# Fourier Analysis of Iterative Algorithms

Chris Jones* Lucas Pesenti†

November 19, 2024

**Abstract**

We study a general class of nonlinear iterative algorithms which includes power iteration, belief propagation and approximate message passing, and many forms of gradient descent. When the input is a random matrix with i.i.d. entries, we use Boolean Fourier analysis to analyze these algorithms as low-degree polynomials in the entries of the input matrix. Each symmetrized Fourier character represents all monomials with a certain shape as specified by a small graph, which we call a *Fourier diagram*.

We prove fundamental asymptotic properties of the Fourier diagrams: over the randomness of the input, all diagrams with cycles are negligible; the tree-shaped diagrams form a basis of *asymptotically independent Gaussian vectors*; and, when restricted to the trees, iterative algorithms exactly follow an idealized Gaussian dynamic. We use this to prove a state evolution formula, giving a "complete" asymptotic description of the algorithm's trajectory.

The restriction to tree-shaped monomials mirrors the assumption of the *cavity method*, a 40-year-old non-rigorous technique in statistical physics which has served as one of the most important techniques in the field. We demonstrate how to implement cavity method derivations by 1) restricting the iteration to its tree approximation, and 2) observing that heuristic cavity method-type arguments hold rigorously on the simplified iteration. Our proofs use combinatorial arguments similar to the trace method from random matrix theory.

Finally, we push the diagram analysis to a number of iterations that scales with the dimension $n$ of the input matrix, proving that the tree approximation still holds for a simple variant of power iteration all the way up to $n^{\Omega(1)}$ iterations.

*Bocconi University. chris.jones@unibocconi.it.
†Bocconi University. lucas.pesenti@phd.unibocconi.it

# Contents

# 1 Introduction

We study nonlinear iterative algorithms which take as input a matrix $A \in \mathbb{R}^{n \times n}$, maintain a vector state $x_t \in \mathbb{R}^n$, and at each step

1. either multiply the state by $A$,
$$x_{t+1} = Ax_t \,,$$
2. or apply the same function $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ to each coordinate of the previous states,
$$x_{t+1} = f_t(x_t, \ldots, x_0) \,.$$

This class of algorithms has been coined *general first-order methods* (GFOM) [CMW20, MW22b]. GFOM algorithms are a simple, widespread, practically efficient, and incredibly powerful computational model. Alternating linear and nonlinear steps can describe first-order optimization algorithms including power iteration and many types of gradient descent (see [CMW20, GTM+22]). This definition also captures belief propagation and other message-passing algorithms which play a central role not only in the design of Bayes-optimal algorithms for planted signal recovery [FVRS22], but also recently in average-case complexity theory for the optimization of random polynomials [AMS23].

In machine learning and artificial intelligence, deep neural networks exhibit a similar structure which alternates multiplying weight matrices and applying nonlinear functions. Remarkably, viewed from this level of generality, the line blurs between neural networks and the gradient descent algorithms used to train them.

Despite the widespread use of GFOM and deep neural networks, developing a mathematical theory for these algorithms continues to be a major challenge. Thus far, it has been difficult to isolate mathematical theorems which describe key phenomena but avoid being too specific to any one setting, model, or algorithm. That being said, one effective theory has emerged at the interface of computer science, physics, and statistics for studying a class of nonlinear iterations known as Belief Propagation (BP) and Approximate Message Passing (AMP) algorithms. This theory is most developed for inputs $A$ that are *dense random matrices with i.i.d entries*, also known as a *mean-field models* in physics, and which can be considered the simplest possible model of random data.

The analysis of BP and AMP algorithms in this setting can be summarized by the *state evolution* formula [DMM09, Bol14]. This is an impressive "complete" description of the trajectory of the iterates $x_t \in \mathbb{R}^n$, in the limit $n \to \infty$. Specifically, state evolution defines a sequence of *scalar* random variables $X_t$ such that for essentially *any* symmetric quantity of interest related to $x_t$, the expectation of a corresponding expression in $X_t$ approximates the quantity with an error that goes to 0 as $n \to \infty$. This yields analytic formulas for quantities such as the loss function or objective value achieved by $x_t$, the norm of $x_t$, the correlation between $x_s$ and $x_t$ across iterations, or the fraction of $x_t$'s coordinates which lie in the interval $[-1, +1]$. The ability to precisely analyze the trajectory and the fixed points of message-passing algorithms (through $X_t$ with large $t$) has been key to their applications.

State evolution for BP/AMP iterations was originally predicted using a powerful and influential technique from statistical physics known as the *cavity method*. Variants of BP

have been studied in physics as "non-linear dynamical systems" as far back as the work of Bethe [Bet35], although the algorithmic perspective came into prominence only later. The cavity method and the related replica method were devised in the 1980s [Par79, Par80, MPV86, MPV87], initially as a tool to compute thermodynamic properties of disordered systems, and later as a tool for analyzing belief propagation algorithms. Since their introduction, the cavity method and the replica method have served as two of the most fundamental tools in the statistical physics toolbox.

The deployment of these techniques has undoubtedly been a huge success; there are many survey articles offering various perspectives from different fields [YFW03, MM09, KF09, ZK16, Gab20, FVRS22, ZY22, CMP+23]. However, the reality is that the situation is not as unified as the above picture would suggest, due to a major issue: *the physical methods are not mathematically rigorous.*

At present, there exists a significant gap between how results are established in the physical and mathematical literature. The two general types of results are: 1) simple non-rigorous arguments based on the cavity/replica method; 2) mathematically rigorous arguments that confirm the physical predictions, but with technically sophisticated proofs that can't closely follow the path of the physical reasoning. For example, the state evolution formula was first proven by Bolthausen [Bol14] using a Gaussian conditioning technique which is fairly technically involved. Although many proofs have been found for predictions of the cavity and replica methods, none can be said to clearly explain the success of the physicists' techniques.

It has appeared that the physicists have some secret advantage yet unmatched by rigorous mathematics. Is there a simple and rigorous mathematical framework that explains why the assumptions made by physicists always seem to work?

## 1.1 Our contributions

We introduce a new method to analyze nonlinear iterations based on Fourier analysis, when the input to the algorithm is a random matrix with i.i.d entries. Our framework gives proofs that are able to closely follow heuristic physics derivations.

Our strategy is to replace the original iteration $(x_t)_{t\geq 0}$ by an idealized version

$$x_t \approx \widehat{x}_t\,,$$

which we call the *tree approximation* to $x_t$. The analysis then follows a two-step structure:

1. The tree approximation $\widehat{x}_t$ tracks the original iteration $x_t$ up to a uniform $\widetilde{O}(n^{-\frac{1}{2}})$ entrywise error. Hence, any reasonable asymptotic result established on $(\widehat{x}_t)_{t\geq 0}$ (such as the joint distribution of their entries) automatically extends to $(x_t)_{t\geq 0}$.

2. Cavity method-type reasoning can be rigorously applied to the tree approximation. In cases where $\widehat{x}_t$ has already been analyzed in physics, one can essentially copy the heuristic physics derivation.

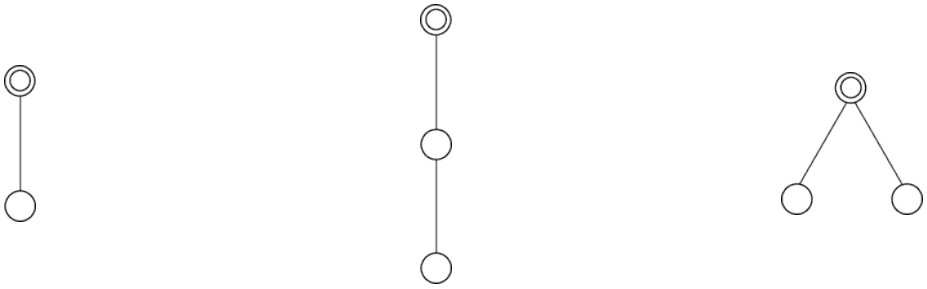
Analyzing $\widehat{x}_t$ is a significant simplification compared to the entire state $x_t$—in fact, we show that the former follows an explicit *Gaussian dynamic*. The simplification directly

yields a *state evolution* formula for GFOM algorithms, as well as rigorous implementations of physics-based cavity method arguments (in the algorithmic or "replica symmetric" setting of the method). In other words, our new notion of tree approximation matches implicit assumptions of the cavity method and gives a way to justify them.

We define the tree approximation $\widehat{x}_t$ essentially as follows: we expand the entries of $x_t$ as polynomials in the entries of the input matrix $A \in \mathbb{R}^{n\times n}$. If we represent the monomials (e.g. $A_{12}A_{23}A_{24}$) as graphs in the natural way, then $\widehat{x}_t$ consists of only the monomials appearing in $x_t$ whose graph is a tree. Hence, we will show that the state of an iterative algorithm can be tightly approximated using the much smaller set of tree monomials.

**Fourier diagrams.** We view iterates of a GFOM with polynomial non-linearities as vector-valued polynomials in the entries of $A$. These polynomials have a special symmetry: they are invariant under permutations of the row/column indices of $A$.

The polynomial representation can be visualized using *Fourier diagrams*, each of which is a small graph representing all the monomials with a given shape. For example, here are three Fourier diagrams along with the vectors associated with them.



$$Z_i := \sum_{\substack{j=1 \\ i,\,j \text{ distinct}}}^{n} A_{ij} \qquad Z'_i := \sum_{\substack{j,k=1 \\ i,\,j,\,k \text{ distinct}}}^{n} A_{ij}A_{jk} \qquad Z''_i := \sum_{\substack{j,k=1 \\ i,\,j,\,k \text{ distinct}}}^{n} A_{ij}A_{ik}$$

In general, a Fourier diagram is an undirected rooted multigraph $\alpha = (V(\alpha), E(\alpha))$ which represents the vector $Z_\alpha \in \mathbb{R}^n$ whose entries are:

$$Z_{\alpha,i} := \sum_{\substack{\text{injective } \varphi\colon V(\alpha)\to[n] \\ \varphi(\circledcirc)=i}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}\,, \qquad \text{for all } i \in [n]\,. \tag{1}$$

We use $\circledcirc \in V(\alpha)$ to notate the root vertex.

The symmetry of the GFOM operations ensures that in the polynomial representation of an iterate $x_t$, all monomials corresponding to the same Fourier diagram come with the same coefficient. Therefore, any iterate $x_t$ of a GFOM with polynomial non-linearities can be expressed as a linear combination of Fourier diagrams, in which case we say that it is written *in the Fourier diagram basis*.

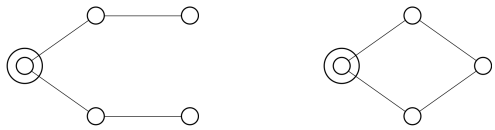
We emphasize that these diagrams are constructed by summing over *injective* embeddings $\varphi\colon V(\alpha) \to [n]$, a crucial detail for the results that follow. The term "Fourier" reflects that this basis of polynomials is a symmetrized version of the standard Fourier basis from Boolean function analysis (see Section 3.4).

**Asymptotic diagram analysis.** It turns out that something special happens to the Fourier diagram basis in the limit $n \to \infty$, when $A$ is a symmetric matrix with independent mean-0, variance-$\frac{1}{n}$ entries. Informally, the entries of the diagrams become mutually independent, and the following properties hold.

- The Fourier diagrams with cycles are negligible.
- The Fourier tree diagrams with one branch from the root are independent Gaussian vectors.
- The Fourier tree diagrams with several branches from the root are Hermite polynomials in the Gaussians represented by the branches.

Most importantly, **the only non-negligible contributions come from the trees**. Based on this classification, we define the *tree approximation* $\widehat{x}_t$ of an expression $x_t$ written in the Fourier diagram basis to be obtained by discarding all diagrams with cycles. That is, as polynomials in $A$, $\widehat{x}_t$ consists of the tree-shaped monomials in $x_t$.

The reason that cyclic Fourier diagrams are negligible is combinatorially intuitive: cyclic diagrams sum over fewer terms than tree-shaped diagrams. For example, the left diagram is a sum over $\approx n^4$ terms, each mean-zero with magnitude $\sim n^{-2}$, so the overall magnitude is $\Theta(1)$. The right diagram is a sum over $\approx n^3$ terms, again of magnitude $n^{-2}$, so the overall of order of the diagram is $\Theta(n^{-1/2})$.



We now state our main theorems. In all of them, we assume that $A$ is a symmetric matrix with independent mean-0 variance-$\frac{1}{n}$ entries (see Assumption 2.1). First, we formalize the classification above by proving that all joint moments of the Fourier diagrams converge to those of the corresponding random variables in a Gaussian space.

**Theorem 1.1** (Classification theorem; see Theorem 4.11)**.** *For any $k \geq 0$ independent of $n$, for all connected Fourier diagrams $\alpha_1, \ldots, \alpha_k$ and $i_1, \ldots, i_k \in [n]$ (allowing repetitions in $\alpha_j$ and $i_j$),*

$$\mathop{\mathbb{E}}_{A}\left[\prod_{j=1}^{k} Z_{\alpha_j, i_j}\right] = \mathbb{E}\left[\prod_{j=1}^{k} Z^{\infty}_{\alpha_j, i_j}\right] + O(n^{-\frac{1}{2}}),$$

*where for any connected Fourier diagram $\alpha$ and $i \in [n]$,*

1. *$Z^{\infty}_{\alpha,i} = 0$, if $\alpha$ has a cycle.*
2. *$Z^{\infty}_{\alpha,i} \sim \mathcal{N}(0, |\mathrm{Aut}(\alpha)|)$ independently, if $\alpha$ is a tree whose root has degree 1.*
3. *$Z^{\infty}_{\alpha,i} = \prod_{\tau} h_{d_\tau}(Z^{\infty}_{\tau,i}\,; |\mathrm{Aut}(\tau)|)$ if $\alpha$ is a tree consisting of $d_\tau$ copies of each tree $\tau$ from case 2 merged at the root, where $h_{d_\tau}$ are the Hermite polynomials (defined in Section 2).*

Next, we prove that the tree approximation of a GFOM closely tracks the original iteration. This addresses the first of the two steps from the overview of our method.

**Theorem 1.2** (Tree approximation of GFOMs; see Theorem 4.14)**.** *Let $t$ be a constant, $x_t \in \mathbb{R}^n$ be the state of a GFOM with polynomial non-linearities, and $\widehat{x}_t \in \mathbb{R}^n$ be the state obtained by performing the GFOM operations on only the tree diagrams. Then with high probability over $A$,*

$$\|x_t - \widehat{x}_t\|_\infty = \widetilde{O}(n^{-\frac{1}{2}}) .$$

The statement of this theorem exactly isolates a key and subtle point: not only are the cyclic diagrams negligible at time $t$, but they will never blow up to affect the state at any future time. The fact that "approximation errors do not propagate" is what gives us the ability to pass the algorithm to an asymptotic limit.[1]

The proof of Theorem 1.2 is intuitive. According to the diagram classification theorem, we can tease out the approximation error for $x_t$ as the monomials with cycles, whereas the approximating quantity $\widehat{x}_t$ consists of the tree monomials. When a GFOM operation is applied, the cycles persist in all cyclic monomials, and hence they continue to be negligible.

As a direct consequence of these results, we deduce a very strong form of state evolution for all GFOM algorithms. The theorem below paints a nearly complete picture of the evolution of $x_t$ as $n$ independent trajectories of a single random variable $X_t$ which is an "idealized Gaussian dynamic" in correspondence with $\widehat{x}_t$.

**Theorem 1.3** (General state evolution; see Theorem 4.18)**.** *Let $t$ be a constant, $x_t \in \mathbb{R}^n$ be the state of a GFOM with polynomial non-linearities, and let $X_t$ be the asymptotic state of $x_t$ (Definition 3.7). Then:*

*(i) For each $i \in [n]$, $(x_{0,i}, \dots, x_{t,i}) \xrightarrow{\mathrm{d}} (X_0, \dots, X_t)$. Furthermore, the coordinates' trajectories $\{(x_{0,i}, \dots, x_{t,i}) : i \in [n]\}$ are asymptotically independent.*

*(ii) With high probability over $A$,*

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

Quantities such as the norm of $x_t$ can be computed using part (ii) along with one additional GFOM iteration that squares $x_t$ componentwise. Without much extra work, Theorem 1.3 also encapsulates other key features of previous state evolution formulas including quantitative error bounds (similar to the main result of [RV18]) and universality (the main result of [BLM15]).[2]

[1] This directly addresses a question raised in the seminal paper of Donoho, Maleki, and Montanari on approximate message passing [DMM10, Section III.E].

[2] Similarly to [BLM15], our technical analysis assumes that the nonlinearities in the GFOM are polynomial functions, but other works have been able to handle the larger class of *pseudo-Lipschitz* non-linearities. We do not find this assumption to be too restrictive since it is known in many cases that we can approximate the non-linearities by polynomials [IS24, Appendix B].

**The cavity method.** To explain the cavity method in one sentence, it allows you to assume that "loopy" message-passing algorithms on random dense graphs behave as if on a tree, gaining extra properties such as the independence of incoming messages. It turns out that the assumption of being on a tree matches the restriction to tree-shaped monomials in $A$, leading to a way to rigorously implement simple cavity method reasoning.

We formalize two types of cavity method arguments. For the first one, we introduce a combinatorial notion of asymptotic equality $\overset{\infty}{=}$ which can rigorously replace heuristic approximations in the cavity method.

**Definition 1.4** ($\overset{\infty}{=}$, informal version of Definition 4.4)**.** *Let $x \overset{\infty}{=} y$ if $x - y$ is a sum of constantly many diagrams with cycles.*

As an application of this definition, we implement the cavity method argument that belief propagation and approximate message passing are asymptotically equivalent for dense random matrices.

**Theorem 1.5** (Equivalence of BP and AMP; see Theorem 5.1)**.** *Let $m_t^{\mathrm{BP}}$ and $m_t^{\mathrm{AMP}}$ be the iterates of respectively the belief propagation and the approximate message passing iterations on the same non-linearities (see Eqs. (11) and (13)). Then with high probability over $A$,*

$$\left\| m_t^{\mathrm{BP}} - m_t^{\mathrm{AMP}} \right\|_\infty = \widetilde{O}(n^{-\frac{1}{2}}) \,.$$

We also use $\overset{\infty}{=}$ to prove a fundamental assumption of the cavity method for belief propagation iterations on dense models, namely that the messages incoming at a vertex are asymptotically independent.

**Theorem 1.6** (Asymptotic independence of incoming messages; see Theorem 5.9)**.** *Let $m_t^{\mathrm{BP}}$ be the iterates of a belief propagation iteration (Eq. (11)). For any $j \in [n]$, the incoming messages at $j$, $\{m_{i\to j}^t : i \in [n], i \neq j\}$, are asymptotically independent.*

The second way that we formalize the cavity method reasoning is through the idealized Gaussian dynamic $X_t$ in Theorem 1.3. We recover the vanilla form of state evolution for approximate message passing, since in this case, $X_t$ has a simple description.

**Theorem 1.7** (Asymptotic state of AMP; see Theorem 5.10)**.** *Consider the AMP iteration*

$$x_{t+1} = A f_t(x_t, \dots, x_0) - \frac{1}{n} \sum_{s=1}^{t} \sum_{i=1}^{n} \frac{\partial f_t}{\partial x_s}(x_{t,i}, \dots, x_{0,i}) f_s(x_s, \dots, x_0) \,. \tag{2}$$

*The asymptotic state of $(x_0, x_1, \dots)$ is a centered Gaussian vector $(X_0, X_1, \dots)$ with covariances given by the recurrence, for all $s, t$,*

$$\mathbb{E}\left[X_s X_t\right] = \mathbb{E}\left[f_{s-1}(X_{s-1}, \dots, X_0) f_{t-1}(X_{t-1}, \dots, X_0)\right] \,.$$

The subtracted term in Eq. (2) is called the *Onsager correction* which, as we show, is carefully designed to cancel out a backtracking term in the asymptotic tree space (Lemma 5.11).

