Diagram Analysis of Iterative Algorithms

Chris Jones* Lucas Pesenti†

April 29, 2024

Abstract

We study a general class of first-order iterative algorithms which includes power iteration, belief propagation and Approximate Message Passing (AMP), and many forms of gradient descent. When the input is a random symmetric matrix with i.i.d. mean-0 variance-1 entries, we present a new way to analyze these algorithms using *combinatorial diagrams*. Each diagram is a small graph, and the operations of the algorithm correspond to simple combinatorial operations on these graphs. The diagrams are derived in a generic way, by symmetry-reducing a Fourier basis.

We prove a fundamental property of the diagrams: asymptotically, we can discard all of the diagrams except for the trees. The mechanics of first-order algorithms simplify dramatically as the algorithmic operations have particularly simple and interpretable effects on the trees. We further show that the tree-shaped diagrams are essentially a basis of *asymptotically independent Gaussian vectors*.

The tree approximation property 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 fundamental techniques in the field. We demonstrate the connection with the replica symmetric cavity method by "implementing" heuristic physics derivations into rigorous proofs. We rigorously establish that belief propagation is asymptotically equal to its associated AMP algorithm and we give a new simple proof of the state evolution formula for AMP.

These results apply when the iterative algorithm runs for constantly many iterations. We then push the diagram analysis to a number of iterations that scales with the dimension n of the input matrix. We prove that for debiased power iteration, the tree diagram representation accurately describes the dynamic all the way up to $n^{\Omega(1)}$ iterations. We conjecture that this can be extended up to $n^{1/2}$ iterations but no further. Our proofs use straightforward combinatorial arguments akin to the trace method from random matrix theory.

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

Contents

1 Introduction **4**

1.1 First-order iterations 5

1.2 Belief propagation, AMP, and the cavity method 6

1.3 Our contributions 8

1.4 Related work 11

1.5 Organization of the paper 13

2 Preliminaries **14**

3 The Diagram Basis **15**

3.1 Example of using diagrams 16

3.2 Properties of the diagram basis 17

3.3 Asymptotic algorithmic operations 19

3.4 Perspective: symmetrized Fourier analysis 21

4 Diagram Analysis of $O(1)$ Iterations **22**

4.1 Equality up to combinatorially negligible diagrams 22

4.2 Classification of constant-size diagrams 24

4.3 Tree approximation of GFOMs 26

4.4 Asymptotic Gaussian space 28

5 Belief Propagation, AMP, and the Cavity Method **30**

5.1 Heuristic derivation of the BP-AMP equivalence 31

5.2 Diagram proof of the BP-AMP equivalence 32

5.3 State evolution for AMP algorithms 36

6 Analyzing $\mathrm{poly}(n)$ Iterations **38**

6.1 Combinatorial phase transitions 39

6.2 Analyzing power iteration via combinatorial walks 40

6.3 Counting combinatorial walks 42

References **43**

A Fourier analytic properties **48**

B Derivation of Operations on the Diagram Basis **50**
B.1 General combinatorial principles 50
B.2 Derivation of the asymptotic operations 51
B.3 Repeated-label diagram basis 53

C Omitted Proofs **54**
C.1 Removing hanging double edges 54
C.2 Omitted proofs for Section 4.1 55
C.3 Scalar diagrams 58
C.4 Classification of diagrams 61
C.5 Handling empirical expectations 63

D High-degree tree diagrams are not Gaussian **64**

1 Introduction

If $A \in \mathbb{R}^{n \times n}$ is a matrix and $\vec{1} \in \mathbb{R}^n$ denotes the all-ones vector, a fundamental observation is that the iterates of power iteration $A^t\vec{1}$ can be expressed as a sum over walks of length t on $[n]$. For example, for any $i \in [n]$, by the definition of matrix multiplication,

$$(A^4\vec{1})_i = \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\sum_{m=1}^{n} A_{ij}A_{jk}A_{k\ell}A_{\ell m}\,.$$

The expression $A^t\vec{1}$ can be viewed as a *linear* iteration in which the matrix A is applied t times to the input $\vec{1}$. In fact, a generalization of this formula also holds for *nonlinear* iterations. Consider the iteration $x_{t+1} = \sigma(Ax_t)$ initialized at $x_0 = \vec{1}$ where $\sigma : \mathbb{R} \to \mathbb{R}$ is a univariate polynomial applied separately to each coordinate. For example, taking $\sigma(x) = x^2$, the first two iterates are:

$$(\sigma(A\vec{1}))_i = \sum_{j_1,j_2=1}^{n} A_{ij_1}A_{ij_2}$$

$$(\sigma(A\sigma(A\vec{1})))_i = \sum_{j_1,j_2=1}^{n}\sum_{k_1,k_2=1}^{n}\sum_{\ell_1,\ell_2=1}^{n} A_{ij_1}A_{ij_2}A_{j_1k_1}A_{j_1\ell_1}A_{j_2k_2}A_{j_2\ell_2}\,.$$

Instead of summing over walks, we sum over "branching walks" on $[n]$ which have higher-degree branching nodes.

We identify a fundamental property of these formulas when A is a **random symmetric matrix with i.i.d. mean-0 variance-1 entries**. In the limit $n \to \infty$, only certain types of walks on $[n]$ contribute to the asymptotic state.

- For the linear iteration, $A^t\vec{1}$ is asymptotically approximated by the walks which may backtrack but otherwise must be self-avoiding (they cannot revisit vertices except for going backwards along the current path).
- For nonlinear iterations, x_t is asymptotically approximated by the branching walks which may have "doubled subtrees" but are otherwise self-avoiding trees.

We call this the *tree approximation* to x_t. The high-level justification of this approximation is that a typical walk on $[n]$ is self-avoiding, so self-avoiding walks are the combinatorially dominant terms. Walks with backtracking steps or doubled subtrees are "Itô-type" terms which also turn out to contribute to the asymptotic state, despite being combinatorially non-dominant.

The consequences of this simple property appear vast, as we will attempt to make clear in the rest of the paper. In the remainder of this introduction, we give some background on first-order iterations, then we describe our contributions in more detail.

1.1 First-order iterations

We study iterative algorithms which alternate multiplying by a matrix $A \in \mathbb{R}^{n \times n}$ and applying a componentwise function f_t. A concrete class that we analyze are *general first-order methods* (GFOM, Definition 4.13) as defined by Celentano, Montanari, and Wu [CMW20, MW22b]. Iterations of this type are simple, widespread, practically efficient, and incredibly powerful. In optimization, it captures power iteration and many types of gradient descent (see [CMW20, GTM+22] for concrete examples of such optimization algorithms). In statistics, it encapsulates belief propagation and message passing algorithms, and the functions f_t are sometimes called *denoisers*. In machine learning, a neural network alternates multiplying by a *weight matrix* A and applying the *nonlinearities* f_t. In signal processing, f_t are *channels* which modulate the signal. First-order iterations are some of the most fundamental algorithms of the current age, both in theory and in practice.

Analyzing even simple iterative methods can be challenging. These algorithms are expressive enough to solve a wide range of tasks, and they have a recursive nature which makes their behavior difficult to analyze. That being said, some key insights were made by statistical physicists in the 1980s studying *belief propagation*, a special class of first-order algorithms. A series of works and a landmark book by Mézard, Parisi, and Virasoro [MPV87] developed a deep theory for *spin glass* models in physics—surprisingly, much of their work is essentially equivalent to analyzing the convergence of variants of the belief propagation algorithm.[1]

The foundational techniques introduced by the physicists in that era proved extremely influential. In the 40 years since then, they have led to hundreds of results and have served as some of the most fundamental tools in the statistical physics toolbox. They have also been extended significantly to understand deep structural properties of random optimization problems, and recently there has been a surge of interest coming from statistical inference and machine learning. See the surveys [MMZ01, ZK16, Gab20], the book [MM09], and the 40 year retrospective [CMP+23].

Viewed from the algorithms perspective, we highlight two major contributions that have developed from this long line of work. First, on the quantitative side, the physical methods give a *complete* mathematical description of how the belief propagation algorithm evolves, now known as the *state evolution* formulas (this is described in more detail in the next subsection). State evolution can be used to compute essentially any desired performance measure of the algorithm, up to $o(1)$ error as the dimension of the input n goes to infinity. Second, on the conceptual side, the methods demonstrate the incredible depth of mathematical formulas and insights that simple iterative algorithms can access. Indeed, belief propagation is believed to be an *optimal* algorithm for many statistical tasks.

However, despite the incredible success of these methods, they are not the end of the story. The largest unsettled point is that the techniques used by physicists are not mathematically rigorous! In a typical physics derivation, almost all the steps are mathematically valid, but at certain points it may be necessary to assume that approximation errors are negligible,

[1] The relationship arises because the fixed points of certain belief propagation algorithms are equal to the candidate approximations to the *free energy* proposed by the physicists.

that a limit exists, etc. Sometimes these assumptions can be rigorously explained later by mathematicians, but in the current context, it has appeared that the physicists know something that the mathematicians do not. The impressive predictions of the non-rigorous *cavity method* [MPV86] and the related *replica method* [Par79, Par80] (which derive the same results in principle [MP03]) have been repeatedly confirmed, but even after 40 years, rigorous mathematical proofs usually cannot follow the heuristic derivations, and they often suffer from very significant technical complications which cannot match the elegance of the physicists' techniques. See the related work in Section 1.4 for more details.

Another shortcoming of existing work is that the methods apply naturally to belief propagation, but not in a satisfying way to other first-order iterations. Although belief propagation algorithms are provably optimal first-order methods for some statistical tasks [BKM+19, CMW20, MW22b, BCMS23], they are not used in practical machine learning nearly as much as other, simpler first-order methods such as stochastic gradient descent.[2] [3]

1.2 Belief propagation, AMP, and the cavity method

Let us review belief propagation (BP) and its state evolution as predicted by physicists.

BP is a first-order iterative algorithm which attempts to recover a hidden statistical signal by iteratively sending messages between nodes of an underlying graph [YFW03, KF09, MM09]. The BP framework can be broadly and flexibly applied to statistical inference and average-case optimization. Typically, we are attempting to compute an n-dimensional vector $x \in \mathbb{R}^n$ (playing the role of the hidden signal) using an input matrix $A \in \mathbb{R}^{n \times n}$ (playing the role of the observations). The messages are passed on the graph given by the nonzero entries of A. In the setting relevant for this paper, the underlying graph is the *complete graph* and all entries of A are on the same scale.

The iteration occurs on a set of messages $m_{i \to j}$ for $i, j \in [n]$ which conceptually are "the belief of vertex i about its own value, disregarding j". It then proceeds by aggregating the messages from neighboring vertices and updating the vertex's belief accordingly:

$$m^0_{i \to j} = 1, \quad m^{t+1}_{i \to j} = \psi\left(\sum_{\substack{k=1 \\ k \neq j}}^{n} A_{ik} m^t_{k \to i}\right), \tag{1}$$
$$m^{t+1}_i = \widetilde{\psi}\left(\sum_{k=1}^{n} A_{ik} m^t_{k \to i}\right),$$

where $\psi : \mathbb{R} \to \mathbb{R}$ and $\widetilde{\psi} : \mathbb{R} \to \mathbb{R}$ are fixed univariate functions used to update the belief. The vector $m^t \in \mathbb{R}^n$ is the output of the algorithm. For example, the choice $\psi(x) =$

[2] Why isn't belief propagation widely used for practical machine learning if it is a statistically optimal algorithm? One downside to belief propagation is that it is not *robust*—the algorithm is specifically tailored to the i.i.d. model for the input. If the model assumption is violated, either due to adversarial perturbations or simply a mismatch between real-world data and the model, then the performance quickly deteriorates [RSFS19, CZK14, VSR+15, MKTZ15].

[3] We study general first-order iterations in this paper but we clarify that the techniques do not immediately apply to stochastic gradient descent.

$\widetilde{\psi}(x) = \tanh(\beta x + h)$ for parameters $h, \beta > 0$ is known as the *replica symmetric BP for the Sherrington{Kirkpatrick model* in physics [ZK16, Section IV.E].[4]

An important point about belief propagation is that its definition "assumes the underlying graph is a tree". The functions $\psi, \widetilde{\psi}$ are chosen as if the messages $m_{k \to i}$ collected from one's neighbors are *independent* (which would be the case if A was actually supported on a tree). When this algorithm is run on a graph which is not a tree, the assumption of independence does not hold, but it is hoped that the different messages are only weakly correlated. This hope seems especially far-fetched in our setting because the underlying graph is the complete graph—not even close to a tree! Yet, here for example, BP runs successfully for the Sherrington–Kirkpatrick model when β is sufficiently small.

An alternative description of this property is that the belief messages $m_{i \to j}$ are computed in a **non-backtracking** way: computing $m_{i \to j}$ uses all of the messages $m_{k \to i}$ except for the reverse message $m_{j \to i}$.

Asymptotic description of BP. Physical methods predict a *complete* description of the messages $m^t_{i \to j}$ and the outputs m^t_i up to $o(1)$ error with respect to the input size n, when A is a random symmetric matrix with i.i.d. entries. In the iteration above, they say that the entries of m^t should be approximately independent and identically distributed as

$$m^t_i \quad \widetilde{\psi}(Z_t)\,, \qquad \text{where} \qquad Z_t \quad \mathsf{N}(0, \sigma_t^2)\,, \quad \sigma_1^2 = 1\,, \qquad \sigma_{t+1}^2 = \mathsf{E}[\psi(Z_t)^2]\,. \tag{2}$$

This result can be approached non-rigorously using the *cavity method* (see [MP03] and [MM09, Part V]). The cavity method extends the tree-based definition of belief propagation into concrete mathematical predictions. Carrying out the replica symmetric cavity approach here, we **assume** that the non-backtracking summation $\sum_{k=1, k \neq j}^n A_{ik} m^t_{k \to i}$ appearing in Eq. (1) has incoming terms $m^t_{k \to i}$ which are independent (by symmetry, they are also identically distributed). Then, we appeal to the central limit theorem to deduce (A is scaled so that $\mathsf{E}[A_{ik}] = 0$ and $\mathsf{E}[A_{ik}^2] = \frac{1}{n}$)

$$\sum_{\substack{k=1 \\ k \neq j}}^n A_{ik} m^t_{k \to i} \quad \mathsf{N}\left(0, \mathsf{E}\left[(m^t_{k \to i})^2\right]\right)\,.$$

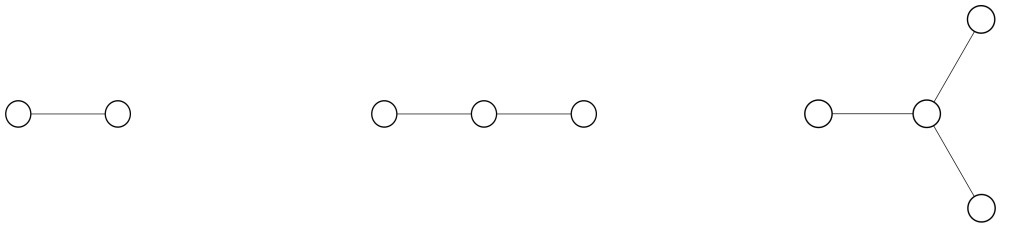
From here, we get that the outgoing message satisfies $m^t_{i \to j} \quad \psi(Z_t)$ for $Z_t \quad \mathsf{N}(0, \sigma_t^2)$ with σ_t^2 defined by the recurrence in Eq. (2). Using a similar argument, we get $m^t_i \quad \widetilde{\psi}(Z_t)$.

Another physics-supported insight is that the BP iteration can be replaced by an asymptotically equivalent *Approximate Message Passing* (AMP) algorithm. The AMP iteration reduces the recursion on the n^2 variables $m^t_{i \to j}$ to a recursion on the n variables m^t_i which can be implemented more efficiently. The correspondence between BP and AMP algorithms is described in Section 5. Eq. (2) is known as the *state evolution* formula for the AMP algorithm. In physics terminology, the AMP iterates are known as the "TAP equation iterates".

[4] In this case, the iteration attempts to compute the marginals of the Gibbs distribution $p(x) \propto e^{H(x)}$ for $x \in \{-1, 1\}^n$, where $H(x) = \beta \sum_{i,j=1}^n A_{ij} x_i x_j + h \sum_{i=1}^n x_i$ is the Hamiltonian of the Sherrington–Kirkpatrick model with inverse temperature β and external field h.

