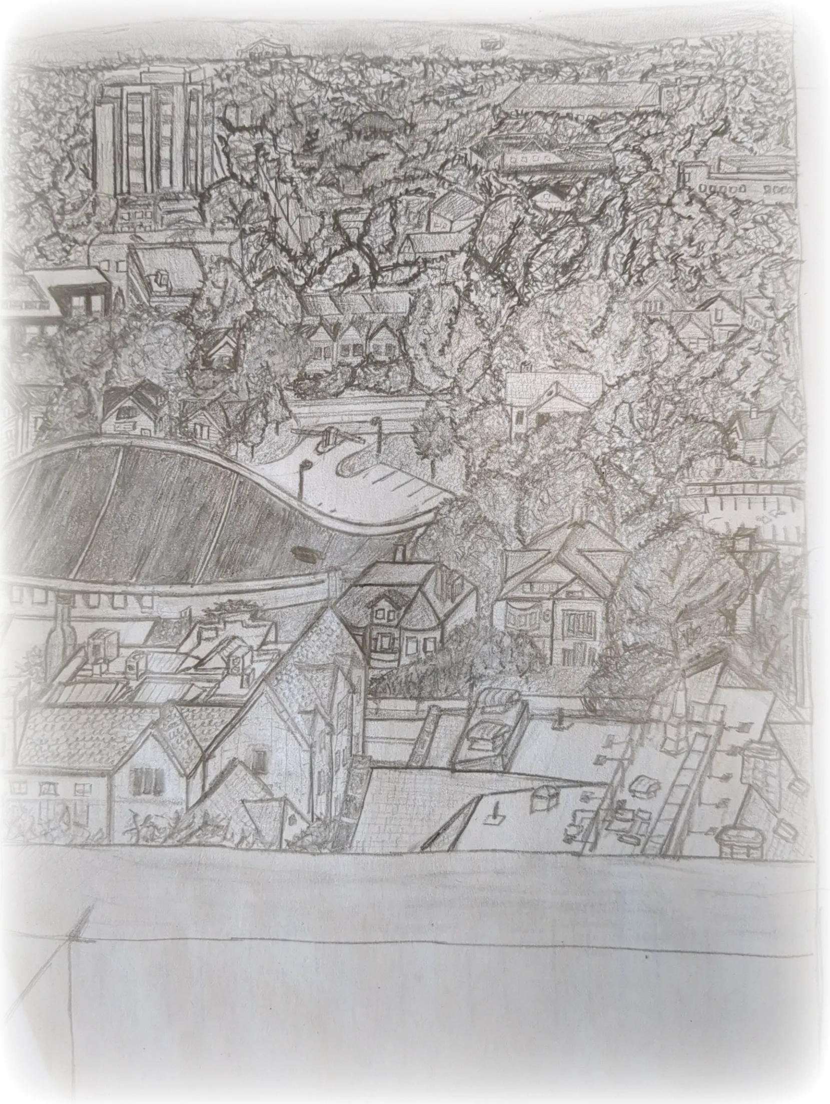
Analyzing High-Dimensional Algorithms and Landscapes



A Prospectus
Presented to the Department of
Statistics and Data Science at
Yale University

In Preparation for a
Ph.D. in Statistics and Data Science

By
Maxwell Lovig

Advisors: Zhou Fan, Ilias Zadik

March 2025

Acknowledgements

I am deeply grateful to my advisors, Ilias Zadik and Zhou Fan. Their patience with my more "outside" ideas and their unerring guidance have been invaluable. A special thanks to Zhou for teaching *High-Dimensional Phenomena in Statistics and Learning*—the first time I saw clearly. I also thank Sat Gupta, Johnathan Williams, and Amy Veprauskas for their mentorship before I came to Yale.

To the friends who have shared in my academic journey: Awni, Christian, Conor, Curtis, Gang, Heejune, Hugo, Jack, Kostas, Megan, Ruixiao, Siyu, Tianhao, Vladamir, Yihan, and Zehao—you have been better friends than I could have asked for. I am also grateful to Carolina, Conor, Emi, Jack, Jing, Kostas, Neil, Pujita, Sadia, Sumaita, and Tianhao for their collaboration.

To my family—Cosmo, Suzanne, and David—thank you for letting me be the oddball who's into math; nothing brings me more joy. And, perhaps most importantly, to Claire and Finn: I love you both.

On The Title Page Drawing

The drawing on the title page was made with pencil by me. I started working on it the summer after my first year at Yale—a whirlwind with the departments move to Kline Tower, an experience I'll never forget (perhaps only partially correlated with the prospect of getting my own office).

One day I was so exhausted of checking base cases for an inductive proof that I looked outside my window and, for the first time, truly appreciated the view. The result took about two years to finish. A majority of the delay was due to the end of fall; the more the leaves fell, the more intricate the drawing had to become. Post completion, a copy was proudly hung on the 10-th floor of Kline Tower.

If only I could grasp a better understanding on the complex landscape known as life—perhaps it wouldn't have taken so long.

Contents

CHAPTER 0

Introduction and Future Research Directions

Introduction

A major focus of statistical and computational research has been related to the study of algorithms and problem landscapes in high-dimensional inference. The term *high-dimensional* refers to a limit where both the number of observations and the dimensionality of the problem grow, but some intrinsic relationship between them is maintained. As real-world datasets enter this regime, understanding the trade-off between algorithmic performance and computational efficiency becomes increasingly important. Moreover, how far can we broaden the set of tools that can be used in high-dimensional inference?

This prospectus investigates three vignettes of high-dimensional inference problems to gradually explore answers to these issues:

(1) Bernoulli group testing and MCMC methods,
(2) Universality results in Approximate Message Passing (AMP) under general assumptions,
(3) The computational-local gap in sparse tensor PCA.

Utilizing techniques from statistical physics, random matrix theory, and probability, we investigate the extent to which the trajectory of a high-dimensional algorithm can be tracked and, therefore, characterized in terms of its performance. Along the way, we encounter statistical-to-computational gaps (where statistically optimal solutions exist but remain computationally intractable), local-to-computational gaps (where locally optimal solutions exist but underperform compared to their global counterparts), and extensions of algorithmic trajectories across a universality class.

Chapter 1 explores the landscape hardness of Bernoulli group testing (BGT). This study is conducted from the perspective of Markov chain dynamics and the Overlap Gap Property (OGP). We show that previous analysis, based on a vanilla "annealed" prediction, misled our understanding of this landscape, and a conditional annealed prediction is rigorously proven to give a more accurate characterization. This not only provides evidence that BGT exhibits a statistical-to-computational gap but also, to our knowledge, marks the first instance where conditioning on a high-probability event "corrects" the annealed prediction.

Chapter 2 explores the universality of an algorithm known as approximate message passing under the assumption of non-separable activation functions. This work broadens the applications of AMP from problems with Gaussian designs to the more general Wigner designs. We also introduce abstract assumptions dependent on the underlying computational graphs representing an AMP algorithm. When these assumptions hold, the trajectory of the AMP algorithm is shown to be identical in the limit with respect to the class of Wigner matrices. These results unify and extend prior work on universality and state evolution in AMP. I hope the results of this chapter will soon be available on *arXiv*.

Chapter 3 explores the computational-local gap in sparse tensor PCA, an in-progress project. These results consider a general framework for analyzing algorithms on Gaussian additive models known as noise-injected querying. We use this framework to analyze a set of local algorithms that can be written as Markov chains and prove that previously believed local-to-computational gaps for sparse tensor PCA do not exist. The results given in this chapter are preliminary and are subject to change.

Taken together, each of these vignettes provides insights into the limits of specific algorithmic classes, offering new perspectives on the study of high-dimensional inference and algorithmic performance. A more comprehensive introduction to each of these problems, along with an overview of the existing literature, is provided in each chapter.

Future Research Directions

Within the next three months, I hope to have finished both manuscripts [27, 50] corresponding to Chapter 2 and Chapter 3.

For future research, I plan to explore models similar to those in this prospectus and examine whether the same techniques apply. I am particularly interested in understanding a class of optimization techniques on generic machine learning architectures. Additionally, I am drawn to the connections between different frameworks for computational hardness, such as the relation between AMP and low-degree polynomials discussed in Chapter 3's introduction. Specifically, I aim to state these results with minimal assumptions on the model's structure. Perhaps a general class of algorithms exists that unifies multiple computational frameworks? Could this class inspire new testing and recovery algorithms?

My approach to these problems is continually evolving, and I look forward to further contributing to this field.

CHAPTER 1

Landscape Hardness Of Bernoulli Group Testing

Disclamer: This project was joint work with **Ilias Zadik**. The following presentation is adapted from the paper [51]. Where appropriate, the exposition follows the original paper. Proof of the below given results are found in [51], the relevant sections and lemmas are provided for the reader's convenience.

Notation. We use standard asymptotic notation. For any two positive sequences $A_n, B_n, n \in \mathbb{N}$, we write $A_n = O(B_n)$ if and only if $\limsup_n A_n/B_n < +\infty$, $A_n = \Omega(B_n)$ if and only if $B_n = O(A_n)$, $A_n = \Theta(B_n)$ if and only if $A_n = O(B_n)$ and $B_n = O(A_n)$, $A_n = o(B_n)$ if and only if $\lim_n A_n/B_n = 0$ and $A_n = \omega(B_n)$ if and only if $B_n = o(A_n)$.

We say that a sequence of events $(A_n)_{n\in\mathbb{N}}$ happen asymptotically almost surely (a.a.s) if and only if $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$ as $n \to +\infty$.

Given a function f of possibly many variables, one of which is γ, define $\partial_\gamma f$ to represent the derivative of f with respect to the variable γ. We also denote for $q_1, q_2 \in [0,1]$, the two point Kullback-Leibler (KL) divergence by

$$D(q_1||q_2) = q_1 \log(q_1/q_2) + (1-q_1)\log((1-q_1)/(1-q_2)). \tag{0.1}$$

Also we denote for any $C > 1$,

$$H_C := h_2^{-1}(2 - 2/C), \tag{0.2}$$

where h_2 is the left branch of the binary entropy function.

Finally, throughout the paper, we denote some important positive constants by $C_i, i \in \mathbb{N}$. Importantly, C_i will represent a specific constant when defined and will never change its value between two instances. There will also be a collection of constants using a different notation (such as $C > 0$) and these constants can vary from context to context.

1. Getting Started: Constraint Satisfaction Problems

Many problems in statistics and optimization fall under the category of (random) constraint satisfaction problems (CSPs). Such a problem is defined by the triple $(\mathcal{X}, \mathcal{D}, \mathcal{C})$, where: (1) $\mathcal{X}$ is a sequence of variables, typically representing a signal to be estimated or some abstract set of variables to optimize over; (2) $\mathcal{D}$ is the domain of each $\mathcal{X}$ variable in the sense that $X_i \in \mathcal{X}$ can only take values in $\mathcal{D}_i$; finally, (3) $\mathcal{C}$ is a set of (possibly random) constraints[1] enforced on the set $\mathcal{X}$.

Solving a CSP entails the demonstration of—either one or all—solutions X^* where every constraint in $\mathcal{C}$ is satisfied. Often time the worst case outcome of the randomized CSPs leads to a problem being significantly harder than the **typical case**, wherein we care about events that occur with (with respect to some notion of size in the problem) probability $1 - o(1)$. Both this chapter and Chapter 3 will study such problems in this typical setting.

A large set of problems that serve as proxies for studying algorithmic efficiency are CSPs. To demonstrate this commonality, we describe three distinct CSPs: (1) Compressed Sensing and

[1] Note, the computer science literature usually assumes some explicit form for these constraints, say written in a <scope, relation> form. For ease of presentation, I omit how a constraint may is explicitly structured and leave their definition abstract.

Sparse Tensor PCA, (2) Graph Coloring and (3) Group testing; the third problem is the main focus of this chapter.

Compressed Sensing and Sparse Tensor PCA: The sparse tensor PCA problem, considered in the matrix case by [43] and the tensor case by [58], is a famed model for demonstrating statistical and computational complexity. We prescribe the matrix version of this problem below, and leave the tensor case to Chapter 3.

Given a vector $\theta \in \{v \in \{0,1\}^n : ||v||_0 = k\}$, with $k \in [n]$, we receive the observation

$$Y = \frac{\lambda}{k}\theta\theta^\top + W \tag{1.1}$$

where $W \in \mathbb{N}$ is an i.i.d. Gaussian matrix and λ can possibly scale with n and k. Converting this problem to a CSP is rather straight forward. Define $\mathcal{X} = \{v \in \{0,1\}^n : ||v||_0 = k\}$, $\mathcal{D} = \{0,1\}^n$ and the constraint set $\mathcal{C}$ can enforce, say, $\langle \theta, v\rangle \geq .8$ or a sufficient amount of posterior mass being given by v conditional on Y.

Much is now understood for this problem, specifically in the study of computational hardness. A smattering of results can be found justifying **statistical-to-computational gaps** [58, 66, 39, 46, 30], meaning that Bayes rule (with some agreed upon "natural" prior) does achieve the desired constraints at some signal-to-noise ratio (SNR), but no polynomial time algorithm can achieve similar feats for the same SNR. The exact definition of the SNR depends on the scaling chosen in model (1.1) but is usually a function of λ. Studies have also demonstrated a secondary **computational-to-local** gap [3, 4, 18], which suggests efficient polynomial time algorithms that achieve an optimal SNR must be "global" in nature and the "local" algorithmic counterparts need a significantly larger SNR. The latter gap is studied for this model in Chapter 3.

Graph Coloring: Another example of a CSPs is the class of graph coloring (or vertex assignment) problems. Perhaps the most well-known graph coloring problem is the *Four Color Theorem* [11]. The general problem setup is as follows:

Given a graph $G = (\mathcal{V}, \mathcal{E})$ with a vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, we are given a number of colors $[a]$ and aim to construct a coloring function $f : \mathcal{V} \to [a]$ where in two adjacent vertices $v_1, v_2 \in \mathcal{V}$ have $f(v_1) \neq f(v_2)$. This problem has close ties to more traditional CSPs and often can be mapped to a SAT problem [36]. Graph coloring can be expressed as a CSP under the choice of $\mathcal{X} = \mathcal{V}$, $\mathcal{D} = (D_i)_{i\in\mathcal{V}}$ where $D_i = [a]$, and where $\mathcal{C}$ enforces the constraint that adjacent vertices must have a different color.

Under the random model that each possible edge (v, u), where $v, u \in \mathcal{V}$, has probability p of being present in $\mathcal{E}$ (i.e. the well known Erdös-Renyi p-model). This model for edge probabilities is studied from the sum of squares perspective in [45] and a statistical physics' framework derived from the cavity method in [6].

Group Testing: Finally, we dial in on the main focus of this chapter, **group testing**. Introduced by [24], consider a vector $\sigma^* \in \{0,1\}^n$ where $||\sigma^*||_0 = k \in [n]$; where the coordinates given ones are chosen uniformly at random. The statistician then assigns N group tests, where, for each group test, a subset $\mathcal{S} \subset [n]$ is proposed and the observation

$$\text{Result}(\mathcal{S}) = \begin{cases} + & \text{if } \mathcal{S} \cap \sigma^* \neq \varnothing \\ - & \text{otherwise.} \end{cases}.$$

is received. Due to its simplicity, the group testing problem has found many applications [47, 65, 73, 63, 64, 54, 5]. A large portion of which rely on the assumption that k is known (at least approximately) and is far smaller than n. For this reason the scaling of $k = \Theta(n^\alpha)$—for some $\alpha \in (0,1)$—is chosen.

Even under this definition, there exists a dichotomy in how (and when) to apply each group test. There is the *adaptive* case where the results of subsequent group tests can inform the assignment of the next group test, or the *non-adaptive* case where the assignment of each group test is fixed, and the results are given all at once. A comprehensive overview of these variants can be found in [1] and a comparison of lower bounds for estimation using Fano's method can be found in [70].

This chapter concerns the non-adaptive version of group testing, where each subset $\mathcal{S}$ is random samples from $[n]$ by including each element with an independent probability q, known as a **Bernoulli q design**. The exact choice of q may depend on n, k or N—as we will see later. Such a design is attractive for its simplicity and ease of implementation. Given these random group test assignments and their corresponding outcomes, the goal is to construct and estimator $\hat{\sigma} \subset [n]$ where $||\sigma||_0 = k$ and

$$\lim_n |\hat{\sigma} \cap \sigma^*|/k = 1, \tag{1.2}$$

asymptotically almost surely (a.a.s.) with respect to the randomness of the prior of σ^* and the design of the group tests. In words, we need to name k individuals of which a $1 - o(1)$ fraction of them are infected with high probability.

1.1. Statistical-To-Computational Gaps In Group Testing. Information theoretic arguments prove that, when $N \leq (1+\varepsilon)\log_2\binom{n}{k}$ for some $\varepsilon > 0$, there **does not** exist designs for group testing that lead to successful recovery [1]. The converse result also holds, when $N \geq (1+\varepsilon)\log_2\binom{n}{k}$, there exist designs (one example of which is a Bernoulli q design) that lead to successful recovery [1] in the sense of (1.2). Indeed, any k subset of the n individuals that is *covers*[2] sufficiently many positive tests almost perfectly recovers σ^* a.a.s. as $n \to +\infty$ (see e.g., [40, Lemma 5]). Thus, a simple brute force search over all k subsets will produce a successful recovery for any such value of N tests.

Although the Bernoulli group test design has favorable properties, a *major challenge* is the existence of a polynomial time algorithm for successful recovery when $N = (1+\varepsilon)\log_2\binom{n}{k}$ for small enough $\varepsilon > 0$. Additionally, the relation of group testing to the $\mathcal{NP}$-hard set-cover problem suggests that a brute force search over all k-subsets may be unavoidable for N close to $\log_2\binom{n}{k}$. An important note is that these results do not imply there is a "fast" way to recover σ^*. The best known algorithm to date is Separate List decoding [1, 69], which requires $N \geq (\log 2)^{-1}\log_2\binom{n}{k}$ tests. It remains unknown if any polynomial time algorithm can beat the constant $(\log 2)^{-1}$ in this setting. As discussed in the introduction to this chapter, the difference in N when a.a.s. **recovery is possible** and when a **polynomial run time algorithm exists** is a statistical-to-computational gap. Confirmation of this gap was found in [20] from the *low-degree* perspective [38]. Specifically, [20] proved that no $O(\log(n))$-degree polynomial estimator can recover σ^* when $N < (\log 2)^{-1}\log_2\binom{n}{k}$ for $k = \Theta(n^\alpha)$ with α sufficiently small. This result is significant due to a conjecture [38] which suggests the class of $O(\log(n))$-degree polynomials is a proxy for the class of all polynomial-time estimators.

An alternative perspective to the low-degree view comes from a direct study on the "landscape" of the group testing. Here, [40] attempted to prove or disprove the existence of the *Overlap Gap Phenomenon for inference* [32], we refer to this property as b-OGP for the remainder of this chapter. B-OGP postulates that some level set of an objective divides the set of candidate solutions with said objective or less (say with the goal of minimization) into two sets, a "good" set aligned well with the true solution and a "bad" with poor alignment. This suggests local algorithms—such as low-temperature MCMC methods (see e.g., [29, 32, 4, 33, 16, 19])—are unable to escape the bad region as they are unable to "jump" over this gap. Moreover, b-OGP is known to coincide with the threshold for the slow-to-fast mixing of low-temperature MCMC methods for: (1) sparse regression [32, 19], (2) planted clique [33], and (3) sparse tensor PCA [4, 19]. See Definition 2.1 for a rigorous definition of b-OGP .

Interestingly, [40] concluded that, **under sufficient concentration of certain key quantities around their expectation**, that a *first moment function* (often referred to as *annealed* in statistical physics [82]) proves that b-OGP **never occurs** for Bernoulli group testing for any

[2] We say that a k-subset of individuals "covers" a given test if at least one of the k individuals took part in this test.

$N \geq (1+\varepsilon)\log_2\binom{n}{k}, \varepsilon > 0$. Of particular surpise is that this threshold corresponds to information-theoretic possible recovery, not the expected algorithmic threshold of [20] discussed previously. Indeed, if this conclusion holds, it would represent the first time that an MCMC method could provably outperform the set of $O(\log(n))$-degree polynomials, contradicting many current predictions in the literature of statistical-to-computational gaps [38]. A major motivation for this chapter is to determine whether such an advantage exists for Bernoulli group testing.

To recap this discussion, we provide a more precise definition of the group testing problem in the form of bipartite graphs. We then further detail a vital post-processing step that is common for group testing [40].

1.2. A Rigorous Definition Of Bernoulli Group Testing. We start with properly defining the Bernoulli group testing instance. Consider n to be the number of individuals. We assume that n grows to infinity and all other growing parameters grow as a function of n. We can represent group testing as a bipartite graph $G = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E})$ where $\mathcal{V}_1 = [n]$, $\mathcal{V}_2 = [N]$ and the edge (i, j) is present in $\mathcal{E}$ if and only if individual i is included in test j. Such a bipartite graph is visualized in Figure 1.

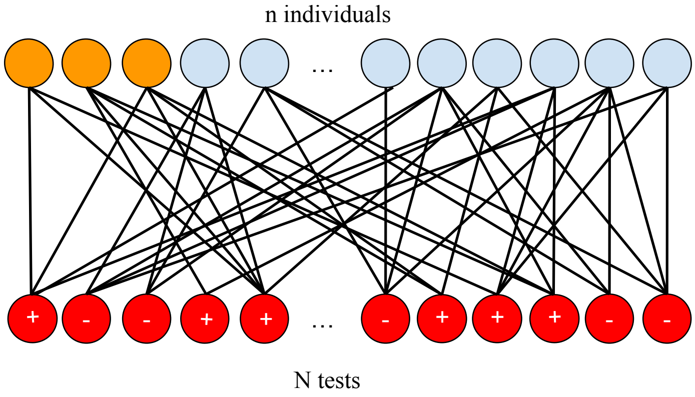


FIGURE 1. A realization for an instance of Bernoulli group testing.

DEFINITION 1.1. Fix some constants $\alpha \in (0, 1)$ and $C > 1$. We call the (α, C)-group instance the following setting. Among the n individuals, we assume there is a subset of $k = \lfloor n^\alpha \rfloor$ infected ones, denoted by σ^*, which are chosen uniformly at random among all k-subsets of $[n]$.

The statistician observes $N = \lfloor C \log_2\binom{n}{k} \rfloor$ group tests, where each individual participates in each test with an assignment probability $q \in (0, 1)$ satisfying

$$(1-q)^k = \frac{1}{2}.$$

As mentioned earlier, we aim to construct an estimator $\hat{\sigma}$ given an (α, C)-instance of group testing where (1.2) holds.

REMARK 1.2. We make a few remarks on the choice of the parameters. First, we have $C > 1$ as the choice of $C = 1$ is the information-theoretic threshold for group testing [1]. Second, the assumption on q is standard in Bernoulli group testing; it is motivated by the information theory idea that it is optimal to "halve" the number of individuals in each test. This heuristic is buttressed by results showing that a (time-inefficient) $\hat{\sigma}$ exists for this q when $C > 1$ [40, Lemma 5]. Third, for convince, we often represent q in the asymptotic form $q = (\log(2) + o(1))/k$ and denote the number of positive tests as $M = (1 + o(1))N/2$ a.a.s. as $n \to +\infty$.

1.3. Post-Processing With COMP. This section describes a post-processing step which is applied as a first step to a vanilla Bernoulli group testing instance, recall this was pictured in Figure 1. This processing follows from the simple observation that if an individual participated in a negative test, then they cannot be infected. Thus, it is natural to restrict the candidate infected individuals to those which did not participate in any negative test, a technique known as Combinatorial Orthogonal Matching Pursuit (COMP) [1]. If $C > 2$ from Definition 1.1 then COMP outputs only the infected individuals a.a.s. as $n \to +\infty$, therefore recovering σ^* [1]. This leads us to reduce our set of parameters to $1 < C < 2$ for the remainder of this chapter.

Moreover, in this regime where $1 < C < 2$, [51, Lemma 6.1] and [51, Lemma 6.2] (which follow from standard concentration of measure inequalities), imply that there are $M = (1 + o(1))N/2$ positive tests and $p = (1 + o(1))n(k/n)^{C/2} + k$ remaining individuals that are possibly infected. Pictorially, this post-processing step when applied to Figure 1, results in Figure 2.

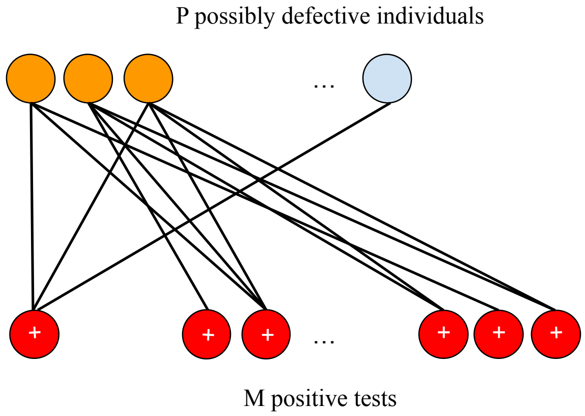


Figure 2. A realization for an instance of Bernoulli group testing, now with the COMP post-processing applied.

2. Main Contributions

Notation: These results will use the notation $|\cdot|$ to refer to $||\cdot||_0$ and for two vectors $\sigma, \sigma' \in \mathbb{R}^n$ we denote $\sigma \cap \sigma'$ as the Hadamard product.

This section formally presents the landscape b-OGP results from [51], resulting in our lower bounds for low-temperature MCMC methods. In all that follows, as explained in Section 1.3 we consider only the p possibly infected individuals and subsets σ of them. Similar to [40], the first key step is to study the following (random) restricted optimization problems over $\ell \in \{0, 1, \ldots, k\}$,

$$\phi(\ell) := \min\{H(\sigma) : |\sigma| = k, |\sigma \cap \sigma^*| = \ell\},$$

where $H : \mathbb{R}^n \to \mathbb{R}$ is the following Hamiltonian,

$$H(\sigma) := \#\text{ of positive tests non-covered by } \sigma/M.$$

The non-monotonicity of $\phi(\ell)$ is known to be linked with b-OGP [32], defined as follows.

Definition 2.1. Let constants $\zeta_1, \zeta_2 \in [0, 1]$ with $\zeta_1 < \zeta_2$, threshold value $r = r_n > 0$ and height value $\delta = \delta_n > 0$. A group testing instance exhibits the bottleneck Overlap Gap Property (b-OGP) for parameters $\zeta_1, \zeta_2, r, \delta$ if the following conditions hold.