We emphasize that Theorems 1.5 and 1.7 are *known*. They were originally predicted with the cavity method, then later confirmed by rigorous proofs (in [BLM15] and [Bol14,

BM11, CMW20], respectively). The main message about our proofs is the new and quite comprehensive perspective obtained through the tree approximation, providing a clear way in which GFOM algorithms on dense random inputs "can be assumed to occur on a tree".

Finally, we provide an exposition in Section 5.5 of the breakthrough *iterative AMP* algorithm devised by Montanari to compute ground states of the Sherrington–Kirkpatrick model [Mon19, AM20, AMS21]. We explain from the diagram perspective how the algorithm is the optimal choice among algorithms which "extract" a Brownian motion from the input.

**Taking the tree approximation farther.** The asymptotic theory above applies to an iterative algorithm running for a *constant* number of iterations. Although this "short-time" setting is used in a large majority of previous works in this area, there is interest in extending the analysis to, say, $O(\log n)$ iterations, which may be enough to capture planted recovery from random initialization and distinct phases of learning algorithms [LFW23].

Can we use the tree-like Fourier characters to analyze the long-time behavior? We show in Section 6 that some care needs to be taken. First, we prove a positive result, that the tree approximation continues to hold for $n^{\Omega(1)}$ iterations for a simple belief propagation algorithm (debiased power iteration, or asymptotically equivalently, power iteration on the non-backtracking walk matrix).

**Theorem 1.8** (See Theorem 6.2)**.** *Generate $x_t \in \mathbb{R}^n$ from the debiased power iteration and let $\widehat{x}_t$ be the tree approximation to $x_t$. Then there exist universal constants $c, \delta > 0$ such that for all $t \le cn^\delta$,*

$$\|x_t - \widehat{x}_t\|_\infty \xrightarrow{\text{a.s.}} 0\,.$$

However, we also identify some problems with the technology which suggest that new ideas will be needed to completely capture the long-time setting. We observe that the asymptotic Gaussian classification theorem is no longer valid for diagrams of size $t \approx \log n$. Finally, we identify a further threshold at $t \approx \sqrt{n}$ iterations beyond which the tree approximation we use seems to break down.

**Conclusion.** We demonstrate that for iterative algorithms running on dense random inputs, trees are all you need. The tree-shaped Fourier diagrams form an asymptotic basis of independent Gaussian vectors associated to an arbitrary Wigner matrix. This basis seems extremely useful, and we are not aware of any previous works on it.

We note that from the outset, it is not at all clear how to find this basis. Individual monomials (i.e. Boolean Fourier characters) such as $A_{12}A_{23}A_{34}$ and $A_{12}A_{23}A_{13}$ have the same magnitude, and the asymptotic negligibility of the cyclic terms including $A_{12}A_{23}A_{13}$ only appears after summing up the total contribution of all monomials with the same shape. Furthermore, grouping terms in a different way does not identify our notion of tree approximation, such as by allowing repeated indices in Eq. (1) (as done in [BLM15, IS24]). In this repeated-label representation, there is no clear notion of tree approximation of iterative algorithms (in fact, with this alternative definition, the iterates of a GFOM can always be represented *exactly* with trees!) or of the simplified Gaussian dynamic on trees, which is central to our approach.

As we show, the Fourier tree approximation leads to streamlined proofs of several arguments based on the cavity method. We believe that this framework has potential to generalize well beyond the Wigner case and to address outstanding open problems in the area—such as the long-time setting mentioned above.

## 1.2 Related work

**Comparison with prior work.** Our analysis is based on the recent "low-degree paradigm" in which algorithms are analyzed as low-degree multivariate polynomial functions of the input [KWB19]. Several recent works have used a similar approach for iterative algorithms [BLM15, MW22a, IS24], with subtle but crucial differences to our work.

Bayati, Lelarge, and Montanari [BLM15] decompose the AMP iterates into certain "non-reversing" labeled trees. They also observe that the Onsager correction corresponds to a backtracking term. Montanari and Wein [MW22a, Section 4.2] introduce an orthogonal diagram basis (similar to our Fourier diagram basis) in their proof that no low-degree polynomial estimator can outperform AMP for rank-1 matrix estimation. Ivkov and Schramm [IS24] use a repeated-label representation to show that AMP can be robustified.

Diagrammatically, the main advantage of our method is the precise choice of the Fourier diagram basis. By summing over injective a.k.a. self-avoiding labelings $\varphi$ in Eq. (1), each diagram exactly describes all monomials with a given shape. When working with other polynomial basis, for example diagrams with repeated labels [BLM15, IS24] (see Appendix A.3), the key properties of the Fourier diagram basis (the family of asymptotically independent Gaussian vectors and the associated Gaussian dynamic) do not seem clearly visible. In particular, previous work does not show the tree approximation.

Our results stated above which are cavity method-based reproofs of existing results are Theorem 1.5, which essentially follows from [BLM15, Proposition 3], and Theorem 1.7, which was first proven by Bayati and Montanari [BM11]. Notably, Bayati, Lelarge, and Montanari [BLM15] use an approach based on the moment method as we do. Their proof is somewhat more technical, it does not use the Fourier diagram basis, and it is not able to clearly follow the simple cavity method argument that we reproduce in Section 5.2.1.

We also compare our state evolution formula for GFOM in Theorem 1.3 with a state evolution formula for GFOM proven by [CMW20]. They give a reduction from GFOM to AMP to derive a state evolution formula for GFOM. The corresponding description of the asymptotic state $X_0, \dots, X_t$ is inside a very compressed probability space generated by $t$ Gaussians with a certain covariance structure.

Our description of the random variables $X_0, \dots, X_t$ (necessarily having the same distribution) has a simpler interpretation inside a larger probability space generated by $(Z_\sigma^\infty)_{\sigma \in \mathcal{S}}$. Both descriptions of the asymptotic state $X_t$ are likely to be valuable for different purposes or explicit calculations. Our formulation of state evolution also includes the asymptotic independence of the trajectories of different coordinates.

**Analyzing algorithms as low-degree polynomials.** Our technical framework is adapted from the average-case analysis of *Sum-of-Squares* algorithms. The Sum-of-Squares algorithm is a powerful meta-algorithm for combinatorial optimization and statistical inference [RSS18, FKP19]. Sum-of-Squares has been successfully analyzed on i.i.d. random inputs using *graph matrices*, which are a Fourier basis for matrix-valued functions of a random matrix $A$ in the same way that our diagram basis is a basis for vector-valued functions of $A$.

The theory appears much more pristine in the current setting, so we hope that the current results can bring some new clarity to the technically challenging works on Sum-of-Squares. Many key ideas on graph matrices are present in a pioneering work by Barak et al. which analyzes the Sum-of-Squares algorithm for the Planted Clique problem [BHK+19] (building on earlier work [DM15, MPW15, HKP+18]). Analytical ideas were subsequently isolated by Ahn, Medarametla, and Potechin [AMP20] and Potechin and Rajendran [PR20, PR22] and developed in several more works [GJJ+20, RT23, JPR+21, JP22, Jon22, JPRX23, KPX24]. Several recent works have made explicit connections between AMP, Sum-of-Squares, and low-degree polynomials [MW22a, IS24, SS24a, SS24b]. Another similar class of diagrammatic techniques are *tensor networks* [MW19, KMW24].

**Statistical physics and the cavity method.** The cavity and replica methods are widely used in statistical physics to compute the free energy, complexity, etc. of Gibbs distributions on large systems, or similarly to compute the satisfiability threshold, number of solutions, etc. for many non-convex random optimization problems. For an introduction to statistical physics methods in computer science, we recommend the surveys [MMZ01, ZK16, Gab20], the book [MM09], and the 40-year retrospective [CMP+23]. The cavity method is described in [MP03] and [MM09, Part V].

Rigorously verifying the predictions of the physical methods has been far from easy for mathematicians. To highlight some major landmarks in the literature over the past decades, tour-de-force proofs of the Parisi formula for the free energy of the SK model were developed by Talagrand [Tal06, Tal10] and Panchenko [Pan13]. Ding, Sly, and Sun [DSS16, DSS14, DSS22] identified the satisfiability threshold for several random optimization problems including $k$-SAT with large $k$. Ding and Sun [DS19] and Huang [Hua24] rigorously analyze the storage capacity of the Ising perceptron, assuming a numerical condition.

Note that the results above are strictly outside the regime of the current work. They require the replica method in "replica symmetry breaking" settings, whereas we study the simpler but related cavity method in the replica symmetric setting. $k$-SAT is also a sparse (a.k.a. dilute) model whereas our results are for dense (a.k.a. mean-field) models. Despite these differences, our results tantalizingly suggest that it may be possible to validate the physical techniques in a more direct and generic way than taken by current approaches.

Other authors have also directly considered the cavity assumption, albeit using a less combinatorial approach. Both proofs of the Parisi formula implement analytic forms of the cavity calculation ([Tal10, Section 1.6] and [Pan13, Section 3.5]). The cavity method can also be partially justified for sparse models in the replica symmetric regime using that the interactions are locally treelike with high probability [BN06, CKPZ17].

Diagrammatic methods are common in physics, and in fact they have been used in

the vicinity of belief propagation even since a seminal 1977 paper by Thouless–Anderson–Palmer [TAP77] which introduced the TAP equations of Eq. (9). A version of the tree approximation actually appears briefly in their diagrammatic formula for the free energy of the SK model in Section 3. However, it has not been clear how or whether these arguments could be made rigorous, and to date, rigorous proofs have not directly followed these approaches.

**Belief propagation and AMP.** Belief propagation originates in computer science and statistics from Pearl [Pea88]. In the current setting, we can view the underlying graphical model as the complete graph, with correlations between the variables induced by the random matrix $A$. State evolution was first predicted for BP algorithms in this setting by Kabashima [Kab03] and Donoho–Maleki–Montanari [DMM09]. Since the first rigorous proof of state evolution by Bolthausen [Bol14], his Gaussian conditioning technique has been extended to prove state evolution for many variants of AMP [BM11, JM13, MRB17, Tak19a, BMN20, Tak19b, AMS21, Tak21, Lu21, FVRS22, Fan22, GB23, HS23].

A notably different proof of state evolution by Bayati, Lelarge, and Montanari [BLM15] uses a moment-based approach which is closer to ours (see also follow-up proofs [CL21, DG21, WZF22, DLS23]). These proofs and also ours show universality statements which the Bolthausen conditioning method cannot handle.

All of the above works restrict themselves to a constant number of iterations, although some recent papers push the analysis of AMP in some settings to a superconstant number of iterations [RV18, CR23, WZ23, WZ24]. Very recently, [LW22, LFW23] managed to analyze $t = \widetilde{\Omega}(n)$ iterations of AMP in the spiked Wigner model. This last line of work is especially intriguing, given that our approach seems to break down at $t \approx \sqrt{n}$ (Section 6.1).

The perspective that we take is slightly different from most of these papers. Whereas previous works analyze the asymptotic *distribution* of the AMP iterates over the randomness of $A$, we give an explicit function $\widehat{x}_t$ which exactly satisfies a "Gaussian dynamics" and asymptotically approximates the iterates. This general approach provides more information and we hope that it has increased potential for generalization.

On first-order iterations which are not BP/AMP algorithms, a smaller number of physical analyses have been performed using the more general techniques of *dynamical mean field theory* [MSR73]. We refer to the survey [Gab20]. Most analyses rely on heuristic arguments, although some more recent works [CCM21, GTM+22, LSS23] prove rigorous results.

Finally, we note that the tree approximation bears similarities to the suppression of crossing partitions in free probability [NS06]. Unlike the traditional viewpoint of free probability, the combinatorial cancellations behind the tree approximation occur directly on the trajectory of random objects (the iterates of the algorithm), and not only for averaged quantities associated with them.

## 1.3 Organization of the paper

After background preliminaries in Section 2, we introduce the diagrams in Section 3 and describe their key properties without proofs.

In Section 4, we present the full diagram analysis: we define the key notion of asymptotic equality $\overset{\infty}{=}$, and we prove three central theorems: the classification of the diagrams (Theorem 4.11), the tree approximation for GFOMs (Theorem 4.14), and a general state evolution formula (Theorem 4.18).

In Section 5, we demonstrate the connection with the cavity method by proving the equivalence between belief propagation and approximate message passing (Theorem 5.1), the independence of incoming messages in belief propagation (Theorem 5.9), the state evolution formula for AMP (Theorem 5.10), and the analysis of the iterative AMP algorithm of Montanari.

Finally, Section 6 investigates algorithms running for a large number of iterations.

Appendices A to C contain omitted proofs and calculations.

**Acknowledgments.** We are pleased to thank numerous colleagues for feedback and discussions as this work developed, including Giorgi Kanchaveli, Carlo Lucibello, Enrico Malatesta, Goutham Rajendran, Alon Rosen, and Riccardo Zecchina. Work supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement Nos. 834861 and 101019547). CJ is also a member of the Bocconi Institute for Data Science and Analytics (BIDSA).

# 2 Preliminaries

To maintain generality, we specify the input (a random matrix) and the algorithm (a first-order iteration), but we do not specify an objective/energy function, and for this reason our results are in the flavor of random matrix theory. While the setting of this paper is a null model without any hidden signal, we expect that our techniques can also be applied to planted recovery problems. A concrete algorithmic application to keep in mind in the null model is the optimization of random degree-2 polynomials that we revisit in Section 5.5.

Our results will apply universally to a Wigner random matrix model (they hold regardless of the specific choice of $\mu, \mu_0$ below).

**Assumption 2.1** (Assumptions on matrix entries)**.** *Let $\mu$ and $\mu_0$ be two subgaussian*[3] *distributions on $\mathbb{R}$ such that $\mathbb{E}_{X\sim\mu}[X] = 0$ and $\mathbb{E}_{X\sim\mu}[X^2] = 1$.*

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

The subgaussian assumption on $\mu$ and $\mu_0$ can be relaxed to require only the existence of the $q$-th moment of $\mu$ for some large enough constant $q \in \mathbb{N}$ that depends only on the number of iterations and the degree of the nonlinearities appearing in the algorithm. In this case, our statements of the form "$\|x_n - y_n\|_\infty = \widetilde{O}(n^{-1/2})$ with high probability"[4] weaken to "$\|x_n - y_n\|_\infty \xrightarrow{\text{a.s.}} 0$".

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

[4] We say a sequence of events $(A_n)_{n\ge 0}$ occurs with high probability if $\Pr(A_n) \ge 1 - 1/\mathrm{poly}(n)$.

**Definition 2.2** (Convergence of random vectors)**.** *Let $(X_n)_{n\in\mathbb{N}}$ and $Z$ be random vectors.*

- *We write $X_n \xrightarrow{\text{a.s.}} Z$ if $X_n$ converges to $Z$ almost surely, i.e. $\lim_{n\to\infty} X_n$ exists and equals $Z$ with probability 1.*
- *We write $X_n \xrightarrow{\text{d}} Z$ if $X_n$ converges to $Z$ in distribution, i.e. for every real-valued bounded continuous function $f$, $\lim_{n\to\infty} \mathbb{E}\, f(X_n)$ exists and equals $\mathbb{E}\, f(Z)$.*

We can derive convergence in distribution of random vectors by computing their moments.

**Lemma 2.3** (Method of moments [Bil95, Theorems 29.4, 30.1, and 30.2])**.** *Let $X_n \in \mathbb{R}^d$ for $n \in \mathbb{N}$ and $Z \in \mathbb{R}^d$ be random vectors such that for any $q_1, \dots, q_d \in \mathbb{N}$,*

$$\mathbb{E}\left[\prod_{i=1}^{d} X_{n,i}^{q_i}\right] \underset{n\to\infty}{\longrightarrow} \mathbb{E}\left[\prod_{i=1}^{d} Z_i^{q_i}\right].$$

*Suppose that for all $i \in [n]$, $Z_i$ has the distribution of a polynomial in Gaussian random variable. Then $X_n \xrightarrow{\text{d}} Z$.*

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

The Gaussian distribution and Hermite polynomials have combinatorial interpretations related to matchings.

**Lemma 2.4.** *For $Z \sim \mathcal{N}(0,\sigma^2)$,*

$$\mathbb{E}\left[Z^q\right] = |\mathcal{PM}(q)|\sigma^{\frac{q}{2}} = \begin{cases} (q-1)!! \cdot \sigma^{\frac{q}{2}} & \text{if } q \text{ is even} \\ 0 & \text{if } q \text{ is odd} \end{cases},$$

*where $\mathcal{PM}(q)$ is the set of perfect matchings on $q$ objects and $(q-1)!! = \frac{q!}{2^{q/2}(q/2)!}$.*

**Lemma 2.5** ([Jan97, Theorem 3.4 and Example 3.18])**.** *For all $k \geq 0$ and $x \in \mathbb{R}$,*

$$h_k(x;\sigma^2) = \sum_{M\in\mathcal{M}(k)} (-1)^{|M|}\sigma^{2|M|}x^{k-2|M|},$$

*where $\mathcal{M}(k)$ is the set of (partial) matchings on $k$ objects (including the empty matching and perfect matchings).*

**Lemma 2.6** ([Jan97, Theorem 3.15 and Example 3.18])**.** *For any $k_1, \dots, k_l \geq 0$ and $x \in \mathbb{R}$,*

$$h_{k_1}(x;\sigma^2)\cdots h_{k_\ell}(x;\sigma^2) = \sum_{M\in\mathcal{M}(k_1,\dots,k_\ell)} h_{k-2|M|}(x;\sigma^2)\sigma^{2|M|},$$

*where $\mathcal{M}(k_1,\dots,k_\ell)$ is the set of (partial) matchings on $k = k_1+\cdots+k_\ell$ objects divided into $\ell$ blocks of sizes $k_1,\dots,k_\ell$ such that no two elements from the same block are matched.*

Finally, we recall:

**Lemma 2.7** (Gaussian integration by parts)**.** *Let $(Z_1, \ldots, Z_k)$ be a centered Gaussian vector. Then for all smooth $f : \mathbb{R}^k \to \mathbb{R}$,*

$$\mathbb{E}\left[Z_1 f(Z_1, \ldots, Z_k)\right] = \sum_{i=1}^{k} \mathbb{E}\left[Z_1 Z_i\right] \mathbb{E}\left[\frac{\partial f}{\partial z_i}(Z_1, \ldots, Z_k)\right].$$

# 3 The Diagram Basis

Here we give the key properties of the Fourier diagrams on a high level, delaying formal statements and proofs to the next section.

- In Section 3.1, we give an example.
- In Section 3.2, we define the class of diagrams and describe their behavior both for fixed $n$ and in the limit $n \to \infty$.
- In Section 3.3, we summarize how iterative algorithms behave asymptotically.
- In Section 3.4, we explain how the diagram basis can be derived from standard discrete Fourier analysis.

## 3.1 Example of using diagrams

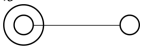
We show how to compute the vector $A(A\vec{1})^2$ in the diagram basis, where $\vec{1} \in \mathbb{R}^n$ denotes the all-ones vector and the square function is applied componentwise. Calculation with diagrams is a bit like a symbolic version of the trace method from random matrix theory [Bor19].

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

We will use rooted multigraphs to represent vectors.[5] Multigraphs may include multi-edges and self-loops. In our figures, the root will be drawn as a circled vertex ◎. The vector $\vec{1}$ will correspond to the singleton graph with one vertex (the root): ◎. Edges will correspond to $A_{ij}$ terms.

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

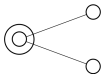
$$(A\vec{1})_i = \sum_{j=1}^{n} A_{ij} = \sum_{\substack{j=1 \\ i,j \text{ distinct}}}^{n} A_{ij} \equiv$$



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

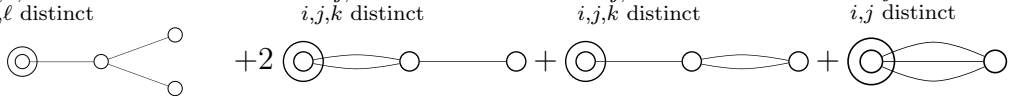
where the second equality uses the assumption that $A$ has zero diagonal. Now to apply the square function componentwise, we can decompose:

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

$$\equiv$$  $$+$$ 

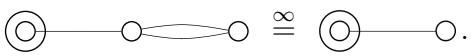
Moving on, we apply $A$ to this representation by casing on whether the new index $i$ matches one of the previous indices. We group terms together using the symmetry of $A$ and the fact that $A_{ii} = 0$.