1.3 Our contributions

Diagram analysis. We present a new way to analyze first-order algorithms using *combinatorial diagrams*. Each diagram is a small graph that represents a particular symmetric low-degree polynomial in the input matrix A. For example, here are three diagrams and the polynomials that they represent.



$$\sum_{\substack{i,j=1\\ \text{distinct}}}^{n} A_{ij} \qquad \sum_{\substack{i,j,k=1\\ \text{distinct}}}^{n} A_{ij}A_{jk} \qquad \sum_{\substack{i,j,k,\ell=1\\ \text{distinct}}}^{n} A_{ij}A_{ik}A_{i\ell}$$

We will express the algorithmic state as a linear combination of diagrams, in which case we say that the state is written "in the diagram basis". The idea of the diagrammatic approach is that the operations of the algorithm correspond to simple combinatorial operations on the diagram basis. Thus we reduce analytic questions to combinatorial questions on these small graphs. The definition of the diagram basis will be given in Section 3.

Several past works have used a diagrammatic approach to analyze iterative algorithms; see Section 1.4. Our main contribution is pinpointing and taking advantage of the many useful properties exhibited by our particular choice of diagrams.

When an iterative algorithm with polynomial non-linearities is run for a constant number of iterations, its state can be expressed as a sum of constantly many diagrams with constant size and constant coefficients. For starters, we would like to know, what quantity does each individual diagram express?

Our first main theorem classifies the constant-size diagrams (Theorem 4.11 in Section 4.2). It turns out that the constant-size diagrams are essentially a basis of *asymptotically independent Gaussian random variables*. In the examples above, the three polynomials are asymptotically independent Gaussians when the entries A_{ij} are i.i.d. ± 1 or sampled independently from any other mean-0 variance-1 distribution.[5] The following describes the classification of diagrams which are unrooted graphs; we will also show a similar classification for rooted graphs.

- A diagram with a *cycle* is asymptotically *negligible*.
- A diagram which is a *tree* is asymptotically an *independent Gaussian variable*.
- A diagram which is a *forest* is asymptotically a *multivariate Hermite polynomial* in these Gaussian variables. The degree of the Hermite polynomial in each tree equals the multiplicity of the tree.

[5]The asymptotic Gaussians are *universal* i.e. they don't depend on the distribution of the entries A_{ij}.

In summary, the diagram expansion is asymptotically the Hermite polynomial expansion of the algorithm with respect to a set of "tree Gaussians" extracted from A. We also see that the asymptotic approximation to an expression written in the diagram basis consists of throwing away all diagrams with cycles. This is what we call the *tree approximation*.

Our second main theorem (Theorem 4.16 in Section 4.3) proves that the errors in the tree approximation do not propagate for a wide class of nonlinear iterations. That is, the cyclic diagrams can be ignored at all times—the algorithm can be completely seen as operating on the trees. Showing that errors do not propagate for "random dynamical systems" cleanly addresses an open problem raised in a seminal paper by Donoho, Maleki, and Montanari [DMM10, Section III.E]. We study a null model without any hidden signal, although we expect that the tree approximation can also be used in the presence of a planted signal, since the null model often captures the difficulty of the problem.

When restricted to the tree-shaped diagrams, the operations of a first-order iteration have much simpler combinatorial effects. In particular, multiplying by the matrix A corresponds in the asymptotic diagram space to summing a "forward step" and a "backward step" on the trees (Section 3.3). Despite the powerful recursive nature of these algorithms, our results demonstrate that a much simpler description of the algorithmic trajectory is possible which discards the complicated, negligible actions taking place on the cyclic diagrams. The fact that a generic nonlinear iteration admits this approximation is extremely striking.

Making statistical physics methods rigorous. It turns out that the properties of the diagrams match the heuristic assumptions used to study belief propagation very closely. For example, consider the message passing update equation,

$$m_{i \,!\, j}^{t+1} = \psi\left(\sum_{\substack{k=1 \\ k \neq j}}^{n} A_{ik} m_{k \,!\, i}^{t}\right).$$

The summation $\sum_{k=1}^{n} A_{ik} m_{k \,!\, i}^{t}$ (which differs from above in that $k = j$ is allowed) is equal to the matrix multiplication of A times $m_{k \,!\, i}^{t}$.[6] By the diagrammatic properties, there are only two asymptotically non-negligible terms: the "forward step" and the "backward step" (Section 3.3). We will show that the "backward step" is just equal to the backtracking term $A_{ij} m_{j \,!\, i}^{t}$. Thus we can see that the crucial part of the belief propagation algorithm in which we remove the backtracking message $j \;!\; i$ is equivalent to using only the "forward step" in the diagram basis. Furthermore, the "forward step" is easy to understand: it is always an asymptotic Gaussian vector!

Using the diagrams, we are able to implement in a completely rigorous way two heuristic arguments based on the cavity method. First, the equivalence of belief propagation and AMP is typically argued heuristically [DMM09, ZK16]. We show (Theorem 5.1 in Section 5) that belief propagation and its corresponding AMP iteration are asymptotically equivalent in a formal sense by going through the standard heuristic derivation and giving a line-by-line diagrammatic justification of each non-rigorous approximation. We define a combinatorial

[6] $m_{k \,!\, i}^{t}$ is indexed by two coordinates i, k. This matrix multiplication is computed along the k coordinate.

notion of equality $\overset{1}{=}$ (Section 4.1) which can be used in place of the heuristic in the physical argument.

Second, we provide a short diagrammatic proof of the state evolution formula for AMP algorithms (Theorem 5.2). In this result and the previous one, we can deduce very strong forms of convergence, although we make the technical assumption that the nonlinearities are polynomials.

These results combine to show that the asymptotic trajectory of the belief propagation iterates follows the state evolution formula (at all temperatures). Note that this does not a priori say anything about whether or not the iteration converges to a fixed point or what kind of fixed point it might reach. Convergence to the "correct" fixed point is a further assumption of physics which we do not consider here.

We also take a look in Example 5.11 at the *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 "extracts" a Brownian motion from the input.

Pushing the tree approximation farther. So far, we have only been discussing diagrams of constant size and first-order algorithms with a constant number of iterations, for which the diagram basis works very well. This "short-time setting" may seem like a major restriction, but in fact many tasks for first-order algorithms can be $(1 \quad \varepsilon)$-achieved within constantly many steps, and constantly many iterations is the setting of a large majority of previous works in this area.

Can we prove that the tree approximation holds until convergence, whenever that might be (maybe never)? We show in Section 6 that some care needs to be taken when addressing this question. To fix a demonstrative setting, we consider a simple belief propagation algorithm, debiased power iteration (equivalently, power iteration on the non-backtracking walk matrix). For this iteration, by a simple counting argument, all of the diagrammatic properties extend up to $t \quad \log n / \log \log n$ iterations, which matches a threshold found in previous work analyzing various instantiations of AMP [RV18, CR23]. We prove (Theorem 6.2) that in fact the tree approximation for debiased power iteration holds until the much later time $t \quad n^{\delta}$ for some $\delta > 0$, which goes much beyond the natural boundary suggested by a naive analysis. Finally, we identify a further threshold at $t \quad \overline{n}$ iterations beyond which the tree approximation we use seems to break down.

Proof techniques. We derived the diagrams as a "symmetrized Fourier basis" when the algorithm is viewed as a function from $A \; 2 \; \mathsf{R}^{n \quad n}$ to $x_t \; 2 \; \mathsf{R}^n$. This approach has been successfully used in the average-case analysis of Sum-of-Squares algorithms, where the elements of the symmetrized Fourier basis are known as *graph matrices* (see the related work in Section 1.4). The Fourier analytic principles are described in more detail in Section 3.4 where it is also made clear how to generalize the techniques to other types of i.i.d. inputs. The symmetrization technique is a perfect philosophical match with the prototypical analysis strategy of statistical physics. We reduce an n-dimensional problem to a finite-dimensional one using symmetry over permutations of n, and then we perform the key arguments in the finite-dimensional diagram space.

Our proof proceeds by developing exact combinatorial formulas for operations in the diagram basis. The proof is almost ludicrously direct. We maintain a complete representation of the algorithmic state throughout the entire algorithm, without taking the limit $n \to \infty$ until the final step. We estimate the magnitude of every term in the diagram expansion using an appropriate combinatorial quantity (see Lemma 4.2 and Definition 4.3).[7] We also devise new variants of diagrams as needed to express the terms that arise. Once the low-level definitions are in place, the overall argument consists of short and direct combinatorial lemmas. As we demonstrate, this approach is extremely flexible and strong.

Pushing the diagram analysis techniques to $n^{\Omega(1)}$ iterations requires more advanced combinatorial arguments. Conceptually, the proof of the result is straightforward: we just sum up the non-tree error diagrams and show that they are small. Technically, it is not so simple, as we employ precise encoding arguments to count the number of terms with each magnitude. The error terms we need to count correspond to even traversals with a fixed number of excess edges (see Definition 6.3). This type of combinatorial analysis is similar to the trace method from random matrix theory [Bor19]. More generally, the diagram representation can be seen as a symbolic trace method.

1.4 Related work

Diagrammatic methods. The diagram basis is inspired by techniques introduced to analyze *Sum-of-Squares* algorithms. The Sum-of-Squares algorithm is a powerful meta-algorithm for combinatorial optimization and statistical inference [RSS18, FKP19]. A recent technology developed for Sum-of-Squares algorithms is *graph matrix analysis*. Graph matrices are a Fourier basis for matrix-valued functions of a random matrix A, in the same way that our diagram basis is a Fourier basis for vector-valued functions of A. 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]). Core analytical ideas were subsequently isolated by Ahn, Medarametla, and Potechin [AMP20] and Potechin and Rajendran [PR20, PR22]. Graph matrix analysis was developed further in several more works [GJJ+20, RT23, JPR+21, JP22, Jon22, JPRX23, KPX24] in which one of the authors took part. They perform sophisticated combinatorial analyses on diagrams (called "shapes") which inspired some of the definitions and techniques of the current work.

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 which introduced the "TAP equations" [TAP77]. It appears that no previous works have identified the asymptotic basis of independent Gaussians vectors, although several pinpoint the importance of some version of tree-shaped diagrams. Bayati, Lelarge, and Montanari [BLM15] use a diagrammatic approach to prove universality of AMP algorithms, showing that the Onsager correction corresponds to a backtracking term. Montanari and Wein [MW22a, Section 3.2] use an orthogonal diagram basis to analyze AMP in the setting of rank-1 matrix estimation. Ivkov and Schramm [IS23] analyze AMP algorithms using the

[7] This also gives a way to identify lower-order error terms in the state. See Remark 4.17.

simpler *non-orthogonal* repeated-label basis (see Appendix B.3). Another similar class of diagrammatic techniques are *tensor networks*, e.g. [MW19].

We note that both Sum-of-Squares and AMP are part of an emerging class of *low-degree algorithms*, which are algorithms whose output can be approximated by low-degree multivariate polynomials in the input. Analyzing degree-d polynomials roughly corresponds to analyzing first-order iterations with d iterations or degree-d Sum-of-Squares relaxations (although this is not 100% clear for either Sum-of-Squares [HKP+17] or AMP [MW22a]). Low-degree methods are conjectured to be *optimal* for many average-case statistical and optimization tasks [KWB19, BAH+22, CM22]. Fourier analysis is a very promising tool for these algorithms, since it expresses them with respect to a natural basis. Several recent works have made explicit connections between AMP, Sum-of-Squares, and low-degree polynomials [MW22a, IS23, SS24a, SS24b]. In comparison, we do not make a direct link but instead bridge the underlying mathematical tool of Fourier analysis.

Statistical physics and the cavity method. For an introduction to statistical physics in computer science, we recommend the surveys [MMZ01, ZK16, Gab20, CMP+23].

Although the techniques of physics are broadly used on many types of random ensembles, the results in this paper will be restricted to the case of dense random matrices, for example with i.i.d. $N(0,1)$ or Rademacher entries. This gives rise to the *Sherrington–Kirkpatrick* (SK) model in physics. The replica method and the cavity method were originally developed to compute the free energy of the SK model [Par79, Par80, MPV86]. The replica method is the original and the more well-known of the two, whereas the cavity method is conceptually simpler; to an extent, the methods have the same power [MP03]. In this paper, we only consider the cavity method.

Rigorously proving arguments based on the cavity method has been a major challenge for mathematicians that has spanned decades [CMP+23]. Formalizing the method into an all-encompassing mathematical statement does not seem like an easy task, as it is a problem-solving technique employed in diverse situations. Prior works verify the *predictions* of the methods, but they largely cannot touch the methods themselves. Two landmark tour-de-force proofs of the *Parisi formula* for the SK model were developed by Talagrand [Tal06, Tal10] and Panchenko [Pan13]. Both works implement analytic forms of the cavity calculation ([Tal10, Section 1.6] and [Pan13, Section 3.5]). In comparison to existing work, our approach identifies combinatorial aspects of the cavity method in a way that directly justifies the method in practice. Our results tantalizingly show that generic validation of the methods may be possible, perhaps even with simple techniques. However, note that in this work, we do not prove anything specific about SK or any other model.

A work of Coja-Oghlan, Krzakala, Perkins, and Zdeborová makes progress on a generic validation of the cavity method for sparse (a.k.a diluted) models in the replica symmetric regime [BN06, CKPZ17]. In sparse inputs, the entries of A have moments depending on n, e.g. the Erdős–Rényi random graph $G(n,p)$ with $p = o(1)$ or $p = \Theta(1/n)$. Their methods are completely different from ours, and the sparse/dense settings are somewhat hard to compare.[8]

[8] Message passing algorithms with constantly many iterations on sparse models can be simpler due to the

Belief propagation and AMP. Statistical physics has a long relationship with belief propagation, although the algorithmic viewpoint is traditionally not the focus. Belief propagation originates in computer science and statistics from Pearl [Pea88] though early ideas were present in physics as far back as Bethe [Bet35]. Recently, AMP algorithms, which are equivalent to belief propagation in the "dense" setting (Section 5), have ushered in a renaissance in algorithmic statistics. Surveys and notes with perspectives coming from different fields are [YFW03, MM09, KF09, ZK16, Gab20, FVRS22, ZY22].

The asymptotic description of the belief propagation algorithm is known in the setting of AMP as the *state evolution* formulas. A fairly modern form of state evolution was predicted for AMP by [Kab03, DMM09] using numerical simulations and the non-rigorous cavity-based approach. These formulas were first rigorously proven by Bayati and Montanari [BM11] using a conditioning technique of Bolthausen [Bol14]. This technique has since been extended to prove state evolution for many variants of AMP [JM13, MRB17, Tak19a, BMN20, Tak19b, AMS21, Tak21, 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 works push the analysis of AMP in some settings to $t = o(\log n / \log\log n)$ iterations [RV18, CR23] and incredibly $t = \widetilde{\Omega}(n)$ iterations [LW22, LFW23, LW24]. This last line of work is intriguing considering that our approach seems to break down at $t \approx \sqrt{n}$ (Section 6.1).

The perspective that we take is 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 of the input A which asymptotically approximates the iterates almost surely. This general approach both provides more information and has increased potential for generalization beyond i.i.d. inputs.

Looking at first-order iterations beyond 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] achieve rigorous results.

1.5 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 useful notion $\overset{1}{=}$, then we prove two central theorems, classification of the diagrams (Theorem 4.11) and the tree approximation for GFOMs (Theorem 4.16). We also define an asymptotic Gaussian space which contains the limiting objects. In the next Section 5, we demonstrate the connection with the cavity method, proving diagrammatically that belief propagation follows the state evolution formula. Section 6 investigates algorithms

fact that the message passing graph is locally treelike around a $1 - o(1)$ fraction of vertices, and hence the iteration behaves in many ways as if it is occurring on a tree. However, analysis beyond constantly many iterations or beyond the replica symmetric regime seems poorly understood.

running for a large number of iterations and proves that debiased power iteration still admits the tree approximation for $n^{\Omega(1)}$ iterations.