(1) There exist size k subsets σ_1, σ_2 with $\frac{1}{k}|\sigma_1 \cap \sigma^*| \leq \zeta_1$, $\frac{1}{k}|\sigma_2 \cap \sigma^*| \geq \zeta_2$, for which it holds $\max\{H(\sigma_1), H(\sigma_2)\} < r$.

(2) For any k-subset σ with $|\sigma \cap \sigma^*| \in [\zeta_1, \zeta_2]$ it holds $H(\sigma) \geq r + \delta$.

It is well-known in the literature that b-OGP is related to the (non)-monotonicity of $\phi(\ell)$. Indeed, [40, Lemma 20] implies that the non-monotonicity of $\phi(\ell)$ is necessary for the existence of b-OGP and a simple argument, used for example in [33, Theorem 2], implies that the non-monotonicity of $\phi(\ell)$ is also sufficient for the existence of b-OGP .

Characterizing $\phi(\ell)$ leads to studying the count of size k subsets σ which have a given overlap ℓ and objective value t.

Definition 2.2. For $t \in \{0, 1, \ldots, M\}, \ell \in \{0, 1, \ldots, k\}$ define $Z_{t,\ell}$ to be the random variable

$$Z_{t,\ell} = |\{\sigma : |\sigma| = k, |\sigma \cap \sigma^*| = \ell, \sigma \text{ leaves at most } t \text{ positive tests uncovered}\}|$$

Notice that $\phi(\ell) \leq t/M$ if and only if $Z_{t,\ell} \geq 1$. Hence, it suffices to find the minimal $t > 0$ such that $Z_{t,\ell} \geq 1$ a.a.s. as $n \to +\infty$. We first detail how—hueristically—we can use $Z_{t,\ell}$ to approximate $\phi(\ell)$, therefore getting a better view of the landscape of group testing.

Following a method proposed by [40], we can define an implicit "first-moment" equation in t of the form,

$$\mathbb{E}[Z_{t,\ell}] = 1. \tag{2.1}$$

The motivation for this choice is two-fold. To explain this, let us fix a $\ell \in \{0, 1, \ldots, k\}$.

(a) If for some $t_1 > 0$ it holds that $\mathbb{E}[Z_{t_1,\ell}] = o(1)$, then by Markov's inequality $Z_{t_1,\ell} = 0$ a.a.s. as $n \to +\infty$, and therefore $\phi(\ell) \geq t_1$. This is customary called the first moment method.

(b) On the other hand, if for some $t_2 > 0$ (ideally relatively "close" to $t_1 > 0$) it holds that $\mathbb{E}[Z_{t_2,\ell}] = \omega(1)$ and the distribution of $Z_{t_2,\ell}$ concentrates, for example with $\text{Var}[Z^2_{t_2,\ell}] = o(\mathbb{E}[Z_{t_2,\ell}]^2)$, then $Z_{t_2,\ell} \geq 1$ a.a.s. as $n \to +\infty$, giving $\phi(\ell) \leq t_2$. This is customary called the second moment method.

Thus, if t_ℓ is the "first-moment" solution for (2.1) with respect to t and one establishes sufficient concentration of $Z_{t,\ell}$ for $t \approx t_\ell$, then one could naturally predict that a.a.s. as $n \to +\infty$, it holds

$$\phi(\ell) \approx t_\ell. \tag{2.2}$$

As an example we plot the solution to (2.1), given implicitly in [40, Definition B.6], for the set of parameters $(n = 2^{75}, C = 1.1, \alpha = .02)$.

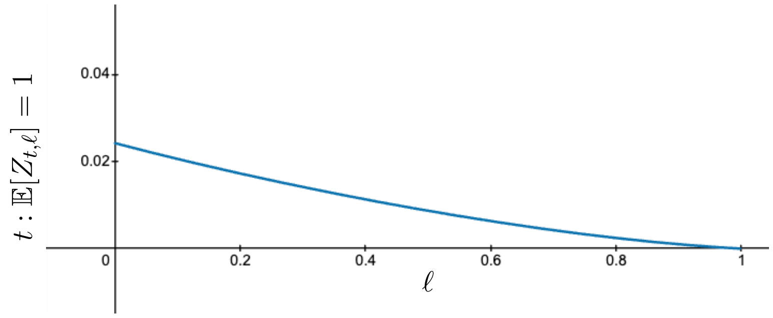


Figure 3. The plot of the unconditional first moment function given in (2.1), the above plot was calculated under the choice of parameters $(n = 2^{75}, C = 1.1, \alpha = .01)$.

Under the assumption that the true minimizer constrained to some level of overlap concentrates around its expectation, we can treat the above curve as a proxy for the function $\phi(\ell)$. The immediate observation that the curve in Figure 3 is decreases as ℓ increases implies that as we would attempt a minimization of $H(\sigma) \approx t$ we naturally will increase the overlap ℓ while decreasing the objective value. Indeed, this was the conclusion of [40]—up to a **conjectured** success of the second moment method [40, Conjecture B.9]; the landscape does not demonstrate b-OGP as for each level set of $t = t_0$ the set of ℓ providing such an objective or lower is connected and contains $\ell = k$. Suggesting that this constraint satisfaction problem is far simpler than other, b-OGP demonstrating problems.

It is worth noting that similar problems have demonstrated that the approximation (2.2) **usually holds**. Examples of sucess for this method include sparse regression [32, 19], planted clique [33] and sparse tensor PCA [4, 19].

2.1. The Conditional First Moment Function. A crucial contribution of this work is demonstrating that in Bernoulli group testing (2.2), as well its conclusion on non-existence of the b-OGP , are **incorrect** due to the presence of rare events. Notice that one can consider a variation of the first-moment equation (2.1) under a conditioned event $\mathcal{A}$,

$$\mathbb{E}[Z_{t,\ell}|\mathcal{A}] = 1. \tag{2.3}$$

The key idea is that a *conditional* first moment method also holds: if $\mathcal{A}$ occurs a.a.s. as $n \to +\infty$, then for any $t_1' > 0$, with $\mathbb{E}[Z_{t_1',\ell}|\mathcal{A}] = o(1)$, it must hold that $\phi(\ell) \geq t_1'$ a.a.s. as $n \to +\infty$, with the potential t_1' being much larger than t_1 coming from the vanilla first moment method. Albeit a natural idea—**no such conditioning** has been required in the analysis of similar sparse problems [32, 33, 4, 19].

We first define the key conditioning event.

Lemma 2.3 ([20], Section 9.2.1 (arxiv version)). *Consider an (α, C) instance of group testing. If a is an element of the set*

$$\left\{a : \log(2)C(a\log(a) - a + 1) > \frac{\alpha}{1-\alpha}\right\}, \tag{2.4}$$

then for

$$\mathcal{A} := \{\deg(i) \leq 2aqM,\ \forall i \in \sigma^*\}$$

it holds that $P(\mathcal{A}) = 1 - o(1)$.

Using this choice of $\mathcal{A}$ in equation (2.3), we denote by $t_\ell' = t_\ell'(\mathcal{A})$ the (now conditional) first moment solution of (2.3) with respect to t given the value of $\ell \in \{0, 1, \ldots, k\}$. One could aim to solve for t_ℓ' and seek to get a simpler formula for it. Using linearity of expectation, standard concentration of measure asymptotics, and a direct computation with (2.3) (Given in [51, Section 5]), we indeed get a simpler (but still implicit) set of equations satisfied by a very close proxy to t_ℓ'.

To explain the derived equations, notice that both t and ℓ take values in growing regions, $\{0, 1, \ldots, M\}$ and $\{0, 1, \ldots, k\}$ respectively. Hence, it is convenient to re-parameterize our setting in terms of the proportional overlap $\frac{\ell}{k} = x \in [0, 1]$. Moreover, we also denote our proxy for the re-scaled quantity $\frac{t_\ell'}{M} = \frac{t_{xk}'}{M}$ by $y(x) \in [0, 1]$. To define $y(x)$ we first remind the reader the definition of the two point KL divergence from (0.1). We now define $y(x)$ as follows.

Definition 2.4. Consider $r(x) := 4 \cdot 2^{-x}(1 - 2^{-x})$, $s(x) := 1 - 2^{x-1}$, with $x \in [0, 1]$, $\alpha \in (0, 1)$, $C \in (1, 2)$, constants $C_1, C_2, C_3 > 0$, and a an element of the set (2.4).

For any $x \in [0, 1]$ define the (C_1, C_2, C_3)-*first moment function* at x, denoted by $y = y(x)$ as the solution to the equation,

$$\frac{1}{M}\log\left(\binom{k}{\lfloor xk \rfloor}\binom{p-k}{\lfloor (1-x)k \rfloor}\right) = (1-y)D\left(\frac{2a\log(2)x}{1-y}\bigg|\bigg|r(x)\right) + D(y||s(x)) \tag{2.5}$$

satisfying the following four constraints,

$$\frac{2a\log(2)x}{1-y} \leq (1 - C_1)r(x) \tag{2.6}$$

$$y \leq (1 - C_2)s(x) \tag{2.7}$$

$$2a\log(2)x \leq (1 - C_1)r(x) \tag{2.8}$$

$$D\left(1-\frac{2a\log(2)x}{(1-C_1)r(x)}\bigg|\bigg|s(x)\right)+\frac{2a\log(2)x}{(1-C_1)r(x)}D((1-C_1)r(x)||r(x))\leq(1-C_3)(1-x)(2-C)\log(2)/C \tag{2.9}$$

Often we will reference the region of x where $(x, y(x))$ satisfy (2.6)-(2.9) , in which there is an implicit choice of $\alpha, C, C_1, C_2, C_3, a$.

The definition of the first moment function is unfortunately quite technical. For this reason, we defer explaining the exact relation between t'_ℓ/M and $y(x)$ to Section [51, Section 5]and proceed with a few high level explanatory remarks.

Remark 2.5. The equation (2.5) turns out to be equivalent to (2.3) up to lower order terms. This is an outcome of a standard concentration of measure argument on the product Bernoulli distribution that constraints (2.6) and (2.7) allow to be applied. Moreover, under constraints (2.6) and (2.7), the additional constraints (2.8) and (2.9) allow us to restrict to values of x that the first moment function $y(x)$ provably exists and is unique. The proof of this fact is given in [51, Section 5.2]. Moreover, as long as the first moment function exists on an interval, a similar argument allows us to conclude the continuous differentiability of $y(x)$ on the interval (see also [51, Section 5.2]).

Remark 2.6 (The role of C_1, C_2 and C_3). The introduction of the constants C_1, C_2 and C_3 in the definition is purely for technical convenience. They do not change the value of the solution to $y(x)$ in (2.5), they simply slightly restrict the region of x where $(x, y(x))$ is defined to avoid certain degeneracies in our arguments in [51, Section 5]. For this reason, we consider them to be arbitrarily small constants.

Remark 2.7. Lastly, we highlight that often in what follows (but not always) we consider the values of x to be restricted on the set $\{0, 1/k, 2/k, \ldots, 1\}$. In those cases, for notational simplicity and when clear from context, we drop the floor function from the binomial coefficients in (2.5).

As a follow-up to Figure 3, we can compare the solution to [40]'s vanilla first moment function and the conditional first moment function from (2.3) in Figure 4. We use the same parameters as Figure 3 with the additional choice of $a = 1.17$, defining the conditional event in Lemma 2.3.

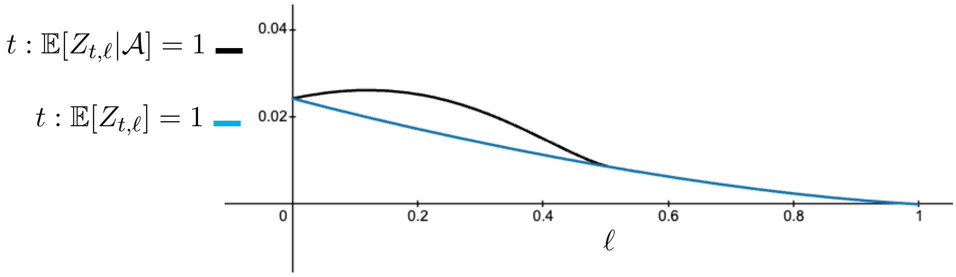
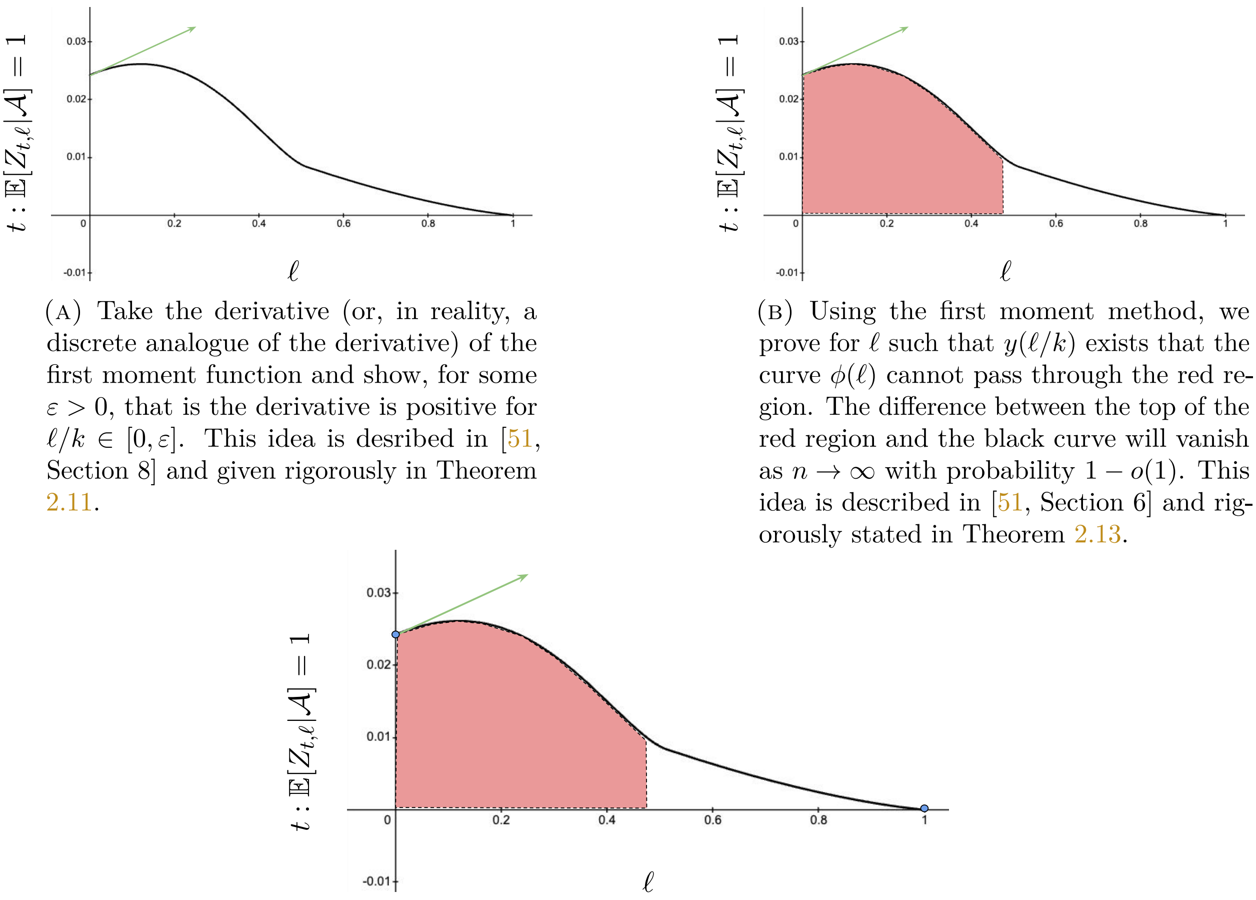


Figure 4. The plot of the conditional first moment function given in (2.3), the above plot was calculated under the choice of parameters $(n = 2^{75}, C = 1.1, \alpha = .01, a = 1.17)$.

Immediately, we see that the conditional first moment function in is no longer decreasing. In fact, when $\phi(\ell) \approx t_\ell$ where t_ℓ is the solution to (2.3), then the non-montonicity of the black curve in the above figure suggests that group testing would display b-OGP for this (α, C) group testing instance. Moreover, if we can prove sufficient concentration then we could pick $t \approx H(\sigma) = .025$ and see that the set of ℓ which achieve such a solution will be separated into the "good" and "bad" sets given in the introduction.

Of course, to actually study the typical case of group testing, we must provide rigorous justification for this annealed heuristic. For visual simplicity we "stitch" together the conditional first moment function (where it exists) with the unconditional first moment function (where the conditional



(A) Take the derivative (or, in reality, a discrete analogue of the derivative) of the first moment function and show, for some $\varepsilon > 0$, that is the derivative is positive for $\ell/k \in [0, \varepsilon]$. This idea is desribed in [51, Section 8] and given rigorously in Theorem 2.11.

(B) Using the first moment method, we prove for ℓ such that $y(\ell/k)$ exists that the curve $\phi(\ell)$ cannot pass through the red region. The difference between the top of the red region and the black curve will vanish as $n \to \infty$ with probability $1 - o(1)$. This idea is described in [51, Section 6] and rigorously stated in Theorem 2.13.

(C) Using the second moment method, we prove that at $\ell = 0$, the difference between $\phi(\ell)$ and the solution to the conditional first moment function will vanish as $n \to \infty$ with probability $1 - o(1)$. This upper bound is represented by the blue dot at $\ell = 0$. The blue dot at $\ell = k$ and $t = 0$ is due to the fact that the true solution always covers every test. This idea is described in [51, Section 7] and rigorously stated in Theorem 2.13.

FIGURE 5. The three steps to verifying b-OGP .

variate does not exist). In Figure 5, we visualize the technical feats that lead to the conclusion of b-OGP for some parameters of Bernoulli group testing.

Upon reinspecting Figure 5 and Definition 2.1, we can see that rigorously verifying these 3 steps is sufficient for b-OGP to occur.

2.2. Local Monotonicity Of A First-Moment Function. Recall that our goal is to prove that $\phi(\ell)$ ($\ell \in \{0, 1, \dots, k\}$) is non-monotonic for some regime of α, C to conclude the existence of b-OGP . Moreover, as we aim to approximate $\phi(\ell)$ using the deterministic $y(\ell/k)$, a natural question is whether $y(\ell/k)$ is non-monotonic. On top of that, following the plots in Figure 5, it is natural to expect that the non monotonicity to take place around $\ell/k \approx 0$. Hence, we now focus on whether there exists a region of $x = \ell/k$ close to 0 where we can prove the non-monotonicity behavior of $y(x)$.

To answer this question, we first naturally need to guarantee that for some $\varepsilon > 0$ the first moment function exists for all $x \in [0, \varepsilon]$ which, as explained in Remark 2.5 it is guaranteed if the constraints (2.6)-(2.9) are satisfied for all $x \in [0, \varepsilon]$. The following assumption suffices to guarantee this part.

ASSUMPTION 2.8. We assume that the parameters (α, C, a, C_1, C_3) satisfy

$$D\left(1 - \frac{a}{2(1-C_1)} \Big|\Big| \frac{1}{2}\right) \leq (1 - C_3)\frac{2-C}{C}\log(2) \tag{2.10}$$

and

$$\frac{a}{2(1-C_1)} < 1, \tag{2.11}$$

where $C_1, C_3 > 0$ and a being a valid choice from (2.4).

Because of the complexity of the assumption, we plot the range of α and C for which Assumption 2.8 holds in Figure 6, by setting a and C_1, C_3 to their lowest possible values. It is worth pointing out that the assumption is satisfied for any $1 < C < 2$ as long as $\alpha > 0$ is small enough.

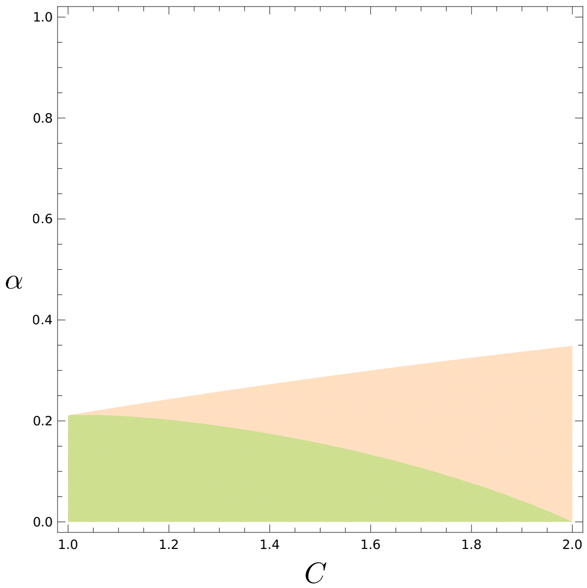


FIGURE 6. The green and orange regions in the above plot represent the values of α and C for which conditions (2.10) and (2.11) from Assumption 2.8 are satisfied under the choice of a from the lower boundary of the set (2.4) and setting $C_1, C_3 = 0$. Note that the region in green is a subset of the region in orange.

Under Assumption 2.8, we have the following result.

LEMMA 2.9. *If (α, C, a, C_1, C_3) satisfy Assumption 2.8, then there exists an $\varepsilon > 0$ such that the first moment function $y(x)$ according to Definition 2.4 for $x \in [0, \varepsilon]$ exists and is unique a.a.s. as $n \to +\infty$ (with respect to the randomness of p, M). Moreover, $y(x)$ is continuous and differentiable over* $[0, \varepsilon]$.

The proof of this result is given in [51, Section 5.2].

Now that we have established that the first moment function exists and is unique around zero, we also make the following assumption on our parameters which allows us to conclude the desired monotonicity of the first moment function at 0.

ASSUMPTION 2.10. Recall H_C from Definition 0.2. We assume that (α, C, a) satisfies

$$C < \frac{1 - \frac{\alpha}{1-\alpha}}{a\left(1 - \log\left(\frac{a}{2(1-H_C)}\right)\right) + H_C - 1},$$

and that a is a valid choice from (2.4).

This cumbersome assumption appears quite naturally by calculating the discrete derivative of $y(\ell/k)$ around $\ell/k \approx 0$ and checking when it is strictly positive (See [51, Section 8]). Given a pair (α, C), if one chooses a to be the lowest feasible value from (2.4), then the pairs (α, C) that satisfy this assumption are given in Figure 7. In particular, we highlight that the condition is valid for all $0 < C < C^* \approx 1.4749$ for $\alpha > 0$ sufficiently small.

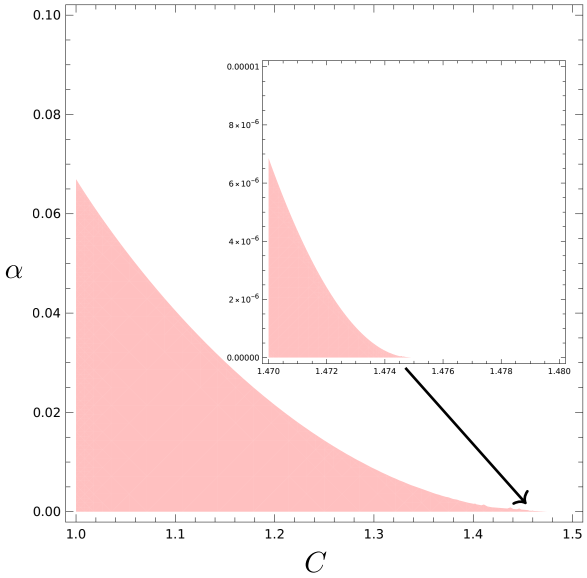


FIGURE 7. The region in red represents the values of α and C for which Assumption 2.10 holds when choosing of a from the lower boundary of the set (2.4).

Now, under the above assumptions we prove that indeed the first moment function must increase near 0.

THEOREM 2.11. *If the parameters (α, C, a, C_1, C_3) satisfy Assumption 2.8 and Assumption 2.10, then, a.a.s. as $n \to +\infty$ (with respect to the randomness of p, M), there exist constants $\varepsilon_1 > 0$ and $\delta_1 > 0$ such that for all $0 \le \ell \le \varepsilon_1 k$ it holds*

$$y(\ell/k) - y(0) \ge \delta_1 \ell/k.$$

The proof of the theorem is given in [51, Section 8].

2.3. Local Monotonicity Of $\phi(\ell)$ Via First Moment Function Approximations.

From Theorem 2.11, we know that $y(\ell/k)$ increases for all $\ell \le \varepsilon k$ for some small $\varepsilon > 0$. We now investigate whether ϕ inherits this monotonic increase near zero from the first moment's functions behavior. To establish this, it suffices to show that $y(\ell/k) - o(1)$ a.a.s. lower bounds $\phi(\ell)$ over the region $\ell/k \in [0, \varepsilon]$ and demonstrate an equivalent $y(0) + o(1)$ a.a.s. upper bound for $\phi(0)$.

Similar to the above result on the first moment function, the following result on $\phi(\ell)$ is subject to a few parameter assumptions. This assumption is again rather cumbersome, an outcome of an involved second moment method argument that leverages it. Crucially, however, this assumption is satisfied for all $1 < C < 2$ when α is sufficiently small (see Figure 8). We also direct the reader to [51, Section 9.1] for more details on this assumption.

ASSUMPTION 2.12. The pair of parameters (α, C) satisfy $\alpha < 28/1000$ and

$$C < 2\frac{1-2\alpha}{1-\alpha}.$$

Moreover, the pair satisfies the following two conditions with H_C from Definition 0.2,

$$C\left[(1-H_C)(1-\log(2(1-H_C))) - \frac{h_2(H_C)}{2} - 7\sqrt{\frac{\alpha}{1-\alpha}}\left(\frac{1}{2}\log(2(1-H_C))\right)\right] > 4\alpha/(1-\alpha) \quad (2.12)$$

and

$$C\left[\frac{h_2(H_c)}{2}+\frac{1}{2}\log\left(\frac{1-H_C}{H_C}\right)\left(1-H_C-5\sqrt{\frac{\alpha}{1-\alpha}}\right)+H_C-1\right]>3\alpha/(1-\alpha). \tag{2.13}$$

Using this assumption we can then get our desired bounds on $\phi(\ell)$.