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

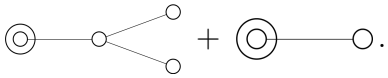
$$\equiv$$ 

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

In the limit $n \to \infty$, only some of these terms contribute to the *asymptotic* Fourier diagram basis representation. Asymptotically, *hanging* double edges can be removed from a diagram[6], so that the third diagram in the representation above satisfies, as $n \to \infty$,



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

$$A(A\vec{1})^2 \overset{\infty}{\equiv}$$ 

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

## 3.2 Properties of the diagram basis

**Definition 3.1.** *A Fourier diagram is an unlabeled undirected multigraph $\alpha = (V(\alpha), E(\alpha))$ with a special vertex labeled ◎ which we call the root. No vertices may be isolated except for the root. We let $\mathcal{A}$ be the set of all Fourier diagrams.*

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

**Definition 3.2** ($Z_\alpha$)**.** *For a Fourier diagram $\alpha \in \mathcal{A}$ with root $\circledcirc$, define the vector $Z_\alpha \in \mathbb{R}^n$ by*

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

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

**Definition 3.3** ($\mathcal{S}$ and $\mathcal{T}$)**.** *Let $\mathcal{S}$ be the set of unlabeled rooted trees such that the root has exactly one subtree (i.e. the root has degree 1). Let $\mathcal{T}$ be the set of all unlabeled rooted trees (non-empty, but allowing the singleton).*

**Definition 3.4** (Proper Fourier diagram)**.** *A proper Fourier diagram is a Fourier diagram with no multiedges or self-loops (i.e. a rooted simple graph).*

For *proper* Fourier diagrams $\alpha \in \mathcal{A}$, the following properties of $Z_\alpha$ hold non-asymptotically i.e. for arbitrary $n$:

(i) $Z_\alpha$ is a multilinear polynomial in the entries of $A$ with degree $|E(\alpha)|$ (or $Z_\alpha = 0$ when $|V(\alpha)| > n$).

(ii) $Z_\alpha$ has the symmetry that $Z_{\alpha,i}(A) = Z_{\alpha,\pi(i)}(\pi(A))$ for all permutations $\pi \in S_n$, where $\pi$ acts on $A$ by permuting the rows and columns simultaneously.

(iii) For each $i \in [n]$, the set $\{Z_{\alpha,i} : \text{proper Fourier diagram } \alpha \in \mathcal{A}\}$ is orthogonal with respect to the expectation over $A$.

(iv) In fact, $Z_\alpha$ is a symmetrized multilinear Fourier character (see Section 3.4). This implies the previous properties and it shows that the proper diagrams are an orthogonal basis for a class of symmetric functions of $A$.

We represent the algorithmic state as a Fourier diagram expression of the form $x = \sum_{\alpha\in\mathcal{A}} c_\alpha Z_\alpha$. To multiply together or apply algorithmic operations on a diagram expression, we case on which indices repeat, like in the example in Section 3.1. See Lemmas A.4 and A.7 in Appendix A.2 for a formal derivation of these rules.

Now we turn to the asymptotic properties. The constant-size tree diagrams $(Z_\tau)_{\tau\in\mathcal{T}}$ exhibit the following key properties in the limit $n \to \infty$ and with respect to the randomness of $A$.

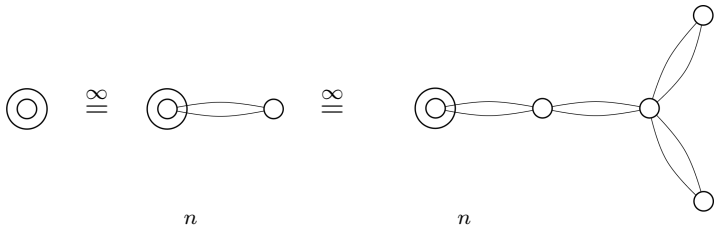
(i) The coordinates of $Z_\tau \in \mathbb{R}^n$ for any $\tau \in \mathcal{T}$ are asymptotically independent and identically distributed.

(ii) The random variables $Z_{\sigma,1}$ for $\sigma \in \mathcal{S}$ (the tree diagrams with one subtree) are asymptotically independent Gaussians with variance $|\mathrm{Aut}(\sigma)|$, where $\mathrm{Aut}(\sigma)$ are the graph automorphisms of $\sigma$ which fix the root.

(iii) The random variable $Z_{\tau,1}$ for $\tau \in \mathcal{T}$ (the tree diagrams with multiple subtrees) is asymptotically equal to the multivariate Hermite polynomial $\prod_{\sigma\in\mathcal{S}} h_{d_\sigma}(Z_{\sigma,1}; |\mathrm{Aut}(\sigma)|)$ where $d_\sigma$ is the number of children of the root whose subtree (including the root) equals $\sigma \in \mathcal{S}$.

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

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



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

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

(vi) The disconnected diagrams have only a minor and negligible role in the algorithms that we consider. See Section 4.2 for the description of these random variables.

To summarize the properties, given a sum $x$ of connected diagrams, by removing the hanging double trees, and then removing all diagrams not in $\mathcal{T}$, the expression admits an *asymptotic* Fourier diagram basis representation of the form

$$x \overset{\infty}{=} \sum_{\tau\in\mathcal{T}} c_\tau Z_\tau \,, \tag{3}$$

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

## 3.3 Asymptotic state evolution

The main appeal of the tree approximation in Eq. (3) is that when restricted to the tree-shaped diagrams, the GFOM operations have a very simple interpretation: they implement an idealized *Gaussian dynamics* which we describe now.

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

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

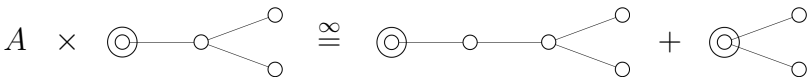
With that prologue, the algorithmic operations restricted to the trees and the corresponding evolution of the asymptotic state $X_t$ are as follows. Two important operations on a tree-shaped diagram are extending/contracting the root by one edge.

**Definition 3.5** (+ and − operators)**.** *We define* $+ : \mathcal{T} \to \mathcal{S}$ *and* $- : \mathcal{S} \to \mathcal{T}$ *by:*

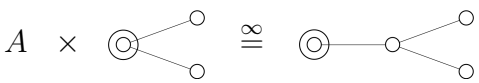
- *If* $\tau \in \mathcal{T}$*, let* $\tau^+$ *be the diagram obtained by extending the root by one edge (i.e. adding one new vertex and one edge connecting it to the root of* $\tau$*, and re-rooting* $\tau^+$ *at this new vertex).*
- *If* $\tau \in \mathcal{S}$*, let* $\tau^-$ *be the diagram obtained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

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

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



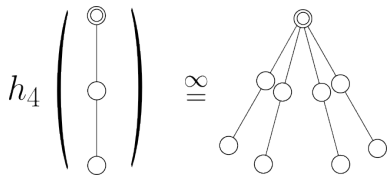
  If $\tau \in \mathcal{T} \setminus \mathcal{S}$, then $AZ_\tau$ is asymptotically only the $\tau^+$ term. For example,



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

$$\prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(Z_\sigma; |\mathrm{Aut}(\sigma)|) \overset{\infty}{=} Z_\tau \,. \tag{4}$$

  Therefore, to compute $f(Z_\sigma : \sigma \in \mathcal{S})$, we expand $f$ in the Hermite polynomial basis associated to $\mathcal{S}$, and apply the rule Eq. (4) to all the terms. For example,



These operations correspond to the following Gaussian dynamic.

**Definition 3.6** (Asymptotic Gaussian space, $\Omega$)**.** *Let $(Z_\sigma^\infty)_{\sigma\in\mathcal{S}}$ be a set of independent centered (one-dimensional) Gaussian random variables with variances* $\mathrm{Var}(Z_\sigma^\infty) = |\mathrm{Aut}(\sigma)|$.

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

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

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

**Definition 3.7** (Asymptotic state)**.** *If $x \in \mathbb{R}^n$ is such that $x \overset{\infty}{=} \sum_{\tau\in\mathcal{T}} c_\tau Z_\tau$, we define the* asymptotic state *of $x$ by*

$$X = \sum_{\tau\in\mathcal{T}} c_\tau Z_\tau^\infty\,.$$

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

- If $x_t$ has asymptotic state $X_t$, then the asymptotic state of $Ax_t$ is $X_t^+ + X_t^-$. Here we extend the $+$ and $-$ operators linearly to sums of $Z_\tau$ or $Z_\tau^\infty$ (let $Z_\tau^- = (Z_\tau^\infty)^- = 0$ if $\tau \in \mathcal{T} \setminus \mathcal{S}$).
- If $x_{t-1}, \dots, x_0$ have asymptotic states $X_{t-1}, \dots, X_0$ and $f$ is any polynomial, then the asymptotic state of $f(x_{t-1}, \dots, x_0)$ is $f(X_{t-1}, \dots, X_0)$.

## 3.4 Perspective: equivariant Fourier analysis

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

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

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

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

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

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

Thus by construction, the diagrams are an orthogonal basis for symmetric low-degree polynomials of $A$. We use this to derive some simple facts in Appendix A.1. Asymptotic independence of the Gaussian diagrams can be predicted based on the fact that the diagrams are an *orthogonal* basis, and orthogonal Gaussians are independent (thus we expect a set of independent Gaussians to appear from other types of i.i.d. inputs as well).

The above discussion was for Boolean matrices with $A_{ij} \sim \{\pm 1\}$. The natural generalization expresses polynomials in the basis of orthogonal polynomials for the entries $A_{ij}$ (e.g. the Hermite polynomials when the $A_{ij} \sim \mathcal{N}(0, 1/n)$ [MW22a, Section 3.2]).

Our results show that for the first-order algorithms we consider, only the multilinear part of the basis matters (i.e. the orthogonal polynomials which are degree 0 or 1 in each variable): up to negligible error, we can approximate $A_{ij}^2 \approx \frac{1}{n}$ and $A_{ij}^k \approx 0$ for $k \geq 3$. We use the monomial basis[7] to represent higher-degree polynomials instead of higher-degree orthogonal polynomials in order to simplify the presentation (except for the degree-2 orthogonal polynomial $A_{ij}^2 - \frac{1}{n}$ which expresses some error terms).

# 4 Diagram Analysis of $O(1)$ Iterations

In this section we develop tools for rigorously analyzing diagrams of constant size, corresponding to first-order algorithms with constantly many iterations. These proofs make formal the intuitive ideas developed in Section 3. Longer proofs in this section are delayed to Appendix B for readability.

- In Section 4.1, we give a rigorous definition of the asymptotic equality $\overset{\infty}{=}$.

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

- In Section 4.2, we prove the classification of the asymptotic behavior of the constant-size diagrams.
- In Section 4.3, we prove the tree approximation for the class of GFOM algorithms.
- In Section 4.4, we prove a general state evolution formula for GFOM algorithms.

## 4.1 Equality up to combinatorially negligible diagrams

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

**Definition 4.1** ($I(\alpha)$)**.** *For a diagram $\alpha \in \mathcal{A}$, let $I(\alpha)$ be the subset of non-root vertices such that every edge incident to that vertex has multiplicity $\geq 2$ or is a self-loop.*

**Lemma 4.2.** *Let $q \in \mathbb{N}$ be a constant independent of $n$ and $\alpha \in \mathcal{A}$ be a constant-size diagram. Then for $i \in [n]$,*

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

A similar norm bound for matrices is a crucial ingredient in Fourier analysis of matrix-valued functions [AMP20]. The proof of Lemma 4.2 is in Appendix B.2.

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

**Definition 4.3** (Combinatorially negligible and order 1)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real-valued coefficients such that $a_n = \Theta(n^{-k})$ for some $k \geq 0$ with $2k \in \mathbb{Z}$. Let $\alpha \in \mathcal{A}$ be a constant-size diagram.*

1. *We say that $a_n Z_\alpha$ is combinatorially negligible if*

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

   *For $a_n = 0$, we also say that $a_n Z_\alpha$ is combinatorially negligible.*

2. *We say that $a_n Z_\alpha$ has combinatorial order 1 if*

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

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

**Definition 4.4** ($\overset{\infty}{=}$). *We say that $x \overset{\infty}{=} y$ if there exists real coefficients $(c_\alpha)_{\alpha \in \mathcal{A}}$ depending on $n$ and supported on diagrams of constant size such that*

$$x - y = \sum_{\alpha \in \mathcal{A}} c_\alpha Z_\alpha \,,$$

*where $c_\alpha Z_\alpha$ is combinatorially negligible for all $\alpha \in \mathcal{A}$.*

Later, we will prove results of the form $x \overset{\infty}{=} \widehat{x}$ where $x$ is the state of an algorithm and $\widehat{x}$ is some asymptotic approximation of $x$. In order to interpret these results, we note that $\overset{\infty}{=}$ implies very strong forms of convergence of the error to 0. The proof of the following lemma can be found in Appendix B.2.

**Lemma 4.5.** *Suppose that $A = A(n)$ is a sequence of random matrices satisfying Assumption 2.1. If $x$ and $y$ are diagram expressions such that $x \overset{\infty}{=} y$, then $\|x - y\|_\infty = \widetilde{O}(n^{-1/2})$ with high probability.*

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

**Lemma 4.6.** *If $x, y$ are diagram expressions with $x \overset{\infty}{=} y$, then*

$$Ax \overset{\infty}{=} Ay \,.$$

*Moreover, if $x_1, \ldots, x_t, y_1, \ldots, y_t$ are diagram expressions with $x_i \overset{\infty}{=} y_i$ for all $i \in [t]$, then*

$$f(x_1, \ldots, x_t) \overset{\infty}{=} f(y_1, \ldots, y_t) \,,$$

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

The proof of Lemma 4.6 is in Appendix B.2. The intuitive view of this lemma is that a diagram with a cycle still has the cycle after the algorithmic operations and thus remains negligible. The proof in Appendix B.2 is a syntactic version.

We can also show combinatorially that the error of removing a hanging double edge from any diagram is negligible. The proof proceeds by extending the definition of diagrams to allow new types of residual edges that are only used in the analysis (see Appendix B.1).

**Lemma 4.7.** *Let $a_n Z_\alpha$ be a term of combinatorial order at most 1 such that $\alpha$ has a hanging double edge. Let $\alpha_0$ be $\alpha$ with the hanging double edge and hanging vertex removed. Then*

$$a_n Z_\alpha \overset{\infty}{=} a_n Z_{\alpha_0} \,.$$

## 4.2 Classification of constant-size diagrams

We classify the asymptotic limits of constant-size diagrams and prove that all of their constant-order joint moments are within $O(n^{-1/2})$ of the asymptotic limit. In addition to the vector Fourier diagrams from Definition 3.1, we will classify *scalar Fourier diagrams*, which are simply unlabeled undirected multigraphs (the only difference with vector diagrams being that they do not have a root). The notation for scalar diagrams is analogous.

**Definition 4.8** (Scalar Fourier diagrams)**.** *Let $\mathcal{A}_{\text{scalar}}$ be the set of all unlabeled undirected multigraphs with no isolated vertices. Let $\mathcal{T}_{\text{scalar}}$ be the set of non-empty unlabeled trees.*

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

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

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

**Definition 4.9** ($\mathcal{F}_{\text{scalar}}$ and $\mathcal{F}$)**.** *Let $\mathcal{F}_{\text{scalar}}$ be the set of unlabeled forests with no isolated vertices. Let $\mathcal{F}$ be the set of unlabeled forests such that one vertex is the special root vertex ◎. No vertices may be isolated except for the root.*

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

We classify the diagrams in $\mathcal{A}$ and $\mathcal{A}_{\text{scalar}}$. First, the next lemma identifies which of the diagrams are non-negligible. This lemma is for *connected* vector diagrams; scalar diagrams and disconnected vector diagrams have a similar characterization in Lemma B.12.

**Lemma 4.10.** *Let $\alpha \in \mathcal{A}$ be a connected Fourier diagram. Then $Z_\alpha$ is either combinatorially negligible or combinatorially order 1. Moreover, it is combinatorially order 1 if and only if the following four conditions hold simultaneously:*

*(i) Every multiedge has multiplicity 1 or 2.*

*(ii) There are no cycles.*

*(iii) The subgraph of multiplicity 1 edges is connected and contains the root if it is nonempty (i.e. the multiplicity 2 edges consist of hanging trees).*

*(iv) There are no self-loops or 2-labeled edges (Appendix B.1).*

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

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

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

As a result, the non-negligible connected diagrams in $\mathcal{A}$ are asymptotically equal to trees in $\mathcal{T}$ after using Lemma 4.7 to remove the hanging double edges (disconnected diagrams $\alpha \in \mathcal{A}$ and scalar diagrams $\alpha \in \mathcal{A}_{\text{scalar}}$ are likewise asymptotically equal to a forest in $\mathcal{F}$ or $\mathcal{F}_{\text{scalar}}$).

The next Theorem 4.11 completes the classification by showing that the non-negligible diagrams in $\mathcal{T}, \mathcal{F}$, and $\mathcal{F}_{\text{scalar}}$ are asymptotically Gaussians and Hermite polynomials. The proof is in Appendix B.4. Also see Theorem B.18 for a version of the theorem in terms of moments.

**Theorem 4.11** (Classification)**.** *Suppose that $A = A(n)$ is a sequence of random matrices satisfying Assumption 2.1.*

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

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

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

  *where $d_\tau$ is the number of copies of $\tau$ in $\rho$.*

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

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

*Moreover, the random variables*

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

*are asymptotically independent (Definition 4.12).*

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

**Definition 4.12** (Asymptotic independence)**.** *A family of real-valued random variables $(X_{n,i})_{n \in \mathbb{N}, i \in \mathcal{I}_n}$ is asymptotically independent if:*

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

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

## 4.3 Tree approximation of GFOMs

The class of general first-order methods is defined as follows.

**Definition 4.13** (General first-order method)**.** *The input is a matrix $A \in \mathbb{R}^{n\times n}$. The state of the algorithm at time $t$ is a vector $x_t \in \mathbb{R}^n$. Initially, $x_0 = \vec{1}$. At each time $t$, we can execute one of the following two operations:*

1. *Multiply by $A$ ($x_{t+1} = Ax_t$).*
2. *Apply coordinatewise a polynomial[8] function independent of $n$, $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ to $(x_t, x_{t-1}, \dots, x_0)$ (for all $i \in [n]$, $x_{t+1,i} = f_t(x_{t,i}, \dots, x_{0,i})$).*

Inductively following the rules given explicitly in Appendix A.2, we may represent the algorithmic state $x_t$ of a GFOM in the diagram basis. Define the *tree approximation* $\widehat{x}_t$ to be the analogous diagram expression obtained by performing the algorithmic operations on only the tree diagrams, removing hanging double edges and removing the cyclic diagrams that appear (see Definition A.8 for the formal definition).

