# Universality of first-order methods on random and deterministic matrices

Nicola Gorini* Chris Jones† Dmitriy Kunisky‡ Lucas Pesenti§

April 21, 2026

**Abstract**

General first-order methods (GFOM) are a flexible class of iterative algorithms which update a state vector by matrix-vector multiplications and entrywise nonlinearities. A long line of work has sought to understand the large-$n$ dynamics of GFOM, mostly focusing on "very random" input matrices and the approximate message passing (AMP) special case of GFOM whose state is asymptotically Gaussian. Yet, it has long remained unknown how to construct iterative algorithms that retain this Gaussianity for more structured inputs, or why existing AMP algorithms can be as effective for some deterministic matrices as they are for random matrices.

We analyze diagrammatic expansions of GFOM via the limiting *traffic distribution* of the input matrix, the collection of all limiting values of permutation-invariant polynomials in the matrix entries, to obtain the following results:

1. We calculate the traffic distribution for the first non-trivial deterministic matrices, including (minor variants of) the Walsh–Hadamard and discrete sine and cosine transform matrices. This determines the limiting dynamics of GFOM on these inputs, resolving parts of longstanding conjectures of Marinari, Parisi, and Ritort (1994).
2. We design a new AMP iteration which unifies several previous AMP variants and generalizes to new input types, whose limiting dynamics are Gaussian conditional on some latent random variables. The asymptotic dynamics hold for a large and natural class of traffic distributions (encompassing both random and deterministic input matrices) and the algorithm's analysis gives a simple combinatorial interpretation of the Onsager correction, answering questions posed recently by Wang, Zhong, and Fan (2022).

*Bocconi University. nicola.gorini@phd.unibocconi.it
†UC Davis. chijones@ucdavis.edu
‡Johns Hopkins University. kunisky@jhu.edu
§ETH Zürich. lpesenti@ethz.ch

# Contents

# 1 Introduction

Complex systems with a large number of simply interacting pieces underlie many natural processes and, more recently, have been studied in computer science in an effort to make sense of how simple machine learning algorithms can learn complex structures latent in large, semi-random input data. Iterative optimization algorithms making sequential updates can be viewed as dynamical systems, with the main task being to understand how the algorithm evolves over time and what properties the eventual output will have.

When the size of these systems grows very large, a key insight from statistical physics is that the macroscopic properties of the system can simplify dramatically:

> *As the size of a random, smoothly-interacting dynamical system grows, the effect of individual particles averages out, and the dynamical system's trajectory approximately follows an asymptotic distributional equation.*

We refer to these distributional equations as *(asymptotic) effective dynamics*. We seek to prove this kind of theorem for discrete-time nonlinear iterative algorithms such as those used in modern optimization, statistics, and machine learning. Concretely, we study *general first-order methods* (GFOM) [CMW20, MW24] which take as input a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, maintain a vector state $\boldsymbol{x} \in \mathbb{R}^n$, and at each step can perform one of two possible operations:

1. either multiply the state by $\boldsymbol{A}$:
$$\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t \,,$$
2. or apply a function $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ componentwise to the previous states:
$$\boldsymbol{x}_{t+1} = f_t(\boldsymbol{x}_t, \dots, \boldsymbol{x}_0), \quad \text{i.e.,} \quad \boldsymbol{x}_{t+1}[i] = f_t(\boldsymbol{x}_t[i], \dots, \boldsymbol{x}_0[i]) \text{ for each } i \in [n] \,.$$

The initial state will be either the deterministic all-ones vector $\boldsymbol{x}_0 = \boldsymbol{1}$, or a random Gaussian vector $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ independent of $\boldsymbol{A}$. Without loss of generality, we may assume that these operations alternate, giving an iteration of the form

$$\boldsymbol{x}_{t+1} = \boldsymbol{A} f_t(\boldsymbol{x}_t, \dots, \boldsymbol{x}_0) \,.$$

We fix some number of iterations $t$ and view $\boldsymbol{x}_t = \boldsymbol{x}_t(\boldsymbol{A})$ as the output of the algorithm.

GFOM is a flexible computational model which is expressive enough to capture many types of gradient descent [CMW20, GTM+24] and message passing algorithms [FVRS22]. It may be viewed as a nonlinear version of the power method for estimating top eigenvectors. The alternation of linear and nonlinear steps also closely matches the structure of a feedforward neural network [CHS24]. One may view the structural restriction on GFOM as forcing $\boldsymbol{x}_t$ viewed as a function of $\boldsymbol{A}$ to be *permutation-equivariant*: if we apply the same permutation to the rows and columns of $\boldsymbol{A}$, then $\boldsymbol{x}_t$ undergoes the same permutation, a natural condition of an algorithm's not depending on the particular indexing of its inputs.

GFOM and their special case of *approximate message passing* (AMP) are very popular algorithms for many statistical inference tasks and are known to perform optimally in various such settings [DMM09, Ran11, Mon12, RFSK16, BM11, FVRS22]. In these cases, an algorithm takes as

input not an arbitrary matrix $\boldsymbol{A}$, but one that contains a corrupted observation of a signal (in a common example, the input $\boldsymbol{A}$ is a low-rank $\boldsymbol{y}\boldsymbol{y}^\top$ plus independent random noise).

GFOM have also been used as optimization algorithms in average-case settings without any such planted structures. For instance, they are the best known algorithms for optimizing quadratic forms with random coefficients over the non-negative orthant [MR15] (the *non-negative PCA* objective function), other convex cones [DMR14], and the hypercube [Mon19] (the *Sherrington–Kirkpatrick Hamiltonian*), all of which are NP-hard problems in the worst case. This situation is the main target of our analysis. We receive an input matrix $\boldsymbol{A}$ without any particular "signal" and wish to output $\boldsymbol{x}$ approximately solving an optimization problem parametrized by $\boldsymbol{A}$, such as

$$\begin{array}{ll} \text{maximize} & \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle \\ \text{subject to} & \boldsymbol{x} \in S \end{array} \tag{1}$$

studied in the above references for various choices of the constraint set $S \subseteq \mathbb{R}^n$.

To view GFOM as an instance of the physical setting sketched above, we consider a growing sequence of matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n \times n}$, and think of the "particles" as being the coordinates $\boldsymbol{x}_t[i]$ of $\boldsymbol{x}_t \in \mathbb{R}^n$. To keep notation reasonable, while all of these objects depend on $n$, we omit the $(n)$ superscript whenever possible. We analyze the *empirical distribution* of our particles, accessed by sampling a random coordinate of a vector:

$$\mathrm{samp}(\boldsymbol{x}) := \boldsymbol{x}[i] \in \mathbb{R} \text{ for } i \sim \mathrm{Unif}([n])\,.$$

In order to study a particle's entire trajectory more generally, we may "stack" several vectors and define $\mathrm{samp}((\boldsymbol{x}_0, \dots, \boldsymbol{x}_t)) := \mathrm{samp}(\boldsymbol{x}_0, \dots, \boldsymbol{x}_t) := (\boldsymbol{x}_0[i], \dots, \boldsymbol{x}_t[i]) \in \mathbb{R}^{t+1}$ for $i \sim \mathrm{Unif}([n])$.

The analysis of GFOM hinges on the observation that these random variables often converge in distribution to certain limiting distributions. That is, for suitably nice test functions $\varphi : \mathbb{R}^{t+1} \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\, \varphi(\mathrm{samp}(\boldsymbol{x}_0^{(n)}, \dots, \boldsymbol{x}_t^{(n)})) = \lim_{n\to\infty} \mathbb{E}\, \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_0^{(n)}[i], \dots, \boldsymbol{x}_t^{(n)}[i]) = \int \varphi \, \mathrm{d}\nu_{\le t}^{\infty}\,,$$

for some probability measures $\nu_{\le t}^{\infty}$. For example, we can analyze the objective function of a problem like Eq. (1) in this way: given a GFOM to run for $t$ iterations producing $\boldsymbol{x}_t = \boldsymbol{x}_t(\boldsymbol{A})$, we extend it to $\boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t$ so that

$$\mathbb{E}\, \frac{1}{n} \langle \boldsymbol{x}_t, \boldsymbol{A}\boldsymbol{x}_t \rangle = \mathbb{E} \frac{1}{n} \langle \boldsymbol{x}_t, \boldsymbol{x}_{t+1} \rangle = \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_t[i] \boldsymbol{x}_{t+1}[i]\,,$$

a quantity accessible in the above formalism by a suitable choice of $\varphi$. We can also study the algorithm's convergence by expanding $\frac{1}{n}\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|_2^2$ in the same way.

The goal of an asymptotic effective dynamics is then to identify the asymptotic measures $\nu_{\le t}^{\infty}$. Such a description is a natural first step to designing *optimal* GFOM for optimization problems: given an explicit description of the limiting performance of any GFOM, we then optimize the performance over all GFOM [CMW20, EMS21, MW25, Pes26].

The goal of this paper is to study the following three questions regarding effective dynamics:

1. **Existence:** What are minimal assumptions on the input matrices and the algorithm that ensure the existence of asymptotic effective dynamics?

2. **Universality:** What properties of the sequence of input matrices $\boldsymbol{A}^{(n)}$ determine the asymptotic effective dynamics? In particular, how can we show that two sequences of $\boldsymbol{A}^{(n)}$ share the same dynamics?

3. **Explicit Calculation:** What are the effective dynamics? In particular, for a given algorithm, how can one describe $\nu_{\leq t}^{\infty}$ for each fixed $t \in \mathbb{N}$?

### 1.1 Approximate message passing and simple effective dynamics

The majority of results to date on effective dynamics for GFOM, including ours, are most useful for *Approximate Message Passing* (AMP) algorithms. Originating from physicists' work on mean-field spin glass models [MPV87, DMM09], AMP algorithms are a special case of GFOM with very simple effective dynamics: each distribution $\nu_t^{\infty}$ (the marginal distribution of $\nu_{\leq t}^{\infty}$ above on the last coordinate) is a Gaussian distribution,

$$\nu_t^{\infty} = \mathcal{N}(\mu_t, \sigma_t^2)\,,$$

and the effective dynamics gives $(\mu_{t+1}, \sigma_{t+1}^2)$ in terms of $(\mu_t, \sigma_t^2), \ldots, (\mu_0, \sigma_0^2)$ via a formula known as the *state evolution* equation. This gives a simple yet complete description of the leading-order behavior of an algorithm as $n \to \infty$. In part due to the power afforded by such a description, AMP (and the closely related *belief propagation*, of which AMP is a limit in a suitable sense) has taken on an indispensable role in statistical physics [MPV87, MM09, CMP+23] and, more recently, in computational statistics [ZK16, FVRS22].

In fact, while the original appearances of AMP in statistical physics were intrinsically motivated, for statistics applications the simplicity of state evolution is so useful that a line of work has emerged trying to *design* GFOM that have Gaussian $\nu_t^{\infty}$ and effective dynamics given by state evolution [JM13, BSK15, VSR+15, Fan22, ZWF24, LWF25]. The term "AMP" is now often used to describe any choice of GFOM for a given family of inputs $\boldsymbol{A}^{(n)}$ that has these properties. While it is not clear that this should be the case *a priori*, a common fortuitous coincidence is that, for various problems, the best GFOM algorithms (in the sense of achieving optimal rates in estimation or inference tasks) happen to be in the special class of AMP. That is, in many cases, the GFOM with the simplest asymptotic effective dynamics are also the most useful in applications.

Given the successes of AMP, it is a longstanding goal in the literature to identify AMP-like algorithms for as many different choices of inputs and input distributions as possible. Yet, even to go slightly beyond the simplest choices of matrices $\boldsymbol{A}^{(n)}$ has proved challenging and subtle (e.g., random matrices with i.i.d. entries [JM13, BLM15], orthogonally invariant distributions [Fan22], or semi-random ensembles [DLS23, WZF22]). Constructing AMP algorithms in such settings involves carefully inserting so-called *Onsager correction terms* into the nonlinearities $f_t$ in ways that remain somewhat mysterious yet are crucial to obtain Gaussian limiting behavior.

Here, we will present an approach to the analysis of GFOM that re-derives different existing variants of AMP in a unified way, derives AMP algorithms for new inputs (both random and deterministic), and offers new conceptual insights into the design of these algorithms and into the proof of their asymptotic effective dynamics, in particular giving a clear combinatorial explanation for the Onsager corrections mentioned above.

## 1.2 Our contributions: Combinatorial method for GFOM

We study GFOM by expressing them as vectors of polynomials in the entries of the input matrix. For this reason we focus on polynomial $f_t$; it is likely possible to treat more general nonlinearities by approximating them by polynomials (see Section 1.3 for some discussion).

**Definition 1.1.** *We call a GFOM as described above a polynomial GFOM (pGFOM) if all nonlinearities $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ are polynomials.*

Our approach is divided into two parts. The first is a "static" analysis of certain symmetric polynomials in the entries of the input $\boldsymbol{A}$. The second translates this to "dynamic" information about vector-valued functions, allowing us to calculate effective dynamics for $O(1)$ iterations of GFOM in a general way.

### 1.2.1 Statics of graph polynomials: Traffic distributions and universality

The basic objects of study for our static analysis are the following *graph polynomials*.

**Definition 1.2** (Diagram classes)**.** *We write $\mathcal{A} = \mathcal{A}_0$ for the set of finite, undirected, connected (multi)graphs. We also write $\mathcal{E} = \mathcal{E}_0 \subseteq \mathcal{A}_0$ for the set of 2-edge-connected (multi)graphs (ones that cannot be disconnected by removing any single edge) and $\mathcal{C} = \mathcal{C}_0 \subseteq \mathcal{E}_0 \subseteq \mathcal{A}_0$ for the set of cactus graphs, ones where every edge belongs to exactly one simple cycle.*[1] *See Figure 1.*

The optional subscript "0" of the diagram classes refers to the outputs of the polynomials being 0-dimensional, i.e., scalars, which will be useful to distinguish them from vector- and matrix-valued polynomials to be defined later (with subscript "1" and "2", respectively).

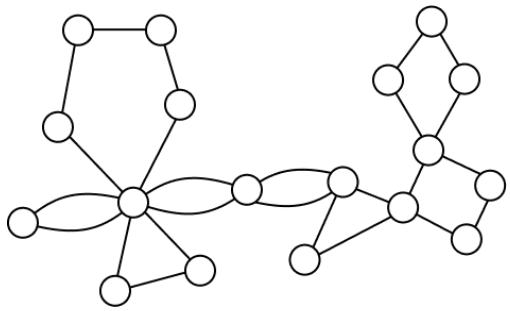


Figure 1: A cactus graph in $\mathcal{C}$. Intuitively, a cactus is a "tree of cycles".

**Definition 1.3** (Scalar graph polynomials)**.** *Given $\alpha \in \mathcal{A}$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, define polynomials $w_\alpha(\boldsymbol{A}), z_\alpha(\boldsymbol{A}) \in \mathbb{R}[\boldsymbol{A}]$ by:*

$$w_\alpha(\boldsymbol{A}) = \sum_{i:V(\alpha)\to[n]} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[i(u), i(v)]\,,$$
$$z_\alpha(\boldsymbol{A}) = \sum_{i:V(\alpha)\hookrightarrow[n]} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[i(u), i(v)]\,.$$

[1] This notion is sometimes more specifically called a *bridgeless cactus*; in this paper we take this to be part of the definition of a cactus.

That is, $w_\alpha(\boldsymbol{A})$ and $z_\alpha(\boldsymbol{A})$ are each multivariate polynomials in the $\frac{n(n+1)}{2}$ entries on and above the diagonal of the matrix $\boldsymbol{A}$ obtained by summing over all labelings of the vertices of $\alpha$ by $[n] = \{1, 2, \ldots, n\}$ and with each edge corresponding to an entry of $\boldsymbol{A}$. The only difference between $w_\alpha(\boldsymbol{A})$ and $z_\alpha(\boldsymbol{A})$ is that the vertex labeling for $z_\alpha(\boldsymbol{A})$ is restricted to be injective by the notation $i : V(\alpha) \hookrightarrow [n]$ whereas labels in $w_\alpha(\boldsymbol{A})$ are allowed to repeat.

Each monomial in the entries of $\boldsymbol{A}$ can be represented as a multigraph on $\{1, 2, \ldots, n\}$. By summing all monomials with the same "shape", the $w_\alpha(\boldsymbol{A})$ and $z_\alpha(\boldsymbol{A})$ give two different spanning sets for a subspace of the $S_n$-invariant polynomials in the entries of $\boldsymbol{A}$, where $S_n$ acts on $\boldsymbol{A}$ by permuting the rows and columns simultaneously. There are only a few possible distinct shapes for monomials with low degree, so analysis on the $w$ or $z$ polynomials is a highly compressed way to analyze $S_n$-invariant low-degree polynomial functions of $\boldsymbol{A}$.

The limiting values of the graph polynomials are a basic set of parameters for the sequence of matrices $\boldsymbol{A}^{(n)}$, introduced in random matrix theory by Male [Mal20], who termed them the *traffic distribution*.

**Definition 1.4** (Traffic distribution)**.** *For a sequence of random*[2] *matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ we say that $\mathcal{D} : \mathcal{A} \to \mathbb{R}$ is the (limiting) traffic distribution of $\boldsymbol{A}$ if*

$$\lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A}) = \mathcal{D}(\alpha) \text{ for all } \alpha \in \mathcal{A}. \tag{2}$$

*We say the (limiting) traffic distribution exists if the limit exists for all $\alpha \in \mathcal{A}$.*[3]

When the limiting traffic distribution exists, it is easy to show that it determines the asymptotic behavior of all constant-time GFOM algorithms with input $\boldsymbol{A}$:

**Claim 1.5.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ have traffic distribution $\mathcal{D}$, and that a pGFOM defines $\boldsymbol{x}_t = \boldsymbol{x}_t(\boldsymbol{A})$ with $\boldsymbol{x}_0 = \mathbf{1}$. Then, for any fixed $t$ and polynomial $\varphi \in \mathbb{R}[x]$,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{A}} \frac{1}{n} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_t[i]) = C,$$

*where $C$ is a constant depending only on $\mathcal{D}$, $(f_s)_{1\le s\le t}$, and $\varphi$.*

Because of this observation, the traffic distribution is a natural way both to show existence of effective dynamics for constant-time GFOM (when the traffic distribution exists then so do effective dynamics) and to characterize the *universality class* of GFOM (when two sequences of matrices have the same traffic distribution then they have the same effective dynamics).

We now reach our first main contribution: by calculating their limiting traffic distributions, we obtain the first analysis of GFOM on non-trivial completely deterministic inputs. Namely, we prove that any delocalized orthogonal matrix, after a slight modification, has the same traffic distribution as a corresponding random matrix model, the *regular random orthogonal model* (r-ROM; see Definition 2.5).

**Theorem 1.6** (See Theorem 5.1)**.** *Let $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(n)} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{H} = \boldsymbol{H}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be a sequence*

[2] Deterministic matrices are also allowed just by taking a constant distribution.

[3] Note that the diagram $\alpha$ cannot depend on $n$. It has constant size as $n \to \infty$.

*of orthogonal matrices such that*

$$\max_{1\le i,j\le n} |\boldsymbol{H}[i,j]| \le n^{-\frac{1}{2}+o(1)} . \tag{3}$$

*Then, the traffic distribution of* $\boldsymbol{\Pi H \Pi}$ *exists and equals that of the* r-ROM*.*

The motivating examples for Theorem 1.6 are "Fourier transform matrices" such as the Walsh–Hadamard matrix (Definition 2.3), discrete sine transform matrix, or discrete cosine transform matrix (Definition 2.4). We call conjugating by the projection matrix $\boldsymbol{\Pi}$ *puncturing* the matrix. Theorem 1.6 implies that, after puncturing, the effective dynamics of GFOM on these matrices are the same as those for the r-ROM, which itself is a punctured version of the *random orthogonal model* (ROM) of [MPR94a]. Explicit state evolution equations for these dynamics are given in Theorem 6.29.

Puncturing is necessary in Theorem 1.6 and is natural for Fourier transform matrices. For the Walsh–Hadamard matrix, puncturing removes the first row and column, all of whose entries are identically $1/\sqrt{n}$. This row/column makes $\boldsymbol{H}\mathbf{1}$ have a single large entry; because of that imbalance, without puncturing the traffic distribution of $\boldsymbol{H}$ does not exist[4] and some GFOMs do not have well-defined asymptotic dynamics. This phenomenon has also been observed experimentally: [Sch20] writes that "structured matrices (e.g., DCT, Hadamard, Fourier) should work as well as i.i.d. random ones. But, in practice, AMP often diverges with such structured matrices." We propose, and our results corroborate, that it is precisely alignment with the all-ones vector that causes this behavior.

Showing that Fourier transform matrices are pseudorandom orthogonal matrices has been a longstanding folklore open problem in the statistical physics and AMP literature. It seems to originate in the work of [MPR94a, MPR94b, PP95] in statistical physics, who proposed these matrices as couplings for spin glass models. Recently (nearly 30 years later), [DLS23] summarized the situation as follows:

> More generally, numerical studies reported in the literature *[...]* suggest that AMP algorithms exhibit universality properties as long as the eigenvectors are generic. Formalizing this conjecture remains squarely beyond existing techniques, and presents a fascinating challenge.

Similar comments have been made in [JM12, RSFS19, BSK15], and relevant numerical experiments can be found in [ÇO19, ABKZ20, DLS23]. Fourier transform matrices are also favored in compressed sensing applications since they admit fast multiplications via the Fast Fourier Transform [WZF22, Example 2.26].

