lesssim \sqrt{\operatorname{tr} \boldsymbol{B}} \, .$$

Our proof of Theorem 11.5 uses the iterative machinery described in §9.3.1. Since $\|\cdot\|_{\boldsymbol{B}}^2$ is already a smooth degree-2 polynomial, no regularization is needed here. However, simply repeating the analysis in §11.1 would incur an extra logarithmic factor.

[1] We thank the anonymous SODA reviewers for pointing this out.

Proof. First, apply a rotation to reduce to the case where $\boldsymbol{B}$ is diagonal. This does not affect the norms of the columns of $\boldsymbol{A}$. Let $b_1 \geqslant \ldots \geqslant b_d \geqslant 0$ be the diagonal elements of $\boldsymbol{B}$.

As usual, we run a sticky walk $\boldsymbol{x} = \boldsymbol{x}(t) \in [-1, 1]^n$. When we make an update $\boldsymbol{\delta}$, the increase in discrepancy is

$$\|\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{\delta})\|_{\boldsymbol{B}}^2 - \|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}}^2 = \|\boldsymbol{A}\boldsymbol{\delta}\|_{\boldsymbol{B}}^2 + 2\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{\delta}\rangle_{\boldsymbol{B}} \,. \tag{11.5}$$

As before, we pick a signing $\pm\boldsymbol{\delta}$ that makes the linear term non-positive, and otherwise pick the update that minimizes $\|\boldsymbol{A}\boldsymbol{\delta}\|_{\boldsymbol{B}}^2$ over all $\boldsymbol{\delta}$ supported on active coordinates.

Suppose that the set of active coordinates is F, and let $k := |F|$. Since $\|\boldsymbol{A}\boldsymbol{\delta}\|_{\boldsymbol{B}}^2 = \langle \boldsymbol{\delta}, \boldsymbol{A}^\top \boldsymbol{B}\boldsymbol{A}\boldsymbol{\delta}\rangle$, it suffices to prove that $(\boldsymbol{A}^F)^\top \boldsymbol{B}\boldsymbol{A}^F$ has a small eigenvalue, where $\boldsymbol{A}^F$ is the matrix $\boldsymbol{A}$ restricted to the columns in F. A standard Gaussian δ in the $(k/2)$-dimensional subspace orthogonal to the rows $\boldsymbol{A}_1^F, \ldots, \boldsymbol{A}_{k/2}^F$ has ℓ_2-norm $\approx \sqrt{k/2}$ and expected quadratic form at most

$$b_{k/2} \sum_{i=1}^{d} \left\|\boldsymbol{A}_i^F\right\|_2^2 \leqslant k b_{k/2} \,.$$

This implies that $(\boldsymbol{A}^F)^\top \boldsymbol{B}\boldsymbol{A}^F$ has an eigenvalue smaller than $O(k b_{k/2})$.

We now need to bound the increase in discrepancy across the entire algorithm. Note that our bound depends on $b_1, \ldots, b_d$, which play the role of the gradients in §11.1. Here we leverage the fact that those gradients stay constant over the duration of the algorithm to provide an amortized analysis.

Let β_k denote the total ℓ_2-squared norm injected into $\boldsymbol{x}(t)$ between times $t_k = \min\{t : |F(t)| \leqslant k\}$ and $T = \min\{t : |F(t)| \leqslant 3\}$. In particular, we have $\beta_k \leqslant k$. We integrate the second-order increase in (11.5) via summation by parts:

$$\begin{aligned}
\sum_{k=2}^{\lfloor d/2 \rfloor} b_k(\beta_{2k+1} - \beta_{2k-1}) &= \sum_{k=3}^{\lfloor d/2 \rfloor} \beta_{2k-1}(b_{k-1} - b_k) + b_{\lfloor d/2 \rfloor}\beta_d \\
&\leqslant \sum_{k=3}^{\lfloor d/2 \rfloor} (2k-1)(b_{k-1} - b_k) + d b_{\lfloor d/2 \rfloor} \\
&\lesssim \operatorname{tr} \boldsymbol{B} \,.
\end{aligned}$$

Since the final step of Algorithm 4 only changes the squared discrepancy by $O(\operatorname{tr} \boldsymbol{B})$, the final coloring $\boldsymbol{x}^*$ satisfies

$$\|\boldsymbol{A}\boldsymbol{x}^*\|_{\boldsymbol{B}}^2 \lesssim \operatorname{tr} \boldsymbol{B} \,.$$

□

11.4. On the optimal constant for graph sparsification

In this section, we explore potential applications of our framework to improving graph sparsification bounds.

We start by recalling the notion of *spectral sparsification* that generalizes cut sparsification (Problem 1.3) [ST11].

Definition 11.6. Let $G = (V, E)$ be a graph with edge weights $w : E \to \mathbb{R}_{\geqslant 0}$. For any edge $e = \{u, v\} \in E$, let $\boldsymbol{L}_e = w_e(\boldsymbol{e}_u - \boldsymbol{e}_v)(\boldsymbol{e}_u - \boldsymbol{e}_v)^\top$ be the Laplacian matrix of e, and $\boldsymbol{L}_G = \sum_{e \in E} \boldsymbol{L}_e$ be the Laplacian matrix of G.

We say that a weighted subgraph H of G is a *spectral sparsifier* of G with error $\varepsilon > 0$ if

$$(1 - \varepsilon)\, \boldsymbol{L}_G \preceq \boldsymbol{L}_H \preceq (1 + \varepsilon)\, \boldsymbol{L}_G\,. \tag{11.6}$$

When $\boldsymbol{x}$ is a $\{0, 1\}$-valued indicator vector of a cut, the quadratic form $\langle \boldsymbol{x}, \boldsymbol{L}_G \boldsymbol{x} \rangle$ is the number of edges of G crossing the cut. This implies that any spectral sparsifier is also a cut sparsifier.

It is believed that the hardest graph to sparsify (in terms of tradeoff between error and number of edges) is the (unweighted) clique. The seminal result of Alon and Boppana [Nil91] shows that any *unweighted* spectral sparsifier of the clique with error ε must have at least $(2 - o(1))n/\varepsilon^2$ edges.[2] Srivastava and Trevisan conjecture that this lower bound still holds for (weighted) sparsifiers of the clique [ST18]. Batson, Spielman, and Srivastava [BSS14] show that any graph has a spectral sparsifier with $(4 + o(1))n/\varepsilon^2$ edges. Finally, Chen, Shi, and Trevisan [CST22] show that the clique has a *cut sparsifier* with $(4/\pi + o(1))n/\varepsilon^2$ edges. This leaves a gap of 2 between the best-known upper and lower bound for spectral sparsification.

The algorithm in Batson, Spielman, and Srivastava implicitly solves a harder online problem. Srivastava and Trevisan [ST18] show that for this harder problem, the constant 4 in the guarantees of [BSS14] is the best one can hope for. This raises the question of whether algorithms inspired by [BSS14] but that avoid solving this harder problem might offer a path to improving this constant.

One such candidate algorithm arises from the reduction from graph sparsification to discrepancy theory described in §1.1.3. Indeed, using this reduction, Reis and Rothvoss [RR20] showed that the construction of linear-size sparsifiers follows from a partial coloring lemma:

Lemma 11.7. *For any symmetric matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^{n \times n}$ such that $\sum_{i=1}^m \boldsymbol{A}_i = \boldsymbol{I}_n$, there exists $\boldsymbol{x} \in [-1, 1]^n$ that has $\Omega(m)$ coordinates equal to ± 1, and*

$$\left\| \sum_{i=1}^m x_i \boldsymbol{A}_i \right\|_2 \lesssim \sqrt{\frac{n}{m}}\,.$$

The connection between Definition 11.6 and Lemma 11.7 is clarified by the following observation. Let $\widetilde{\boldsymbol{L}}_e = \boldsymbol{L}_G^{\dagger/2} \boldsymbol{L}_e \boldsymbol{L}_G^{\dagger/2}$ denote the normalized Laplacian of edge e, where $\dagger$

[2] In this section, we use $o(1)$ to denote an error term going to 0 in the double limit $n \to \infty$, then $\varepsilon \to 0$.

is the Moore–Penrose pseudoinverse. Then, $\sum_{e\in E} \widetilde{\boldsymbol{L}}_e = \boldsymbol{I}_n$, and in this notation, (11.6) is equivalent to

$$(1-\varepsilon)\,\boldsymbol{I}_n \preceq \boldsymbol{L}_G^{\dagger/2}\boldsymbol{L}_H\boldsymbol{L}_G^{\dagger/2} \preceq (1+\varepsilon)\,\boldsymbol{I}_n\,.$$

Building on the framework introduced in this thesis, Lau, Wang, and Zhou [LWZ25] showed how to prove Lemma 11.7 using ℓ_q-regularization. Given the discussion earlier in this chapter, one might hope that our framework could help break the long-standing sparsification barrier.

One issue is that our approach does not distinguish between sums of arbitrary matrices like in Lemma 11.7, and the case where these matrices arise as normalized Laplacians of graphs. However, some general matrices may be harder to sparsify than the clique itself:

Theorem 11.8. *Let $\varepsilon > 0$ be a constant. For any $n \ll m \ll n^2$, there exist vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ such that $(1-o(1))\boldsymbol{I}_n \preceq \sum_{i=1}^m \boldsymbol{u}_i\boldsymbol{u}_i^\top \preceq (1+o(1))\boldsymbol{I}_n$, but for any $S \subseteq [m]$ of size $|S| = n/\varepsilon^2$,*

$$\frac{\lambda_{\max}\left(\sum_{s\in S}\boldsymbol{u}_s\boldsymbol{u}_s^\top\right)}{\lambda_{\min}\left(\sum_{s\in S}\boldsymbol{u}_s\boldsymbol{u}_s^\top\right)} \geqslant 1 + (4-O(\varepsilon))\varepsilon\,.$$

Theorem 11.8 does not entirely rule out our strategy, because it only excludes the existence of unweighted sparsifiers. However, the situation seems similar to the prediction of Srivastava and Trevisan. We conjecture that allowing weights does not make sparsification of rank-one matrices any easier.