**Theorem 4.14** (Tree approximation of GFOMs)**.** *Let $t \geq 0$ be a constant independent of $n$ and $A = A(n)$ be a sequence of random matrices satisfying Assumption 2.1. Let $x_t \in \mathbb{R}^n$ be the state of a GFOM and let $\widehat{x}_t$ be its tree approximation. Then $x_t \overset{\infty}{=} \widehat{x}_t$. In particular,*

$$\|x_t - \widehat{x}_t\|_\infty = \widetilde{O}(n^{-1/2}) \text{ with high probability}. \tag{5}$$

*Proof.* We can prove $x_t \overset{\infty}{=} \widehat{x}_t$ inductively. By Lemma 4.6, each of the combinatorially negligible diagrams in $x_t$ remains combinatorially negligible at time $t+1$. Meanwhile, the combinatorially non-negligible tree diagrams in $\widehat{x}_t$ get updated to $\widehat{x}_{t+1}$. The error bound Eq. (5) follows from Lemma 4.5. □

**Remark 4.15.** *The leading order guarantee of Theorem 4.14 is best possible in general (up to logarithmic factors). Similar but more complicated equations can be given for the lower-order error terms in Eq. (5). For example, since the other connected diagrams with $E$ edges and $V$ vertices have magnitude $n^{(V-1-E)/2}$, the first lower-order term of order $n^{-1/2}$ consists of connected diagrams with exactly one cycle. The GFOM operations on this set of diagrams describe how the error evolves at this order.*

**Remark 4.16.** *One technical caveat of our analysis is that many nonlinearities used in applications are not polynomial functions (e.g.* ReLU, tanh*). We note that existing polynomial approximation arguments in the literature (see for example [MW22a, IS24]) should apply here to prove that the tree approximation holds for GFOMs with Lipschitz denoisers $f_t$ up to arbitrarily small $\frac{1}{\sqrt{n}}\|\cdot\|_2$ error. This is however strictly weaker than the guarantees of Theorem 4.14.*

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

## 4.4 General state evolution

From the ideas established so far, we directly deduce *state evolution* for GFOM algorithms, capturing that the coordinates of $x_t$ are asymptotically independent trajectories of an explicit random variable $X_t$. Recall the definition of the asymptotic state $X_t$ from Definition 3.7.

To state the theorem, asymptotic independence is extended from Definition 4.12 to $\mathbb{R}^{t+1}$-valued random variables in the natural way.

**Definition 4.17** (Asymptotic independence of trajectories)**.** *A family of random variables $(X_{n,i})_{n\in\mathbb{N},i\in\mathcal{I}_n}$ taking values in $\mathbb{R}^{t+1}$ is asymptotically independent if:*

$$\forall q \in \mathbb{N}.\ \exists \varepsilon = \varepsilon(q) \underset{n\to\infty}{\longrightarrow} 0.\ \forall k \in \mathbb{N}^{\mathcal{I}_n \times [t+1]} : \sum_{\substack{i\in\mathcal{I}_n \\ j\in[t+1]}} k_{ij} = q.$$

$$\left| \mathbb{E}\left[ \prod_{i\in\mathcal{I}_n, j\in[t+1]} X_{n,i,j}^{k_{ij}} \right] - \prod_{i\in\mathcal{I}_n} \mathbb{E}\left[ \prod_{j\in[t+1]} X_{n,i,j}^{k_{ij}} \right] \right| \le \varepsilon(q)\,.$$

**Theorem 4.18** (General state evolution)**.** *Let $t$ be a constant and $A = A(n)$ be a sequence of random matrices satisfying Assumption 2.1. Let $x_t \in \mathbb{R}^n$ be the state of a GFOM and let $X_t$ be the asymptotic state of $x_t$. Then:* [9]

*(i) For each $i \in [n]$, $(x_{0,i}, \dots, x_{t,i}) \xrightarrow{\mathrm{d}} (X_0, \dots, X_t)$. Furthermore, the coordinates' trajectories $\{(x_{0,i}, \dots, x_{t,i}) : i \in [n]\}$ are asymptotically independent.*

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

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

*(iii) $X_t$ satisfies the explicit recurrence defined at the end of Section 3.3.*

*Proof.* For (i), by Lemma 2.3, it suffices to check that all of the constant-order joint moments of $x_{t,i}$ converge to the joint moments of $X_t$. This follows from convergence of the moments of every diagram $Z_\alpha$ to those of $Z_\alpha^\infty$ in the diagram classification Theorem 4.11.

Part (ii) will be proven in Appendix B.5 as the following lemma.

**Lemma 4.19.** *Let $x$ be a vector diagram expression with asymptotic state $X \in \Omega$. Then as scalar diagrams, $\frac{1}{n}\sum_{i=1}^n x_i \overset{\infty}{=} \mathbb{E}[X]$.*

For (iii), the tree approximation $x_t = \widehat{x}_t$ holds by Theorem 4.14. The asymptotic state $X_t$ corresponding to $\widehat{x}_t$ then satisfies the explicit recursion on trees presented in Section 3.3. □

[9] It is natural to wonder whether (ii) can be derived as a consequence of (i) in Theorem 4.18. The answer is a resounding no. Even given good control over the distribution of individual coordinates $x_{t,i}$ it is crucial to ensure that errors do not correlate adversarially when summed. This is precisely the kind of heuristic assumption made when using the cavity method.

We conclude this section by working out a few lemmas which help compute asymptotic states. We will use them in Section 5.4 to compute the state evolution of approximate message passing.

The set of asymptotic states $\Omega$ has an inner product coming from the expectation over the Gaussians $(Z_\sigma^\infty)_{\sigma\in\mathcal{S}}$. Since these random variables are independent Gaussians, the multivariate Hermite polynomials $(Z_\tau^\infty)_{\tau\in\mathcal{T}}$ form an orthogonal basis of $\Omega$ with respect to this inner product. Recall the $+$ and $-$ operators from Definition 3.5.

**Fact 4.20.** *$+$ and $-$ are bijections between $\mathcal{T}$ and $\mathcal{S}$ which are inverses of each other and preserve $|\mathrm{Aut}(\tau)|$.*

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

**Fact 4.21.** *For all $X \in \Omega$, $(X^+)^- = X$ and $(X^-)^+$ is the orthogonal projection of $X$ to the subspace spanned by $\mathcal{S}$.*

We deduce that $+$ and $-$ are adjoint operators on $\Omega$:

**Lemma 4.22.** *For all $X, Y \in \Omega$, $\mathbb{E}[X^+Y] = \mathbb{E}[XY^-]$.*

*Proof.* Since $(Z_\tau^\infty)_{\tau\in\mathcal{T}}$ is a basis of the vector space $\Omega$, it suffices to check this for each pair of basis elements $\tau, \rho \in \mathcal{T}$. By orthogonality, $\mathbb{E}\left[Z_{\tau^+}^\infty Z_\rho^\infty\right]$ is nonzero if and only if $\tau^+ = \rho$ and in this case it takes value $|\mathrm{Aut}(\tau^+)|$. By Fact 4.20, this occurs if and only if $\rho \in \mathcal{S}$ and $\tau = \rho^-$. Moreover, in this case the value is also $|\mathrm{Aut}(\tau^+)| = |\mathrm{Aut}(\tau)|$, as needed. □

**Lemma 4.23.** *For all $X, Y \in \Omega$, $\mathbb{E}[XY] = \mathbb{E}[X^+Y^+]$ and $\mathbb{E}[(X^-)^2] \le \mathbb{E}[X^2]$.*

*Proof.* For the first statement, apply Lemma 4.22 on $X$ and $Y^+$, then use Fact 4.21. For the second statement, apply Lemma 4.22 on $X^-$ and $X$ to get $\mathbb{E}[(X^-)^+X] = \mathbb{E}[(X^-)^2]$. Since $(X^-)^+$ projects away some terms from $X$ by Fact 4.21, the left-hand side is upper bounded by $\mathbb{E}[X^2]$. □

# 5 Belief Propagation, AMP, and the Cavity Method

With the tree approximation in hand, we describe how to use it to implement cavity method reasoning about nonlinear iterative algorithms.

- In Section 5.1, we give background on the cavity method and how it can be used to predict the asymptotic trajectory of message-passing algorithms.
- In Section 5.2, we prove the asymptotic equivalence of BP and AMP on mean-field models (Theorem 5.1). The proof precisely follows the structure described earlier: we reproduce a folklore physics argument in Section 5.2.1 and make it directly rigorous in Section 5.2.2.

- In Section 5.3, we use the same technology to prove a fundamental assumption of the cavity method: the asymptotic independence of messages incoming at a vertex (Theorem 5.9).
- In Section 5.4, we give a new proof of the state evolution formula for BP/AMP (Theorem 5.10).
- In Section 5.5, we reinterpret Montanari's algorithm for optimizing spin glass Hamiltonians through the lens of the asymptotic tree space.

## 5.1 Background on the cavity method

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

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

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

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

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

$$m_{i\to j} = f\left(\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m_{k\to i}\right), \quad m_i \approx f\left(\sum_{k=1}^n A_{ik} m_{k\to i}\right), \tag{6}$$

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

$$m_{i\to j}^{t+1} = f\left(\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m_{k\to i}^t\right), \quad m_i^{t+1} = f\left(\sum_{k=1}^n A_{ik} m_{k\to i}^t\right), \tag{7}$$

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

The physical techniques predict the asymptotic trajectory of the messages $m_{i\to j}^t$ and the outputs $m_i^t$ in Eq. (7) with respect to the randomness of the matrix $A$. They say that $m^t$ will have approximately independent and identically distributed entries,

$$m_i^t \sim f(Z_t)\,, \qquad \text{where } Z_t \sim \mathcal{N}(0, \sigma_t^2)\,, \quad \sigma_1^2 = 1\,, \quad \sigma_{t+1}^2 = \mathbb{E}\, f(Z_t)^2\,. \tag{8}$$

A heuristic replica symmetric cavity approach for proving Eq. (8) would go as follows. We make an **independence assumption** that the incoming terms $m_{k\to i}^t$ in the non-backtracking summation $\sum_{k=1,k\neq j}^n A_{ik} m_{k\to i}^t$ of Eq. (7) are independent, as if the messages were coming up from disjoint branches of a tree. By symmetry, the messages are identically distributed. Then, we appeal to the central limit theorem to deduce

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

From here, we get that the outgoing message satisfies $m_{i\to j}^t \sim f(Z_t)$ for $Z_t \sim \mathcal{N}(0, \sigma_t^2)$ with $\sigma_t^2$ defined by the recurrence in Eq. (8). Using a similar argument, we get $m_i^t \sim f(Z_t)$.

[MPV87] also derived from Eq. (6) a simpler form of self-consistent equations involving only the marginals themselves, known as the Thouless–Anderson–Palmer equations [TAP77],

$$m_i \approx f\left(\sum_{\substack{k=1\\k\neq i}}^n A_{ik} m_k - \beta\left(1 - \frac{1}{n}\sum_{k=1}^n m_k^2\right) m_i\right)\,. \tag{9}$$

The subtracted term on the right-hand side in which $m_i$ re-occurs is the *Onsager reaction term*. In the same way that belief propagation Eq. (7) tries to compute solutions to the cavity equations Eq. (6), an *approximate message passing* algorithm can be iterated to compute approximate solutions to Eq. (9),

$$m_i^{t+1} = f\left(\sum_{\substack{k=1\\k\neq i}}^n A_{ik} m_k^t - \beta\left(1 - \frac{1}{n}\sum_{k=1}^n (m_k^t)^2\right) m_i^{t-1}\right)\,. \tag{10}$$

The approximate equivalence between the BP iteration Eq. (7) and the AMP iteration Eq. (10) is a folklore cavity method argument which we elaborate next.

## 5.2 Equivalence between message-passing iterations

**Belief propagation.** We consider BP iterations on $A$ of the form

$$m_{i\to j}^0 = 1\,, \qquad m_{i\to j}^t = f_t\left(\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m_{k\to i}^{t-1}, \dots, \sum_{\substack{k=1\\k\neq j}}^n A_{ik} m_{k\to i}^0,\ m_{i\to j}^0\right)\,, \tag{11}$$
$$m_i^t = \widetilde{f}_t\left(\sum_{k=1}^n A_{ik} m_{k\to i}^{t-1}, \dots, \sum_{k=1}^n A_{ik} m_{k\to i}^0,\ m_{i\to j}^0\right)\,,$$

for a sequence of functions $f_t, \widetilde{f}_t : \mathbb{R}^{t+1} \to \mathbb{R}$. Eq. (11) is a generalization of Eq. (7) to iterations "with memory" i.e. that can use all the previous messages. At any timestep $t$, the $(m_{i\to j}^t)_{1\le i,j\le n}$ are *cavity messages* that try to compute some information about the $i$-th variable by ignoring the edge between $i$ and $j$, while the $(m_i^t)_{1\le i\le n}$ are the output of the algorithm.

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

$$w^0 = \vec{1}\,, \qquad w^{t+1} = A f_t(w^t, \dots, w^0) - \sum_{s=1}^t b_{s,t} f_{s-1}(w^{s-1}, \dots, w^0)\,, \tag{12}$$
$$m^t = \widetilde{f}_t(w^t, \dots, w^0)\,, \tag{13}$$

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

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

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

**Theorem 5.1** (Equivalence of BP and AMP)**.** *Let $T \ge 1$ be a constant independent of $n$, $f_t, \widetilde{f}_t : \mathbb{R}^{t+1} \to \mathbb{R}$ for $t \le T$ be a sequence of polynomials independent of $n$, and $A = A(n)$ be a sequence of random matrices satisfying Assumption 2.1. Generate $m^{t,\mathrm{BP}}$ according to Eq. (11) and $m^{t,\mathrm{AMP}}$ according to Eq. (13). Then*

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

*so in particular, with high probability,*

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

### 5.2.1 Heuristic derivation of Theorem 5.1

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

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

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

Hence it suffices to show that $w^t$ asymptotically follows the AMP iteration Eq. (12). First, Eq. (11) can be rewritten

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

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

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

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

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

This shows that $w_i^{t+1}$ approximately satisfies the AMP recursion Eq. (12), as desired.

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

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

### 5.2.2 Diagram proof of Theorem 5.1

In fact, the previous heuristic argument can be made directly rigorous by working with the tree approximation. It suffices to justify Eq. $(*)$ and Eq. $(**)$ in order to prove Theorem 5.1.

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

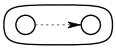
**Definition 5.2** (Cavity diagrams)**.** *A cavity diagram is an unlabeled undirected graph $\alpha = (V(\alpha), E(\alpha))$ with two distinct, ordered root vertices*  *. No vertices may be isolated except for the roots.*

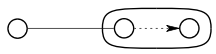
*For any cavity diagram $\alpha$, we define $Z_\alpha \in \mathbb{R}^{n(n-1)}$ by*

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

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

Below is the representation of the first iterate of Eq. (11) with cavity diagrams. In the pictures, we draw an arrow from the first root to the second root to indicate the order. If a (multi)edge exists in the graph between the roots, then the arrow is on the edge; otherwise we use a dashed line to indicate that there is no edge.

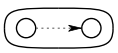
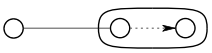
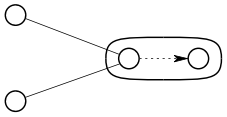
$$m^0_{i\to j} = \text{  }$$

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

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

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

To apply $f_1$, we need to multiply $i \to j$ diagrams componentwise, which is achieved by fixing/merging the roots $i, j$ and summing over the part outside the roots. For some coefficients $c_0, c_1, c_2, \ldots$ we have[10]

$$m^1_{i\to j} = f_1\left(\sum_{\substack{k=1\\k\neq j}}^{n} A_{ik} m^0_{k\to i}\right) = c_0 \text{  } + c_1 \text{  } + c_2 \text{  } + \cdots$$

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

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

**Claim 5.3** (Conversion of cavity diagrams)**.** *For any cavity diagram $\alpha$ and $i \in [n]$,*

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

*where $\alpha'$ is the diagram (in the sense of Definition 3.1) obtained from $\alpha$ by adding an edge between the two roots of $\alpha$ and unrooting the first root.*

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

**Definition 5.4.** *A cavity diagram $\alpha$ is combinatorially negligible if the diagram $\alpha'$ obtained in Claim 5.3 is combinatorially negligible. We naturally extend the $\overset{\infty}{=}$ notation to cavity diagrams as in Definition 4.4.*

**Claim 5.5.** *Let $x$ and $x'$ be in the span of the cavity diagrams such that $x \overset{\infty}{=} x'$. If we let*

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

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

*If $x_1, \dots, x_t, x'_1, \dots, x'_t$ are in the span of cavity diagrams, $x_i \overset{\infty}{=} x'_i$ for all $i \in [n]$, and $f : \mathbb{R}^t \to \mathbb{R}$ is a polynomial function applied componentwise, then*

$$f(x_1, \dots, x_t) \overset{\infty}{=} f(x'_1, \dots, x'_t)\,.$$

Claim 5.5 follows directly from Lemma 4.6.

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

**Lemma 5.6** (Eq. (∗))**.**

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

*Proof.* Since $f_t$ is a polynomial, it has an exact Taylor expansion. The terms of degree higher than 1 in the Taylor expansion create at least 2 edges between the roots $i$ and $j$. All cavity diagrams with 2 edges between the roots are combinatorially negligible because the unrooting operation of Claim 5.3 adds one more edge between $i$ and $j$, and diagrams with multiedges of multiplicity $> 2$ are combinatorially negligible (Lemma 4.10). □

**Lemma 5.7** (Eq. (∗∗))**.**

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

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

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

The difference between this and $f_{s-1}(w_i^{s-1},\ldots,w_i^0)$ are the backtracking terms $A_{ik}m_{k\to i}^r$. All terms in the entire Taylor expansion of the polynomial on the right-hand side around $w_i^{s-1},\ldots,w_i^0$ will introduce at least one additional factor of $A_{ik}$, which combines with the $A_{ik}^2$ present in the summation over $k$ to become a negligible multiplicity $> 2$ edge (Lemma 4.10). This shows that

$$\sum_{k=1}^{n} A_{ik}^2 m_{i\to k}^{s-1} \frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_k^0) \ \overset{\infty}{=}\ f_{s-1}(w_i^{s-1},\ldots,w_i^0) \sum_{k=1}^{n} A_{ik}^2 \frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_k^0). \tag{14}$$

Second, we argue about the replacement of $A_{ik}^2$. This double edge is only non-negligible if it is hanging (Lemma 4.10). Among the diagrams in $\frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_0^t)$ the only one which does not attach something to $k$ is the singleton diagram ◎. The coefficient of this diagram is the expected value (Corollary A.3),

$$\mathbb{E}\left[\frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_k^0)\right].$$

The expected value is equal to the empirical expectation up to negligible terms (Lemma 4.19),

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

This implies

$$\sum_{k=1}^{n} A_{ik}^2 \frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_k^0) \ \overset{\infty}{=}\ \frac{1}{n}\sum_{k=1}^{n} \frac{\partial f_t}{\partial w^s}(w_k^t,\ldots,w_k^0). \tag{15}$$

The desired statement follows from combining Eq. (14) and Eq. (15). □

*Proof of Theorem 5.1.* Replace the $\approx$ signs in the heuristic argument from Section 5.2.1 by $\overset{\infty}{=}$ and use Claim 5.5 repeatedly. □

## 5.3 Proving the cavity assumptions

We examine the belief propagation iteration Eq. (11) more closely. The BP iterates have the following asymptotic structure.

**Lemma 5.8.** *$m_{i\to j}^t$ is asymptotically equivalent to a linear combination of cavity diagrams which have a tree hanging off of $i$, no edges between the roots, and nothing attached to $j$.*

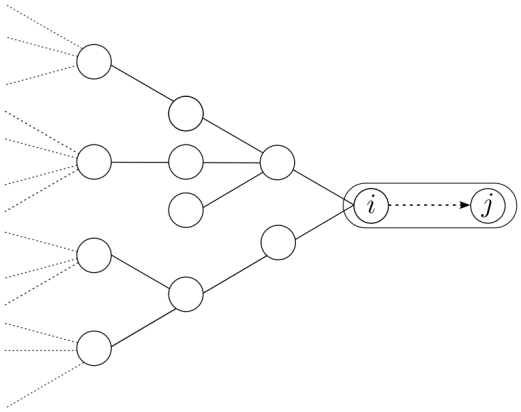



Figure 1: Diagram representation of the cavity messages $m^t_{i\to j}$. Each cavity diagram in the asymptotic cavity diagram representation of $m^t_{i\to j}$ is a tree rooted at $i$.