Theorem 2.13. *If the parameters (α, C, a, C_1, C_3) satisfy Assumption 2.8 and Assumption 2.12, then there exists an $\varepsilon' > 0$ such that, for all $x = \ell/k \in [0, \varepsilon']$, we have a.a.s. as $n \to +\infty$ that,*

$$\phi(\ell) \geq y(\ell/k) - O(1/k).$$

Moreover, a.a.s. as $n \to +\infty$,

$$\phi(0) = y(0) + o(1) = H_C + o(1).$$

This result combines an a.a.s, as $n \to +\infty$ lower bound on $\phi(\ell)$ for all $\ell = 0, 1, \ldots, k$ as well as an a.a.s, as $n \to +\infty$ upper bound on $\phi(0)$, both of which are shown in [51, Section 6] and [51, Section 7]. The former relies on a relatively straightforward application of a conditional first moment method. The latter part is highly non-trivial to prove. We prove it via an elaborate conditional second moment method and is far more technical due to the necessity for delicate control over shared positive tests between two non-infected individuals.

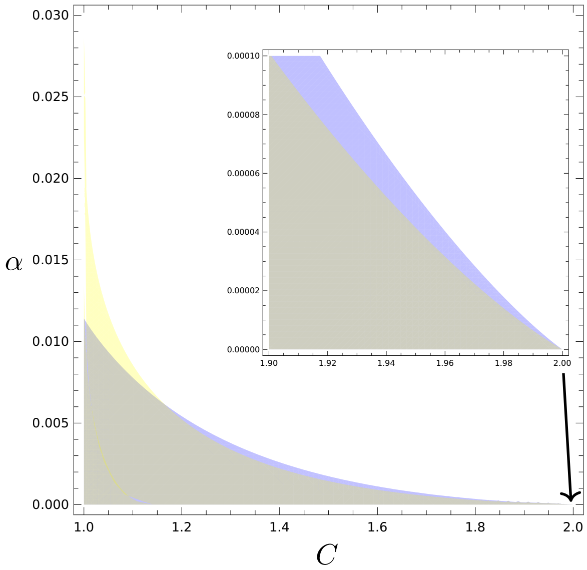


Figure 8. A visual representation for when Assumption 2.12 holds. The x-axis represents the value of C and the y-axis represents the value of α. The blue region contains the values for which the condition (2.12) holds, and the yellow region contains the values for which the condition (2.13) holds. The intersection of both colors represents the region where both conditions are satisfied.

2.4. B-OGP In Bernoulli Group Testing. Combining Theorem 2.13 with Theorem 2.11 lets us directly conclude that $\phi(\ell)$ is increasing for small ℓ/k. Moreover, notice that $\phi(k) = 0$ by the definition of σ^*. Combining this fact with Theorems 2.11 and 2.13, with α and C satisfying Assumptions 2.8, 2.10, 2.12, we can conclude that $\phi(\ell)$ is non-monotonic and in particular, using standard arguments in the literature, that b-OGP appears.

Theorem 2.14. *For an (α, C) instance of group testing, a valid choice of a from (2.4) and arbitrarily small $C_1, C_3 > 0$ satisfy Assumptions 2.8, 2.10, 2.12, then there exists $\delta > 0$ and $0 < \varepsilon_1 < \varepsilon_2$ such that for all ℓ with $\ell/k \in [\varepsilon_1, \varepsilon_2]$, we have a.a.s. as $n \to +\infty$, $\phi(\ell) - \phi(0) \geq \delta$. In particular, as $\phi(k) = 0$, b-OGP holds in this regime.*

Proof of Theorem 2.14. Using Assumption 2.8 we invoke Lemma 2.9 to conclude the existence of an $\varepsilon > 0$ such that the first moment function $y(x)$ exists for all $x \in [0, \varepsilon]$.

Setting $\varepsilon_1 = \varepsilon/3$ and $\varepsilon_2 = 2\varepsilon/3$, Assumption 2.8, Assumption 2.10 and Assumption 2.12 allows us to invoke Theorem 2.11 and Theorem 2.13 to give for some $C_4 > 0$ that, for any $\ell \in \{\ell : \ell/k \in (\varepsilon_1, \varepsilon_2)\}$, a.a.s. as $n \to +\infty$,

$$\begin{aligned} &\text{Using Theorem 2.13,} && \phi(\ell) - \phi(0) \geq y(\ell/k) - y(0) - o(1) \\ &\text{Using Theorem 2.11,} && \geq C_4 \frac{\ell}{k} \end{aligned}$$

We can set $\delta = C_4 \varepsilon_1 > 0$ to prove that for $\ell/k \in [\varepsilon_1, \varepsilon_2]$, we have $\phi(\ell) - \phi(0) \geq \delta$ a.a.s. as $n \to +\infty$. Hence we can conclude by Theorem 2.13 that a.a.s. as $n \to +\infty$

$$\min_{\ell:\ell/k \in [\varepsilon_1, \varepsilon_2]} \phi(\ell) \geq \phi(0) + \delta \geq y(0) + \delta/2.$$

Furthermore, choosing $\zeta_1 = \varepsilon_1$, $\zeta_2 = \varepsilon_2$ and $r = y(0) + \delta/2$ gives the b-OGP since $\phi(k) = 0$ and therefore, a.a.s. as $n \to +\infty$, it holds $\max\{\phi(0), \phi(k)\} = \phi(0) \leq r$. ■

Recall that in the Figures 6, 7, 8 we plotted the regions of α and C such that the required Assumptions 2.8, 2.10, 2.12 for Theorem 2.14 hold. Meaning that any pair (α, C) in each of these colored regions satisfies Theorem 2.14.

2.5. Implied Failure Of Markov Chains.

The primary motivation of this work is the performance of Markov chains in constructing an estimator $\hat{\sigma}$. Recall that our goal is to minimize the (normalized) Hamiltonian,

$$H(\sigma) = \#\text{ of positive tests non-covered by } \sigma/M,$$

over all k-subsets σ. We focus on "local" Markov chains, meaning the underlying neighborhood graph on the k-subsets of $[n]$ connects two subsets if and only if their Hamming distance equals to 2, i.e., the chain swaps one individual at every step. This neighborhood graph is also commonly referred to as the Johnson graph [37, p. 300]. A common stationary distribution from such a process is given by $\pi_\beta(\sigma) \propto \exp(-\beta H(\sigma))$, for a sufficiently large choice of β.

Definition 2.15. Let d_H be the Hamming distance on k-subsets. Given a group testing instance, we define the Glauber Dynamics over k-subsets and inverse temperature β to have transition kernel $P_\beta(\sigma, \sigma')$ given by,

$$P_\beta(\sigma, \sigma') = \begin{cases} \frac{1}{k(p-k)} \frac{\exp(-\beta H(\sigma'))}{\exp(-\beta H(\sigma')) + \exp(-\beta H(\sigma))} & \text{if } d_H(\sigma, \sigma') = 2, |\sigma| = k, \\ \sum_{\sigma': d_H(\sigma, \sigma') = 2} \frac{1}{k(p-k)} \frac{\exp(-\beta H(\sigma))}{\exp(-\beta H(\sigma')) + \exp(-\beta H(\sigma))} & \text{if } \sigma = \sigma' \\ 0 & \text{otherwise.} \end{cases}$$

Using now also standard bottleneck arguments in the literature [33, 49], we conclude via the existence of b-OGP that all local MCMC methods sampling from π_β for β large enough, take a super-polynomial time to recover σ^*. This result is formally described in the following theorem and is the main contribution of this work, answering the main question of [40].

Corollary 2.16. *For an (α, C) instance of group testing, a valid choice of a from (2.4) and arbitrarily small $C_1, C_3 > 0$ satisfy Assumptions 2.8, 2.10, 2.12, then there exists $\varepsilon_1, \varepsilon_2 \in (0, 1)$ with $\varepsilon_1 < \varepsilon_2$ and an ε_1 dependent constant $C_\varepsilon > 0$ such that if $\beta \geq C_\varepsilon k \log(p/k)$ the following holds a.a.s. as $n \to +\infty$.*

For any local Markov chain on the Johnson graph with stationary distribution π_β, there exists an initialization for which the Markov chain requires at least $\exp(\Omega(k \log(p/k)))$ iterations to reach any k-subset σ with $|\sigma \cap \sigma^*| \geq \varepsilon_2 k$.

The proof of the corollary is given in [51, Section 10].

CHAPTER 2

Universality Of Non-Seperable Approximate Message Passing

Disclamer: This project is ongoing joint work with **Zhou Fan** and **Tianhao Wang**. This manuscript is in the final stages of editing and will—hopefully—be submitted soon. For proofs of many of these results I will simply cite [27]. I pray the location of these references do not change too much.

Notation. Scalars are denoted in regular font, while vectors, matrices, and tensors are bold, with lowercase for vectors and uppercase for higher-order tensors. For example, an element of a vector $\mathbf{v}$ is v_1.

For a function $f : \mathbb{R}^{n\times q} \to \mathbb{R}^n$, we write $\partial_{i,j} f(X) := \frac{\partial f(X)_i}{\partial X_{ij}}$, i.e., the derivative of the i-th component of f with respect to the (i,j) entry of X. Given partitions π, τ of $[n]$, we write $\tau \geq \pi$ (or equivalently $\pi \leq \tau$) if every block of π is contained in a block of τ. We denote the number of blocks in π by $|\pi|$, and for any $i \in [n]$, let $[i]$ be the block in π containing i. The identity matrix is denoted by Id.

For a matrix $\mathbf{X}$, we define the projection onto its column space as $P_{\mathbf{X}} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top$, with $P_{\mathbf{X}}^\perp = \mathrm{Id} - P_{\mathbf{X}}$ for its orthogonal complement.

To simplify indexing when subscripts become cumbersome, we use Python-style notation: entries of $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{W} \in \mathbb{R}^{n\times n}$ are written as $x_i \equiv x[i]$ and $W_{ij} \equiv W[i,j]$. More generally, for a matrix $\mathbf{\Sigma}$:

$$\Sigma_{i_1:i_2, j_1:j_2} = \Sigma[i_1{:}i_2, j_1{:}j_2] = \begin{pmatrix} \Sigma_{i_1 j_1} & \cdots & \Sigma_{i_1 j_2} \\ \vdots & \ddots & \vdots \\ \Sigma_{i_2 j_1} & \cdots & \Sigma_{i_2 j_2} \end{pmatrix} \in \mathbb{R}^{(i_2-i_1)\times(j_2-j_1)}.$$

The notation $\sigma_{\min}$ or $\lambda_{\min}$ refers to the minimum singular value of a matrix, while $\sigma_{\max}$ or $\lambda_{\max}$ refer to the maximum. We use the shorthand $\mathbf{x}_{1:k} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and $\mathbf{X}_{1:k} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$. Finally, we denote the number of connected components in a graph G as $\mathbf{c}(G)$ and use $\overset{d}{=}$ to signify equality in distribution.

1. A Brief Reflection On Approximate Message Passing

Approximate Message Passing (AMP) is a class of iterative algorithms whose study has been a major focus in statistical and computational research in the past 15 years. AMP originates from applications in compressed sensing [23] and relies on an ingenious conditioning technique found by [12] to study a set of iterative solutions for the TAP equations [74] on the Sherrington-Kirkpatrick model (see [59] for an introduction to this model and the mean-field approach).

To define an AMP algorithm, let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a symmetric random matrix, $\mathbf{u}_1 \in \mathbb{R}^n$ a deterministic initialization vector, and $\mathcal{F} = \{f_2, f_3, f_4, \ldots\}$ a sequence of functions where $f_{t+1} : \mathbb{R}^{n\times t} \to \mathbb{R}^n$. We then consider the following iteration,

$$\begin{aligned} \mathbf{z}_t &= \mathbf{W}\mathbf{u}_t - \sum_{s=1}^{t-1} b_{ts}\mathbf{u}_s \\ \mathbf{u}_{t+1} &= f_{t+1}(\mathbf{z}_1, \ldots, \mathbf{z}_t). \end{aligned} \tag{1.1}$$

The coefficients $\{b_{ts}\}_{s<t}$ are scalar *Onsager correction* terms whose values are deferred to Definition 2.1. Although iteration (1.1) seems rather abstract, it can be cast into a wide variety of inference and optimization problems, as we see soon.

The benefit of using algorithm (1.1) is that under the assumptions on the random matrix $\mathbf{W}$ and the functions $f_2, f_3, f_4, \ldots$, the empirical distribution of the iterates $z_{1,i}, z_{2,i}, \cdots$ "look" like i.i.d draws from a mean zero Gaussian random variable with variance determined by the choice of $\mathbf{u}_1$ and $\mathcal{F}$.

Definition 1.1. If $\mathbf{W} \sim \mathrm{GOE}(n)$ has $W[i,j] \sim \mathcal{N}(0, 1/n)$ for $i < j$ and $W[i,i] \sim \mathcal{N}(0, 2/n)$ then we say it has $\mathrm{GOE}(n)$ law.

Under the assumption that each $f \in \mathcal{F}$ is Lipschitz and separable, meaning, with $\mathbf{Z} \in \mathbb{R}^{n\times t}$, that $f(\mathbf{Z}) = (\mathring{f}(\mathbf{Z}[1,:]), \ldots, \mathring{f}(\mathbf{Z}[n,:]))$ for a Lipschitz function $\mathring{f} : \mathbb{R}^t \to \mathbb{R}$, and $\mathbf{W}$ being $\mathrm{GOE}(n)$ in law, convergence of this empirical distribution is proven by [9] and later extended by [41]. An overview of these techniques can be found in [28].

Since its introduction, AMP has been used to rigorously analyze many inference problems in the—so called—**thermodynamic** limit where the number of observations and the number of parameters are divergent, yet their ratio is fixed. Alongside the previously mentioned compressed sensing [23], there has been applications in generalized linear models under the GAMP [67] and VAMP [68] extensions to AMP, analysis of high dimensional robust M-estimation [22], and many more [28]. Moreover, AMP is the backbone of the analysis of more general classes of algorithms, two examples are: (1) the set of *first order methods* [14, 13, 61] where AMP can be mapped injectively to any such algorithm, and (2) the famed series of *tensor programs* [78, 79, 80, 81] which characterizes the limiting weights of multi-layer machine learning architectures in the "mean-field" limit. In addition, AMP is conjectured to be an optimal algorithm for many inference problems, previous work has verified the connection of AMP to Bayes optimal estimation [48], low-degree methods [60], and—even when AMP is found to be non-optimal—higher order message passing analogues close this gap [76]. It has also been shown in [44] that unrolled denoising networks converge to a special choice of $\mathcal{F}$ called *Bayes-AMP* [28].

Due to the vast set of problems AMP can be used to solve, many lines of research have been devoted to relaxing the conditions on both the **function class** $\mathcal{F}$ and the **random matrix** $\mathbf{W}$. We discuss the former now and return to the latter in a moment.

1.1. Relaxation of the Function Class $\mathcal{F}$.

Differing assumptions on $\mathcal{F}$ can be found as early as [8] where they considered $\mathcal{F}$ to be a subset of the set of separable polynomials. To our knowledge, The most general set of assumptions—before the results provided in this chapter—were first found in [10], considered general Lipschitz functions $f : \mathbb{R}^n \to \mathbb{R}^n$ (which do not need to be separable) to make up to class $\mathcal{F}$. These results were further extended to graph valued functions in [34]. An application of this **non-separable** AMP on the analysis of sliding window convolution algorithms can be found in [52].

1.2. Relaxation of the Random Matrix $\mathbf{W}$.

Perhaps the most prevalent relaxation of the random matrix assumption is that $\mathbf{W}$ is a Wigner matrix.

Definition 1.2. $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a **Wigner matrix** if $\mathbf{W}$ is symmetric, $(W[i,j] : 1 \le i \le j \le n)$ are independent, and there is a constant $C_k > 0$ for each $k = 2, 3, \ldots$ (not depending on n) such that

- $\mathbb{E}W[i,j] = 0$ for all $i, j \in [n]$
- $\mathbb{E}W[i,j]^2 = 1/n$ for all $i \ne j \in [n]$, and $\mathbb{E}W[i,i]^2 \le C_2/n$ for all $i \in [n]$
- $\mathbb{E}|W[i,j]|^k \le C_k n^{-k/2}$ for each $k \ge 3$ and all $i, j \in [n]$.

Recall, as a special case, $\mathbf{W} \sim \mathrm{GOE}(n)$ when $W[i,j] \sim \mathcal{N}(0, 1/n)$ for $i < j$ and $W[i,i] \sim \mathcal{N}(0, 2/n)$.

Results on the limiting behavior of the iterates $\mathbf{Z} = \mathbf{z}_1, \mathbf{z}_2, \dots$ from (1.1) have shown that the empirical law of the rows of $\mathbf{Z}$ converges to the identical limit if $\mathbf{W}$ was GOE(n) in law. For this reason, results of this kind are called **universality** results. The first instance of such an argument dates back to the seminal work of [8] which was the first to prove the universality of separable polynomial AMP. This work introduced the now common technique of **unrolling** the algorithm (1.1) into a linear combination of computational graphs. Years later, the separable Lipschitz case was then argued by [15] using an interpolation technique.

Different relaxations of $\mathbf{W}$ exists for *orthogonally invariant* matrices [26], for *semi-random* matrices [25], and *rotationally invariant* matrices [75]. The third results involved a method similar to [8] where they consider a polynomial approximation of an AMP algorithm and then show the unrolled polynomial AMP has a limiting value only dependent on the first and second moment of $\mathbf{W}$. These sequence of universality results also share a strong connection to computational graphs of matrices [55] and the literature on (traffic) free probability [56, 53].

The benefit of this relaxation on $\mathbf{W}$ is the added set of applications that can now be analyzed using AMP. This idea was used, implicitly[1], in [21] and has found applications in Qualitative group testing (a generalization of the problem from Chapter 1) analyzing the limiting number of tests to achieve some desired level of estimation error [71, 72].

The main goal of this work is to “glue” these two lines of research together, i.e. we answer the following question:

Does there exist abstract conditions on $\mathcal{F}$ and $\mathbf{W}$ that encompasses both the relaxations of the function class and the random matrix described above?

For the case of $\mathcal{F}$ being non-separable satisfying Definition 2.18 and $\mathbf{W}$ being Wigner, we answer the question in the affirmative.

2. Main Contributions

It is shown in [10] that when $\mathbf{W} \sim \text{GOE}(n)$, under certain asymptotic conditions for the sequence of functions $\{f_t\}$, the iterates $\{\mathbf{z}_t\}$ of the above AMP algorithm are characterized by a Gaussian state evolution. A version of this state evolution is reviewed in the following definition.

Definition 2.1. Let $\mathbf{\Sigma}_1 = \|\mathbf{u}_1\|_2^2/n \in \mathbb{R}^{1\times 1}$. For each $t \geq 1$, given $\mathbf{\Sigma}_t \in \mathbb{R}^{t\times t}$, let $[\mathbf{Z}_1, \dots, \mathbf{Z}_t] \in \mathbb{R}^{n\times t}$ have i.i.d. rows with distribution $\mathcal{N}(0, \mathbf{\Sigma}_t)$, and define $\mathbf{\Sigma}_{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}$ by

$$\mathbf{\Sigma}_{t+1}[1{:}t, 1{:}t] = \mathbf{\Sigma}_t, \qquad \mathbf{\Sigma}_{t+1}[1, t{+}1] = \mathbf{\Sigma}_{t+1}[t{+}1, 1] = \frac{1}{n}\mathbb{E}[f_{t+1}(\mathbf{Z}_{1:t})^\top \mathbf{u}_1]$$

$$\mathbf{\Sigma}_{t+1}[s{+}1, t{+}1] = \mathbf{\Sigma}_{t+1}[t{+}1, s{+}1] = \frac{1}{n}\mathbb{E}[f_{t+1}(\mathbf{Z}_{1:t})^\top f_{s+1}(\mathbf{Z}_{1:s})] \text{ for } j = 1, \dots, t.$$

Then the state evolution corresponding to (1.1) is given by the sequence of laws $\mathcal{N}(0, \mathbf{\Sigma}_t)$.

The Onsager correction terms $\{b_{ts}\}_{s<t}$ in (1.1) are defined as

$$b_{ts} = \frac{1}{n}\mathbb{E}[\text{div}_s\, f_t(\mathbf{Z}_1, \dots, \mathbf{Z}_{t-1})],$$

where div_s is the divergence with respect to the s-th column of the input.

The above definitions differ slightly from [10] in that we define $\mathbf{\Sigma}_t$ and $\{b_{ts}\}_{s<t}$ to be n-dependent, instead of assuming that these quantities have asymptotic limits as $n \to \infty$.

[1] This paper essentially proved a very small universality result comparing a specific AMP algorithm when changing the law of $\mathbf{W}$ from GOE(n) to a two point valued distribution.

2.1. State Evolution For Gaussian matrices. We first provide a stronger form of the state evolution guarantee shown in [10] when $\mathbf{W} \sim \mathrm{GOE}(n)$, for functions $\{f_t\}$ that may be non-Lipschitz and instead have polynomial growth.

REMARK 2.2. Throughout, for notational convenience, we will identify the initialization $\mathbf{u}_1 \equiv f_1(\cdot)$ as the output of a constant function f_1, and understand $f_{s+1}(\mathbf{z}_{1:s})$ for $s = 0$ as $\mathbf{u}_1$.

ASSUMPTION 2.3. For each fixed $t \geq 1$:

(a) There are constants $C_t, c_t > 0$ such that $c_t < \lambda_{\min}(\mathbf{\Sigma}_t) \leq \lambda_{\max}(\mathbf{\Sigma}_t) < C_t$ and $\max_{s<t} |b_{ts}| < C_t$ for all n.
(b) If $\mathbf{Z}_{1:t} \in \mathbb{R}^{n\times t}$ has i.i.d. rows with distribution $\mathcal{N}(0, \mathbf{\Sigma}_t)$ and $\mathbf{E}_{1:t} \in \mathbb{R}^{n\times t}$ is any random matrix in the probability space of $\mathbf{Z}_{1:t}$ such that

$$\mathbb{P}[\|\mathbf{E}_{1:t}\|_F \geq (\log n)^{C_0}] \leq n^{-(1+\varepsilon)}$$

for some constants $C_0, \varepsilon > 0$, then for each $s = 0, \ldots, t$, with probability at least $1 - n^{-d}$ (for any $d > 1$), there exists a constant C (only dependent on d, C_0, ε, t) such that

$$\frac{1}{n}\Big|f_{t+1}(\mathbf{Z}_{1:t}+\mathbf{E}_{1:t})^\top f_{s+1}(\mathbf{Z}_{1:s}+\mathbf{E}_{1:s}) - \mathbb{E}[f_{t+1}(\mathbf{Z}_{1:t})^\top f_{s+1}(\mathbf{Z}_{1:s})]\Big| \leq \frac{C\log^C(n)}{\sqrt{n}}, \tag{2.1}$$
$$\frac{1}{n}\Big\|(\mathbf{Z}_{1:t}+\mathbf{E}_{1:t})^\top f_{s+1}(\mathbf{Z}_{1:s}+\mathbf{E}_{1:s}) - \mathbb{E}[\mathbf{Z}_{1:t}^\top f_{s+1}(\mathbf{Z}_{1:s})]\Big\|_2 \leq \frac{C\log^C(n)}{\sqrt{n}}.$$

The following proposition shows that Assumption 2.3 holds in the case where $\{f_t\}$ are Lipschitz, or more generally, in many settings where $\{f_t\}$ are only locally Lipschitz with polynomial growth.

PROPOSITION 2.4. *Suppose $\|\mathbf{u}_1\|_2 \leq (\log n)^{C_1}\sqrt{n}$. Suppose furthermore, for each $t \geq 1$ and any $C_0, \varepsilon > 0$, there exists a n-dependent convex set $\mathcal{A} \subseteq \mathbb{R}^{n\times\infty}$ and constants $C_1, \delta > 0$ depending only on C_0, ε, t such that, for each $\mathcal{A}_t \subset \{\mathbf{x} \in \mathbb{R}^{n\times t} : \exists \mathbf{y} \in \mathbb{R}^{n\times\infty}$ with $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}\}$,*

(1) With $\mathbf{z} \in \mathbb{R}^{n\times t}$, $\max_{\mathbf{z}\in\mathcal{A}_t} \|\mathbf{z}\|_2 \leq (\log n)^{C_1}\sqrt{n}$.
(2) $\|f_{t+1}(\mathbf{z})\|_2 \leq (\log n)^{C_1}\sqrt{n}$ and $\|f_{t+1}(\mathbf{z}) - f_{t+1}(\mathbf{z}')\|_2 \leq (\log n)^{C_1}\|\mathbf{z}-\mathbf{z}'\|_2$ for all $\mathbf{z}, \mathbf{z}' \in \mathcal{A}_t$.
(3) If $\mathbf{Z} \in \mathbb{R}^{n\times t}$ has i.i.d. rows with distribution $\mathcal{N}(0, \mathbf{\Sigma})$ for any $\mathbf{\Sigma} \in \mathbb{R}^{t\times t}$ satisfying $\|\mathbf{\Sigma}\|_{\mathrm{op}} \leq C_0$, and $\mathbf{E} \in \mathbb{R}^{n\times t}$ is any matrix in the probability space of $\mathbf{Z}$ satisfying

$$\mathbb{P}[\|\mathbf{E}\|_F \geq (\log n)^{C_0}] \leq n^{-(1+\varepsilon)},$$

then

$$\mathbb{E}\|f_{t+1}(\mathbf{Z})\|_2^4 \leq C_1 n^2, \qquad \mathbb{P}[\mathbf{Z}+\mathbf{E} \notin \mathcal{A}_t] \leq n^{-(1+\delta)}.$$

Then Assumption 2.3 holds.

The proof of this result is given in [27, Apppendix A].

EXAMPLE 2.5. *Below is an example for how to check the conditions of Proposition 2.4. Let $g(z) = (z+\sin(5z))^2$ and let $f : \mathbb{R}^n \to \mathbb{R}^n$ be the function applying g to each coordinate of the input. Let $\mathcal{A} = [-C\log^C(n), C\log^C(n)]^n$, where $C > 0$ (dependent on C_0 from Proposition 2.4) can be chosen large enough to satisfy $\mathbb{P}(\mathbf{Z}+\mathbf{E}) \in \mathcal{A}$ described in Proposition 2.4. From this choice of $\mathcal{A}$, $\max_{\mathbf{z}\in\mathcal{A}} \|\mathbf{z}\|_2 \leq C\log^C(n)\sqrt{n}$ and $\|f(\mathbf{z})\|_2 \leq (C\log^C(n)+1)^2\sqrt{n}$ for all $\mathbf{z} \in \mathcal{A}$. Moreover, $\frac{d}{dz_i}f(z)_i = 2(z_i+\sin(5z_i))(1+5\cos(5z_i))$ which is bounded by a poly-log(n) factor for any $\mathbf{z} \in \mathcal{A}$. With $\mathbf{Z}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Sigma)$, the expectation $\mathbb{E}[\|f(\mathbf{Z})\|_2^4] \leq \sum_{i,j=1}^n \mathbb{E}[(1+Z_i)^2(1+Z_j)^2] \leq C'n^2$, for a constant $C' > 0$ dependent on C_0, leading to each condition in Proposition 2.4 satisfied.*

Under Assumption 2.3, for $\mathbf{W} \sim \mathrm{GOE}(n)$, we can relate the iterates $\mathbf{z}_1, \ldots, \mathbf{z}_t$ to random variables defined by the state evolution matrix $\mathbf{\Sigma}_t$.