Proof. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_m$ be i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for some $n \ll m \ll n^2$. Note that $\mathbb{E}\,\frac{1}{m}\sum_i \boldsymbol{v}_i\boldsymbol{v}_i^\top = \boldsymbol{I}_n$.

Fix $S \subseteq [m]$ of size $|S| = n/\varepsilon^2$. Concentration bounds for Wishart matrices (see, e.g., [VMB07]) imply that

$$\Pr\left(\lambda_{\max}\left(\frac{1}{|S|}\sum_{s\in S}\boldsymbol{v}_s\boldsymbol{v}_s^\top\right) \leqslant (1+\varepsilon)^2 - \varepsilon^2\right) \leqslant e^{-n^2C(\varepsilon)}$$

and

$$\Pr\left(\lambda_{\min}\left(\frac{1}{|S|}\sum_{s\in S}\boldsymbol{v}_s\boldsymbol{v}_s^\top\right) \geqslant (1-\varepsilon)^2 + \varepsilon^2\right) \leqslant e^{-n^2C(\varepsilon)}\,,$$

for some constant $C(\varepsilon) > 0$ independent of n. Hence, with probability at least $1 - 2e^{-n^2C(\varepsilon)}$, we have

$$\frac{\lambda_{\max}\left(\sum_{s\in S}\boldsymbol{v}_s\boldsymbol{v}_s^\top\right)}{\lambda_{\min}\left(\sum_{s\in S}\boldsymbol{v}_s\boldsymbol{v}_s^\top\right)} \geqslant \frac{1+2\varepsilon}{1-2\varepsilon+2\varepsilon^2} \geqslant 1 + 4\varepsilon - O(\varepsilon^2)\,.$$

Then, we take a union bound over all $\binom{m}{n/\varepsilon^2} \leqslant 2^{O(n\log n/\varepsilon^2)}$ possible choices for S. Moreover, when $n \ll m \ll n^2$, we have

$$(1-o(1))\boldsymbol{I}_n \preceq \frac{1}{m}\sum_{i=1}^m \boldsymbol{v}_i\boldsymbol{v}_i^\top \preceq (1+o(1))\boldsymbol{I}_n\,,$$

with high probability, so by another union bound we satisfy both the condition number property and the normalization one, with high probability. □

11.5. Summary

In this final chapter, we showed that our proof of Spencer's theorem can be strengthened to yield a tighter constant, and we suggested several avenues for further improvement.

A key insight from the study of random polynomial optimization in Part I and Part II is that finding the "optimal regularizer" leads to the best-performing algorithm for the problem [Mon19, JSS25].

Does there exist a regularizer that achieves the optimal constant in Spencer's theorem?

We leave this as our closing open problem.

Bibliography

[AB10] Jean-Yves Audibert and Sébastien Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010. 177

[ABH+05] Sanjeev Arora, Eli Berger, Elad Hazan, Guy Kindler, and Muli Safra. On Non-Approximability for Quadratic Programs. In *Symposium on Foundations of Computer Science (FOCS)*, pages 206–215, 2005. 131

[AG11] Sanjeev Arora and Rong Ge. New Tools for Graph Coloring. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, volume 6845, pages 1–12, 2011. 22

[AGK04] Noga Alon, Gregory Gutin, and Michael Krivelevich. Algorithms with large domination ratio. *Journal of Algorithms*, 50(1):118–131, 2004. 104, 134, 153

[AGKM23] Omar Alrabiah, Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar. A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation. In *Symposium on Theory of Computing (STOC)*, pages 1438–1448, 2023. 17

[AIM14] Scott Aaronson, Russell Impagliazzo, and Dana Moshkovitz. AM with multiple Merlins. In *Computational Complexity Conference (CCC)*, pages 44–55, 2014. 18, 23

[AIS19] Dimitris Achlioptas, Fotis Iliopoulos, and Alistair Sinclair. Beyond the Lovász Local Lemma: Point to Set Correlations and Their Algorithmic Applications. In *Symposium on Foundations of Computer Science (FOCS)*, pages 725–744, 2019. 203

[ALO15] Zeyuan Allen-Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral Sparsification and Regret Minimization Beyond Matrix Multiplicative Upyears. In *Symposium on Theory of Computing (STOC)*, pages 237–245, 2015. 177, 183

[AMP20] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: Norm bounds and applications. *arXiv preprint arXiv:1604.03423*, 2020. 57

[AN06] Noga Alon and Assaf Naor. Approximating the Cut-Norm via Grothendieck's Inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006. 22, 133

[AOW15] Sarah R. Allen, Ryan O'Donnell, and David Witmer. How to refute a random CSP. In *Symposium on Foundations of Computer Science (FOCS)*, pages 689–708, 2015. 29, 112

[Ban90] Wojciech Banaszczyk. A Beck-Fiala-type Theorem for Euclidean Norms. *European Journal of Combinatorics*, 11(6):497–500, 1990. 213

[Ban98] Wojciech Banaszczyk. Balancing Vectors and Gaussian Measures of n-dimensional Convex Bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998. 17, 33, 189, 190, 191

[Ban10] Nikhil Bansal. Constructive Algorithms for Discrepancy Minimization. In *Symposium on Foundations of Computer Science (FOCS)*, pages 3–10, 2010. 17, 31, 167

[Ban19] Nikhil Bansal. On a generalization of iterated and randomized rounding. In *Symposium on Theory of Computing (STOC)*, pages 1125–1135, 2019. 32

[BBH+12] Boaz Barak, Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Symposium on Theory of Computing (STOC)*, pages 307–326, 2012. 18, 20, 23

[BBP05] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, 33(5):1643–1697, 2005. 85

[BBvH23] Afonso S. Bandeira, March T. Boedihardjo, and Ramon van Handel. Matrix concentration inequalities and free probability. *Inventiones mathematicae*, 234:419–487, 2023. 185, 187

[BDG19] Nikhil Bansal, Daniel Dadush, and Shashwat Garg. An Algorithm for Komlós Conjecture Matching Banaszczyk's Bound. *SIAM Journal on Computing*, 48(2):534–553, 2019. 17, 33, 190, 191, 199

[BDGL19] Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The Gram-Schmidt Walk: A Cure for the Banaszczyk Blues. *Theory of Computing*, 15:1–27, 2019. 17, 190, 213

[Bel13] Adrian William Belshaw. *Strong Normality, Modular Normality, and Flat Polynomials: Applications of Probability in Number Theory and Analysis.* PhD thesis, Simon Fraser University, Department of Mathematics, 2013. 17, 208

[Ber99] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999. 176

[Ber01] Gregory Berkolaiko. Spectral Gap of Doubly Stochastic Matrices Generated from Equidistributed Unitary Matrices. *Journal of Physics A: Mathematical and General*, 34(22):L319, 2001. 201

[BF81] József Beck and Tibor Fiala. "Integer-making" Theorems. *Discrete Applied Mathematics*, 3(1):1–8, 1981. 190, 205

[BG17] Nikhil Bansal and Shashwat Garg. Algorithmic discrepancy beyond partial coloring. In *Symposium on Theory of Computing (STOC)*, pages 914–926, 2017. 17, 190

[BGG+17] Vijay Bhattiprolu, Mrinalkanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani. Weak Decoupling, Polynomial Folds and Approximate Optimization over the Sphere. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1008–1019, 2017. 22, 30, 31, 127, 151

[BGJ+25] Afonso S. Bandeira, Sivakanth Gopi, Haotian Jiang, Kevin Lucca, and Thomas Rothvoss. Tensor Concentration Inequalities: A Geometric Approach. In *Symposium on Theory of Computing (STOC)*, pages 822–832, 2025. 29, 113

[Bil95] Patrick Billingsley. *Probability and Measure.* John Wiley and Sons, Third edition, 1995. 231

[BJ25] Nikhil Bansal and Haotian Jiang. Decoupling via Affine Spectral-Independence: Beck-Fiala and Komlós Bounds Beyond Banaszczyk. *arXiv preprint arXiv:2508.03961*, 2025. 33

[BJM23] Nikhil Bansal, Haotian Jiang, and Raghu Meka. Resolving Matrix Spencer Conjecture

Up to Poly-logarithmic Rank. In *Symposium on Theory of Computing (STOC)*, pages 1814–1819, 2023. 17, 185

[BK96] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *Symposium on Theory of Computing (STOC)*, pages 47–55, 1996. 15

[BKMR25] Afonso S. Bandeira, Anastasia Kireeva, Antoine Maillard, and Almut Rödder. Randomstrasse101: Open Problems of 2024. *arXiv preprint arXiv:2504.20539*, 2025. 208

[BKS17] Boaz Barak, Pravesh K. Kothari, and David Steurer. Quantum entanglement, sum of squares, and the log rank conjecture. In *Symposium on Theory of Computing (STOC)*, pages 975–988, 2017. 18, 136

[BL06] Yonatan Bilu and Nathan Linial. Lifts, Discrepancy and Nearly Optimal Spectral Gap. *Combinatorica*, 26:495–519, 2006. 29, 114, 115

[BLV22] Nikhil Bansal, Aditi Laddha, and Santosh S. Vempala. A Unified Approach to Discrepancy Minimization. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, volume 245, pages 1:1–1:22, 2022. 17, 167, 170, 183

[BM20] Nikhil Bansal and Raghu Meka. On the Discrepancy of Random Low Degree Set Systems. *Random Structures & Algorithms*, 57(3):695–705, 2020. 17

[BMO+15] Boaz Barak, Ankur Moitra, Ryan O'Donnell, Prasad Raghavendra, Oded Regev, David Steurer, Luca Trevisan, Aravindan Vijayaraghavan, David Witmer, and John Wright. Beating the Random Assignment on Constraint Satisfaction Problems of Bounded Degree. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, volume 40, pages 110–123, 2015. 106, 108

[BN11] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011. 86

[Boe24] March T. Boedihardjo. Injective norm of random tensors with independent entries. *arXiv preprint arXiv:2412.21193*, 2024. 113

[Bog98] Vladimir I. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998. 134

[Bol14] Erwin Bolthausen. An Iterative Construction of Solutions of the TAP Equations for the Sherrington–Kirkpatrick Model. *Communications in Mathematical Physics*, 325(1):333–366, 2014. 27