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

All diagrams of type (3) are negligible (and stay so when applying further operations to them), because they create a cycle of length $> 2$. The backtracking step in (2) is canceled by summing over $k \neq j$ in the belief propagation iteration. What asymptotically remains is the forward step (1) which again has the stated form. Additionally, componentwise functions preserve the stated form. □

**Theorem 5.9.** *For any $j \in [n]$, the incoming messages at $j$, $\{m^t_{i\to j} : i \in [n] \setminus \{j\}\}$, are asymptotically independent (Definition 4.12).*

*Proof.* When $j$ is ignored, the cavity diagrams in the asymptotic representation of $m^t_{i\to j}$ in Lemma 5.8 are equivalent to non-cavity diagrams (replacing $n$ by $n-1$). From the classification theorem (Theorem 4.11), these are asymptotically independent. □

## 5.4 State evolution formula for BP/AMP

We show how to simplify the asymptotic state appearing in Theorem 4.18 for the special case of approximate message passing. Recall the $+$ and $-$ operators from Section 3.3.

**Theorem 5.10** (Asymptotic state for AMP)**.** *Under the same assumptions as Theorem 5.1, the asymptotic state of $(w_t)_{t\le T}$ satisfies the recursion*

$$W_0 = 1\,, \qquad W_{t+1} = f_t(W_t, \dots, W_0)^+\,. \tag{16}$$

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

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

Combining Theorem 5.10 and part (ii) of Theorem 4.18 recovers the typical formulation of state evolution for AMP algorithms. We note that while the formula for computing iterates of AMP (Eq. (13)) might look mysterious at first sight, the AMP recursion in the asymptotic space (Eq. (16)) is much easier to interpret.

We now prove Theorem 5.10. Note that Eq. (12) is not directly captured by the definition of a GFOM because $b_{s,t}$ requires computing an average over coordinates. This is only a technical issue: by Lemma 4.19, empirical expectations are concentrated up to combinatorially negligible terms. Hence, the following inductive definition of a GFOM for $w_t \in \mathbb{R}^n$ and its corresponding asymptotic state $W_t$ is asymptotically equivalent to Eq. (12):

$$w_0 = \vec{1}\,, \quad w_{t+1} = Af_t(w_t, \dots, w_0) - \sum_{s=1}^{t} \mathbb{E}\left[\frac{\partial f_t}{\partial w_t}(W_t, \dots, W_0)\right] f_{s-1}(w_{s-1}, \dots, w_0)\,. \tag{17}$$

The Onsager correction term in Eq. (17) will be rigorously interpreted as a backtracking term using diagrams.

**Lemma 5.11.** *Let $W_1, \dots, W_t \in \Omega$ be Gaussian (i.e. each $W_s$ is in the span of $(Z_\sigma^\infty)_{\sigma\in\mathcal{S}}$). Then for any polynomial function $f : \mathbb{R}^t \to \mathbb{R}$,*

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

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

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

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

$$\begin{aligned} c_\sigma|\mathrm{Aut}(\sigma)| &= \mathbb{E}\left[Z_\sigma^\infty f(W_1, \dots, W_t)\right] && \text{(orthogonality)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_\sigma^\infty W_s\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] && \text{(Lemma 2.7)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_{\sigma^-}^\infty W_s^-\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right]\,. && \text{(Lemma 4.22)} \end{aligned}$$

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

Now we complete the proof of Theorem 5.10.

*Proof of Theorem 5.10.* We prove by induction on $t$ that $W_{t+1} = f_t(W_t, \dots, W_0)^+$. For $t = 0$, we have $w_1 = Af_0(\vec{1})$ so $W_1 = f_0(W_0)^+$ and the statement holds.

Now suppose that the statement holds for $W_1, \dots, W_t$ for some $t < T$. The asymptotic state of $Af_t(w_t, \dots, w_0)$ is $f_t(W_t, \dots, W_0)^+ + f_t(W_t, \dots, W_0)^-$. By the induction hypothesis and Fact 4.21, for any $s \le t$,

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

Combining this with Lemma 5.11, we see that the asymptotic state of the Onsager correction term equals $f_t(W_t, \dots, W_0)^-$. This concludes the induction.

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

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

where the last equality follows from Lemma 4.23. This completes the proof. □

## 5.5 Montanari's iterative AMP algorithm

A special type of approximate message passing iterations, called *iterative AMP*, was introduced by Montanari to optimize Ising spin glass Hamiltonians [Mon19, AM20, AMS21]. Here we reinterpret iterative AMP and its analysis in the asymptotic space.

The problem considered in [Mon19] is to optimize a degree-2 polynomial with random coefficients over the hypercube (an average-case variant of the Max-Cut-Gain problem), i.e. given $A$ satisfying Assumption 2.1, find $x \in \{-1, 1\}^n$ (approximately) solving

$$\frac{1}{n} \max_{x \in \{-1,1\}^n} \langle x, Ax \rangle = \frac{1}{n} \max_{x \in \{-1,1\}^n} \sum_{i,j=1}^n A_{ij} x_i x_j\,. \tag{18}$$

The value of Eq. (18) is known to concentrate around the constant $2P_* \approx 1.52$ [Tal06, CH06]. Montanari gave an algorithm running in time $n^{O_\varepsilon(1)}$ that, with high probability over $A$ (as $n \to \infty$), finds an assignment $x \in \{-1, 1\}^n$ achieving a $(1 - \varepsilon)$-approximation to Eq. (18). This result is conditional on the widely believed conjecture [Mon19, Assumption 2] that the problem exhibits no overlap gap.

Montanari's algorithm is an AMP iteration with non-polynomial nonlinearities, although Ivkov and Schramm [IS24, Lemma B.4] proved that it can be well-approximated by AMP with polynomial nonlinearities.[11] Iterative AMP [Mon19] uses Eq. (12) with the functions

$$f_t(w_t, \dots, w_0) = w_t \odot u_t(w_{t-1}, \dots, w_0) \tag{19}$$

for chosen functions $u_t : \mathbb{R}^t \to \mathbb{R}$ applied componentwise, where $\odot$ denotes componentwise multiplication. The candidate output of the algorithm is $x_T = \sum_{t=1}^T w_t \odot u_t(w_{t-1}, \dots, w_0) = \sum_{t=1}^T f_t(w_t, \dots, w_0)$.

The special property of iterative AMP is that it sums up *independent* Gaussian vectors $w_t$ scaled componentwise by the functions $u_t$. The independence of the Gaussian vectors $w_t$

[11] The assignment constructed by the latter iteration is not precisely Boolean but it can be rounded to $\{\pm 1\}^n$ with $o(1)$ loss in the objective value.

is contained in the state evolution for AMP as follows. By Theorem 5.10, the asymptotic states $W_t, U_t, X_t$ of $w_t, u_t, x_t$ satisfy $U_0 = W_0 = 1$,

$$U_t = u_t(W_{t-1}, \dots, W_0)\,, \quad W_{t+1} = (U_t W_t)^+\,, \quad X_t = \sum_{s=1}^{t} U_s W_s\,.$$

**Claim 5.12.** *$U_t$ is in the span of trees in $\mathcal{T}$ with depth at most $t-1$ and $W_t$ is in the span of trees in $\mathcal{S}$ with depth exactly $t$.*

*Proof.* Arguing inductively, as componentwise functions do not increase the depth, $U_t$ is in the span of trees from $\mathcal{T}$ of depth at most $t-1$. In the product $U_t W_t$, the trees of depth $t$ in $W_t$ cannot be cancelled by any trees of lower depth from $U_t$. Therefore all trees in $U_t W_t$ and $W_{t+1} = (U_t W_t)^+$ have depth exactly $t$ and $t+1$ respectively, as needed. □

Claim 5.12 provides a very clear explanation of where the independent Gaussians in iterative AMP are coming from: the $W_t$ have different depths, and Gaussian diagrams of different depths are asymptotically independent Gaussian vectors.

**Optimality via state evolution.** The objective value achieved by the iteration can also be computed using state evolution:

$$\begin{aligned}
\frac{1}{n}\langle x_T, Ax_T\rangle &\overset{\infty}{=} \mathbb{E}\left[X_T(X_T^+ + X_T^-)\right] && \text{(Lemma 4.19)}\\
&= 2\,\mathbb{E}\left[X_T X_T^+\right] && \text{(Lemma 4.22)}\\
&= 2\sum_{s,t=1}^{T} \mathbb{E}\left[U_s W_s (U_t W_t)^+\right]\\
&= 2\sum_{s,t=1}^{T} \mathbb{E}\left[U_s W_s W_{t+1}\right]\\
&= 2\sum_{t=2}^{T} \mathbb{E}\left[U_t W_t^2\right] && \text{(Independence of the } W_t\text{)}\\
&= 2\sum_{t=2}^{T} \mathbb{E}\left[U_t\right]\mathbb{E}\left[W_t^2\right] && \text{(Claim 5.12 and independence of the } W_t\text{)}
\end{aligned}$$

This gives an asymptotic description of the iterates $x_t$ (as asymptotically independent trajectories of $X_t$) and the objective value achieved by the algorithm (as above). We can now try to optimize the best choice of the functions $u_t$ subject to the constraint that the output point is Boolean. This yields the following program for the value achievable by an iteration with $T$ steps (selecting $u_t(w_{t-1}, \dots, w_0)$ is equivalent to selecting $U_t$ which is measurable with respect to $W_{t-1}, \dots, W_0$):

$$\max\ 2\sum_{t=1}^{T} \mathbb{E}\left[U_t\right]\mathbb{E}[W_t^2]$$

$$\begin{aligned} \text{s.t.} \quad & U_t \text{ is measurable w.r.t. } W_0, \dots, W_{t-1} \\ & X_T = \sum_{t=1}^{T} U_t W_t \in \{-1, +1\} \text{ a.s.} \end{aligned}$$

Note that by Claim 5.12, the trajectory $X_T = \sum_{t=1}^{T} U_t W_t$ is a martingale. As written, the Boolean constraint "$X_T \in \{-1, 1\}$ a.s." is infeasible because each step $W_t$ is a (potentially unbounded) Gaussian. In the actual algorithm, we make $X_T$ close to $\pm 1$ then apply a final rounding step.

The remaining key step used by [Mon19] is to take $T$ large in order to approach a continuous time stochastic process $\mathrm{d}X_t = U_t \mathrm{d}B_t$. The limiting Brownian motion only appears if we add a constraint that $\mathbb{E}[U_t^2] = 1$ so that $\mathbb{E}[W_t^2] = \mathbb{E}[U_t^2]\,\mathbb{E}\left[W_{t-1}^2\right] = \mathbb{E}\left[W_{t-1}^2\right]$ for all $t$.

This yields a continuous optimization problem for the best achievable value:

$$\begin{aligned} \max \quad & 2 \int_0^1 \mathbb{E}[U_t] \, \mathrm{d}t \\ \text{s.t.} \quad & (U_t)_{t \in [0,1]} \text{ is progressively measurable w.r.t. a Brownian motion } (B_t)_{t \in [0,1]} \\ & \mathbb{E}\left[U_t^2\right] = 1 \text{ for all } t \in [0, 1] \\ & X_1 = \int_0^1 U_t \mathrm{d}B_t \in \{-1, +1\} \text{ a.s.} \end{aligned}$$

This continuous optimization problem is convex in $(U_t)_{t \in [0,1]}$ and dual to an "extended Parisi formula" for the optimal value of the SK model [AMS21, Section 4]. The remaining important technical step is to show that this program is well-posed, and that the maximizer of this program, which can be written in terms of the solution to the Parisi PDE, is smooth enough that it can be discretely approximated by the limit $T \to \infty$.

# 6 Analyzing $\mathrm{poly}(n)$ Iterations

In summary so far, we have completely described the asymptotic trajectory of first-order algorithms for a *constant* number of iterations. We now discuss extensions to a number of iterations that scales with the dimension $n$ of the matrix.

A motivation for studying longer iterations is that for problems with a hidden planted signal, it has been observed empirically that first-order iterations initialized at random can learn the planted signal. However, the standard machinery is only able to prove that these algorithms achieve recovery from an *informative* initialization which has positive correlation with the planted signal. The underlying reason appears to be that "picking up" the signal and escaping the random initialization takes $\omega(1)$ steps, which is beyond what most previous works can handle.

## 6.1 Combinatorial phase transitions

In order to show that this is a delicate question, we compute in Appendix C that some diagrams of $\omega(1)$ size are no longer asymptotically Gaussian, breaking the classification Theorem 4.11. Larger-degree vertices in a diagram can access high moments of the entries of other diagrams, which will detect that these quantities are not exactly Gaussian.

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

$$x_0 = \vec{1}\,, \quad x_{t+1} = Ax_t - x_{t-1}\,. \tag{20}$$

Eq. (20) has a very simple tree approximation (the $t$-path diagram). Note that by Theorem 5.1, for constantly many iterations this algorithm is asymptotically equivalent to power iteration on the non-backtracking walk matrix, which is the algorithm

$$m_0 = \vec{1}\,, \qquad m_{t+1} = Bm_t\,,$$
$$x_{t+1,i} = \sum_{k=1}^{n} A_{ik} m_{t,k\to i}\,,$$

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

We distinguish several regimes of $T = T(n)$ depending on the obstacles that arise when trying to generalize the tree approximation for Eq. (20) to a larger number of iterations.

- When $T \ll \frac{\log n}{\log\log n}$, we expect the proofs of Theorem 4.11 and Theorem 4.14 to generalize with minimal changes. The total number of diagrams that arise can be bounded by $T^{O(T)}$ which is $n^{o(1)}$ in this regime.
- When $T \approx \frac{\log n}{\log\log n}$, there are $T^{O(T)} = \mathrm{poly}(n)$ many diagrams to keep track of. This could overpower the magnitude of some cyclic diagrams, and make the naive union bound argument fail. This barrier is also the one of previous non-asymptotic analyses of AMP [RV18, CR23].
- When $T \ll n^{\delta}$ for some small constant $\delta > 0$, we show in the next subsections that the tree approximation of debiased power iteration still holds by a more careful accounting of the error terms. We predict that this can be extended up to $T \ll \sqrt{n}$.
- When $T \approx \sqrt{n}$, the tree diagrams with $T$ vertices are exponentially small in magnitude (see Lemma A.2) and the number of non-tree diagrams starts to become overwhelmingly large. At the conceptual level, random walks of length $>\sqrt{n}$ in an $n$-vertex graph are likely to collide. Therefore, it is unclear whether or not the tree diagrams of size $>\sqrt{n}$ are significantly different from diagrams with cycles. This threshold also appears in recent analyses of AMP [LFW23], although it is not a barrier for their result.

## 6.2 Analyzing power iteration via combinatorial walks

For constantly many iterations of debiased power iteration, by Theorem 4.14, we know that $x_t$ is well-approximated by the $t$-path diagram, denoted $Z_{t\text{-path}}$. Here we prove that this approximation holds much longer. To simplify the calculation, we assume:

**Assumption 6.1.** *Let $A$ be a random $n \times n$ symmetric matrix with $A_{ii} = 0$ and $A_{ij}$ drawn independently from the uniform distribution over $\left\{-\frac{1}{\sqrt{n-1}}, \frac{1}{\sqrt{n-1}}\right\}$ for all $i < j$.*

We prove that for this iterative algorithm we can extend Theorem 4.14 to a polynomial number of iterations, hence overcoming some obstructions mentioned in Section 6.1. A similar argument can also show that $Z_{t\text{-path}}$ remain approximately independent Gaussians for $t$ in the same regime. Taken together, we see that the "usual" state evolution formula for constantly many iterations continues to hold much longer, up to conjecturally $\sqrt{n}$ iterations.

**Theorem 6.2.** *Suppose that $A = A(n)$ satisfies Assumption 6.1 and generate $x_t$ according to Eq. (20). Then there exist universal constants $c, \delta > 0$ such that for all $t \le cn^{\delta}$,*

$$\|x_t - Z_{t\text{-path}}\|_{\infty} \xrightarrow{a.s.} 0\,.$$

To obtain the tree approximation of algorithms with $\mathrm{poly}(n)$ many iterations, we need to very carefully count combinatorial factors that were neglected in Section 4. The total number of diagrams in the unapproximated diagram expansion is very large, and furthermore, each diagram can arise in many different ways if it has high-degree vertices. To perform the analysis, we decompose $x_t$ in terms of walks of length $t$; we need to track walks instead of diagrams so that we do not throw away additional information about high-degree vertices.

Our goal is to show that the walk without any back edge (the $t$-path) dominates asymptotically. We will proceed as in the proof of Theorem 4.11 by bounding the $q$-th moment of $x_t - Z_{t\text{-path}}$. This moment can be represented diagrammatically using $q$-tuples of non-backtracking walks with at least one back edge.

**Definition 6.3.** *A $(q,t)$-traversal $\gamma = (\gamma_1, \dots, \gamma_q)$ is an ordered sequence of $q$ walks, each of length $t$ and starting from the same vertex:*

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

*Each traversal $\gamma$ is naturally associated to an improper diagram $(V(\gamma), E(\gamma))$ with $V(\gamma) = \{u_{i,j} : i \in [q], j \in [t]\}$ and $E(\gamma) = \{(u_{i,j}, u_{i,j+1}) : i \in [q], j \in [t-1]\}$ (viewed as a multiset). We use the notation $Z_\gamma = Z_{(V(\gamma),E(\gamma))}$ following Definition 3.2.*

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

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

Definition 6.3 is motivated by the following decomposition:

**Claim 6.4.** *Suppose that $x_t$ is generated according to Eq. (20) and $A$ satisfies Assumption 6.1. Then,*

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

We now proceed to proving Theorem 6.2. We will bound the magnitude of $\mathbb{E}\left[Z_{\gamma,i}\right]$ for $\gamma \in \mathcal{W}_t^q$, then count the number of traversals in $\mathcal{W}_t^q$. Both bounds will depend on $\frac{E}{2} - V + 1$ (where $V$ is the number of vertices of the traversal and $E$ the number of edges), which quantifies how close the traversal is to a tree of double edges.

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

**Lemma 6.5.** *For any $\gamma \in \mathcal{W}_t^q$ with $V$ vertices and $E$ edges, $\frac{E}{2} - V + 1 \geq \frac{q}{2}$.*

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

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

as needed. □

**Lemma 6.6.** *For any $i \in [n]$ and $\gamma \in \mathcal{W}_t^q$ with $V$ vertices and $E$ edges,*

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

*Proof.* Using Assumption 6.1, we can directly count

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

□

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

**Lemma 6.7.** *The number of $\gamma \in \mathcal{W}_t^q$ with $V$ vertices and $E$ edges is at most*

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

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

*Proof of Theorem 6.2.* We decompose the sum over $\gamma \in \mathcal{W}_t^q$ according to the value of $r = \frac{E}{2} - V + 1$ using Lemma 6.6 and Lemma 6.7:

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

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

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

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

## 6.3 Counting combinatorial walks

Our goal here is to prove Lemma 6.7.

In the extreme case $V \approx \frac{E}{2}$ where the moment bound Lemma 6.6 is the weakest, typical traversals $\gamma \in \mathcal{W}_t^q$ look like trees of double edges with a constant number of back edges. In this regime, most vertices will have degree exactly 4. Following this intuition, our encoding will proceed by compressing the long paths of degree-4 vertices connected by double edges.

**Definition 6.8.** *For $\gamma \in \mathcal{W}_t^q$, let $\gamma_c$ be the traversal obtained by replacing all maximally long paths of degree-4 vertices in $\gamma$ by a single special marked edge between the endpoints of the paths, and removing the internal vertices of the path. (The paths should be broken at the root so that it is not removed.)*

Note that these operations can create self-loops in $\gamma_c$.