Among the appendices, Appendix B contains useful general combinatorial principles for diagram analysis and a derivation of the algorithmic operations on diagrams. Appendices A, C, D contain omitted proofs and calculations.

Acknowledgements. We are pleased to acknowledge Carlo Lucibello, Enrico Malatesta, and the members of the Bocconi Computing Sciences Department for discussions on physics. Goutham Rajendran and Giorgi Kanchaveli provided comments on a draft. This research was supported in part by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 834861). CJ is also a member of the Bocconi Institute for Data Science and Analytics (BIDSA).

2 Preliminaries

Our results will apply universally to a Wigner random matrix model (they hold regardless of the specific choice of μ, μ_0):

Assumption 2.1 (Assumptions on matrix entries)**.** *Let μ and μ_0 be two distributions on $\mathbb{R}$ such that*

(i) all moments of μ_0 are finite and independent of n;

(ii) all moments of μ are finite and independent of n; $\mathbb{E}_{X\sim\mu}[X] = 0$ 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.

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

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

- *We write $X_n \xrightarrow{a.s.} Z$ 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 $X_n \xrightarrow{d} Z$ if X_n converges to Z in distribution, i.e. for every real-valued bounded continuous function f, $\mathbb{E} f(X_n) \to \mathbb{E} f(Z)$.*

We will derive convergence in distribution by working with the moments of the random variables.

Lemma 2.3 (Method of moments [Dur19, Theorem 3.3.26])**.** *Let $(X_n)_{n\geq 0}$ and Z be random variables such that for all constant $q \in \mathbb{N}$,*

$$\mathbb{E}\left[X_n^q\right] \to_{n\to\infty} \mathbb{E}\left[Z^q\right].$$

Suppose that Z has a Gaussian distribution. Then $X_n \xrightarrow{d} Z$.

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

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

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

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

where $PM(q)$ 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 $x \in \mathbb{R}$,*

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

where $M(k)$ 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 $x \in \mathbb{R}$,*

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

where $M(k_1,\dots,k_\ell)$ is the set of (partial) matchings on $k = k_1+\cdots+k_\ell$ objects divided into ℓ 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 $(Z_1,\dots,Z_k)$ be a centered Gaussian vector. Then for all smooth $f:\mathbb{R}^k \to \mathbb{R}$,*

$$\mathbb{E}[Z_1 f(Z_1,\dots,Z_k)] = \sum_{i=1}^k \mathbb{E}[Z_1Z_i]\,\mathbb{E}\left[\frac{\partial f}{\partial z_i}(Z_1,\dots,Z_k)\right].$$

3 The Diagram Basis

In this section, we define the diagram basis and state their key properties on a high level. We delay formal statements and proofs to later sections.

- In Section 3.1, we give an example.
- In Section 3.2, we define the concept of diagram and describe their behavior both for fixed n and in the limit $n \to \infty$.

- In Section 3.3, we show the utility of the diagram basis by observing the simple effects of algorithmic operations on the diagram representation.
- In Section 3.4, we explain how the diagram basis can be derived from standard discrete Fourier analysis.

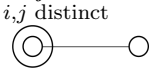
3.1 Example of using diagrams

We show how to represent 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. 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. Multigraphs may include multiedges 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.

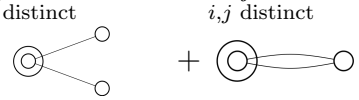
The vector $A\vec{1}$ will be represented by the graph consisting of a single edge, with one of the endpoints being the root:

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



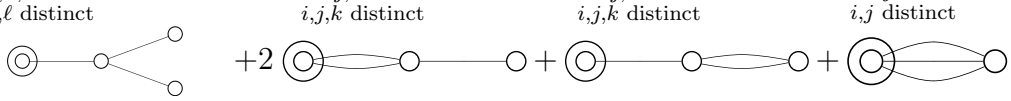
where the second equality uses the assumption that A has zero diagonal. Now to apply the square function componentwise, we can decompose:

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



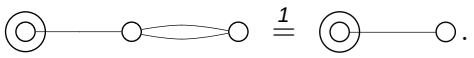
Moving on, we apply A to this representation by casing on whether the new index i matches one of the previous indices. We group terms together using the symmetry of A and the fact that $A_{ii} = 0$.

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

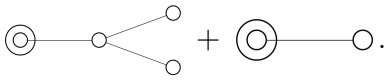


This is the (non-asymptotic) diagram basis representation of $A(A\vec{1})^2$. From the examples, one can see that in each diagram, the root is fixed to the vector index i, and we sum over all possible distinct labels for the non-root vertices.

In the limit $n \to \infty$, only some of the terms contribute to the *asymptotic* diagram basis representation. We will see that asymptotically, *hanging* double edges can be removed from every diagram (Combinatorial Principle 2 in Appendix B.1). Hence, as $n \to \infty$ the third diagram in the representation above satisfies



As we will see, 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. To conclude,

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

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 mutually independent. It can be verified numerically that for large n, $A(A\vec{1})^2$ indeed 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 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 diagrams.*

Each diagram represents a specific vector.

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

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

These diagrams represent vector quantities; unrooted graphs can be used to represent scalar quantities, which we introduce in Section 4.2.[9] Among all diagrams, the ones corresponding to trees play a special role. They will constitute the *asymptotic 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 diagram)**.** *A proper diagram is a diagram with no multiedges or self-loops (i.e. a rooted simple graph).*

[9] Graphs with multiple roots can be used to represent matrices and tensors, although we will not need those here. See also Section 6.

The diagrams Z_α turn out to have many beautiful and useful properties as we now describe. For *proper* diagrams $\alpha \in A$, the following properties of Z_α hold non-asymptotically i.e. for arbitrary n (see Section 3.4 for an extended discussion and Appendix A for the proofs):

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

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

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

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

Now we turn to the asymptotic properties. Formal statements are found in Section 4, and additional intuition is given in Appendix B.1. The constant-size diagrams $(Z_\tau)_{\tau \in T}$ exhibit the following key properties in the limit $n \to \infty$ and with respect to the randomness of A.

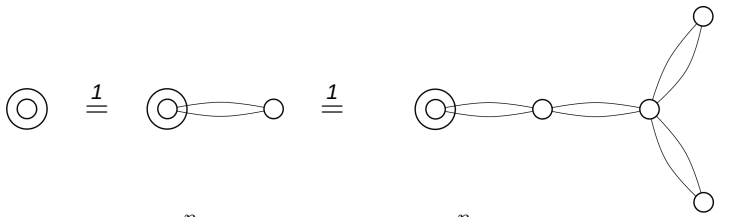
(i) For any $\tau \in T$ (the tree diagrams), the coordinates of $Z_\tau \in \mathbb{R}^n$ are asymptotically independent and identically distributed.

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

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

The remaining diagrams not in T can be understood using the further asymptotic properties:

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



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

(v) For any *connected* $\alpha \in A$, if removing the hanging trees of double edges from α creates a diagram in T, then by the previous property, Z_α is asymptotically equal to that diagram. If the result is not in T, then Z_α 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 T, the expression admits an *asymptotic* diagram basis representation of the form

$$x \overset{1}{=} \sum_{\tau \in T} c_\tau Z_\tau ,$$

for some coefficients $c_\tau \in \mathbb{R}$. We call this the *tree approximation* to x.

Mathematically, this is the Hermite polynomial expansion of the algorithm with respect to the Gaussian random variables $(Z_\sigma)_{\sigma \in S}$. The underlying probability space is a system of independent Gaussian vectors $(Z_\sigma)_{\sigma \in S}$ with one vector for each rooted tree in S.

3.3 Asymptotic algorithmic operations

Generalizing the example in Section 3.1, the operations of a first-order algorithm can be represented in the diagram basis. When restricted to the asymptotic diagram basis, the combinatorial operations of the algorithm become significantly simpler. We maintain an expression of the algorithm state $x_t \in \mathbb{R}^n$ asymptotically in the basis of treelike diagrams,

$$x_t \overset{1}{=} \sum_{\tau \in T} c_\tau^t Z_\tau ,$$

for some coefficients $c_\tau^t \in \mathbb{R}$. In this section, we study the following two core operations of first-order algorithms on a vector $x \in \mathbb{R}^n$.

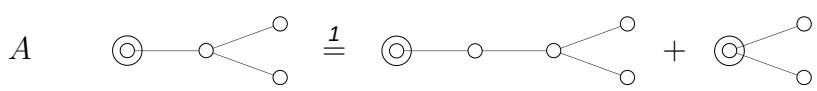
(i) Multiply x by A.

(ii) Update x to $f(x)$ where $f : \mathbb{R} \to \mathbb{R}$ is a polynomial function applied componentwise.

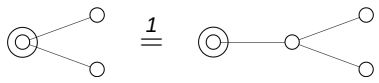
These operations have the following explicit asymptotic effects on the tree basis (see Appendix B for the derivation).

(i) **A takes forward and backward steps on the trees.**

If $\sigma \in S$, then AZ_σ is asymptotically the sum of the diagrams σ^+ and σ^- obtained by extending and contracting the root by one, respectively. For example:



If $\tau \in T_n S$, then AZ_τ is asymptotically only the τ^+ term. For example:

A  $\overset{1}{=}$

This rule can be written concisely as

$$Ax \overset{1}{=} x^+ + x$$

where the $+$ and $\cdot$ operations are extended linearly to sums of diagrams.

(ii) **Componentwise functions create several subtrees at the root.**

The expression $f\left(\sum_{\tau \in T} c_\tau Z_\tau\right)$ where f is a polynomial applied componentwise can be computed by applying each monomial of f separately and summing the results.

To componentwise multiply two tree-shaped diagrams, we first merge the two roots of the diagrams, and then sum over all possible partial matchings of isomorphic subtrees, under the constraint no two subtrees from the same diagram can be matched. Whenever we match up two copies of a subtree σ, we delete them and multiply by $|\mathrm{Aut}(\sigma)|$. See Fig. 1 for an example and Appendix B for a more detailed explanation.

Algebraically, a diagram $\tau \in T$ represents a multivariate Hermite polynomial. When two or more diagrams are multiplied, this operation corresponds to multiplying their Hermite polynomials and re-expressing the result as a sum of Hermite polynomials, exactly as in Lemma 2.6.

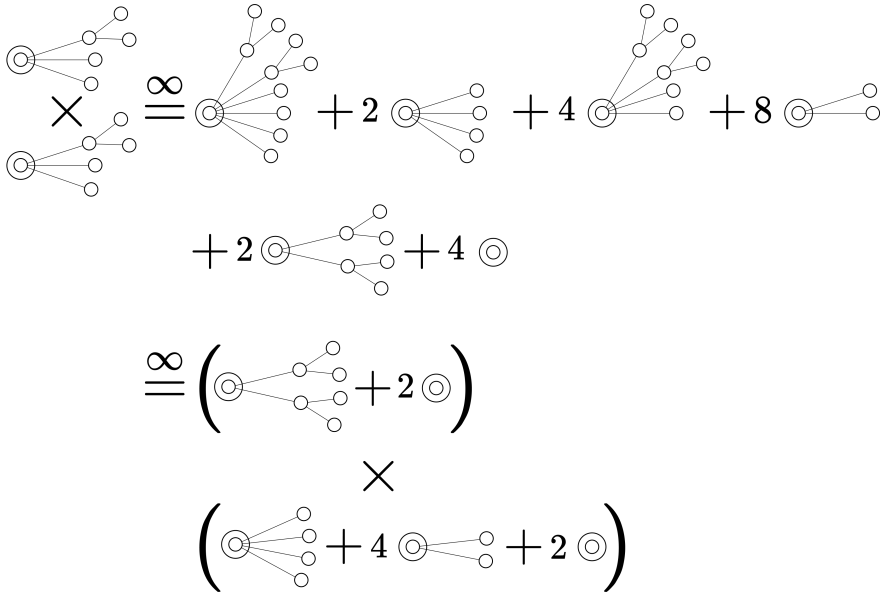


Figure 1: Computing the componentwise square of $Z_{\text{edge, edge, (1,2)-tree}}$. The subtrees of the root are a "$(1,2)$-tree" and two single edges. The (1,2)-tree has $|\mathrm{Aut}(\sigma)| = 2$ which yields its factor of 2 when paired. For the single-edge subtrees, we can choose one edge from each diagram in 4 ways, or pair up both edges in 2 ways, which yields the factors of 4 and 2.

We have characterized the tree diagrams as being the only non-negligible parts of the asymptotic state. If α is a negligible diagram, we will show that any operation involving α creates only negligible diagrams (Lemma 4.6). This allows us to discard the negligible diagrams any time they appear. In other words, we can asymptotically assume that the algorithm is operating only on the tree-like diagrams.

3.4 Perspective: symmetrized Fourier analysis

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

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_α 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 α can be interpreted combinatorially as graphs on the vertex set $[n]$.

The key observation that allows us to significantly simplify the representation is that many of the Fourier coefficients are guaranteed to be equal for algorithms which are symmetric under permutations of $[n]$, a property which holds for many common algorithms.[10] Considering the permutation action of S_n on $[n]$ (the vertex set of the Fourier characters), any two Fourier characters α, β which are in the same orbit will have the same Fourier coefficient. Equivalently, if α and β are isomorphic as graphs, then their Fourier coefficients are the same. By grouping together all isomorphic Fourier characters, we obtain the symmetry-reduced representation which defines the 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. We also point out that 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 discussion above applies to Boolean matrices A with i.i.d. Rademacher entries. In general, the natural way to express a function of A as a polynomial is in basis of orthogonal polynomial basis for the entries A_{ij} (e.g. the Hermite polynomials when the A_{ij} are

[10] This property holds for any function/algorithm $f : \mathbb{R}^{n \times n} \to \mathbb{R}^n$ which "acts the same at all indices": $f(A)_i = f(\pi(A))_{(i)}$ for all $\pi \in S_n$, where π acts on A by permuting the rows and columns simultaneously.

Gaussian). This is for example the convention used in [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, every term A_{ij}^2 can be approximated by $\frac{1}{n}$, and every term involving A_{ij}^k for k 3 can be discarded. In our case, it turns out that using the monomial basis[11] to represent higher-degree polynomials will simplify the presentation (except we will use the degree-2 orthogonal polynomial A_{ij}^2 $\frac{1}{n}$ to express 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 C for readability.

- In Section 4.1, we give a rigorous definition of the asymptotic equality $\overset{1}{=}$. This definition will let us reason at a high level in a completely rigorous way.
- In Section 4.2, we prove our first main theorem that classifies the asymptotic behavior of the constant-size diagrams.
- In Section 4.3, we prove our second main theorem about the validity of the tree approximation for the class of general first-order methods.
- In Section 4.4, we define an asymptotic probability space which can be used to describe the n ! 1 limit of diagram expressions.

4.1 Equality up to combinatorially negligible diagrams

The idea behind $\overset{1}{=}$ is to make a purely combinatorial definition for which we can utilize combinatorial arguments on the diagrams. First, we can estimate the magnitude in n of a diagram Z_α with the following combinatorial formula.

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

Lemma 4.2. *Let q 2* N *be a constant independent of n and α 2 A be a constant-size diagram. Then for i 2 $[n]$,*

$$\left|\mathsf{E}\left[Z_{\alpha,i}^q\right]\right| \quad O\left(n^{\frac{q}{2}(jV(\alpha)j \;\; 1 \;\; jE(\alpha)j+jI(\alpha)j)}\right).$$

[11] 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.

This bound is analogous to the "graph matrix norm bounds" used in Fourier analysis of matrix-valued functions [AMP20]. The proof of Lemma 4.2 is in Appendix C.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 $\alpha \in A$ 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 ($\overset{1}{=}$)**.** *We say that $x \overset{1}{=} y$ if there exists real coefficients $(c_\alpha)_{\alpha \in A}$ depending on n and supported on diagrams of constant size such that*

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

where $c_\alpha Z_\alpha$ is combinatorially negligible for all $\alpha \in A$.

In later sections, we will prove results of the form $x \overset{1}{=} \widehat{x}$ 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 $\overset{1}{=}$ implies a very strong form of probabilistic convergence of the error to 0. The proof of the following lemma can be found in Appendix C.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 $x \overset{1}{=} y$, then $\|x - y\|_1 \xrightarrow{a.s.} 0$.*

Next, we state a very important property of $\overset{1}{=}$. The combinatorially negligible diagrams remain combinatorially negligible after applying additional algorithmic operations.