Although Theorem 1.6 concerns orthogonal matrices, we also prove generally that after puncturing, any sequence of delocalized matrices has the same traffic distribution as the orthogonally invariant ensemble with the same eigenvalue distribution, assuming stronger delocalization properties than Eq. (3). See Theorem 5.3 for the formal statement.

### 1.2.2 Cactus properties: conditions for simple traffic distributions

The traffic distribution is a complicated object in general, just because its indexing set $\mathcal{A}$ is very large. Fortunately, traffic distributions of many common matrices are much simpler. Specifically,

[4] For example, when $\boldsymbol{H}$ is the Walsh–Hadamard matrix, the degree-$D$ star diagram $\sigma_D$ satisfies $\frac{1}{n}|w_{\sigma_D}(\boldsymbol{H})| = \Theta(n^{D/2-1})$, which diverges for $D > 2$.

they often satisfy a *cactus property*: almost all of the graph polynomials $z_\alpha(\boldsymbol{A})$ are asymptotically negligible as $n \to \infty$, with the only exceptions being the cactus graphs $\alpha \in \mathcal{C} \subsetneq \mathcal{A}$ (in the $z$ basis, but *not* in the $w$ basis).

**Definition 1.7** (Cactus properties and cactus type)**.** *For a sequence of symmetric matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}$, we say that:*

*(i)* *$\boldsymbol{A}$ has the strong cactus property if $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A}) = 0$ for all $\alpha \in \mathcal{A} \setminus \mathcal{C}$.*

*(ii)* *$\boldsymbol{A}$ has the weak cactus property if $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A}) = 0$ for all $\alpha \in \mathcal{E} \setminus \mathcal{C}$.*

*(iii)* *$\boldsymbol{A}$ has the factorizing (strong or weak) cactus property if it has the (strong or weak) cactus property, and for each $\sigma \in \mathcal{C}$ we have $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\sigma(\boldsymbol{A}) = \prod_{\rho\in \mathrm{cyc}(\sigma)} \kappa_{|\rho|}$ for some real numbers $\kappa_q$, where $\mathrm{cyc}(\sigma)$ is the set of cycles of a cactus and $|\rho|$ is the length of a cycle.*[5]

The idea that the non-negligible diagrams for many random matrix models are cactuses appeared in the physics literature as early as the 1990s [PP95, MFC+19] and we will show in Appendix A how it can be derived from the *Feynman diagram expansion* widely used in physics. More recent mathematical work [Mal20, CDM24] reviewed in Section 4 has rigorously established the strong cactus property for Wigner matrices and unitarily invariant matrices whose eigenvalue distributions converge weakly. In fact, the factorizing strong cactus property is essentially equivalent to $\boldsymbol{A}$ having the same limiting traffic distribution as some orthogonally invariant random matrix model.

The strong cactus property implies that the traffic distribution is specified only by the limiting values associated to $\sigma \in \mathcal{C}$, a much smaller set of graphs than $\mathcal{A}$. Another way to say this is that, under the strong cactus property, the traffic distribution contains no extra information beyond the considerably simpler *diagonal distribution*, introduced by [WZF22].

**Definition 1.8** (Diagonal distribution)**.** *For a sequence of symmetric matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}$, we say that $\mathcal{D} : \mathcal{C} \to \mathbb{R}$ is the limiting diagonal distribution of $\boldsymbol{A}$ if*

$$\lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} w_\sigma(\boldsymbol{A}) = \mathcal{D}(\sigma) \text{ for all } \sigma \in \mathcal{C}.$$

*We say the diagonal distribution exists if the limit exists for all $\sigma \in \mathcal{C}$.*

Let us make several important observations about the definitions of the traffic distribution, the diagonal distribution, and the cactus properties.

First, note that Definition 1.7 is stated in the $z$-polynomial basis, whereas Definitions 1.4 and 1.8 are stated in the $w$-polynomial basis. Throughout the paper, it will be helpful to move back and forth between these bases, since some properties are most natural (or even are only true) in one basis or the other. This can be done via Möbius inversion, as described in Section 3.3.

Second, neither the diagonal distribution nor the traffic distribution is an actual probability distribution. Instead, they should be interpreted as specifying limiting moments of certain empirical distributions, namely, the empirical distributions of the entries of vector graph polynomials.[6]

[5] In the traffic probability literature, the factorizing strong cactus property has been referred to as a traffic distribution being *of cactus type* [CDM24]. The parameters $\kappa_q$ are the *free cumulants* appearing in free probability theory.

[6] The reason for the name of the diagonal distribution $\mathcal{D}$ is that it can also be interpreted as specifying the moments of the empirical distribution over the diagonal of certain matrices, namely those that can be formed from $\boldsymbol{A}$ by matrix multiplication and the operation of zeroing out the off-diagonal entries of a matrix [WZF22].

Third, one can view the diagonal and traffic distributions as generalizations of the limiting spectral distribution of a sequence of matrices. The spectral moments are $\frac{1}{n}\operatorname{Tr}(\boldsymbol{A}^q) = \frac{1}{n} w_\alpha(\boldsymbol{A})$, where $\alpha$ is the $q$-cycle diagram, so they are included in both the diagonal and traffic distributions:

$$\text{“ spectral distribution } \subsetneq \text{ diagonal distribution } \subsetneq \text{ traffic distribution ”}$$

Just as the empirical spectral distribution characterizes the limiting behavior of all polynomials in $\boldsymbol{A}$ that are invariant under the action of the orthogonal group $O(n)$ (acting by $\boldsymbol{Q} \cdot \boldsymbol{A} = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top$), the traffic distribution characterizes the limiting behavior of the larger space of polynomials invariant under the smaller symmetric group $S_n$, i.e., where $\boldsymbol{Q}$ is restricted to be a permutation matrix.

Finally, the strong cactus properties describe when these inclusions can be reversed: if the strong cactus property holds, then the traffic distribution contains no more information than the diagonal distribution. If the factorizing strong cactus property holds, then the diagonal distribution, in turn, contains no more information than the spectral distribution.

Due to the effect of the puncturing operation, the strong cactus property actually is *not* satisfied by the pseudorandom matrices or r-ROM matrices appearing in our Theorem 1.6. But, these matrices satisfy the weak cactus property, and establishing this is a key step in the analysis of these matrices (in fact, the weak cactus property holds for the Fourier transform matrices without puncturing, as we show in Part 1 of Theorem 5.3).

### 1.2.3 Dynamics of graph polynomials: asymptotic GFOM state and treelike AMP

Recall that our final goal is to describe the state $\boldsymbol{x}_t = \boldsymbol{x}_t(\boldsymbol{A})$ of a GFOM. Since $\boldsymbol{x}_t \in \mathbb{R}^n$ we use vector diagrams for this task. Compared to the scalar diagrams in $\mathcal{A}_0$, the only extra information in these diagrams is that one of the vertices is specially marked as the “root”, whose label specifies the coordinate of the vector output.

**Definition 1.9** (Vector diagram classes)**.** *We write $\mathcal{A}_1$ and $\mathcal{C}_1$ for the set of graphs in $\mathcal{A}$ and $\mathcal{C}$ respectively, further decorated with a distinguished root vertex. For $\alpha \in \mathcal{A}_1$, we write* $\operatorname{root}(\alpha) \in V(\alpha)$ *for its root vertex.*

**Definition 1.10** (Vector graph polynomials)**.** *Given $\alpha \in \mathcal{A}_1$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, define vectors of polynomials $w_\alpha(\boldsymbol{A}), z_\alpha(\boldsymbol{A}) \in (\mathbb{R}[\boldsymbol{A}])^n$ by,*

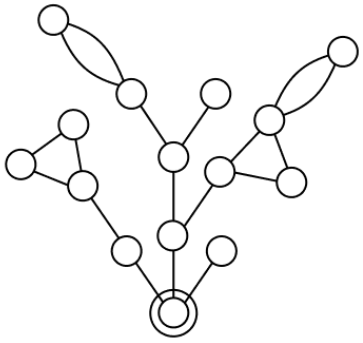
$$\boldsymbol{w}_\alpha(\boldsymbol{A})[i] := \sum_{\substack{j:V(\alpha)\to[n] \\ j(\operatorname{root}(\alpha))=i}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[j(u), j(v)]\,,$$

$$\boldsymbol{z}_\alpha(\boldsymbol{A})[i] := \sum_{\substack{j:V(\alpha)\hookrightarrow[n] \\ j(\operatorname{root}(\alpha))=i}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[j(u), j(v)]\,,$$

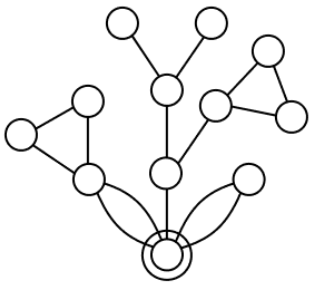
*for all $i \in [n]$.*

To analyze the vector graph polynomials, we compute the moments of the empirical distribution of their entries. We will see that these are matched (asymptotically) by a family of scalar random variables $Z^\infty_\alpha$, so the empirical distribution of the entries of $\boldsymbol{z}_\alpha(\boldsymbol{A})$ converges in a suitable sense to $Z^\infty_\alpha$ as $n \to \infty$. Further, when $\boldsymbol{A}$ has the strong cactus property, an analogous property is inherited by these limiting distributions, only a small number of $\alpha \in \mathcal{A}_1$ having a non-negligible limit.

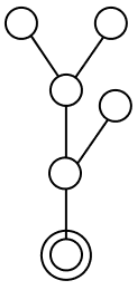
**Definition 1.11** (Treelike diagrams)**.** *We say that $\alpha \in \mathcal{A}_1$ is treelike if it is a tree with hanging cactuses attached to the leaves of the tree (see Figure 2). We denote the set of treelike diagrams by $\mathcal{T}_1$, and denote by $\mathcal{G}_1 \subseteq \mathcal{T}_1$ the set of treelike diagrams in which, after removing hanging cactuses, the root has degree exactly 1.*



(a) A treelike diagram in $\mathcal{T}_1$



(b) A Gaussian diagram in $\mathcal{G}_1$



(c) The diagram in $\mathcal{G}_1$ after removing its hanging cactus.

Figure 2: Examples of treelike and Gaussian diagrams. The root vertex is circled.

**Theorem 1.12** (Vector polynomial limits; see Theorem 6.2)**.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ has the strong cactus property with limiting diagonal distribution $\mathcal{D}$. Assume also that the sequence of random variables $(\|\boldsymbol{A}^{(n)}\|)_{n\geq 1}$ is tight,*[7] *i.e., that*

$$\text{for all } \varepsilon > 0 \text{ there exists } K > 0 \text{ such that } \sup_{n\geq 1} \Pr\Big(\|\boldsymbol{A}^{(n)}\| > K\Big) \leq \varepsilon\,. \tag{4}$$

*Write $\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}) \in (\mathbb{R}^{\mathcal{A}_1})^n$ for the stacking of values of all $\boldsymbol{z}_\alpha(\boldsymbol{A})$ for $\alpha \in \mathcal{A}_1$. Then,*

$$\operatorname{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A})) \xrightarrow[n\to\infty]{\text{(d)}} (Z^\infty_\alpha)_{\alpha\in\mathcal{A}_1}\,,$$

*for a family of (partially dependent) random variables $(Z^\infty_\alpha)_{\alpha\in\mathcal{A}_1}$ such that $Z^\infty_\alpha = 0$ for all $\alpha$ not treelike, and which can be sampled as follows for $\alpha \in \mathcal{T}_1$:*

1. *Draw $(Z^\infty_\sigma)_{\sigma\in\mathcal{C}_1}$ from a distribution determined by $\mathcal{D}$.*
2. *Draw $(Z^\infty_\gamma)_{\gamma\in\mathcal{G}_1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^\infty)$ from a centered Gaussian distribution with countably infinite covariance matrix $\boldsymbol{\Sigma}^\infty$ depending on $(Z^\infty_\sigma)_{\sigma\in\mathcal{C}_1}$.*
3. *Set $(Z^\infty_\alpha)_{\alpha\in\mathcal{T}_1\setminus(\mathcal{G}_1\cup\mathcal{C}_1)}$ to be certain deterministic polynomial functions of $(Z^\infty_\alpha)_{\alpha\in\mathcal{G}_1\cup\mathcal{C}_1}$.*

We note that $\operatorname{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}))$ is a random variable taking values in $\mathbb{R}^{\mathcal{A}_1}$, a countable product space. Thus, its convergence in distribution is the same as convergence in distribution of any finite-dimensional projection; see Appendix C.

The application to pGFOM is as follows. Analogously to Claim 1.5, it is easy to see that the

[7] If the matrices $\boldsymbol{A}^{(n)}$ are deterministic, this should be understood as $(\|\boldsymbol{A}^{(n)}\|)_{n\geq 1}$ being bounded.

iterates $\boldsymbol{x}_t(\boldsymbol{A})$ of a pGFOM admit a diagrammatic expansion of the form

$$\boldsymbol{x}_t(\boldsymbol{A}) = \sum_{\alpha \in \mathcal{A}_1} c_{t,\alpha} \boldsymbol{z}_\alpha(\boldsymbol{A}) \,, \tag{5}$$

for finitely supported coefficients $(c_{t,\alpha})_{\alpha \in \mathcal{A}_1}$. Given the limits of the individual diagrams above, for a given GFOM, number of iterations $t$, and coefficients as in Eq. (5), we write

$$(X_0^\infty, \dots, X_t^\infty) := \left( \sum_{\alpha \in \mathcal{A}_1} c_{0,\alpha} Z_\alpha^\infty, \dots, \sum_{\alpha \in \mathcal{A}_1} c_{t,\alpha} Z_\alpha^\infty \right) ,$$

a random variable in $\mathbb{R}^{t+1}$ that describes the joint empirical distribution of the first $t$ steps of the GFOM. We call this the *asymptotic state* of a GFOM (Definition 6.16). By Theorem 1.12, the asymptotic state describes limiting empirical averages over the GFOM states, in the sense that

$$\lim_{n \to \infty} \mathop{\mathbb{E}}_{\boldsymbol{A}} \frac{1}{n} \sum_{i=1}^n \varphi(\boldsymbol{x}_0[i], \dots, \boldsymbol{x}_t[i]) = \mathbb{E}\, \varphi(X_0^\infty, \dots, X_t^\infty)$$

for any $\varphi : \mathbb{R}^{t+1} \to \mathbb{R}$ either a polynomial or a bounded continuous function (Lemma 6.17).

In particular, if the only nonzero $c_{t,\alpha}$ in Eq. (5) are non-treelike $\alpha$ or treelike $\alpha \in \mathcal{G}_1$, then the GFOM has an asymptotic state that is Gaussian conditional on $(Z_\sigma^\infty)_{\sigma \in \mathcal{C}_1}$. This observation leads to our second main contribution: a new family of *treelike AMP* algorithms simultaneously generalizing Orthogonal Approximate Message Passing (OAMP) algorithms [RSF19, Fan22] for orthogonally invariant matrices, and Generalized Approximate Message Passing (GAMP) algorithms [Ran11, JM13] for matrices with independent entries that are not necessarily identically distributed.[8]

**Theorem 1.13** (Treelike AMP; see Theorem 6.18)**.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies the assumptions of Theorem 1.12. Given polynomial functions $f_t : \mathbb{R} \to \mathbb{R}$, define the pGFOM:*

$$\boldsymbol{x}_0 := \boldsymbol{1} \,, \qquad \boldsymbol{x}_t := \boldsymbol{A} \boldsymbol{f}_{t-1} - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot \boldsymbol{f}_s \,, \qquad \textit{(The product } \boldsymbol{b}_{s,t} \cdot \boldsymbol{f}_s \textit{ is entrywise.)}$$

$$\boldsymbol{f}_t := f_t(\boldsymbol{x}_t) \,, \qquad \boldsymbol{f}_t' := f_t'(\boldsymbol{x}_t) \,.$$

$$\boldsymbol{b}_{s,t}[i] := \sum_{\substack{i_s, \dots, i_{t-1} = 1 \\ \text{distinct} \\ i_s = i}}^n \boldsymbol{A}[i_s, i_{t-1}] \boldsymbol{f}_{t-1}'[i_{t-1}] \boldsymbol{A}[i_{t-1}, i_{t-2}] \boldsymbol{f}_{t-2}'[i_{t-2}] \cdots \boldsymbol{f}_{s+1}'[i_{s+1}] \boldsymbol{A}[i_{s+1}, i_s] \,.$$

*Then, for any fixed $t$ as $n \to \infty$, the asymptotic state $(X_1^\infty, \dots, X_t^\infty)$, conditional on $(Z_\sigma^\infty)_{\sigma \in \mathcal{C}_1}$, is a centered Gaussian vector. A formula for its covariance is given in Proposition 6.26.*

The subtracted terms $\sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot \boldsymbol{f}_s$ generalize the "Onsager correction terms" appearing in different variants of AMP. Theorem 1.13 and its proof address two questions posed in [WZF22], namely (1) to obtain a combinatorial interpretation of the Onsager correction for OAMP algorithms, and (2) to identify a more general class of AMP algorithms whose state evolution is characterized by the diagonal distribution of the input matrix. Theorem 1.13 shows that (2) is possible for arbitrary

[8] The second comparison is with the caveat that GAMP uses a certain class of "non-separable" nonlinearities (applying a different function $f_t$ to each coordinate of $\boldsymbol{x}_t$) which are not directly covered by our result [Ran11, JM13].

matrices satisfying the strong cactus property, and explicitly describes such an algorithm and its conditionally Gaussian asymptotic states. We show in Section 6.3 how the treelike AMP iteration simultaneously generalizes several variants of AMP introduced in prior work.

We emphasize that, in contrast to all existing state evolution results we are aware of, we derive an Onsager correction and state evolution formula *without assuming an explicit random model for* $\boldsymbol{A}$. The iteration in Theorem 1.13 is the same regardless of the limiting diagonal distribution of $\boldsymbol{A}$, provided that these matrices (random or deterministic) satisfy the strong cactus property and have *some* limiting diagonal distribution (which will affect the covariance formula in Proposition 6.26). Note that the matrices in our universality result (Theorem 1.6) and their random counterparts (the r-ROM), satisfy the weak cactus property instead of the strong cactus one. Nevertheless, the Onsager correction and the state evolution can still be determined by a reduction to the strong-cactus-property setting, as we explain in Section 6.3.2.

### 1.3 Related work

**Moment method for AMP.** Our overall approach to graph polynomials generalizes prior work for the case of Wigner matrices [JP25]. Similar techniques have also appeared in prior works using the moment method to study AMP algorithms [BLM15, WZF22, MW25, DLS23, IS24, DSL24]. The $w$ and $z$ polynomials are rather fundamental objects which, along with their vector, matrix, and tensor generalizations, have variously been called "graph monomials" or "traffics" in free probability, "graph matrices" in computer science, "graph homomorphism polynomials" in combinatorics, and are also related to "tensor networks" and "Feynman diagrams" in physics.

**Polynomial vs. non-polynomial GFOM.** In random and semi-random models, general first-order methods with a constant number of iterations using (1) only polynomial nonlinearities or (2) arbitrary Lipschitz nonlinearities are generally expected to have the same computational power. Using polynomial approximation arguments, this has been made precise in several previous works [MW25, IS24, WZF22]. For example, [WZF22, Lemma 2.12] gives an abstract reduction showing that if state evolution for AMP on rotationally-invariant matrices holds for polynomial nonlinearities, then it also holds for arbitrary Lipschitz nonlinearities. While we study more general matrix models, we expect the assumption of polynomial nonlinearities is not essential.

**AMP vs. GFOM.** A simple reduction shows that every algorithm in the GFOM class can be expressed as a certain post-processing of an AMP algorithm (allowing "memory terms") [CMW20]. Therefore, these two classes of algorithms are equivalent from the standpoint of computational power. In our analysis, this is mirrored by the fact that, in Theorem 1.12, all possible non-Gaussian limits after conditioning on the draw of $(Z_\sigma^\infty)_{\sigma\in\mathcal{C}_1}$ are deterministic functions of the possible Gaussian limits.

**GFOM on independent entry matrices.** The analysis of GFOM and AMP on Wigner matrices or inhomogeneous versions thereof was the first case widely considered in the literature, and goes back to the origins of the mathematical analysis of AMP in the statistical physics literature on spin glasses [Bol14, DMM09, BM11, Mon12, BSK15, RV18, LW22]. See [FVRS22] for a survey of many of these works. Further, see [BLM15, CL21] for universality results over such models allowing for different entry distributions (but still requiring entrywise independence), [DJM13, JM13] for results

on block-structured variance profiles along the lines of our block GOE model, and [GHN26, BHX25] for recent progress on more general variance profiles.

**GFOM on orthogonally invariant matrices.** The correct form of AMP (to ensure Gaussian limiting distributions) in orthogonally invariant models was first predicted non-rigorously for physics applications by [OÇW16] using dynamical mean-field theory (DMFT), and then proved by [Fan22]. Precursors for special "divergence-free" forms of AMP were also obtained by [ÇO19, MP17, RSF19, Tak19] under the names of *Vector AMP* and *Orthogonal AMP*. Related calculations for a more general statistical physics framework subsuming these AMP variants are carried out in [MFC+19]; in particular, this work includes special cases of and discusses the more general form of the calculations we detail in Appendix B. See the discussion in [Fan22] for a more thorough overview of these distinctions.