**Lemma 6.9.** *For any $\gamma \in \mathcal{W}_t^q$,*

$$|E(\gamma_c)| \leq 3|E(\gamma)| - 6(|V(\gamma)| - 1) + 2q .$$

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

$$2|V_2(\gamma)| + 4|V_4(\gamma)| + 6\Big(|V(\gamma)| - |V_2(\gamma)| - |V_4(\gamma)| - 1\Big) \leq 2|E(\gamma)| .$$

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

$$|E(\gamma)| - |E(\gamma_c)| = 2|V_4(\gamma)| \geq 6(|V(\gamma)| - 1) - 4|V_2(\gamma)| - 2|E(\gamma)| .$$

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

We are now ready to prove Lemma 6.7.

*Proof of Lemma 6.7.* We encode a traversal $\gamma \in \mathcal{W}_t^q$ as follows:

1. We first encode $\gamma_c$. We write down the sequence of vertices of each walk and indicate whether each step should be the first step of a marked edge (Definition 6.8). Every time we traverse a marked edge for the second time, instead of recording the next vertex of the walk, we record the identifier of the marked edge. We also add a single bit of information to each edge to indicate whether it is the last edge of its walk. The target space of the encoding has size $O(|E(\gamma_c)|)^{|E(\gamma_c)|}$.

2. We then expand the marked edges in $\gamma_c$ of which there are at most $|E(\gamma_c)|/2$. For each marked edge, we write down the length of the path that it replaced. This can be encoded using "stars and bars". Initially allocating 2 edges to each marked edge, there are at most $\binom{E}{|E(\gamma_c)|/2}$ such objects.

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

$$O(|E(\gamma_c)|)^{|E(\gamma_c)|}\binom{E}{|E(\gamma_c)|/2} \leq O(|E(\gamma_c)|)^{|E(\gamma_c)|}O\left(\frac{E}{|E(\gamma_c)|}\right)^{|E(\gamma_c)|/2} = O_q(t)^{|E(\gamma_c)|}\,.$$

The proof follows after plugging in the bound of Lemma 6.9. □

# References


[AM20] Ahmed El Alaoui and Andrea Montanari. Algorithmic Thresholds in Mean Field Spin Glasses. *arXiv preprint arXiv:2009.11481*, 2020. 7, 36

[AMP20] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: Norm bounds and applications. *arXiv preprint arXiv:1604.03423*, 2020. 9, 20

[AMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *Annals of Probability*, 49(6):2922–2960, 2021. 7, 10, 36, 38

[AMS23] Antonio Auffinger, Andrea Montanari, and Eliran Subag. Optimization of Random High-Dimensional Functions: Structure and Algorithms. In *Spin Glass Theory and Far Beyond*, chapter 29, pages 609–633. World Scientific, 2023. 1

[Bet35] Hans A. Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 150(871):552–575, 1935. 2

[BHK+19] Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem. *SIAM Journal on Computing*, 48(2):687–735, 2019. 9

[Bil95] Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, Third edition, 1995. 12

[BLM15] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *Annals of Applied Probability*, 25(2):753–822, 2015. 5, 6, 7, 8, 10

[BM11] 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, 2011. 7, 8, 10

[BMN20] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020. 10

[BN06] Mohsen Bayati and Chandra Nair. A rigorous proof of the cavity method for counting matchings. In *Proceedings of the 44th Annual Allerton Conference on Communication, Control and Computing*, 2006. 9

[Bol14] Erwin Bolthausen. An Iterative Construction of Solutions of the TAP Equations for the Sherrington–Kirkpatrick Model. *Communications in Mathematical Physics*, 325(1):333–366, 2014. 1, 2, 7, 10

[Bor19] Charles Bordenave. Lecture notes on random matrix theory, 2019. 13

[CCM21] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021. 10

[CH06] Philippe Carmona and Yueyun Hu. Universality in Sherrington–Kirkpatrick's spin glass model. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 42(2):215–222, 2006. 36

[CKPZ17] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. Information-theoretic thresholds from the cavity method. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017*, pages 146–157. ACM, 2017. 9

[CL21] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021. 10

[CMP+23] Patrick Charbonneau, Enzo Marinari, Giorgio Parisi, Federico Ricci-Tersenghi, Gabriele Sicuro, Francesco Zamponi, and Marc Mézard. *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking after 40 Years*. World Scientific, 2023. 2, 9

[CMW20] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory, COLT 2020*, pages 1078–1141. PMLR, 2020. 1, 7, 8

[CR23] Collin Cademartori and Cynthia Rush. A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. *arXiv preprint arXiv:2302.00088*, 2023. 10, 39

[DG21] Amir Dembo and Reza Gheissari. Diffusions interacting through a random matrix: universality via stochastic Taylor expansion. *Probability Theory and Related Fields*, 180:1057–1097, 2021. 10

[DLS23] Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *Annals of Probability*, 51(5):1616–1683, 2023. 10

[DM15] Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Conference on Learning Theory, COLT 2015*, pages 523–562. PMLR, 2015. 9

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009. 1, 10, 30

[DMM10] David L. Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing: I. motivation and construction. In *IEEE Information Theory Workshop on Information Theory, ITW 2010*, pages 1–5. IEEE, 2010. 5

[DS19] Jian Ding and Nike Sun. Capacity lower bound for the Ising perceptron. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, pages 816–827, 2019. 9

[DSS14] Jian Ding, Allan Sly, and Nike Sun. Satisfiability threshold for random regular NAE-SAT. In *Proceedings of the forty-sixth Annual ACM Symposium on Theory of Computing, STOC 2014*, pages 814–822, 2014. 9

[DSS16] Jian Ding, Allan Sly, and Nike Sun. Maximum independent sets on random regular graphs. *Acta Mathematica*, 217(2):263–340, 2016. 9

[DSS22] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large $k$. *Annals of Mathematics*, 196(1):1–388, 2022. 9

[Fan22] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *Annals of Statistics*, 50(1):197–224, 2022. 10

[FKP19] Noah Fleming, Pravesh Kothari, and Toniann Pitassi. Semialgebraic proofs and efficient algorithm design. *Foundations and Trends in Theoretical Computer Science*, 14(1-2):1–221, 2019. 9

[FVRS22] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A Unifying Tutorial on Approximate Message Passing. *Foundations and Trends in Machine Learning*, 15(4):335–536, 2022. 1, 2, 10

[Gab20] Marylou Gabrié. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020. 2, 9, 10

[GB23] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023. 10

[GJJ+20] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for Sherrington-Kirkpatrick via planted affine planes. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 954–965. IEEE, 2020. 9

[GTM+22] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022. 1, 10

[HKP+18] Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, and Tselil Schramm. On the Integrality Gap of Degree-4 Sum of Squares for Planted Clique. *ACM Transactions on Algorithms*, 14(3):1–31, 2018. 9

[HS23] Brice Huang and Mark Sellke. Optimization Algorithms for Multi-Species Spherical Spin Glasses. *arXiv preprint arXiv:2308.09672*, 2023. 10

[Hua24] Brice Huang. Capacity threshold for the Ising perceptron. *arXiv preprint arXiv:2404.18902*, 2024. 9

[IS24] Misha Ivkov and Tselil Schramm. Semidefinite programs simulate approximate message passing robustly. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024*, pages 348–357. ACM, 2024. 5, 7, 8, 9, 24, 36, 52

[Jan97] Svante Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997. 12

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013. 10

[Jon22] Christopher Jones. *Symmetrized Fourier Analysis of Convex Relaxations for Combinatorial Optimization Problems*. PhD thesis, The University of Chicago, 2022. 9

[JP22] Chris Jones and Aaron Potechin. Almost-orthogonal bases for inner product polynomials. In *Proceedings of the 13th Conference on Innovations in Theoretical Computer Science, ITCS 2022*, volume 215, pages 89:1–89:21, 2022. 9

[JPR+21] Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. Sum-of-Squares Lower Bounds for Sparse Independent Set. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021*, pages 406–416. IEEE, 2021. 9

[JPRX23] Chris Jones, Aaron Potechin, Goutham Rajendran, and Jeff Xu. Sum-of-Squares Lower Bounds for Densest $k$-Subgraph. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023*, pages 84–95. ACM, 2023. 9

[Kab03] Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111, 2003. 10

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009. 2

[KMW24] Dmitriy Kunisky, Cristopher Moore, and Alexander S. Wein. Tensor cumulants for statistical inference on invariant distributions. In *65th Annual Symposium on Foundations of Computer Science, FOCS 2024*, 2024. 9

[KPX24] Pravesh K. Kothari, Aaron Potechin, and Jeff Xu. Sum-of-Squares Lower Bounds for Independent Set on Ultra-Sparse Random Graphs. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024*, pages 1923–1934. ACM, 2024. 9

[KWB19] 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. 8

[LFW23] Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to $\mathbb{Z}_2$-synchronization. *Proceedings of the National Academy of Sciences*, 120(31):e2302930120, 2023. 7, 10, 39

[LSS23] Tengyuan Liang, Subhabrata Sen, and Pragya Sur. High-dimensional Asymptotics of Langevin Dynamics in Spiked Matrix Models. *Information and Inference: A Journal of the IMA*, 12(4):2720–2752, 2023. 10

[Lu21] Yue M. Lu. Householder Dice: A Matrix-Free Algorithm for Simulating Dynamics on Gaussian and Random Orthogonal Ensembles. *IEEE Transactions on Information Theory*, 67(12):8264–8272, 2021. 10

[LW22] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022. 10

[MM09] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. 2, 9

[MMZ01] Olivier C. Martin, Rémi Monasson, and Riccardo Zecchina. Statistical mechanics methods and phase transitions in optimization problems. *Theoretical computer science*, 265(1-2):3–67, 2001. 9

[Mon19] Andrea Montanari. Optimization of the Sherrington-Kirkpatrick Hamiltonian. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019*, pages 1417–1433. IEEE, 2019. 7, 36, 38

[MP03] Marc Mézard and Giorgio Parisi. The cavity method at zero temperature. *Journal of Statistical Physics*, 111:1–34, 2003. 9

[MPV86] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. SK Model: The Replica Solution without Replicas. *Europhysics Letters*, 1(2):77, 1986. 2

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific, 1987. 2, 27, 28

[MPW15] Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares Lower Bounds for Planted Clique. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, pages 87–96, 2015. 9

[MRB17] Yanting Ma, Cynthia Rush, and Dror Baron. Analysis of approximate message passing with a class of non-separable denoisers. In *IEEE International Symposium on Information Theory, ISIT 2017*, pages 231–235. IEEE, 2017. 10

[MSR73] Paul C. Martin, Eric D. Siggia, and Harvey A. Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973. 10

[MW19] Ankur Moitra and Alexander S. Wein. Spectral methods from tensor networks. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC 2019*, pages 926–937, 2019. 9

[MW22a] Andrea Montanari and Alexander S. Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *arXiv preprint arXiv:2212.06996*, 2022. 8, 9, 19, 24

[MW22b] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022. 1

[NS06] Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*. Cambridge University Press, 2006. 10

[Pan13] Dmitry Panchenko. *The Sherrington–Kirkpatrick model*. Springer Science & Business Media, 2013. 9

[Par79] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979. 2

[Par80] Giorgio Parisi. A sequence of approximated solutions to the SK model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980. 2

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. 10

[PR20] Aaron Potechin and Goutham Rajendran. Machinery for Proving Sum-of-Squares Lower Bounds on Certification Problems. *arXiv preprint arXiv:2011.04253*, 2020. 9

[PR22] Aaron Potechin and Goutham Rajendran. Sub-exponential time Sum-of-Squares lower bounds for Principal Components Analysis. In *Advances in Neural Information Processing Systems, NeurIPS 2022*, volume 35, pages 35724–35740, 2022. 9

[RSS18] Prasad Raghavendra, Tselil Schramm, and David Steurer. High dimensional estimation via sum-of-squares proofs. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3389–3423. World Scientific, 2018. 9

[RT23] Goutham Rajendran and Madhur Tulsiani. Concentration of polynomial random matrices via Efron-Stein inequalities. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA 2023*, pages 3614–3653. SIAM, 2023. 9

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018. 5, 10, 39

[SS24a] Juspreet Singh Sandhu and Jonathan Shi. Sum-of-Squares & Gaussian Processes I: Certification. *arXiv preprint arXiv:2401.14383*, 2024. 9

[SS24b] Juspreet Singh Sandhu and Jonathan Shi. Sum-of-Squares & Gaussian Processes II: Rounding. *To appear*, 2024. 9

[Tak19a] Keigo Takeuchi. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386, 2019. 10

[Tak19b] Keigo Takeuchi. A unified framework of state evolution for message-passing algorithms. In *IEEE International Symposium on Information Theory, ISIT 2019*, pages 151–155. IEEE, 2019. 10

[Tak21] Keigo Takeuchi. Bayes-optimal convolutional AMP. *IEEE Transactions on Information Theory*, 67(7):4405–4428, 2021. 10

[Tal06] Michel Talagrand. The Parisi formula. *Annals of Mathematics*, pages 221–263, 2006. 9, 36

[Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010. 9

[TAP77] David J. Thouless, Philip W. Anderson, and Robert G. Palmer. Solution of 'Solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977. 10, 28

[WZ23] Yuchen Wu and Kangjie Zhou. Lower Bounds for the Convergence of Tensor Power Iteration on Random Overcomplete Models. In *Conference on Learning Theory, COLT 2023*, volume 195, pages 3783–3820. PMLR, 2023. 10

[WZ24] Yuchen Wu and Kangjie Zhou. Sharp Analysis of Power Iteration for Tensor PCA. *arXiv preprint arXiv:2401.01047*, 2024. 10

[WZF22] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022. 10

[YFW03] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236-239):0018–9448, 2003. 2

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. 2, 9, 30

[ZY22] Qiuyun Zou and Hongwen Yang. A Concise Tutorial on Approximate Message Passing. *arXiv preprint arXiv:2201.07487*, 2022. 2

# A Non-asymptotic Diagram Analysis

## A.1 Fourier analytic properties

In Definition 3.2, for a proper $\alpha \in \mathcal{A}$ (a graph instead of a multigraph), $Z_\alpha$ has entries which are homogeneous multilinear polynomials in the entries of the matrix $A$. The next lemma shows that the proper diagrams with size at most $n$ form an orthogonal basis of symmetric polynomials in $A$ with respect to the expectation over $A$.

**Lemma A.1.** *For all $i, j \in [n]$ and distinct proper diagrams $\alpha, \beta \in \mathcal{A}$, $\mathbb{E}\left[Z_{\alpha,i} Z_{\beta,j}\right] = 0$.*

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

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

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

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

**Lemma A.2.** *For all $i \in [n]$ and proper $\alpha \in \mathcal{A} \setminus \{\circledcirc\}$ we have $\mathbb{E}[Z_{\alpha,i}] = 0$ and*

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

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

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

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

Since $\mathbb{E}[A_{jk}] = 0$ for $j \neq k$, the only terms with nonzero expectation have each $A_{jk}$ occurring at least twice. As $\varphi_1$ are $\varphi_2$ are injective, each $A_{jk}$ can only occur at most twice. Therefore, if we fix $\varphi_1$ the embeddings $\varphi_2$ that contribute a nonzero value are exactly graph isomorphisms onto $\mathrm{im}(\varphi_1)$. The total number of choices for $\varphi_1$ and $\varphi_2$ is $(n-1)\cdots(n-|V(\alpha)|+1)\cdot|\mathrm{Aut}(\alpha)|$ and the expectation of a nonzero term is

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

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

$$\begin{aligned}\left|\log\left(\frac{(n-1)\ldots(n-|V(\alpha)|+1)}{n^{|V(\alpha)|-1}}\right)\right| &\leq \sum_{i=1}^{|V(\alpha)|-1} \left|\log\left(1-\frac{i}{n}\right)\right| \\ &\leq \sum_{i=1}^{|V(\alpha)|-1} \frac{i}{n} \underset{n\to\infty}{\longrightarrow} 0\,.\end{aligned}$$

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

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

We will use orthogonality repeatedly in the sequel through the following direct consequence of Lemma A.1 and Lemma A.2:

**Corollary A.3.** *Let $x = \sum_{\text{proper } \alpha \in \mathcal{A}} c_\alpha Z_\alpha$. Then for any $\tau \in \mathcal{T}$,*

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

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

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

## A.2 Operations on the diagram representation

We compute the diagrammatic effect of multiplying by $A$.

**Lemma A.4.** *For all diagrams $\alpha \in \mathcal{A}$,*

$$AZ_\alpha = Z_{\alpha^+} + \sum_{v \in V(\alpha)} Z_{\text{contract } v \text{ and } \circledcirc \text{ in } \alpha^+}\,.$$

*Proof.*

$$\begin{aligned}(AZ_\alpha)_i &= \sum_{j=1}^{n} A_{ij} \sum_{\substack{\varphi: V(\alpha)\to[n] \\ \varphi \text{ injective} \\ \varphi(\circledcirc)=j}} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)} \\ &= \sum_{\substack{\varphi: V(\alpha)\to[n] \\ \varphi \text{ injective}}} A_{i,\varphi(\circledcirc)} \prod_{\{u,v\}\in E(\alpha)} A_{\varphi(u)\varphi(v)}\,.\end{aligned}$$

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

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

**Definition A.5** (Intersection pattern, $P \in \mathcal{P}(\alpha_1, \ldots, \alpha_k)$)**.** *Let $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$. Let $\alpha$ be the diagram obtained by putting all $\alpha_i$ at the same root. An intersection pattern $P$ is a partition of $V(\alpha) \setminus \{\circledcirc\}$ such that for all $i \in [k]$ and $v, w \in V(\alpha_i) \setminus \{\circledcirc\}$, $v$ and $w$ are not in the same block of the partition.*

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

**Definition A.6** (Intersection diagram, $\alpha_P$)**.** *Let $\alpha \in \mathcal{A}$. Given a partition $P$ of $V(\alpha)$, let $\alpha_P$ be the diagram obtained by contracting each block of $P$ into a single vertex. Keep all edges (hence there may be new multiedges or self-loops).*

By casing on which vertices are equal among the embeddings of $\alpha_1, \ldots, \alpha_k$ as in the proof of Lemma A.4, we have:

**Lemma A.7.** *For $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$, the componentwise product of $Z_{\alpha_1}, \ldots, Z_{\alpha_k}$ is*

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

Next we consider these operations when restricted to the tree diagrams. Suppose we start from $\tau \in \mathcal{T}$ and compute $AZ_\tau$. Which diagrams appearing in Lemma A.4 are non-negligible? Following the asymptotic classification of non-negligible diagrams (Section 4.2), it is only $\tau^+$ and $\tau^-$ (the latter only appears if the root of $\tau$ has degree 1, in which case $\tau^-$ is the result of intersecting $\circledcirc$ and the child of the root then removing a double edge). Hence we conclude

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

Given tree diagrams $\tau_1, \dots, \tau_k \in \mathcal{T}$, the asymptotically non-negligible terms in the product in Lemma A.7 are identified as follows. Let $\widetilde{\tau}$ be a non-negligible diagram appearing in the result, i.e. $\widetilde{\tau}$ is a tree with hanging trees of double edges. Since $\tau_1, \dots, \tau_k$ are connected, the hanging double trees must hang off the root vertex of $\widetilde{\tau}$ in order to avoid cycles. Additionally, they must arise as the overlap of two complete copies of the tree. Thus the asymptotically non-negligible terms are the partial matchings between isomorphic branches of the roots of the $\tau_i$. Two copies of a branch $\sigma \in \mathcal{S}$ can be matched up into a tree of double edges in $|\mathrm{Aut}(\sigma)|$ ways.