Lemma 4.6. *If x, y are diagram expressions with $x \overset{1}{=} y$, then*

$$Ax \overset{1}{=} Ay .$$

Moreover, if $x_1, \ldots, x_t, y_1, \ldots, y_t$ are diagram expressions with $x_i \overset{1}{=} y_i$ for all $i \in [t]$, then

$$f(x_1, \ldots, x_t) \overset{1}{=} f(y_1, \ldots, y_t) ,$$

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

The proof of Lemma 4.6 is postponed to Appendix C.2.

As a further demonstration of the power of this definition, we show combinatorially in Appendix C.2 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 also Appendix C.1).

Lemma 4.7. *Let $a_n Z_\alpha$ be a term of combinatorial order at most 1 such that α has a hanging double edge. Let α_0 be α with the hanging double edge and hanging vertex removed. Then*

$$a_n Z_\alpha \overset{1}{=} a_n Z_{\alpha_0}.$$

4.2 Classification of constant-size diagrams

In addition to the vector diagrams from Definition 3.1, we will classify *scalar diagrams*, which are simply unlabeled undirected multigraphs (the only difference with vector diagrams being that they do not have a root). We introduce analogous notation for scalar diagrams:

Definition 4.8 (Scalar diagrams)**.** *Let A_{scalar} be the set of all unlabeled undirected multigraphs with no isolated vertices. Let T_{scalar} be the set of non-empty unlabeled trees.*

Given a scalar diagram $\alpha \in A_{\text{scalar}}$, we analogously 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 diagram which represents the constant 1.

Just as the trees T are the non-negligible (connected) vector diagrams, forests are the non-negligible scalar diagrams (and disconnected vector diagrams).

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

The scalar diagrams are not normalized "correctly" by default. Z_ρ for $\rho \in F_{\text{scalar}}$ has order $n^{c/2}$ where c is the number of connected components in ρ. The proper normalization divides by $n^{c/2}$ to put all the diagrams on the same scale. The notion of $\overset{1}{=}$ and combinatorial negligibility also extends in a natural way to scalar diagrams. See Appendix C.3 for full definitions.

We classify the diagrams in A and A_{scalar}. First, we characterize the diagrams which are combinatorially non-negligible as follows. The following lemma is for *connected* vector diagrams; scalar diagrams and disconnected vector diagrams have a similar characterization in Lemma C.12.

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

(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 C.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

$$|V(\alpha)| \quad |I(\alpha)| \quad 1) + 2|I(\alpha)| \quad |E(\alpha)| .$$

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

Continuing, we describe the behavior of the non-negligible terms. If $\alpha \in A$ is connected and non-negligible, then by Lemma 4.10, it is asymptotically equal to a tree in T, after using Lemma 4.7 to remove the hanging double edges. For disconnected diagrams $\alpha \in A$ and scalar diagrams $\alpha \in A_{\text{scalar}}$, likewise Lemma C.12 describes the non-negligible diagrams as asymptotically equal to a diagram in F or F_{scalar} after removing hanging double edges. The next Theorem 4.11 describes the trees and forests, to be proven in Appendix C.4.

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

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

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

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

- *If $\sigma \in S$ and $i \in [n]$, then $Z_{\sigma,i} \xrightarrow{d} N(0, |\text{Aut}(\sigma)|)$.*
- *If $\tau \in T$, then $Z_\tau \overset{1}{=} \prod_{\sigma \in S} h_{d_\sigma}(Z_\sigma; |\text{Aut}(\sigma)|)$ where d_σ is the number of isomorphic copies of σ starting from the root of τ, and the Hermite polynomial is applied componentwise.*

- *If α 2 F has c oating components (connected components which are not the component of the root), letting $\alpha_{\circledcirc}$ be the component of the root (a vector diagram) and α_{float} be the oating part (a scalar diagram), then* $n^{\frac{c}{2}} Z_\alpha \stackrel{1}{=} n^{\frac{c}{2}} Z_{\alpha \text{ oat}} Z_{\alpha_{\circledcirc}}$.

Moreover, the random variables

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

are asymptotically independent (De nition 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 random vectors $(X_{n,i})_{n \in \mathbb{N}, i \in I_n}$ is asymptotically independent if:*

$$\forall q \in \mathbb{N}.\ \exists \varepsilon = \varepsilon(q) > 0.\ \forall k \in \mathbb{N}^{I_n} : \sum_{i \in I_n} k_i = q.\ \left| \mathbb{E}\left[\prod_{i \in I_n} X_{n,i}^{k_i}\right] - \prod_{i \in I_n} \mathbb{E}\left[X_{n,i}^{k_i}\right] \right| \le \varepsilon(q)\,.$$

Note that I_n may be in nite.

The proof of Theorem 4.11 can be found in Appendix C.4.

4.3 Tree approximation of GFOMs

The previous discussion suggests that to asymptotically describe first-order iterative algorithms, it suffices to analyze their projection on the tree diagrams. Our main result in this section (Theorem 4.16) formalizes this intuition for a class of iterative algorithms known as "general first-order methods" (GFOM) defined by Celentano, Montanari, and Wu [CMW20, MW22b].

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 any 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[12] function independent of n, $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ to $(x_t, x_{t-1}, \ldots, x_0)$ ($x_{t+1} = f_t(x_t, \ldots, x_0)$).*

Note that the algorithmic state x_t of a GFOM can be represented in the diagram basis. This can be seen by induction, or more efficiently by representing the state in the repeated-label diagram basis (Appendix B.3) and converting it to the diagram basis (Lemma B.6) at the end.

[12] Although the restriction to polynomial functions is absent from the original GFOM definition, it is a technical assumption that we need here.

We now introduce formally the *tree approximation* of the state of a GFOM. This part follows the rules presented in Section 3.3. We first define the + and operators acting on a diagram expression:

Definition 4.14 (+ and operators)**.** *Fix τ 2 T .*

- *We de ne τ^+ to be the diagram obtained by extending the root by 1, i.e. by adding one new vertex with one edge connecting it to the root of τ, and re-rooting τ^+ at this new vertex.*
- *If τ 2 S, we de ne τ to be the diagram obtained by contracting the root by 1, i.e. by removing the root vertex and the edge connecting it to its only subtree, and re-rooting τ at the endpoint of that edge. If τ 2 T n S, we let $\tau = 0$ by convention.*

Finally, we extend the operators + and linearly to the span of tree diagrams.

Definition 4.15 (Tree approximation of a GFOM, $\widehat{x}_t$)**.** *Let x_t 2 R^n be the state of a GFOM. We recursively de ne the tree approximation of x_t, denoted by $\widehat{x}_t$, to be a diagram expression in the span of $(Z_\tau)_{\tau 2 T}$.*

1. *Initially, $\widehat{x}_0 = Z_{\circledcirc}$.*
2. *If $x_{t+1} = Ax_t$, de ne $\widehat{x}_{t+1} = (\widehat{x}_t)^+ + (\widehat{x}_t)$.*
3. *If $x_{t+1} = f_t(x_t, \dots, x_0)$ coordinatewise for some polynomial $f_t : \mathsf{R}^t \to \mathsf{R}$, de ne $\widehat{x}_{t+1}$ 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 output of multiplying q diagrams $\tau_1, \dots, \tau_q$ 2 T is*

$$\sum_{M 2 \mathcal{M}(\tau_1,\dots,\tau_q)} c_M Z_{\tau_M} ,$$

where:

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

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

   (c) *$c_M = \prod_{\{\sigma,\sigma'\} 2 M} |\mathrm{Aut}(\sigma)|$.*

Theorem 4.16 (Tree approximation of GFOMs)**.** *Let $T \geq 1$ be a constant independent of n and $A = A(n)$ be a sequence of random matrices satisfying Assumption 2.1. Let x_T 2 R^n be the state of a GFOM and let $\widehat{x}_T$ be its tree approximation. Then $x_T \overset{1}{=} \widehat{x}_T$. So in particular,*

$$\|x_T - \widehat{x}_T\|_1 \xrightarrow{a.s.} 0 . \tag{3}$$

Proof. For each of the two operations of the GFOM, we have expressed the result in terms of diagrams and identified the dominant terms in Appendix B.2 (using Theorem 4.11). They correspond exactly to the definition of $\widehat{x}_t$. The negligible terms remain negligible by Lemma 4.6. Eq. (3) then follows from Lemma 4.5. □

Remark 4.17. *The direct nature of the proof yields a complete description of the speed of convergence in Eq. (3). The other connected diagrams with E edges and V vertices have magnitude $n^{(V\ 1\ E)/2}$. For example, the rst lower-order term of order $n^{\ 1/2}$ consists of connected diagrams with exactly one cycle. We note that the number of vertices in a diagram does not directly impact its magnitude, which can be counter-intuitive.*

Remark 4.18. *One technical caveat of our analysis is that many denoisers used in applications are not polynomial functions. The usual workaround is to approximate these functions by polynomials. We note that existing approximation arguments in the literature (see for example [MW22a, IS23]) should apply here to prove that the tree approximation holds for GFOMs with Lipschitz denoisers f_t up to arbitrarily small $\frac{1}{n}k\ k_2$ error. This is however strictly weaker than the guarantees of Theorem 4.16.*

4.4 Asymptotic Gaussian space

One important feature of Theorem 4.11 we did not exploit so far is that the entries of the tree diagrams are asymptotically independent and identically distributed. This means that in the limit $n\ !\ 1$, the distributional information provided by the tree approximation in Definition 4.15 can be compressed down to a one-dimensional object.

Definition 4.19 (Asymptotic Gaussian space)**.** *Let $(Z_\sigma^{\,1})_{\sigma 2 S}$ be a set of independent centered Gaussian random variables with variances $\mathrm{Var}(Z_\sigma^{\,1}) = j\mathrm{Aut}(\sigma)j$. For τ 2 T consisting of d_σ copies of each σ 2 S, let $Z_\tau^{\,1} = \prod_{\sigma 2 S} h_{d_\sigma}(Z_\sigma^{\,1}; j\mathrm{Aut}(\sigma)j)$.*

Let $\Omega = \mathsf{R}[\mathsf{Z}_\sigma^{\,1} : \sigma$ 2 S$]$ be the set of polynomials in a set of indeterminates $(\mathsf{Z}_\sigma^{\,1})_{\sigma 2 S}$.[13]

Note that this probability space is independent of n and A.[14] Although we have made a small notational distinction here between the formal indeterminates $\mathsf{Z}_\sigma^{\,1}$ and the random variables $Z_\sigma^{\,1}$, we will always think of evaluating polynomials in Ω on $\mathsf{Z}_\sigma^{\,1} = Z_\sigma^{\,1}$.

The space Ω has an inner product coming from the expectation over the $Z_\sigma^{\,1}$. Since these random variables are independent Gaussians, the multivariate Hermite polynomials $(Z_\tau^{\,1})_{\tau 2 T}$ form an orthogonal basis of Ω with respect to this inner product.

Definition 4.20 (Asymptotic state)**.** *Let x 2 R^n such that $x \overset{1}{=} \sum_{\tau 2 T} c_\tau Z_\tau$. The asymptotic state of x is $X = \sum_{\tau 2 T} c_\tau Z_\tau^{\,1}$.*

The distribution of X describes the asymptotic distribution of a single coordinate of x with respect to the randomness of A. Since the coordinates of x are identically distributed

[13] Although there are infinitely many $\mathsf{Z}^{\,1}$, by definition each polynomial consists of a finite number of monomials.

[14] There is a natural obstruction to constructing an actual asymptotic diagram basis from no additional randomness, even for the single edge diagram [Dur19, Exercise 3.4.2].

and asymptotically independent, this also describes the empirical distribution of the coordinates of x for a generic fixing of A and large n. We prove that any polynomial test function of the empirical distribution is concentrated.

Theorem 4.21 (State evolution for GFOMs)**.** *Under the same assumptions as in Theorem 4.16, for any polynomial of constant degree $\psi : \mathsf{R}^{T+1} \to \mathsf{R}$,*

$$\frac{1}{n}\sum_{i=1}^{n} \psi(x_{T,i}, \dots, x_{0,i}) \overset{a.s.}{\to} \mathsf{E}\left[\psi(X_T, \dots, X_0)\right] ,$$

where $X_T, \dots, X_0$ are the asymptotic states of $x_T, \dots, x_0$ and the expectation on the right-hand side is with respect to the randomness of the asymptotic Gaussian space.

For AMP algorithms, this has been a standard statement of state evolution (see also [CMW20, Appendix B]). We refer to the discussion following Theorem 5.2 for additional comparison with Theorem 4.16.

To prove Theorem 4.21, we combine Theorem 4.16 with the following combinatorial concentration lemma whose proof is in Appendix C.5. Almost sure convergence follows using Lemma C.10.

Lemma 4.22. *Let x be a vector diagram expression with asymptotic state $X \in \Omega$. Then as scalar diagrams, $\frac{1}{n}\sum_{i=1}^{n} x_i \overset{1}{=} \mathsf{E}\left[X\right]$.*

We conclude this section by working out the effects of the $+$ and $*$ operators from Definition 4.14 on the asymptotic Gaussian space. The most important fact about the $+$ operator is that X^+ is always a Gaussian random variable for any $X \in \Omega$. The second most important fact is that the variance of X and X^+ is the same, as we develop now.

Fact 4.23. *$+$ and $*$ are bijections between T and S which are inverses of each other and preserve $|\mathrm{Aut}(\tau)|$.*

Fact 4.24. *For all $X \in \Omega$, $(X^+)^* = X$ and $(X^*)^+$ is the orthogonal projection of X to the subspace spanned by S.*

We deduce that $+$ and $*$ are adjoint operators on Ω:

Lemma 4.25. *For all $X, Y \in \Omega$, $\mathsf{E}\left[X^+ Y\right] = \mathsf{E}\left[X Y^*\right]$.*

Proof. Since $(Z_\tau^{\mathbf{1}})_{\tau \in T}$ is a basis of the vector space Ω, it suffices to check this for each pair of basis elements $\tau, \rho \in T$. By orthogonality, $\mathsf{E}\left[Z_{\tau^+}^{\mathbf{1}} Z_\rho^{\mathbf{1}}\right]$ is nonzero if and only if $\tau^+ = \rho$ and in this case it takes value $|\mathrm{Aut}(\tau^+)|$. By Fact 4.23, this occurs if and only if $\rho \in S$ and $\tau = \rho^*$. Moreover, in this case the value is also $|\mathrm{Aut}(\tau^+)| = |\mathrm{Aut}(\tau)|$, as needed. □

Lemma 4.26. *For all $X, Y \in \Omega$, $\mathsf{E}\left[XY\right] = \mathsf{E}\left[X^+ Y^+\right]$ and $\mathsf{E}\left[(X^*)^2\right] \le \mathsf{E}\left[X^2\right]$.*

Proof. For the first statement, apply Lemma 4.25 on X and Y^+, then use Fact 4.24. For the second statement, apply Lemma 4.25 on X^* and X to get $\mathsf{E}\left[(X^*)^+ X\right] = \mathsf{E}\left[(X^*)^2\right]$. Since $(X^*)^+$ projects away some terms from X by Fact 4.24, the left-hand side is upper bounded by $\mathsf{E}\left[X^2\right]$. □

5 Belief Propagation, AMP, and the Cavity Method

We now explore formally the connection between the tree approximation and the replica symmetric cavity method. We will do that by proving the *state evolution* formula for belief propagation in a simple way that mimics the heuristic argument.

Belief propagation. A general message passing algorithm on A is an iterative algorithm of the form

$$m^0_{i\to j} = 1\,, \qquad m^t_{i\to j} = f_t\left(\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m^{t-1}_{k\to i},\dots,\sum_{\substack{k=1\\k\neq j}}^n A_{ik} m^0_{k\to i},\ m^0_{i\to j}\right), \tag{4}$$
$$m^t_i = \widetilde{f}_t\left(\sum_{k=1}^n A_{ik} m^{t-1}_{k\to i},\dots,\sum_{k=1}^n A_{ik} m^0_{k\to i},\ m^0_{i\to j}\right),$$

for a sequence of functions $f_t, \widetilde{f}_t : \mathbb{R}^{t+1} \to \mathbb{R}$. Eq. (4) is a generalization of Eq. (1) to iterations "with memory" i.e. that can use all the previous messages. At any timestep t, the $(m^t_{i\to j})_{1\le i,j\le n}$ are *cavity messages* that try to compute some information about the i-th variable by ignoring the edge between i and j, while the $(m^t_i)_{1\le i\le n}$ correspond to the informative output of the algorithm (such as the estimate of the marginals of some Gibbs distribution).