Theorem 2.6 (Strong GOE State Evolution). *Suppose* $\mathbf{W} \sim \mathrm{GOE}(n)$*, and Assumption 2.3 holds for the state evolution sequence* $\{\mathbf{\Sigma}_t\}$ *and Onsager terms* $\{b_{ts}\}$ *of Definition 2.1.*

Then for any fixed $t \geq 1$*, the AMP iterates* $\mathbf{z}_1, \ldots, \mathbf{z}_t$ *may be written as*

$$[\mathbf{z}_1, \ldots, \mathbf{z}_t] = [\mathbf{Z}_1, \ldots, \mathbf{Z}_t] + [\mathbf{E}_1, \ldots, \mathbf{E}_t],$$

where $[\mathbf{Z}_1, \ldots, \mathbf{Z}_t] \in \mathbb{R}^{n \times t}$ *has i.i.d. rows with distribution* $\mathcal{N}(0, \mathbf{\Sigma}_t)$*, and the residual term satisfies* $\|[\mathbf{E}_1, \ldots, \mathbf{E}_t]\|_{\mathrm{F}} \prec 1$*.*

We call Theorem 2.6 a "strong state evolution", as the error bound $\|[\mathbf{E}_1, \ldots, \mathbf{E}_t]\|_{\mathrm{F}} \prec 1$ is stronger than what is needed to show the usual state evolution guarantee of approximation of the empirical distribution of rows of $\mathbf{z}_{1:t}$ with vanishing error under a metric of weak convergence. (For this, a bound of $\|[\mathbf{E}_1, \ldots, \mathbf{E}_t]\|_{\mathrm{F}} \prec n^{1/2-\varepsilon}$ would suffice.) This result is proven in [27, Section 3]. This strong state evolution implies the convergence of a large class of test functions evaluated at the AMP iterates, stated in the following corollary.

Corollary 2.7. *Under the setting of Theorem 2.6, fix any* $t \geq 1$ *and consider a function* $\phi : \mathbb{R}^{n \times t} \to \mathbb{R}$ *given by*

$$\phi(\mathbf{z}_1, \ldots, \mathbf{z}_t) = \frac{1}{n} \phi_1(\mathbf{z}_1, \ldots, \mathbf{z}_t)^\top \phi_2(\mathbf{z}_1, \ldots, \mathbf{z}_t), \qquad \phi_1, \phi_2 : \mathbb{R}^{n \times t} \to \mathbb{R}^n, \tag{2.2}$$

where the functions ϕ_1 *and* ϕ_2 *satisfy (2.1) from Assumption 2.3 with* ϕ_1 *and* ϕ_2 *in the place of* f_{t+1} *and* f_{s+1} *respectively. Then*

$$\phi(\mathbf{z}_1, \ldots, \mathbf{z}_t) - \mathbb{E}[\phi(\mathbf{Z}_1, \ldots, \mathbf{Z}_t)] \prec n^{-1/2}.$$

2.2. Universality For Tensor Networks. When the functions $\{f_t\}$ of (1.1) and test functions ϕ_1, ϕ_2 of (2.2) are polynomials, the value

$$\frac{1}{n} \phi_1(\mathbf{z}_1, \ldots, \mathbf{z}_t)^\top \phi_2(\mathbf{z}_1, \ldots, \mathbf{z}_t)$$

may be expressed as a linear combination of contracted values of tensor networks, constructed from a class of deterministic tensors $\mathcal{T}$ and the Wigner matrix $\mathbf{W}$. We describe in this section an abstract definition of such a network and a condition for universality of its contracted value.

Let $p : \mathbb{R}^{n \times t} \to \mathbb{R}^n$ be a polynomial function of maximum total degree $D \geq 0$ (i.e. each coordinate of the output of p is a multivariate polynomial of total degree at most D in its inputs in $\mathbb{R}^{n \times t}$). For each $d = 1, \ldots, D$, let $\mathcal{S}_{t,d}$ be the collection of all mappings $\sigma : [t] \to [d]$. Then there exist tensors $\mathbf{T}^{(0)} \in \mathbb{R}^n$ and $\mathbf{T}^{(\sigma)} \in (\mathbb{R}^n)^{\otimes(d+1)}$ for each $d = 1, \ldots, D$ and each $\sigma \in \mathcal{S}_{t,d}$, such that

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_t) = \mathbf{T}^{(0)} + \sum_{d=1}^{D} \sum_{\sigma \in \mathcal{S}_{t,d}} \mathbf{T}^{(\sigma)}[\mathbf{z}_{\sigma(1)}, \ldots, \mathbf{z}_{\sigma(d)}, \cdot] \tag{2.3}$$

We write $\mathbf{T}[\mathbf{z}_1, \ldots, \mathbf{z}_k, \cdot] \in \mathbb{R}^n$ to denote the partial contraction whose j^{th} coordinate is given by $\sum_{i_1, \ldots, i_k \in [n]} T[i_1, \ldots, i_k, j] z_1[i_1] \ldots z_k[i_k]$. Thus $\mathbf{T}^{(0)}$ is the constant term of p, and $\{\mathbf{T}^{(\sigma)}\}_{\sigma \in \mathcal{S}_{t,d}}$ represent the terms of p of degree d.

Definition 2.8. Given a collection of tensors

$$\mathcal{T} \subseteq \bigsqcup_{k \geq 1} (\mathbb{R}^n)^{\otimes k}$$

of any orders $k \geq 1$, a polynomial $p : \mathbb{R}^{n \times t} \to \mathbb{R}^n$ is **$\mathcal{T}$-representable** if it admits a representation (2.3) where $\mathbf{T}^{(0)} \in \mathbb{R}^n \cap \mathcal{T}$ and $\mathbf{T}^{(\sigma)} \in (\mathbb{R}^n)^{\otimes(d+1)} \cap \mathcal{T}$ for each $d = 1, \ldots, D$ and $\sigma \in \mathcal{S}_{t,d}$.

We note that these tensors $\mathbf{T}^{(0)}, \mathbf{T}^{(\sigma)}$ are, in general, not symmetric in their arguments. Furthermore, for any given polynomial function p, the choice of tensors $\{\mathbf{T}^{(\sigma)}\}_{\sigma \in \mathcal{S}_{t,d}}$ that represent it in (2.3) is not unique.

Definition 2.9. An **ordered multigraph** $G = (\mathcal{V}, \mathcal{E})$ is an undirected multigraph with vertices $\mathcal{V}$ and edges $\mathcal{E}$, having no self-loops and no isolated vertices, and having a specified ordering $e_1, \ldots, e_{\deg(v)}$ of the edges incident to each vertex $v \in \mathcal{V}$. Here, $\deg(v)$ is the **degree** of v, defined as the total number of edges incident to v counting multiplicity. G is **connected** if it consists of a single connected component.

A **tensor labeling** $\mathcal{L}$ of G is an assignment of a tensor $\mathbf{T}_v \in (\mathbb{R}^n)^{\otimes \deg(v)}$ to each vertex $v \in \mathcal{V}$, where the order of $\mathbf{T}_v$ equals the degree of v. We call $(G, \mathcal{L})$ a **tensor network**. The **value** of this tensor network is

$$\mathrm{val}_G(\mathcal{L}) = \sum_{\mathbf{i} \in [n]^{\mathcal{E}}} \prod_{v \in \mathcal{V}} \mathbf{T}_v[i_e : e \sim v] \tag{2.4}$$

where $[i_e : e \sim v]$ denotes the ordered tuple of indices $[i_{e_1}, \ldots, i_{e_{\deg(v)}}]$, and $e_1, \ldots, e_{\deg(v)}$ are the ordered edges incident to v.

When G is connected, the value $\mathrm{val}_G(\mathcal{L})$ may be understood as the scalar value obtained by contracting (in any order) the tensor-tensor product associated to each edge. When G consists of multiple connected components, $\mathrm{val}_G(\mathcal{L})$ factorizes as the product of these values across the different components. We clarify that the specification of an edge ordering is needed to define this value, because the tensors $\{\mathbf{T}_v\}_{v \in \mathcal{V}}$ may not be symmetric.

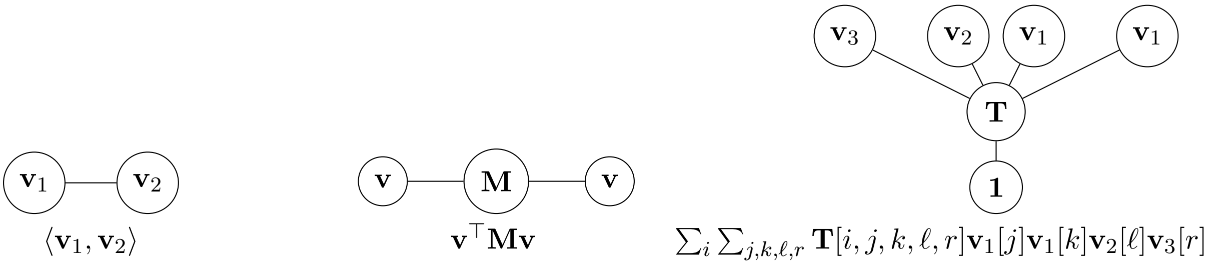

$\mathbf{v}_1$
$\mathbf{v}_2$
$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$
$\mathbf{v}$
$\mathbf{M}$
$\mathbf{v}$
$\mathbf{v}^\top \mathbf{M} \mathbf{v}$
$\mathbf{v}_3$
$\mathbf{v}_2$
$\mathbf{v}_1$
$\mathbf{v}_1$
$\mathbf{T}$
$\mathbf{1}$
$\sum_i \sum_{j,k,\ell,r} \mathbf{T}[i,j,k,\ell,r]\mathbf{v}_1[j]\mathbf{v}_1[k]\mathbf{v}_2[\ell]\mathbf{v}_3[r]$

Figure 1. Some example tensor networks and their corresponding value (note the rightmost image is under some implicit ordering of the edges).

In our asymptotic analyses, the multigraphs G will always be fixed and independent of n, while the tensor labels $\mathbf{T}_v$ will be n-dependent.

Definition 2.10. Given a collection of deterministic tensors $\mathcal{T} \subseteq \bigsqcup_{k \geq 1} (\mathbb{R}^n)^{\otimes k}$ and a random matrix $\mathbf{W} \in \mathbb{R}^{n \times n} \equiv (\mathbb{R}^n)^{\otimes 2}$, a $(\mathcal{T}, \mathbf{W})$**-labeling** of G is a tensor labeling of G such that each tensor label $\mathbf{T}_v$ belongs to $\mathcal{T} \cup \{\mathbf{W}\}$ (where only vertices with $\deg(v) = 2$ are labeled by $\mathbf{W}$), and such that no two adjacent vertices connected by an edge are both labeled by $\mathbf{W}$.

The relation of Definition 2.9 to the AMP algorithm (1.1) is clarified by the following consequence of (2.3), which we prove in [27, Appendix B]. (Here f_1 is a constant function identified with the initialization $\mathbf{u}_1$, c.f. Remark 2.2.)

Lemma 2.11. *Fix any $t \geq 1$, and suppose that $f_1, \ldots, f_t$ and the test functions ϕ_1, ϕ_2 defining ϕ in (2.2) are $\mathcal{T}$-representable polynomial functions with maximum total degree D independent of n. Suppose also* $\max_{r<s\leq t} |b_{sr}| < C_t$ *for a constant $C_t > 0$.*

Then there exist constants $C, M > 0$ depending only on (t, D, C_t), a list of connected ordered multigraphs $G_1, \ldots, G_M$ independent of n, and $(\mathcal{T}, \mathbf{W})$-labelings $\mathcal{L}_1, \ldots, \mathcal{L}_M$ of $G_1, \ldots, G_M$ together with coefficients $a_1, \ldots, a_M \in \mathbb{R}$ with each $|a_m| < C$, such that

$$\phi(\mathbf{z}_1, \ldots, \mathbf{z}_t) = \sum_{m=1}^{M} \frac{a_m \mathrm{val}_{G_m}(\mathcal{L}_m)}{n}.$$

Let us denote by $\mathrm{Id}_k \in (\mathbb{R}^n)^{\otimes k}$ the "identity" diagonal tensor with entries

$$\mathrm{Id}_k[i_1, \ldots, i_k] = \mathbb{1}\{i_1 = \cdots = i_k\}.$$

We omit the subscript $\mathrm{Id} \equiv \mathrm{Id}_k$ when its order is clear from context. We consider the following Bounded Composition Property for a class of deterministic tensors $\mathcal{T}$.

Definition 2.12. An ordered multigraph $G = (\mathcal{V}_{\mathrm{Id}} \sqcup \mathcal{V}_T, \mathcal{E})$ is **bipartite** if its vertex set is the disjoint union of two sets $\mathcal{V}_{\mathrm{Id}}, \mathcal{V}_T$, and each edge of $\mathcal{E}$ connects a vertex of $\mathcal{V}_{\mathrm{Id}}$ with a vertex of $\mathcal{V}_T$.

A $(\mathrm{Id}, \mathcal{T})$**-labeling** $\mathcal{L}$ of a bipartite ordered multigraph G is a tensor labeling of G such that each vertex $u \in \mathcal{V}_{\mathrm{Id}}$ is labeled with $\mathrm{Id}_{\deg(u)}$, and each vertex $v \in \mathcal{V}_T$ has a label $\mathbf{T}_v \in \mathcal{T}$.

Definition 2.13. A collection of deterministic tensors $\mathcal{T} \subseteq \bigsqcup_{k\geq 1}(\mathbb{R}^n)^{\otimes k}$ satisfies the **Bounded Composition Property (BCP)** if the following holds:

Let $G = (\mathcal{V}_{\mathrm{Id}} \sqcup \mathcal{V}_T, \mathcal{E})$ be any bipartite ordered multigraph (independent of n) such that G is connected and all vertices in $\mathcal{V}_{\mathrm{Id}}$ have even degree. Then there exists a constant $C := C(G) > 0$ independent of n such that for any $(\mathrm{Id}, \mathcal{T})$-labeling $\mathcal{L}$ of G,

$$|\mathrm{val}_G(\mathcal{L})| \leq Cn.$$

Remark 2.14. The BCP is a nuanced property of the class of tensors $\mathcal{T}$, of particular interest is the fact that BCP implies that Assumption 2.3 holds for whatever class of functions that are $\mathcal{T}$-representable. This result is proven in [27, Section 5.1].

Example 2.15. *If the functions $f_1, \ldots, f_t$ and ϕ_1, ϕ_2 in Lemma 2.11 are coordinate-separable polynomial functions with all coefficients bounded in magnitude by $B > 0$, then they are $\mathcal{T}$-representable by the class*

$$\mathcal{T} = \bigsqcup_{k\geq 1} \Big\{ \textit{diagonal tensors } \mathbf{T} \in (\mathbb{R}^n)^{\otimes k} \textit{ with } \max_{i=1}^{n} |\mathbf{T}[i, \ldots, i]| \leq B \Big\}.$$

This class satisfies the BCP: For any connected bipartite ordered multigraph $G = (\mathcal{V}_{\mathrm{Id}} \sqcup \mathcal{V}_T, \mathcal{E})$ and any $(\mathrm{Id}, \mathcal{T})$-labeling of G, it is easily checked that

$$\mathrm{val}_G(\mathcal{L}) = \sum_{i=1}^{n} \prod_{v \in \mathcal{V}_T} \mathbf{T}_v[i, \ldots, i]$$

and hence $|\mathrm{val}_G(\mathcal{L})| \leq B^{|\mathcal{V}_T|} n$. We will discuss examples of non-coordinate-separable functions in the following sections.

The following theorem establishes universality of $\mathrm{val}_G(\mathcal{L})$ under the above BCP condition for the tensor class $\mathcal{T}$. Its proof is given in Section 3.

Theorem 2.16 (Universality of tensor network value). *Let $\mathcal{T} \subseteq \bigsqcup_{k\geq 1}(\mathbb{R}^n)^{\otimes k}$ be a class of tensors satisfying BCP, and let $\mathbf{W}, \mathbf{W}'$ be two Wigner matrices satisfying Definition 1.2. Fix any connected ordered multigraph G independent of n, let $\mathcal{L}$ be a $(\mathcal{T}, \mathbf{W})$-labeling of G, and let $\mathcal{L}'$ be the $(\mathcal{T}, \mathbf{W}')$-labeling that replaces $\mathbf{W}$ by $\mathbf{W}'$. Then almost surely,*

$$\lim_{n\to\infty} \left(\frac{1}{n} \mathrm{val}_G(\mathcal{L}) - \frac{1}{n} \mathrm{val}_G(\mathcal{L}') \right) = 0.$$

The following is an immediate corollary of the above theorem and Lemma 2.11.

Corollary 2.17. *Let $\mathcal{T} \subseteq \bigsqcup_{k\geq 1}(\mathbb{R}^n)^{\otimes k}$ be a class of tensors satisfying BCP, and let $\mathbf{W}, \mathbf{W}'$ be two Wigner matrices satisfying Definition 1.2.*

Fix any $t \geq 1$, and suppose that $f_1, \ldots, f_t$ and the test functions ϕ_1, ϕ_2 defining ϕ in (2.2) are $\mathcal{T}$-representable polynomial functions with maximum total degree D independent of n. Let $\mathbf{z}_{1:t}, \mathbf{z}'_{1:t}$ be the iterates of (1.1) defined by these functions and $\mathbf{W}, \mathbf{W}'$ respectively. Then almost surely,

$$\lim_{n\to\infty} \phi(\mathbf{z}_{1:t}) - \phi(\mathbf{z}'_{1:t}) = 0.$$

2.3. Universality Of Approximate Message Passing. We now apply the preceding universality of tensor networks to deduce universality for AMP algorithms (1.1) defined by functions satisfying the following polynomial approximability condition.

Definition 2.18. Let $\mathcal{F} = \bigsqcup_{t\geq 0} \mathcal{F}_t$ be a class of functions, where $\mathcal{F}_t$ consists of functions $f : \mathbb{R}^{n\times t} \to \mathbb{R}^n$ and $\mathcal{F}_0$ consists of constant vectors in $\mathbb{R}^n$. $\mathcal{F}$ is **BCP-approximable** if, for any fixed $C_0 > 0$, there exists a collection of tensors $\mathcal{T} \subseteq \bigsqcup_{k\geq 1} \mathbb{R}^{n^{\otimes k}}$ satisfying BCP for which the following holds:

Choose any $t \geq 0$, $f_t \in \mathcal{F}_t$, $\mathbf{\Sigma} \in \mathbb{R}^{t\times t}$ with $\|\mathbf{\Sigma}\|_{\text{op}} < C_0$, and any $\varepsilon > 0$. Let $\mathbf{Z} \in \mathbb{R}^{n\times t}$ have i.i.d. rows with distribution $\mathcal{N}(0, \mathbf{\Sigma})$. Then

(1) For a constant $D \geq 0$ depending only on C_0, ε, t, there exists a ($\mathbf{\Sigma}, n$-dependent) polynomial function $p_t : \mathbb{R}^{n\times t} \to \mathbb{R}^n$ of maximum total degree D that is $\mathcal{T}$-representable, such that

$$\frac{1}{n}\,\mathbb{E}\|f_t(\mathbf{Z}) - p_t(\mathbf{Z})\|_2^2 < \varepsilon.$$

(2) Let $\mathbf{z} \in \mathbb{R}^{n\times t}$ be any random matrix satisfying, for any fixed constant $D \geq 0$ and two (n-dependent) $\mathcal{T}$-representable polynomial functions $q_1, q_2 : \mathbb{R}^{n\times t} \to \mathbb{R}^n$ of maximum total degree D,

$$\lim_{n\to\infty} \frac{1}{n}\,\mathbb{E}\,q_1(\mathbf{z})^\top q_2(\mathbf{z}) - \frac{1}{n}\,\mathbb{E}\,q_1(\mathbf{Z})^\top q_2(\mathbf{Z}) = 0 \text{ almost surely.}$$

Then, for the above functions f_t and p_t,

$$\frac{1}{n}\,\mathbb{E}\|f_t(\mathbf{z}) - p_t(\mathbf{z})\|_2^2 < \varepsilon \text{ almost surely for all large } n.$$

Theorem 2.19. *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a Wigner matrix satisfying Definition 1.2. Let $\mathcal{F} = \bigsqcup_{t\geq 0} \mathcal{F}_t$ be a BCP-approximable class of functions. Suppose, for each $t \geq 1$, that $f_t \in \mathcal{F}_t$ and f_t is L-Lipschitz for some constant $L > 0$ independent of n.*

For any fixed $t \geq 1$, let $\phi_1, \phi_2 : \mathbb{R}^{n\times t} \to \mathbb{R}^n$ also satisfy $\phi_1, \phi_2 \in \mathcal{F}_t$ and ϕ_1, ϕ_2 are L-Lipschitz for some constant $L > 0$ independent of n. Let $\mathbf{\Sigma}_t \in \mathbb{R}^{t\times t}$ be the state evolution matrix in Definition 2.1, let $\mathbf{Z}_{1:t} \in \mathbb{R}^{n\times t}$ have i.i.d. rows with distribution $\mathcal{N}(0, \mathbf{\Sigma}_t)$, and define

$$\phi(\mathbf{z}_{1:t}) = \frac{1}{n}\,\phi_1(\mathbf{z}_{1:t})^\top \phi_2(\mathbf{z}_{1:t}).$$

Then almost surely

$$\lim_{n\to\infty} \phi(\mathbf{z}_{1:t}) - \mathbb{E}\phi(\mathbf{Z}_{1:t}) \to 0.$$

2.4. Applications To Three Function Classes. This section presents three function classes within each of which the state evolution of the AMP algorithm (1.1) is universal for Wigner $\mathbf{W}$. We prove that these functions are BCP-approximable in [27, Section 6] and thus, by Remark 2.14, these functions satisfy Assumption 2.3. When writing these functions, we consider them mapping a set of m input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ to a single outpur vector in $\mathbb{R}^n$. The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ represent both side information vectors (including the nromalization vector $\mathbf{u}_1$) and the AMP iterates $\mathbf{z}_1, \ldots, \mathbf{z}_T$. More details on the side information vectors is given in [27, Section 6].

2.4.1. *Local Functions.* First, we consider a natural extension of seperable AMP algorithms, where, instead of assuming that class $\{f_{t+1}\}_{t\in[T-1]}$ is row-wise separable, we assume that each output index has a finite number of input row depeandencies. A similar class was analyzed in [**?**] where they considered "sliding window denoisers" which demonstrate such a dependency.

Definition 2.20. Consider the set of functions, for each $m \in \mathbb{N}$, where $f : \mathbb{R}^{n\times m} \to \mathbb{R}^n$, such that for each $i, j \in [n] \times [m]$, uniformly over $\mathbf{X} = (x_1, \ldots, x_m) \in \mathbb{R}^{n\times m}$ where there exists a constant $C > 0$ satisfying the following conditions:

(1) **Locality:** Define $\mathbf{arg}_f(j) = \{i : \mathbf{X}[i,:] \text{ is an argument in } f(\mathbf{X})[j]\}$. Similarly, define $\mathbf{arg}_f^{-1}(i) = \{j : i \in \mathbf{arg}_f(j)\}$. The function f satisfies $\max_{ij} |\mathbf{arg}_f^{-1}(i)| + |\mathbf{arg}_f(j)| \leq C$.

(2) **Polynomial Growth:** We have that $f(\mathbf{X})[j] \leq C\left(1 + \sum_{i\in\mathbf{arg}(j)} \|\mathbf{X}_{i\cdot}\|_1\right)^C$ and, for each $(k,\ell) \in [n] \times [t+k]$, $\partial_{k,\ell} f(\mathbf{X})[j] \leq C\left(1 + \sum_{i\in\mathbf{arg}(j)} \|\mathbf{X}_{i\cdot}\|_1\right)^C$, where $\partial_{k,\ell} f$ is the partial derivative of f with respect to the (k,ℓ)-th entry of $\mathbf{X}$.

This set of functions is named the **local functions**.

2.4.2. *Anisotropic Functions.* Another application of AMP algorithms arises with the existence of correlations in the random matrix, let $\hat{\mathbf{W}} \in \mathbb{R}^{n\times n}$ be such a matrix. Assume that $\hat{\mathbf{W}}$ has the form $\hat{\mathbf{W}} = \mathbf{K}_1\mathbf{W}\mathbf{K}_2$ where $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times n}$ are known matrices. Providing a state evolution guarentee for this matrix in (1.1) is difficult as the entries of $\hat{\mathbf{W}}$ are no longer independent. To remedie this issue, we pass the matrix multiplication in $\hat{\mathbf{W}}$ to the activations $\{f_{t+1}\}_{t\in[T-1]}$. For a given fucntion $f_{t+1} : \mathbb{R}^{n\times t} \to \mathbb{R}^n$, define $\hat{f}(\mathbf{z}_1,\ldots,\mathbf{z}_t) = \mathbf{K}_2 f_{t+1}(\mathbf{K}_1\mathbf{z}_1,\ldots,\mathbf{K}_1\mathbf{z}_t)$. Such a choice elicits the following AMP algorithm,

$$\begin{aligned} \mathbf{z}_t &= \mathbf{W}\mathbf{u}_t - \sum_{s=1}^t b_{ts}\mathbf{u}_s, \\ \mathbf{u}_{t+1} &= \hat{f}_{t+1}(\mathbf{z}_t). \end{aligned} \tag{2.5}$$

Under the change of variables $\hat{\mathbf{z}}_t = \mathbf{K}_1\mathbf{z}_t, \hat{\mathbf{u}}_t = \mathbf{K}_2\mathbf{u}_t$ and $\hat{b}_{ts} = b_{ts}\mathbf{K}_1\mathbf{K}_2$ give an equivalent algorithm to (2.5),

$$\begin{aligned} \hat{\mathbf{z}}_t &= \hat{\mathbf{W}}\mathbf{u}_t - \sum_{s=1}^t \hat{b}_{ts}\hat{\mathbf{u}}_s, \\ \hat{\mathbf{u}}_{t+1} &= f_{t+1}(\mathbf{z}_1,\ldots,\mathbf{z}_t,\boldsymbol{\gamma}). \end{aligned} \tag{2.6}$$

Thus, we can endow algorithm (2.6) with the equivalent state evolution of algorithm (2.5). Representing the conversion between these two algorithms is the following set of functions.