[Bor19] Charles Bordenave. Lecture notes on random matrix theory, 2019. 43

[BR16] Jop Briët and Shravas Rao. Arithmetic expanders and deviation bounds for random tensors. *arXiv preprint arXiv:1610.03428*, 2016. 29

[BRR23] Rainie Bozzai, Victor Reis, and Thomas Rothvoss. The vector balancing constant for zonotopes. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1292–1300, 2023. 16

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding Semidefinite Programming Hierarchies via Global Correlation. In *Symposium on Foundations of Computer Science (FOCS)*, pages 472–481, 2011. 22, 31

[BS16] Boaz Barak and David Steurer. Proofs, beliefs, and algorithms through the lens of sum-of-squares. Available at https://www.sumofsquares.org/, 2016. 129

[BSS14] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan Sparsifiers. *SIAM Review*, 56(2):315–334, 2014. 16, 17, 32, 170, 181, 183, 215

[BSST13] Joshua Batson, Daniel A. Spielman, Nikhil Srivastava, and Shang-Hua Teng. Spectral sparsification of graphs: theory and algorithms. *Communications of the ACM*, 56(8):87–94, 2013. 15

[BY88] Zhidong Bai and Yuanqi Yin. Necessary and Sufficient Conditions for Almost Sure Convergence of the Largest Eigenvalue of a Wigner Matrix. *Annals of Probability*, 16:1729–1741, 1988. 14

[CH06] Philippe Carmona and Yueyun Hu. Universality in Sherrington–Kirkpatrick's spin glass model. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 42:215–222, 2006. 13

[CM25] Hong-Bin Chen and Jean-Christophe Mourrat. On the free energy of vector spin glasses with non-convex interactions. *Probability and Mathematical Physics*, 6:1–80, 2025. 14

[CMW20] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory (COLT)*, pages 1078–1141, 2020. 24, 42

[CR24] Collin Cademartori and Cynthia Rush. A Non-Asymptotic Analysis of Generalized Vector Approximate Message Passing Algorithms With Rotationally Invariant Designs. *IEEE Transactions on Information Theory*, 70(8):5811–5856, 2024. 79

[CST22] Antares Chen, Jonathan Shi, and Luca Trevisan. Cut Sparsification of the Clique Beyond the Ramanujan Bound: A Separation of Cut Versus Spectral Sparsification. In *Symposium on Discrete Algorithms (SODA)*, pages 3693–3731, 2022. 215

[CW04] Moses Charikar and Anthony Wirth. Maximizing Quadratic Programs: Extending Grothendieck's Inequality. In *Symposium on Foundations of Computer Science (FOCS)*, pages 54–60, 2004. 22, 131, 132

[DFKO06] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the fourier tails of bounded functions over the discrete cube. In *Symposium on Theory of Computing (STOC)*, pages 437–446, 2006. 105, 107

[DJR22] Daniel Dadush, Haotian Jiang, and Victor Reis. A New Framework for Matrix Discrepancy: Partial Coloring Bounds via Mirror Descent. In *Symposium on Theory of Computing (STOC)*, pages 649–658, 2022. 17, 181, 185

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009. 69

[DMS17] Amir Dembo, Andrea Montanari, and Subhabrata Sen. Extremal cuts of sparse random graphs. *Annals of Probability*, 45:1190–1217, 2017. 13

[DPP+18] Andrzej Dudek, Xavier Pérez-Giménez, Pawel Pralat, Hao Qi, Douglas B. West, and Xuding Zhu. Randomly Twisted Hypercubes. *European Journal in Combinatorics*, 70:364–373, 2018. 205

[dT23] Tommaso d'Orsi and Luca Trevisan. A Ihara-Bass Formula for Non-Boolean Matrices and Strong Refutations of Random CSPs. In *Computational Complexity Conference (CCC)*, volume 264, pages 27:1–27:16, 2023. 113

[EM20] Ahmed El Alaoui and Andrea Montanari. Algorithmic Thresholds in Mean Field Spin Glasses. *arXiv preprint arXiv:2009.11481*, 2020. 24

[EMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *Annals of Probability*, 49(6):2922–2960, 2021. 24, 27, 103

[ES18] Ronen Eldan and Mohit Singh. Efficient Algorithms for Discrepancy Minimization in Convex Sets. *Random Structures & Algorithms*, 53(2):289–307, 2018. 17, 167, 204

[EYY12] László Erdős, Horng-Tzer Yau, and Jun Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Advances in Mathematics*, 229(3):1435–1515, 2012. 94

[Fei02] Uriel Feige. Relations between average case complexity and approximation complexity. In *Symposium on Theory of Computing (STOC)*, pages 534–543, 2002. 112

[FK81] Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1:233–241, 1981. 14, 19

[FK08] Alan Frieze and Ravi Kannan. A new approach to the planted clique problem. In *Foundations of Software Technology and Theoretical Computer Science*, 2008. 18, 23

[FKP19] Noah Fleming, Pravesh K. Kothari, and Toniann Pitassi. Semialgebraic Proofs and Efficient Algorithm Design. *Foundations and Trends in Theoretical Computer Science*, 14(1-2):1–221, 2019. 129

[FKS89] Joel Friedman, Jeff Kahn, and Endre Szemerédi. On the Second Eigenvalue in Random Regular Graphs. In *Symposium on Theory of Computing (STOC)*, pages 587–598, 1989. 19, 29, 120

[FL06] Uriel Feige and Michael Langberg. The RPR^2 rounding technique for semidefinite programs. *Journal of Algorithms*, 60(1):1–23, 2006. 131

[FO05] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005. 19, 120

[FP07] Delphine Féral and Sandrine Péché. The Largest Eigenvalue of Rank One Deformation of Large Wigner Matrices. *Communications in Mathematical Physics*, 272:185–228, 2007. 86

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. 98

[FW95] Joel Friedman and Avi Wigderson. On the second eigenvalue of hypergraphs. *Combinatorica*, 15(1):43–65, 1995. 28, 29, 111

[Gia97] Apostolos Giannopoulos. On some Vector Balancing Problems. *Studia Mathematica*, 122(3):225–234, 1997. 167

[GJJ+20] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for Sherrington-Kirkpatrick via planted affine planes. In *Symposium on Foundations of Computer Science (FOCS)*, pages 954–965, 2020. 21

[GJKP25] Nicola Gorini, Chris Jones, Dmitriy Kunisky, and Lucas Pesenti. Polynomial Universality and Dynamics of General First-Order Methods for Random and Deterministic Matrices. *manuscript*, 2025. 28

[GK21] Alejandro Ginory and Jongwon Kim. Weingarten Calculus and the IntHaar Package for Integrals over Compact Matrix Groups. *Journal of Symbolic Computation*, 103:178–200, 2021. 201

[Glu89] Efim D. Gluskin. Extremal Properties of Orthogonal Parallelepipeds and their Applications to the Geometry of Banach Spaces. *Mathematics of the USSR-Sbornik*, 64(1):85–96, 1989. 167

[GS11] Venkatesan Guruswami and Ali Kemal Sinop. Lasserre Hierarchy, Higher Eigenvalues, and Approximation Schemes for Graph Partitioning and Quadratic Integer Programming with PSD Objectives. In *Symposium on Foundations of Computer Science (FOCS)*, pages 482–491, 2011. 22

[GS12] Venkatesan Guruswami and Ali Kemal Sinop. Faster SDP Hierarchy Solvers for Local Rounding Algorithms. In *Symposium on Foundations of Computer Science (FOCS)*, pages 197–206, 2012. 31

[GT02] Francesco Guerra and Fabio L. Toninelli. The Thermodynamic Limit in Mean Field Spin Glass Models. *Communications in Mathematical Physics*, 230:71–79, 2002. 14

[Gue03] Francesco Guerra. Broken Replica Symmetry Bounds in the Mean Field Spin Glass Model. *Communications in Mathematical Physics*, 233:1–12, 2003. 14

[GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995. 22, 130

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001. 12

[HGN+24] Mohammad Hashemi, Shengbo Gong, Juntong Ni, Wenqi Fan, B. Aditya Prakash, and Wei Jin. A comprehensive survey on graph reduction: sparsification, coarsening, and condensation. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2024. 15

[HKM23] Jun-Ting Hsieh, Pravesh K. Kothari, and Sidhanth Mohanty. A simple and sharper proof of the hypergraph Moore bound. In *Symposium on Discrete Algorithms (SODA)*, pages 2324–2344, 2023. 17

[HKPT24] Jun-Ting Hsieh, Pravesh K. Kothari, Lucas Pesenti, and Luca Trevisan. New SDP Roundings and Certifiable Approximation for Cubic Optimization. In *Symposium on Discrete Algorithms (SODA)*, pages 2337–2362, 2024. 20, 35, 128

[HLZ10] Simai He, Zhening Li, and Shuzhong Zhang. Approximation algorithms for homogeneous polynomial optimization with quadratic constraints. *Mathematical Programming*, 125:353–383, 2010. 20, 133

[HRS22] Samuel B. Hopkins, Prasad Raghavendra, and Abhishek Shetty. Matrix Discrepancy from Quantum Communication. In *Symposium on Theory of Computing (STOC)*, pages 637–648, 2022. 17, 185, 204

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory (COLT)*, volume 40, pages 956–1006, 2015. 20

[HV04] Johan Håstad and Srinivasan Venkatesh. On the advantage over a random assignment. *Random Structures & Algorithms*, 25(2):117–149, 2004. 12, 30, 153

[IS24] Misha Ivkov and Tselil Schramm. Semidefinite programs simulate approximate message passing robustly. In *Symposium on Theory of Computing (STOC)*, pages 348–357, 2024. 49, 62, 102

[Jan97] Svante Janson. *Gaussian Hilbert spaces*. Cambridge University Press, 1997. 231

[JLLS23] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying Sums of Norms. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1953–1962, 2023. 15

[JLLS24] Arun Jambulapati, James R. Lee, Yang P. Liu, and Aaron Sidford. Sparsifying Generalized Linear Models. In *Symposium on Theory of Computing (STOC)*, pages 1665–1675, 2024. 15