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

**Definition A.8** (Tree approximation of a GFOM, $\widehat{x}_t$)**.** *Let $x_t \in \mathbb{R}^n$ be the state of a GFOM. We recursively define the tree approximation of $x_t$, denoted by $\widehat{x}_t$, to be a diagram expression in the span of $(Z_\tau)_{\tau \in \mathcal{T}}$.*

1. *Initially, $\widehat{x}_0 = Z_{\circledcirc}$.*
2. *If $x_{t+1} = Ax_t$, define $\widehat{x}_{t+1} = (\widehat{x}_t)^+ + (\widehat{x}_t)^-$.*
3. *If $x_{t+1} = f_t(x_t, \dots, x_0)$ coordinatewise for some polynomial $f_t : \mathbb{R}^t \to \mathbb{R}$, define $\widehat{x}_{t+1}$ by applying each monomial of $f_t$ to $\widehat{x}_t, \dots, \widehat{x}_0$ separately and summing the results. To apply a monomial on $\widehat{x}_t, \dots, \widehat{x}_0$, expand each $\widehat{x}_s$ in the diagram basis and sum all the cross product terms. The result of multiplying $q$ tree diagrams $\tau_1, \dots, \tau_q \in \mathcal{T}$ is*

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

*where:*

   (a) *$\mathcal{M}(\tau_1, \dots, \tau_q)$ is the set of (partial) matchings of isomorphic branches of $\tau_1, \dots, \tau_q$ such that no two branches from the same $\tau_i$ are matched.*

   (b) *$\tau_M$ is the tree obtained by merging the roots of $\tau_1, \dots, \tau_q$ and removing all subtrees matched in $M$.*

   (c) *$c_M = \prod_{\{\sigma, \sigma'\} \in M} |\mathrm{Aut}(\sigma)|$.*

## A.3 Repeated-label diagram basis

An alternative basis for the diagram space consists of diagrams in which labels are allowed to repeat. This representation has been defined by Ivkov and Schramm [IS24, Section 3.5].

**Definition A.9** ($\widetilde{Z}_\alpha$)**.** *For a diagram $\alpha$ with root $\circledcirc$, define $\widetilde{Z}_\alpha \in \mathbb{R}^n$ by*

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

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

**Lemma A.10.** *For $\alpha \in \mathcal{A}$,*

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

*where $\mathcal{P}(\alpha)$ is the set of partitions of $V(\alpha)$ and $\alpha_P$ contracts the blocks of $P$ (Definition A.6).*

*Proof.* We have

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

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

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

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

(i) **Multiplying by $A$ extends the root.**

We have $A\widetilde{Z}_\alpha = \widetilde{Z}_{\alpha^+}$ where $\alpha^+$ is obtained by extending the root by one edge.

(ii) **Componentwise products graft trees together.**

To componentwise multiply $\widetilde{Z}_\alpha$ and $\widetilde{Z}_\beta$, we "graft" $\alpha$ and $\beta$ by merging their roots.

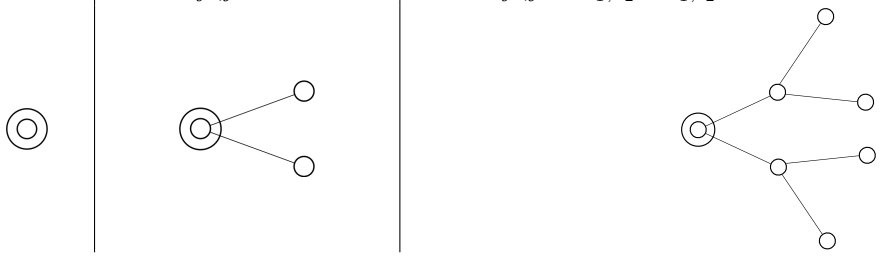
**Example A.11.** *Consider the example,*

$$x_{t+1} = (Ax_t)^2 \qquad\qquad x_0 = \vec{1}$$

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

$x_0 = \vec{1}$ | $x_1 = (A\vec{1})^2$ | $x_2 = (A(A\vec{1})^2)^2$

$x_{0,i} = 1$ | $x_{1,i} = \sum_{j_1,j_2=1}^{n} A_{ij_1}A_{ij_2}$ | $x_{2,i} = \sum_{j_1,j_2=1}^{n}\sum_{k_1,k_2=1}^{n}\sum_{\ell_1,\ell_2=1}^{n} A_{ij_1}A_{ij_2}A_{j_1k_1}A_{j_1\ell_1}A_{j_2k_2}A_{j_2\ell_2}$



# B Omitted Proofs

## B.1 Removing hanging double edges

In order to implement the removal of hanging double edges, we introduce an additional diagrammatic construct to track the error, *2-labeled edges*. These terms are equal to zero when $A$ is a Rademacher matrix and it is recommended to ignore them on a first read.

**Definition B.1** (Edge-labeled diagram)**.** *An edge-labeled diagram is a diagram in which some of the edges are labeled "2".*

*We let $E(\alpha)$ denote the entire multiset of labeled and unlabeled edges of $\alpha$, $E_2(\alpha)$ the multiset of 2-labeled edges and $E_1(\alpha) = E \setminus E_2(\alpha)$ the multiset of non-labeled edges.*

*We use the convention that $|E(\alpha)|$ counts each 2-labeled edge twice, so that $|E(\alpha)|$ continues to equal the degree of the polynomial $Z_{\alpha,i}$.*

**Definition B.2** (Edge-labeled $Z_\alpha$)**.** *For an edge-labeled diagram $\alpha$, we define $Z_\alpha \in \mathbb{R}^n$ by*

$$Z_{\alpha,i} = \sum_{\substack{\text{injective } \varphi:V(\alpha)\to[n] \\ \varphi(\circledcirc)=i}} \prod_{\{u,v\}\in E_1(\alpha)} A_{\varphi(u)\varphi(v)} \prod_{\{u,v\}\in E_2(\alpha)} \left(A^2_{\varphi(u)\varphi(v)} - \frac{1}{n}\right).$$

The set of diagrams $\mathcal{A}$ is extended to allow diagrams which may have 2-labeled edges. The definition of $I(\alpha)$ from Definition 4.1 must also be updated to incorporate labeled edges (because a labeled edge is mean-0, it is treated like a single edge).

**Definition B.3** (Updated definition of $I(\alpha)$)**.** *For a diagram $\alpha \in \mathcal{A}$, let $I(\alpha)$ be the subset of non-root vertices such that every edge incident to that vertex has multiplicity $\geq 2$ or is a self-loop, treating 2-labeled edges as if they were normal edges.*

The following is an exact decomposition for removing hanging double edges.

**Lemma B.4.** *Let $\alpha \in \mathcal{A}$ be a diagram with a hanging (unlabeled) double edge. Let $\alpha_0$ be $\alpha$ with both the hanging double edge and corresponding hanging vertex removed, and $\alpha_2$ be $\alpha$ with the hanging double edge replaced by a single 2-labeled edge. Then,*

$$Z_\alpha = Z_{\alpha_0} - \frac{|V(\alpha)| - 1}{n} \cdot Z_{\alpha_0} + Z_{\alpha_2}.$$

*Proof.* We write:

$$\begin{aligned} Z_{\alpha,i} &= \sum_{\substack{\text{injective } \varphi:V(\alpha)\to[n] \\ \varphi(\circledcirc)=i}} A_{u,v}^2 \prod_{\{x,y\}\in E(\alpha)\setminus\{\{u,v\},\{u,v\}\}} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{1}{n} \sum_{\substack{\text{injective } \varphi:V(\alpha)\to[n] \\ \varphi(\circledcirc)=i}} \prod_{\{x,y\}\in E(\alpha)\setminus\{\{u,v\},\{u,v\}\}} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{n-|V(\alpha)|+1}{n}\cdot Z_{\alpha_0,i} = Z_{\alpha_0,i} - \frac{|V(\alpha)|-1}{n} Z_{\alpha_0,i} + Z_{\alpha_2,i}\,. \end{aligned}$$

The additional $n-|V(\alpha)|+1$ scaling factor comes from removing the hanging vertex. □

## B.2 Omitted proofs for Section 4.1

We prove a more specific version of Lemma 4.2.

**Lemma B.5.** *Let $q\in\mathbb{N}, \alpha\in\mathcal{A}$, and $i\in[n]$. Then,*

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

*where $M_k$ is a bound on the $k$-th moment of the entries of $A$ (recall the notations of Assumption 2.1),*

$$M_k = \max\left(\mathop{\mathbb{E}}_{X\sim\mu}\left[|X|^k\right], \mathop{\mathbb{E}}_{X\sim\mu_0}\left[|X|^k\right]\right).$$

*When $q$ and $|V(\alpha)|$ are $O(1)$, the overall bound reduces to*

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

*Proof.* We expand $\mathbb{E}\left[Z_{\alpha,i}^q\right]$ as

$$\sum_{\substack{\text{injective } \varphi_1,\dots,\varphi_q:V(\alpha)\to[n] \\ \varphi_1(\circledcirc)=\cdots=\varphi_q(\circledcirc)=i}} \mathbb{E}\left[\prod_{p=1}^{q}\left(\prod_{\{u,v\}\in E_1(\alpha)} A_{\varphi_p(u)\varphi_p(v)}\right)\left(\prod_{\{u,v\}\in E_2(\alpha)}\left(A_{\varphi_p(u)\varphi_p(v)}^2 - \frac{1}{n}\right)\right)\right].$$

This is a polynomial of degree $q|E(\alpha)|$ in $A$ (by convention every 2-labeled edge contributes 2 to $|E(\alpha)|$). We first estimate the magnitude of any summand of the sum over $\varphi_1,\dots,\varphi_q$ with nonzero expectation. Each such summand can be decomposed into $2^{q|E_2(\alpha)|}$ terms by expanding out[12] the $A_{ij}^2-\frac{1}{n}$. This leaves monomials in the entries of $A$ of total degree at most $q|E(\alpha)|$. We bound the expected value of each of these monomials by $M_{q|E(\alpha)|}n^{-q|E(\alpha)|/2}$ using Hölder's inequality. This shows that any nonzero term in the summation has magnitude at most $2^{q|E_2(\alpha)|}M_{q|E(\alpha)|}n^{-q|E(\alpha)|/2}$.

To bound the number of nonzero terms, we observe that every edge $A_{jk}$ for $j\neq k$ must occur zero times or at least twice in order to have nonzero expectation (the self-loops $A_{jj}$

[12]The factor $2^{q|E_2(\alpha)|}$ may be removed with a tighter argument.

can occur any number of times, and the 2-labeled edges $A_{jk}^2 - \frac{1}{n}$ must overlap at least one additional edge in order to have nonzero expectation). Each vertex in $V(\alpha) \setminus I(\alpha) \setminus \{\circledcirc\}$ is incident to an edge of multiplicity 1 or a 2-labeled edge, and so it must occur in at least two embeddings in order for that edge $A_{jk}$ to overlap and not make the expectation 0. This implies that the number of distinct non-root vertices among the embeddings is at most $q\left(|V(\alpha)| - 1 + |I(\alpha)|\right)/2$ where the $-1$ is used to avoid counting the root.

Hence, there are at most $n^{q(|V(\alpha)|-1+|I(\alpha)|)/2}$ ways to choose the entire image $\mathrm{im}(\varphi_1) \cup \ldots \cup \mathrm{im}(\varphi_q)$. Once this is fixed, there are at most $(q|V(\alpha)|)^{q|V(\alpha)|}$ $q$-tuples of embeddings that map to these vertices. We conclude by combining the bound on the number of nonzero terms and the bound on the magnitude of each of these terms. □

**Lemma 4.5.** *Suppose that $A = A(n)$ is a sequence of random matrices satisfying Assumption 2.1. If $x$ and $y$ are diagram expressions such that $x \overset{\infty}{=} y$, then $\|x - y\|_\infty = \widetilde{O}(n^{-1/2})$ with high probability.*

*Proof.* By assumption, $x - y$ is a sum of combinatorially negligible terms. We first focus on a single one of them, say $a_n Z_\alpha$. For any $\varepsilon > 0\,, q \in \mathbb{N}$ and $i \in [n]$, we have

$$\begin{aligned}
\Pr\left(|a_n Z_{\alpha,i}| \geq \varepsilon\right) &\leq \frac{\mathbb{E}\,|a_n Z_{\alpha,i}|^q}{\varepsilon^q} && \text{(Markov's inequality)} \\
&\leq \frac{1}{\varepsilon^q} M_{q|E(\alpha)|} 2^{q|E(\alpha)|} (q|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{-\frac{q}{2}} && \text{(Lemma B.5)} \\
&\leq \frac{1}{\varepsilon^q} (q|E(\alpha)|)^{O(q)} 2^{q|E(\alpha)|} (q|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{-\frac{q}{2}} && \text{(subgaussianity of } A_{ij}) \\
&= \exp\left(O(q \log q) - \frac{q}{2} \log n + q \log(1/\varepsilon)\right)\,.
\end{aligned}$$

Picking $q = \log n$ and $\varepsilon = q^C n^{-1/2}$ and taking the constant $C$ large enough we can make the probability an arbitrarily small inverse polynomial in $n$. Then we take a union bound over all $i \in [n]$ and all combinatorially negligible term appearing in $x - y$ (there are constantly many such terms by definition). □

**Lemma 4.6.** *If $x, y$ are diagram expressions with $x \overset{\infty}{=} y$, then*

$$Ax \overset{\infty}{=} Ay\,.$$

*Moreover, if $x_1, \ldots, x_t, y_1, \ldots, y_t$ are diagram expressions with $x_i \overset{\infty}{=} y_i$ for all $i \in [t]$, then*

$$f(x_1, \ldots, x_t) \overset{\infty}{=} f(y_1, \ldots, y_t)\,,$$

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

*Proof.* It suffices to prove that for a combinatorially negligible term $n^{-k} Z_\alpha$:

(i) All terms in the diagram representation of $n^{-k} A Z_\alpha$ are combinatorially negligible.

(ii) Let $n^{-\ell} Z_\beta$ be any term of combinatorial order 1 or combinatorially negligible. Then all terms in the diagram representation of the componentwise product $n^{-(k+\ell)} Z_\alpha \odot Z_\beta$ are combinatorially negligible, where $\odot$ is the componentwise product.

For (i), the diagram representation of $AZ_\alpha$ is given by Lemma A.4. In the term $\alpha^+$ without intersections,

$$|V(\alpha^+)| = |V(\alpha)| + 1\,, \qquad |I(\alpha^+)| = |I(\alpha)|\,, \qquad |E(\alpha^+)| = |E(\alpha)| + 1\,.$$

From this we can check that $n^{-k}Z_{\alpha^+}$ is still combinatorially negligible.

In a term $\beta$ corresponding to an intersection between the new root and a vertex of $\alpha$,

$$|V(\beta)| = |V(\alpha)|\,, \qquad |I(\beta)| \leq |I(\alpha)| + 1\,, \qquad |E(\beta)| = |E(\alpha)| + 1\,.$$

The second inequality follows from the observation that the only vertices from $\alpha$ whose neighborhood structure can be affected by the intersection are the root of $\alpha$ (which does not contribute to $|I(\alpha)|$) and the intersected vertex. Hence, $n^{-k}Z_\beta$ is also combinatorially negligible.

For (ii), the diagram representation of $Z_\alpha \odot Z_\beta$ is given by Lemma A.7. Fix an intersection pattern $P \in \mathcal{P}(\alpha, \beta)$ that has $b$ blocks and denote by $\gamma$ the resulting diagram. Then,

$$\begin{aligned}
&|V(\gamma)| = b + 1\,,\\
&|E(\gamma)| = |E(\alpha)| + |E(\beta)|\,,\\
&|I(\gamma)| \leq |I(\alpha)| + |I(\beta)| + |V(\alpha)| + |V(\beta)| - b - 2\,.
\end{aligned}$$

The last inequality is proven by observing that for a non-root vertex that is neither in $I(\alpha)$ nor $I(\beta)$ to contribute to $I(\gamma)$, it must intersect another vertex. Moreover, there are at most $|V(\alpha)| + |V(\beta)| - b - 2$ intersected non-root vertices in $\gamma$.

Putting everything together,

$$\begin{aligned}
&|V(\gamma)| - 1 - |E(\gamma)| + |I(\gamma)|\\
\leq\ &|V(\alpha)| - 1 - |E(\alpha)| + |I(\alpha)| + |V(\beta)| - 1 - |E(\beta)| + |I(\beta)|\\
<\ &2(k + l)\,,
\end{aligned}$$

since $n^{-k}Z_\alpha$ is combinatorially negligible and $n^{-\ell}Z_\beta$ is at most order 1. This concludes the proof. □

Using the 2-labeled edges introduced in Appendix B.1, we can implement the removal of hanging double edges.

**Lemma 4.7.** *Let $a_n Z_\alpha$ be a term of combinatorial order at most 1 such that $\alpha$ has a hanging double edge. Let $\alpha_0$ be $\alpha$ with the hanging double edge and hanging vertex removed. Then*

$$a_n Z_\alpha \overset{\infty}{=} a_n Z_{\alpha_0}\,.$$

*Proof.* Starting from the decomposition of Lemma B.4,

$$a_n Z_\alpha = a_n Z_{\alpha_0} - a_n \frac{|V(\alpha)| - 1}{n} Z_{\alpha_0} + a_n Z_{\alpha_2}\,,$$

we claim that the first term is combinatorially order 1, and the second and third terms are combinatorially negligible. Comparing $\alpha_0$ to $\alpha$, two edges and one vertex in $I(\alpha)$ are removed.

This does not change the combinatorial order. The second term scales down by $n$ and this becomes negligible (by assumption $|V(\alpha)|$ is constant). In the third term, $|I(\alpha_2)| < |I(\alpha)|$ to take into account the hanging vertex, while $|V(\alpha)| = |V(\alpha_2)|$ and $|E(\alpha)| = |E(\alpha_2)|$ remain unchanged, making the term negligible. We remind the reader that $|E(\alpha)| = |E(\alpha_2)|$ because $|E(\alpha_2)|$ counts 2-labeled edges twice. □

Definition 4.3 includes the coefficient $a_n$ in the definition in order to incorporate factors of $\frac{1}{n}$ on some error terms such as those in the proof above.

## B.3 Scalar diagrams

We collect the properties of scalar diagrams (Definition 4.8) which naturally generalize those of vector diagrams. We omit the proofs of the results in this section, as they are direct modifications of their vector analogs.

First, the scalar diagrams are an orthogonal basis for scalar functions of $A$.

**Lemma B.6.** *For any proper $\alpha \in \mathcal{A}_{\text{scalar}}$:*

- *For any proper $\beta \in \mathcal{A}_{\text{scalar}}$ such that $\beta \neq \alpha$, $\mathbb{E}[Z_\alpha Z_\beta] = 0$.*
- *$\mathbb{E}[Z_\alpha] = 0$ if $\alpha$ is not a singleton.*
- *The second moment of $Z_\alpha$ is*

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

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

*Proof.* Analogous to Lemma A.1 and Lemma A.2. □

When scalar and vector diagrams are multiplied together, the result can be expressed in terms of diagrams by extending the notion of intersection patterns $\mathcal{P}(\alpha_1, \ldots, \alpha_k)$ (Definition A.5) and intersection diagrams (Definition A.6) to allow scalar and vector diagrams simultaneously. The "unintersected" diagram consists of adding all the scalar diagrams as floating components to the vector diagrams, which are put at the same root. Intersection patterns are partitions of this vertex set such that no two vertices from the same diagram are matched.

**Lemma B.7.** *Let $\alpha_1, \ldots, \alpha_k$ be either scalar or vector diagrams. Then*

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

*where the product is componentwise for the vector diagrams.*

*Proof.* Analogous to Lemma A.7. □

We define $I(\alpha)$ for scalar diagrams exactly as in Definition 4.1.