Definition 2.21. Consider the set, for each $m \in \mathbb{N}$, of functions $f : \mathbb{R}^{n\times m} \to \mathbb{R}^n$ such that, for each $i, j \in [n] \times [m]$ uniformly over $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_m) \in \mathbb{R}^{n\times m}$, there exists a constant $C > 0$ and function $g_j : \mathbb{R}^m \to \mathbb{R}, \mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{n\times n}$ satisfying the following conditions:

(1) **Anisotropic Form:** There exists a vector[2] $\mathbf{a} \in \{0,1\}^m$ such that f can be represented as,

$$f(\mathbf{X}) = \mathbf{K}_2 g((\mathbf{a}_1\mathbf{K}_1 + (1-\mathbf{a}_1)\mathrm{Id})\mathbf{X}[:,1],\ldots,(\mathbf{a}_m\mathbf{K}_1 + (1-\mathbf{a}_m)\mathrm{Id})\mathbf{X}[:,m]),$$

with $g = \{g_j\}_{j\in[n]}$ applied row-wise seperably.

(2) **Polynomial Growth:** We have that $|g_j(\mathbf{X}[i,:])| \leq C(1+\|\mathbf{X}_{i\cdot}\|_\infty)^C$ and $|\partial_j g_i(\mathbf{X}[j,:])| \leq C(1+\|\mathbf{X}_{i\cdot}\|_\infty)^C$. Moreover, assume that $C' \leq \|\mathbf{K}_1\|_{\mathrm{op}} \vee \|\mathbf{K}_2\|_{\mathrm{op}} \leq C$ for some constant $C' > 0$.

(3) $\|\mathbf{K}_1\|_1 \vee \|\mathbf{K}_1\|_\infty \vee \|\mathbf{K}_2\|_1 \vee \|\mathbf{K}_2\|_\infty \leq C$, or more generally, Assumption 6.2 in [27, Section 6].

This set of functions is named the **anisotropic functions**.

2.4.3. *Spectral Functions.* A final application of an AMP algorithm is when the iterates of (1.1) have a latent matrix structure. A clever application of Definition 2.20 or Definition 2.21 permits the use of convolution type operations, however, this class covers activations applied to the spectrum of said matrix.

[2] The role played by $\mathbf{a}$ in this definition is to allow some inputs to be multiplied by $\mathbf{K}_1$ and for some to be left unchanged.

Consider a vector iterate $\mathbf{x} \in \mathbb{R}^n$, we assume the existence of a matricization map, $\mathrm{mat}(\cdot)$ defined as

$$\mathrm{mat}(\mathbf{x}) = [\mathbf{x}_{1:N}^\top, \mathbf{x}_{(N+1):2N}^\top, \ldots, x_{(N(M-1)+1):NM}^\top] \in \mathbb{R}^{N\times M},$$

where $N, M \in \mathbb{N}$, $N \cdot M = n$ and $N \leq M$. We also define the inverse operation, i.e. vectorization, as $\mathrm{vec}(\mathbf{X}) = \mathrm{mat}^{-1}(\mathbf{X})$. Under the assumptions that the singular values of $\mathrm{mat}(\mathbf{x})$ are equi-order (as they would be if the matrix was i.i.d Gaussian), then we have

$$\|\mathrm{mat}(\mathbf{x})\|_{\mathrm{op}}^2 = \max_i \sigma_i^2 = \Theta\left(\frac{\sum_{i\in[\sqrt{n}]} \sigma_i^2}{\sqrt{n}}\right) = \Theta\left(\frac{\|\mathrm{mat}(\mathbf{x})\|_F^2}{\sqrt{n}}\right).$$

Thus, as (1.1) utilized iterates with squared two-norm of order n, $\|\mathrm{mat}(\mathbf{x})\|_F^2 = \|\mathbf{x}\|_2^2 = \Theta(n)$, and thus, $\|\mathrm{mat}(\mathbf{x})\|_{\mathrm{op}} = \Theta(n^{1/4})$. This motivates applying the normalization of $n^{-1/4}\mathrm{mat}(\mathbf{x})$ to each input of a spectral function leading the input matrix to have a maximum singular value of order $\Theta(1)$. This allows one to write the spectral decomposition $n^{-1/4}\mathrm{mat}(\mathbf{x}) = \sum_{i\in[\sqrt{n}]} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ where, by convention, $\sigma_i \geq 0$ and $||\mathbf{u}_i||_2 = ||\mathbf{v}_i||_2 = 1$ represent the singular vectors. To incorporate multiple iterates and side information vectors, the input $\mathbf{x}$ is written as $\mathbf{x} = \sum_{i=1}^{\ell} \mathbf{a}_i \mathbf{x}_i$, $\mathbf{a} \in \mathbb{R}^\ell$ for a set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell \in \mathbb{R}^n$ and $\mathbf{a} \in \mathbb{R}^\ell$.

Definition 2.22. Consider the set of functions $f : \mathbb{R}^n \to \mathbb{R}^n$, such that there exists a constant $C > 0$ where the following conditions hold:

(1) **Spectral Form:** There exists a function $f_{\mathrm{spec}} : \mathbb{R}_+ \to \mathbb{R}$ applied seperably to the spectrum of a matrix, an $\ell \in \mathbb{N}$ and vector $\mathbf{a} \in [\ell]$ such that, where f is represented as,

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_m) = n^{1/4}\mathrm{vec}\left(f_{\mathrm{spec}}\left(\mathrm{mat}\left(n^{-1/4}\sum_{i=1}^{\ell} \mathbf{a}_i\mathbf{x}_i\right)\right)\right).$$

(2) **Polynomial Growth:** The above function f_{spec} satisfies $|f_{\mathrm{spec}}(x)| \leq C(1+|x|)^C$ and $\left|\frac{d}{dx} f_{\mathrm{spec}}(x)\right| \leq C(1+|x|)^C$.

The finite span over the above set of functions is named the **spectral functions**.

We end this section with the following theorem. This follows from each function class being BCP-approximable, a fact shown in [27, Section 6].

Theorem 2.23. *Consider a sequence of Lipschitz activations $\mathcal{F} = \{f_{t+1}\}_{t\geq 1}$, and two Lipschitz test functions ϕ_1, ϕ_2. If $\mathcal{F} \cup \{\phi_1, \phi_2\}$ are subsets of the local, anisotropic or spectral function class, then the result from Theorem 2.19 holds.*

3. Universality Of Tensor Networks

As the results on BCP implying universality are the most novel in this work, we present this analysis here.

3.1. Universality In Expectation.

In this section, we first show the following lemma.

Lemma 3.1. *Let the ordered connected multigraph G and tensor labelings $\mathcal{L}, \mathcal{L}'$ be as in Theorem 2.16. Then there is a constant $C > 0$ for which*

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] - \mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L}')\right] \leq Cn^{-1/2}.$$

Proof. Throughout the proof, we fix the ordered multigraph $G = (\mathcal{V}, \mathcal{E})$ and a decomposition of its vertex set $\mathcal{V} = \mathcal{V}_W \sqcup \mathcal{V}_T$, where no two vertices of $\mathcal{V}_W$ are connected by an edge. It suffices to prove the result for tensor labelings $\mathcal{L}$ that assign label $\mathbf{W}$ to $\mathcal{V}_W$ and labels in $\mathcal{T}$ to $\mathcal{V}_T$, for each fixed decomposition $\mathcal{V} = \mathcal{V}_W \sqcup \mathcal{V}_T$.

For any such labeling $\mathcal{L}$, taking expectation over $\mathbf{W}$ in the definition of the value (2.4),

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = \frac{1}{n^{1+|\mathcal{V}_W|/2}} \sum_{\mathbf{i}\in[n]^{\mathcal{E}}} \mathbb{E}\left[\prod_{v\in\mathcal{V}_W} n^{1/2}\mathbf{W}[i_e : e \sim v]\right] \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_e : e \sim v].$$

Let $\mathcal{P}(\mathcal{E})$ be the set of all partitions of the edge set $\mathcal{E}$. Let $\pi_{\mathbf{i}} \in \mathcal{P}(\mathcal{E})$ denote the partition that is induced by the index tuple $\mathbf{i} \in [n]^{\mathcal{E}}$: edges $e, e' \in \mathcal{E}$ belong to the same block of $\pi_{\mathbf{i}}$ if and only if $i_e = i_{e'}$. We write $[e]$ for the block of π that contains edge e. Then the above summation may be decomposed as

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = \sum_{\pi\in\mathcal{P}(\mathcal{E})} \frac{1}{n^{1+|\mathcal{V}_W|/2}} \sum_{\mathbf{i}\in[n]^{\pi}}^{*} \mathbb{E}\left[\prod_{v\in\mathcal{V}_W} n^{1/2}\mathbf{W}[i_{[e]} : e \sim v]\right] \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{[e]} : e \sim v]. \tag{3.1}$$

Here, the first summation is over all possible edge partitions $\pi = \pi(\mathbf{i})$, and the second summation $\sum^*_{\mathbf{i}\in[n]^\pi}$ is over a distinct index $i_{[e]} \in [n]$ for each block $[e] \in \pi$, where $*$ denotes that indices $i_{[e]}, i_{[e']}$ must be distinct for different blocks $[e] \neq [e'] \in \pi$.

Let $\mathcal{P}(\mathcal{V}_W)$ be the set of all partitions of the vertex subset $\mathcal{V}_W$. Given a partition $\pi \in \mathcal{P}(\mathcal{E})$, we associate to it a partition $\pi_W(\pi) \in \mathcal{P}(\mathcal{V}_W)$ where $u, v \in \mathcal{V}_W$ belong to the same block of $\pi_W(\pi)$ if their incident edges belong to the same two blocks of π. More precisely:

Definition 3.2. For any $v, u \in \mathcal{V}_W$, let e, e' be the two edges incident to v, and f, f' the two edges incident to u. The partition $\pi_W(\pi) \in \mathcal{P}(\mathcal{V}_W)$ **associated to** π is such that v, u belong to the same block of $\pi_W(\pi)$ if and only if

$$\{[e], [e']\} = \{[f], [f']\}$$

(as equality of unordered sets, where possibly $[e] = [e']$ and $[f] = [f']$).

Writing $[v] \in \pi_W(\pi)$ for the block of $\pi_W(\pi)$ containing v, we say that these blocks $[e], [e'] \in \pi$ are **incident to** the block $[v] \in \pi_W(\pi)$ and denote this by $[e] \sim [v]$.

This definition is such that for any $\mathbf{i} \in [n]^\pi$ of the summation $\sum^*_{\mathbf{i}\in[n]^\pi}$, the entries $\mathbf{W}[i_{[e]} : e \sim v]$ and $\mathbf{W}[i_{[e]} : e \sim u]$ of $\mathbf{W}$ are equal if v, u belong to the same block of $\pi_W(\pi)$, and are independent otherwise. Thus each block $[v] \in \pi_W(\pi)$ corresponds to a different independent entry of $\mathbf{W}$. For each $k \geq 1$, define $\mathbf{M}_k \in \mathbb{R}^{n\times n}$ as the matrix with entries

$$\mathbf{M}_k[i, j] = \mathbb{E}[n^{k/2}\mathbf{W}[i, j]^k], \tag{3.2}$$

where Definition 1.2 guarantees that $\mathbf{M}_k$ is symmetric and $|\mathbf{M}_k[i, j]| < C_k$ for a constant $C_k > 0$. Then evaluating the expectation over $\mathbf{W}$ in (3.1) gives

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = \sum_{\pi\in\mathcal{P}(\mathcal{E})} \frac{1}{n^{1+|\mathcal{V}_W|/2}} \sum_{\mathbf{i}\in[n]^{\pi}}^{*} \prod_{[v]\in\pi_W(\pi)} \mathbf{M}_{k[v]}[i_{[e]} : [e] \sim [v]] \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{[e]} : e \sim v].$$

Here, the first product is over all blocks $[v] \in \pi_W(\pi)$, $k[v]$ denotes the number of vertices of $\mathcal{V}_W$ in the block $[v]$, and $[i_{[e]} : [e] \sim [v]]$ is the index pair $[i_{[e]}, i_{[e']}]$ for the blocks $[e], [e']$ incident to $[v]$.

Definition 3.3. $\pi \in \mathcal{P}(\mathcal{E})$ is **single** if some block $[v] \in \pi_W(\pi)$ has $k[v] = 1$. A block $[v] \in \pi_W(\pi)$ is **paired** if $k[v] = 2$ and if its incident blocks $[e], [e'] \in \pi$ are such that $[e] \neq [e']$.

(Thus if π is not single and $[v] \in \pi_W(\pi)$ is not paired, then either $k[v] \geq 3$ or $k[v] = 2$ and $[e] = [e']$.)

By the vanishing of first moments of $\mathbf{W}[i, j]$ in Definition 1.2, if π is single then there is some $[v] \in \pi_W(\pi)$ for which $k[v] = 1$ and hence $\mathbf{M}_{k[v]} = 0$. By the assumption for second moments of off-diagonal entries $\mathbf{W}[i, j]$, if $[v] \in \pi_W(\pi)$ is paired then $M_{k[v]}[i_{[e]} : [e] \sim [v]] = 1$. Applying these observations above,

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = \sum_{\substack{\pi\in\mathcal{P}(\mathcal{E})\\ \text{not single}}} \frac{1}{n^{1+|\mathcal{V}_W|/2}} \sideset{}{^*}\sum_{\mathbf{i}\in[n]^\pi} \prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{[e]}:[e]\sim[v]] \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{[e]}:e\sim v]. \tag{3.3}$$

Next, we apply an inclusion-exclusion argument followed by Cauchy-Schwarz to bound the difference of (3.3) between $\mathcal{L}$ and $\mathcal{L}'$. Endow $\mathcal{P}(\mathcal{E})$ with ordering by refinement: $\tau \geq \pi$ if each block of τ is a union of one or more blocks of π. We will use $\langle e\rangle \in \tau$ to denote the block of τ containing edge e, to avoid notational confusion with the block $[e]\in\pi$. Note that if $v,u\in\mathcal{V}_W$ belong to the same block of $\pi_W(\pi)$, then the two edges incident to v and those incident to u belong to the same blocks $[e],[e']\in\pi$, and hence also the same blocks $\langle e\rangle,\langle e'\rangle\in\tau$ since $\tau\geq\pi$. Analogous to Definition 3.2, we continue to say that $\langle e\rangle,\langle e'\rangle\in\tau$ are the blocks **incident to** $[v]\in\pi_W(\pi)$ and denote this by $\langle e\rangle\sim[v]$.

Let $\mu(\pi,\tau)$ be the inclusion-exclusion (i.e. Möbius inversion) coefficients such that, for any fixed $\pi\in\mathcal{P}(\mathcal{E})$ whose blocks we denote momentarily by $[e_1],\ldots,[e_m]$ (where $e_1,\ldots,e_m$ are any choices of a representative edge in each block), and for any function $f:[n]^\pi\to\mathbb{R}$,

$$\sideset{}{^*}\sum_{\mathbf{i}\in[n]^\pi} f(i_{[e_1]},\ldots,i_{[e_m]}) = \sum_{\tau\in\mathcal{P}(\mathcal{E}):\tau\geq\pi} \mu(\pi,\tau) \sum_{\mathbf{i}\in[n]^\tau} f(i_{\langle e_1\rangle},\ldots,i_{\langle e_m\rangle}).$$

The sum $\sum_{\mathbf{i}\in[n]^\tau}$ on the right side is over one index $i_{\langle e\rangle}\in[n]$ for each block $\langle e\rangle\in\tau$, and no longer restricts indices for different blocks $\langle e\rangle\in\tau$ to be distinct. Applying this inclusion-exclusion relation to (3.3),

$$\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = \sum_{\substack{\pi\in\mathcal{P}(\mathcal{E})\\ \text{not single}}} \sum_{\tau\in\mathcal{P}(\mathcal{E}):\tau\geq\pi} \frac{\mu(\pi,\tau)}{n^{1+|\mathcal{V}_W|/2}} \underbrace{\sum_{\mathbf{i}\in[n]^\tau} \prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{\langle e\rangle}:\langle e\rangle\sim[v]] \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{\langle e\rangle}:e\sim v]}_{:=\mathrm{val}_{\check{G}}(\check{\mathcal{L}})}. \tag{3.4}$$

We clarify that here, $\pi_W(\pi)$ in the first product of $\mathrm{val}_{\check{G}}(\check{\mathcal{L}})$ continues to be defined by the partition π (not by τ), and $[i_{\langle e\rangle}:\langle e\rangle\sim[v]]$ is the index tuple $[i_{\langle e\rangle},i_{\langle e'\rangle}]$ for the blocks $\langle e\rangle,\langle e'\rangle\in\tau$ that are incident to $[v]\in\pi_W(\pi)$. For later reference in the proof, it is helpful to interpret $\mathrm{val}_{\check{G}}(\check{\mathcal{L}})$ in (3.4) as the value of a (π,τ)-dependent tensor network $(\check{G},\check{\mathcal{L}})$ constructed as follows:

- $\check{G}=(\check{\mathcal{V}},\check{\mathcal{E}})$ has three disjoint sets of vertices $\check{\mathcal{V}}=\check{\mathcal{V}}_W\sqcup\check{\mathcal{V}}_{\mathrm{Id}}\sqcup\check{\mathcal{V}}_T$, and each edge $e\in\check{\mathcal{E}}$ connects a vertex of $\check{\mathcal{V}}_{\mathrm{Id}}$ with a vertex of either $\check{\mathcal{V}}_W$ or $\check{\mathcal{V}}_T$.
- The vertices of $\check{\mathcal{V}}_{\mathrm{Id}}$ are the blocks of τ. Each vertex $\langle e\rangle\in\check{\mathcal{V}}_{\mathrm{Id}}\equiv\tau$ is labeled by Id, and the ordering of its edges is arbitrary (as the tensor Id is symmetric).
- The vertices of $\check{\mathcal{V}}_W$ are the blocks of $\pi_W(\pi)$. Each vertex $[v]\in\check{\mathcal{V}}_W\equiv\pi_W(\pi)$ is labeled by $\mathbf{M}_{k[v]}$, and has two edges (ordered arbitrarily) connecting to the blocks $\langle e\rangle,\langle e'\rangle\in\check{\mathcal{V}}_{\mathrm{Id}}\equiv\tau$ that are incident to $[v]$.
- $\check{\mathcal{V}}_T$ is the same as the vertex set $\mathcal{V}_T$ of G, with the same tensor labels. For each vertex $v\in\mathcal{V}_T$ with ordered edges $e_1,\ldots,e_m$ in G, the vertex $v\in\check{\mathcal{V}}_T\equiv\mathcal{V}_T$ has ordered edges connecting to $\langle e_1\rangle,\ldots,\langle e_m\rangle\in\check{\mathcal{V}}_{\mathrm{Id}}\equiv\tau$.

An example of this construction of $(\check{G},\check{\mathcal{L}})$ from $(G,\mathcal{L},\pi,\tau)$ is depicted in Figure 2. It is direct to check that the quantity $\mathrm{val}_{\check{G}}(\check{\mathcal{L}})$ defined in (3.4) indeed equals the value of this tensor network as defined in (2.4), where the label Id on each vertex $\langle e\rangle\in\check{\mathcal{V}}_{\mathrm{Id}}$ ensures that only summands which have the same index value $i_{\langle e\rangle}\in[n]$ for all edges incident to $\langle e\rangle$ contribute to (2.4).

Then, defining $\mathbf{M}'_k$ and $\mathrm{val}_{\check{G}}(\check{\mathcal{L}}')$ as in (3.2) and (3.4) with $\mathbf{W}'$ in place of $\mathbf{W}$, we have

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] - \mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L}')\right]\right| \le \sum_{\substack{\pi\in\mathcal{P}(\mathcal{E})\\ \text{not single}}} \sum_{\tau\in\mathcal{P}(\mathcal{E}):\tau\ge\pi} \frac{|\mu(\pi,\tau)|}{n^{1+|\mathcal{V}_W|/2}} \times \\ &\left|\underbrace{\sum_{\mathbf{i}\in[n]^\tau}\left(\prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{\langle e\rangle}:\langle e\rangle\sim[v]] - \prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}'_{k[v]}[i_{\langle e\rangle}:\langle e\rangle\sim[v]]\right)\prod_{v\in\mathcal{V}_T}\mathbf{T}_v[i_{\langle e\rangle}: e\sim v]}_{=\mathrm{val}_{\check{G}}(\check{\mathcal{L}})-\mathrm{val}_{\check{G}}(\check{\mathcal{L}}')}\right|.\end{aligned} \tag{3.5}$$

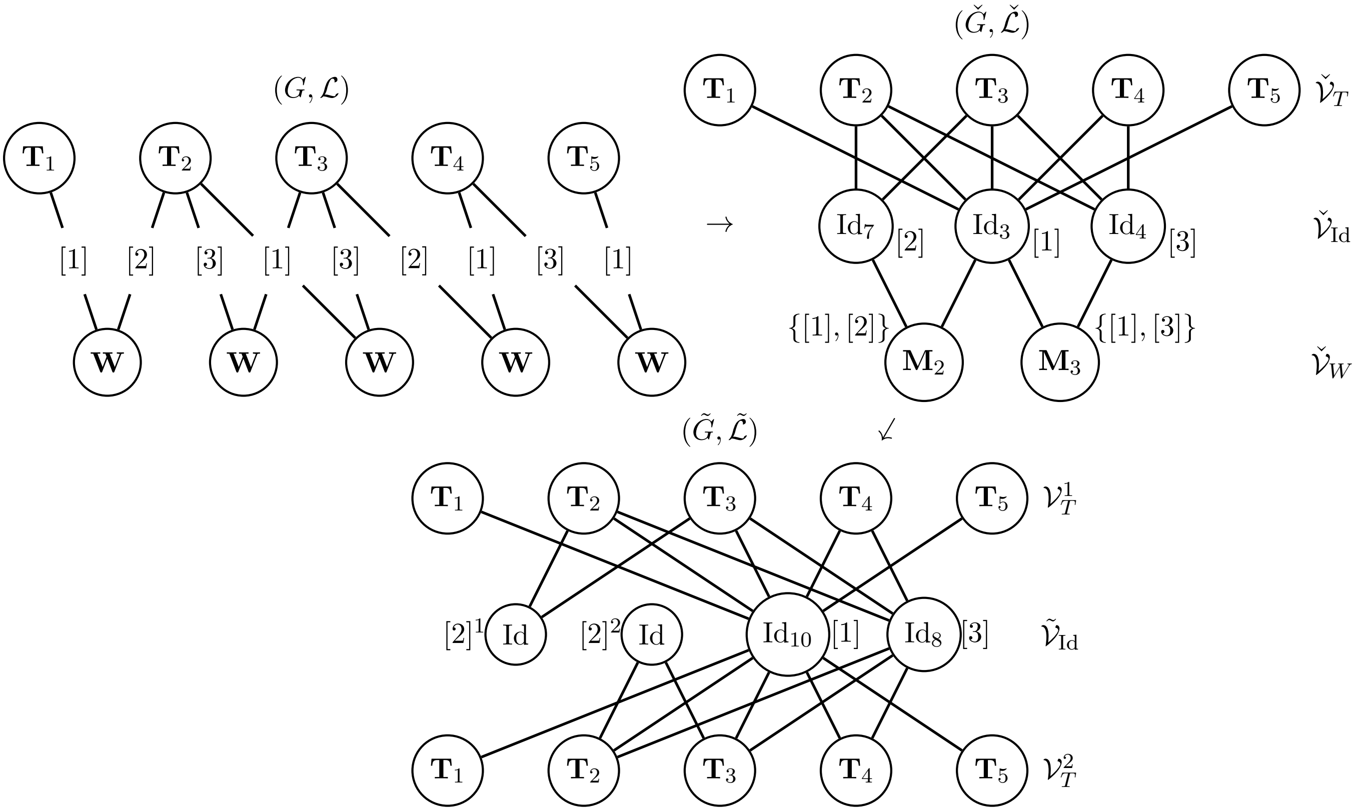

$(G, \mathcal{L})$
$\mathbf{T}_1$ $\mathbf{T}_2$ $\mathbf{T}_3$ $\mathbf{T}_4$ $\mathbf{T}_5$
[1] [2] [3] [1] [3] [2] [1] [3] [1]
$\mathbf{W}$ $\mathbf{W}$ $\mathbf{W}$ $\mathbf{W}$ $\mathbf{W}$
$\rightarrow$
$(\check{G}, \check{\mathcal{L}})$
$\mathbf{T}_1$ $\mathbf{T}_2$ $\mathbf{T}_3$ $\mathbf{T}_4$ $\mathbf{T}_5$ $\check{\mathcal{V}}_T$
Id_7 [2] Id_3 [1] Id_4 [3] $\check{\mathcal{V}}_{\mathrm{Id}}$
$\{[1],[2]\}$ $\mathbf{M}_2$ $\mathbf{M}_3$ $\{[1],[3]\}$ $\check{\mathcal{V}}_W$
$\swarrow$
$(\tilde{G}, \tilde{\mathcal{L}})$
$\mathbf{T}_1$ $\mathbf{T}_2$ $\mathbf{T}_3$ $\mathbf{T}_4$ $\mathbf{T}_5$ $\mathcal{V}_T^1$
$[2]^1$ Id $[2]^2$ Id Id_{10} [1] Id_8 [3] $\tilde{\mathcal{V}}_{\mathrm{Id}}$
$\mathbf{T}_1$ $\mathbf{T}_2$ $\mathbf{T}_3$ $\mathbf{T}_4$ $\mathbf{T}_5$ $\mathcal{V}_T^2$

FIGURE 2. An example conversion from $(G, \mathcal{L}) \to (\check{G}, \check{\mathcal{L}}) \to (\tilde{G}, \tilde{\mathcal{L}})$, nodes are labeled by the tensor assigned to them. (Top Left) $(G, \mathcal{L})$ is the initial graph we start with, depending on the choice of partition, we assign each edge in the graph to some block. (Top Right) $(\check{G}, \check{\mathcal{L}})$ is the graph G restructured to represent that each block of edges must have the identical index; specific blocks in π and π_W are shown beside the node which represents them. (Bottom) $(\tilde{G}, \tilde{\mathcal{L}})$ is the graph $(\check{G}, \check{\mathcal{L}})$ after removing the nodes in $\check{\mathcal{V}}_W$ and then copying the graph along $\check{\mathcal{V}}_{\mathrm{Id}}$. Again beside each node is the block they represent, with multiplicity in the superscript if the block is good, as [2] is.