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

$$w^0 = \vec{1}\,, \qquad w^{t+1} = Af_t(w^t,\dots,w^0) - \sum_{s=1}^t b_{s,t} f_{s-1}(w^{s-1},\dots,w^0)\,, \tag{5}$$

$$m^t = \widetilde{f}_t(w^t,\dots,w^0)\,, \tag{6}$$

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

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

The subtracted term in Eq. (5) is known as the *Onsager correction*.

Our results in this section are twofold. First, we show that the BP and AMP iterations Eq. (4) and Eq. (6) asymptotically coincide in a rigorous sense.

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. (4) and $m^{t,\mathrm{AMP}}$ according to Eq. (6). Then $m^{t,\mathrm{AMP}} \overset{1}{=} m^{\mathrm{BP}}$, so in particular*

$$\left\|m^{t,\mathrm{AMP}} - m^{t,\mathrm{BP}}\right\|_1 \overset{a.s.}{\to} 0\,.$$

Several heuristic justifications of the equivalence between BP and AMP exist in the literature, and our proof takes the heuristic argument and makes it rigorous.

Apart from a smaller number of messages to track, the other main advantage of the AMP form is the existence of an asymptotic description of the distribution of its trajectory known as *state evolution*. The second main result of this section deduces state evolution from the theory developed in the previous section.

Theorem 5.2 (Asymptotic state for AMP)**.** *Under the same assumptions as Theorem 5.1, the asymptotic state of* $(w_t)_{t \quad T}$ *satis es the recursion*

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

In particular, W_t *is a centered Gaussian and for all* $s, t \quad T$*,*

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

Combining Theorem 5.2 and Theorem 4.21 recovers the typical formulation of state evolution for AMP algorithms, which says that every empirical expectation is concentrated. This is sufficient to compute "averaged" quantities such as the norm $\|m^t\|_2$ or the loss achieved by m^t. The formulation in Theorem 5.2 is more powerful since it gives an explicit approximation to w_t which can also approximate some quantities that do not concentrate, such as the first coordinate of w_t.

The Gaussianity of the w_t may appear surprising at first sight. Theorem 5.2 gives a diagrammatic interpretation of this property. First, $f_t(W_t, \dots, W_0)^+$ is guaranteed to be Gaussian since applying the + operator sends all the tree diagrams to S. Second, we will see in the proof that the Onsager correction term perfectly cancels out the backtracking term $f_t(W_t, \dots, W_0)$.

The plan for the rest of the section is as follows. In Section 5.1, we give a heuristic derivation of the equivalence between BP and AMP on dense models [DMM09, ZK16]. In Section 5.2, we justify formally the approximate steps to prove Theorem 5.1. In Section 5.3 we prove Theorem 5.2.

5.1 Heuristic derivation of the BP-AMP equivalence

The following argument can be found in e.g. [ZK16, Appendix IV.E] or [DMM09, Appendix A]. We start by rewriting the BP iteration by letting $w^0 = \vec{1}$ and $w_i^{t+1} = \sum_{k=1}^n A_{ik} m^t_{k \to i}$. 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. (5). First, Eq. (4) can be rewritten

$$m^{t+1}_{i \to j} = f_{t+1}\left(w_i^{t+1} \quad A_{ij} m^t_{j \to i}, \dots, w^1_{k \to i} \quad A_{ij} m^0_{j \to i}, w_i^0\right)\,.$$

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):

$$m_{i \to j}^{t+1} \quad f_{t+1}\left(w_i^{t+1}, \ldots, w_i^1, w_i^0\right) \quad A_{ij} \sum_{s=1}^{t+1} m_{j \to i}^{s} \frac{\partial f_{t+1}}{\partial w^s}\left(w_i^{t+1}, \ldots, w_i^1, w_i^0\right) . \qquad (\ )$$

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

$$\begin{aligned} w_i^{t+1} \quad & \sum_{k=1}^{n} A_{ik} f_t(w_k^t, \ldots, w_k^0) \quad \sum_{k=1}^{n} A_{ik}^2 \sum_{s=1}^{t} m_{i \to k}^{s} \frac{\partial f_t}{\partial w^s}(w_k^t, \ldots, w_k^0) && (\ ) \\ & \sum_{k=1}^{n} A_{ik} f_t(w_k^t, \ldots, w_k^0) \quad \sum_{k=1}^{n} \frac{1}{n} \sum_{s=1}^{t} f_s(w_i^s, \ldots, w_i^0) \frac{\partial f_t}{\partial w^s}(w_k^t, \ldots, w_k^0) && (\ ) \\ = & \sum_{k=1}^{n} A_{ik} f_t(w_k^t, \ldots, w_k^0) \quad \sum_{s=1}^{t} b_{s,t} f_s(w_i^s, \ldots, w_i^0) . \end{aligned}$$

This shows that w_i^{t+1} approximately satisfies the AMP recursion Eq. (5), as desired.

A heuristic explanation of the approximation in Eq. ( ) is that because we are summing over k, we may expand A_{ik}^2 and $m_{i \to k}^{s}$ on the first order and replace them by averages which do not depend on k:

$$\begin{aligned} A_{ik}^2 \quad & \mathsf{E}\left[A_{ik}^2\right] = \frac{1}{n} , \\ m_{i \to k}^{s} = f_s & \left(w_i^s \quad A_{ik} m_{k \to i}^t, \ldots, w_i^1 \quad A_{ik} m_{k \to i}^0, w_i^0\right) \quad f_s\left(w_i^s, \ldots, w_i^0\right) . \end{aligned}$$

5.2 Diagram proof of the BP-AMP equivalence

We now justify Eq. ( ) and Eq. ( ) in order to prove Theorem 5.1.

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

Definition 5.3 (Cavity diagrams)**.** *A cavity diagram is an unlabeled undirected graph $\alpha = (V(\alpha), E(\alpha))$ with two distinct, ordered root vertices $\textcircled{\scriptsize 1}\textcircled{\scriptsize 2}$. No vertices may be isolated except for the root.*

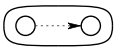
For any cavity diagram α, we define $Z_\alpha \in \mathsf{R}^{n(n-1)}$ by

$$Z_{\alpha, i \to j} = \sum_{\substack{\varphi : V(\alpha) \to [n] \\ \varphi \text{ injective} \\ \varphi(\textcircled{\scriptsize 1}\textcircled{\scriptsize 2}) = (i,j)}} \prod_{\{u,v\} \in E(\alpha)} A_{\varphi(u), \varphi(v)} ,$$

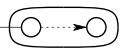
for any distinct $i, j \in [n]$.

We show as an example how to represent the first iterate of Eq. (4) 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.

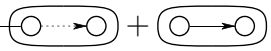
$$m^0_{i\,!\,j} =$$



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



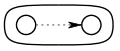
$$\sum_{k=1}^{n} A_{ik} m^0_{k\,!\,i} =$$



When we multiply together $A_{ik} m^t_{k\,!\,i}$ we "fill in" the edge between k and 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 $k \in \{i, j\}$. 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".

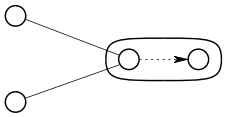
To apply f_1, we need to multiply $i \;!\; j$ diagrams componentwise, which is achieved by fixing/merging the roots i, j and summing over the part outside the roots. For some coefficients $c_0, c_1, c_2, \ldots$ we have[15]

$$m^1_{i\,!\,j} = f_1\left(\sum_{\substack{k=1\\ k\neq j}}^{n} A_{ik} m^0_{k\,!\,i}\right) = c_0$$



$$+ c_1$$

$$+ c_2$$



$$+$$

The output m^{t+1}_i uses the non-cavity quantities $\sum_{k=1}^{n} A_{ik} m^t_{k\,!\,i}$ which do not really depend on the first root in $m^t_{k\,!\,i}$. The cavity diagrams are converted back to the usual diagram basis as follows.

Claim 5.4 (Conversion of cavity diagrams)**.** *For any cavity diagram α and $i \in [n]$,*

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

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

Since all cavity diagrams are eventually converted back to regular diagrams using the previous claim, the definition of combinatorial negligibility and the $\overset{1}{=}$ notation can be extended to cavity diagrams. The cavity diagrams that will contribute to the tree approximation of the output of the message-passing algorithm are precisely those whose diagram obtained by performing the previous unrooting operation is combinatorially order 1. The following definition and claim extend $\overset{1}{=}$ to cavity diagrams.

[15] The exact values of the coefficients c_i are not necessary to compute, since the final state will be derived using the high-level cavity method reasoning.

Definition 5.5. *A cavity diagram α is combinatorially negligible if the diagram $\alpha^{\emptyset}$ obtained in Claim 5.4 is combinatorially negligible. We naturally extend the $\overset{1}{=}$ notation to cavity diagrams as in Definition 4.4.*

Claim 5.6. *Let x and $x^{\emptyset}$ be in the span of the cavity diagrams such that $x \overset{1}{=} x^{\emptyset}$. If we let*

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

then $y \overset{1}{=} y^{\emptyset}$.

If $x_1, \ldots, x_t, x_1^{\emptyset}, \ldots, x_t^{\emptyset}$ are in the span of cavity diagrams and $x_i \overset{1}{=} x_i^{\emptyset}$ and $f : \mathsf{R}^t \to \mathsf{R}$ is a polynomial function applied componentwise, then

$$f(x_1, \ldots, x_t) \overset{1}{=} f(x_1^{\emptyset}, \ldots, x_t^{\emptyset})\,.$$

Claim 5.6 follows directly from Lemma 4.6.

Although we will not explicitly need it in the sequel, we make the following connection between the cavity diagrams and the cavity method. When applying the message-passing recursion Eq. (4), the asymptotic diagram representation of the cavity messages $m^t_{i \to j}$ are trees rooted at the vertex labelled i. This precisely mimics the assumptions of the cavity method (see Fig. 2).

Fact 5.7. *$m^t_{i \to j}$ is asymptotically a linear combination of cavity diagrams which have a tree hanging out of i, nothing attached to j, and no edges between i and j.*

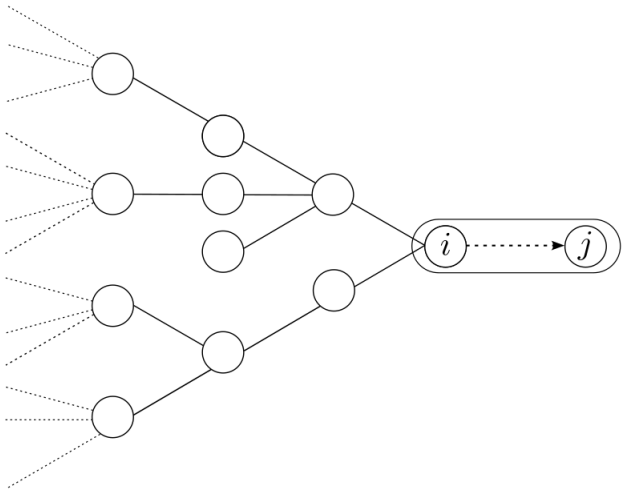


Figure 2: Diagram representation of the cavity messages $m^t_{i \to j}$. Each cavity diagram in the asymptotic cavity diagram representation of $m^t_{i \to j}$ is a tree rooted at i.

This completes the diagrammatic description of the belief propagation algorithm. Finally, we use diagrams to justify rigorously the approximations made during the heuristic argument.

Lemma 5.8 (Eq. ( )).

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

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 5.4 adds one more edge between i and j, and regular connected diagrams with multiedges of multiplicity > 2 are combinatorially negligible (Lemma 4.10). □

Lemma 5.9 (Eq. ( )).

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

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

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

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

$$\sum_{k=1}^{n} A^2_{ik} m^{s\ 1}_{i\,!\,k} \frac{\partial f_t}{\partial w^s}(w^t_k, \dots, w^0_k) \overset{1}{=} f_{s\ 1}(w^{s\ 1}_i, \dots, w^0_i) \sum_{k=1}^{n} A^2_{ik} \frac{\partial f_t}{\partial w^s}(w^t_k, \dots, w^0_k) . \tag{7}$$

Second, we argue about the replacement of A^2_{ik}. The double edge is only non-negligible if it is hanging (Lemma 4.10). Among the diagrams in $\frac{\partial f_t}{\partial w^s}(w^t_k, \dots, w^t_0)$ the only one which does not attach something to k is the singleton diagram ◎. The coefficient of this diagram is the expected value (Corollary A.3),

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

The expected value is equal to the empirical expectation up to negligible terms (Lemma 4.22),

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

This implies

$$\sum_{k=1}^{n} A^2_{ik} \frac{\partial f_t}{\partial w^s}(w^t_k, \dots, w^0_k) \overset{1}{=} \frac{1}{n} \sum_{k=1}^{n} \frac{\partial f_t}{\partial w^s}(w^t_k, \dots, w^0_k) . \tag{8}$$

The desired statement follows from combining Eq. (7) and Eq. (8). □

Proof of Theorem 5.1. The proof follows by replacing the signs in the heuristic argument from Section 5.1 by $\stackrel{1}{=}$ and using Claim 5.6 repeatedly. □

5.3 State evolution for AMP algorithms

We now give a more explicit description of the tree approximation for the BP/AMP iterations Eq. (4) and Eq. (5). As mentioned before we expect w_t to be asymptotically Gaussian. Our main result here is Theorem 5.2, which describes the asymptotic covariance structure of these Gaussians. Such result is known in the literature under the name of state evolution.

Note that Eq. (5) is not directly captured by the definition of a GFOM (Definition 4.15) because $b_{s,t}$ requires computing an average over coordinates. This is only a technical issue: such quantities are concentrated up to combinatorially negligible terms, so we can replace them by their limiting value without asymptotically altering the iteration (this is verified by Lemma 4.22). 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:

$$w_0 = \vec{1}, \quad w_{t+1} = Af_t(w_t, \dots, w_0) \quad \sum_{s=1}^{t} \mathbb{E}\left[\frac{\partial f_t}{\partial w_t}(W_t, \dots, W_0)\right] f_{s\ 1}(w_{s\ 1}, \dots, w_0)\,. \tag{9}$$

The Onsager correction term from Eq. (9) has a simple diagrammatic interpretation:

Lemma 5.10. *Let $W_1, \dots, W_t \in \Omega$ be Gaussian (i.e. each W_s is in the span of $(Z_\sigma^{\mathbf{1}})_{\sigma \in S}$). Then for any polynomial function $f : \mathbb{R}^t \to \mathbb{R}$,*

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

In other words, the Onsager correction term precisely cancels out the backtracking term appearing in item (i) of Section 3.3.

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

$$\begin{aligned} f(W_1, \dots, W_t) \quad &= \sum_{\sigma \in S} c_\sigma Z_\sigma^{\mathbf{1}} + \sum_{\tau \in T \setminus S} c_\tau Z_\tau^{\mathbf{1}} \ , \\ f(W_1, \dots, W_t) \quad &= \sum_{\sigma \in S} c_\sigma Z_\sigma^{\mathbf{1}} \ , \end{aligned}$$

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

$$\begin{aligned} c_\sigma |\mathrm{Aut}(\sigma)| &= \mathbb{E}\left[Z_\sigma^{\mathbf{1}} f(W_1, \dots, W_t)\right] && \text{(orthogonality)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_\sigma^{\mathbf{1}} W_s\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] && \text{(Lemma 2.7)} \\ &= \sum_{s=1}^{t} \mathbb{E}\left[Z_\sigma^{\mathbf{1}} W_s\right] \mathbb{E}\left[\frac{\partial f}{\partial W_s}(W_1, \dots, W_t)\right] . && \text{(Lemma 4.25)} \end{aligned}$$

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

Now we complete the proof of Theorem 5.2.