**Universality principles for GFOM.** Beyond the above results, the main ones we are aware of that reduce the amount of randomness required for AMP are the recent works [WZF22, DLS23], which, modulo technical differences, both prove universality results over random matrices whose distribution is invariant under signed permutations. In other words, they treat broad classes of matrices provided that these are conjugated by random signed permutations, a considerable reduction in randomness from, e.g., conjugating by random Haar-distributed orthogonal or unitary matrices as in OAMP. Numerous experimental works have found universality phenomena for "sufficiently pseudorandom" deterministic matrices, but we are not aware of any rigorous results for completely deterministic matrices prior to our work. See discussion in [ÇO19, Sch20, ABKZ20, DLS23].

### 1.4 Organization of the paper

We give preliminaries on the matrices considered in this work and modes of convergence for our limiting theorem in Section 2. We introduce our definitions of diagrams and consequences of Möbius inversions for the traffic distribution in Section 3. In Section 4, to build intuition on traffic distributions, we describe them for several random matrix ensembles. Section 5 is dedicated to the proof of our first main result, the polynomial universality of delocalized deterministic matrices (Theorem 1.6). Section 6 details and proves the effective dynamics of GFOM under the strong cactus property (Theorems 1.12 and 1.13).

We illustrate two viable approaches to computing the traffic distribution of orthogonally invariant matrix models: Appendix A is based on Feynman diagrams and Appendix B relies on Weingarten calculus. Appendix C provides background on convergence of stochastic processes, and Appendix D contains omitted proofs.

### 1.5 Acknowledgments

Thanks to Zhou Fan, Cynthia Rush, and Subhabrata Sen for helpful discussions over the course of this project. CJ was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101019547). LP's work was supported by the Swiss National Science Foundation (SNSF), grant no. 10004947.

# 2 Preliminaries

## 2.1 Matrix notation

Given matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, we will use:

- $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ to specify that $\boldsymbol{A}$ is symmetric.
- $\boldsymbol{A} \in O(n) \subseteq \mathbb{R}^{n\times n}$ to specify that $\boldsymbol{A}$ is orthogonal.
- $\boldsymbol{A}[i,j]$ to denote its $(i,j)$-th entry for $i,j \in [n] := \{1,\dots,n\}$.
- $\|\boldsymbol{A}\| := \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$ to denote its spectral or operator norm.
- $\|\boldsymbol{A}\|_{\mathrm{F}}^2 := \sum_{i,j=1}^n \boldsymbol{A}[i,j]^2$ to denote its Frobenius norm.
- $\mathrm{Tr}(\boldsymbol{A}) := \sum_{i=1}^n \boldsymbol{A}[i,i]$ to denote its trace.
- $\lambda_1(\boldsymbol{A}) \geq \ldots \geq \lambda_n(\boldsymbol{A})$ to denote its eigenvalues when $\boldsymbol{A}$ is symmetric.
- $\boldsymbol{A} \odot \boldsymbol{B}$ to denote the entrywise or Hadamard product with entries $(\boldsymbol{A}[i,j]\boldsymbol{B}[i,j])_{i,j\in[n]}$.

**Definition 2.1** (Puncturing)**.** *Let $\boldsymbol{H} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and $\boldsymbol{\Pi} := \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$ be the projection orthogonal to the all-ones direction. The puncturing of $\boldsymbol{H}$ is the matrix $\boldsymbol{A} = \boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi}$.*

**Definition 2.2** (GOE)**.** *The (normalized) Gaussian Orthogonal Ensemble* GOE *is the distribution of random matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ with $\boldsymbol{A}[i,j] = \boldsymbol{A}[j,i] \sim \mathcal{N}(0,1/n)$ independently for all $1 \leq i < j \leq n$, and $\boldsymbol{A}[i,i] \sim \mathcal{N}(0,2/n)$ independently for all $i \in [n]$.*

**Definition 2.3** (Hadamard matrices)**.** *When $n$ is a power of 2, the (normalized) Walsh–Hadamard matrix $\boldsymbol{H}^{(n)}_{\mathrm{had}} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is defined recursively by*

$$\boldsymbol{H}^{(1)}_{\mathrm{had}} = \begin{bmatrix}1\end{bmatrix}, \qquad \boldsymbol{H}^{(2n)}_{\mathrm{had}} := \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{H}^{(n)}_{\mathrm{had}} & \boldsymbol{H}^{(n)}_{\mathrm{had}} \\ \boldsymbol{H}^{(n)}_{\mathrm{had}} & -\boldsymbol{H}^{(n)}_{\mathrm{had}}\end{bmatrix}.$$

$\boldsymbol{H}^{(n)}_{\mathrm{had}}$ is a symmetric orthogonal matrix with entries in $\pm 1/\sqrt{n}$.

**Definition 2.4** (DST and DCT matrices)**.** *The discrete sine transform matrices $\boldsymbol{H}^{(n)}_{\mathrm{sin}} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ are*

$$\boldsymbol{H}^{(n)}_{\mathrm{sin}}[i,j] := \sqrt{\frac{2}{n+1}}\sin\left(\frac{\pi i j}{n+1}\right) \quad \forall i,j \in [n].$$

*The discrete cosine transform matrices $\boldsymbol{H}^{(n)}_{\mathrm{cos}} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ are*

$$\boldsymbol{H}^{(n)}_{\mathrm{cos}}[i,j] := \sqrt{\frac{2}{n}}\cos\left(\frac{\pi(i-\frac{1}{2})(j-\frac{1}{2})}{n}\right) \quad \forall i,j \in [n].$$

$\boldsymbol{H}^{(n)}_{\mathrm{cos}}$ and $\boldsymbol{H}^{(n)}_{\mathrm{sin}}$ are symmetric orthogonal matrices with entries at most $O(1/\sqrt{n})$ in magnitude.

**Definition 2.5** (ROM and r-ROM)**.** *The Random Orthogonal Model* ROM *is the distribution of random matrices $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} \in O(n)$ is Haar-distributed, and $\boldsymbol{D}$ is a diagonal matrix*

*with i.i.d.* $\mathrm{Unif}(\{-1,1\})$ *entries, independent from* $\boldsymbol{Q}$*. The Regular Random Orthogonal Model* r-ROM *is the distribution of the puncturing of* $\boldsymbol{H}$*, when* $\boldsymbol{H}$ *is sampled from the* ROM*.*

Random matrices from the ROM are symmetric orthogonal matrices, satisfying $\boldsymbol{H}^2 = \boldsymbol{I}$. They are a special case of the orthogonally invariant models we discuss in Section 4.2.

## 2.2 Modes of convergence

We will use a few standard modes of convergence from scalar-valued probability theory.

**Definition 2.6** (Modes of convergence: scalars)**.** *For a sequence of random variables* $x^{(n)} \in \mathbb{R}$*, we say that:*

- $x^{(n)}$ *converge in expectation if, for some* $c \in \mathbb{R}$*,* $\lim_{n\to\infty} \mathbb{E}x^{(n)} = c$*.*
- $x^{(n)}$ *converge in probability if, for some* $c \in \mathbb{R}$*, for all* $\varepsilon > 0$*,* $\lim_{n\to\infty} \mathbb{P}[|x^{(n)} - c| > \varepsilon] = 0$*.*
- $x^{(n)}$ *converge in* $L^2$ *if they converge in expectation and* $\lim_{n\to\infty} \mathbb{E}(x^{(n)} - c)^2 = 0$*, or equivalently if they converge in expectation and* $\lim_{n\to\infty} \operatorname{Var} x^{(n)} = 0$*.*

*We write a symbol* $\mathcal{M} \in \{\mathbb{E}, \mathbb{P}, L^2\}$ *to indicate these modes of convergence, and in this notation say that the* $x^{(n)}$ *converge in* $\mathcal{M}$*.*

Moreover, we say a sequence of random vectors $\boldsymbol{x}^{(n)} \in \mathbb{R}^d$ in fixed dimension $d \geq 1$ *converges in distribution* to a random vector $\boldsymbol{x} \in \mathbb{R}^d$ if for every bounded continuous function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\,\varphi(\boldsymbol{x}^{(n)}) \underset{n\to\infty}{\longrightarrow} \mathbb{E}\,\varphi(\boldsymbol{x})\,,$$

in which case we write $\boldsymbol{x}^{(n)} \xrightarrow{(\mathrm{d})} \boldsymbol{x}$. See Appendix C for a generalization to random variables indexed by a countably infinite index set.

**Definition 2.7** (Modes of convergence: tracial moments)**.** *For a mode of convergence* $\mathcal{M}$*, we say that a sequence of random matrices* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *converges in tracial moments in* $\mathcal{M}$ *if, for every* $k \geq 1$*,* $\frac{1}{n}\operatorname{Tr}\boldsymbol{A}^k$ *converges in* $\mathcal{M}$*. We say that it converges in tracial moments in* $\mathcal{M}$ *to a probability measure* $\mu$ *over* $\mathbb{R}$ *if*

$$\frac{1}{n}\operatorname{Tr}\boldsymbol{A}^k \to \int x^k \,\mathrm{d}\mu(x)$$

*in the mode of convergence* $\mathcal{M}$*.*

## 2.3 Matchings and Wick calculus

Given a set $S$, let $\mathcal{M}(S)$ denote the set of matchings on $S$. Let $\mathcal{M}_{\mathrm{perf}}(S)$ denote the subset of perfect matchings. The elements of $M \in \mathcal{M}(S)$ are written as pairs $\{i, j\} \subseteq S$. For several sets $S_1, \ldots, S_k$, denote by $\mathcal{M}(S_1, \ldots, S_k)$ the set of matchings on the disjoint union $S_1 \sqcup \cdots \sqcup S_k$ that do not match any two elements of the same $S_i$. For two sets $S_1, S_2$ of the same size, denote by $\mathcal{M}_{\mathrm{perf}}(S_1, S_2)$ the bipartite perfect matchings of $S_1 \sqcup S_2$ that only match elements of $S_1$ to ones of $S_2$. We will abbreviate $\mathcal{M}(\{1, 2, \ldots, q\})$ as $\mathcal{M}(q)$.

**Lemma 2.8** (Wick lemma)**.** *Let $X_1, \dots, X_q$ be jointly Gaussian random variables with mean zero. Then:*

$$\mathbb{E}[X_1 \cdots X_q] = \sum_{M \in \mathcal{M}_{\mathrm{perf}}(q)} \prod_{ij \in M} \mathbb{E}[X_i X_j] \,.$$

The Wick products are the multivariate generalization of the Hermite polynomials to correlated Gaussians [Jan97, Chapter 3].

**Definition 2.9** (Wick product)**.** *Let $I$ be an index set, $\boldsymbol{X} = (X_i)_{i \in I}$ be formal variables, and $\boldsymbol{\Sigma} \in \mathbb{R}^{I \times I}_{\mathrm{sym}}$. The* Wick products *are defined by, for each finitely supported $\alpha \in \mathbb{N}^I$,*

$$\mathrm{He}_\alpha(\boldsymbol{X} \,;\, \boldsymbol{\Sigma}) := \sum_{M \in \mathcal{M}(\alpha)} (-1)^{|M|} \prod_{uv \in M} \boldsymbol{\Sigma}[u, v] \prod_{u \notin M} X_u \,,$$

*where $\mathcal{M}(\alpha)$ denotes the set of matchings on a collection consisting of $\alpha_i$ copies of each $i \in I$.*

When $|I| = 1$, $X \sim \mathcal{N}(0, 1)$, and $\Sigma = 1$, then $\mathrm{He}_{(p)}(X \,;\, \Sigma)$ equals the $p$th Hermite polynomial.

When the $X_i$ are mean-zero Gaussian random variables and $\boldsymbol{\Sigma}$ is their covariance matrix, the Wick products satisfy the (partial) orthogonality property that for each finitely supported $\alpha, \beta \in \mathbb{N}^I$ with $\sum_i \alpha_i \neq \sum_i \beta_i$,

$$\mathbb{E}\left[\mathrm{He}_\alpha(\boldsymbol{X} \,;\, \boldsymbol{\Sigma}) \, \mathrm{He}_\beta(\boldsymbol{X} \,;\, \boldsymbol{\Sigma})\right] = 0 \,.$$

In general, we have

$$\mathbb{E}\left[\mathrm{He}_\alpha(\boldsymbol{X} \,;\, \boldsymbol{\Sigma}) \, \mathrm{He}_\beta(\boldsymbol{X} \,;\, \boldsymbol{\Sigma})\right] = \sum_{M \in \mathcal{M}_{\mathrm{perf}}(\alpha, \beta)} \prod_{uv \in M} \boldsymbol{\Sigma}[u, v] \,.$$

Since by the Wick lemma $\mathbb{E}\left[\prod_{i \in \alpha} X_i \cdot \prod_{j \in \beta} X_j\right]$ equals the same sum over all matchings of $\alpha \sqcup \beta$, the Wick products achieve a general "partial orthogonalization" that removes all terms from this covariance where any pairs within $\alpha$ or within $\beta$ are matched.

For each choice of $\boldsymbol{\Sigma} \in \mathbb{R}^{I \times I}_{\mathrm{sym}}$, the Wick products are a basis for polynomials in the $X_i$. Multiplication of polynomials gives an algebra structure to this space which we call the *Wick algebra* of $\boldsymbol{X}$. Below is a combinatorial formula for multiplication in the Wick algebra.

**Proposition 2.10** ([Jan97, Theorem 3.15])**.** *Let $I$ be an index set, $\boldsymbol{X} = (X_i)_{i \in I}$ be formal variables, and $\boldsymbol{\Sigma} \in \mathbb{R}^{I \times I}_{\mathrm{sym}}$. Let $\alpha^1, \dots, \alpha^k \in \mathbb{N}^I$. Then:*

$$\prod_{j=1}^{k} \mathrm{He}_{\alpha^j}(\boldsymbol{X}; \boldsymbol{\Sigma}) = \sum_{M \in \mathcal{M}(\alpha^1, \dots, \alpha^k)} \prod_{uv \in M} \boldsymbol{\Sigma}[u, v] \, \mathrm{He}_{U(M)}(\boldsymbol{X} \,;\, \boldsymbol{\Sigma}) \,,$$

*where $\alpha^j$ is a multiset of size $|\alpha^j|$ with $\alpha^j_i$ copies of each $i \in I$. Here $U(M) \in \mathbb{N}^I$ for $M$ a matching of $\alpha^1 \sqcup \cdots \sqcup \alpha^k$ counts the number of unmatched elements of each type.*

In the special case where each group $\alpha^j$ consists of a single element, we obtain:

**Corollary 2.11.** *For every $i_1, \dots, i_k \in I$,*

$$\prod_{j=1}^{k} X_{i_j} = \sum_{M \in \mathcal{M}(k)} \prod_{uv \in M} \boldsymbol{\Sigma}[i_u, i_v] \, \mathrm{He}_{U(M)}(\boldsymbol{X} \,;\, \boldsymbol{\Sigma}) \,.$$

# 3 Diagrams and the $w$- and $z$-Bases of Polynomials

All graphs considered in this paper are *multigraphs* (loops and multiedges are allowed) and will be denoted by Greek letters $(\alpha, \beta, \gamma, \ldots)$. We use the terms *graphs* and *diagrams* interchangeably in this paper. Given a diagram $\alpha$, we use $V(\alpha)$ to denote its vertex set and $E(\alpha)$ to denote its edge set. We denote by $\alpha[S]$ the subgraph of $\alpha$ induced by $S \subseteq V(\alpha)$. We count self-loops as contributing 2 to the degree of a vertex.

## 3.1 Classes of diagrams

Each diagram can have either 0, 1, or an ordered pair of 2 special vertices called its *root(s)*. With the exception of the class of graphs defined in Definition 5.4, the roots of a graph can be arbitrary vertices (in particular, they might be equal if there are two of them).

**Notation 3.1.** *Let $\mathcal{A} = \mathcal{A}_0$ (resp. $\mathcal{A}_1$ or $\mathcal{A}_2$) be the set of all connected graphs with no root (resp. 1 root or 2 roots). We also refer to such graphs as scalar (resp. vector or matrix) diagrams.*

Given $\alpha \in \mathcal{A}$, an edge $e \in E(\alpha)$ is a *bridge* of $\alpha$ if deleting $e$ would disconnect the graph. $\alpha \in \mathcal{A}$ is *2-edge-connected* if it contains no bridge. In general, $\alpha \in \mathcal{A}$ can be decomposed into a tree of 2-edge-connected components connected by bridges.

**Notation 3.2.** *Let $\mathcal{E} = \mathcal{E}_0 \subseteq \mathcal{A}$ (resp. $\mathcal{E}_1 \subseteq \mathcal{A}_1$ or $\mathcal{E}_2 \subseteq \mathcal{A}_2$) be the set of all 2-edge-connected scalar (resp. vector or matrix) diagrams.*

Given $\alpha \in \mathcal{A}$, a vertex $u \in V(\alpha)$ is an *articulation point* of $\alpha$ if removing $u$ and its incident edges disconnects the graph. $\alpha$ is *2-vertex-connected* if it has no articulation point. Any $\alpha \in \mathcal{A}$ decomposes into its 2-vertex-connected components (blocks), which refine the 2-edge-connected components. The block-cut graph (whose vertices are the articulation points and the blocks, with edges for incidence) is a tree.

A connected graph is a *cactus* if every edge lies on exactly one simple cycle. Thus, cactuses are in a sense the minimal 2-edge-connected graphs.

**Notation 3.3.** *Let $\mathcal{C} = \mathcal{C}_0 \subseteq \mathcal{A}$ (resp. $\mathcal{C}_1 \subseteq \mathcal{A}_1$) be the set of all scalar (resp. vector) cactus diagrams.*

For a cactus $\sigma$, we will denote by $\mathrm{cyc}(\sigma)$ the set of (unrooted) cycles of $\sigma$.

Finally, as in Definition 1.11, we will denote the treelike diagrams by $\mathcal{T}_1$ and the treelike diagrams such that the root has degree 1 after deleting all hanging cactuses by $\mathcal{G}_1$.

## 3.2 Graph polynomials

Each diagram represents different scalar-, vector-, or matrix-valued polynomials in a matrix input, depending on whether it is viewed in the $w$-basis or the $z$-basis. In the following definitions, we fix $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, $\alpha$ to be a scalar, vector, or matrix diagram, and $i, j \in [n]$.

**Definition 3.4.** *Define* $w_\alpha(\boldsymbol{A}) \in \mathbb{R}$, $\boldsymbol{w}_\alpha(\boldsymbol{A}) \in \mathbb{R}^n$, *and* $\boldsymbol{W}_\alpha(\boldsymbol{A}) \in \mathbb{R}^{n\times n}$ *by*

$$w_\alpha(\boldsymbol{A}) = \sum_{\varphi\colon V(\alpha)\to[n]} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a scalar diagram,}$$

$$\boldsymbol{w}_\alpha(\boldsymbol{A})[i] = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(r)=i}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a vector diagram with root } r,$$

$$\boldsymbol{W}_\alpha(\boldsymbol{A})[i,j] = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(r_1)=i,\varphi(r_2)=j}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a matrix diagram with roots } (r_1,r_2).$$

**Definition 3.5.** *Define* $z_\alpha(\boldsymbol{A}) \in \mathbb{R}$, $\boldsymbol{z}_\alpha(\boldsymbol{A}) \in \mathbb{R}^n$, *and* $\boldsymbol{Z}_\alpha(\boldsymbol{A}) \in \mathbb{R}^{n\times n}$ *by*

$$z_\alpha(\boldsymbol{A}) = \sum_{\varphi\colon V(\alpha)\hookrightarrow[n]} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a scalar diagram,}$$

$$\boldsymbol{z}_\alpha(\boldsymbol{A})[i] = \sum_{\substack{\varphi\colon V(\alpha)\hookrightarrow[n]\\ \varphi(r)=i}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a vector diagram with root } r,$$

$$\boldsymbol{Z}_\alpha(\boldsymbol{A})[i,j] = \sum_{\substack{\varphi\colon V(\alpha)\hookrightarrow[n]\\ \varphi(r_1)=i,\varphi(r_2)=j}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \qquad \text{if } \alpha \text{ is a matrix diagram with roots } (r_1,r_2).$$

The only difference between the $w$- and $z$-bases is the summation domain: Definition 3.5 sums over injective embeddings $\varphi$, whereas Definition 3.4 sums over all embeddings.

Finally, we define two extensions of Definition 3.4 that we will need in the proofs. The following allows us to use a different matrix on each edge of the graph:

**Definition 3.6.** *Let* $\alpha$ *be a matrix diagram with roots* $(r_1,r_2)$ *and* $\boldsymbol{\mathcal{A}} = (\boldsymbol{A}_e)_{e\in E(\alpha)}$ *be such that* $\boldsymbol{A}_e \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *for all* $e \in E(\alpha)$. *Define* $\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}}) \in \mathbb{R}^{n\times n}$ *by*

$$\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}})[i,j] = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(r_1)=i,\varphi(r_2)=j}} \prod_{e=\{u,v\}\in E(\alpha)} \boldsymbol{A}_e[\varphi(u),\varphi(v)]\,.$$

The following is an intermediate quantity between Definition 3.4 and Definition 3.5 which only restricts the sum over injective labelings on two vertices:

**Definition 3.7.** *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, $\alpha$ *be a scalar/vector/matrix diagram,* $i,j \in [n]$, *and* $s,t \in V(\alpha)$.