[JM20] Chris Jones and Matt McPartlon. Spherical Discrepancy Minimization and Algorithmic Lower Bounds for Covering the Sphere. In *Symposium on Discrete Algorithms (SODA)*, pages 874–891, 2020. 180

[Jof74] Anatole Joffe. On a Set of Almost Deterministic k-Independent Random Variables. *Annals of Probability*, 2(6):161–162, 1974. 143

[JP25] Chris Jones and Lucas Pesenti. Fourier Analysis of Iterative Algorithms. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 334, pages 102:1–102:21, 2025. 35, 41, 53, 65, 77

[JRT24] Arun Jambulapati, Victor Reis, and Kevin Tian. Linear-sized sparsifiers via near-linear time discrepancy theory. In *Symposium on Discrete Algorithms (SODA)*, pages 5169–5208, 2024. 16, 32

[JSS25] David Jekel, Juspreet Singh Sandhu, and Jonathan Shi. Potential Hessian Ascent: The Sherrington-Kirkpatrick Model. In *Symposium on Discrete Algorithms (SODA)*, pages 5307–5387, 2025. 31, 217

[JST23] Fernando Granha Jeronimo, Shashank Srivastava, and Madhur Tulsiani. List Decoding of Tanner and Expander Amplified Codes from Distance Certificates. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1682–1693, 2023. 22

[Kar72] Richard M. Karp. Reducibility Among Combinatorial Problems. *Complexity of Computer Computations*, pages 85–103, 1972. 11

[KB21] Dmitriy Kunisky and Afonso S. Bandeira. A tight degree 4 sum-of-squares lower bound for the Sherrington–Kirkpatrick Hamiltonian. *Mathematical Programming*, 190:721–759, 2021. 21

[KMW24] Dmitriy Kunisky, Cristopher Moore, and Alexander S. Wein. Tensor Cumulants for Statistical Inference on Invariant Distributions. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1007–1026, 2024. 24

[KN08] Subhash Khot and Assaf Naor. Linear Equations Modulo 2 and the L_1 Diameter of Convex Bodies. *SIAM Journal on Computing*, 38(4):1448, 2008. 21, 23, 31, 106, 133, 136

[KP17] Anna R. Karlin and Yuval Peres. *Game Theory, Alive*. American Mathematical Society, 2017. 211, 213

[KS05] Robert Kohn and Sylvia Serfaty. A deterministic-control-based approach motion by curvature. *Communications on Pure and Applied Mathematics*, 59(3), 2005. 171

[Kun21] Dmitriy Kunisky. *Spectral Barriers in Certification Problems*. PhD thesis, New York University, 2021. 21

[Kun23] Dmitriy Kunisky. The Discrepancy of Unsatisfiable Matrices and a Lower Bound for the Komlós Conjecture Constant. *SIAM Journal on Discrete Mathematics*, 37(2):586–603, 2023. 190

[KW92] Jacek Kuczyński and Henryk Wozniakowski. Estimating the Largest Eigenvalue by the Power and Lanczos Algorithms with a Random Start. *SIAM Journal on Matrix Analysis and Applications*, 13:993–1313, 1992. 101

[Lat24] Rafał Latała. On the spectral norm of Rademacher matrices. *arXiv preprint arXiv:2405.13656*, 2024. 114

[Lee23] James R. Lee. Spectral Hypergraph Sparsification via Chaining. In *Symposium on Theory of Computing (STOC)*, pages 207–218, 2023. 15, 29

[LFW23] Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to $\mathbb{Z}_2$-synchronization. *Proceedings of the National Academy of Sciences*, 120(31):e2302930120, 2023. 28, 79

[Lie73] Elliot H. Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973. 186

[LM15] Shachar Lovett and Raghu Meka. Constructive Discrepancy Minimization by Walking on the Edges. *SIAM Journal on Computing*, 44(5):1573–1582, 2015. 17, 31, 167, 191

[LRR17] Avi Levy, Harishchandra Ramadas, and Thomas Rothvoss. Deterministic Discrepancy Minimization via the Multiplicative Weight Update Method. In *Conference on Integer Programming and Combinatorial Optimization (IPCO)*, pages 380–391, 2017. 17, 167, 190, 193, 194, 196, 199

[LvHY18] Rafal Latala, Ramon van Handel, and Pierre Youssef. The dimension-free structure of nonhomogeneous random matrices. *Inventiones mathematicae*, 214:1031–1080, 2018. 29

[LW22] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiV preprint arXiv:2208.03313*, 2022. 28, 77

[LWZ25] Lap Chi Lau, Robert Wang, and Hong Zhou. Spectral Sparsification by Deterministic Discrepancy Walk. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 315–340, 2025. 16, 181, 216

[Mon90] Stephen Montgomery-Smith. The Distribution of Rademacher Sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990. 139

[Mon19] Andrea Montanari. Optimization of the Sherrington-Kirkpatrick Hamiltonian. In *Symposium on Foundations of Computer Science (FOCS)*, pages 1417–1433, 2019. 23, 24, 31, 42, 91, 102, 103, 217

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific, 1987. 66, 67

[MR16] Andrea Montanari and Emile Richard. Non-Negative Principal Component Analysis: Message Passing Algorithms and Sharp Asymptotics. *IEEE Transactions on Information Theory*, 62(3), 2016. 24, 27

[MRX20] Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Symposium on Theory of Computing (STOC)*, pages 840–853, 2020. 21

[MS65] Theodore S. Motzkin and Ernst G. Strauss. Maxima for Graphs and a New Proof of a Theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540, 1965. 20

[MSS15] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. *Annals of Mathematics*, 182:327–350, 2015. 17, 183

[MSS18] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing Families IV: Bipartite Ramanujan Graphs of All Sizes. *SIAM Journal on Computing*, 47(6):2488–2509, 2018. 170

[MSS22] Adam W. Marcus, Daniel A. Spielman, and Nikhil Srivastava. Finite free convolutions of polynomials. *Probability Theory and Related Fields*, 182:807–848, 2022. 170

[MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 49:321–345, 2021. 28, 77

[MV22] Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114003, 2022. 28, 77

[MW25] Andrea Montanari and Alexander S. Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *Probability Theory and Related Fields*, 191:181–233, 2025. 46, 62

[Nes03] Yurii Nesterov. Random walk in a simplex and quadratic optimization over convex polytopes. *preprint*, 2003. 19

[Nil91] A. Nilli. On the second eigenvalue of a graph. *Discrete Mathematics*, 91:207–210, 1991. 215

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. 134

[Par80] Giorgio Parisi. A sequence of approximated solutions to the S-K model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4), 1980. 14

[Par00] Pablo A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000. 18

[Pot20] Aditya Potukuchi. A Spectral Bound on Hypergraph Discrepancy. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 93:1–93:14, 2020. 190, 191, 192, 193

[PR22] Aaron Potechin and Goutham Rajendran. Sub-exponential time Sum-of-Squares lower

bounds for Principal Components Analysis. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 35724–35740, 2022. 20

[PV23] Lucas Pesenti and Adrian Vladu. Discrepancy Minimization via Regularization. In *Symposium on Discrete Algorithms (SODA)*, pages 1734–1758, 2023. 35, 166, 189, 207

[Rag08] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *Symposium on Theory of Computing (STOC)*, pages 245–254, 2008. 21

[RF12] Sundeep Rangan and Alyson K. Fletcher. Iterative estimation of constrained rank-one matrices in noise. In *International Symposium on Information Theory (ISIT)*, pages 1246–1250, 2012. 24, 43

[Rot17] Thomas Rothvoss. Constructive Discrepancy Minimization for Convex Sets. *SIAM Journal on Computing*, 46(1):224–234, 2017. 17, 167

[RR20] Victor Reis and Thomas Rothvoss. Linear size sparsifier and the geometry of the operator norm ball. In *Symposium on Discrete Algorithms (SODA)*, pages 2337–2348, 2020. 16, 215

[RR22] Victor Reis and Thomas Rothvoss. Approximate Carathéodory bounds via Discrepancy Theory. *arXiv preprint arXiv:2207.03614*, 2022. 16

[RT12] Prasad Raghavendra and Ning Tan. Approximating CSPs with global cardinality constraints using SDP hierarchies. In *Symposium on Discrete Algorithms (SODA)*, pages 373–387, 2012. 22

[Rud99] Mark Rudelson. Random Vectors in the Isotropic Position. *Journal of Functional Analysis*, 164:60–72, 1999. 29

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018. 28, 79

[Sel24] Mark Sellke. Optimizing mean field spin glasses with external field. *Electronic Journal of Probability*, 29:1–47, 2024. 24

[SK75] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975. 13

[Spe85] Joel Spencer. Six Standard Deviations Suffice. *Transactions of the American Mathematical Society*, 289(2):679–706, 1985. 17, 165, 167, 190, 208

[SS11] Daniel A. Spielman and Nikhil Srivastava. Graph Sparsification by Effective Resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. 15

[ST11] Daniel A. Spielman and Shang-Hua Teng. Spectral Sparsification of Graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011. 15, 215

[ST14] Daniel A. Spielman and Shang-Hua Teng. Nearly Linear Time Algorithms for Preconditioning and Solving Symmetric, Diagonally Dominant Linear Systems. *SIAM Journal on Matrix Analysis and Applications*, 35:835–885, 2014. 15

[ST18] Nikhil Srivastava and Luca Trevisan. An Alon-Boppana Type Bound for Weighted Graphs and Lowerbounds for Spectral Sparsification. In *Symposium on Discrete Algorithms (SODA)*, pages 1306–1315, 2018. 215

[ST21] David Steurer and Stefan Tiegel. SoS Degree Reduction with Applications to Clustering and Robust Moment Estimation. In *Symposium on Discrete Algorithms (SODA)*, pages 374–393, 2021. 31

[Sub20] Eliran Subag. Following the Ground States of Full-RSB Spherical Spin Glasses. *Communications on Pure and Applied Mathematics*, 74(5):1021–1044, 2020. 24, 31, 32

[Tal06] Michel Talagrand. The Parisi formula. *Annals of Mathematics*, 163:221–263, 2006. 14