DEFINITION 3.4. Given partitions $\pi, \tau \in \mathcal{P}(\mathcal{E})$ with $\tau \ge \pi$, a block $\langle e\rangle \in \tau$ is **bad** if there exists at least one block $[v] \in \pi_W(\pi)$ that is not paired and that is incident to $\langle e\rangle$, and **good** otherwise. We write $\tau = \tau^b \sqcup \tau^g$ where τ^b and τ^g are the sets of bad and good blocks, respectively.

Note that if $|\tau^b| = 0$, i.e. all blocks of τ are good, then every block $[v] \in \pi_W(\pi)$ must be paired, so the products $\prod_{[v]\in\pi_W(\pi):\text{not paired}}$ defining $\mathrm{val}_{\check{G}}(\check{\mathcal{L}}), \mathrm{val}_{\check{G}}(\check{\mathcal{L}}')$ are both trivial and equal to 1, and $\mathrm{val}_{\check{G}}(\check{\mathcal{L}}) - \mathrm{val}_{\check{G}}(\check{\mathcal{L}}') = 0$. When $|\tau^b| \neq 0$, these products involve only indices corresponding to $\langle e\rangle \in \tau^b$ and not $\langle e\rangle \in \tau^g$. Thus

$$\mathrm{val}_{\check{G}}(\check{\mathcal{L}}) - \mathrm{val}_{\check{G}}(\check{\mathcal{L}}') = \sum_{\mathbf{i}\in[n]^{\tau^b}} \Bigg[\Bigg(\prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{\langle e\rangle} : \langle e\rangle \sim [v]] - \prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}'_{k[v]}[i_{\langle e\rangle} : \langle e\rangle \sim [v]]\Bigg) \times \sum_{\mathbf{i}\in[n]^{\tau^g}} \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{\langle e\rangle} : e \sim v]\Bigg] \mathbb{1}\{|\tau^b| \neq 0\}.$$

Applying Cauchy-Schwarz over the outer summation $\sum_{\mathbf{i}\in[n]^{\tau^b}}$,

$$|\mathrm{val}_{\check{G}}(\check{\mathcal{L}}) - \mathrm{val}_{\check{G}}(\check{\mathcal{L}}')| \leq \Bigg[\sum_{\mathbf{i}\in[n]^{\tau^b}} \Bigg(\prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{\langle e\rangle} : \langle e\rangle \sim [v]] - \prod_{\substack{[v]\in\pi_W(\pi)\\ \text{not paired}}} \mathbf{M}'_{k[v]}[i_{\langle e\rangle} : \langle e\rangle \sim [v]]\Bigg)^2\Bigg]^{1/2} \times \Bigg[\sum_{\mathbf{i}\in[n]^{\tau^b}} \Bigg(\sum_{\mathbf{i}\in[n]^{\tau^g}} \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{\langle e\rangle} : e \sim v]\Bigg)^2\Bigg]^{1/2} \mathbb{1}\{|\tau^b| \neq 0\}.$$

Then applying that $|\mathbf{M}_k[i,j]| \leq C_k$ for a constant $C_k > 0$ and all $i, j \in [n]$, there exists a constant $C(\pi,\tau) > 0$ for which the first factor is at most $C(\pi,\tau)n^{|\tau^b|/2}$, so

$$|\mathrm{val}_{\check{G}}(\check{\mathcal{L}}) - \mathrm{val}_{\check{G}}(\check{\mathcal{L}}')| \leq \mathbb{1}\{|\tau^b| \neq 0\} C_{\pi,\tau} n^{|\tau^b|/2} \underbrace{\Bigg[\sum_{\mathbf{i}\in[n]^{\tau^b}} \Bigg(\sum_{\mathbf{i}\in[n]^{\tau^g}} \prod_{v\in\mathcal{V}_T} \mathbf{T}_v[i_{\langle e\rangle} : e \sim v]\Bigg)^2\Bigg]}_{:=\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})}^{1/2}. \tag{3.6}$$

We interpret the quantity $\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})$ in (3.6) as the value of a (π,τ)-dependent bipartite tensor network $\tilde{G} = (\tilde{\mathcal{V}}_{\mathrm{Id}} \sqcup \tilde{\mathcal{V}}_T, \tilde{\mathcal{E}})$ with $(\mathrm{Id}, \mathcal{T})$-labeling $\tilde{\mathcal{L}}$, constructed as follows:

- $\tilde{\mathcal{V}}_{\mathrm{Id}}$ has one vertex for each block $\langle e\rangle \in \tau^b$, which we denote also by $\langle e\rangle \in \tilde{\mathcal{V}}_{\mathrm{Id}}$, and two vertices for each block $\langle e\rangle \in \tau^g$, which we denote by $\langle e\rangle^1, \langle e\rangle^2 \in \tilde{\mathcal{V}}_{\mathrm{Id}}$. These are labeled by Id, and the ordering of their edges is arbitrary.
- $\tilde{\mathcal{V}}_T = \mathcal{V}_T^1 \sqcup \mathcal{V}_T^2$ consists of two copies of the original vertex set $\mathcal{V}_T$ of G, with the same tensor labels. For each $v \in \mathcal{V}_T$, we denote its copies by $v^1 \in \mathcal{V}_T^1$ and $v^2 \in \mathcal{V}_T^2$. Suppose $v \in \mathcal{V}_T$ has ordered edges $e_1, \ldots, e_m$ in the original graph G. If $\langle e_i\rangle \in \tau^b$, then the i^{th} edge of both $v^1 \in \mathcal{V}_T^1$ and $v^2 \in \mathcal{V}_T^2$ connect to $\langle e_i\rangle \in \tilde{\mathcal{V}}_{\mathrm{Id}}$. If $\langle e_i\rangle \in \tau^g$ then the i^{th} edge of $v^1 \in \mathcal{V}_T^1$ connects to $\langle e_i\rangle^1 \in \tilde{\mathcal{V}}_{\mathrm{Id}}$, and the i^{th} edge of $v^2 \in \mathcal{V}_T^2$ connects to $\langle e_i\rangle^2 \in \tilde{\mathcal{V}}_{\mathrm{Id}}$.

An example of this construction is also illustrated in Figure 2. Note that since each edge $e \in \mathcal{E}$ of the original graph $G = (\mathcal{V}, \mathcal{E})$ is incident to at least one vertex $v \in \mathcal{V}_T$ (because no two vertices of $\mathcal{V}_W$ are adjacent), each block $\langle e\rangle \in \tau^b \sqcup \tau^g$ has also at least one vertex $v \in \mathcal{V}_T$ that is incident to an edge of that block. Then it is direct to check that the quantity $\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})$ of (3.6) is indeed the value of this tensor network as defined in (2.4).

Finally, we bound $\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})$ using the given BCP property of $\mathcal{T}$ and a combinatorial argument. Fixing any $\pi \in \mathcal{P}(\mathcal{E})$ that is not single, we categorize the possible types of blocks $[v] \in \pi_W(\pi)$ based on $k[v]$ (the number of vertices belonging to $[v]$) and on its incident blocks $[e], [e'] \in \pi$:

- Let N_3 be the number of blocks $[v]$ with $k[v] \geq 3$
- Let N_2 be the number of paired blocks $[v]$, i.e. with $k[v] = 2$ and $[e] \neq [e']$
- Let N_1 be the number of blocks $[v]$ with $k[v] = 2$ and $[e] = [e']$.

Let $\mathbf{c}(\tilde{G})$ be the number of connected components of $\tilde{G}$. We claim the following combinatorial properties:

(1) The number of vertices of $\mathcal{V}_W$ satisfies $|\mathcal{V}_W| \geq 3N_3 + 2N_2 + 2N_1$.
(2) The number of blocks of τ^b satisfies $|\tau^b| \leq 2N_3 + N_1$.
(3) The degree of each vertex of $\tilde{\mathcal{V}}_{\mathrm{Id}}$ in $\tilde{G}$ is even.
(4) If $|\tau^b| \neq 0$, then the number of connected components of $\tilde{G}$ satisfies $\mathbf{c}(\tilde{G}) \leq 1+2N_2+N_3$.

Let us verify each of these claims: (1) holds because each block $[v] \in \pi_W(\pi)$ counted by N_1 or N_2 contains exactly $k[v] = 2$ vertices of $\mathcal{V}_W$, and each block counted by N_3 contains $k[v] \geq 3$ vertices.

(2) holds because any block of τ^b must be incident to some block $[v] \in \pi_W(\pi)$ that is not paired. Each non-paired block $[v] \in \pi_W(\pi)$ that is counted by N_3 is incident to two distinct blocks $[e], [e'] \in \pi$ — hence at most two blocks in τ^b because $\tau \geq \pi$ — and each non-paired block counted by N_1 is incident to one distinct block $[e] \in \pi$ — hence also one block in τ^b.

For (3), consider first a bad block $\langle e \rangle \in \tau^b$. By construction, the edges of its corresponding vertex $\langle e \rangle \in \tilde{\mathcal{V}}_{\mathrm{Id}}$ come in pairs, connecting to pairs of vertices (v^1, v^2). Thus $\langle e \rangle$ has even degree. Now consider a good block $\langle e \rangle \in \tau^g$ and its corresponding vertices $\langle e \rangle^1, \langle e \rangle^2 \in \tilde{\mathcal{V}}_{\mathrm{Id}}$. Let $e_1, \ldots, e_m$ be the edges of G that belong to this block $\langle e \rangle \in \tau^g$. If such an edge e_i connects two vertices of $\mathcal{V}_T$, then there are two corresponding edges in $\tilde{G}$ that connect these vertices of $\mathcal{V}_T^1$ with $\langle e \rangle^1$. Otherwise e_i connects a vertex in $u \in \mathcal{V}_T$ with a vertex $v \in \mathcal{V}_W$. Since $\langle e \rangle \in \tau^g$ is good, the block $[v] \in \pi_W(\pi)$ containing this vertex $v \in \mathcal{V}_W$ must be paired — thus, there is exactly one other vertex $v' \in \pi_W(\pi)$ that belongs to $[v]$. If v is incident to exactly one edge in this block $\langle e \rangle$, then so is v', and if v is incident to two edges both in $\langle e \rangle$ (which may occur if its incident blocks $[e] \neq [e'] \in \pi$ are merged into a single block $\langle e \rangle \in \tau$) then so is v'. This shows that the edges among $e_1, \ldots, e_m$ that connect $\mathcal{V}_T$ to $\mathcal{V}_W$ come in pairs, and each pair contributes two edges of $\tilde{G}$ between $\mathcal{V}_T^1$ and $\langle e \rangle^1$. So $\langle e \rangle^1$ has even degree. Similarly $\langle e \rangle^2$ has even degree, which shows (3).

For (4), note that $(\tilde{G}, \tilde{\mathcal{L}})$ may be obtained from $(\check{G}, \check{\mathcal{L}})$ by removing all vertices of $\check{\mathcal{V}}_W$ and their incident edges from $\check{G}$, duplicating the remaining graph on the vertex set $\check{\mathcal{V}}_{\mathrm{Id}} \cup \check{\mathcal{V}}_T$ into two disjoint copies on $\check{\mathcal{V}}_{\mathrm{Id}}^1 \cup \check{\mathcal{V}}_T^1$ and $\check{\mathcal{V}}_{\mathrm{Id}}^2 \cup \check{\mathcal{V}}_T^2$, and merging the vertices of $\check{\mathcal{V}}_{\mathrm{Id}}^1$ representing bad blocks $\langle e \rangle \in \tau^b$ with their copies in $\check{\mathcal{V}}_{\mathrm{Id}}^2$ while keeping the remaining vertices of $\check{\mathcal{V}}_{\mathrm{Id}}^1, \check{\mathcal{V}}_{\mathrm{Id}}^2$ (representing good blocks $\langle e \rangle \in \tau^g$) distinct. We may then bound $\mathbf{c}(\tilde{G})$ via the following observations:

- $\check{G}$ is a connected graph, because the original graph G is connected by assumption.
- For any connected subgraph K of $\check{G}$, call it *good* if all vertices of $K \cap \check{\mathcal{V}}_{\mathrm{Id}}$ represent good blocks $\langle e \rangle \in \tau^g$, and *bad* if at least one vertex of $K \cap \check{\mathcal{V}}_{\mathrm{Id}}$ represents a bad block $\langle e \rangle \in \tau^b$. We track the number N_g of good connected components and N_b of bad connected components as we sequentially remove vertices of $\check{\mathcal{V}}_W$ from $\check{G}$ one at a time:

  Supposing that $|\tau^b| \neq 0$ as assumed in claim (4), the starting connected graph $\check{G}$ is bad, so $N_g = 0$ and $N_b = 1$. Each vertex $[v] \in \check{\mathcal{V}}_W$ counted by N_1 can be connected to only one vertex of $\check{\mathcal{V}}_{\mathrm{Id}}$, so its removal does not change (N_g, N_b). Each vertex $[v] \in \check{\mathcal{V}}_W$ counted by N_3 is connected to at most 2 vertices of $\check{\mathcal{V}}_{\mathrm{Id}}$, both of which are bad by definition, so its removal does not change N_g and increases N_b by at most 1. Each vertex $[v] \in \check{\mathcal{V}}_W$ counted by N_2 is connected to at most 2 vertices of $\check{\mathcal{V}}_{\mathrm{Id}}$ which may be either good or bad, so its removal increases the total number of connected components $N_b + N_g$ by at most 1. Thus, after removing all vertices of $\check{\mathcal{V}}_W$ from $\check{G}$, we have

  $$N_b + N_g \leq 1 + N_2 + N_3, \qquad N_g \leq N_2.$$

- By the above process of obtaining $\tilde{G}$ from $\check{G}$, after removing all vertices of $\check{\mathcal{V}}_W$, each component counted by N_b results in one connected component of $\tilde{G}$, while each component counted by N_g results in two connected components of $\tilde{G}$. Thus

  $$\mathbf{c}(\tilde{G}) = N_b + 2N_g,$$

  and applying the above bounds gives $\mathbf{c}(\tilde{G}) \leq 1 + 2N_2 + N_3$ which is claim (4).

We apply these combinatorial claims and the BCP property to conclude the proof: Suppose $\pi, \tau \in \mathcal{P}(\mathcal{E})$ are such π is not single, $\tau \geq \pi$, and $|\tau^b| \neq 0$. Recalling that $\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})$ factorizes as the product of the values across connected components, and applying claims (3–4) and BCP to each connected component of $\tilde{G}$, we have

$$\mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}}) \leq C(\tilde{G}) n^{\mathbf{c}(\tilde{G})} \leq C(\tilde{G}) n^{1+2N_2+N_3} \tag{3.7}$$

for a constant $C(\tilde{G}) > 0$. Since $\tilde{G}$ is determined by π and τ, applying (3.7) and claim (2) to (3.6) gives, for some different constant $C(\pi, \tau) > 0$,

$$|\mathrm{val}_{\check{G}}(\check{\mathcal{L}}) - \mathrm{val}_{\check{G}}(\check{\mathcal{L}}')| \leq C(\pi, \tau) \cdot n^{\frac{2N_3+N_1}{2}} \cdot n^{\frac{1+2N_2+N_3}{2}}.$$

Applying this and claim (1) back to (3.5), and noting that the number of such partitions $\pi, \tau \in \mathcal{P}(\mathcal{E})$ is a constant independent of n, we obtain as desired

$$\left|\mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] - \mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L}')\right]\right| \leq C \cdot \frac{1}{n^{1+\frac{3N_3+2N_2+2N_1}{2}}} \cdot n^{\frac{2N_3+N_1}{2}} \cdot n^{\frac{1+2N_2+N_3}{2}} \leq C n^{-1/2}.$$

■

3.2. Almost-Sure Convergence. To complete the proof of Theorem 2.16, we show the following fourth moment bound for concentration of the tensor network value around its mean.

LEMMA 3.5. *Let the ordered multigraph G and tensor labeling $\mathcal{L}$ be as in Theorem 2.16. Then there is a constant $C > 0$ for which*

$$\mathbb{E}\left[\left(\frac{1}{n}\mathrm{val}_G(\mathcal{L}) - \frac{1}{n}\mathbb{E}\mathrm{val}_G(\mathcal{L})\right)^4\right] \leq \frac{C}{n^2}.$$

PROOF. We again fix the ordered multigraph $G = (\mathcal{V}, \mathcal{E})$ and a decomposition $\mathcal{V} = \mathcal{V}_W \sqcup \mathcal{V}_T$ of its vertices, and consider a labeling $\mathcal{L}$ that assigns $\mathbf{W}$ to $\mathcal{V}_W$ and elements of $\mathcal{T}$ to $\mathcal{V}_T$.

Let $G^{\sqcup 4} = (\mathcal{V}^{\sqcup 4}, \mathcal{E}^{\sqcup 4})$ be the ordered multigraph consisting of four disjoint copies of G, where $\mathcal{V}^{\sqcup 4} = \mathcal{V}^1 \sqcup \mathcal{V}^2 \sqcup \mathcal{V}^3 \sqcup \mathcal{V}^4$ are the four copies of $\mathcal{V}$ decomposed as $\mathcal{V}_j = \mathcal{V}_W^j \sqcup \mathcal{V}_T^j$ for $j = 1, 2, 3, 4$, and $\mathcal{E}^{\sqcup 4} = \mathcal{E}^1 \sqcup \mathcal{E}^2 \sqcup \mathcal{E}^3 \sqcup \mathcal{E}^4$ are the four copies of $\mathcal{E}$. Let $\mathbf{W}^1, \ldots, \mathbf{W}^4$ be four independent copies of the Wigner matrix $\mathbf{W}$. For any word $a = a_1a_2a_3a_4$ with letters $a_1, a_2, a_3, a_4 \in \{1, 2, 3, 4\}$, define $\mathcal{L}_a$ as the tensor labeling of $G^{\sqcup 4}$ such that for each $j = 1, 2, 3, 4$, vertices of $\mathcal{V}_W^j$ are labeled by the matrix $\mathbf{W}^{a_j}$, and vertices of $\mathcal{V}_T^j$ have the same labels as $\mathcal{V}_T$ under $\mathcal{L}$. Then

$$\begin{aligned}
&\mathbb{E}[(\mathrm{val}_G(\mathcal{L}) - \mathbb{E}\mathrm{val}_G(\mathcal{L}))^4] \\
&= \mathbb{E}[\mathrm{val}_G(\mathcal{L})^4] - 4\mathbb{E}[\mathrm{val}_G(\mathcal{L})^3]\mathbb{E}[\mathrm{val}_G(\mathcal{L})] + 6\mathbb{E}[\mathrm{val}_G(\mathcal{L})^2]\mathbb{E}[\mathrm{val}_G(\mathcal{L})]^2 - 3\mathbb{E}[\mathrm{val}_G(\mathcal{L})]^4 \\
&= \mathbb{E}[\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_{1111}) - 4\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_{1112}) + 6\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_{1123}) - 3\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_{1234})]
\end{aligned}$$

where the expectation on the last line is over the independent Wigner matrices $\mathbf{W}^1, \ldots, \mathbf{W}^4$.

Let $\mathcal{P}(\mathcal{E}^{\sqcup 4})$ be the set of all partitions of the combined edge set $\mathcal{E}^{\sqcup 4}$. For any $a = a_1a_2a_3a_4$, we have analogously to (3.1)

$$\mathbb{E}\left[\frac{1}{n^4}\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_a)\right] = \sum_{\pi \in \mathcal{P}(\mathcal{E}^{\sqcup 4})} \underbrace{\frac{1}{n^{4+2|\mathcal{V}_W|}} \sum_{\mathbf{i} \in [n]^\pi}^{*} \mathbb{E}\left[\prod_{j=1}^{4} \prod_{v \in \mathcal{V}_W^j} n^{1/2}\mathbf{W}^{a_j}[i_{[e]} : e \sim v]\right] \prod_{j=1}^{4} \prod_{v \in \mathcal{V}_T^j} \mathbf{T}_v[i_{[e]} : e \sim v]}_{:= V_a(\pi)}. \tag{3.8}$$

Let us split $\mathcal{P}(\mathcal{E}^{\sqcup 4})$ into three disjoint sets:

- $\mathcal{A}$: Partitions π such that every block $[e] \in \pi$ satisfies $[e] \subseteq \mathcal{E}^j$ for a single copy $j = 1, 2, 3, 4$.
- $\mathcal{B}$: Partitions π for which there is a decomposition $\{1, 2, 3, 4\} = \{j_1, j_2\} \sqcup \{k_1, k_2\}$ such that every block $[e] \in \pi$ satisfies either $[e] \subseteq \mathcal{E}^{j_1}$, $[e] \subseteq \mathcal{E}^{j_2}$, or $[e] \subseteq \mathcal{E}^{k_1} \cup \mathcal{E}^{k_2}$, and at least one block $[e] \in \pi$ has a nonempty intersection with both $\mathcal{E}^{k_1}$ and $\mathcal{E}^{k_2}$.

– $\mathcal{C}$: All remaining partitions of $\mathcal{P}(\mathcal{E}^{\sqcup 4})$.

We write correspondingly

$$V_a(\mathcal{A}) = \sum_{\pi \in \mathcal{A}} V_a(\pi), \qquad V_a(\mathcal{B}) = \sum_{\pi \in \mathcal{B}} V_a(\pi), \qquad V_a(\mathcal{C}) = \sum_{\pi \in \mathcal{C}} V_a(\pi)$$

so that $\mathbb{E}[n^{-4}\mathrm{val}_{G^{\sqcup 4}}(\mathcal{L}_a)] = V_a(\mathcal{A}) + V_a(\mathcal{B}) + V_a(\mathcal{C})$. Then

$$\mathbb{E}\left[\left(\frac{1}{n}\mathrm{val}_G(\mathcal{L}) - \frac{1}{n}\mathbb{E}\mathrm{val}_G(\mathcal{L})\right)^4\right] = \sum_{\mathcal{S} \in \{\mathcal{A},\mathcal{B},\mathcal{C}\}} V_{1111}(\mathcal{S}) - 4V_{1112}(\mathcal{S}) + 6V_{1123}(\mathcal{S}) - 3V_{1234}(\mathcal{S}). \tag{3.9}$$

We now analyze separately the terms of (3.9) for $\mathcal{S} = \mathcal{A}, \mathcal{B}, \mathcal{C}$: For $\mathcal{A}$, observe that for any $\pi \in \mathcal{A}$, since the edge sets $\mathcal{E}^1, \mathcal{E}^2, \mathcal{E}^3, \mathcal{E}^4$ are unions of disjoint blocks of π, the indices of each of the matrices $\mathbf{W}^1, \mathbf{W}^2, \mathbf{W}^3, \mathbf{W}^4$ are distinct in (3.8). Then $V_a(\pi)$ has the same value for all words $a = a_1a_2a_3a_4$, so $V_{1111}(\pi) = V_{1112}(\pi) = V_{1123}(\pi) = V_{1234}(\pi)$, and hence

$$V_{1111}(\mathcal{A}) - 4V_{1112}(\mathcal{A}) + 6V_{1123}(\mathcal{A}) - 3V_{1234}(\mathcal{A}) = 0. \tag{3.10}$$

For $\mathcal{B}$, recall that each $\pi \in \mathcal{B}$ corresponds to a (unique) associated decomposition $\{1,2,3,4\} = \{j_1, j_2\} \sqcup \{k_1, k_2\}$ where each block $[e] \in \pi$ belongs to $\mathcal{E}^{j_1}$, $\mathcal{E}^{j_2}$, or $\mathcal{E}^{k_1 \cup k_2}$. We further decompose

$$V_{a_1a_2a_3a_4}(\mathcal{B}) = V_{\underline{a_1a_2}a_3a_4} + V_{\underline{a_1}a_2\underline{a_3}a_4} + V_{\underline{a_1}a_2a_3\underline{a_4}} + V_{a_1\underline{a_2a_3}a_4} + V_{a_1\underline{a_2}a_3\underline{a_4}} + V_{a_1a_2\underline{a_3a_4}}$$

where each term is a summation over those $\pi \in \mathcal{B}$ corresponding to a single such decomposition $\{1,2,3,4\} = \{j_1, j_2\} \sqcup \{k_1, k_2\}$, and the underlined positions indicate the indices $\{k_1, k_2\}$ while the non-underlined positions indicate the indices $\{j_1, j_2\}$. So for instance, $V_{\underline{a_1}a_2\underline{a_3}a_4}$ is the summation of $V_{a_1a_2a_3a_4}(\pi)$ over those $\pi \in \mathcal{B}$ for which each block $[e] \in \pi$ belongs to either $\mathcal{E}^1 \cup \mathcal{E}^3$, $\mathcal{E}^2$, or $\mathcal{E}^4$. Note that for any such π, the indices of $\mathbf{W}^2$ and $\mathbf{W}^4$ in (3.8) are distinct from those of $\{\mathbf{W}^1, \mathbf{W}^3\}$, and hence for any $a_1, a_3 \in \{1,2,3,4\}$, the value $V_{\underline{a_1}a_2\underline{a_3}a_4}$ is the same for all choices of a_2, a_4. This type of observation, together with symmetry of $V_{\underline{a_1}a_2\underline{a_3}a_4}$ under permutations of the four indices and relabelings of the copies $\{1,2,3,4\}$, yields the identities

$$\begin{aligned}
V_{1111}(\mathcal{B}) &= 6V_{\underline{11}11} = 6V_{\underline{11}23}\\
V_{1112}(\mathcal{B}) &= 3V_{\underline{11}12} + 3V_{\underline{1}11\underline{2}} = 3V_{\underline{11}23} + 3V_{\underline{12}34}\\
V_{1123}(\mathcal{B}) &= V_{\underline{11}23} + 2V_{\underline{1}1\underline{2}3} + 2V_{\underline{1}12\underline{3}} + V_{11\underline{23}} = V_{\underline{11}23} + 5V_{\underline{12}34}\\
V_{1234}(\mathcal{B}) &= 6V_{\underline{12}34}.
\end{aligned}$$

Applying these identities shows

$$V_{1111}(\mathcal{B}) - 4V_{1112}(\mathcal{B}) + 6V_{1123}(\mathcal{B}) - 3V_{1234}(\mathcal{B}) = 0. \tag{3.11}$$