*Define* $w_\alpha^{s\neq t} \in \mathbb{R}$, $\boldsymbol{w}_\alpha^{s\neq t} \in \mathbb{R}^n$, *and* $\boldsymbol{W}_\alpha^{s\neq t} \in \mathbb{R}^{n\times n}$ *by*

$$w_\alpha^{s\neq t}(\boldsymbol{A}) = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(s)\neq\varphi(t)}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \quad \text{if } \alpha \text{ is a scalar diagram,}$$

$$\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{A})[i] = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(s)\neq\varphi(t)\\ \varphi(r)=i}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \quad \text{if } \alpha \text{ is a vector diagram with root } r,$$

$$\boldsymbol{W}_\alpha^{s\neq t}(\boldsymbol{A})[i,j] = \sum_{\substack{\varphi\colon V(\alpha)\to[n]\\ \varphi(s)\neq\varphi(t)\\ \varphi(r_1)=i\\ \varphi(r_2)=j}} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{A}[\varphi(u),\varphi(v)] \quad \text{if } \alpha \text{ is a matrix diagram with roots } (r_1,r_2).$$

### 3.3 Partitions, change of basis, and Möbius inversion

While $(z_\alpha(\boldsymbol{A}))_{\alpha\in\mathcal{A}}$ and $(w_\alpha(\boldsymbol{A}))_{\alpha\in\mathcal{A}}$ span the same space of $S_n$-invariant polynomials in the entries of $\boldsymbol{A}$, some properties are better expressed in one basis than the other. Here we take a closer look at these bases and derive change-of-basis formulas.

Given a set $S$, let $\mathcal{P}(S)$ denote the set of all *partitions* of $S$, sets of non-empty disjoint subsets of $S$ whose union is all of $S$. We call the parts of a partition *blocks*. Each block is a set, and $P$ is the set of blocks, so we denote the blocks by $b \in P$.

For a (scalar, vector, or matrix) diagram $\alpha$ and a partition $P \in \mathcal{P}(V(\alpha))$, we define a new diagram $\alpha_P$ by identifying the vertices within each block of $P$ into a single vertex. The vertices of $\alpha_P$ may thus be identified with the blocks of $P$. $\alpha_P$ retains all edges of $\alpha$, which may become multiedges or self-loops. The status of being one of the (0, 1, or 2) roots of $\alpha$ is inherited by the block containing that root.

To change from the $w$- to the $z$-basis, we then simply sum over all partitions:

**Claim 3.8.** *For all (scalar, vector, or matrix) diagrams* $\alpha$*,*

$$w_\alpha(\boldsymbol{A}) = \sum_{P\in\mathcal{P}(V(\alpha))} z_{\alpha_P}(\boldsymbol{A})\,.$$

Define the relation $\alpha \preceq \beta$ on scalar diagrams if there exists a partition $P \in \mathcal{P}(V(\beta))$ such that $\alpha = \beta_P$. It is easy to check that this relation gives a partial ordering, inherited from the standard partial ordering on partitions. We write $\alpha \prec \beta$ as a shorthand for $\alpha \preceq \beta$ and $\alpha \neq \beta$.

**Lemma 3.9.** *There exist* $(c_{\alpha,\beta})_{\alpha,\beta\in\mathcal{A}}$ *and* $(c'_{\alpha,\beta})_{\alpha,\beta\in\mathcal{A}}$ *not depending on* $n$ *such that* $c_{\alpha,\beta} \in \mathbb{N}$*,* $c'_{\alpha,\beta} \in \mathbb{Z}$ *and for any* $\alpha,\beta \in \mathcal{A}$*,*

$$w_\beta(\boldsymbol{A}) = \sum_{\alpha\preceq\beta} c_{\alpha,\beta} z_\alpha(\boldsymbol{A})\,, \qquad z_\beta(\boldsymbol{A}) = \sum_{\alpha\preceq\beta} c'_{\alpha,\beta} w_\alpha(\boldsymbol{A})\,.$$

*Proof.* The coefficients in the left equation count symmetries in Claim 3.8, i.e., $c_{\alpha,\beta}$ equals the number of ways to choose a partition $P \in \mathcal{P}(V(\beta))$ such that $\beta_P$ is isomorphic to $\alpha$. Reciprocally, since $\preceq$ is a partial ordering, this transformation can be inverted using Möbius inversion [Rot64] on

this poset. Although an explicit formula for $c'_{\alpha,\beta}$ is available in terms of the combinatorial structure of the graphs, we will not need it in this paper. □

### 3.4 The example of cycles: Moments versus free cumulants

The difference between the $w$- and $z$-bases is illustrated nicely by the special case of the diagrams $\sigma_q$ which are cycles of length $q \geq 1$. In this case, $\frac{1}{n} w_{\sigma_q}(\boldsymbol{A})$ and $\frac{1}{n} z_{\sigma_q}(\boldsymbol{A})$ are versions of the limiting spectral moments and free cumulants, respectively, for finite-dimensional matrices.

Let $\mathcal{P}(q)$ denote the set of partitions of $\{1, 2, \ldots, q\}$ and let $\mathrm{NC}(q)$ denote the subset of non-crossing partitions (partitions such that there does not exist $i < j < k < \ell$ with $i, k$ in the same block and $j, \ell$ in the same block, different from the one $i, k$ are in). It is convenient to view these as partitions of the vertices of the $q$-cycle so that the term *non-crossing* may be interpreted visually: in a non-crossing partition, the blocks do not intersect one another when drawn as "blobs" inside the cycle.

In the $w$-basis, we have

$$\frac{1}{n} w_{\sigma_q}(\boldsymbol{A}) = \frac{1}{n} \sum_{i_1,\ldots,i_q=1}^{n} \boldsymbol{A}[i_1, i_2]\boldsymbol{A}[i_2, i_3] \ldots \boldsymbol{A}[i_q, i_1] = \frac{1}{n} \operatorname{Tr}(\boldsymbol{A}^q) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i(\boldsymbol{A})^q \,. \tag{6}$$

Suppose that the expression in Eq. (6) converges as $n \to \infty$ to the $q$th moment $m_q \in \mathbb{R}$ of a limiting spectral distribution, $m_q = \int \lambda^q \, \mathrm{d}\mu(\lambda)$.

The free cumulants are defined from the moments by a formula similar to the classical cumulants vis-à-vis the moments of a random variable:

**Definition 3.10** (Free cumulant)**.** *The free cumulants $(\kappa_q)_{q\geq 1}$ corresponding to $(m_q)_{q\geq 1}$ are defined implicitly by:*

$$m_q = \sum_{\sigma \in \mathrm{NC}(q)} \prod_{b \in \sigma} \kappa_{|b|} \,. \tag{7}$$

The $\kappa_q$ can be computed explicitly in terms of the $m_q$ by applying Möbius inversion to Eq. (7); see Eq. (59).

Analogous to Eq. (6) which is in the $w$-basis, it appears to be folklore[9] that if $\boldsymbol{A}$ is drawn from an orthogonally invariant matrix ensemble with free cumulants $(\kappa_q)_{q\geq 1}$, then

$$\frac{1}{n} \mathbb{E}\, z_{\sigma_q}(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} \kappa_q \,. \tag{8}$$

The quantity $\frac{1}{n} z_{\sigma_q}(\boldsymbol{A})$ has also been called the $q$th *injective trace* of $\boldsymbol{A}$. Below in Lemma 3.12, we prove Eq. (8) using a change of basis from $w$ to $z$.

For example, below are the parameters $m_q$ and $\kappa_q$ for the GOE and the ROM, whose limiting empirical spectral distribution are the Wigner semicircle distribution and the Rademacher distribution, respectively.

**Claim 3.11.** *Let* $\mathrm{Cat}(k) := \frac{1}{k+1}\binom{2k}{k}$ *be the $k$th Catalan number. For the* GOE*, the limiting spectral*

[9] This is for example explicitly stated in [MFC+19, Theorem 1 and Appendix D.1].

*moments and free cumulants are:*

$$m_q = \left\{ \begin{array}{ll} \mathrm{Cat}(q/2) & \text{if } q \text{ is even} \\ 0 & \text{if } q \text{ is odd} \end{array} \right\}, \qquad \kappa_q = \left\{ \begin{array}{ll} 1 & \text{if } q = 2 \\ 0 & \text{otherwise} \end{array} \right\}.$$

*For the* ROM*, the limiting spectral moments and free cumulants are:*

$$m_q = \left\{ \begin{array}{ll} 1 & \text{if } q \text{ is even} \\ 0 & \text{if } q \text{ is odd} \end{array} \right\}, \qquad \kappa_q = \left\{ \begin{array}{ll} (-1)^{q/2-1}\mathrm{Cat}(q/2-1) & \text{if } q \text{ is even} \\ 0 & \text{if } q \text{ is odd} \end{array} \right\}. \tag{9}$$

### 3.5 Solving equations in the traffic distribution

The traffic distribution is defined as the limiting values of all $w$-basis polynomials, but we show now how it can be derived from various combinations of limits of $w$- and $z$-basis polynomials. In our other arguments, we will also find it convenient to describe the traffic distribution of sequences of matrices (random or deterministic) using the two bases simultaneously. While Lemma 3.9 shows that we could in principle express all these results in a single basis, this would involve precisely tracking very complicated combinatorial coefficients (in fact, this was a major technical obstacle in previous diagrammatic analyses of AMP).

As we have discussed, when a matrix satisfies the strong cactus property, its traffic distribution is determined by its values on the cactus diagrams (equivalently, by the diagonal distribution), and when it satisfies the factorizing strong cactus property, its traffic distribution is determined by the spectral distribution. We show that one can use either the $w$-basis or $z$-basis for these determinations.

**Lemma 3.12.** *Suppose that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies the weak cactus property, i.e., for all $\alpha \in \mathcal{E} \setminus \mathcal{C}$,*

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0\,.$$

*Then the following are equivalent:*

*(i) For all $\sigma \in \mathcal{C}$ there exists $m_\sigma \in \mathbb{R}$ such that $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, w_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} m_\sigma$.*

*(ii) For all $\sigma \in \mathcal{C}$ there exists $k_\sigma \in \mathbb{R}$ such that $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} k_\sigma$.*

*Furthermore, when they exist, $(m_\sigma)_{\sigma\in\mathcal{C}}$ and $(k_\sigma)_{\sigma\in\mathcal{C}}$ determine each other. The following are also equivalent:*

*(i) There exist real numbers $(m_q)_{q\in\mathbb{N}}$ such that for all $\sigma \in \mathcal{C}$, $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, w_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} \prod_{\rho\in\mathrm{cyc}(\sigma)} m_{|\rho|}$.*

*(ii) There exist real numbers $(\kappa_q)_{q\in\mathbb{N}}$ such that for all $\sigma \in \mathcal{C}$, $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} \prod_{\rho\in\mathrm{cyc}(\sigma)} \kappa_{|\rho|}$.*

*Furthermore, when they exist, $(m_q)_{q\in\mathbb{N}}$ and $(\kappa_q)_{q\in\mathbb{N}}$ are related by Eq. (7).*

We use the following observation which will be used repeatedly in Section 5:

**Lemma 3.13.** *If $\alpha \in \mathcal{E}$ and $\beta \preceq \alpha$, then $\beta \in \mathcal{E}$.*

*Proof of Lemma 3.13.* By Menger's theorem, a graph is 2-edge-connected if and only if there exist two edge-disjoint paths between every pair of distinct vertices. These paths are maintained when $\alpha$ is contracted into $\beta$. □

*Proof of Lemma 3.12. (ii) $\implies$ (i).* Using Claim 3.8,

$$\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\sigma(\boldsymbol{A}) = \frac{1}{n}\sum_{\beta \preceq \sigma} c_{\beta,\sigma} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_\beta(\boldsymbol{A})\,.$$

Every diagram $\beta \preceq \sigma$ remains 2-edge-connected by Lemma 3.13. There are only finitely many terms in the sum, so we can directly take the $n \to \infty$ limit and use the assumptions to obtain that $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, w_\sigma(\boldsymbol{A})$ converges to $\sum_{\beta\preceq\sigma} c_{\beta,\sigma}k_\beta$.

Note that by the weak cactus property, the only asymptotically nonzero $\beta \preceq \sigma$ are when $\beta$ is a cactus. Assuming furthermore that $k_\beta = \prod_{\rho\in\mathrm{cyc}(\beta)} \kappa_{|\rho|}$ factors over the cycles of each cactus $\beta$ we will derive the second part of the lemma.

Using the more specific result of Claim 3.8, we have

$$\lim_{n\to\infty}\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\sigma(\boldsymbol{A}) = \sum_{P\in\mathcal{P}(V(\sigma))} \lim_{n\to\infty}\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_{\sigma_P}(\boldsymbol{A})$$

Since $\boldsymbol{A}$ has the weak cactus property and $\sigma$ is a cactus, only the terms where $\sigma_P$ is a cactus contribute. These are precisely the terms where $P$ restricted to each cycle of $\sigma$ is non-crossing. Given $P_\rho \in \mathrm{NC}(V(\rho))$ for each $\rho \in \mathrm{cyc}(\sigma)$, let us write $P(P_\rho : \rho \in \mathrm{cyc}(\sigma))$ for the partition obtained by composing these partitions of each cycle, and let us write, following our previous notation, $\mathrm{cyc}(\rho_{P_\rho})$ for the set of cycles created when the single cycle $\rho$ is contracted according to $P_\rho$. Then, we have

$$\begin{aligned}
&= \sum_{\substack{P_\rho\in\mathrm{NC}(V(\rho))\\ \text{for each } \rho\in\mathrm{cyc}(\sigma)}} \lim_{n\to\infty}\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_{\sigma_{P(P_\rho:\rho\in\mathrm{cyc}(\sigma))}}(\boldsymbol{A})\\
&= \sum_{\substack{P_\rho\in\mathrm{NC}(V(\rho))\\ \text{for each } \rho\in\mathrm{cyc}(\sigma)}} \prod_{\rho\in\mathrm{cyc}(\sigma)} \prod_{\pi\in\mathrm{cyc}(\rho_{P_\rho})} \kappa_{|\pi|}\\
&= \prod_{\rho\in\mathrm{cyc}(\sigma)} \left(\sum_{P\in\mathrm{NC}(V(\rho))} \prod_{\pi\in\mathrm{cyc}(\rho_P)} \kappa_{|\pi|}\right)\\
&= \prod_{\rho\in\mathrm{cyc}(\sigma)} m_{|\rho|}.
\end{aligned}$$

Thus we have the claimed factorization. Further, the coefficients $m_q$ and $\kappa_q$ indeed have the relation between moments and free cumulants from Eq. (7):

$$m_q = \sum_{\sigma\in\mathrm{NC}(q)} \prod_{b\in\sigma} \kappa_{|b|}\,.$$

*(i) $\implies$ (ii).* This direction uses a recursive change of basis technique that will be very useful

in Section 5. Using Lemma 3.9 in both directions, we get

$$\begin{aligned}\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\sigma(\boldsymbol{A}) &= \frac{1}{n}\sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{C}}} c'_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\beta(\boldsymbol{A}) + \frac{1}{n}\sum_{\substack{\beta\prec\sigma\\ \beta\in\mathcal{E}\setminus\mathcal{C}}} c'_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\beta(\boldsymbol{A})\\ &= \frac{1}{n}\sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{C}}} c'_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\beta(\boldsymbol{A}) + \frac{1}{n}\sum_{\substack{\beta\prec\sigma\\ \beta\in\mathcal{E}\setminus\mathcal{C}}} c'_{\beta,\sigma}\sum_{\alpha\preceq\beta} c_{\alpha,\beta}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A})\\ &= \frac{1}{n}\sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{C}}} c'_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\beta(\boldsymbol{A}) + \frac{1}{n}\sum_{\alpha\prec\sigma}\left(\sum_{\substack{\beta\in\mathcal{E}\setminus\mathcal{C}\\ \alpha\preceq\beta\prec\sigma}} c'_{\beta,\sigma}c_{\alpha,\beta}\right)\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A})\end{aligned}$$

Note that every diagram in this expansion remains 2-edge-connected by Lemma 3.13.

Every contraction identifying a non-empty subset of vertices decreases the number of vertices in the graph, and the $w$ and $z$ bases coincide for 1-vertex graphs. Therefore, we can apply the same steps inductively on terms for which $\alpha \in \mathcal{C}$ to finally obtain

$$\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\sigma(\boldsymbol{A}) = \frac{1}{n}\sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{C}}} c''_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} w_\beta(\boldsymbol{A}) + \frac{1}{n}\sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{E}\setminus\mathcal{C}}} c''_{\beta,\sigma}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\beta(\boldsymbol{A})\,.$$

for some coefficients $\{c''_{\alpha,\beta}\}$ independent of $n$. Take the $n\to\infty$ limit to obtain

$$\lim_{n\to\infty}\frac{1}{n}\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\sigma(\boldsymbol{A}) = \sum_{\substack{\beta\preceq\sigma\\ \beta\in\mathcal{C}}} c''_{\beta,\sigma} m_\beta\,,$$

which finishes the proof of the first equivalence. Assuming furthermore that $m_\beta$ factors over the cycles of each cactus $\beta$, then $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\sigma(\boldsymbol{A})$ also asymptotically factors over its cycles: $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\sigma(\boldsymbol{A}) \longrightarrow \prod_{\rho\in\mathrm{cyc}(\sigma)}\kappa_{|\rho|}$ for some numbers $\kappa_q$. This is because the cactuses $\beta\preceq\sigma$ still only arise by contracting a separate non-crossing partition for each cycle of $\sigma$, and so we can perform the above recursive analysis separately inside each cycle. □

The following lemma shows that the properties of graph polynomials we will establish for delocalized deterministic matrices in Section 5 characterize their traffic distribution. We emphasize our use of a combination of assumptions on limits of the $w$- and $z$-bases that makes this formulation convenient.

**Lemma 3.14.** *Suppose that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies:*

1. *The weak cactus property, i.e., that for all $\alpha\in\mathcal{E}\setminus\mathcal{C}$, $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$.*
2. *For all $\alpha\in\mathcal{A}\setminus\mathcal{E}$, $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, w_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$.*
3. *For all $\sigma\in\mathcal{C}$, there exists $m_\sigma\in\mathbb{R}$ such that $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, w_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} m_\sigma$.*

*Then the traffic distribution of $\boldsymbol{A}$ exists and only depends on $\{m_\sigma : \sigma\in\mathcal{C}\}$.*

*Proof.* We want to show that for every $\alpha \in \mathcal{A}$, $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A})$ exists and only depends on $\{m_\sigma : \sigma \in \mathcal{C}\}$. By assumption, it suffices to prove it for $\alpha \in \mathcal{E} \setminus \mathcal{C}$. By Lemma 3.9,

$$\frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A}) = \frac{1}{n} \sum_{\beta \preceq \alpha} c_{\beta,\alpha} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_\beta(\boldsymbol{A}) \,.$$

By Lemma 3.13, every $\beta$ in the support of the sum is 2-edge-connected. If $\beta \in \mathcal{C}$, then the value of $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\beta(\boldsymbol{A})$ exists and only depends on $\{m_\sigma : \sigma \in \mathcal{C}\}$ by Lemma 3.12. Otherwise, $\beta \in \mathcal{E} \setminus \mathcal{C}$, and $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\beta(\boldsymbol{A}) = 0$ by assumption. This implies that $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A})$ exists and only depends on $\{m_\sigma : \sigma \in \mathcal{C}\}$, which concludes the proof. □

Note that, more generally, by Lemma 3.12, the same statement will hold with Condition 3 of Lemma 3.14 taken in terms of either the $w$- or $z$-basis.

### 3.6 Products and concentration of traffic observables

Recall that the traffic distribution specifies the limits of $\frac{1}{n} \mathbb{E}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A})$ for all $\alpha \in \mathcal{A}$. In all of the random matrix models we consider, these expectations are highly concentrated. We say that *the traffic distribution concentrates for* $\boldsymbol{A}$ if the following property holds, studied in [Mal20].

**Definition 3.15.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and assume that the traffic distribution of $\boldsymbol{A}$ exists. We say that the traffic distribution concentrates for $\boldsymbol{A}$ if for all $k \geq 2$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{A}} \left[ \prod_{j=1}^{k} \frac{1}{n} w_{\alpha_j}(\boldsymbol{A}) \right] = \prod_{j=1}^{k} \lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} w_{\alpha_j}(\boldsymbol{A}) \,.$$

The case $k = 2$ and $\alpha_1 = \alpha_2 = \alpha$ of the definition specializes to the statement:

**Lemma 3.16.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ have traffic distribution $\mathcal{D}$. If the traffic distribution concentrates for $\boldsymbol{A}$, then $\frac{1}{n} \mathbb{E}_{\boldsymbol{A}} w_\alpha(\boldsymbol{A})$ converges to $\mathcal{D}(\alpha)$ in $L^2$.*

The full condition may be viewed as a strengthening of this straightforward notion of concentration. We note that the product of several $w$-basis polynomials is equivalent to taking the disjoint union of their diagrams:

$$w_{\alpha_1}(\boldsymbol{A}) \cdots w_{\alpha_k}(\boldsymbol{A}) = w_{\alpha_1 \sqcup \cdots \sqcup \alpha_k}(\boldsymbol{A}).$$

Therefore, Definition 3.15 says that the values of disconnected diagrams asymptotically factor over the components. This justifies defining $\mathcal{A}$ and the traffic distribution to include only connected diagrams. The following shows that concentration may equally well be considered in the $z$-basis.