[Tal21] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes : Decomposition Theorems*, volume 60. Springer, second edition, 2021. 14, 29, 113, 114

[Tao12] Terence Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012. 202

[TAP77] David J. Thouless, Philip W. Anderson, and Robert G. Palmer. Solution of 'Solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977. 67

[Tre17a] Luca Trevisan. Lecture Notes on Beyond Worst-Case Analysis. Available at https://lucatrevisan.github.io/teaching/bwca17/index.html, 2017. 19, 111

[Tre17b] Luca Trevisan. Lecture Notes on Graph Partitioning, Expanders and Spectral Methods. Available at https://lucatrevisan.github.io/books/expanders-2016.pdf, 2017. 92

[Tre19] Luca Trevisan. Online Optimization for Complexity Theorists. Available at https://lucatrevisan.wordpress.com/2019/04/17/online-optimization-for-complexity-theorists/, 2019. 171

[Tro20] Joel A. Tropp. *Randomized Algorithms for Matrix Computations*. Caltech CMS Lecture Notes, 2020. 100, 101

[Ver18] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. 113

[VMB07] Pierpaolo Vivo, Satya N. Majumdar, and Oriol Bohigas. Large deviations of the maximum eigenvalue in Wishart random matrices. *Journal of Physics A: Mathematical and Theoretical*, 40(16):4317, 2007. 216

[Voi91] Dan Voiculescu. Limit laws for Random matrices and free products. *Inventiones mathematicae*, 104:201–220, 1991. 28

[Wea04] Nik Weaver. The Kadison–Singer problem in discrepancy theory. *Discrete Mathematics*, 278:227–239, 2004. 183

[Wei25] Alexander S. Wein. Computational Complexity of Statistics: New Insights from Low-Degree Polynomials. *arXiv preprint arXiv:2506:10748*, 2025. 24

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. 69

[Zou12] Anastasios Zouzias. A Matrix Hyperbolic Cosine Algorithm and Applications. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 846–858, 2012. 181

APPENDIX A.

Additional material on the Fourier diagram basis

A.1. Gaussian distribution and combinatorics

We can derive convergence in distribution of random vectors by computing their moments.

Lemma A.1 (Method of moments [Bil95, Theorems 29.4, 30.1, and 30.2])**.** *Let $\boldsymbol{X}_n \in \mathbb{R}^d$ for $n \in \mathbb{N}$ and $\boldsymbol{Z} \in \mathbb{R}^d$ be random vectors such that for any $q_1, \dots, q_d \in \mathbb{N}$,*

$$\mathbb{E}\left[\prod_{i=1}^{d} X_{n,i}^{q_i}\right] \underset{n\to\infty}{\longrightarrow} \mathbb{E}\left[\prod_{i=1}^{d} Z_i^{q_i}\right].$$

Suppose that for all $i \in [d]$, Z_i has a Gaussian distribution. Then $X_n \xrightarrow{\mathrm{d}} Z$.

The Gaussian distribution and their orthogonal polynomials (the Hermite polynomials) have combinatorial interpretations related to matchings.

Lemma A.2. *Let $\mathcal{M}_{\text{perfect}}(q)$ be the set of perfect matchings on q objects. Then,*

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,\sigma^2)}[Z^q] = \left|\mathcal{M}_{\text{perfect}}(q)\right| \sigma^q = \begin{cases} \frac{q!\sigma^q}{2^{q/2}(q/2)!} & \text{if } q \text{ is even} \\ 0 & \text{if } q \text{ is odd} \end{cases}$$

Lemma A.3 ([Jan97, Theorem 3.4 and Example 3.18])**.** *For all $q \geqslant 0$ and $x \in \mathbb{R}$,*

$$h_q(x;\sigma^2) = \sum_{M\in\mathcal{M}(q)} (-1)^{|M|}\sigma^{2|M|}x^{q-2|M|},$$

where $\mathcal{M}(q)$ is the set of (partial) matchings on q objects (including the empty matching and perfect matchings).

Lemma A.4 ([Jan97, Theorem 3.15 and Example 3.18])**.** *For any $q_1, \dots, q_\ell \geqslant 0$ and $x \in \mathbb{R}$,*

$$h_{q_1}(x;\sigma^2)\cdots h_{q_\ell}(x;\sigma^2) = \sum_{M\in\mathcal{M}(q_1,\dots,q_\ell)} h_{q-2|M|}(x;\sigma^2)\sigma^{2|M|},$$

where $\mathcal{M}(q_1, \ldots, q_\ell)$ is the set of (partial) matchings on $q = q_1 + \cdots + q_\ell$ objects divided into ℓ blocks of sizes $q_1, \ldots, q_\ell$ such that no two elements from the same block are matched.

Finally, we recall:

Lemma A.5 (Gaussian integration by parts)**.** *Let $(Z_1, \ldots, Z_k)$ be a centered Gaussian vector. Then for all smooth $f : \mathbb{R}^k \to \mathbb{R}$,*

$$\mathbb{E}\left[Z_1 f(Z_1, \ldots, Z_k)\right] = \sum_{i=1}^{k} \mathbb{E}\left[Z_1 Z_i\right] \mathbb{E}\left[\frac{\partial f}{\partial z_i}(Z_1, \ldots, Z_k)\right].$$

A.2. Omitted Proofs

A.2.1. Removing hanging double edges

In order to implement the removal of hanging double edges, we introduce an additional diagrammatic construct to track the error, *2-labeled edges*. These terms are equal to zero when $\boldsymbol{A}$ is a Rademacher matrix and it is recommended to ignore them on a first read.

Definition A.6 (Edge-labeled diagram)**.** An edge-labeled diagram is a diagram in which some of the edges are labeled "2".

We let $E(\alpha)$ denote the entire multiset of labeled and unlabeled edges of α, $E_2(\alpha)$ the multiset of 2-labeled edges and $E_1(\alpha) = E \setminus E_2(\alpha)$ the multiset of non-labeled edges.

We use the convention that $|E(\alpha)|$ counts each 2-labeled edge twice, so that $|E(\alpha)|$ continues to equal the degree of the polynomial $Z_{\alpha,i}$.

Definition A.7 (Edge-labeled $\boldsymbol{Z}_\alpha$)**.** For an edge-labeled diagram α, we define $\boldsymbol{Z}_\alpha \in \mathbb{R}^n$ by

$$Z_{\alpha,i} = \sum_{\substack{\varphi : V(\alpha) \hookrightarrow [n] \\ \varphi(\circledcirc) = i}} \prod_{\{u,v\} \in E_1(\alpha)} A_{\varphi(u)\varphi(v)} \prod_{\{u,v\} \in E_2(\alpha)} \left(A^2_{\varphi(u)\varphi(v)} - \frac{1}{n}\right).$$

The set of diagrams $\mathcal{A}$ is extended to allow diagrams which may have 2-labeled edges. The definition of $I(\alpha)$ from Definition 3.4 must also be updated to incorporate labeled edges (because a labeled edge is mean-0, it is treated like a single edge).

Definition A.8 (Updated definition of $I(\alpha)$)**.** For a diagram $\alpha \in \mathcal{A}$, let $I(\alpha)$ be the subset of non-root vertices such that every edge incident to that vertex has multiplicity $\geqslant 2$ or is a self-loop, treating 2-labeled edges as if they were normal edges.

The following is an exact decomposition for removing hanging double edges.

Lemma A.9. *Let $\alpha \in \mathcal{A}$ be a diagram with a hanging (unlabeled) double edge. Let α_0 be α with both the hanging double edge and corresponding hanging vertex removed, and α_2 be α with the hanging double edge replaced by a single 2-labeled edge. Then,*

$$Z_\alpha = Z_{\alpha_0} - \frac{|V(\alpha)| - 1}{n} \cdot Z_{\alpha_0} + Z_{\alpha_2} .$$

Proof. We write:

$$\begin{aligned} Z_{\alpha,i} &= \sum_{\substack{\varphi \colon V(\alpha) \hookrightarrow [n] \\ \varphi(\circledcirc)=i}} A_{u,v}^2 \prod_{\{x,y\} \in E(\alpha) \setminus \{\{u,v\},\{u,v\}\}} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{1}{n} \sum_{\substack{\varphi \colon V(\alpha) \hookrightarrow [n] \\ \varphi(\circledcirc)=i}} \prod_{\{x,y\} \in E(\alpha) \setminus \{\{u,v\},\{u,v\}\}} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{n - |V(\alpha)| + 1}{n} \cdot Z_{\alpha_0,i} = Z_{\alpha_0,i} - \frac{|V(\alpha)| - 1}{n} Z_{\alpha_0,i} + Z_{\alpha_2,i} . \end{aligned}$$

The additional $n - |V(\alpha)| + 1$ scaling factor comes from removing the hanging vertex. □

A.2.2. Omitted proofs for §3.3

We prove a more specific version of Lemma 3.5.