Proof of Theorem 5.2. We first 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 4.24, for any s t,

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

Combining with Lemma 5.10, 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 S, so it has a centered Gaussian distribution. Moreover, for all s, t T,

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

where the last equality follows from Lemma 4.26. This completes the proof. □

Example 5.11 (Iterative AMP)**.** *A special type of iterative AMP (or martingale AMP) was introduced to optimize Ising spin glass Hamiltonians up to arbitrary approximation error [Mon19, AMS21]. Iterative AMP [Mon19] uses Eq. (5) with the functions*

$$f_t(w_t, \dots, w_0) = w_t \quad u_t(w_{t-1}, \dots, w_0) \tag{10}$$

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

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 is contained in the state evolution for AMP as follows.[16] *By Theorem 5.2, the asymptotic states W_t, U_t, X_t of w_t, u_t, x_t satisfy $U_0 = W_0 = 1$,*

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

Claim 5.12. *U_t is in the span of trees in T with depth at most t 1 and W_t is in the span of trees in S with depth exactly t.*

Proof of Claim 5.12. Arguing inductively, as componentwise functions do not increase the depth, U_t is in the span of trees from 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. □

[16] The algorithm of [Mon19] uses a non-polynomial f_t which is not directly covered by our state evolution proof. However, Ivkov and Schramm [IS23] prove that this AMP can be approximated by polynomial f_t.

Claim 5.12 provides a very clear explanation of where the independent Gaussians are coming from: the W_t have di erent depths, and Gaussian diagrams of di erent depths are asymptotically independent Gaussian vectors.

The ingenious next step used by [Mon19] is to observe that when the number of steps t is taken large, the point x_T heuristically approaches a martingale process $\mathrm{d}X_t = U_t \mathrm{d}B_t$ with the steps w_t converging to the Brownian motion. Based on this description, the functions u_t can be chosen in an optimal way [Mon19, AMS21].

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

In summary, so far we have completely described the trajectory of first-order algorithms for a *constant* number of iterations using their projections on the tree diagrams. We now discuss extensions to a number of iterations that scales with the dimension n of the matrix. A primary motivation for this is to study how first-order iterations converge, or whether they are stuck searching endlessly for a fixed point that cannot be found.

A second motivation is to study iterations with a warm start such as a spectral initialization [MV21, MV22, LW22] which occurs commonly in practice. If the initialization can be computed by a first-order method (e.g. power iteration for the spectral initialization) then we might hope to analyze the composite algorithm diagrammatically. This is demonstrated for algorithms with constantly many iterations in Fig. 3.

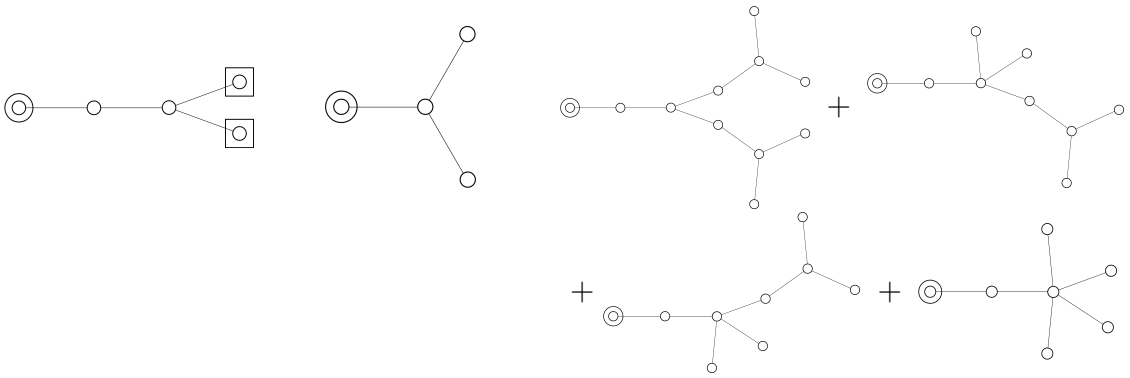


Figure 3: Composing diagram representations. The leaves of a diagram access the entries of the input, so we draw a box around each one to indicate that the input's entries are not yet fixed. When another diagram is used as input, it is duplicated at each leaf. Following the principles in Appendix B.1, the treelike diagrams in the result are a sum over i of contracting i path edges from both sides of the merged root/leaf. Note that here the second and third output diagram are the same.

6.1 Combinatorial phase transitions

Unfortunately, the main tree approximation theorems (Theorem 4.11 and Theorem 4.16) are not generically true for a superconstant number of iterations. 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. See Appendix D for more explanation.

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

$$x_0 = \vec{1}, \quad x_{t+1} = Ax_t \quad x_t \ _1 . \tag{11}$$

Equation (11), which appears for example in the PCA literature [RM14, Section 5], 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

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

where $B \ 2 \ \mathsf{R}^{n(n \ 1)}$ is the weighted non-backtracking walk matrix defined by $B_{i \ ! \ j,k \ ! \ \ell} = A_{k\ell}$ if $j = k$ and $i \ 6= \ \ell$, and $B_{i \ ! \ j,k \ ! \ \ell} = 0$ otherwise.

We distinguish several regimes of $T = T(n)$ depending on the obstacles that arise when trying to generalize the tree approximation for Eq. (11) to a larger number of iterations.

- When $T \quad \frac{\log n}{\log \log n}$, we expect the proofs of Theorem 4.11 and Theorem 4.16 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 $T \quad \frac{\log n}{\log \log n}$, 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 $T \quad n^{\delta}$ 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 $T \quad \sqrt{n}$.
- When $T \quad \sqrt{n}$, 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 4.16, 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 $n \quad n$ symmetric matrix with $A_{ii} = 0$ and A_{ij} for all $i < j$ independently drawn from the uniform distribution over $\left\{ \quad \frac{1}{n \quad 1}, \frac{1}{n \quad 1} \right\}$.*

We prove that for this iterative algorithm we can extend Theorem 4.16 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 this iteration does not converge in a strong way, 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. (11). Then there exist universal constants $c, \delta > 0$ such that for all $t \quad cn^{\delta}$,*

$$\| x_t \quad Z_{t\text{-path}} \|_1 \xrightarrow{a.s.} 0 .$$

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 $x_t \quad Z_{t\text{-path}}$. 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 $\gamma = (\gamma_1, \ldots, \gamma_q)$ is an ordered sequence of q walks, each of length t and starting from the same vertex:*

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

Each traversal γ is naturally associated to an improper diagram $(V(\gamma), E(\gamma))$ with $V(\gamma) = \{u_{i,j} : i \in [q], j \in [t]\}$ and $E(\gamma) = \{(u_{i,j}, u_{i,j+1}) : i \in [q], j \in [t \quad 1]\}$ (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. $u_{i,j+1} \neq u_{i,j \quad 1}$ for all $i \in [q]$ and $j \in \{2, \ldots, t \quad 1\}$.*
- *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 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. (11) and A satis es Assumption 6.1. Then,*

$$\mathsf{E}\left[(x_t \quad Z_{t\text{-path}})^q\right] = \sum_{\gamma 2 W_t^q} \mathsf{E}\left[Z_\gamma\right] .$$

We now proceed to proving Theorem 6.2. We will bound the magnitude of $\mathsf{E}\left[Z_{\gamma,i}\right]$ for γ 2 W_t^q, then count the number of traversals in W_t^q. Both bounds will depend on $\frac{E}{2} \quad V+1$ (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 $(x_t \quad Z_{t\text{-path}})^q$ become further from trees as q increases because each walk must have a back edge.

Lemma 6.5. *For any γ 2 W_t^q with V vertices and E edges, $\frac{E}{2} \quad V+1 \quad \frac{q}{2}$.*

Proof. Assign to each vertex all the edges going into it in γ. Each non-root vertex must have at least 2 incoming edges: the edge which explores it, and since γ is even and non-backtracking, an edge which revisits it a second time. Since γ is non-full-forward, each γ_i has a back edge; the first back edge in each γ_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 *rst* back edge in γ_i, it cannot cover both incident edges from the first visit). We have

$$E \quad 2(V \quad 1) + q ,$$

as needed. □

Lemma 6.6. *For any i 2 $[n]$ and γ 2 W_t^q with V vertices and E edges,*

$$j\mathsf{E}\left[Z_{\gamma,i}\right]j \quad O\left(n^{\left(\frac{E}{2} \quad V+1\right)}\right) .$$

Proof. Using Assumption 6.1, we can directly count

$$\begin{aligned} j\mathsf{E}\left[Z_{\gamma,i}\right]j \quad & O(1) \quad \frac{(n \quad 1)(n \quad 2) \quad (n \quad V+1)}{n^{\frac{E}{2}}} \\ & = O\left(n^{V \quad 1 \quad \frac{E}{2}}\right) . \end{aligned}$$

□

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

Lemma 6.7. *The number of γ 2 W_t^q with V vertices and E edges is at most*

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

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

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

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

If t satisfies $t \le cn^\delta$ 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 $\varepsilon > 0$,

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

This is enough to imply that $\|x_t - Z_{t\text{-path}}\|_1 \xrightarrow{a.s.} 0$ 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 $V = \frac{E}{2} + 1$ where the moment bound Lemma 6.6 is the weakest, typical traversals $\gamma \in W_t^q$ 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 $\gamma \in W_t^q$, let γ_c be the traversal obtained by replacing all maximally long paths of degree-4 vertices in γ 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 γ_c.

Lemma 6.9. *For any $\gamma \in W_t^q$,*

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

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

$$2|V_2(\gamma)| + 4|V_4(\gamma)| + 6\left(|V(\gamma)| - |V_2(\gamma)| - |V_4(\gamma)| - 1\right) \le 2|E(\gamma)|.$$

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

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

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

We are now ready to prove Lemma 6.7.

Proof of Lemma 6.7. We encode a traversal $\gamma \in W_t^q$ as follows:

1. We first encode γ_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 γ_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 γ, and its length can be bounded by

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

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. 10

[AMP20] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: Norm bounds and applications. *arXiv preprint arXiv:1604.03423*, 2020. 11, 23

[AMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *Annals of Probability*, 49(6):2922–2960, 2021. 10, 13, 37, 38

[BAH+22] Afonso S. Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S. Wein, and Ilias Zadik. The Franz-Parisi criterion and computational trade-offs in high dimensional statistics. In *Advances in Neural Information Processing Systems, NeurIPS 2021*, volume 35, pages 33831–33844, 2022. 12

[BCMS23] Jean Barbier, Francesco Camilli, Marco Mondelli, and Manuel Sáenz. Fundamental limits in structured principal component analysis and how to reach them. *Proceedings of the National Academy of Sciences*, 120(30), 2023. 6

[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. 13

[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. 11

[BKM+19] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. 6

[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. 11, 13

[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. 13

[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. 13

[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. 12

[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. 13

[Bor19] Charles Bordenave. Lecture notes on random matrix theory, 2019. 11

[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. 13

[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. 12

[CL21] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021. 13

[CM22] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *Annals of Statistics*, 50(1):170–196, 2022. 12

[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. 5, 12

[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. 5, 6, 26, 29

[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, 13, 39

[CZK14] Francesco Caltagirone, Lenka Zdeborová, and Florent Krzakala. On convergence of approximate message passing. In *IEEE International Symposium on Information Theory, ISIT 2014*, pages 1812–1816. IEEE, 2014. 6

[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. 13

[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. 13

[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. 11

[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. 9, 13, 31

[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. 9

[Dur19] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2019. 14, 28

[Fan22] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *Annals of Statistics*, 50(1):197–224, 2022. 13

[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. 11

[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. 13

[Gab20] Marylou Gabrié. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020. 5, 12, 13

[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. 13

[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. 11

[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. 5, 13

[HKP+17] Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The Power of Sum-of-Squares for Detecting Hidden Structures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 720–731. IEEE, 2017. 12

[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. 11

[HS23] Brice Huang and Mark Sellke. Optimization Algorithms for Multi-Species Spherical Spin Glasses. *arXiv preprint arXiv:2308.09672*, 2023. 13

[IS23] Misha Ivkov and Tselil Schramm. Semidefinite programs simulate approximate message passing robustly. *arXiv preprint arXiv:2311.09017*, 2023. 11, 12, 28, 37, 53

[Jan97] Svante Janson. *Gaussian Hilbert spaces*, volume 129 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1997. 15

[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. 13

[Jon22] Christopher Jones. *Symmetrized Fourier Analysis of Convex Relaxations for Combinatorial Optimization Problems*. PhD thesis, The University of Chicago, 2022. 11

[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. 11

[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. 11

[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. 11

[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. 13

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009. 6, 13

[KPX24] Pravesh Kothari, Aaron Potechin, and Jeff Xu. Sum-of-squares lower bounds for ultra-sparse random graphs: Independent set and coloring. *To appear*, 2024. 11

[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. 12

[LFW23] Gen Li, Wei Fan, and Yuting Wei. Approximate message passing from random initialization with applications to Z_2-synchronization. *Proceedings of the National Academy of Sciences*, 120(31):e2302930120, 2023. 13, 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. 13

[LW22] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiV preprint arXiv:2208.03313*, 2022. 13, 38

[LW24] Gen Li and Yuting Wei. A non-asymptotic distributional theory of approximate message passing for sparse and robust regression. *To appear*, 2024. 13

[MKTZ15] Andre Manoel, Florent Krzakala, Eric W. Tramel, and Lenka Zdeborová. Swept Approximate Message Passing for Sparse Estimation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37, pages 1123–1132, 2015. 6

[MM09] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. 5, 6, 7, 13

[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. 5, 12

[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. 10, 37, 38

[MP03] Marc Mézard and Giorgio Parisi. The cavity method at zero temperature. *Journal of Statistical Physics*, 111:1–34, 2003. 6, 7, 12

[MPV86] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. SK Model: The Replica Solution without Replicas. *Europhysics Letters*, 1(2):77, 1986. 6, 12

[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. 5

[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. 11

[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. 13

[MSR73] Paul C. Martin, Eric D. Siggia, and Harvey A. Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973. 13

[MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 49:321–345, 2021. 38

[MV22] Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114003, 2022. 38

[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*, pages 926–937, 2019. 12

[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. 11, 12, 22, 28

[MW22b] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022. 5, 6, 26

[Pan13] Dmitry Panchenko. *The Sherrington–Kirkpatrick model*. Springer Science & Business Media, 2013. 12

[Par79] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979. 6, 12

[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. 6, 12

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. 13

[PR20] Aaron Potechin and Goutham Rajendran. Machinery for Proving Sum-of-Squares Lower Bounds on Certification Problems. *arXiv preprint arXiv:2011.04253*, 2020. 11

[PR22] Aaron Potechin and Goutham Rajendran. Sub-exponential time Sum-of-Squares lower bounds for Principal Components Analysis. *Advances in Neural Information Processing Systems, NeurIPS 2022*, 35:35724–35740, 2022. 11

[RM14] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems, NIPS 2014*, pages 2897–2905, 2014. 39

[RSFS19] Sundeep Rangan, Philip Schniter, Alyson K. Fletcher, and Subrata Sarkar. On the convergence of approximate message passing with arbitrary matrices. *IEEE Transactions on Information Theory*, 65(9):5339–5351, 2019. 6

[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. 11

[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. 11

[RV18] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018. 10, 13, 39

[SS24a] Juspreet Singh Sandhu and Jonathan Shi. A sum-of-squares hierarchy in the absence of pointwise proofs i: Energy certificates. *arXiv preprint arXiv:2401.14383*, 2024. 12

[SS24b] Juspreet Singh Sandhu and Jonathan Shi. A sum-of-squares hierarchy in the absence of pointwise proofs ii: Rounding high-entropy steps. *To appear*, 2024. 12

[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. 13

[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. 13

[Tak21] Keigo Takeuchi. Bayes-optimal convolutional AMP. *IEEE Transactions on Information Theory*, 67(7):4405–4428, 2021. 13

[Tal06] Michel Talagrand. The Parisi formula. *Annals of Mathematics*, pages 221–263, 2006. 12

[Tal10] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010. 12

[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. 11

[VSR+15] Jeremy Vila, Philip Schniter, Sundeep Rangan, Florent Krzakala, and Lenka Zdeborová. Adaptive damping and mean removal for the generalized approximate message passing algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015*, pages 2021–2025. IEEE, 2015. 6

[WZF22] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022. 13

[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. 6, 13

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. 5, 7, 9, 12, 13, 31

[ZY22] Qiuyun Zou and Hongwen Yang. A Concise Tutorial on Approximate Message Passing. *arXiv preprint arXiv:2201.07487*, 2022. 13

A Fourier analytic properties

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

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

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

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

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 $\alpha \in A \setminus \{\circledcirc\}$ we have $\mathsf{E}[Z_{\alpha,i}] = 0$ and*

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

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

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

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

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

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

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

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

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 $\alpha \in T$ is a tree, then the variance of $Z_{\alpha,i}$ is $\Theta(1)$, whereas if α is a connected graph with a cycle, then the variance of $Z_{\alpha,i}$ is $o(1)$.