**Lemma B.8.** *Let $q \in \mathbb{N}, \alpha \in \mathcal{A}_{\text{scalar}}$, and $i \in [n]$. Then,*

$$|\mathbb{E}[Z_\alpha^q]| \le M_{q|E(\alpha)|} 2^{q|E(\alpha)|} (q|V(\alpha)|)^{q|V(\alpha)|} \cdot n^{\frac{q}{2}(|V(\alpha)|-|E(\alpha)|+|I(\alpha)|)},$$

*where $M_k$ is defined as in Lemma B.5. When $q$ and $|V(\alpha)|$ are $O(1)$, this reduces to*

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

*Proof.* Analogous to Lemma B.5. □

**Definition B.9** (Combinatorially negligible and order 1 scalar)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence of real-valued coefficients with $a_n = \Theta(n^{-k})$, where $k \ge 0$ is such that $2k \in \mathbb{Z}$. Let $\alpha \in \mathcal{A}_{\text{scalar}}$ be a scalar diagram.*

- *We say that $a_n Z_\alpha$ is combinatorially negligible if*

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

- *We say that $a_n Z_\alpha$ has combinatorial order 1 if*

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

*We define $\overset{\infty}{=}$ for scalar diagram expressions exactly as in Definition 4.4.*

**Lemma B.10.** *Let $x$ and $y$ be scalar diagram expressions with $x \overset{\infty}{=} y$. Then $|x - y| = \widetilde{O}(n^{-1/2})$ with high probability.*

*Proof.* Analogous to Lemma 4.5. □

**Lemma B.11.** *Let $a_n Z_\alpha$ be a combinatorially negligible scalar term. Let $b_n Z_\beta$ be any scalar or vector term of combinatorial order at most 1. Then all terms in the product $a_n b_n Z_\alpha Z_\beta$ are combinatorially negligible.*

*Proof.* Analogous to Lemma 4.6. □

In Lemma 4.10, we characterized the connected vector diagrams which are combinatorially order 1. We now similarly characterize the order 1 scalar diagrams.

**Lemma B.12.** *Let $\alpha \in \mathcal{A}_{\text{scalar}}$ be a scalar diagram with $c$ connected components, $c_I$ of which contain only vertices in $I(\alpha)$. Then $n^{-(c+c_I)/2} Z_\alpha$ is combinatorially negligible or combinatorially order 1, and it is combinatorially order 1 if and only if the following conditions hold simultaneously:*

*(i) Every multiedge has multiplicity 1 or 2.*

*(ii) There are no cycles.*

*(iii) In each component, the subgraph of multiplicity 1 edges is empty or a connected graph (i.e. the multiplicity 2 edges consist of hanging trees)*

*(iv) There are no self-loops or 2-labeled edges (Appendix B.1).*

*Proof.* We proceed as in the proof of Lemma 4.10. In each connected component $C$ containing at least one vertex $s \in V(\alpha) \setminus I(\alpha)$, we run a breadth-first search from $s$, assigning the multiedges used to explore a vertex to that vertex. This assigns at least one edge to every vertex in $C \setminus \{s\}$, and at least two edges to every vertex in $I(\alpha) \cap C$. This encoding argument shows that

$$2|I(\alpha) \cap C| + |(V(\alpha) \setminus I(\alpha)) \cap C| - 1 \le |E(C)|\,, \tag{21}$$

where $E(C)$ denotes the set of edges in the connected component $C$.

In each connected component $C$ composed only of vertices in $I(\alpha)$, we run a breadth-first search from an arbitrary vertex, and obtain

$$2(|I(\alpha) \cap C| - 1) = |V(\alpha) \cap C| + |I(\alpha) \cap C| - 2 \le |E(C)|\,. \tag{22}$$

Summing Eq. (21) and Eq. (22) over all connected components, we obtain

$$|V(\alpha)| - |E(\alpha)| + |I(\alpha)| \le (c - c_I) + 2c_I = c + c_I\,.$$

This shows that $n^{-(c+c_I)/2} Z_\alpha$ is combinatorially negligible or combinatorially order 1, and it is combinatorially order 1 if and only if equality holds in the argument. This happens if and only if there is no cycle, multiplicity >2 edges, self-loops, or 2-labeled edges anywhere; and if the graph induced by the multiplicity 1 multiedges is connected. □

With this result in hand, we can now characterize the order-1 vector diagrams with several connected components:

**Corollary B.13.** *Let $\alpha \in \mathcal{A}$ be a vector diagram with $c$ floating components, $c_I$ of which consist only of vertices in $I(\alpha)$. Then $n^{-(c+c_I)/2} Z_\alpha$ is combinatorially order 1 if and only if both the floating components (viewed as one scalar diagram) scaled by $n^{-(c+c_I)/2}$ and the component of the root are combinatorially order 1.*

*Proof.* Definition 4.3 sums across the root and floating components, so we may apply both Lemma 4.10 and Lemma B.12. □

## B.4 Classification of diagrams

**Lemma B.14.** *For all $\sigma \in \mathcal{S}$ and $i \in [n]$, $Z_{\sigma,i} \xrightarrow{d} \mathcal{N}(0, |\mathrm{Aut}(\sigma)|)$. Similarly, for all $\tau \in \mathcal{T}_{\text{scalar}}$, $n^{-\frac{1}{2}} Z_\tau \xrightarrow{d} \mathcal{N}(0, |\mathrm{Aut}(\tau)|)$.*

*Proof.* We prove that the moments $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ match the Gaussian moments and use Lemma 2.3.

Let $q \in \mathbb{N}$ be a constant independent of $n$. First, we expand the product $Z_{\sigma,i}^q$ in the diagram basis using Lemma A.7. Using Lemma 4.10, the only combinatorially order 1 terms occur when there are no cycles, all multiedges have multiplicity 1 or 2, and the multiplicity 2 edges form hanging trees. Any term with an edge of multiplicity 1 disappears when we take the expectation $\mathbb{E}\left[Z_{\sigma,i}^q\right]$, while the diagrams which are entirely hanging trees are equal to ◎ up to combinatorially negligible terms (Lemma 4.7). Further, ◎ has expectation 1, and by Lemma B.5 each of the combinatorially negligible terms has expectation $O(n^{-1/2})$. Thus, $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ equals the number of ways to create hanging trees of double edges, up to a term that converges to 0 as $n \to \infty$.

For each of the $q$ copies of $\sigma$, the single edge incident to the root must be paired with another such edge. This extends to an automorphism of the entire subtree. In conclusion, $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ converges to $|\mathrm{Aut}(\sigma)|^{q/2}$ times the number of perfect matchings on $q$ objects, and we conclude by Lemma 2.4 and Lemma 2.3. The proof for the scalar case is analogous. □

**Lemma B.15.** *If $\tau \in \mathcal{T}$ consists of $d_\sigma$ copies of the subtrees $\sigma \in \mathcal{S}$, then*

$$Z_\tau \overset{\circ}{=} \prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(Z_\sigma; |\mathrm{Aut}(\sigma)|)\,.$$

*For $\rho \in \mathcal{F}_{\mathrm{scalar}}$ with $c$ components and consisting of $d_\tau$ copies of each tree $\tau \in \mathcal{T}_{\mathrm{scalar}}$,*

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

*Proof.* We first expand $h_d(Z_\sigma; |\mathrm{Aut}(\sigma)|)$ in the diagram basis using Lemma A.7 and identify the dominant terms, i.e. those which are combinatorially order 1. As in the proof of Lemma B.14, the combinatorially order 1 terms in each monomial $Z_{\sigma,i}^k$ consist of pairing up copies of the tree $\sigma$:

$$Z_\sigma^k \overset{\circ}{=} \sum_{M \in \mathcal{M}(k)} |\mathrm{Aut}(\sigma)|^{|M|} Z_{k-2|M| \text{ copies of } \sigma}\,,$$

where $\mathcal{M}(k)$ is the set of partial matchings on $k$ objects. Now we use the combinatorial interpretation of Hermite polynomials (Lemma 2.5),

$$\begin{aligned}
h_d(Z_\sigma; |\mathrm{Aut}(\sigma)|) &= \sum_{N \in \mathcal{M}(d)} (-1)^{|N|} |\mathrm{Aut}(\sigma)|^{|N|} Z_\sigma^{d-2|N|} \\
&\overset{\circ}{=} \sum_{N \in \mathcal{M}(d)} (-1)^{|N|} |\mathrm{Aut}(\sigma)|^{|N|} \sum_{M \in \mathcal{M}(d-2|N|)} |\mathrm{Aut}(\sigma)|^{|M|} Z_{d-2|N|-2|M| \text{ copies of } \sigma} \\
&= \sum_{M' \in \mathcal{M}(d)} |\mathrm{Aut}(\sigma)|^{|M'|} Z_{d-2|M'| \text{ copies of } \sigma} \sum_{N \subseteq M'} (-1)^{|N|} \\
&= Z_{d \text{ copies of } \sigma}\,.
\end{aligned}$$

This completes the argument when $\tau$ consists of several copies of a single $\sigma \in \mathcal{S}$. If $\sigma, \sigma' \in \mathcal{S}$ are distinct, using again Lemma A.7 and Lemma 4.10, we can check that

$$Z_{d \text{ copies of } \sigma} \odot Z_{d' \text{ copies of } \sigma'} \overset{\circ}{=} Z_{d \text{ copies of } \sigma \text{ and } d' \text{ copies of } \sigma'}\,.$$

The proof then follows by applying these arguments inductively, and extends analogously to scalar diagrams. □

**Lemma B.16.** *Let $\alpha \in \mathcal{F}$ have $c$ floating components. Let $\alpha_{\circledcirc}$ be the component of the root and $\alpha_{\text{float}}$ be the floating components. Then $n^{-\frac{c}{2}} Z_\alpha \overset{\infty}{=} n^{-\frac{c}{2}} Z_{\alpha_{\text{float}}} Z_{\alpha_{\circledcirc}}$.*

*Proof.* The product $n^{-\frac{c}{2}} Z_{\alpha_{\text{float}}} Z_{\alpha_{\circledcirc}}$ can be expanded in the diagram basis using Lemma B.7. We claim that the only non combinatorially negligible diagram is the one without intersections, which equals $n^{-\frac{c}{2}} Z_\alpha$. When an intersection occurs, it can only be between the root component and a floating component. The new component of the root is at most combinatorially order 1 (this is a property of all connected vector diagrams, Lemma 4.10), so there is an "extra" factor of $\frac{1}{\sqrt{n}}$ from the lost component which makes the intersection term negligible. □

**Lemma B.17.** $\{Z_{\sigma,i} : \sigma \in \mathcal{S}, i \in [n]\} \cup \left\{ n^{-\frac{1}{2}} Z_\tau : \tau \in \mathcal{T}_{\text{scalar}} \right\}$ *are asymptotically independent.*

*Proof.* Fix constants $q, r \in \mathbb{N}$. We proceed by computing the moment of a set of diagrams $\sigma_1, \dots, \sigma_q \in \mathcal{S}$ rooted at $i_1, \dots, i_q \in [n]$ and $\tau_1, \dots, \tau_r \in \mathcal{T}_{\text{scalar}}$:

$$\mathbb{E}\left[ \prod_{p=1}^{q} Z_{\sigma_p, i_p} \prod_{p=1}^{r} n^{-\frac{1}{2}} Z_{\tau_p} \right]. \tag{23}$$

Let $|V| = \sum_{p=1}^{q} |V(\sigma_p)| + \sum_{p=1}^{r} |V(\tau_p)|$ and $|E| = \sum_{p=1}^{q} |E(\sigma_p)| + \sum_{p=1}^{r} |E(\tau_p)|$. Let $q_{\text{distinct}}$ be the number of distinct roots, i.e. the number of distinct elements in $\{i_1, \dots, i_q\}$.

Expanding Eq. (23) gives a sum over embeddings of the diagrams. We will prove that the dominant terms factor across the distinct $(\sigma_p, i_p)$ and $\tau_p$; they correspond to pairing up isomorphic $\sigma_p$ at each distinct root and isomorphic $\tau_p$.

Each nonzero term in the expansion of Eq. (23) equals $n^{-(|E|+r)/2}$ (when every edge appears exactly twice) or $O(n^{-(|E|+r)/2})$ (in general) by Assumption 2.1. We partition the summation based on the intersection pattern as in Definition A.5. For a given intersection pattern, letting $I$ be the union of the images of the embeddings, the number of terms with this pattern is $(1 - o(1)) \cdot n^{|I| - q_{\text{distinct}}}$ because the $q_{\text{distinct}}$ root vertices are fixed. In an embedding with nonzero expectation, every edge appears at least twice, so every non-root vertex is in at least two embeddings. Applying this bound to all of the non-root vertices in $I$,

$$|I| \le q_{\text{distinct}} + \frac{|V| - q}{2}.$$

Multiplying the value of each term times the number of terms, the total contribution of this intersection pattern is

$$n^{|I| - q_{\text{distinct}} - \frac{|E|+r}{2}} \le n^{\frac{1}{2}(|V| - q - |E| - r)}.$$

Since the individual diagrams are connected, the exponent is nonpositive. The dominant terms occur exactly when $|I| = q_{\text{distinct}} + (|V| - q)/2$, equivalently all of the non-root vertices intersect exactly one other non-root vertex. Each edge must occur at least twice, and this condition implies that each edge occurs exactly twice in the dominant terms.

We claim that the only way that each edge and vertex can be in exactly two embeddings is if isomorphic $\sigma_p$ and $\tau_p$ are paired. Indeed, by connectivity of $\sigma_p$ and $\tau_p$, sharing one edge extends to an isomorphism. Furthermore, because non-root vertices must intersect other non-root vertices in the dominant terms, we have that no pairs can be made between $\sigma_p$ and $\tau_{p'}$, or between $\sigma_p$ and $\sigma_{p'}$ which have distinct roots. □

Theorem 4.11 follows from Lemma B.14, Lemma B.15, Lemma B.16, and Lemma B.17. The constant-order joint moments of all the diagrams are summarized into the next theorem which generalizes Theorem 1.1.

**Theorem B.18.** *Suppose that $A = A(n)$ is a sequence of random matrices satisfying Assumption 2.1. For all $\alpha_1, \dots, \alpha_k \in \mathcal{A}$, $i_1, \dots, i_k \in [n]$ and $\beta_1, \dots, \beta_\ell \in \mathcal{A}_{\text{scalar}}$ (allowing repetitions anywhere),*

$$\mathbb{E}\left[\prod_{j=1}^{k} n^{-C(\alpha_j)/2} Z_{\alpha_j, i_j} \prod_{j=1}^{\ell} n^{-C(\beta_j)/2} Z_{\beta_j}\right] = \mathbb{E}\left[\prod_{j=1}^{k} Z^{\infty}_{\alpha_j, i_j} \prod_{j=1}^{\ell} Z^{\infty}_{\beta_j}\right] + O(n^{-\frac{1}{2}}),$$

*where $C(\alpha)$ is the number of floating components of $\alpha$, and where the asymptotic random variables $(Z^{\infty}_{\alpha,i})_{\alpha \in \mathcal{A}, i \in [n]}$ and $(Z^{\infty}_{\beta})_{\beta \in \mathcal{A}_{\text{scalar}}}$ are defined as:*

$$\begin{cases} Z^{\infty}_{\sigma,i} \sim \mathcal{N}(0, |\mathrm{Aut}(\sigma)|) \text{ independently} & \text{if } \sigma \in \mathcal{S} \\ Z^{\infty}_{\tau} \sim \mathcal{N}(0, |\mathrm{Aut}(\tau)|) \text{ independently} & \text{if } \tau \in \mathcal{T}_{\text{scalar}} \\ Z^{\infty}_{\rho,i} = \displaystyle\prod_{\sigma \in \mathcal{S}} h_{d_\sigma}(Z^{\infty}_{\sigma,i}; |\mathrm{Aut}(\sigma)|) \prod_{\tau \in \mathcal{T}_{\text{scalar}}} h_{d_\tau}(Z^{\infty}_{\tau}; |\mathrm{Aut}(\tau)|) & \text{if } \rho \in \mathcal{F} \\ Z^{\infty}_{\rho} = \displaystyle\prod_{\tau \in \mathcal{T}_{\text{scalar}}} h_{d_\tau}(Z^{\infty}_{\tau}; |\mathrm{Aut}(\tau)|) & \text{if } \rho \in \mathcal{F}_{\text{scalar}} \\ Z^{\infty}_{\alpha,i} = Z^{\infty}_{\alpha_0,i} \text{ and } Z^{\infty}_{\beta} = Z^{\infty}_{\beta_0} & \text{if removing hanging double edges} \\ & \text{creates } \alpha_0 \in \mathcal{F} \text{ or } \beta_0 \in \mathcal{F}_{\text{scalar}} \\ Z^{\infty}_{\alpha,i} = Z^{\infty}_{\beta} = 0 & \text{if removing hanging double edges} \\ & \text{is not in } \mathcal{F} \text{ or } \mathcal{F}_{\text{scalar}} \end{cases}$$

## B.5 Handling empirical expectations

Empirical expectations are highly concentrated and the following lemma confirms this. Note that the empirical expectations in the Onsager correction for AMP (Section 5.4) will create floating components in the diagrams of the algorithmic state, but all such diagrams will be negligible.

**Lemma 4.19.** *Let $x$ be a vector diagram expression with asymptotic state $X \in \Omega$. Then as scalar diagrams, $\frac{1}{n}\sum_{i=1}^{n} x_i \overset{\infty}{=} \mathbb{E}[X]$.*

*Proof.* The effect of summing a vector diagram $Z_\alpha = (Z_{\alpha,i})_{i \in [n]}$ over $i$ is to unroot $\alpha$, converting it to a scalar diagram. We prove this operation makes every diagram combinatorially negligible, except for the constant term. For $k \geq 0$ and a vector diagram $\alpha \in \mathcal{A}$:

(i) If $a_n Z_\alpha$ is combinatorially negligible, then $\frac{a_n}{n}\sum_{i=1}^n Z_{\alpha,i}$ is a combinatorially negligible scalar term.

(ii) If $a_n Z_\alpha$ has combinatorial order 1, and the root of $\alpha$ is incident to at least one edge of multiplicity 1, then $\frac{a_n}{n}\sum_{i=1}^n Z_{\alpha,i}$ is a combinatorially negligible scalar term.

Unrooting a vector diagram does not change the number of vertices nor the number of edges. During this operation, the number of vertices in $I(\alpha)$ stays the same if the root is adjacent to an edge of multiplicity 1; otherwise it increases by at most 1. We readily check from the definition that the extra $\frac{1}{n}$ makes the resulting scalar terms combinatorially negligible.

Now let $\widehat{x} \overset{\infty}{=} x$ be the tree approximation. The difference $x - \widehat{x}$ consists of combinatorially negligible terms which stay negligible by part (i) above. The trees in $\mathcal{T}$ become negligible by part (ii) above with the exception of the singleton tree which becomes 1. The singleton has coefficient $\mathbb{E}[\widehat{x}_1] = \mathbb{E}[X]$ since the other trees are mean-zero (Corollary A.3). □

# C High-degree tree diagrams are not Gaussian

We compute that the star-shaped diagram with $\log n$ leaves and the root at a leaf is not Gaussian (its fourth moment is significantly larger than the square of its second moment), demonstrating that care must be taken when studying diagrams of superconstant size.[13] This diagram appears after only $T = O(\log\log n)$ iterations in the recursion

$$x_1 = A\vec{1} \qquad x_{t+1} = (x_t)^2 \qquad x_{T+1} = Ax_T\,.$$

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

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

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

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

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

**Lemma C.1.** *Up to a multiplicative* $\mathrm{poly}(d)$ *factor,* $|\mathcal{M}(d,d,d,d)| \gtrsim 3^d(d!)^2$.

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

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

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

**Claim C.2.** *Let $d, k \in \mathbb{N}$ such that $\frac{d}{k-1}$ is an integer. Counting the matchings between $d$ objects and a subset of $(k-1)d$ objects in $k-1$ groups, as a function of the number of objects matched in each group, the number of matchings is maximized when there are $\frac{d}{k-1}$ matched elements per group.*

*Proof of Claim C.2.* Letting $n_1, \ldots, n_{k-1}$ be the number of matched elements per group, we may directly compute this number as $\prod_{i=1}^{k-1}(d)_{n_i}$ where $(d)_k = d(d-1)\cdots(d-k+1)$ is the falling factorial. When $n_i$ and $n_j$ are replaced by $n_i - 1$ and $n_j + 1$, the ratio of new to old values is

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

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

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

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

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

**Fact C.3** (Stirling's approximation)**.** *Up to a multiplicative* $\mathrm{poly}(d)$ *factor,* $d! \asymp \left(\frac{d}{e}\right)^d$.

Recursing via the same principle,

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

In total,

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