Lemma A.10. *Let $q \in \mathbb{N}$, $\alpha \in \mathcal{A}$, and $i \in [n]$. Then,*

$$\left| \mathbb{E}\left[Z_{\alpha,i}^q \right] \right| \leqslant M_{q|E(\alpha)|} 2^{q|E(\alpha)|} (q\,|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{\frac{q}{2}(|V(\alpha)|-1-|E(\alpha)|+|I(\alpha)|)} ,$$

where M_k is a bound on the k-th moment of the entries of $\boldsymbol{A}$ (recall the notations of Assumption 2.1),

$$M_k = \max\left(\mathop{\mathbb{E}}_{X \sim \mu} \left[|X|^k \right], \mathop{\mathbb{E}}_{X \sim \mu_0} \left[|X|^k \right] \right) .$$

When q and $|V(\alpha)|$ are $O(1)$, the overall bound reduces to

$$\left| \mathbb{E}\left[Z_{\alpha,i}^q \right] \right| \leqslant O\left(n^{\frac{q}{2}(|V(\alpha)|-1-|E(\alpha)|+|I(\alpha)|)} \right) .$$

Proof. We expand $\mathbb{E}\left[Z_{\alpha,i}^q \right]$ as

$$\sum_{\substack{\varphi_1,\ldots,\varphi_q \colon V(\alpha) \hookrightarrow [n] \\ \varphi_1(\circledcirc)=\cdots=\varphi_q(\circledcirc)=i}} \mathbb{E}\left[\prod_{p=1}^{q} \left(\prod_{\{u,v\} \in E_1(\alpha)} A_{\varphi_p(u)\varphi_p(v)} \right) \left(\prod_{\{u,v\} \in E_2(\alpha)} \left(A_{\varphi_p(u)\varphi_p(v)}^2 - \frac{1}{n} \right) \right) \right] .$$

This is a polynomial of degree $q|E(\alpha)|$ in $\boldsymbol{A}$ (by convention every 2-labeled edge contributes 2 to $|E(\alpha)|$). We first estimate the magnitude of any summand of the sum over $\varphi_1, \ldots, \varphi_q$ with nonzero expectation. Each such summand can be decomposed into $2^{q|E_2(\alpha)|}$ terms by expanding out[1] the $A_{ij}^2 - \frac{1}{n}$. This leaves monomials in the entries of A of total degree at most $q|E(\alpha)|$. We bound the expected value of each of these monomials by $M_{q|E(\alpha)|} n^{-q|E(\alpha)|/2}$ using Hölder's inequality. This shows that any nonzero term in the summation has magnitude at most $2^{q|E_2(\alpha)|} M_{q|E(\alpha)|} n^{-q|E(\alpha)|/2}$.

To bound the number of nonzero terms, we observe that every edge A_{jk} for $j \neq k$ must occur zero times or at least twice in order to have nonzero expectation (the self-loops A_{jj} can occur any number of times, and the 2-labeled edges $A_{jk}^2 - \frac{1}{n}$ must overlap at least one additional edge in order to have nonzero expectation). Each vertex in $V(\alpha) \setminus I(\alpha) \setminus \{\circledcirc\}$ is incident to an edge of multiplicity 1 or a 2-labeled edge, and so it must occur in at least two embeddings in order for that edge A_{jk} to overlap and not make the expectation 0. This implies that the number of distinct non-root vertices among the embeddings is at most $q\left(|V(\alpha)| - 1 + |I(\alpha)|\right)/2$ where the -1 is used to avoid counting the root.

Hence, there are at most $n^{q(|V(\alpha)|-1+|I(\alpha)|)/2}$ ways to choose the entire image $\operatorname{img}(\varphi_1) \cup \ldots \cup \operatorname{img}(\varphi_q)$. Once this is fixed, there are at most $(q|V(\alpha)|)^{q|V(\alpha)|}$ q-tuples of embeddings that map to these vertices. We conclude by combining the bound on the number of nonzero terms and the bound on the magnitude of each of these terms. □

Proof of Lemma 3.8. By assumption, $x - y$ is a sum of combinatorially negligible terms. We first focus on a single one of them, say $a_n \boldsymbol{Z}_\alpha$. For any $\varepsilon > 0$, $q \in \mathbb{N}$ and $i \in [n]$, we have

$$\begin{aligned}
&\Pr\left(|a_n Z_{\alpha,i}| \geqslant \varepsilon\right) \\
&\leqslant \frac{\mathbb{E}\,|a_n Z_{\alpha,i}|^q}{\varepsilon^q} && \text{(Markov's inequality)} \\
&\leqslant \frac{1}{\varepsilon^q} M_{q|E(\alpha)|} 2^{q|E(\alpha)|} (q|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{-\frac{q}{2}} && \text{(Lemma A.10)} \\
&\leqslant \frac{1}{\varepsilon^q} (q|E(\alpha)|)^{O(q)} 2^{q|E(\alpha)|} (q|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{-\frac{q}{2}} && \text{(subgaussianity of } A_{ij}) \\
&= \exp\left(O(q \log q) - \frac{q}{2} \log n + q \log(1/\varepsilon)\right).
\end{aligned}$$

Picking $q = \log n$ and $\varepsilon = q^C n^{-1/2}$ and taking the constant C large enough we can make the probability an arbitrarily small inverse polynomial in n. Then we take a union bound over all $i \in [n]$ and all combinatorially negligible term appearing in $\boldsymbol{x} - \boldsymbol{y}$ (there are constantly many such terms by definition). □

Proof of Lemma 3.9. It suffices to prove that for a combinatorially negligible term $n^{-k} \boldsymbol{Z}_\alpha$:

[1] The factor $2^{q|E_2(\alpha)|}$ may be removed with a tighter argument.

1. All terms in the diagram representation of $n^{-k}\boldsymbol{A}\boldsymbol{Z}_\alpha$ are combinatorially negligible.
2. Let $n^{-\ell}\boldsymbol{Z}_\beta$ be any term of combinatorial order 1 or combinatorially negligible. Then all terms in the diagram representation of the componentwise product $n^{-(k+\ell)}\boldsymbol{Z}_\alpha \odot \boldsymbol{Z}_\beta$ are combinatorially negligible, where $\odot$ is the componentwise product.

For 1, the diagram representation of $\boldsymbol{A}\boldsymbol{Z}_\alpha$ is given by Lemma 2.11. In the term α^+ without intersections,

$$|V(\alpha^+)| = |V(\alpha)| + 1\,, \qquad |I(\alpha^+)| = |I(\alpha)|\,, \qquad |E(\alpha^+)| = |E(\alpha)| + 1\,.$$

From this we can check that $n^{-k}\boldsymbol{Z}_{\alpha^+}$ is still combinatorially negligible.

In a term β corresponding to an intersection between the new root and a vertex of α,

$$|V(\beta)| = |V(\alpha)|\,, \qquad |I(\beta)| \leqslant |I(\alpha)| + 1\,, \qquad |E(\beta)| = |E(\alpha)| + 1\,.$$

The second inequality follows from the observation that the only vertices from α whose neighborhood structure can be affected by the intersection are the root of α (which does not contribute to $|I(\alpha)|$) and the intersected vertex. Hence, $n^{-k}\boldsymbol{Z}_\beta$ is also combinatorially negligible.

For (ii), the diagram representation of $\boldsymbol{Z}_\alpha \odot \boldsymbol{Z}_\beta$ is given by Lemma 2.14. Fix an intersection pattern $P \in \mathcal{P}(\alpha, \beta)$ that has b blocks and denote by γ the resulting diagram. Then,

$$\begin{aligned} |V(\gamma)| &= b + 1\,, \\ |E(\gamma)| &= |E(\alpha)| + |E(\beta)|\,, \\ |I(\gamma)| &\leqslant |I(\alpha)| + |I(\beta)| + |V(\alpha)| + |V(\beta)| - b - 2\,. \end{aligned}$$

The last inequality is proven by observing that for a non-root vertex that is neither in $I(\alpha)$ nor $I(\beta)$ to contribute to $I(\gamma)$, it must intersect another vertex. Moreover, there are at most $|V(\alpha)| + |V(\beta)| - b - 2$ intersected non-root vertices in γ.

Putting everything together,

$$\begin{aligned} &|V(\gamma)| - 1 - |E(\gamma)| + |I(\gamma)| \\ &\leqslant |V(\alpha)| - 1 - |E(\alpha)| + |I(\alpha)| + |V(\beta)| - 1 - |E(\beta)| + |I(\beta)| \\ &< 2(k + \ell)\,, \end{aligned}$$

since $n^{-k}\boldsymbol{Z}_\alpha$ is combinatorially negligible and $n^{-\ell}\boldsymbol{Z}_\beta$ is at most order 1. This concludes the proof. □

Using the 2-labeled edges introduced in Appendix A.2.1, we can implement the removal of hanging double edges.

Proof of Lemma 3.10. Starting from the decomposition of Lemma A.9,

$$a_n Z_\alpha = a_n Z_{\alpha_0} - a_n \frac{|V(\alpha)| - 1}{n} Z_{\alpha_0} + a_n Z_{\alpha_2},$$

we claim that the first term is combinatorially order 1, and the second and third terms are combinatorially negligible. Comparing α_0 to α, two edges and one vertex in $I(\alpha)$ are removed. This does not change the combinatorial order. The second term scales down by n and this becomes negligible (by assumption $|V(\alpha)|$ is constant). In the third term, $|I(\alpha_2)| < |I(\alpha)|$ to take into account the hanging vertex, while $|V(\alpha)| = |V(\alpha_2)|$ and $|E(\alpha)| = |E(\alpha_2)|$ remain unchanged, making the term negligible. We remind the reader that $|E(\alpha)| = |E(\alpha_2)|$ because $|E(\alpha_2)|$ counts 2-labeled edges twice. □

Definition 3.6 includes the coefficient a_n in the definition in order to incorporate factors of $\frac{1}{n}$ on some error terms such as those in the proof above.

A.3. Scalar diagrams

We collect the properties of scalar diagrams (Definition 3.11) which naturally generalize those of vector diagrams. We omit the proofs of the results in this section, as they are direct modifications of their vector analogs.

First, the scalar diagrams are an orthogonal basis for scalar functions of A.

Lemma A.11. *For any proper $\alpha \in \mathcal{A}_{\text{scalar}}$:*

- *For any proper $\beta \in \mathcal{A}_{\text{scalar}}$ such that $\beta \neq \alpha$, $\mathbb{E}\left[Z_\alpha Z_\beta\right] = 0$.*
- *$\mathbb{E}\left[Z_\alpha\right] = 0$ if α is not a singleton.*
- *The second moment of Z_α is*

$$\begin{aligned}\mathbb{E}\left[Z_\alpha^2\right] &= |\mathrm{Aut}(\alpha)| \cdot \frac{n(n-1)\cdots(n-|V(\alpha)|+1)}{n^{|E(\alpha)|}} \\ &\underset{n\to\infty}{=} |\mathrm{Aut}(\alpha)| \cdot n^{|V(\alpha)|-|E(\alpha)|}(1+o(1)),\end{aligned}$$

where the last estimate holds whenever $|V(\alpha)| = o(\sqrt{n})$.