Finally, for $\mathcal{C}$, we claim that there is a constant $C > 0$ such that for any $a = a_1a_2a_3a_4$, we have

$$|V_a(\mathcal{C})| \le Cn^{-2}.$$

The proof is similar to the analysis in Lemma 3.1: Fix any $a = a_1a_2a_3a_4$. Associated to any edge partition $\pi \in \mathcal{C}$, consider the vertex partition $\pi_W(\pi) \in \mathcal{P}(\mathcal{V}_W^1 \sqcup \mathcal{V}_W^2 \sqcup \mathcal{V}_W^3 \sqcup \mathcal{V}_W^4)$ such that v, u belong to the same block of $\pi_W(\pi)$ if and only if their incident edges belong to the same two incident blocks of π and, in addition, $v \in \mathcal{V}_W^j$ and $u \in \mathcal{V}_W^k$ for two indices $j, k \in \{1,2,3,4\}$ such that $a_j = a_k$ (i.e. v, u correspond to the same Wigner matrix $\mathbf{W}^{a_j} = \mathbf{W}^{a_k}$). Let $k[v]$ be the number of vertices in the block $[v] \in \pi_W(\pi)$, call π single if some block $[v] \in \pi_W(\pi)$ has $k[v] = 1$, and call $[v] \in \pi_W(\pi)$ paired if $k[v] = 2$ and its incident blocks $[e], [e'] \in \pi$ satisfy $[e] \neq [e']$. Then evaluating the expectation over $\mathbf{W}^1, \ldots, \mathbf{W}^4$ in (3.8), we get analogously to (3.3) and (3.4)

$$V_a(\mathcal{C}) = \sum_{\substack{\pi \in \mathcal{C} \\ \text{not single}}} \frac{1}{n^{4+2|\mathcal{V}_W|}} \sum_{\mathbf{i} \in [n]^\pi}^{*} \prod_{\substack{[v] \in \pi_W(\pi) \\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{[e]} : [e] \sim [v]] \prod_{j=1}^{4} \prod_{v \in \mathcal{V}_T^j} \mathbf{T}_v[i_{[e]} : e \sim v]$$

$$= \sum_{\substack{\pi \in \mathcal{C} \\ \text{not single}}} \sum_{\tau \in \mathcal{P}(\mathcal{E}) : \tau \geq \pi} \frac{\mu(\pi, \tau)}{n^{4+2|\mathcal{V}_W|}} \underbrace{\sum_{\mathbf{i} \in [n]^\tau} \prod_{\substack{[v] \in \pi_W(\pi) \\ \text{not paired}}} \mathbf{M}_{k[v]}[i_{\langle e \rangle} : \langle e \rangle \sim [v]] \prod_{j=1}^{4} \prod_{v \in \mathcal{V}_T^j} \mathbf{T}_v[i_{\langle e \rangle} : e \sim v]}_{\mathrm{val}_{\check{G}}(\check{\mathcal{L}})} . \tag{3.12}$$

Let τ^b, τ^g denote the sets of bad and good blocks of τ defined in the same way as Definition 3.4. Then applying Cauchy-Schwarz over $\sum_{\mathbf{i} \in [n]^{\tau^b}}$, we obtain analogously to (3.6)

$$|\mathrm{val}_{\check{G}}(\check{\mathcal{L}})| \leq C(\pi, \tau) n^{|\tau^b|/2} \Bigg[\underbrace{\sum_{\mathbf{i} \in [n]^{\tau^b}} \Bigg(\sum_{\mathbf{i} \in [n]^{\tau^g}} \prod_{j=1}^{4} \prod_{v \in \mathcal{V}_T^j} \mathbf{T}_v[i_{\langle e \rangle} : e \sim v] \Bigg)^2}_{:= \mathrm{val}_{\tilde{G}}(\tilde{\mathcal{L}})} \Bigg]^{1/2} . \tag{3.13}$$

Now let N_3, N_2, and N_1 be the numbers of blocks $[v] \in \pi_W(\pi)$ with $k[v] \geq 3$, with $k[v] = 2$ and incident blocks $[e] \neq [e'] \in \pi$, and with $k[v] = 2$ and incident blocks $[e] = [e'] \in \pi$, respectively. Then the same arguments as in Lemma 3.1 show that

(1) $4|\mathcal{V}_W| \geq 3N_3 + 2N_2 + 2N_1$.
(2) $|\tau^b| \leq 2N_3 + N_1$.
(3) The degree of each vertex of $\check{\mathcal{V}}_{\mathrm{Id}}$ in $\tilde{G}$ is even.

Furthermore we may count the number of connected components $\mathbf{c}(\tilde{G})$ of $\tilde{G}$ by the following extension of the argument in Lemma 3.1: Analogous to Lemma 3.1, $\check{G}$ above is an ordered multigraph with three disjoint sets of vertices $\check{\mathcal{V}}_W \equiv \pi_W(\pi)$, $\check{\mathcal{V}}_{\mathrm{Id}} \equiv \tau$, and $\check{\mathcal{V}}_T \equiv \mathcal{V}_T^1 \sqcup \mathcal{V}_T^2 \sqcup \mathcal{V}_T^3 \sqcup \mathcal{V}_T^4$, and $\tilde{G}$ is again obtained from $\check{G}$ by removing all vertices of $\check{\mathcal{V}}_W$, duplicating the resulting graph on $\check{\mathcal{V}}_{\mathrm{Id}} \cup \check{\mathcal{V}}_T$, and merging the two copies of vertices in $\check{\mathcal{V}}_{\mathrm{Id}}$ that correspond to bad blocks $\langle e \rangle \in \tau^b$. Observe that:

– By definition, $G^{\sqcup 4}$ consists of 4 connected components. For any $\pi \in \mathcal{C}$, there are at least two different pairs of indices $1 \leq j < k \leq 4$ for which a block of π has non-empty intersection with both $\mathcal{E}^j$ and $\mathcal{E}^k$. (Otherwise, we would have $\pi \in \mathcal{A}$ or $\pi \in \mathcal{B}$.) Then $\check{G}$ has at most 2 connected components.
– Call a connected subgraph K of $\check{G}$ good if all vertices $K \cap \check{\mathcal{V}}_{\mathrm{Id}}$ represent good blocks $\langle e \rangle \in \tau^g$, and bad otherwise. We again track the numbers N_g and N_b of good and bad connected components of $\check{G}$ as we sequentially remove vertices of $\check{\mathcal{V}}_W$. The 1 or 2 connected components of the starting graph $\check{G}$ can be either good or bad. Removing a vertex $[v] \in \check{\mathcal{V}}_W$ counted by N_1 does not change (N_g, N_b), removing a vertex $[v] \in \check{\mathcal{V}}_W$ counted by N_3 does not change N_g and increases N_b by at most 1, and removing a vertex counted by N_2 increases $N_b + N_g$ by at most 1. Hence, after removing all vertices of $\check{\mathcal{V}}_W$ from $\check{G}$, we have

$$N_b + N_g \leq 2 + N_2 + N_3, \qquad N_g \leq 2 + N_2.$$

– After removing all vertices of $\check{\mathcal{V}}_W$, we have $\mathbf{c}(\tilde{G}) = N_b + 2N_g$.

Thus we have also

(4) $\mathbf{c}(\tilde{G}) \leq 4 + 2N_2 + N_3$.

Applying these properties (1–4) and the BCP condition to (3.12) and (3.13),

$$|V_a(\mathcal{C})| \leq C \cdot \frac{1}{n^{4 + \frac{3N_3 + 2N_2 + 2N_1}{2}}} \cdot n^{\frac{2N_3 + N_1}{2}} \cdot n^{\frac{4 + 2N_2 + N_3}{2}} \leq C n^{-2}$$

as claimed. Thus, for a different constant $C > 0$,

$$|V_{1111}(\mathcal{C}) - 4V_{1112}(\mathcal{C}) + 6V_{1123}(\mathcal{C}) - 3V_{1234}(\mathcal{C})| \leq C n^{-2}. \tag{3.14}$$

Applying (3.10), (3.11) and (3.14) to (3.9) proves the lemma. ■

PROOF OF THEOREM 2.16. Applying Lemma 3.5 and Markov's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\mathrm{val}_G(\mathcal{L}) - \frac{1}{n}\mathbb{E}\mathrm{val}_G(\mathcal{L})\right| > \varepsilon\right) \leq \frac{C}{\varepsilon^4 n^2}.$$

This bound is summable in n, so by the Borel-Cantelli Lemma, almost surely

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{val}_G(\mathcal{L}) - \mathbb{E}\left[\frac{1}{n}\mathrm{val}_G(\mathcal{L})\right] = 0.$$

The same statement holds for $\mathcal{L}'$, and combining this with Lemma 3.1 gives Theorem 2.16. ■

CHAPTER 3

Analyzing Markov Chain Dynamics Using Noise Injected Querying, Closing the Local-Computational Gap In Sparse Tensor PCA

Disclamer: This project is ongoing joint work with **Conor Sheehan**, **Kostas Tsirkas**, and **Ilias Zadik**. New results on this project are still in development, but the main arguments on absence of a local-computational gap are reasonably finished. Similar to Chapter 2, we will reference the (in progress) paper [50].

Notation. The notation $A_n = \tilde{\Omega}(B_n)$ is used for statements that hold when $A_n \geq C \log^s(n) B_n$ for some sufficiently large $C > 0$ and $s \in \mathbb{N}$. The standard asymptotic notations O, o, Ω, ω are used with respect to n. The notation $C_n = \text{Poly}(n)$ refers to any function satisfying $n^s \leq C_n \leq n^{s+1}$ for some $s \in \mathbb{N}$. The symbols d_H and d_1 denote the Hamming distance and the ℓ_1 distance, respectively. Finally, we often refer to the angle between two vectors as $\cos(\sigma, \theta) = \frac{\langle \sigma, \theta \rangle}{\|\sigma\|_2 \|\theta\|_2}$ for two vectors $\sigma, \theta \in \mathbb{R}^n$.

1. Computational-Local Gaps And Markov Chains

In contrast to Chapter 1, this chapter establishes a class of iterative algorithms on Gaussian additive models (GAMs) which (a) have a **tractable analysis** in the sense that we can explicitly write the law of many macroscopic properties of our iterates and (b) is expressive enough to achieve the **algorithmic thresholds** predicted by low-degree methods or other type of computational frameworks. As an application, we prove positive results for a restricted class of GAMs known as *sparse tensor PCA*, which was briefly introduced in Chapter 1.

These results provide a refutation of the indictment of Markov chain Monte Carlo (MCMC) with respect to local-to-computational gaps. In fact, a similar story has played out in literature on the planted clique problem.

1.1. Planted Clique And The Resurrection Of MCMC. Given a signal $\theta \in \{0,1\}^n$ with k ones, consider a random matrix $A \in \mathbb{R}^{n \times n}$ with entry-wise distribution,

$$A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \text{ has } \theta_i = \theta_j = 1 \\ \text{Bernoulli}(1/2) & \text{otherwise} \end{cases}.$$

The goal of recovering θ given A is known as the *planted clique problem.* This model has, perhaps, the most famous computational-to-statistical gap; simple arguments are able to prove that when $k = 0$, the largest clique (i.e. any vector $x \in \{0,1\}^n$ where $\frac{x^\top A x}{\|x\|_0^2} = 1$) grows at a rate of $2\log_2(n)$. It is then intuitive that maximum likelihood estimation will recover cliques of size $(2+\varepsilon)\log_2(n)$, albeit solved by brute force search, yet—embarrassingly—no known polynomial time algorithm has been demonstrated recovery of cliques $o(\sqrt{n})$. Evidence for this representing this threshold being "hard" has been proved from both the sum of squares perspective [7] and from the overlap gap perspective [31]. This threshold was also confirmed from the Markov chain perspective [42] where the metropolis process fails to find cliques of size $\sqrt{n}$ or smaller.

It was then of relative surprise that [17] proved that any Markov chain process with stationary distribution $e^{-\beta|C|}$, where C is any clique and $\beta > 0$, could not find cliques up to $k = o(n)$ in size. And thus, the entire class of Markov chain algorithms came into question. *Are they* *strictly worse than other classes of polynomial-time algorithms? Is there any way to correct these failures?*

Yes, there is. Consider a greedy algorithm on the Hamiltonian $H(v) = -|E(v)| + \gamma(\binom{|v|}{2} - |E(\sigma)|)$ where v is a subset of the vertex set $[n]$ and $E(v)$ is the number of edges connecting two vertices in v. In lieu of the hard constraints (i.e. restricting to the domain of H to only cliques or to v with $|v| = k$), [35] introduced the regularization term seen above with parameter $\gamma > 0$ and relaxed the domain of H to all $v \subset [n]$. For a specific $\gamma > 0$, it was proven that such an algorithm initialized at $\mathbf{1}$ (and therefore a randomized low-temperature Markov chain version) will output the planted clique θ when $k \geq C\sqrt{n}$ for some constant C. This result was able to resurrect Markov chain methods through two major ideas: (1) considering **soft constraints** to encourage sparsity, and (2) **warm initializations** can be vital for MCMC to reach the performance of other polynomial time algorithms. Indeed, these tools will be vital to s similar story surrounding the sparse tensor PCA model.

1.2. Sparse Tensor PCA (Again). We return to the sparse tensor PCA model introduced in Chapter 1, giving a brief refresher now. Given vector $\theta \in \mathbb{R}^n$, say with $\|\theta\|_0 = k$, we receive the observation,

$$Y = \frac{\lambda}{k^{r/2}}\theta^{\otimes r} + G,$$

where $G \in \mathbb{R}^{n^{\otimes r}}$ is a tensor with i.i.d standard normal entries. The goal is to recover θ from Y with high probability. As mentioned previously, this algorithm demonstrates a local-to-computational gap, suggesting that efficient algorithms for this problem must be "global" in nature [3, 4, 18].

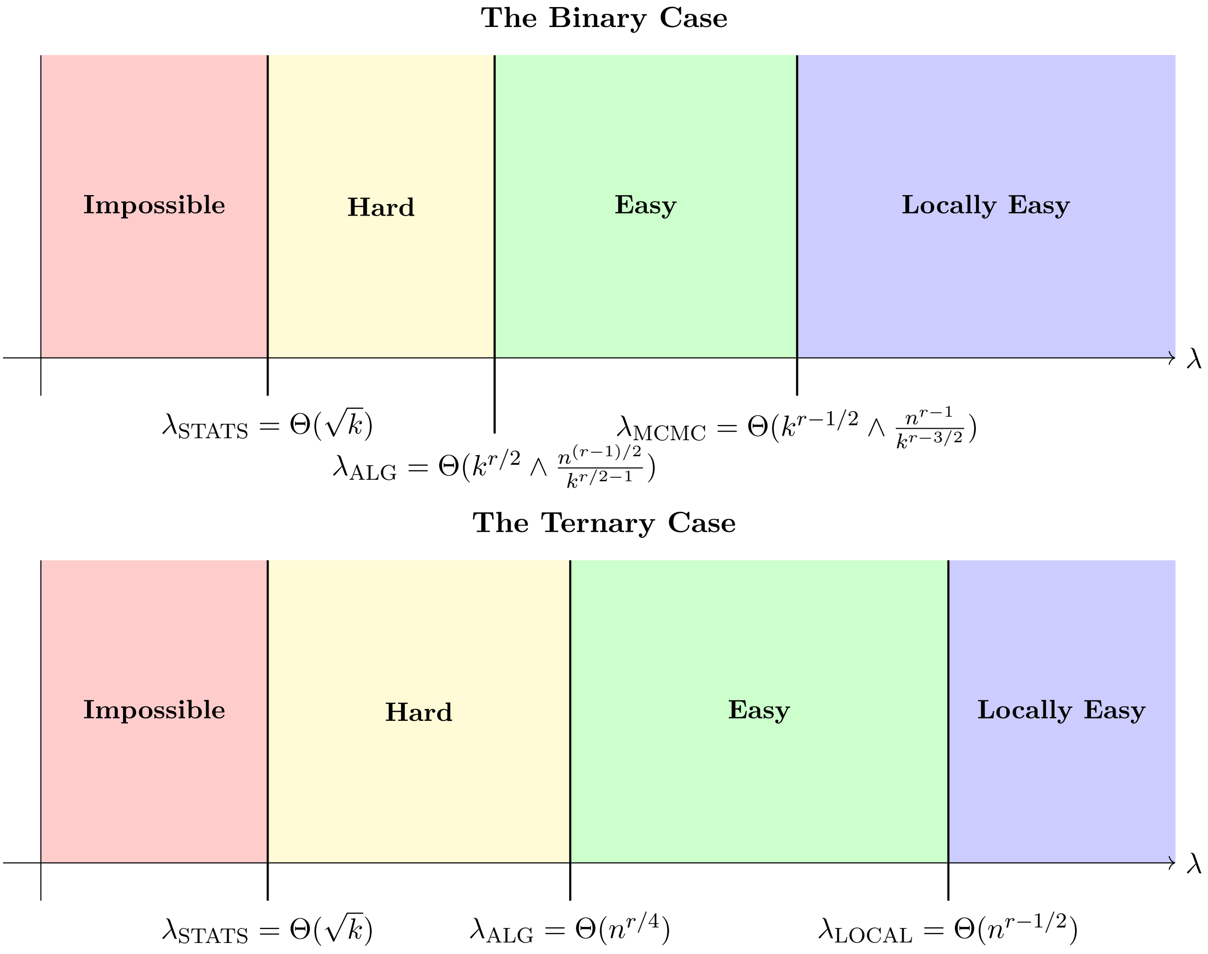

The Binary Case
Impossible
Hard
Easy
Locally Easy
λ
$\lambda_{\text{STATS}} = \Theta(\sqrt{k})$
$\lambda_{\text{MCMC}} = \Theta(k^{r-1/2} \wedge \frac{n^{r-1}}{k^{r-3/2}})$
$\lambda_{\text{ALG}} = \Theta(k^{r/2} \wedge \frac{n^{(r-1)/2}}{k^{r/2-1}})$
The Ternary Case
Impossible
Hard
Easy
Locally Easy
λ
$\lambda_{\text{STATS}} = \Theta(\sqrt{k})$
$\lambda_{\text{ALG}} = \Theta(n^{r/4})$
$\lambda_{\text{LOCAL}} = \Theta(n^{r-1/2})$

FIGURE 1. Differing computational thresholds for sparse tensor PCA.

We review these gaps in the *binary* case of $\theta \in \{0,1\}^n$ and the *ternary* case of $\theta \in \{-1,0,1\}^n$ with $\|\theta\|_0 = k \in [n]$. Thresholds for the *binary* case [17] (i.e. when $\theta \in \{0,1\}^n$) and the ternary case [3] (i.e. when $\theta \in \{-1,0,1\}^n$) are provided in Figure 1.

The λ_{MCMC} threshold from [17] was proven for a set of Markov chains with a hard sparsity constraint, wherein they restrict the state space to $\|\sigma\|_0 = k$. Meaning that ideas from [35] could be used to improve the performance of Markov chains. **We prove this is the case**, demonstrating that state space relaxations, soft constraints and a warm initialization allows MCMC algorithms to achieve the threshold $\sqrt{\log(n)}\lambda_{\text{ALG}}$ for both the binary and ternary case of θ.

The analysis of these algorithms is also of independent interest. We consider a set of noise injected querying methods which have a simple Gaussian law. This leads to the ability to track macroscopic properties of Markov chain algorithms, drastically simplifying the analysis. An interesting future direction would be to prove this class of algorithms can provide negative results on Gaussian additive models and perhaps can be linked to low-degree polynomial estimation in a manner similar to AMP methods from [60].

2. Main Contributions

This section will briefly describe the sparse tensor PCA problem and our main contributions to the binary signal case. We present the framework that drives this analysis in Section 3 and provide a proof of Theorem 2.4 in Section 4. We leave the results for the ternary case to [50].

Definition 2.1. Given $n \in \mathbb{N}, \alpha \in (0,1), r \in \mathbb{N}, k = \Theta(n^\alpha) \in \mathbb{N}$ and λ (scaling with respect to n and k). Let $\theta \in \{0,1\}^n$ be such that $\|\theta\|_0 = k$ (without loss of generality, we assume that $\theta = [k]$). We generate the tensor $Y \in \mathbb{R}^{n^{\otimes r}}$ as

$$Y = \frac{\lambda}{k^{r/2}}\theta^{\otimes r} + G, \tag{2.1}$$

where G is the standard $\mathcal{N}(0,1)$ Gaussian tensor with i.i.d. entries.

Definition 2.2. Define algorithm $\mathcal{MC} : \mathbb{R}^{n^{\otimes r}} \times \{0,1\}^n \to \{0,1\}^n$ given $Y \in \mathbb{R}^{n^{\otimes r}}$ and initialization $\sigma^1 \in \{0,1\}^n$ as follows:

Set $\xi = 25\log(n)$ and $\gamma = 2(\xi^* \log(\xi^* n))^{1/2}$, let $P \subset [n]^{\mathbb{N}}$ be a sequence of i.i.d. uniform samples of $[n]$, and initialize auxiliary variables $t = 0$ and $t_i = 0$ for all $i \in [n]$. For each $i \in [n]$, generate ξ independent mean zero Gaussian random variables G_i^j, for $j \in [\xi]$, further set $\bar{G}_i = \frac{1}{\xi}\sum_{j=1}^{\xi} G_i^j$. Algorithm, $\mathcal{MC}$ then runs the following iterative loop,

(1) Increment t by one. When $P_t = i$, also increment t_i by one.
(2) Calculate,

$$D_t = (1 - 2\sigma_i^t)\left(\langle e_i \otimes (\sigma^t)^{\otimes(r-1)}, Y\rangle - \gamma\|\sigma^t\|_0^{r-1} + (G_i^{t_i} - \bar{G}_i)\|\sigma^t\|_0^{(r-1)/2}\right). \tag{2.2}$$

If $D_t > 0$ then accept the transition of setting $\sigma_i^{t+1} = 1 - \sigma_i^t$ and keep all other coordinates the same value.
(3) Return σ^{t+1} the first time that $t > \xi n/2$.

Remark 2.3. It can be shown that $\mathcal{MC}$ is a Markov chain algorithm in the variables $\sigma, \{t_i : i \in [n]\}$. Meaning the following theorem dispels the idea that MCMC is strictly worse than other algorithmic classes for the model from Definition 2.1.

Theorem 2.4. *Given σ^1, consider $\mathcal{MC}(Y, \sigma^1)$ where $Y \in \mathbb{R}^{n^{\otimes r}}$ is generated as in Definition 2.1. The following holds with probability* $1 - o(1)$.

(1) When $\sigma^1 = \mathbf{1}$, then $\mathcal{MC}(Y, \mathbf{1})$ outputs θ at the end of run-time for some $\lambda = \tilde{\Omega}(n^{(r-1)/2}/k^{r/2-1})$.
(2) When $\sigma^1 = e_i$, for some i where $\theta_i = 1$, then $\mathcal{MC}(Y, e_i)$ outputs θ at the end of run-time for some $\lambda = \tilde{\Omega}(k^{r/2})$.

As mentioned at the beginning of this section, Theorem 2.4 is proven in Section 4.

3. Noise Injected Querying

Definition 3.1. Let n be a generating parameter for the following algorithm. Consider a sequence of observation $\{Y_i\}_{i\in[N]}$ where $N = \text{Poly}(n)$ and each $Y_i = \mu_i^* + Z_i$ with $\mu^* \in \mathbb{R}^N$ and[1] $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i \in [N]$. Given a sequence of sets $\Delta \subset [N]^{\mathbb{N}}$, $\xi \in \mathbb{N}$, and a sequence of functions $\mathcal{F} = \{f_t\}_{t\in\mathbb{N}}$ where $f_t : \mathbb{R}^{|\Delta_t|} \to \mathbb{R}$. We sample $\xi \cdot N$ independent mean zero Gaussian random variables G_i^j for $i \in [N]$ and $j \in [\xi]$, further setting $\bar{G}_i = \frac{1}{\xi}\sum_{j=1}^{\xi} G_i^j$ and initializing auxiliary variables $b = (b_i)_{i\in[N]} = 0$ for all $i \in [N]$. Then—for each iterative step t—we consider the querying set Δ_t and for all $i \in \Delta_t$ we increment b_i by one. We then return the value of

$$f_t(Y_{\Delta_t}, b_{\Delta_t}) = f(Y_{\Delta_t} + (G_{\Delta_t}^{b_{\Delta_t}} - \bar{G}_{\Delta_t})).$$

We iterate this process in t until there exists an $i \in [N]$ where $b_i > \xi$. This final run-time is denoted T_ξ. We refer to this algorithm as **Noise Injected Querying**, abbreviated as $\mathcal{NIQ}(Y, \mathcal{F}, \Delta, \xi)$.

This general set of algorithms was inspired by the noise injection methods from [2]. Their results involve a spectral method on an order three Gaussian tensor plus a rank one spike. In each step of their tensor power method, they consider a query $\Delta_t = [N] = [n^3]$ and run their algorithm for $\xi = \log(\log(n))$ iterations. We extend this technique to a more general version of noising that can be run for much longer times by limiting the access to the amount of Gaussian observations at each time step.

3.1. $\mathcal{NIQ}$ Has An Independent Gaussian Law.

Although this technique of noising seems rather simple, it represents a general algorithmic principle: The algorithm $\mathcal{NIQ}$ can represent a proxy for the algorithm $f_t(Y_{\Delta_t})$ for $t \in [T_\xi]$ while having a simple technical analysis. Indeed, the following theorem dictates a simple law for the returned values of $f_t(Y_{\Delta_t} + (G_{\Delta_t}^{b_{\Delta_t}} - \bar{G}_{\Delta_t}))$.

Theorem 3.2 (Noise Injected Querying Has An Independent Gaussian Law). *Consider the algorithm $\mathcal{NIQ}(Y, \mathcal{F}, \Delta, \xi)$ given in Definition 3.1 where each $f_t \in \mathcal{F}$ is Borel measurerable with respect to the set of random varaibles Y_{Δ_t}, $G_{\Delta_t}^{b_{\Delta_t}}$ and $\bar{G}_{\Delta_t}$. Then, the following equivalence in law holds:*

$$(f_t(Y_{\Delta_t} + (G_{\Delta_t}^{b_{\Delta_t}} - \bar{G}_{\Delta_t}))_{t\in[T_\xi]}) \overset{\mathcal{L}}{=} f_t(X_{\Delta_t})_{t\in[T_\xi]},$$

where $X_{\Delta_t} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_{\Delta_t}^*, \xi I_{|\Delta_t|})$ for each $t \in [T_\xi]$.