**Lemma 3.17** ([Mal20, Lemma 2.9])**.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and assume that the traffic distribution of $\boldsymbol{A}$ exists. The traffic distribution concentrates for $\boldsymbol{A}$ if and only if, for all $k \geq 2$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{A}} \left[ \prod_{j=1}^{k} \frac{1}{n} z_{\alpha_j}(\boldsymbol{A}) \right] = \prod_{j=1}^{k} \lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_{\alpha_j}(\boldsymbol{A}) \,.$$

For vector diagrams, the componentwise or Hadamard product is

$$\boldsymbol{w}_{\alpha_1}(\boldsymbol{A}) \cdots \boldsymbol{w}_{\alpha_k}(\boldsymbol{A}) = \boldsymbol{w}_{\alpha_1 \oplus \cdots \oplus \alpha_k}(\boldsymbol{A}) \,,$$

where $\alpha_1 \oplus \cdots \oplus \alpha_k$ is the diagram formed by taking the disjoint union of $\alpha_1$ through $\alpha_k$ and then identifying the roots together into a single root. We sometimes refer to this operation as *grafting* $\alpha_1, \ldots, \alpha_k$ at the root.

# 4 Traffic Distributions of Random Matrices

As both a technical preliminary for our results and useful background, this section describes the traffic distributions of several common random matrix ensembles. A common theme is that all of these classical models satisfy the strong cactus property. Most of these results have appeared previously in the literature, though we provide some extensions and new interpretations.

## 4.1 Wigner random matrices

A *Wigner matrix* is a random symmetric matrix with i.i.d. entries on and above the diagonal. Changes to the diagonal entries such as setting them to zero (which is the convention used in some works), or taking the diagonal variances to be twice the off-diagonal ones (as in the GOE model), do not affect the results.

The limiting traffic distribution of a sequence of Wigner matrices was derived by Male [Mal20], by generalizing the combinatorial proof of the semicircle limit theorem for the limiting spectral distribution [AGZ10]. The same result was re-discovered in [JP25] in the context of analyzing pGFOM on such matrices.

**Theorem 4.1** (Traffic distribution of Wigner matrices)**.** *Let $\nu$ be a probability measure on $\mathbb{R}$ with all moments finite, mean 0, and variance 1. For all $n \geq 1$, let $\widetilde{\boldsymbol{A}}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ have entries on and above the diagonal drawn i.i.d. from $\nu$. Define $\boldsymbol{A}^{(n)} := \frac{1}{\sqrt{n}}\widetilde{\boldsymbol{A}}^{(n)}$. Then, for all $\alpha \in \mathcal{A}$,*

$$\lim_{n\to\infty} \frac{1}{n}\,\mathbb{E}\, z_\alpha(\boldsymbol{A}^{(n)}) = \begin{cases} 1 & \text{if } \alpha \text{ is a cactus of 2-cycles,} \\ 0 & \text{otherwise.} \end{cases}$$

The same result holds for normalized GOE matrices. Note that a cactus of 2-cycles may equivalently be viewed as a "doubled tree", a tree where every edge is repeated exactly twice, which is the formulation used in the previous works [Mal20, JP25].

Thus, sequences of Wigner matrices have the factorizing strong cactus property, with the especially simple sequence of free cumulants $\kappa_2 = 1$ and $\kappa_q = 0$ for all $q \neq 2$. These are also the free cumulants of the semicircle law, which is the limiting eigenvalue distribution of $\boldsymbol{A}^{(n)}$.

## 4.2 Orthogonally invariant random matrices

Let the orthogonal group $O(n)$ act on $\mathbb{R}^{n\times n}_{\mathrm{sym}}$ by conjugation, with $\boldsymbol{Q} \in O(n)$ acting as $\boldsymbol{Q}\cdot\boldsymbol{A} := \boldsymbol{Q}^\top \boldsymbol{A}\boldsymbol{Q}$. Let $\mu$ denote a probability measure on $\mathbb{R}^{n\times n}_{\mathrm{sym}}$ that is invariant under this action of $O(n)$. In this case, we call $\boldsymbol{A} \sim \mu$ an *orthogonally invariant random matrix*.

If $\mu$ has a density on $\mathbb{R}^{n\times n}_{\mathrm{sym}}$, an equivalent condition is that the density at $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ depends only on the unordered multiset of eigenvalues of $\boldsymbol{A}$. An important class of examples in physics is given by *matrix models with potential* $V : \mathbb{R} \to \mathbb{R}$, whose density is proportional to $\exp(-\operatorname{Tr} V(\boldsymbol{A}))$.

For example, the GOE model corresponds to $V(t) = t^2/2$. We will come back to these examples in Appendix A.

For the complex-valued variant where $O(n)$ is replaced by the unitary group $U(n)$, the limiting traffic distribution of such *unitarily invariant* random matrices is described in [CDM24, Theorem 1.1]. The same description holds in the orthogonal case. The proof is a straightforward generalization of the unitarily invariant case, but for the sake of completeness we present it in detail in Appendix B.

**Theorem 4.2** (Traffic distribution of orthogonally invariant random matrices)**.** *Let $\boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be a sequence of orthogonally invariant random matrices that converges in tracial moments in $L^2$ to a probability measure $\mu$. Then, for all $\alpha \in \mathcal{A}$,*

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\, z_\alpha(\boldsymbol{A}^{(n)}) = \begin{cases} \prod_{\sigma\in \mathrm{cyc}(\alpha)} \kappa_{|\sigma|} & \text{if } \alpha \in \mathcal{C}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

*where $\kappa_q$ is the $q$th free cumulant of $\mu$ (Definition 3.10), and $|\sigma|$ denotes the length of the cycle.*

Eq. (10) shows that the factorizing strong cactus property holds for orthogonally invariant random matrices, and in particular their limiting traffic distribution is supported only on cactus diagrams in the $z$-basis.

Actually, in this case the strong cactus property is non-trivial only for the Eulerian diagrams, since the non-Eulerian ones have identically zero expectation for each fixed dimension $n$:

**Claim 4.3.** *Let $\boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be an orthogonally invariant random matrix. Then for all $\alpha \in \mathcal{A}$ which are not Eulerian, $\mathbb{E}\, z_\alpha(\boldsymbol{A}^{(n)}) = 0$.*

We show this at the beginning of our proof in Appendix B.

Both the proof of [CDM24, Theorem 1.1] and our proof of Theorem 4.2 are based on the *Weingarten calculus*, a combinatorial description of the entrywise moments of Haar-distributed matrices from a matrix group. In Appendix A, we present an alternative (albeit non-rigorous) derivation of Theorem 4.2 using the *Feynman diagram* method from physics. Arguably, the combinatorics of the Feynman diagram method is simpler than that of the Weingarten calculus proof.

### 4.3 Block-structured random matrices

Wigner random matrices and orthogonally invariant random matrices both extend the GOE in different directions, while still satisfying the factorizing strong cactus property. We now consider a third generalization, block matrices, which typically do *not* satisfy the factorizing property.

Fix $q \in \mathbb{N}$. For $r, c \in [q]$, let $\boldsymbol{A}_{r,c} = \boldsymbol{A}^{(n)}_{r,c} \in \mathbb{R}^{n/q\times n/q}_{\mathrm{sym}}$ be a sequence of random matrices with $\boldsymbol{A}_{r,c} = \boldsymbol{A}_{c,r}$. The corresponding *block matrix model* is the symmetric $n$-by-$n$ matrix whose rows and columns are partitioned into blocks of sizes $n/q$ which has blocks $(\boldsymbol{A}_{r,c})_{r,c\in[q]}$. We let $\mathrm{block}(i) \in [q]$ denote the block label of $i \in [n]$.

The simplest example of a block matrix model is the *block GOE* model, which has previously been studied in the context of the Generalized AMP algorithm [JM13].[10]

[10] In this paper, we study a slightly more symmetric variant, in which the blocks themselves are symmetric. This

**Definition 4.4** (Block GOE model)**.** *Let $q \in \mathbb{N}$ and let $\mathbf{\Sigma} \in \mathbb{R}^{q \times q}$ be a symmetric with nonnegative entries. For $1 \le r \le c \le q$, let $\boldsymbol{A}_{r,c} \in \mathbb{R}_{\mathrm{sym}}^{n/q \times n/q}$ be a symmetric random matrix whose entries on and above the diagonal are independent Gaussians with mean 0 and variance $\mathbf{\Sigma}[r,c]/n$, and let $\boldsymbol{A}_{c,r} = \boldsymbol{A}_{r,c}$ for $q \ge r > c \ge 1$. The block GOE model $\boldsymbol{A} \sim \mathsf{BlockGOE}(n, \mathbf{\Sigma})$ is the block matrix with blocks $(\boldsymbol{A}_{r,c})_{r,c \in [q]}$.*

Following the arguments of [Mal20, JP25], one can prove that the block GOE model with fixed parameter $\mathbf{\Sigma}$ satisfies the strong cactus property. Indeed, as in Theorem 4.1, it is still only the doubled trees or cactuses of 2-cycles that have non-zero value in the traffic distribution. However, these values depend non-trivially on $\mathbf{\Sigma}$, and in general the block GOE model does not satisfy the factorizing strong cactus property.[11]

**Traffic independence.** We study block models through the notion of *traffic independence*. Traffic independence was introduced by Male [Mal20] as a generalization of free independence of matrices. Free independence is a property of the mixed traces of several random matrices (in our notation, these traces are represented by cycle diagrams), whereas traffic independence is a property of *all* diagrams. Using this concept, below we prove a general result that block-structured matrices have the strong cactus property provided that *(i)* each of the blocks separately has the strong cactus property, and *(ii)* those blocks are asymptotically traffic independent.

For a sequence of symmetric matrices $(\boldsymbol{A}_1, \dots, \boldsymbol{A}_k) \in (\mathbb{R}_{\mathrm{sym}}^{n \times n})^k$, we generalize the graph polynomials to $w_\alpha(\boldsymbol{A}_1, \dots, \boldsymbol{A}_k)$ and $z_\alpha(\boldsymbol{A}_1, \dots, \boldsymbol{A}_k)$, where $\alpha$ is a multigraph whose edges are additionally colored by $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$. The graph polynomial defined by $\alpha$ uses the entries of $\boldsymbol{A}_i$ on each edge whose color is $\boldsymbol{A}_i$, as in Definition 3.6.

Define a *colored component* to be a maximal connected subgraph of $\alpha$ whose edges all have the same label $\boldsymbol{A}_i$. Let $\mathrm{CC}(\alpha)$ denote the set of colored components. Define the *graph of colored components* $\mathrm{GCC}(\alpha)$ to be the bipartite graph $\chi$ with:

$$V(\chi) = \mathrm{CC}(\alpha) \cup \{u \in V(\alpha) : u \text{ belongs to at least two colored components}\},$$
$$E(\chi) = \{(\mathcal{C}, u) : u \text{ belongs to the colored component } \mathcal{C}\}.$$

**Definition 4.5** (Traffic independence)**.** *Let $(\boldsymbol{A}_1, \dots, \boldsymbol{A}_k) = (\boldsymbol{A}_1^{(n)}, \dots, \boldsymbol{A}_k^{(n)}) \in (\mathbb{R}_{\mathrm{sym}}^{n \times n})^k$ be sequences of symmetric random matrices, with respective limiting traffic distributions $\mathcal{D}_1, \dots, \mathcal{D}_k$. We say that $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are asymptotically traffic independent if, for all connected undirected multigraphs $\alpha$ with edges labeled by $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$,*

$$\lim_{n \to \infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}_1, \dots, \boldsymbol{A}_k} z_\alpha(\boldsymbol{A}_1, \dots, \boldsymbol{A}_k) = \begin{cases} \displaystyle\prod_{\mathcal{C} \in \mathrm{CC}(\alpha)} \mathcal{D}_{i(\mathcal{C})}(\mathcal{C}) & \text{if } \mathrm{GCC}(\alpha) \text{ is a tree} \\ 0 & \text{otherwise} \end{cases}$$

*Here, $i(\mathcal{C})$ denotes the matrix label associated with the colored component $\mathcal{C}$.*

Next, we prove that traffic independence of the blocks preserves the strong cactus property:

**Proposition 4.6.** *Let $q \in \mathbb{N}$. For $r, c \in [q]$, let $\boldsymbol{A}_{r,c} = \boldsymbol{A}_{r,c}^{(n)} \in \mathbb{R}_{\mathrm{sym}}^{n/q \times n/q}$ be a sequence of symmetric*

modification is made purely for technical reasons, since we work in our other definitions only with symmetric matrices.

[11] If the row sums of $\mathbf{\Sigma}$ are constant, yielding what is sometimes called a *generalized Wigner matrix*, then up to rescaling the traffic distribution is again that of the GOE and the factorizing property does hold.

*random matrices such that* $\boldsymbol{A}_{r,c} = \boldsymbol{A}_{c,r}$. *Assume that each* $\boldsymbol{A}_{r,c}$ *has a limiting traffic distribution that satisfies the strong cactus property and* $(\boldsymbol{A}_{r,c})_{1\le r\le c\le q}$ *are asymptotically traffic independent. Then, the block matrix* $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *with blocks* $(\boldsymbol{A}_{r,c})_{r,c\in[q]}$ *also has a limiting traffic distribution that satisfies the strong cactus property.*

*Proof.* Let $\alpha \in \mathcal{A}$. In the graph polynomial $z_\alpha(\boldsymbol{A})$ we partition the sum based on the block of each vertex:

$$\frac{1}{n} z_\alpha(\boldsymbol{A}) = \frac{1}{n} \sum_{\chi: V(\alpha)\to[q]} \sum_{i: V(\alpha)\to[\frac{n}{q}]} \prod_{uv\in E(\alpha)} \boldsymbol{A}_{\chi(u),\chi(v)}[i(u), i(v)]\,.$$

We can interpret the inner summation as a generalized graph polynomial whose edges are labeled by the matrices $\boldsymbol{A}_{r,c}$. Call this diagram $\alpha_\chi$ and write:

$$\frac{1}{n} z_\alpha(\boldsymbol{A}) = \sum_{\chi: V(\alpha)\to[q]} \frac{1}{n} z_{\alpha_\chi}((\boldsymbol{A}_{r,c})_{r,c\in[q]})\,.$$

Taking the expectation and the limit $n \to \infty$, by traffic independence, all limits exist (so the block matrix has a limiting traffic distribution), and the nonzero terms on the right-hand side are those for which $\mathrm{GCC}(\alpha_\chi)$ is a tree. By the strong cactus property for each $\boldsymbol{A}_{rc}$, each colored component must be a cactus. Therefore, any nonzero $\alpha$ is formed by gluing several cactuses along a tree, which forms a bigger cactus. □

Finally, traffic independence is shown in [Mal20] to hold quite generally for independent random matrices $\boldsymbol{A}_i$, each of which has a permutation-invariant distribution.

**Theorem 4.7** ([Mal20, Theorem 1.8])**.** *Let* $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *be independent random matrices such that for each* $i \in [k]$,

*(i) The law of* $\boldsymbol{A}_i \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *is* $S_n$*-invariant (i.e., invariant under the simultaneous action of* $S_n$ *on the rows and columns of* $\boldsymbol{A}_i$*).*

*(ii) The limiting traffic distribution of* $\boldsymbol{A}_i$ *exists.*

*(iii) The traffic distribution concentrates for* $\boldsymbol{A}_i$ *(Definition 3.15).*

*Then* $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ *are asymptotically traffic independent.*

Together with Proposition 4.6, Theorem 4.7 implies that block-structured matrices with independent blocks, each satisfying the strong cactus property and Conditions *(i), (ii), (iii)* also satisfy the strong cactus property (such as the block GOE matrix). We note that Condition *(i)* can be ensured by applying an independent random permutation to the rows and columns of each $\boldsymbol{A}_i$. Condition *(iii)* is proven for orthogonally invariant random matrices in Lemma B.7.

## 5 Universality for Deterministic Matrices

Recall the definition of puncturing (Definition 2.1) and of the r-ROM (Definition 2.5). Our main theorem in this section is:

**Theorem 5.1.** *Let $\boldsymbol{H} = \boldsymbol{H}^{(n)} \in \mathbb{R}^{n\times n}_{\text{sym}}$ be a sequence of symmetric orthogonal matrices such that*

$$\max_{1\le i\le j\le n} |\boldsymbol{H}[i,j]| \le n^{-\frac{1}{2}+o(1)}. \tag{11}$$

*Then, the limiting traffic distribution of the puncturing of $\boldsymbol{H}$ exists and equals that of the* r-ROM.

Theorem 5.1 directly applies to $\boldsymbol{H}$ being the sequence of Walsh-Hadamard matrices, discrete sine transform matrices, or discrete cosine transform matrices. Theorem 5.1 follows from the more general Theorem 5.3 below, which applies to symmetric matrices that are not necessarily orthogonal, but have a limiting diagonal distribution and satisfy a generalized delocalization assumption.

**Assumption 5.2.** *Let $\boldsymbol{H} = \boldsymbol{H}^{(n)} \in \mathbb{R}^{n\times n}_{\text{sym}}$ and $\varepsilon = \varepsilon^{(n)} > 0$. We introduce the assumptions:*

$$\|\boldsymbol{H}\| \le 1, \tag{12}$$

$$\max_{1\le i<j\le n} |\boldsymbol{W}_\alpha(\boldsymbol{H})[i,j]| \le \varepsilon \qquad \text{for each open cactus } \alpha \text{ (Definition 5.4)}, \tag{13}$$

$$\frac{1}{\sqrt{n}}\|\boldsymbol{\Pi}\boldsymbol{w}_\sigma(\boldsymbol{H})\|_2 \le \varepsilon \qquad \text{for all } \sigma \in \mathcal{C}_1, \tag{14}$$

*where $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(n)} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ denotes the projection orthogonal to the all-ones direction.*

For example, one of the constraints of Eq. (13) is that $|\boldsymbol{H}^k[i,j]| \le \varepsilon$ uniformly for all $k, n \in \mathbb{N}$ and distinct $i, j \in [n]$ (a bound which is uniform in $n, i, j$ but may depend on $k$ would also be sufficient, but we omit this for simplicity).

**Theorem 5.3** (Universality)**.** *Let $\boldsymbol{H} = \boldsymbol{H}^{(n)} \in \mathbb{R}^{n\times n}_{\text{sym}}$, $\boldsymbol{A}$ be the puncturing of $\boldsymbol{H}$, and $\varepsilon^{(n)} > 0$.*

1. *If $\boldsymbol{H}$ satisfies Eqs. (12) and (13), then for all $\alpha \in \mathcal{E} \setminus \mathcal{C}$,*

$$\frac{1}{n}|z_\alpha(\boldsymbol{H})| \le O_\alpha\left(\varepsilon^{(n)} + \frac{1}{\sqrt{n}}\right) \qquad \text{and} \qquad \frac{1}{n}|z_\alpha(\boldsymbol{A})| \le O_\alpha\left(\varepsilon^{(n)} + \frac{1}{\sqrt{n}}\right).$$

   *In particular, if $\varepsilon^{(n)} = o(1)$, then both $\boldsymbol{H}$ and $\boldsymbol{A}$ satisfy the weak cactus property.*

2. *If $\boldsymbol{H}$ satisfies Eqs. (12) to (14), then for all $\alpha \in \mathcal{A} \setminus \mathcal{E}$,*

$$\frac{1}{n}|w_\alpha(\boldsymbol{A})| \le \frac{1}{\sqrt{n}} \cdot \left(1 + \varepsilon^{(n)}\sqrt{n}\right)^{O_\alpha(1)}.$$

   *In particular, if $\varepsilon^{(n)} = n^{-\frac{1}{2}+o(1)}$, then the right-hand side is $n^{-\frac{1}{2}+o_\alpha(1)}$.*

*Hence, if $\boldsymbol{H}$ satisfies Eqs. (12) to (14) with $\varepsilon^{(n)} = n^{-\frac{1}{2}+o(1)}$, and the diagonal distribution of $\boldsymbol{H}$ exists, then the traffic distribution of $\boldsymbol{A}$ exists and is determined by the diagonal distribution of $\boldsymbol{H}$.*

We emphasized in the statement that all constants in the $O$ notations depend on $\alpha$. We will drop this dependency in the rest of the section.

**Comparison with prior work.** In [WZF22, Theorem 2.8], the authors assume *(i)* delocalization of open cactuses (Eq. (13)) and *(ii)* the existence of a limiting diagonal distribution. They show that, after conjugation by a randomly signed permutation matrix, the resulting "semi-random" matrix

lies in the same universality class (in the sense of AMP dynamics) as an orthogonally invariant matrix with the same diagonal distribution. Theorem 5.3 shows that the same conclusion holds for deterministic matrices, if we replace random conjugation with puncturing.

The universality result of [WZF22] can also be extended in a black-box way to deterministic matrices, but only for GFOM with *odd* nonlinearities [DLS23, ZWF24]. This assumption lets one only consider the limiting traffic distribution evaluated on *Eulerian* diagrams. Under the same assumption, our proof would also significantly simplify. Indeed, in Theorem 5.1, the number of monomials appearing in $w_\alpha(\boldsymbol{H})$ is $O(n^{|V(\alpha)|})$, and each term has magnitude $\max_{i,j\in[n]} |\boldsymbol{H}[i,j]|^{|E(\alpha)|} \le n^{-|E(\alpha)|/2+o(1)}$, giving the upper bound $|w_\alpha(\boldsymbol{H})| \le n^{o(1)}$ if $\alpha$ has minimum degree 4. It only remains to incorporate paths of degree-2 vertices, which simply compute $\boldsymbol{H}^k \in \{\boldsymbol{I}, \boldsymbol{H}\}$ for some $k \ge 1$.

### 5.1 Calculation of cactus diagrams and diagonal distribution

To apply Theorem 5.3, one needs to compute the diagonal distribution of $\boldsymbol{H}$ and small strengthenings of it in order to verify Assumption 5.2. Notice that the only diagrams involved in the assumptions are cactuses, so this is a much simpler task than calculating the entire traffic distribution. In this subsection, we do this calculation directly to prove Theorem 5.1 assuming Theorem 5.3.