Proof. Analogous to Lemma 2.8 and Lemma 2.9. □

When scalar and vector diagrams are multiplied together, the result can be expressed in terms of diagrams by extending the notion of intersection patterns $\mathcal{P}(\alpha_1, \ldots, \alpha_k)$ (Definition 2.12) and intersection diagrams (Definition 2.13) to allow scalar and vector diagrams simultaneously. The "unintersected" diagram consists of adding all the scalar diagrams as floating components to the vector diagrams, which are put at the same root. The intersection patterns are partitions of this vertex set such that no two vertices from the same diagram are matched.

Lemma A.12. *Let $\alpha_1, \dots, \alpha_k$ be either scalar or vector diagrams. Then*

$$Z_{\alpha_1} \cdots Z_{\alpha_k} = \sum_{P \in \mathcal{P}(\alpha_1, \dots, \alpha_k)} Z_{\alpha_P},$$

where the product is componentwise for the vector diagrams.

Proof. Analogous to Lemma 2.14. □

We define $I(\alpha)$ for scalar diagrams exactly as in Definition 3.4.

Lemma A.13. *Let $q \in \mathbb{N}, \alpha \in \mathcal{A}_{\text{scalar}}$, and $i \in [n]$. Then,*

$$\left|\mathbb{E}\left[Z_\alpha^q\right]\right| \leqslant M_{q|E(\alpha)|} 2^{q|E(\alpha)|} (q\,|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{\frac{q}{2}(|V(\alpha)|-|E(\alpha)|+|I(\alpha)|)},$$

where M_k is defined as in Lemma A.10. When q and $|V(\alpha)|$ are $O(1)$, this reduces to

$$\left|\mathbb{E}\left[Z_\alpha^q\right]\right| \leqslant O\left(n^{\frac{q}{2}(|V(\alpha)|-|E(\alpha)|+|I(\alpha)|)}\right).$$

Proof. Analogous to Lemma A.10. □

Definition A.14 (Combinatorially negligible and order 1 scalar)**.** Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real-valued coefficients with $a_n = \Theta(n^{-k})$, where $k \geqslant 0$ is such that $2k \in \mathbb{Z}$. Let $\alpha \in \mathcal{A}_{\text{scalar}}$ be a scalar diagram.

- We say that $a_n Z_\alpha$ is *combinatorially negligible* if

$$|V(\alpha)| - |E(\alpha)| + |I(\alpha)| \leqslant 2k - 1.$$

- We say that $a_n Z_\alpha$ has *combinatorial order 1* if

$$|V(\alpha)| - |E(\alpha)| + |I(\alpha)| = 2k.$$

We define $\overset{\infty}{=}$ for scalar diagram expressions exactly as in Definition 3.7.

Lemma A.15. *Let x and y be scalar diagram expressions with $x \overset{\infty}{=} y$. Then $|x - y| \xrightarrow{\text{a.s.}} 0$.*

Proof. Analogous to Lemma 3.8. □

Lemma A.16. *Let $a_n Z_\alpha$ be a combinatorially negligible scalar term. Let $b_n Z_\beta$ be any scalar or vector term of combinatorial order at most 1. Then all terms in the product $a_n b_n Z_\alpha Z_\beta$ are combinatorially negligible.*

Proof. Analogous to Lemma 3.9. □

In Lemma 3.13, we characterized the connected vector diagrams which are combinatorially order 1. We now similarly characterize the order 1 scalar diagrams.

Lemma A.17. *Let $\alpha \in \mathcal{A}_{\text{scalar}}$ be a scalar diagram with c connected components, c_I of which contain only vertices in $I(\alpha)$. Then $n^{-(c+c_I)/2} Z_\alpha$ is combinatorially negligible or combinatorially order 1, and it is combinatorially order 1 if and only if the following conditions hold simultaneously:*

1. *Every multiedge has multiplicity 1 or 2.*
2. *There are no cycles.*
3. *In each component, the subgraph of multiplicity 1 edges is empty or a connected graph (i.e. the multiplicity 2 edges consist of hanging trees)*
4. *There are no self-loops or 2-labeled edges (Appendix A.2.1).*

Proof. We proceed as in the proof of Lemma 3.13. In each connected component C containing at least one vertex $s \in V(\alpha) \setminus I(\alpha)$, we run a breadth-first search from s, assigning the multiedges used to explore a vertex to that vertex. This assigns at least one edge to every vertex in $C \setminus \{s\}$, and at least two edges to every vertex in $I(\alpha) \cap C$. This encoding argument shows that

$$2\,|I(\alpha) \cap C| + |(V(\alpha) \setminus I(\alpha)) \cap C| - 1 \leqslant |E(C)|\,, \tag{A.1}$$

where $E(C)$ denotes the set of edges in the connected component C.

In each connected component C composed only of vertices in $I(\alpha)$, we run a breadth-first search from an arbitrary vertex, and obtain

$$2(|I(\alpha) \cap C| - 1) = |V(\alpha) \cap C| + |I(\alpha) \cap C| - 2 \leqslant |E(C)|\,. \tag{A.2}$$

Summing (A.1) and (A.2) over all connected components, we obtain

$$|V(\alpha)| - |E(\alpha)| + |I(\alpha)| \leqslant (c - c_I) + 2c_I = c + c_I\,.$$

This shows that $n^{-(c+c_I)/2} Z_\alpha$ is combinatorially negligible or combinatorially order 1, and it is combinatorially order 1 if and only if equality holds in the argument. This happens if and only if there is no cycle, multiplicity $\geqslant 2$ edges, self-loops, or 2-labeled edges anywhere; and if the graph induced by the multiplicity 1 multiedges is connected. □

With this result in hand, we can now characterize the order-1 vector diagrams with several connected components:

Corollary A.18. *Let $\alpha \in \mathcal{A}$ be a vector diagram with c floating components, c_I of which consist only of vertices in $I(\alpha)$. Then $n^{-(c+c_I)/2} \boldsymbol{Z}_\alpha$ is combinatorially order 1 if and only if both the floating components (viewed as one scalar diagram) scaled by $n^{-(c+c_I)/2}$ and the component of the root are combinatorially order 1.*

Proof. Definition 3.6 sums across the root and floating components, so we may apply both Lemma 3.13 and Lemma A.17. □

A.4. Proof of classification of diagrams

Lemma A.19. *For all $\sigma \in \mathcal{S}$ and $i \in [n]$, $Z_{\sigma,i} \xrightarrow{d} \mathcal{N}(0, |\mathrm{Aut}(\sigma)|)$. Similarly, for all $\tau \in \mathcal{T}_{\mathrm{scalar}}$, $n^{-\frac{1}{2}} Z_\tau \xrightarrow{d} \mathcal{N}(0, |\mathrm{Aut}(\tau)|)$.*

Proof. We show that the moments $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ match the Gaussian ones, and use Lemma A.1.

Let $q \in \mathbb{N}$ be a constant independent of n. First, we expand the product $Z_{\sigma,i}^q$ in the diagram basis using Lemma 2.14. Using Lemma 3.13, the only combinatorially order 1 terms occur when there are no cycles, all multiedges have multiplicity 1 or 2, and the multiplicity 2 edges form hanging trees. Any term with an edge of multiplicity 1 disappears when we take the expectation $\mathbb{E}\left[Z_{\sigma,i}^q\right]$, while the diagrams which are entirely hanging trees are equal to ◎ up to combinatorially negligible terms (Lemma 3.10). Further, ◎ has expectation 1, and by Lemma A.10 each of the combinatorially negligible terms has expectation $O(n^{-1/2})$. Thus, $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ equals the number of ways to create hanging trees of double edges, up to a term that converges to 0 as $n \to \infty$.

For each of the q copies of σ, the single edge incident to the root must be paired with another such edge. This extends to an automorphism of the entire subtree. In conclusion, $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ converges to $|\mathrm{Aut}(\sigma)|^{q/2}$ times the number of perfect matchings on q objects, and we conclude by Lemma A.2 and Lemma A.1. The proof for the scalar case is analogous. □

Lemma A.20. *If $\tau \in \mathcal{T}$ consists of d_σ copies of the subtrees $\sigma \in \mathcal{S}$, then*

$$Z_\tau \overset{\infty}{=} \prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(Z_\sigma; |\mathrm{Aut}(\sigma)|)\,.$$

For $\rho \in \mathcal{F}_{\mathrm{scalar}}$ with c components and consisting of d_τ copies of each tree $\tau \in \mathcal{T}_{\mathrm{scalar}}$,

$$n^{-\frac{c}{2}} Z_\rho \overset{\infty}{=} \prod_{\tau \in \mathcal{T}_{\mathrm{scalar}}} h_{d_\tau}\left(n^{-\frac{1}{2}} Z_\tau; |\mathrm{Aut}(\tau)|\right)\,.$$

Proof. We first expand $h_d(Z_\sigma; |\mathrm{Aut}(\sigma)|)$ in the diagram basis using Lemma 2.14 and identify the dominant terms, i.e. those which are combinatorially order 1. As in the proof of Lemma A.19, the combinatorially order 1 terms in each monomial $Z_{\sigma,i}^k$ consist of pairing up copies of the tree σ:

$$Z_\sigma^k \overset{\infty}{=} \sum_{M \in \mathcal{M}(k)} |\mathrm{Aut}(\sigma)|^{|M|}\, Z_{k-2|M| \text{ copies of } \sigma}\,,$$

where $\mathcal{M}(k)$ is the set of partial matchings on k objects. Now we use the combinatorial interpretation of Hermite polynomials (Lemma A.3),

$$\begin{aligned} h_d(Z_\sigma; |\mathrm{Aut}(\sigma)|) &= \sum_{N \in \mathcal{M}(d)} (-1)^{|N|} \, |\mathrm{Aut}(\sigma)|^{|N|} \, Z_\sigma^{d-2|N|} \\ &\overset{\infty}{=} \sum_{N \in \mathcal{M}(d)} (-1)^{|N|} \, |\mathrm{Aut}(\sigma)|^{|N|} \sum_{M \in \mathcal{M}(d-2|N|)} |\mathrm{Aut}(\sigma)|^{|M|} \, Z_{d-2|N|-2|M| \text{ copies of } \sigma} \\ &= \sum_{M' \in \mathcal{M}(d)} |\mathrm{Aut}(\sigma)|^{|M'|} \, Z_{d-2|M'| \text{ copies of } \sigma} \sum_{N \subseteq M'} (-1)^{|N|} \\ &= Z_{d \text{ copies of } \sigma} \, . \end{aligned}$$