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

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

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

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

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

B Derivation of Operations on the Diagram Basis

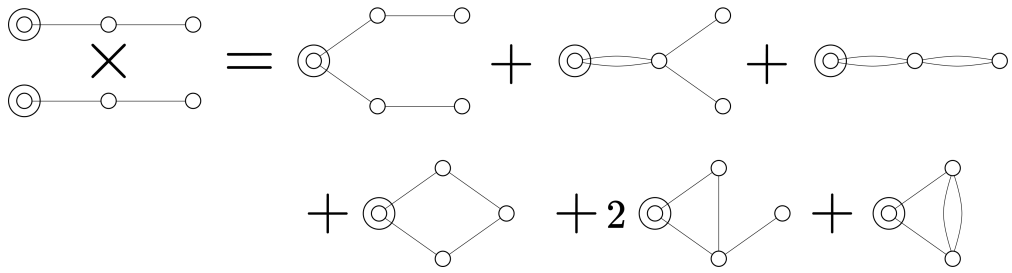
The plan of this section is the following.

- Appendix B.1 is a cheat sheet describing three important combinatorial principles that underlie diagram calculus.
- In Appendix B.2, we derive the effects of GFOM operations on the full diagram expansion. We then deduce their asymptotic effects on tree diagrams.
- In Appendix B.3, we define the repeated-label diagram basis, which is the simplest way to represent a first-order iteration. We do not use the repeated-label diagram basis directly in this paper.

B.1 General combinatorial principles

We demonstrate combinatorial principles for diagram analysis through an example of squaring the 2-path diagram componentwise.

When we multiply two diagrams τ and ρ coordinatewise, by virtue of Definition 3.2 we obtain a sum over pairs of injective embeddings of the vertices of τ and ρ into $[n]$. The joint embedding may not be injective, so we case on the overlap between the two embeddings. In this example, we have the following (non-asymptotic) diagrammatic equality:



which corresponds to the algebraic identity:[17]

$$\left(\sum_{\substack{j,k=1\\ i,j,k \text{ distinct}}}^{n} A_{ij}A_{jk}\right)^2 = \sum_{\substack{j,k,\ell,m=1\\ i,j,k,\ell,m \text{ distinct}}}^{n} A_{ij}A_{jk}A_{i\ell}A_{\ell m} + \sum_{\substack{j,k,\ell=1\\ i,j,k,\ell \text{ distinct}}}^{n} A_{ij}^2 A_{jk}A_{j\ell} + \sum_{\substack{j,k=1\\ i,j,k \text{ distinct}}}^{n} A_{ij}^2 A_{jk}^2$$
$$+ \sum_{\substack{j,k,\ell=1\\ i,j,k,\ell \text{ distinct}}}^{n} A_{ij}A_{jk}A_{i\ell}A_{\ell k} + 2\sum_{\substack{j,k,\ell=1\\ i,j,k,\ell \text{ distinct}}}^{n} A_{ij}A_{jk}A_{ik}A_{k\ell} + \sum_{\substack{j,k=1\\ i,j,k \text{ distinct}}}^{n} A_{ij}A_{ik}A_{jk}^2$$

This is summarized in the following general principle:

[17] The factor of 2 is a "symmetry factor" which appears when performing combinatorics on symmetry-reduced objects such as the diagrams.

Combinatorial Principle 1. *A joint summation over two or more embeddings can be partitioned based on which vertices overlap.*

This re-expresses the product of two diagrams in the diagram basis. Notice that new multiedges may be created.

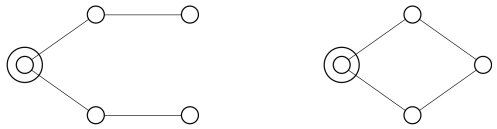
Combinatorial Principle 2. *A hanging double edge can be asymptotically removed. Inductively, a hanging tree of double edges can be asymptotically removed.*

The intuitive justification for this is as follows. If a hanging double edge attaches a vertex j to the hanging vertex k, then the contribution of k to the diagram is A_{jk}^2 summed over $n \quad jV(\alpha)j + 1$ choices for k. We have $\mathsf{E}\left[A_{jk}^2\right] = \frac{1}{n}$ so that in expectation the contribution is $1 \quad \frac{jV(\alpha)j \ 1}{n} \quad 1$. We can remove the hanging double edge up to a negligible multiplicative factor and the negligible fluctuation of the summation over k.

As for the remaining intersection terms, it turns out that they are all negligible.

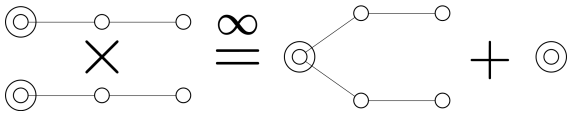
Combinatorial Principle 3. *Let α be a rooted, connected diagram. If removing the hanging double edges creates a tree of single edges, then α is non-negligible. Otherwise, α is negligible.*

The informal explanation is simply that cyclic diagrams sum over fewer terms than tree-shaped diagrams. For example, the left diagram is a sum over n^4 terms while the right diagram is a sum over n^3 terms.[18]



In almost all cases (cycles, non-hanging double edges, higher multiplicity edges, etc), the smaller number of terms makes the diagram negligible. Trees of single edges with hanging trees of double edges are the only non-negligible cases.

Following the combinatorial principles, the asymptotic representation of the componentwise square of the 2-path is:



B.2 Derivation of the asymptotic operations

First, we exactly express the effect of multiplying a diagram by A:

Lemma B.1. *For all diagrams α 2 A,*

$$AZ_\alpha = Z_{\alpha^+} + \sum_{v2V(\alpha)} Z_{\text{contract } v \text{ and } \circledcirc \text{ in } \alpha^+} \cdot$$

[18] In the message passing viewpoint, a vertex v_0 receives $\Theta(n^4)$ messages along paths of distinct vertices (v_0, v_1, v_2) and (v_0, v_3, v_4), whereas it only receives $\Theta(n^3)$ "cyclic" messages along paths $(v_0, v_1, v_2, v_3, v_0)$.

Proof. We write:

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

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

In the asymptotic diagram basis, we start from $\tau \in \mathcal{T}$. Following the asymptotic classification of non-negligible diagrams (Combinatorial Principle 3 or Section 4.2), which diagrams are non-negligible? The only diagrams that do not introduce a cycle of length ≥ 3 are τ^+ and intersections between $\circledcirc$ in τ^+ and the children of the old root. The latter case introduces a double edge, which can only be hanging if $\tau \in \mathcal{S}$ (and in this case it can be removed asymptotically). Hence we conclude

$$AZ_\tau \overset{1}{=} \begin{cases} Z_{\tau^+} + Z_\tau & \text{if } \tau \in \mathcal{S} \\ Z_{\tau^+} & \text{if } \tau \in \mathcal{T} \setminus \mathcal{S} . \end{cases}$$

We now switch to describe the effect of componentwise product. To capture the combinatorics, we define the concepts of intersection patterns and intersection diagrams.

Definition B.2 (Intersection pattern, $P \in \mathcal{P}(\alpha_1, \ldots, \alpha_k)$)**.** *Let $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$. Let α be the diagram obtained by putting all α_i at the same root. An intersection pattern P is a partition of $V(\alpha) \setminus \{\circledcirc\}$ 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 B.3 (Intersection diagram, α_P)**.** *Let $\alpha \in \mathcal{A}$. Given a partition P of $V(\alpha)$, let α_P be the diagram obtained by contracting each block of P into a single vertex. Keep all edges (hence there may be new multiedges).*

By casing on which vertices are equal among the embeddings of $\alpha_1, \ldots, \alpha_k$ as in the proof of Lemma B.1, we have:

Lemma B.4. *For $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$, the componentwise product of $Z_{\alpha_1}, \ldots, Z_{\alpha_k}$ is*

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

Given tree diagrams $\tau_1, \ldots, \tau_k \in \mathcal{T}$, we can identify the asymptotically non-negligible terms in the product 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 $\tau_1, \ldots, \tau_k$ 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 we see that the asymptotically non-negligible terms are the partial matchings between isomorphic subtrees of the root. Two copies of a subtree $\sigma \in S$ can be matched up into a tree of double edges in $|\mathrm{Aut}(\sigma)|$ ways.

B.3 Repeated-label diagram basis

An alternative basis for the diagram space consists of diagrams in which labels are allowed to repeat. This representation was also defined by Ivkov and Schramm [IS23, Section 3.5].

Definition B.5 ($\widetilde{Z}_\alpha$)**.** *For a diagram α 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_α is that the embedding φ must be injective in Z_α. To perform the change of basis in one direction is as easy as replacing $\widetilde{Z}_\alpha$ by a sum of Z_α based on which labels are repeated.

Lemma B.6. *For $\alpha \in A$,*

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

where $\mathcal{P}(\alpha)$ is the set of partitions of $V(\alpha)$ and α_P contracts the blocks of P (Definition B.3).

Proof. We have

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

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

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

Given the current representation $x_t = \sum_{\tau \in T} c_\tau \widetilde{Z}_\tau$ 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 α^+ 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" α and β by merging their roots.

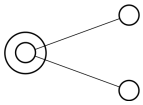
Example B.7. *Returning to the running example,*

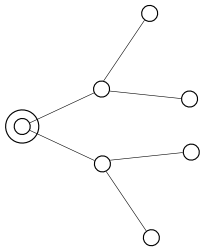
$$x_{t+1} = (Ax_t)^2 \qquad x_0 = \vec{1}$$

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

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

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





C Omitted Proofs

C.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 C.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* α*,* $E_2(\alpha)$ *the multiset of 2-labeled edges and* $E_1(\alpha) = E \setminus E_2(\alpha)$ *the multiset of non-labeled edges.*

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

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

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

We extend the set of diagrams $\mathcal{A}$ to allow diagrams which may have 2-labeled edges. We update the definition of $I(\alpha)$ from Definition 4.1 to incorporate labeled edges (because a labeled edge is mean-0, it is treated like a single edge).

Definition C.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* $\neq 2$ *or is a self-loop, treating 2-labeled edges as if they were normal edges.*

We show the following exact decomposition for removing hanging double edges:

Lemma C.4. *Let α 2 A be a diagram with a hanging (unlabeled) double edge. Let α_0 be α with both the hanging double edge and corresponding hanging vertex removed, and α_2 be α with the hanging double edge replaced by a single 2-labeled edge. Then,*

$$Z_\alpha = Z_{\alpha_0} \quad \frac{jV(\alpha)j \quad 1}{n} \quad Z_{\alpha_0} + Z_{\alpha_2}\,.$$

Proof. We write:

$$\begin{aligned} Z_{\alpha,i} &= \sum_{\substack{\text{injective } \varphi:V(\alpha)\ !\ [n] \\ \varphi(\circledcirc)=i}} A^2_{u,v} \prod_{fx,yg2E(\alpha)nffu,vg,fu,vgg} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{1}{n} \sum_{\substack{\text{injective } \varphi:V(\alpha)\ !\ [n] \\ \varphi(\circledcirc)=i}} \prod_{fx,yg2E(\alpha)nffu,vg,fu,vgg} A_{\varphi(x)\varphi(y)} \\ &= Z_{\alpha_2,i} + \frac{n \quad jV(\alpha)j+1}{n} \quad Z_{\alpha_0,i} = Z_{\alpha_0,i} \quad \frac{jV(\alpha)j \quad 1}{n} Z_{\alpha_0,i} + Z_{\alpha_2,i}\,. \end{aligned}$$

The additional $n \quad jV(\alpha)j+1$ scaling factor comes from removing the hanging vertex. □

C.2 Omitted proofs for Section 4.1

We prove a more specific version of Lemma 4.2

Lemma C.5. *Let q 2 N, α 2 A, and i 2 $[n]$. Then,*

$$\left|\mathsf{E}\left[Z^q_{\alpha,i}\right]\right| \quad M_{qjE(\alpha)j} 2^{qjE(\alpha)j} (qjV(\alpha)j)^{qjV(\alpha)j} \quad n^{\frac{q}{2}(jV(\alpha)j \quad 1 \quad jE(\alpha)j+jI(\alpha)j)}\,,$$

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

$$M_k = \max\left({}_X\mathsf{E}_\mu\left[jXj^k\right],\ {}_X\mathsf{E}_{\mu_0}\left[jXj^k\right]\right)\,.$$

When q and $jV(\alpha)j$ are $O(1)$, the overall bound reduces to

$$\left|\mathsf{E}\left[Z^q_{\alpha,i}\right]\right| \quad O\left(n^{\frac{q}{2}(jV(\alpha)j \quad 1 \quad jE(\alpha)j+jI(\alpha)j)}\right)\,.$$

Proof. We expand $\mathsf{E}\left[Z^q_{\alpha,i}\right]$ as

$$\sum_{\substack{\text{injective } \varphi_1,\ldots,\varphi_q:V(\alpha)\ !\ [n] \\ \varphi_1(\circledcirc)= \quad =\varphi_q(\circledcirc)=i}} \mathsf{E}\left[\prod_{p=1}^{q}\left(\prod_{fu,vg2E_1(\alpha)} A_{\varphi_p(u)\varphi_p(v)}\right)\left(\prod_{fu,vg2E_2(\alpha)}\left(A^2_{\varphi_p(u)\varphi_p(v)} \quad \frac{1}{n}\right)\right)\right]\,.$$

This is a polynomial of degree $qjE(\alpha)j$ in A (by convention every 2-labeled edge contributes 2 to $jE(\alpha)j$). We first estimate the magnitude of any summand of the sum over $\varphi_1,\ldots,\varphi_q$ with nonzero expectation. We decompose each such summand into $2^{qjE_2(\alpha)j}$ terms by expanding

out[19] the $A_{ij}^2 \quad \frac{1}{n}$. We are left with monomials in the entries of A of total degree at most $qjE(\alpha)j$. We bound the expected value of each of these monomials by $M_{qjE(\alpha)j}n^{\ qjE(\alpha)j/2}$ using Hölder's inequality. This shows that any nonzero term in the summation has magnitude at most $2^{qjE_2(\alpha)j}M_{qjE(\alpha)j}n^{\ qjE(\alpha)j/2}$.

To bound the number of nonzero terms, we observe that every edge A_{jk} for $j \neq k$ must occur zero times or at least twice in order to have nonzero expectation (the self-loops A_{jj} can occur any number of times, and the 2-labeled edges $A_{jk}^2 \quad \frac{1}{n}$ must overlap at least one additional edge in order to have nonzero expectation). Each vertex in $V(\alpha) \cap I(\alpha) \cap f\circledcirc g$ 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 $q\left(jV(\alpha)j \quad 1 + jI(\alpha)j\right)/2$ where the $\quad 1$ is used to avoid counting the root.