This theorem prescribes that each returned value of $f_t(Y_{\Delta_t} + (G_{\Delta_t}^{b_{\Delta_t}} - \bar{G}_{\Delta_t}))$ is equivalent to viewing $f_t(\mu_{\Delta_t}^* + \text{noise})$ where the noise is Gaussian, independent across the indexes of the observations in Y and independent across the run-time of the algorithm. This leads to a far easier technical analysis of $\mathcal{NIQ}$ algorithms at the cost of inflating the noise by a ξ factor. Many times we consider ξ to be poly-logarithmic in n, for such a case we only need a modest poly-logarithmic increase in the signal-to-noise ratio λ.

The Proof of Theorem 3.2. This proof follows from calculating the mean and covariance of $A_t = Y_{\Delta_t} + (G_{\Delta_t}^{b_{\Delta_t}} - \bar{G}_{\Delta_t})$ between each coordinate of A_t and $A_{t'}$ for two times $t, t' \in [T_\xi]$. Clearly $\{A_t\}_{t\in[T_\xi]}$ is a set of Gaussian random variables so this characterization of the mean and covariance exactly determines the law of A_t for $t \in [T_\xi]$. The extension of this statement to all borel measurable functions f_t is immediate by considering the push forward measure under f_t.

[1] Notice that we could also consider $Z \sim \mathcal{N}(0, \Sigma)$ for known covariance $\Sigma \in \mathbb{R}^{N\times N}$ by considering $\Sigma^{-1/2}Y$ and $\Sigma^{-1/2}\mu^*$.

First, we check the mean. This follows from the fact that $Y_{\Delta_t} = \mu^*_{\Delta_t} + Z_{\Delta_t}$ where $Z_{\Delta_t} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{|\Delta_t|})$ and the fact that both $G^{b_{\Delta_t}}_{\Delta_t}$ and $\bar{G}_{\Delta_t}$ are centered Russians. Thus, we have the trivial calculation that $\mathbb{E}[X_t] = \mu^*_{\Delta_t}$—the desired mean.

Second, we check the covariance. We calculate this value on the centered random variable $\tilde{A}_t = Z_{\Delta_t} - G^{b_{\Delta_t}}_{\Delta_t} + \bar{G}_{\Delta_t}$ as it has the same covariance structure as A_t. As $\tilde{A}_t$ is centered, we need just calculate the product of each $\tilde{A}_{t,i}$ and $\tilde{A}_{t',i'}$. To do this, we require that for $t, t' \in [T_\xi]$, $i \in \Delta_t$ and $i' \in \Delta_{t'}$. Moreover, we let $j_t = b_{\Delta_t}$ and $j_{t'} = b_{\Delta_{t'}}$. Calculate,

$$\begin{aligned}\mathbb{E}[\tilde{A}_{t,i}\tilde{A}_{t',i'}] &= \mathbb{E}[(Z_{t,i} + G^{j_t}_i - \bar{G}_i)(Z_{t',i'} + G^{j_{t'}}_{i'} - \bar{G}_{i'})] \\ &= \mathbb{E}[Z_{t,i}Z_{t',i'}] + \mathbb{E}[G^{j_t}_i G^{j_{t'}}_{i'}] - \mathbb{E}[G^{j_t}_i \bar{G}_{i'}] - \mathbb{E}[\bar{G}_i G^{j_{t'}}_{i'}] + \mathbb{E}[\bar{G}_i \bar{G}_{i'}]. \qquad (3.1)\end{aligned}$$

We calculate the above summation term by term.

(1) $\mathbb{E}[Z_{t,i}Z_{t',i'}]$: This term is zero if $i \neq i'$ and one if $i = i'$ by the assumption on Z_t in Definition 3.1.
(2) $\mathbb{E}[G^{j_t}_i G^{j_{t'}}_{i'}]$: This is ξ if $t = t'$ and $i = i'$ and zero otherwise as we assume each G^j_i are independent Guassians for each $i \in [N]$ and $j \in [\xi]$.
(3) $\mathbb{E}[G^{j_t}_i \bar{G}_{i'}]$: By the definition of $\bar{G}_i$ from Definition 3.1, $\mathbb{E}[G^{j_t}_i \frac{1}{\xi}\sum_{j=1}^{\xi} G^j_{i'}] = \frac{1}{\xi}\mathbb{E}[G^{j_t}_i G^{j_t}_{i'}]$, where the last equality is because $G^{j_t}_i$ and $G^{j_t}_{i'}$ are independent if $j \neq j_t$. It then follows that $\mathbb{E}[G^{j_t}_i \bar{G}_{i'}]$ is one if and only if $i = i'$.
(4) $\mathbb{E}[\bar{G}_i G^{j_{t'}}_{i'}]$: By symmetry, this term is also one if and only if $i = i'$.
(5) $\mathbb{E}[\bar{G}_i \bar{G}_{i'}]$: We have that $\mathbb{E}[\bar{G}_i \bar{G}_{i'}] = \frac{1}{\xi^2}\sum_{j=1}^{\xi}\sum_{j'=1}^{\xi} \mathbb{E}[G^j_i G^{j'}_{i'}]$. This is equal to $\frac{1}{\xi}\sum_{j=1}^{\xi} \xi = 1$ if $i = i'$ and zero otherwise.

Plugging each term into (3.1) gives,

$$\mathbb{E}[X_{t,i}X_{t',i'}] = 1\{i = i'\} + \xi \cdot 1\{t = t', i = i'\} - 2 \cdot 1\{i = i'\} + 1\{i = i'\} = \xi \cdot 1\{t = t', i = i'\}.$$

This is the desired covariance. ■

We will see in the subsequent sections that each algorithm defined in Section 2 is a special case of $\mathcal{NIQ}$ with polynomial number of Gaussian observations depending on the value of r given in Definition 2.1.

4. Application: Binary Signal Sparse PCA.

We first prove our results on a general value of ξ and then later specify the value ξ^* which guarantees complete recovery of θ for algorithm $\mathcal{MC}$.

Lemma 4.1. *Given a general $\xi \in \mathbb{N}$, the algorithm $\mathcal{MC}$ and its generated sequence $(D_t)_{t\in[T_\xi]}$ from (2.2) has the following equivalence in law,*

$$(D_t)_{t\in[T_\xi]} \overset{\mathcal{L}}{=} \left((1 - 2\sigma^t_i)\left(\theta_i \frac{\lambda}{k^{r/2}}\langle\sigma^t, \theta\rangle^{r-1} - \gamma\|\sigma^t\|_0^{(r-1)/2} + \|\sigma\|_0^{(r-1)/2} Z_t\right)\right)_{t\in[T_\xi]},$$

where $Z_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \xi)$ for each $t \in [T_\xi]$.

Proof. The proof of this theorem follows from demonstrating a choice of function sequence $\mathcal{F}$ and querying sequence Δ, such that $\mathcal{MC}$ is equivalent to $\mathcal{NIQ}(Y, \mathcal{F}, \Delta, \xi)$ algorithm for Y from (2.1) and each $\xi \in \mathbb{N}$.

Algorithm $\mathcal{MC}$ (from Definition 2.2) will propose a single index $i \in [n]$ in which D_t is calculated from. To include this type of sampling in the querying sequence Δ, we set $\mathcal{S}_i = \{(i, j_1, \ldots, j_{r-1}) : j_1, \ldots, j_{r-1} \in [n]\}$ and then define Δ as a sequence of uniformly random draws from the set

$\{\mathcal{S}_i\}_{i\in[n]}$. Under this querying, each function $f_t \in \mathcal{F}$ will have $f_t : \mathbb{R}^{n^{r-1}} \to \mathbb{R}$. The function corresponding to D_t is then,

$$f_t(Y_{\mathcal{S}_i}) = (1-2\sigma_i^t)\left(\sum_{j_1,\dots,j_{r-1}\in[n]} Y_{i,j_1,\dots,j_{r-1}}\sigma_{j_1}^t\cdots\sigma_{j_{r-1}}^t - \gamma\|\sigma^t\|_0^{(r-1)/2}\right).$$

By Theorem 3.2, we have that

$$f_t(Y_{\mathcal{S}_i} + G_{\mathcal{S}_i}^{b_{\mathcal{S}_i}} - \bar{G}_{\mathcal{S}_i}) \overset{\mathcal{L}}{=} f_t(X_{\mathcal{S}_i})$$

where $X_{\mathcal{S}_i} \overset{\text{iid}}{\sim} \mathcal{N}(\frac{\lambda}{k^{r/2}}\theta_i\Theta_{r-1}, \xi\text{Id})$ with $\Theta_{r-1,j_1,\dots,j_{r-1}} = \theta_{j_1}\dots\theta_{j_{r-1}}$. Expanding out $f_t(X_{\mathcal{S}_i})$, let $\bar{X}_{\mathcal{S}_i}$ be the centered random variable of $X_{\mathcal{S}_i}$, we have,

$$\begin{aligned}
f_t(X_{\mathcal{S}_i}) &\overset{\mathcal{L}}{=} (1-2\sigma_i^t)\left(\sum_{j_1,\dots,j_{r-1}\in[n]} X_{i,j_1,\dots,j_{r-1}}\sigma_{j_1}^t\cdots\sigma_{j_{r-1}}^t - \gamma\|\sigma^t\|_0^{(r-1)/2}\right)\\
&\overset{\mathcal{L}}{=} (1-2\sigma_i^t)\left(\sum_{j_1,\dots,j_{r-1}\in[n]} \left(\frac{\lambda}{k^{r/2}}\theta_i\theta_{j_1}\cdots\theta_{j_{r-1}} + \bar{X}_{i,j_1,\dots,j_{r-1}}\right)\sigma_{j_1}^t\cdots\sigma_{j_{r-1}}^t - \gamma\|\sigma^t\|_0^{(r-1)/2}\right)\\
&\overset{\mathcal{L}}{=} (1-2\sigma_i^t)\left(\frac{\lambda}{k^{r/2}}\theta_i\langle\theta,\sigma\rangle^{r-1} + \|\sigma\|_0^{(r-1)/2}Z_t - \gamma\|\sigma^t\|_0^{(r-1)/2}\right),
\end{aligned}$$

where $Z_t \overset{\text{iid}}{\sim} \mathcal{N}(0,\xi)$. Thus, we complete the proof.

■

Using Lemma 4.1, we then have the following guarentee on algorithm $\mathcal{MC}$ from Definition 2.2.

LEMMA 4.2. *Given an arbitrary $\xi \in \mathbb{N}$, recall T_ξ from Definition 3.1. The following then holds with probability $1-o(1)$:*

Let $C^* = (3\xi\log(\xi n))^{1/2}$, if $\sigma^1 \in \{0,1\}^n$ satisfies $\cos(\sigma^1,\theta)^{r-1} > \frac{\sqrt{k}}{\lambda}(C^*+\gamma)$ and $\gamma > C^*$, then $\text{Sign}(D_t) = \text{Sign}(d_H(\sigma^t,\theta) - d_H(\sigma^{t+1},\theta))$ for all $t \in [T_\xi]$. In words, we always accept (or reject) transition that decrease (or increase) the Hamming distance.

PROOF. Consider the equivalence in law from Lemma 4.1, with probability $1-o(1)$, by Mills' upper bound on the right tail over the $\xi \cdot n$ mean zero variance ξ Gaussian variables prescribed by Lemma 4.1, we have that $\max_t |Z_t| \le (3\xi\log(\xi n))^{1/2} = C^*$. Conditioning on this bound, we now prove inductively that if $\cos(\sigma^t,\theta) \ge \cos(\sigma^1,\theta)$ then $D_t > 0$ if and only if the proposed transition decreases the Hamming distance and thus cos increases monotonically. The base case of $t=1$ holds as the cosine holds by the assumption of the Lemma and there is no proposed transition leading to σ^1, making the statement vacuous.

Therefore, assume that σ_t satisfies the inductive hypothesis, as a consequence we have that $\cos(\sigma^t,\theta) \ge \cos(\sigma^1,\theta)$, we now consider each possible transition for our algorithm,

(a) $\theta_i = 1$ and $\sigma_i^t = 0$:

$$D_t \ge \frac{\lambda}{k^{r/2}}\langle\sigma^t,\theta\rangle^{r-1} - C^*\|\sigma\|_0^{(r-1)/2} - \gamma\|\sigma^t\|_0^{(r-1)/2}.$$

This transition is always accepted as $\cos(\sigma^t,\theta)^{r-1} \ge \cos(\sigma,\theta)^{r-1} > \frac{\sqrt{k}}{\lambda}(C^*+\gamma)$.

(b) $\theta_i = 1$ and $\sigma_i^t = 0$:

$$D_t \le -\frac{\lambda}{k^{r/2}}\langle\sigma^t,\theta\rangle^{r-1} + C^*\|\sigma\|_0^{(r-1)/2} - \gamma\|\sigma^t\|_0^{(r-1)/2}.$$

With the same condition as (a), this transition is always rejected.

(c) $\theta_i = 0$ and $\sigma_i^t = 1$:

$$D_t \geq \gamma \|\sigma\|_0^{(r-1)/2} - C^* \|\sigma\|_0^{(r-1)/2}.$$

This transition is always accepted as $\gamma > C^*$.

(d) $\theta_i = 0$ and $\sigma_i^t = 0$:

$$D_t \leq -\gamma \|\sigma\|_0^{(r-1)/2} + C^* \|\sigma\|_0^{(r-1)/2}.$$

With the same condition as (c), this transition is always rejected.

Thus, with the stated conditions on $\langle \sigma^t, \theta \rangle$ and γ, the statement on D_t holds for time t and trivially the inequality on the cosine holds as well as we must have reduced the Hamming distance to the true solution, increasing the cosine of the angle. Thus, the statement is proven by induction up to time T_ξ, the maximal run time for which we can invoke Lemma 4.1. ■

The above lemma prove monotonic growth of $\cos(\sigma^t, \theta)$ when one can initialize in the set of $\cos(\sigma^1, \theta)$ large enough. It is then immediate that—for a good initialization–$\mathcal{MC}$ will output θ at the end of run-time if **every coordinate** $i \in [n]$ **has been proposed at least once**. The following lemma proves that the proposal sequence P from Definition 2.2 satisfies this condition with high probability. This result, combined with Lemma 4.2, gives a proof for Theorem 2.4 so long as we can show the initializations detailed in this theorem satisfy the conditions of Lemma 4.2.

LEMMA 4.3. *Let $P = (p_t)_{t \in \mathbb{N}}$ be a random sequence which sample each coordinate uniformly from the set $[n]$. Let T be the first time that each element of $[n]$ is in $P = (p_1, \ldots, p_T)$ and let T_i be the random variable $T_i = \sum_{j=1}^{\xi \cdot n/2} 1\{p_j = i\}$. With $\xi = \xi^* = C\log(n)$ for any constant $C \geq 25$, the following event holds with probability $1 - o(1)$:*

$$\{\Delta : T \leq \xi^* n/2 \cup T_i \leq \xi^* \text{ for all } i \in [n]\}. \tag{4.1}$$

Before proving this result we provide a simple interpretation of the above. The uniformly random proposal sequence will output each coordinate at least once before the end of run-time $\xi^* n/2$ with probability $1 - o(1)$. Moreover, we have that T_ξ, the run-time for which we can invoke Lemma 4.1, must be larger than $\xi^* n/2$ with probability $1 - o(1)$, meaning that we can interchange a statement with $t \in [T_\xi]$ with $t \in [\xi^* n/2]$.

PROOF. We prove that $\mathbb{P}(\mathcal{A}^c) = o(1)$ by showing both conditions in (4.1) have complements with probability $o(1)$ for the desired ξ^*. A union bound then gives the statement.

First, using a classic bound on the coupon collector problem from [77] (a sharper version of this bound is from Theorem 5.13 from [57]), we have that

$$\mathbb{P}(T > \xi^* n/2) = \mathbb{P}(T > Cn\log(n)/2) \leq n^{-C/2+1}.$$

Second, using that $T_i \sim \text{Binomial}(\xi^* n/2, \frac{1}{n})$ and a Chernoff bound (see Section 4.1 of [62]), we have that

$$\mathbb{P}(\cup_{i \in [n]} T_i > \xi^*) \leq n\mathbb{P}(T_i > .75\xi^*) \leq ne^{(.5)^2 \xi^*/6} = ne^{\xi^*/24} = n^{-C/24+1}.$$

Choosing any $C \geq 25$ gives the proof. ■

All of these results have now reduced the proof of Theorem 2.4 to demonstrating that the initialization $\mathbf{1}$ and e_i satisfy condition (4.2), which we recall as,

$$\cos(\sigma^1, \theta)^{r-1} > \frac{\sqrt{k}}{\lambda}(C^* + \gamma) \tag{4.2}$$

for some C^* from Lemma 4.2 for their stated values of λ. For simplicity, as $\xi^* = 25\log(n)$, we assume that $\gamma = \Theta(C^*) = \Theta(\log(n))$ (but still $\gamma > C^*$). Then, condition (4.2) is equivalent to

$$\lambda \geq C\log(n)\frac{\sqrt{k}}{\cos(\sigma^1, \theta)^{r-1}}, \tag{4.3}$$

for some $C > 0$. If we consider $\sigma^1 = \mathbf{1}$, then condition (4.3) is satisfied when,

$$\lambda \geq C \log(n) \frac{\sqrt{k}}{\sqrt{k/n}^{r-1}} = C \log(n) \frac{n^{(r-1)/2}}{k^{r/2-1}}. \tag{4.4}$$

Moreover, considering $\sigma^1 = e_i$, then condition (4.3) is satisfied when,

$$\lambda \geq C \log(n) \frac{\sqrt{k}}{\sqrt{1/k}^{r-1}} = C \log(n) k^{r/2}. \tag{4.5}$$

Inspecting equations (4.4) and (4.5), we see that running both $\mathcal{MC}(Y, \mathbf{1})$ and $\mathcal{MC}(Y, e_i)$, for our specified choice of ξ^* recovers the true solution for at least one of the two runs when $\lambda \geq C \log(n) \min\left(\frac{n^{(r-1)/2}}{k^{r/2-1}}, k^{r/2}\right)$, this is the famed Algorithm Threshold (up to a log factor) for the sparse tensor PCA problem from [18], which they previously claimed was likely impossible for Markov chains (or any other local algorithm) to succeed at.

Bibliography

[1] Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 15(3–4):196–392, 2019.

[2] Anima Anandkumar, Yuan Deng, Rong Ge, and Hossein Mobahi. Homotopy analysis for tensor pca, 2017.

[3] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor pca. *The Annals of Probability*, 48(4):2052–2087, 2020.

[4] Gérard Ben Arous, Alexander S Wein, and Ilias Zadik. Free energy wells and overlap gap property in sparse pca. *Communications on Pure and Applied Mathematics*, 76(10):2410–2473, 2023.

[5] Ned Augenblick, Jonathan Kolstad, Ziad Obermeyer, and Ao Wang. Pooled testing efficiency increases with test frequency. *Proceedings of the National Academy of Sciences*, 119(2):e2105180119, 2022.

[6] Victor Bapst, Amin Coja-Oghlan, Samuel Hetterich, Felicia Ra[m]ann, and Dan Vilenchik. The condensation phase transition in random graph coloring. *Communications in Mathematical Physics*, 341(2):543–606, October 2015.

[7] Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem, 2016.

[8] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2), April 2015.

[9] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, February 2011.

[10] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *arXiv preprint*, 2017.

[11] George David Birkhoff. The reducibility of maps. *American Journal of Mathematics*, 35:115, 1913.

[12] Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington-kirkpatrick model, 2012.

[13] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, 2021.

[14] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods, 2020.

[15] Wei Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:36, 2021.

[16] Zongchen Chen, Elchanan Mossel, and Ilias Zadik. Almost-linear planted cliques elude the metropolis process. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4504–4539. SIAM, 2023.

[17] Zongchen Chen, Elchanan Mossel, and Ilias Zadik. *Almost-Linear Planted Cliques Elude the Metropolis Process*, pages 4504–4539. ACM-SIAM, 2023.

[18] Zongchen Chen, Conor Sheehan, and Ilias Zadik. On the low-temperature mcmc threshold: the cases of sparse tensor pca, sparse regression, and a geometric rule, 2024.

[19] Zongchen Chen, Conor Sheehan, and Ilias Zadik. On the low-temperature mcmc threshold: the cases of sparse tensor pca, sparse regression, and a geometric rule. *arXiv preprint arXiv:2408.00746*, 2024.

[20] Amin Coja-Oghlan, Oliver Gebhard, Max Hahn-Klimroth, Alexander S. Wein, and Ilias Zadik. Statistical and computational phase transitions in group testing. *Proceedings of Machine Learning Research (COLT)*, 178:1-18, 2022.

[21] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the two-groups stochastic block model, 2015.

[22] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing, 2013.

[23] 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, November 2009.

[24] R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.

[25] Rishabh Dudeja, Subhabrata Sen, and Yue M Lu. Spectral universality of regularized linear regression with nearly deterministic sensing matrices. *arXiv preprint arXiv:2208.02753*, 2022.

[26] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices, 2021.

[27] Zhou Fan, Max Lovig, and Tianhao Wang. Untitled work corresponding to chapter 2. In preparation, 2025.

[28] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing, 2021.

[29] David Gamarnik, Aukosh Jagannath, and Subhabrata Sen. The overlap gap property in principal submatrix recovery. *Probability Theory and Related Fields*, 181:757–814, 2021.

[30] David Gamarnik, Cristopher Moore, and Lenka Zdeborová. Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114015, 2022.

[31] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property, 2019.

[32] David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. estimating squared error and a phase transition. *The Annals of Statistics*, 50(2):880–903, 2022.

[33] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *The Annals of Applied Probability*, 34(4):3375 – 3434, 2024.

[34] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations, 2022.

[35] Reza Gheissari, Aukosh Jagannath, and Yiming Xu. Finding planted cliques using markov chain monte carlo, 2023.

[36] Pavol Hell. From graph colouring to constraint satisfaction: There and back again. In Martin Klazar, Jan Kratochvíl, Martin Loebl, Jiří Matoušek, Pavel Valtr, and Robin Thomas, editors, *Topics in Discrete Mathematics*, pages 407–432, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[37] D. A. Holton and J. Sheehan. *The Petersen Graph*. Australian Mathematical Society Lecture Series. Cambridge University Press, 1993.

[38] S. Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

[39] Samuel B. Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-squares proofs, 2015.

[40] Fotis Iliopoulos and Ilias Zadik. Group testing and local search: is there a computational-statistical gap? *Proceedings of Machine Learning Research (COLT)*, 138:1-53, 2021.

[41] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling, 2012.

[42] Mark Jerrum. Large cliques elude the metropolis process. *Random Structures & Algorithms*, 3(4):347–359, 1992.

[43] Iain M. Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009. PMID: 20617121.

[44] Aayush Karan, Kulin Shah, Sitan Chen, and Yonina C. Eldar. Unrolled denoising networks provably learn optimal bayesian inference, 2024.

[45] Pravesh K. Kothari and Peter Manohar. A stress-free sum-of-squares lower bound for coloring, 2021.

[46] Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019.

[47] H. Kwang-Ming and D. Ding-Zhu. Pooling designs and nonadaptive group testing: important tools for DNA sequencing. *World Scientific*, 2006.

[48] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation, 2017.

[49] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, 2006.

[50] Max Lovig, Conor Sheehan, Kostas Tsirkas, and Ilias Zadik. Untitled work corresponding to chapter 3. In preparation, 2025.

[51] Maxwell Lovig and Ilias Zadik. On the mcmc performance in bernoulli group testing and the random max set-cover problem, 2024.

[52] Yanting Ma, Cynthia Rush, and Dror Baron. Analysis of approximate message passing with non-separable denoisers and markov random field priors, 2019.

[53] Camille Male. Traffic distributions and independence: permutation invariant random matrices and the three notions of independence, 2018.

[54] C. McMahan, J. Tebbs, and C. Bilder. Informative Dorfman screening. *Journal of the International Biometric Socienty*, 68:287–296, 2012.

[55] James A Mingo and Roland Speicher. Sharp bounds for sums associated to graphs of matrices. *Journal of Functional Analysis*, 262(5):2272–2288, 2012.

[56] James A. Mingo and Roland Speicher. *Free probability and random matrices.* Springer New York, 2017.

[57] Michael Mitzenmacher and Eli Upfal. *Probability and computing: randomization and probabilistic techniques in algorithms and data.* Combridge University Press, 2017.

[58] Andrea Montanari and Emile Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics, 2014.

[59] Andrea Montanari and Subhabrata Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists, 2024.

[60] Andrea Montanari and Alexander S. Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation, 2024.

[61] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization, 2022.

[62] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[63] R. Mourad, Z. Dawy, and F. Morcos. Designing pooling systems for noisy high-throughput protein-protein interaction experiments using boolean compressed sensing. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10:1478–1490, 2013.

[64] Leon Mutesa, Pacifique Ndishimye, Yvan Butera, Jacob Souopgui, Annette Uwineza, Robert Rutayisire, Ella Larissa Ndoricimpaye, Emile Musoni, Nadine Rujeni, Thierry Nyatanyi, et al. A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature*, 589(7841):276–280, 2021.

[65] H. Ngo and D. Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. *Discrete Mathematical Problems with Medical Applications*,

7:171–182, 2000.

[66] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models, 2017.

[67] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing, 2012.

[68] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. Vector approximate message passing, 2018.

[69] Jonathan Scarlett and Volkan Cevher. Near-optimal noisy group testing via separate decoding of items. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):902–915, 2018.

[70] Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation, 2019.

[71] Nelvin Tan, Pablo Pascual Cobo, Jonathan Scarlett, and Ramji Venkataramanan. Approximate message passing with rigorous guarantees for pooled data and quantitative group testing. *SIAM Journal on Mathematics of Data Science*, 6(4):1027–1054, October 2024.

[72] Nelvin Tan, Pablo Pascual Cobo, and Ramji Venkataramanan. Quantitative group testing and pooled data in the linear regime with sublinear tests, 2024.

[73] N. Thierry-Mieg. A new pooling strategy for high-throughput screening: the shifted transversal design. *BMC Bioinformatics*, 7:28, 2006.

[74] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[75] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks, 2024.

[76] Alexander S. Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca, 2019.

[77] Wikipedia contributors. Coupon collector's problem — Wikipedia, the free encyclopedia, 2024. [Online; accessed 6-February-2025].

[78] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020.

[79] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, 2021.

[80] Greg Yang. Tensor programs iii: Neural matrix laws, 2021.

[81] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks, 2023.

[82] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.