Let $\boldsymbol{H}$ be a delocalized orthogonal matrix satisfying the assumption of Theorem 5.1. Note that it satisfies $\boldsymbol{H}^2 = \boldsymbol{I}$. Hence, Eq. (12) is automatic. Next, we define the notion of *open cactus* appearing in Eq. (13). An open cactus is a matrix diagram with two roots such that merging the roots yields a cactus.

**Definition 5.4.** *An open cactus is a graph obtained from a simple path by attaching vertex-disjoint cactuses to each vertex of the path. Formally, $\alpha = (V(\alpha), E(\alpha))$ is an open cactus if there exist $k \ge 2$, vertex-disjoint cactuses $\beta_1, \dots, \beta_k$, and distinct vertices $u_1 \in V(\beta_1), \dots, u_k \in V(\beta_k)$ with*

$$V(\alpha) = \bigcup_{i=1}^{k} V(\beta_i)\,, \quad E(\alpha) = \{\{u_i, u_{i+1}\} : i \in \{1, \dots, k-1\}\} \cup \bigcup_{i=1}^{k} E(\beta_i)\,.$$

*We call $(u_1, u_k)$ the endpoints of $\alpha$, and $(u_1, \dots, u_k)$ the base path of $\alpha$. Unless specified otherwise, we will view an open cactus $\alpha \in \mathcal{A}_2$ as a matrix diagram rooted at its two ordered endpoints.*

In general, if $\alpha$ is a matrix diagram and $\alpha'$ is the scalar diagram formed by merging the roots of $\alpha$, then $\mathrm{Tr}(\boldsymbol{W}_\alpha(\boldsymbol{A})) = w_{\alpha'}(\boldsymbol{A})$. For an open cactus $\alpha$, this $\alpha'$ is a cactus, and so $w_{\alpha'}(\boldsymbol{A})$ is one of the quantities whose limit is included in the diagonal distribution of $\boldsymbol{A}$; further, all values of the diagonal distribution can be obtained in this way from the *diagonal* entries of open cactus matrices. From this perspective, Eq. (13) is a natural counterpart to the diagonal distribution since it concerns all of the *off-diagonal* entries of the open cactus matrices.

We compute the open cactus matrices for $\boldsymbol{H}$ in the following lemma.

**Lemma 5.5.** *Let $\sigma$ be an open cactus and let $\boldsymbol{H}$ satisfy Eq. (11). If all cycles in all of the hanging cactuses have even length, then $\boldsymbol{W}_\sigma(\boldsymbol{H}) = \boldsymbol{I}$ if the base path has even length and $\boldsymbol{W}_\sigma(\boldsymbol{H}) = \boldsymbol{H}$ if the base path has odd length. Otherwise, $\|\boldsymbol{W}_\sigma(\boldsymbol{H})\| \le n^{-\frac{1}{2}+o(1)}$.*

*Proof.* First, the leaf 2-vertex-connected components of $\sigma$ consisting of cycles of even length can be iteratively removed without changing the value of $\boldsymbol{W}_\sigma(\boldsymbol{H})$. This is because a hanging cycle of even length $k$ contributes $\mathrm{diag}(\boldsymbol{H}^k) = \mathrm{diag}(\boldsymbol{I}) = \boldsymbol{1}$ in the definition of $\boldsymbol{W}_\sigma$. Therefore, if all cycles in all

hanging cactuses have even length, then $\boldsymbol{W}_\sigma(\boldsymbol{H}) = \boldsymbol{H}^\ell \in \{\boldsymbol{I}, \boldsymbol{H}\}$ where $\ell$ is the length of the base path.

In the remaining case where $\sigma$ has an odd cycle, we use induction. Let $\beta_1, \dots, \beta_k$ be the hanging cactuses of $\sigma$. We convert each $\beta_i$ into an open cactus diagram $\beta_i'$ by splitting the vertex at which $\beta_i$ meets $\sigma$. With this notation, we have the matrix factorization:

$$\boldsymbol{W}_\sigma(\boldsymbol{H}) = \operatorname{diag}(\boldsymbol{W}_{\beta_1'}(\boldsymbol{H}))\boldsymbol{H} \operatorname{diag}(\boldsymbol{W}_{\beta_2'}(\boldsymbol{H}))\boldsymbol{H} \dots \boldsymbol{H}\operatorname{diag}(\boldsymbol{W}_{\beta_k'}(\boldsymbol{H}))\,.$$

The odd cycle in $\sigma$ has either become an odd-length base path in some $\beta_i'$ or it continues to be an odd cycle in some $\beta_i'$. In the second case, by sub-multiplicativity of the spectral norm,

$$\|\boldsymbol{W}_\sigma(\boldsymbol{H})\| \le \left\|\operatorname{diag}(\boldsymbol{W}_{\beta_i'}(\boldsymbol{H}))\right\| \le \left\|\boldsymbol{W}_{\beta_i'}(\boldsymbol{H})\right\| \le n^{-\frac{1}{2}+o(1)}$$

with the last inequality by induction. In the first case, we have $\boldsymbol{W}_{\beta_i'}(\boldsymbol{H}) = \boldsymbol{H}$. Then

$$\left\|\operatorname{diag}(\boldsymbol{W}_{\beta_i'}(\boldsymbol{H}))\right\| = \|\operatorname{diag}(\boldsymbol{H})\| \le n^{-\frac{1}{2}+o(1)}$$

by the delocalization assumption, and this case is also complete. □

We use the lemma to complete the proof of Theorem 5.1.

*Proof of Theorem 5.1 from Theorem 5.3.* Eq. (12) holds automatically for $\boldsymbol{H}$ a symmetric orthogonal matrix. Verifying Eq. (13), Lemma 5.5 implies that the off-diagonal entries of all open cactus matrices satisfy

$$\max_{1\le i<j\le n} |\boldsymbol{W}_\sigma(\boldsymbol{H})[i,j]| \le \|\boldsymbol{W}_\sigma(\boldsymbol{H})\| \le n^{-\frac{1}{2}+o(1)}$$

when $\sigma$ has an odd cycle, and the remaining cases $\boldsymbol{W}_\sigma(\boldsymbol{H}) = \boldsymbol{H}$ or $\boldsymbol{W}_\sigma = \boldsymbol{I}$ are easily checked.

Next, each vector cactus diagram $\sigma \in \mathcal{C}_1$ satisfies $\boldsymbol{w}_\sigma(\boldsymbol{H}) = \operatorname{diag}(\boldsymbol{W}_{\sigma'}(\boldsymbol{H}))$ where $\sigma'$ is an open cactus obtained by splitting the root of $\sigma$. By Lemma 5.5 the diagonal of an open cactus matrix is either $\mathbf{1}$ (in which case Eq. (14) is satisfied with $\varepsilon = 0$) or it satisfies

$$\frac{1}{\sqrt{n}}\|\operatorname{diag}(\boldsymbol{W}_{\sigma'}(\boldsymbol{H}))\|_2 \le \|\operatorname{diag}(\boldsymbol{W}_{\sigma'}(\boldsymbol{H}))\|_\infty \le n^{-\frac{1}{2}+o(1)}\,,$$

in which case Eq. (14) is satisfied with $\varepsilon = n^{-\frac{1}{2}+o(1)}$.

The diagonal distribution is computed by averaging the diagonal entries of open cactus matrices:

$$\lim_{n\to\infty} \frac{1}{n} w_\sigma(\boldsymbol{H}) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \boldsymbol{W}_{\sigma'}[i,i] = \begin{cases} 1 & \text{if all cycles in } \sigma \text{ have even length} \\ 0 & \text{otherwise} \end{cases}$$

where on the left-hand side, we convert $\sigma \in \mathcal{C}_0$ to an open cactus diagram $\sigma'$ by rooting it arbitrarily and splitting the root. The right-hand side is by Lemma 5.5. That is, the diagonal distribution of $\boldsymbol{H}$ is just the indicator function that all cycles of the cactus are even.

Thus, we showed that Eqs. (12) to (14) hold and the diagonal distribution converges to the same fixed limit for any orthogonal matrix with delocalized entries. By Theorem 5.3, the traffic distribution of such matrices exists and is always the same.

Finally, we show that the r-ROM is also in this class, by showing that, after conditioning on a suitable high-probability event, the above argument applies to an r-ROM matrix as well. Let $\boldsymbol{H}_{\mathsf{ROM}} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$, where $\boldsymbol{Q}$ is Haar-distributed and $\boldsymbol{D}$ is diagonal with i.i.d. $\pm 1$ entries, independent of $\boldsymbol{Q}$.

**Claim 5.6.** *There exists $c > 0$ such that for any $t > 0$,*

$$\max_{i,j\in[n]} |\boldsymbol{H}_{\mathsf{ROM}}[i,j]| \le t^2 n^{-\frac{1}{2}} \tag{15}$$

*holds with probability at least $1 - n^2 e^{-ct^2}$ .*

*Proof.* Since every entry of $\boldsymbol{Q}$ is $O(n^{-1/2})$-subgaussian, by a union bound

$$\max_{i,j\in[n]} |\boldsymbol{Q}[i,j]| \le t n^{-\frac{1}{2}}$$

holds with probability at least $1 - n^2 e^{-\Omega(t^2)}$. Next, we have $\boldsymbol{H}_{\mathsf{ROM}}[i,j] = \sum_{k=1}^n \boldsymbol{D}[k,k]\boldsymbol{Q}[i,k]\boldsymbol{Q}[j,k]$, which, conditioned on $\boldsymbol{Q}$, is a sum of independent random variables. By Hoeffding's bound, any fixed entry of $\boldsymbol{H}_{\mathsf{ROM}}$ is $O(\sigma)$-subgaussian with parameter

$$\sigma^2 := \sum_{k=1}^n \boldsymbol{Q}[i,k]^2 \boldsymbol{Q}[j,k]^2 \le \max_{i,j\in[n]} \boldsymbol{Q}[i,j]^2 ,$$

since every row of $\boldsymbol{Q}$ has $\ell_2$-norm 1. The conclusion follows from a union bound over all entries. □

Fix $\alpha \in \mathcal{A}$. Let $E_n$ denote the event Eq. (15), with $t = n^{o(1)}$. By the law of total expectation, we decompose

$$\frac{1}{n}\mathbb{E}\, w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) = \frac{1}{n}\mathbb{E}[w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) \mid E_n]\Pr(E_n) + \frac{1}{n}\mathbb{E}[w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) \mid E_n^c]\Pr(E_n^c) .$$

The left-hand side converges to the traffic distribution of the r-ROM evaluated at $\alpha$. Moreover, since $\|\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}\| \le 1$, we may crudely bound the second term by

$$\frac{1}{n}\mathbb{E}\left[w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) \mid E_n^c\right] \cdot \Pr(E_n^c) \le n^{|V(\alpha)|-1}\Pr(E_n^c) \underset{n\to\infty}{\longrightarrow} 0 .$$

Since $\Pr(E_n) \underset{n\to\infty}{\longrightarrow} 1$, we deduce that

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) \mid E_n\right] = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\, w_\alpha(\boldsymbol{\Pi}\boldsymbol{H}_{\mathsf{ROM}}\boldsymbol{\Pi}) .$$

Finally, on the event $E_n$, the matrix $\boldsymbol{H}_{\mathsf{ROM}}$ satisfies the assumptions of Theorem 5.1. Consequently, the traffic distribution of punctured delocalized orthogonal matrices coincides with that of the r-ROM, as desired. □

As a consequence of the above argument, the traffic distribution of the r-ROM is specified implicitly as the solution to the following equations:

1. For every $\alpha \in \mathcal{A} \setminus \mathcal{E}$, $\frac{1}{n}\mathbb{E}\, w_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$.

2. For every $\alpha \in \mathcal{E} \setminus \mathcal{C}$, $\frac{1}{n}\mathbb{E}\, z_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$.

3. For every $\sigma \in \mathcal{C}$, $\frac{1}{n}\mathbb{E}\, w_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 1$ if all cycles of $\sigma$ are even and 0 otherwise.

These equations determine a unique traffic distribution by Lemma 3.14. It is possible to give an explicit but much more complicated description using the Weingarten calculus, which we do in Appendix B.6. However, the above characterization is arguably the conceptually clearer one, and we emphasize that it involves both the $w$- and $z$-bases.

We note also as a point of reference that the last part, the limiting values of cactuses in the $w$-basis, are the same as those for the (unpunctured) ROM, as follows from combining Claim 3.11 with Lemma 3.12, and corresponds simply to the moments of the Rademacher distribution being 1 for moments of even order and 0 for ones of odd order.

## 5.2 The fundamental theorem of graph polynomials

The main proof of Theorem 5.3 throughout the rest of the section relies on the "fundamental theorem of graph polynomials" of Bai and Silverstein [BS10]. This result can be used to easily bound 2-edge-connected graph polynomials expressed in the $w$-basis, which is one reason that it is convenient to restrict to such diagrams in our definition of the weak cactus property. The proof of the fundamental theorem uses a spectral bound on tensor powers of $\boldsymbol{A}$; see [MS12] for another related result.

**Theorem 5.7** ([BS10, Theorems A.31 and A.32])**.** *For every $n \geq 1$, $\alpha \in \mathcal{E} \cup \mathcal{E}_1 \cup \mathcal{E}_2$ and collection of $n \times n$ symmetric matrices $\boldsymbol{\mathcal{A}} = (\boldsymbol{A}_e)_{e\in E(\alpha)}$,*

$$\begin{aligned}
\frac{1}{n}|w_\alpha(\boldsymbol{\mathcal{A}})| &\leq \prod_{e\in E(\alpha)} \|\boldsymbol{A}_e\| \quad \text{if } \alpha \in \mathcal{E}, \\
\|\boldsymbol{w}_\alpha(\boldsymbol{\mathcal{A}})\|_\infty &\leq \prod_{e\in E(\alpha)} \|\boldsymbol{A}_e\| \quad \text{if } \alpha \in \mathcal{E}_1, \\
\|\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}})\| &\leq \prod_{e\in E(\alpha)} \|\boldsymbol{A}_e\| \quad \text{if } \alpha \in \mathcal{E}_2.
\end{aligned}$$

The result of [BS10] only covers scalar and matrix diagrams, but we provide a quick reduction of the vector case to the scalar case.

*Proof of vector case of Theorem 5.7.* For all $q \geq 1$, we can diagrammatically express $\|\boldsymbol{w}_\alpha(\boldsymbol{\mathcal{A}})\|_{2q}^{2q}$ as the diagram formed by merging $2q$ copies of $\alpha$ at the root, and then forgetting the identity of the root to obtain a scalar diagram. Let $\alpha_{2q} = \alpha^{\oplus 2q}$ denote this diagram. The graph $\alpha_{2q}$ remains 2-edge-connected, therefore by the scalar case of the result we have:

$$\|\boldsymbol{w}_\alpha(\boldsymbol{\mathcal{A}})\|_{2q}^{2q} = w_{\alpha_{2q}}(\boldsymbol{\mathcal{A}}) \leq n \cdot \left(\prod_{e\in E(\alpha)} \|\boldsymbol{A}_e\|\right)^{2q}.$$

Taking $q \to \infty$ with $n$ fixed, we obtain $\|\boldsymbol{w}_\alpha(\boldsymbol{\mathcal{A}})\|_\infty \leq \prod_{e\in E(\alpha)} \|\boldsymbol{A}_e\|$. □

We will apply the fundamental theorem by decomposing a general graph into its 2-edge-connected components, which are joined together by a tree of bridge edges. Decomposing diagrams into their 2-edge-connected components is also a fundamental idea in physics, where a 2-edge-connected Feynman diagram is called a "1-particle-irreducible diagram".

### 5.3 Main structural lemma: Open cactus decomposition

To prove the weak cactus property of Theorem 5.3, we begin by observing that any 2-edge-connected non-cactus graph contains three edge-disjoint paths between some pair of vertices. How can we quantify that such a graph is a cactus plus excess edges? We answer this question by introducing the *open cactus decomposition*. Our main structural result is that one can identify an "extra" open cactus subgraph inside any 2-edge-connected graph which is not a cactus, in the sense that the subgraph can be removed without spoiling 2-edge-connectedness.

**Proposition 5.8.** *For any $\alpha \in \mathcal{E}_1 \setminus \mathcal{C}_1$, there exist distinct $s, t \in V(\alpha)$ and an induced subgraph $\beta$ of $\alpha$ such that*

1. *$\beta$ is an open cactus with endpoints $\{s, t\}$.*
2. *$\alpha\,[V(\alpha) \setminus (V(\beta) \setminus \{s, t\})]$ is 2-edge-connected.*
3. *$\mathrm{root}(\alpha) \notin V(\beta) \setminus \{s, t\}$.*

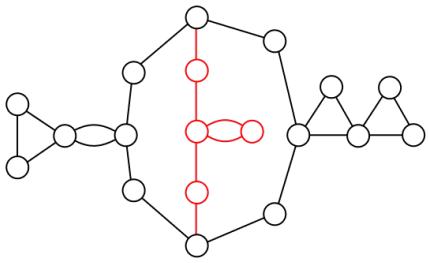


Figure 3: Example for Proposition 5.8 of a 2-edge-connected graph which is not a cactus. If the open cactus in red is removed, the graph remains 2-edge-connected.

To prove Proposition 5.8, we will consider the last ear in an *ear decomposition* of $\alpha$. We prove a small variant of the classical ear decomposition (see [Rob39] or [BM08, §5.3]) which lets us exclude a specified vertex from the internal vertices of the last ear.

**Lemma 5.9.** *Let $\alpha \in \mathcal{E}_1$ be 2-edge-connected with at least 2 vertices. There exists a path $\pi = (u_1, \ldots, u_k)$ in $\alpha$ with $k \geq 2$ such that:*

1. *Each internal vertex $u_2, \ldots, u_{k-1}$ has degree 2 in $\alpha$.*
2. *Each internal vertex $u_2, \ldots, u_{k-1}$ satisfies $u_i \neq \mathrm{root}(\alpha)$.*
3. *$u_1, \ldots, u_k$ are pairwise distinct, except possibly $u_1 = u_k$.*
4. *Removing internal vertices and edges of $\pi$ from $\alpha$ leaves $\alpha$ 2-edge-connected.*

*Proof of Lemma 5.9.* Consider the following sequence $(\alpha_t)_{t \geq 0}$ of 2-edge connected subgraphs of $\alpha$:

1. Start from $\alpha_0$ being any cycle of $\alpha$ containing $\mathrm{root}(\alpha)$.
2. Let $t \geq 0$. If $\alpha_t$ spans all vertices of $\alpha$, then stop.
3. Otherwise, there exists $\{u_1, u_2\} \in E(\alpha)$ such that $u_1 \in V(\alpha_t)$ and $u_2 \notin V(\alpha_t)$. Since $\alpha$ is 2-edge-connected, there exists a simple path $(u_2, \ldots, u_k)$ in $\alpha \setminus \{\{u_1, u_2\}\}$ such that $u_i \notin V(\alpha_t)$ for all $2 \leq i \leq k-1$, and $u_k \in V(\alpha_t)$. Set

$$\alpha_{t+1} = (V(\alpha_t) \cup \{u_2, \ldots, u_{k-1}\}, E(\alpha_t) \cup \{\{u_i, u_{i+1}\} : 1 \leq i < k\}).$$

For any $t \geq 0$, $\alpha_t$ is 2-edge-connected. Therefore, if at the end of the algorithm $V(\alpha_t) = V(\alpha)$ but $E(\alpha_t) \neq E(\alpha)$, then any edge in $E(\alpha) \setminus E(\alpha_t)$ is a length-1 path that satisfies the conclusion of the lemma. Otherwise, this means that $\alpha$ is obtained from $\alpha_{t-1}$ (which is 2-edge-connected) by adding a path of internal degree-2 vertices in $\alpha$ which must all be distinct from $\mathrm{root}(\alpha) \in V(\alpha_0) \subseteq V(\alpha_{t-1})$. This concludes the proof. □

*Proof of Proposition 5.8.* Starting with the graph $\alpha$, consider the following procedure:

1. Delete all self-loops in $\alpha$.
2. If no leaf 2-vertex-connected component (i.e., a 2-vertex-connected component meeting the rest of the graph at a single articulation point) consists of a single cycle, then stop.
3. Otherwise, choose an arbitrary such component. Let $v$ be the articulation point connecting this component to the rest of graph. Delete all edges of this component from the graph.
4. Delete newly isolated vertices; exactly one vertex of the component remains, namely $v$. Since $\alpha \notin \mathcal{C}_1$, the procedure does not delete the entire graph.
5. If the root was removed in Step 4, set $v$ as the new root of the diagram.
6. Return to Step 1.

Call $\beta \in \mathcal{A}_1$ the resulting rooted graph. Note that $\beta$ is still 2-edge-connected, so by Lemma 5.9, we can find a path $\pi = (u_1, \ldots, u_k)$ in $\beta$ with internal degree-2 vertices. $\pi$ cannot be a cycle because of our initial step of removing cyclic 2-vertex-connected components. Therefore, $\pi$ is a simple path and the root of $\beta$ is not an internal vertex of $\pi$.