Hence, there are at most $n^{q(jV(\alpha)j \ \ 1+jI(\alpha)j)/2}$ 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 $(qjV(\alpha)j)^{qjV(\alpha)j}$ 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 $x \overset{1}{=} y$, then $kx \quad yk_1 \overset{a.s.}{!} 0$.*

Proof. We first focus on a single combinatorially negligible term $a_n Z_\alpha$. For any $\varepsilon > 0$ and $i \in [n]$, by Lemma C.5 and Markov's inequality, we have

$$\Pr\left(ja_n Z_{\alpha,i}j \quad \varepsilon\right) \quad \frac{\mathsf{E}\left[(a_n Z_{\alpha,i})^6\right]}{\varepsilon^6} \quad O\left(\frac{1}{n^3\varepsilon^6}\right).$$

By a union bound over all coordinates $i \in [n]$, $\Pr\left(ka_n Z_\alpha k_1 \quad \varepsilon\right) \quad O(n^{\ 2}\varepsilon^{\ 6})$. Applying this to every combinatorially negligible term appearing in $x \quad y$ (there are constantly many such terms by definition), if $x \overset{1}{=} y$ we also have $\Pr\left(kx \quad yk_1 \quad \varepsilon\right) \quad O(n^{\ 2}\varepsilon^{\ 6})$. The conclusion follows from the Borel-Cantelli lemma. □

Lemma 4.6. *If x, y are diagram expressions with $x \overset{1}{=} y$, then*

$$Ax \overset{1}{=} Ay\,.$$

Moreover, if $x_1, \ldots, x_t, y_1, \ldots, y_t$ are diagram expressions with $x_i \overset{1}{=} y_i$ for all $i \in [t]$, then

$$f(x_1, \ldots, x_t) \overset{1}{=} f(y_1, \ldots, y_t)\,,$$

for any polynomial function $f : \mathsf{R}^t \ ! \ \mathsf{R}$ 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 $n^{\ k}AZ_\alpha$ are combinatorially negligible.

[19] The factor $2^{qjE_2(\ )j}$ may be removed with a tighter argument.

(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 $n \ {}^{(k+\ell)}Z_\alpha \quad Z_\beta$ are combinatorially negligible, where $\quad$ is the componentwise product.

For (i), the diagram representation of AZ_α is given by Lemma B.1. In the term α^+ without intersections,

$$\boldsymbol{j}V(\alpha^+)\boldsymbol{j} = \boldsymbol{j}V(\alpha)\boldsymbol{j} + 1\,, \qquad \boldsymbol{j}I(\alpha^+)\boldsymbol{j} = \boldsymbol{j}I(\alpha)\boldsymbol{j}\,, \qquad \boldsymbol{j}E(\alpha^+)\boldsymbol{j} = \boldsymbol{j}E(\alpha)\boldsymbol{j} + 1\,.$$

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

In a term β corresponding to an intersection between the new root and a vertex of α,

$$\boldsymbol{j}V(\beta)\boldsymbol{j} = \boldsymbol{j}V(\alpha)\boldsymbol{j}\,, \qquad \boldsymbol{j}I(\beta)\boldsymbol{j} \quad \boldsymbol{j}I(\alpha)\boldsymbol{j} + 1\,, \qquad \boldsymbol{j}E(\beta)\boldsymbol{j} = \boldsymbol{j}E(\alpha)\boldsymbol{j} + 1\,.$$

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

For (ii), the diagram representation of $Z_\alpha \quad Z_\beta$ is given by Lemma B.4. Fix an intersection pattern $P \in \boldsymbol{P}(\alpha, \beta)$ that has b blocks and denote by γ the resulting diagram. Then,

$$\begin{aligned}
&\boldsymbol{j}V(\gamma)\boldsymbol{j} = b + 1\,,\\
&\boldsymbol{j}E(\gamma)\boldsymbol{j} = \boldsymbol{j}E(\alpha)\boldsymbol{j} + \boldsymbol{j}E(\beta)\boldsymbol{j}\,,\\
&\boldsymbol{j}I(\gamma)\boldsymbol{j} \quad \boldsymbol{j}I(\alpha)\boldsymbol{j} + \boldsymbol{j}I(\beta)\boldsymbol{j} + \boldsymbol{j}V(\alpha)\boldsymbol{j} + \boldsymbol{j}V(\beta)\boldsymbol{j} \quad b \quad 2\,.
\end{aligned}$$

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 $\boldsymbol{j}V(\alpha)\boldsymbol{j} + \boldsymbol{j}V(\beta)\boldsymbol{j} \quad b \quad 2$ intersected non-root vertices in γ.

Putting everything together,

$$\begin{aligned}
&\boldsymbol{j}V(\gamma)\boldsymbol{j} \quad 1 \quad \boldsymbol{j}E(\gamma)\boldsymbol{j} + \boldsymbol{j}I(\gamma)\boldsymbol{j}\\
&\boldsymbol{j}V(\alpha)\boldsymbol{j} \quad 1 \quad \boldsymbol{j}E(\alpha)\boldsymbol{j} + \boldsymbol{j}I(\alpha)\boldsymbol{j} + \boldsymbol{j}V(\beta)\boldsymbol{j} \quad 1 \quad \boldsymbol{j}E(\beta)\boldsymbol{j} + \boldsymbol{j}I(\beta)\boldsymbol{j}\\
&< 2(k+l)\,,
\end{aligned}$$

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 C.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 α has a hanging double edge. Let α_0 be α with the hanging double edge and hanging vertex removed. Then*

$$a_n Z_\alpha \overset{1}{=} a_n Z_{\alpha_0}\,.$$

Proof. Starting from the decomposition of Lemma C.4,

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

we claim that the first term is combinatorially order 1, and the second and third terms are combinatorially negligible. Comparing α_0 to α, two edges and one vertex in $I(\alpha)$ are removed. This does not change the combinatorial order. The second term scales down by n and this becomes negligible (by assumption $|V(\alpha)|$ is constant). In the third term, $|I(\alpha_2)| < |I(\alpha)|$ to take into account the hanging vertex, while $|V(\alpha)| = |V(\alpha_2)|$ and $|E(\alpha)| = |E(\alpha_2)|$ remain unchanged, making the term negligible. We remind the reader that $|E(\alpha)| = |E(\alpha_2)|$ because $|E(\alpha_2)|$ counts 2-labeled edges twice. □

Definition 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.

C.3 Scalar diagrams

We now 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 C.6. *For any proper α 2 $\mathcal{A}_{\text{scalar}}$:*

- *For any proper β 2 $\mathcal{A}_{\text{scalar}}$ such that β 6= α,* $\mathbb{E}[Z_\alpha Z_\beta] = 0$.
- $\mathbb{E}[Z_\alpha] = 0$ *if α is not a singleton.*
- *The second moment of Z_α is*

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

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 B.2) and intersection diagrams (Definition B.3) to allow scalar and vector diagrams simultaneously. The "unintersected" diagram consists of adding all the scalar diagrams as floating components to the vector diagrams, which are put at the same root. The intersection patterns are partitions of this vertex set such that no two vertices from the same diagram are matched.

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

$$Z_{\alpha_1} \quad Z_{\alpha_k} = \sum_{P 2 P(\alpha_1,\ldots,\alpha_k)} Z_{\alpha_P} ,$$

where the product is componentwise for the vector diagrams.

Proof. Analogous to Lemma B.4. □

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

Lemma C.8. *Let q 2 N, α 2 A_{scalar}, and i 2 $[n]$. Then,*

$$\mathsf{E}\,[Z_\alpha^q] \quad M_{qE(\alpha)} 2^{qE(\alpha)} (qV(\alpha))^{qV(\alpha)} \quad n^{\frac{q}{2}(V(\alpha) \quad E(\alpha)+I(\alpha))} ,$$

where M_k is de ned as in Lemma C.5. When q and $V(\alpha)$ are $O(1)$, this reduces to

$$\mathsf{E}\,[Z_\alpha^q] \quad O\left(n^{\frac{q}{2}(V(\alpha) \quad E(\alpha)+I(\alpha))}\right) .$$

Proof. Analogous to Lemma C.5. □

Definition C.9 (Combinatorially negligible and order 1 scalar)**.** *Let $(a_n)_{n2\mathbb{N}}$ be a sequence of real-valued coe cients with $a_n = \Theta(n^{\ k})$, where $k \quad 0$ is such that $2k$ 2 Z. Let α 2 A_{scalar} be a scalar diagram.*

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

$$V(\alpha) \quad E(\alpha) + I(\alpha) \quad 2k \quad 1 .$$

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

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

We de ne $\overset{1}{=}$ for scalar diagram expressions exactly as in De nition 4.4.

Lemma C.10. *Let x and y be scalar diagram expressions with $x \overset{1}{=} y$. Then $x \quad y \overset{a.s.}{!} 0$.*

Proof. Analogous to Lemma 4.5. □

Lemma C.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 C.12. *Let α 2 A_{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 C.1).

Proof. We proceed as in the proof of Lemma 4.10. In each connected component C containing at least one vertex s 2 $V(\alpha)$ n $I(\alpha)$, 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 n fsg, and at least two edges to every vertex in $I(\alpha) \setminus C$. This encoding argument shows that

$$2|I(\alpha) \setminus C| + |(V(\alpha) \cap I(\alpha)) \setminus C| \quad 1 \quad |E(C)|, \tag{12}$$

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

$$2(|I(\alpha) \setminus C| \quad 1) = |V(\alpha) \setminus C| + |I(\alpha) \setminus C| \quad 2 \quad |E(C)|. \tag{13}$$

Summing Eq. (12) and Eq. (13) over all connected components, we obtain

$$|V(\alpha)| \quad |E(\alpha)| + |I(\alpha)| \quad (c \quad c_I) + 2c_I = c + c_I .$$

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 C.13. *Let α 2 A be a vector diagram with c oating 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 oating 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 C.12. □

C.4 Classification of diagrams

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

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 B.4. 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 C.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 σ, the single edge incident to the root must be paired with another such edge. This extends to an automorphism of the entire subtree. In conclusion, $\mathbb{E}\left[Z_{\sigma,i}^q\right]$ converges to $|\mathrm{Aut}(\sigma)|^{q/2}$ times the number of perfect matchings on q objects, and we conclude by Lemma 2.4 and Lemma 2.3. The proof for the scalar case is analogous. □

Lemma C.15. *If $\tau \in T$ consists of d_σ copies of the subtrees $\sigma \in S$, then*

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

For $\rho \in F_{\mathrm{scalar}}$ with c components and consisting of d_τ copies of each tree $\tau \in T_{\mathrm{scalar}}$,

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

Proof. We first expand $h_d(Z_\sigma; |\mathrm{Aut}(\sigma)|)$ in the diagram basis using Lemma B.4 and identify the dominant terms, i.e. those which are combinatorially order 1. As in the proof of Lemma C.14, the combinatorially order 1 terms in each monomial $Z_{\sigma,i}^k$ consist of pairing up copies of the tree σ:

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

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),

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

$$= Z_{d \text{ copies of } \sigma}.$$

This completes the argument when τ consists of several copies of a single $\sigma \in S$. If $\sigma, \sigma' \in S$ are distinct, using again Lemma B.4 and Lemma 4.10, we can check that

$$Z_{d \text{ copies of } \sigma} \quad Z_{d' \text{ copies of } \sigma'} \overset{1}{=} Z_{d \text{ copies of } \sigma \text{ and } d' \text{ copies of } \sigma'}.$$

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

Lemma C.16. *Let $\alpha \in F$ have c floating components. Let $\alpha_\circledcirc$ be the component of the root and α_{float} be the floating components. Then $n^{\frac{c}{2}} Z_\alpha \overset{1}{=} n^{\frac{c}{2}} Z_{\alpha_{\text{float}}} Z_{\alpha_\circledcirc}$.*

Proof. The product $n^{\frac{c}{2}} Z_{\alpha_{\text{float}}} Z_{\alpha_\circledcirc}$ can be expanded in the diagram basis using Lemma C.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 $\frac{1}{n}$ from the lost component which makes the intersection term negligible. □

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

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

$$\mathbb{E}\left[\prod_{p=1}^{q} Z_{\sigma_p, i_p} \prod_{p=1}^{r} n^{\frac{1}{2}} Z_{\tau_p} \right]. \tag{14}$$

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

Expanding Eq. (14) gives a sum over embeddings of the diagrams. We will prove that the dominant terms factor across the distinct (σ_p, i_p) and τ_p; they correspond to pairing up isomorphic σ_p at each distinct root and isomorphic τ_p.

Each nonzero term in the expansion of Eq. (14) 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 B.2. For a given intersection pattern, letting I be the union of the images of the embeddings, the number of terms with this pattern is $(1 \quad o(1))\ n^{|I|\ q_{\text{distinct}}}$ because the q_{distinct} root vertices are fixed. In an embedding with nonzero expectation, every edge appears at least twice, so every non-root vertex is in at least two embeddings. Applying this bound to all of the non-root vertices in I,

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

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

$$n^{jIj \quad q_{\text{distinct}} \quad \frac{jEj+r}{2}} \qquad n^{\frac{1}{2}(jVj \quad q \quad jEj \quad r)}.$$

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

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

Theorem 4.11 follows from Lemma C.14, Lemma C.15, Lemma C.16, and Lemma C.17.

C.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 5.3) will create floating components in the diagrams of the algorithmic state, but all such diagrams will be negligible.

Lemma 4.22. *Let x be a vector diagram expression with asymptotic state X 2 Ω. Then as scalar diagrams, $\frac{1}{n}\sum_{i=1}^{n} x_i \overset{1}{=} \mathsf{E}[X]$.*

Proof. The effect of summing a vector diagram $Z_\alpha = (Z_{\alpha,i})_{i2[n]}$ over i is to unroot α, converting it to a scalar diagram. We prove this operation makes every diagram combinatorially negligible, except for the constant term. For $k \quad 0$ and a vector diagram α 2 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 α is incident to at least one edge of multiplicity 1, then $\frac{a_n}{n}\sum_{i=1}^{n} Z_{\alpha,i}$ is a combinatorially negligible scalar term.

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

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

D High-degree tree diagrams are not Gaussian

Care must be taken when studying the diagrams of superconstant size. In this section we compute that the star-shaped diagram with $\log n$ leaves and the root at a leaf is not Gaussian (its fourth moment is significantly larger than the square of its second moment).[20] 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 γ denote (d-star graph)$^+$. We compute that $\mathsf{E}\left[Z_{\gamma,1}^4\right]$ $\mathsf{E}\left[Z_{\gamma,1}^2\right]^2$ when d $\log n$. By Lemma A.2, the variance is

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

When computing the fourth moment $\mathsf{E}\left[Z_{\gamma,1}^4\right]$ 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 $M(d,d,d,d)$.

Lemma D.1. *Up to a multiplicative* $\mathrm{poly}(d)$ *factor,* $\jmath M(d,d,d,d)\jmath$ & $3^d(d!)^2$.

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

Proof of Lemma D.1. We establish a recursion. There are $(3d)(3d \quad 1) \quad (2d+1)$ 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 D.2. *Let* d, k 2 N *such that* $\frac{d}{k \quad 1}$ *is an integer. Counting the matchings between* d *objects and a subset of* $(k \quad 1)d$ *objects in* $k \quad 1$ *groups, as a function of the number of objects matched in each group, the number of matchings is maximized when there are* $\frac{d}{k \quad 1}$ *matched elements per group.*

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

Proof of Claim D.2. Letting $n_1, \dots, n_{k\ 1}$ be the number of matched elements per group, we may directly compute this number as

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

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

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

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

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

$$\begin{aligned} \jmath M(d,d,d,d)\jmath \ \&\ (3d)(3d\quad 1)\quad (2d+1)\jmath M(2d/3,2d/3,2d/3)\jmath \\ \left(\frac{3d}{e}\right)^{3d}\left(\frac{e}{2d}\right)^{2d}\jmath M(2d/3,2d/3,2d/3)\jmath \end{aligned}$$

where the second equality holds up to a poly(d) factor by Stirling's approximation:

Fact D.3 (Stirling's approximation)**.** *Up to a multiplicative* poly(d) *factor,* $d!\quad \left(\frac{d}{e}\right)^d$*.*

Recursing via the same principle,

$$\begin{aligned} \jmath M(2d/3,2d/3,2d/3)\jmath \ \&\ (4d/3)(4d/3\quad 1)\quad (2d/3+1)\jmath M(d/3,d/3)\jmath \\ = (4d/3)(4d/3\quad 1)\quad (2d/3+1)(d/3)! \\ \left(\frac{4d}{3e}\right)^{4d/3}\left(\frac{3e}{2d}\right)^{2d/3}\left(\frac{d}{3e}\right)^{d/3} \qquad \text{(Fact D.3)} \end{aligned}$$

In total,

$$\jmath M(d,d,d,d)\jmath \ \&\ 3^d\left(\frac{d}{e}\right)^{2d}.$$

□