This completes the argument when τ consists of several copies of a single $\sigma \in \mathcal{S}$. If $\sigma, \sigma' \in \mathcal{S}$ are distinct, using again Lemma 2.14 and Lemma 3.13, we can check that

$$Z_{d \text{ copies of } \sigma} \odot Z_{d' \text{ copies of } \sigma'} \overset{\infty}{=} Z_{d \text{ copies of } \sigma \text{ and } d' \text{ copies of } \sigma'} \, .$$

The proof then follows by applying these arguments inductively, and extends analogously to scalar diagrams. □

Lemma A.21. *Let $\alpha \in \mathcal{F}$ have c floating components. Let $\alpha_{\circledcirc}$ be the component of the root and α_{float} be the floating components. Then $n^{-\frac{c}{2}} Z_\alpha \overset{\infty}{=} n^{-\frac{c}{2}} Z_{\alpha_{\mathrm{float}}} Z_{\alpha_{\circledcirc}}$.*

Proof. The product $n^{-\frac{c}{2}} Z_{\alpha_{\mathrm{float}}} Z_{\alpha_{\circledcirc}}$ can be expanded in the diagram basis using Lemma A.12. We claim that the only non combinatorially negligible diagram is the one without intersections, which equals $n^{-\frac{c}{2}} Z_\alpha$. When an intersection occurs, it can only be between the root component and a floating component. The new component of the root is at most combinatorially order 1 (this is a property of all connected vector diagrams, Lemma 3.13), so there is an "extra" factor of $\frac{1}{\sqrt{n}}$ from the lost component which makes the intersection term negligible. □

Lemma A.22. $\left\{ Z_{\sigma,i} : \sigma \in \mathcal{S}, i \in [n] \right\} \cup \left\{ n^{-\frac{1}{2}} Z_\tau : \tau \in \mathcal{T}_{\mathrm{scalar}} \right\}$ *are asymptotically independent.*

Proof. Fix constants $q, r \in \mathbb{N}$. We proceed by computing the moment of a set of diagrams $\sigma_1, \ldots, \sigma_q \in \mathcal{S}$ rooted at $i_1, \ldots, i_q \in [n]$ and $\tau_1, \ldots, \tau_r \in \mathcal{T}_{\mathrm{scalar}}$:

$$\mathbb{E}\left[\prod_{p=1}^{q} Z_{\sigma_p, i_p} \prod_{p=1}^{r} n^{-\frac{1}{2}} Z_{\tau_p} \right] . \tag{A.3}$$

Let $|V| = \sum_{p=1}^{q} |V(\sigma_p)| + \sum_{p=1}^{r} |V(\tau_p)|$ and $|E| = \sum_{p=1}^{q} |E(\sigma_p)| + \sum_{p=1}^{r} |E(\tau_p)|$. Let q_{distinct} be the number of distinct roots, i.e. the number of distinct elements in $\{i_1, \ldots, i_q\}$.

Expanding (A.3) gives a sum over embeddings of the diagrams. We will prove that the dominant terms factor across the distinct (σ_p, i_p) and τ_p; they correspond to pairing up isomorphic σ_p at each distinct root and isomorphic τ_p.

Each nonzero term in the expansion of (A.3) equals $n^{-(|E|+r)/2}$ (when every edge appears exactly twice) or $O(n^{-(|E|+r)/2})$ (in general) by Assumption 2.1. We partition the summation based on the intersection pattern as in Definition 2.12. For a given intersection pattern, letting I be the union of the images of the embeddings, the number of terms with this pattern is $(1-o(1)) \cdot n^{|I|-q_{\text{distinct}}}$ because the q_{distinct} root vertices are fixed. In an embedding with nonzero expectation, every edge appears at least twice, so every non-root vertex is in at least two embeddings. Applying this bound to all of the non-root vertices in I,

$$|I| \leqslant q_{\text{distinct}} + \frac{|V|-q}{2}.$$

Multiplying the value of each term times the number of terms, the total contribution of this intersection pattern is

$$n^{|I|-q_{\text{distinct}}-\frac{|E|+r}{2}} \leqslant n^{\frac{1}{2}(|V|-q-|E|-r)}.$$

Since the individual diagrams are connected, the exponent is nonpositive. The dominant terms occur exactly when $|I| = q_{\text{distinct}} + (|V|-q)/2$, equivalently all of the non-root vertices intersect exactly one other non-root vertex. Each edge must occur at least twice, and this condition implies that each edge occurs exactly twice in the dominant terms.

We claim that the only way that each edge and vertex can be in exactly two embeddings is if isomorphic σ_p and τ_p are paired. Indeed, by connectivity of σ_p and τ_p, sharing one edge extends to an isomorphism. Furthermore, because non-root vertices must intersect other non-root vertices in the dominant terms, we have that no pairs can be made between σ_p and $\tau_{p'}$, or between σ_p and $\sigma_{p'}$ which have distinct roots. □

Theorem 3.14 follows from Lemma A.19, Lemma A.20, Lemma A.21, and Lemma A.22.

The constant-order joint moments of all the diagrams are summarized into the next theorem which generalizes Theorem 3.14.

Theorem A.23. *Suppose that $\boldsymbol{A} = \boldsymbol{A}(n)$ is a sequence of random matrices satisfying Assumption 2.1. For all $\alpha_1, \dots, \alpha_k \in \mathcal{A}$, $i_1, \dots, i_k \in [n]$ and $\beta_1, \dots, \beta_\ell \in \mathcal{A}_{\text{scalar}}$ (allowing repetitions anywhere),*

$$\mathbb{E}\left[\prod_{j=1}^{k} n^{-C(\alpha_j)/2} Z_{\alpha_j, i_j} \prod_{j=1}^{\ell} n^{-C(\beta_j)/2} Z_{\beta_j}\right] = \mathbb{E}\left[\prod_{j=1}^{k} Z^{\infty}_{\alpha_j, i_j} \prod_{j=1}^{\ell} Z^{\infty}_{\beta_j}\right] + O(n^{-\frac{1}{2}}),$$

where $C(\alpha)$ is the number of floating components of α, and where the asymptotic random variables $(Z^\infty_{\alpha,i})_{\alpha\in\mathcal{A},i\in[n]}$ and $(Z^\infty_\beta)_{\beta\in\mathcal{A}_{\text{scalar}}}$ are defined as:

$$\begin{cases} Z^\infty_{\sigma,i} \sim \mathcal{N}(0,|\mathrm{Aut}(\sigma)|) \text{ independently} & \text{if } \sigma\in\mathcal{S} \\ Z^\infty_{\tau} \sim \mathcal{N}(0,|\mathrm{Aut}(\tau)|) \text{ independently} & \text{if } \tau\in\mathcal{T}_{\text{scalar}} \\ Z^\infty_{\rho,i} = \prod_{\sigma\in\mathcal{S}} h_{d_\sigma}(Z^\infty_{\sigma,i};|\mathrm{Aut}(\sigma)|) \prod_{\tau\in\mathcal{T}_{\text{scalar}}} h_{d_\tau}(Z^\infty_\tau;|\mathrm{Aut}(\tau)|) & \text{if } \rho\in\mathcal{F} \\ Z^\infty_{\rho} = \prod_{\tau\in\mathcal{T}_{\text{scalar}}} h_{d_\tau}(Z^\infty_\tau;|\mathrm{Aut}(\tau)|) & \text{if } \rho\in\mathcal{F}_{\text{scalar}} \\ Z^\infty_{\alpha,i} = Z^\infty_{\alpha_0,i} \text{ and } Z^\infty_\beta = Z^\infty_{\beta_0} & \text{if removing hanging double edges creates } \alpha_0\in\mathcal{F} \text{ or } \beta_0\in\mathcal{F}_{\text{scalar}} \\ Z^\infty_{\alpha,i} = Z^\infty_\beta = 0 & \text{if removing hanging double edges is not in } \mathcal{F} \text{ or } \mathcal{F}_{\text{scalar}} \end{cases}$$

A.5. Handling empirical expectations

Empirical expectations are highly concentrated and Lemma 3.21 confirms this. Note that the empirical expectations in the Onsager correction for AMP (§4.4) will create floating components in the diagrams of the algorithmic state, but all such diagrams will be negligible.

Proof of Lemma 3.21. The effect of summing a vector diagram $\boldsymbol{Z}_\alpha = (Z_{\alpha,i})_{i\in[n]}$ over i is to unroot α, converting it to a scalar diagram. We prove this operation makes every diagram combinatorially negligible, except for the constant term. For $k\geqslant 0$ and a vector diagram $\alpha\in\mathcal{A}$:

1. If $a_n\boldsymbol{Z}_\alpha$ is combinatorially negligible, then $\frac{a_n}{n}\sum_{i=1}^n Z_{\alpha,i}$ is a combinatorially negligible scalar term.
2. If $a_n\boldsymbol{Z}_\alpha$ has combinatorial order 1, and the root of α is incident to at least one edge of multiplicity 1, then $\frac{a_n}{n}\sum_{i=1}^n Z_{\alpha,i}$ is a combinatorially negligible scalar term.

Unrooting a vector diagram does not change the number of vertices nor the number of edges. During this operation, the number of vertices in $I(\alpha)$ stays the same if the root is adjacent to an edge of multiplicity 1; otherwise it increases by at most 1. We readily check from the definition that the extra $\frac{1}{n}$ makes the resulting scalar terms combinatorially negligible.

Now let $\widehat{\boldsymbol{x}}$ be the tree approximation to $\boldsymbol{x}$. The difference $\boldsymbol{x}-\widehat{\boldsymbol{x}}$ consists of combinatorially negligible terms which stay negligible by part 1 above. The trees in $\mathcal{T}$ become negligible by part 2 above with the exception of the singleton tree which becomes 1. The singleton has coefficient $\mathbb{E}[\widehat{x}_1]=\mathbb{E}[X]$ since the other trees are mean-zero (Corollary 2.10). □