**Observation 5.10.** *For $2 \leq i < k$, let $\sigma_i$ be the connected component of $u_i$ in $\alpha \setminus E(\pi)$. Then $\alpha' := \pi \cup \sigma_2 \cup \ldots \cup \sigma_{k-1}$ is an open cactus in $\alpha$ with endpoints $u_1, u_k$. Moreover, $\mathrm{root}(\alpha)$ is not an internal vertex of the open cactus.*

*Proof.* $\pi$ is a simple path in $\beta$, and adding back loops and cyclic 2-vertex-connected components we removed from $\alpha$, we obtain an open cactus. The recursive pruning procedure we used to transfer the root ensures that $\mathrm{root}(\alpha)$ is not in any of the cyclic 2-vertex-connected components that are added to $\pi$. □

**Observation 5.11.** *$\alpha\,[V(\alpha) \setminus (V(\alpha') \setminus \{u_1, u_k\})]$ is 2-edge-connected.*

*Proof.* By Lemma 5.9, $\beta\,[V(\beta) \setminus \{u_2, \ldots, u_{k-1}\}]$ is 2-edge-connected. Adding 2-vertex-connected cyclic components to this graph preserves 2-edge-connectivity. □

Observation 5.10 and Observation 5.11 conclude the proof of Proposition 5.8. □

### 5.4 The effect of puncturing

The main result of this subsection is:

**Proposition 5.12.** *Let $\boldsymbol{H} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ such that $\|\boldsymbol{H}\| \le 1$ and $\boldsymbol{u} \in \mathbb{R}^n$ be a unit vector. Denote by $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top)$. Then for any open cactus $\alpha \in \mathcal{A}_2$,*

$$\|\boldsymbol{W}_\alpha(\boldsymbol{A}) - \boldsymbol{W}_\alpha(\boldsymbol{H})\|_{\mathrm{F}} \le |E(\alpha)| \cdot \|\boldsymbol{A} - \boldsymbol{H}\|_{\mathrm{F}} \le 3|E(\alpha)|\,.$$

We deduce in the following that puncturing does not change the diagonal distribution. In particular, matrices such as the ROM and the r-ROM have the same diagonal distribution.

**Corollary 5.13.** *Let $\boldsymbol{H}$ and $\boldsymbol{A}$ be as in Proposition 5.12. Then for any $\sigma \in \mathcal{C}_1$*

$$\|\boldsymbol{w}_\sigma(\boldsymbol{H}) - \boldsymbol{w}_\sigma(\boldsymbol{A})\|_2 \le O(1)\,,$$

*and for any $\sigma \in \mathcal{C}$,*

$$\frac{1}{n}|w_\sigma(\boldsymbol{H}) - w_\sigma(\boldsymbol{A})| \le O\left(\frac{1}{\sqrt{n}}\right)\,.$$

*Proof of Corollary 5.13 from Proposition 5.12.* Except for the case where $\sigma \in \mathcal{C}_1$ has one vertex (in which case the statement holds because the diagonal entries are bounded), $\mathrm{root}(\sigma)$ has degree $\ge 2$. Create two copies $r_1, r_2$ of $\mathrm{root}(\sigma)$ and re-assign the edges incident to $\mathrm{root}(\sigma)$ to $r_1$ or $r_2$ in such a way that $r_1$ and $r_2$ have degree at least 1. The resulting graph is an open cactus $\alpha$ with endpoints $r_1$ and $r_2$ such that merging these endpoints yields back $\sigma$. Hence,

$$\|\boldsymbol{w}_\sigma(\boldsymbol{H}) - \boldsymbol{w}_\sigma(\boldsymbol{A})\|_2 = \|\mathrm{diag}(\boldsymbol{W}_\alpha(\boldsymbol{H})) - \mathrm{diag}(\boldsymbol{W}_\alpha(\boldsymbol{A}))\|_{\mathrm{F}} \le O(1)\,.$$

The second statement then follows from Cauchy-Schwarz:

$$|w_\sigma(\boldsymbol{H}) - w_\sigma(\boldsymbol{A})| = |\langle \mathbf{1}, \boldsymbol{w}_\sigma(\boldsymbol{H}) - \boldsymbol{w}_\sigma(\boldsymbol{A})\rangle| \le \sqrt{n} \cdot \|\boldsymbol{w}_\sigma(\boldsymbol{H}) - \boldsymbol{w}_\sigma(\boldsymbol{A})\|_2 \le O(\sqrt{n})\,.$$

This concludes the proof. □

However, $\boldsymbol{H}$ and its punctured version $\boldsymbol{A}$ may *not* have the same traffic distribution, even on scalar open cactuses. Thus, the diagonal distribution (i.e., the values of cactus diagrams) is not sensitive to the behavior of $\boldsymbol{H}$ in any single direction $\boldsymbol{u}$, while some diagrams in the traffic distribution *are* sensitive to the behavior in the $\mathbf{1}$ direction.

**Example 5.14** (Puncturing of the Walsh-Hadamard matrix)**.** *Let $\boldsymbol{H}^{(n)}$ be the normalized Walsh-Hadamard matrices (Definition 2.3). Then for the 2-path diagram $\alpha$ (which is an open cactus),*

$$\frac{1}{n}(w_\alpha(\boldsymbol{H}) - w_\alpha(\boldsymbol{A})) = \frac{1}{n}\langle \mathbf{1}, (\boldsymbol{H}^2 - \boldsymbol{A}^2)\mathbf{1}\rangle \underset{n\to\infty}{\longrightarrow} 1\,.$$

*This does not contradict Proposition 5.12: $\boldsymbol{E} = \boldsymbol{W}_\alpha(\boldsymbol{H}) - \boldsymbol{W}_\alpha(\boldsymbol{A})$ indeed satisfies*

$$\sum_{i,j=1}^{n} \boldsymbol{E}[i,j]^2 \le O(1) \quad \text{and} \quad \left| \sum_{i,j=1}^{n} \boldsymbol{E}[i,j] \right| = \Omega(n)\,.$$

In general, as the following example demonstrates, the off-diagonal structure of the error matrix $\boldsymbol{E} = \boldsymbol{W}_\alpha(\boldsymbol{H}) - \boldsymbol{W}_\alpha(\boldsymbol{A})$ in Proposition 5.12 may be intricate. In the following example, $\boldsymbol{E}$ has entries of magnitude $\Omega(1)$, even though its Frobenius norm remains bounded.

**Example 5.15** (Puncturing of the DST matrix)**.** *Let $\boldsymbol{H}^{(n)}$ be the discrete sine transform matrices (Definition 2.4). Then for any fixed odd $i \ge 1$, the normalized sum of the $i$th row of $\boldsymbol{H}^{(n)}$ is*

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \boldsymbol{H}[i,j] = (\sqrt{2} + o(1)) \int_0^1 \sin(\pi i t)\,\mathrm{d}t \underset{n\to\infty}{\longrightarrow} \frac{2\sqrt{2}}{i\pi}\,.$$

*Consider the 2-path diagram $\alpha$. While the off-diagonal entries of $\boldsymbol{W}_\alpha(\boldsymbol{H}) = \boldsymbol{H}^2$ vanish (since $\boldsymbol{H}$ is a symmetric orthogonal matrix), on the other hand, for any fixed distinct odd numbers $i, j \ge 1$,*

$$\boldsymbol{W}_\alpha(\boldsymbol{A})[i,j] = (\boldsymbol{A}^2)[i,j] \underset{n\to\infty}{\longrightarrow} -\frac{8}{ij\pi^2}\,,$$

*which is $\Omega(1)$ for constant $i \ne j$.*

The proof of Proposition 5.12 relies on expanding $\boldsymbol{A}$ in terms of $\boldsymbol{u}\boldsymbol{u}^\top$ and $\boldsymbol{H}$. All rank-1 terms can be neglected thanks to the following lemma:

**Lemma 5.16.** *Let $\alpha$ be an open cactus, $e^* \in E(\alpha)$, and $\boldsymbol{\mathcal{A}} = (\boldsymbol{A}_e)_{e\in E(\alpha)}$ be a collection of matrices such that $\|\boldsymbol{A}_e\| \le 1$ for all $e \in E(\alpha) \setminus \{e^*\}$. Then,*

$$\|\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}})\|_{\mathrm{F}} \le \|\boldsymbol{A}_{e^*}\|_{\mathrm{F}}\,.$$

*Proof.* We first run a pruning procedure that iteratively removes parts of $\alpha$ not containing $e^*$, without decreasing the Frobenius norm of $\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}})$ during the procedure. To this end, we use repeatedly the standard inequalities:

**Claim 5.17.** $\|\boldsymbol{M}_1\boldsymbol{M}_2\|_{\mathrm{F}} \le \|\boldsymbol{M}_1\|_{\mathrm{F}}\|\boldsymbol{M}_2\|$.

**Claim 5.18.** $\|\boldsymbol{M}_1 \odot \boldsymbol{M}_2\|_{\mathrm{F}} \le \|\boldsymbol{M}_1\|_{\mathrm{F}} \cdot \max_{1\le i,j\le n} |\boldsymbol{M}_2[i,j]| \le \|\boldsymbol{M}_1\|_{\mathrm{F}}\|\boldsymbol{M}_2\|$, *where $\odot$ denotes entrywise or Hadamard product.*

Initially, let $u_{\mathrm{L}}$ be one of the endpoints of $\alpha$.

1. If $e^*$ belongs to a cactus hanging from $u_{\mathrm{L}}$, then stop.
2. Otherwise, remove the cactus hanging from $u_{\mathrm{L}}$ from the diagram. Using Claim 5.18 and Theorem 5.7 (the spectral norm of the cactus matrix diagram with a double root at $u_{\mathrm{L}}$ is at most 1), this does not decrease the Frobenius norm.
3. At this point, $u_{\mathrm{L}}$ must have degree equal to 1 in the current graph. If $e$ is the edge adjacent to $u_{\mathrm{L}}$, then stop.

4. Otherwise, remove the edge adjacent to $u_{\mathrm{L}}$. By Claim 5.17 and the assumption, this does not decrease the Frobenius norm. Set $u_{\mathrm{L}}$ to be the vertex that was adjacent to $u_{\mathrm{L}}$, and go back to the first step.

Then, apply the symmetric procedure from the other endpoint $u_{\mathrm{R}}$ of $\alpha$. At this point, there are two cases. If $u_{\mathrm{L}} \neq u_{\mathrm{R}}$, then the resulting graph must consist of the single edge $e^* = \{u_{\mathrm{L}}, u_{\mathrm{R}}\}$, so we get the desired upper bound on the Frobenius norm. Therefore, we assume from now on that $u_{\mathrm{L}} = u_{\mathrm{R}}$.

The resulting graph must be a cactus rooted at $u_{\mathrm{L}} = u_{\mathrm{R}}$, and $e^*$ is one of the edges of this cactus. If there are several cycles incident to $u_{\mathrm{L}}$, we use again Claim 5.18 and Theorem 5.7 to remove all such cycles not containing $e^*$ without decreasing the Frobenius norm.

Finally, we bound the Frobenius norm of the diagonal cactus matrix rooted at $u_{\mathrm{L}}$ by the Frobenius norm of an open cactus obtained by creating two copies of the root and turning the unique cycle hanging at $u_{\mathrm{L}}$ into a simple path between these two copies (we used a similar procedure in Corollary 5.13). We claim that this open cactus has strictly less edges than the one we started with before running the pruning procedure. Indeed, the base path had at least one edge, which was removed during the pruning stage when $u_{\mathrm{L}} = u_{\mathrm{R}}$ at the end. We conclude by induction on the number of edges of the open cactus. □

*Proof of Proposition 5.12.* We replace iteratively $\boldsymbol{H}$ by $\boldsymbol{A}$ in the graph polynomial $\boldsymbol{W}_\alpha(\boldsymbol{H})$: let $e_1, \dots, e_{|E(\alpha)|}$ be the edges of $\alpha$, and write

$$\boldsymbol{W}_\alpha(\boldsymbol{A}) - \boldsymbol{W}_\alpha(\boldsymbol{H}) = \sum_{i=1}^{|E(\alpha)|} \boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}}_i)\,,$$

where $\boldsymbol{\mathcal{A}}_i[e_j] = \boldsymbol{H}$ if $j < i$, $\boldsymbol{\mathcal{A}}_i[e_j] = \boldsymbol{A}$ if $j > i$, and $\boldsymbol{\mathcal{A}}_i[e_i] = \boldsymbol{A} - \boldsymbol{H}$. For each $i \in [|E(\alpha)|]$, we apply Lemma 5.16 with $e^* = e_i$. We have $\|\boldsymbol{A}\| \leq 1$ and $\|\boldsymbol{H}\| \leq 1$ so the assumptions of the lemma are satisfied, and we deduce

$$\|\boldsymbol{W}_\alpha(\boldsymbol{\mathcal{A}}_i)\|_{\mathrm{F}} \leq \|\boldsymbol{A} - \boldsymbol{H}\|_{\mathrm{F}}\,,$$

and by the triangle inequality

$$\|\boldsymbol{W}_\alpha(\boldsymbol{A}) - \boldsymbol{W}_\alpha(\boldsymbol{H})\|_{\mathrm{F}} \leq |E(\alpha)| \cdot \|\boldsymbol{A} - \boldsymbol{H}\|_{\mathrm{F}}\,.$$

Finally, we have

$$\boldsymbol{A} - \boldsymbol{H} = \langle \boldsymbol{u}, \boldsymbol{H}\boldsymbol{u}\rangle \boldsymbol{u}\boldsymbol{u}^\top - (\boldsymbol{H}\boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{H})\,.$$

Since $\|\boldsymbol{H}\| \leq 1$ and $\boldsymbol{u}$ is a unit vector, we have $|\langle \boldsymbol{u}, \boldsymbol{H}\boldsymbol{u}\rangle| \leq 1$ and $\|\boldsymbol{H}\boldsymbol{u}\|_2 \leq 1$, so $\|\boldsymbol{A} - \boldsymbol{H}\|_{\mathrm{F}} \leq 3$. □

### 5.5 Support of the $z$-basis

Let $\boldsymbol{H} = \boldsymbol{H}^{(n)}$ be a family of matrices satisfying Eqs. (12) and (13) and $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ be their puncturing. The main result of this subsection is that $\boldsymbol{A}$ and $\boldsymbol{H}$ satisfy the weak cactus property, that is, their traffic distribution in the $z$-basis is supported on cactuses and graphs with bridges.

**Proposition 5.19.** *For any $\alpha \in \mathcal{E} \setminus \mathcal{C}$,*

$$\frac{1}{n}|z_\alpha(\boldsymbol{H})| \leq O\left(\varepsilon + \frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \frac{1}{n}|z_\alpha(\boldsymbol{A})| \leq O\left(\varepsilon + \frac{1}{\sqrt{n}}\right)\,.$$

The fundamental theorem of graph polynomials can be used to show that these quantities are $O(1)$ (after converting between the $z$ and $w$-bases). The idea of Proposition 5.19 is to isolate an open cactus in $\alpha$ by Proposition 5.8 and apply Assumption 5.2 to gain an additional $\varepsilon$ factor.

We emphasize that analogous bounds in the $w$-basis are false in general; summation over some distinct indices is necessary to prove Proposition 5.19. We prove that, using the notation in Section 3.2:

**Lemma 5.20.** *Let $\alpha \in \mathcal{E}_1 \setminus \mathcal{C}_1$ and let $s, t$ be the endpoints of an open cactus in $\alpha$ satisfying the guarantees of Proposition 5.8. Then*

$$\frac{1}{\sqrt{n}}\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{A})\|_2 \le O\left(\varepsilon + \frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \frac{1}{\sqrt{n}}\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{H})\|_2 \le O\left(\varepsilon + \frac{1}{\sqrt{n}}\right). \tag{16}$$

The constraint $s \neq t$ in Eq. (16) ensures that we only use off-diagonal entries of the open cactuses in the graph polynomial. These are the only entries assumed to be small in Assumption 5.2 (and indeed, the diagonal entries of $\boldsymbol{W}_\alpha(\boldsymbol{H})$ can be large, for example, in the 2-path diagram).

*Proof of Proposition 5.19 from Lemma 5.20.* Let $\boldsymbol{M} \in \{\boldsymbol{A}, \boldsymbol{H}\}$ and $s, t$ be two distinct vertices of $\alpha$ to be fixed later. Using Möbius inversion (Lemma 3.9) recursively, we can expand

$$z_\alpha(\boldsymbol{M}) = c_\alpha w_\alpha^{s\neq t}(\boldsymbol{M}) + \sum_{\beta \prec \alpha} c_\beta z_\beta(\boldsymbol{M}),$$

for some constant coefficients $c_\beta \in \mathbb{R}$. Since all $\beta \prec \alpha$ are 2-edge-connected by Lemma 3.13 and have strictly less vertices than $\alpha$, by induction on the number of vertices of $\alpha$, it suffices to prove:

$$\frac{1}{n}|w_\alpha^{s\neq t}(\boldsymbol{M})| \le O\left(\varepsilon + \frac{1}{\sqrt{n}}\right). \tag{17}$$

But Eq. (17) follows from Lemma 5.20: pick $s, t$ to be the endpoints of an open cactus decomposition provided by Proposition 5.8, so that by Cauchy-Schwarz

$$\frac{1}{n}|w_\alpha^{s\neq t}(\boldsymbol{M})| = \frac{1}{n}\left|\langle \boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M}), \mathbf{1}\rangle\right| \le \frac{1}{\sqrt{n}}\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2 \le O\left(\varepsilon + \frac{1}{\sqrt{n}}\right),$$

which concludes the proof. □

We now move to the proof of Lemma 5.20. A useful concept will be the following graphical interpretation of squaring the polynomial expressed by a diagram:

**Definition 5.21** (Lift)**.** *Let $\alpha \in \mathcal{A}$ and $T \subseteq V(\alpha)$. Let $S_1$ and $S_2$ be two new disjoint sets of size $|V(\alpha)| - |T|$ (also disjoint from $V(\alpha)$). For $i \in \{1, 2\}$, let $p_i$ be a bijection between $V(\alpha) \setminus T$ and $S_i$, which is extended to $V(\alpha)$ by $p_i(u) = u$ for all $u \in T$.*

*The* lift *of $\alpha$ with respect to $T$ is the graph* $\mathrm{Lift}_T(\alpha)$ *with*

$$V(\mathrm{Lift}_T(\alpha)) = T \cup S_1 \cup S_2, \quad E(\mathrm{Lift}_T(\alpha)) = \{\{p_i(u), p_i(v)\} : i \in \{1, 2\}, \{u, v\} \in E(\alpha)\}.$$

**Claim 5.22.** *Let $\alpha \in \mathcal{A}_2$ with roots $(s, t)$, and $T \subseteq V(\alpha)$ be such that $\{s, t\} \subseteq T$. Then for any*

$\boldsymbol{M} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$,

$$\boldsymbol{W}_{\mathrm{Lift}_T(\alpha)}(\boldsymbol{M})[i,j] = \sum_{\substack{\varphi\colon T\to[n]\\ \varphi(s)=i,\varphi(t)=j}} \left( \sum_{\varphi\colon V(\alpha)\setminus T\to[n]} \prod_{\{u,v\}\in E(\alpha)} \boldsymbol{M}[\varphi(u),\varphi(v)] \right)^2 .$$

**Lemma 5.23.** *Let $\alpha \in \mathcal{E}$, let $\pi$ be a connected subgraph of $\alpha$, let $\alpha_1$ be any connected component of $\alpha \setminus E(\pi)$, and let $\alpha_2$ the graph spanned by $E(\alpha) \setminus E(\alpha_1)$. Then for all $j \in \{1,2\}$, $\mathrm{Lift}_{V(\alpha_1)\cap V(\alpha_2)}(\alpha_j)$ is 2-edge-connected.*

*Proof.* First, $\alpha_1$ is connected by definition. Since $\alpha$ is connected, every connected component in $(V(\alpha), E(\alpha) \setminus E(\pi))$ must be connected to $\pi$. Together with the fact that $\pi$ itself is connected, we get that $\alpha_2$ is connected. In particular, the lifts of $\alpha_1$ and $\alpha_2$ are connected.

Fix $j \in \{1,2\}$ and an edge $e'$ in the lift of $\alpha_j$. We need to show that $e'$ belongs to at least one simple cycle in the lift of $\alpha_j$. There exist $i \in \{1,2\}$ and $e = \{x,y\} \in E(\alpha_j)$ such that $e' = \{p_i(x), p_i(y)\}$ (where $p_1, p_2$ are the lift maps from Definition 5.21). Since $\alpha$ is 2-edge-connected, $e$ belongs to a simple cycle in $\alpha$. Consider the longest subpath of this cycle containing $e$ and consisting only of vertices in $V(\alpha_j)$. If this subpath is the entire cycle, then we have found a cycle containing $e$ in $\alpha_j$, and so a cycle containing $e'$ in its lift. Otherwise, the endpoints of this path must be in $V(\alpha_1) \cap V(\alpha_2)$. The images of this path through the lift maps $p_1$ and $p_2$ are disjoint, so their union forms a cycle in the lift of $\alpha_j$ containing $e'$. □

**Lemma 5.24.** *Let $\alpha \in \mathcal{E}_2$ have two distinct roots. Let $\beta$ be a leaf 2-vertex-connected component of $\alpha$ (i.e., removing internal vertices of $\beta$ leaves $\alpha$ connected) that does not contain the roots of $\alpha$. We view $\beta \in \mathcal{E}_1$ as a vector diagram rooted at the articulation point connecting $\beta$ to the rest of $\alpha$. For any distinct $s', t' \in V(\beta)$ and $\boldsymbol{M} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ such that $\|\boldsymbol{M}\| \le 1$,*

$$\sum_{i,j=1}^{n} \left| \boldsymbol{W}_\alpha^{s'\neq t'}(\boldsymbol{M})[i,j] \right| \le \sqrt{n} \cdot \|\boldsymbol{w}_\beta^{s'\neq t'}(\boldsymbol{M})\|_2 .$$

*Proof.* Let $(s,t)$ be the roots of $\alpha$. Since $\alpha$ is 2-edge-connected, there exist two edge-disjoint simple paths between $s$ and $t$. Let $\pi$ be one of them. Let $\alpha_1$ be the connected component of $s$ in $(V(\alpha), E(\alpha) \setminus E(\pi))$, and $\alpha_2$ be the graph spanned by $E(\alpha) \setminus E(\alpha_1)$ (including only the vertices incident with one of these edges). Finally, let $S = V(\alpha_1) \cap V(\alpha_2)$.

**Claim 5.25.** $\{s,t\} \subseteq S$.

*Proof.* On the one hand, $E(\pi) \subseteq E(\alpha_2)$ and $\{s,t\}$ are the endpoints of $\pi$, so $\{s,t\} \subseteq V(\alpha_2)$. On the other hand, $s \in V(\alpha_1)$ by definition, and there is an $s$–$t$ path in $\alpha \setminus E(\pi)$, so $t \in V(\alpha_1)$. □

**Claim 5.26.** *For any $\{u,v\} \in E(\alpha)$ with $u \in V(\alpha_1)$ and $v \in V(\alpha_2)$, we have $u \in S$ or $v \in S$.*

*Proof.* Suppose that $v \notin V(\alpha_1)$. Since $u \in V(\alpha_1)$, $u$ is connected to $s$ by edges of $E(\alpha) \setminus E(\pi)$, and since $v \notin V(\alpha_1)$, $v$ is not connected to $s$ by these edges. But, $\{u,v\} \in E(\alpha)$, so it must be that $\{u,v\} \in E(\pi)$. And, $E(\pi) \subseteq E(\alpha_2)$, so $u \in V(\alpha_2)$. □

As $\pi$ is a simple path and $\beta$ is connected to the rest of $\alpha$ at an articulation vertex, $\pi$ does not contain any edge of $\beta$, so it must be that either $E(\beta) \subseteq E(\alpha_1)$ or $E(\beta) \subseteq E(\alpha_2)$. Assume without loss of generality that this holds for $\alpha_1$ (the argument will be exactly symmetric for $\alpha_2$, as we will only use the fact that these subgraphs satisfy the conclusion of Lemma 5.23). In particular, we then have $s', t' \in V(\alpha_1)$.

We first use the triangle inequality to push the absolute value inside the sum over labelings of vertices in $S$:

$$\begin{aligned}
&\sum_{\varphi(s),\varphi(t)=1}^{n} \left| \boldsymbol{W}_{\alpha}^{s' \neq t'}(\boldsymbol{M})[\varphi(s), \varphi(t)] \right| \\
&\leq \sum_{\varphi\colon S \to [n]} \left| \sum_{\substack{\varphi\colon V(\alpha)\setminus S \to [n] \\ \varphi(s') \neq \varphi(t')}} \prod_{\{u,v\} \in E(\alpha)} \boldsymbol{M}[\varphi(u), \varphi(v)] \right| \qquad (18) \\
&= \sum_{\varphi\colon S \to [n]} \left| \prod_{j=1}^{2} \sum_{\substack{\varphi\colon V(\alpha_j)\setminus S \to [n] \\ \varphi(s') \neq \varphi(t') \text{ if } j=1}} \prod_{\{u,v\} \in E(\alpha_j)} \boldsymbol{M}[\varphi(u), \varphi(v)] \right| \\
&\leq \left[ \prod_{j=1}^{2} \sum_{\varphi\colon S \to [n]} \left( \sum_{\substack{\varphi\colon V(\alpha_j)\setminus S \to [n] \\ \varphi(s') \neq \varphi(t') \text{ if } j=1}} \prod_{\{u,v\} \in E(\alpha_j)} \boldsymbol{M}[\varphi(u), \varphi(v)] \right)^2 \right]^{\frac{1}{2}}, \qquad (19)
\end{aligned}$$

where we applied Cauchy-Schwarz in the second inequality. Note that Eq. (18) is well-defined by Claim 5.26.

By Lemma 5.23 and Claim 5.22, the term for $j = 2$ in Eq. (19) is a 2-edge-connected graph polynomial, so by Theorem 5.7 and the assumption $\|\boldsymbol{M}\| \leq 1$, this term is bounded by

$$\sum_{\varphi\colon S \to [n]} \left( \sum_{\varphi\colon V(\alpha_2)\setminus S \to [n]} \prod_{\{u,v\} \in E(\alpha_2)} \boldsymbol{M}[\varphi(u), \varphi(v)] \right)^2 \leq n\,.$$

We now switch to the term $j = 1$ in Eq. (19). This graph polynomial can be interpreted as

$$\sum_{\varphi\colon S \to [n]} \left( \sum_{\substack{\varphi\colon V(\alpha_1)\setminus S \to [n] \\ \varphi(s') \neq \varphi(t')}} \prod_{\{u,v\} \in E(\alpha_1)} \boldsymbol{M}[\varphi(u), \varphi(v)] \right)^2 = \langle \boldsymbol{w}_{\beta}^{s' \neq t'}(\boldsymbol{M}), \boldsymbol{W}_{\alpha'}(\boldsymbol{M}) \boldsymbol{w}_{\beta}^{s' \neq t'}(\boldsymbol{M}) \rangle\,,$$

where $\alpha'$ is the lift of $\alpha\,[V(\alpha) \setminus (V(\beta) \setminus \{r\})]$ with respect to $S$ (here $r$ denotes the root of $\beta$, the articulation vertex connecting $\beta$ to the rest of $\alpha$), and we add two roots in $\alpha'$ at the two copies of $r$ created during the lift operation.

Hence,

$$\sum_{\varphi\colon S\to[n]}\left(\sum_{\substack{\varphi\colon V(\alpha_1)\setminus S\to[n]\\ \varphi(s')\neq\varphi(t')}}\prod_{\{u,v\}\in E(\alpha_j)}\boldsymbol{M}[\varphi(u),\varphi(v)]\right)^2\leq\|\boldsymbol{W}_{\alpha'}(\boldsymbol{M})\|\cdot\|\boldsymbol{w}_\beta^{s'\neq t'}(\boldsymbol{M})\|_2^2\,.$$

Note that $\alpha'$ is 2-edge-connected by Lemma 5.23, so that $\|\boldsymbol{W}_{\alpha'}(\boldsymbol{M})\|\leq 1$ by Theorem 5.7. Putting everything together, we obtain

$$\sum_{i,j=1}^n\left|\boldsymbol{W}_\alpha^{s'\neq t'}(\boldsymbol{M})[i,j]\right|\leq\sqrt{n}\cdot\|\boldsymbol{w}_\beta^{s'\neq t'}(\boldsymbol{M})\|_2\,,$$

as desired. □

*Proof of Lemma 5.20.* Let $\boldsymbol{M}\in\{\boldsymbol{A},\boldsymbol{H}\}$. Consider $\beta\in\mathcal{A}_2$ defined by:

1. Start from the lift of $\alpha$ with respect to its root. Let $p_1$ and $p_2$ be the lift maps.
2. Delete the edges and internal vertices of the image under $p_1$ of the open cactus in $\alpha$.
3. Root the resulting graph at $p_1(s)$ and $p_1(t)$.

Recall that $s$ and $t$ are the endpoints of the "extra" open cactus in $\alpha$. Thus, $\beta$ is, in short, $\alpha$ grafted to its mirror image at the roots, with just *one* of the copies of that extra open cactus deleted except for its endpoints, and those endpoints made the roots of the matrix diagram $\beta$. See Figure 4 for an illustration of this and the rest of the proof.

Let $\sigma$ be the image of the open cactus in $\alpha$ under the lift map $p_2$, and let $s'$ and $t'$ be the images of the endpoints of this open cactus through the lift map $p_2$. Thus $s'$ and $t'$ are the mirror images of the vertices chosen to be the roots of $\beta$ above. We can then rewrite

$$\begin{aligned}
&\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2^2\\
&=\sum_{\substack{i,j=1\\ i\neq j}}^n\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})[i,j]\boldsymbol{W}_\sigma(\boldsymbol{M})[i,j] &(20)\\
&=\sum_{\substack{i,j=1\\ i\neq j}}^n\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})[i,j]\boldsymbol{W}_\sigma(\boldsymbol{H})[i,j]+\sum_{\substack{i,j=1\\ i\neq j}}^n\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})[i,j](\boldsymbol{W}_\sigma(\boldsymbol{M})-\boldsymbol{W}_\sigma(\boldsymbol{H}))[i,j]\\
&\leq\max_{1\leq i<j\leq n}|\boldsymbol{W}_\sigma(\boldsymbol{H})[i,j]|\sum_{i,j=1}^n\left|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})[i,j]\right|+\|\boldsymbol{W}_\sigma(\boldsymbol{M})-\boldsymbol{W}_\sigma(\boldsymbol{H})\|_{\mathrm{F}}\|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})\|_{\mathrm{F}}\,, &(21)
\end{aligned}$$

using Hölder on the first term and Cauchy-Schwarz on the second. We further bound the first term with Assumption 5.2 and Lemma 5.24:

$$\max_{1\leq i<j\leq n}|\boldsymbol{W}_\sigma(\boldsymbol{H})[i,j]|\cdot\sum_{i,j=1}^n\left|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})[i,j]\right|\leq\varepsilon\sqrt{n}\cdot\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{H})\|_2\,.$$

For the second term, observe that by Proposition 5.12, we know that the change due to puncturing is small in Frobenius norm, i.e., $\|\boldsymbol{W}_\sigma(\boldsymbol{M}) - \boldsymbol{W}_\sigma(\boldsymbol{H})\|_{\mathrm{F}} \le O(1)$ for $\boldsymbol{M} \in \{\boldsymbol{A}, \boldsymbol{H}\}$. Moreover, in the other factor, $\|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})\|_{\mathrm{F}}^2$ is nothing but the lift of $\beta$ with respect to $\{p_1(s), p_1(t)\}$. This lift can be interpreted as:

$$\|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})\|_{\mathrm{F}}^2 = \langle \boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M}), \boldsymbol{W}_{\beta'}(\boldsymbol{M}) \boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M}) \rangle \,, \tag{22}$$

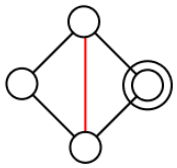
where $\beta'$ is the lift of $\alpha\,[V(\alpha) \setminus (V(\sigma) \setminus \{s,t\})]$ with respect to $\{s,t\}$. By the guarantees of Proposition 5.8, $\alpha\,[V(\alpha) \setminus (V(\sigma) \setminus \{s,t\})]$ is already 2-edge-connected, and therefore so is $\beta'$. As a result, by Theorem 5.7,

$$\begin{aligned}\|\boldsymbol{W}_\sigma(\boldsymbol{M}) - \boldsymbol{W}_\sigma(\boldsymbol{H})\|_{\mathrm{F}} \cdot \|\boldsymbol{W}_\beta^{s'\neq t'}(\boldsymbol{M})\|_{\mathrm{F}} &\le O(1) \cdot \|\boldsymbol{W}_{\beta'}(\boldsymbol{M})\|^{\frac{1}{2}} \|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2 \\ &\le O(1) \cdot \|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2 \,.\end{aligned}$$

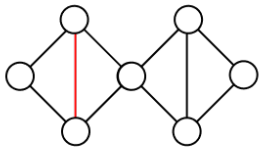
We obtain

$$\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2^2 \le O(1+\varepsilon\sqrt{n}) \cdot \|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{M})\|_2 \,,$$

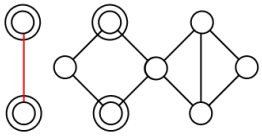
and the result follows after rearranging the inequality. □



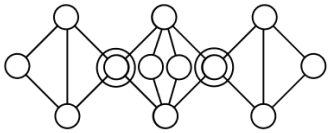
(a) Example of vector diagram $\alpha$ with extra open cactus in red



(b) Lift of $\alpha$ at the root (Eq. (20))



(c) Separate out open cactus (Eq. (21))



(d) Lift again. The left and right sides are copies of $\alpha$ while the inner part is 2-edge-connected (Eq. (22)).

Figure 4: Illustration of the main diagrammatic manipulations in the proof of Lemma 5.20.

### 5.6 Support of the $w$-basis

In this subsection we prove the second part of Theorem 5.3:

**Proposition 5.27.** *Suppose that $\boldsymbol{H}$ satisfies Eqs. (12) to (14). Then for any $\alpha \in \mathcal{A} \setminus \mathcal{E}$,*

$$\frac{1}{n}|w_\alpha(\boldsymbol{A})| \le \frac{1}{\sqrt{n}} \cdot (1+\varepsilon\sqrt{n})^{O(1)} \,.$$

These calculations are simpler than the previous ones, but this is also the point in the proof of Theorem 5.3 where puncturing is essential (note that it was not used at all in the previous section, and indeed those results apply equally well to the original $\boldsymbol{H}$ or the puncturing $\boldsymbol{A}$).

But, without puncturing, the values of graph polynomials that contain bridges can fail to be universal. For instance, when $\alpha$ is the degree-$d$ star, Walsh–Hadamard matrices $\boldsymbol{H} = \boldsymbol{H}^{(n)}$ satisfy $w_\alpha(\boldsymbol{H}) = \Theta(n^{d/2})$, so the limiting traffic distribution does not even exist when $d \geq 3$. As Proposition 5.27 shows, puncturing effectively forces all such diagrams to vanish in the traffic distribution.

To prove Proposition 5.27, we will isolate a bridge edge in the graph, and show by induction over the tree of 2-edge-connected components that:

**Lemma 5.28.** *For all $\alpha \in \mathcal{A}_1$,*

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 \leq (1+\varepsilon\sqrt{n})^{O(1)}\,.$$

*Proof of Proposition 5.27 from Lemma 5.28.* Decompose $\alpha = \alpha_1 \sqcup \alpha_2 \sqcup \{u,v\}$, where $\{u,v\} \in E(\alpha)$ is a bridge edge, $\alpha_1 \in \mathcal{E}_1$ is rooted at $u$, and $\alpha_2 \in \mathcal{A}_1$ is rooted at $v$. Then,

$$|w_\alpha(\boldsymbol{A})| = |\langle \boldsymbol{w}_{\alpha_1}(\boldsymbol{A}), \boldsymbol{A}\boldsymbol{w}_{\alpha_2}(\boldsymbol{A})\rangle| \leq \|\boldsymbol{w}_{\alpha_1}(\boldsymbol{A})\|_2 \|\boldsymbol{A}\boldsymbol{w}_{\alpha_2}(\boldsymbol{A})\|_2 \leq \sqrt{n}\cdot(1+\varepsilon\sqrt{n})^{O(1)}\,,$$

using Theorem 5.7 on the first term and Lemma 5.28 on the second. □

We prove Lemma 5.28 by first treating the cactus special case (Lemma 5.29), then the 2-edge-connected special case (Lemma 5.30), and then finally the general case by the induction mentioned above.

**Lemma 5.29.** *For any $\alpha \in \mathcal{C}_1$,*

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 \leq O(1+\varepsilon\sqrt{n})\,.$$

*Proof.* We first decompose $\boldsymbol{w}_\alpha(\boldsymbol{A})$ as:

$$\boldsymbol{w}_\alpha(\boldsymbol{A}) = (\boldsymbol{w}_\alpha(\boldsymbol{A}) - \boldsymbol{w}_\alpha(\boldsymbol{H})) + \boldsymbol{\Pi}\boldsymbol{w}_\alpha(\boldsymbol{H}) + \frac{1}{n}\langle \mathbf{1}, \boldsymbol{w}_\alpha(\boldsymbol{H})\rangle \mathbf{1}\,.$$

Since $\boldsymbol{A}\mathbf{1} = 0$ and $\|\boldsymbol{A}\| \leq 1$ by assumption, by the triangle inequality we have

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 \leq \|\boldsymbol{w}_\alpha(\boldsymbol{A}) - \boldsymbol{w}_\alpha(\boldsymbol{H})\|_2 + \|\boldsymbol{\Pi}\boldsymbol{w}_\alpha(\boldsymbol{H})\|_2\,.$$

By Corollary 5.13, the first term is $O(1)$, and by our assumption in Eq. (14), the second term is at most $\varepsilon\sqrt{n}$. □

**Lemma 5.30.** *For all $\alpha \in \mathcal{E}_1$,*

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 \leq O(1+\varepsilon\sqrt{n})\,.$$

*Proof.* We proceed by induction on $|V(\alpha)|$. For $\alpha \in \mathcal{C}_1$ (in particular, if $\alpha$ has only one vertex, which is our base case), the claim follows from Lemma 5.29. For $\alpha \in \mathcal{E}_1 \setminus \mathcal{C}_1$, we apply Proposition 5.8: there is an open cactus $\sigma$ induced in $\alpha$ such that removing the internal vertices and edges from $\sigma$

leaves $\alpha$ rooted and 2-edge-connected. Let $\{s,t\}$ be the endpoints of $\sigma$, and $\beta$ be the graph obtained from $\alpha$ by merging $s$ and $t$. Then, we can decompose:

$$\boldsymbol{w}_\alpha(\boldsymbol{A}) = \boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{A}) + \boldsymbol{w}_\beta(\boldsymbol{A})\,.$$

On the one hand, $\beta \in \mathcal{E}_1$ by Lemma 3.13 and has strictly less vertices than $\alpha$, so by induction

$$\|\boldsymbol{A}\boldsymbol{w}_\beta(\boldsymbol{A})\|_2 \leq O(1+\varepsilon\sqrt{n})\,.$$

On the other hand, by Lemma 5.20,

$$\|\boldsymbol{w}_\alpha^{s\neq t}(\boldsymbol{A})\|_2 \leq O(1+\varepsilon\sqrt{n})\,.$$

Putting everything together and using $\|\boldsymbol{A}\| \leq 1$ and the triangle inequality, we obtain

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 \leq O(1+\varepsilon\sqrt{n})\,,$$

which concludes the induction. □

**Lemma 5.31.** *Let $\alpha \in \mathcal{E}$, $v \in V(\alpha)$, and $S$ the set of edges adjacent to $v$ in $\alpha$. Then there exists $\beta \in \mathcal{E}$ and $v_1, v_2 \in V(\beta)$ such that*

$$\begin{aligned} V(\beta) &= (V(\alpha)\setminus\{v\}) \cup \{v_1, v_2\}\,,\\ E(\beta) &= (E(\alpha)\setminus S) \cup \phi(S) \cup \{\{v_1, v_2\}\}\,, \end{aligned}$$

*where $\phi(e) \in \{\{v_1,u\},\{v_2,u\}\}$ for all $e = \{v,u\} \in S$.*

*Proof.* We use the ear decomposition construction from the proof of Lemma 5.9. Consider the step of the ear decomposition which adds $v$. During this step, $v$ is a new interior vertex of a path or cycle added to $\alpha$. We define $\beta$ by splitting $v$ into two vertices $v_1, v_2$ with a new edge between them. When other ears attach to $v$ in $\alpha$, we can attach them to either $v_1$ or $v_2$ in $\beta$. This process yields an ear decomposition for $\beta$, hence $\beta$ is also 2-edge-connected. □

*Proof of Lemma 5.28.* We proceed by induction on the number of 2-edge-connected components in $\alpha$. If $\alpha$ is 2-edge-connected, then the result follows by Lemma 5.30. We assume from now on that $\alpha$ is not 2-edge-connected.

Let $C$ be the 2-edge-connected component of the root of $\alpha$. Let $\beta_1, \dots, \beta_k$ ($k \geq 1$) be the connected components disjoint from $\beta$ in the graph obtained after removing $E(C)$ and all bridges incident to $C$. We root $\beta_i$ at the (unique) vertex of $V(\beta_i)$ adjacent to $\beta$. We also consider $u_1 \in V(\alpha)$, the unique vertex in $V(C)$ that is adjacent to $V(\beta_1)$.

Let $\beta \in \mathcal{A}_2$ be the graph obtained from $\alpha$ by adding a second root at $u_1$, and deleting $V(\beta_1)$, $E(\beta_1)$, and the bridge between $u_1$ and $\beta_1$. Then, for $i = 2, \dots, k$, we iteratively apply the graph transformation from Lemma 5.31, label the new edge $e = \{v_1, v_2\}$ by $\boldsymbol{A}_e = \operatorname{diag}(\boldsymbol{A}\boldsymbol{w}_{\beta_i}(\boldsymbol{A}))$, and transfer the old labels for all other edges. In this way, we obtain a 2-edge-connected graph $\beta' \in \mathcal{E}_2$ and a family of matrices $\boldsymbol{\mathcal{A}} = (\boldsymbol{A}_e)_{e\in E(\beta')}$ such that

$$\boldsymbol{W}_{\beta'}(\boldsymbol{\mathcal{A}}) = \boldsymbol{W}_\beta(\boldsymbol{A})\,.$$

All involved matrices $\boldsymbol{A}_e \in \boldsymbol{\mathcal{A}}$ are either $\boldsymbol{A}$ or of the form $\mathrm{diag}(\boldsymbol{A}\boldsymbol{w}_{\beta_i}(\boldsymbol{A}))$ for some $i \in \{2, \dots, k\}$, so they satisfy $\|\boldsymbol{A}_e\| \le (1 + \varepsilon\sqrt{n})^{O(1)}$ by induction. Next, applying Theorem 5.7, we get

$$\|\boldsymbol{W}_\beta(\boldsymbol{A})\| = \|\boldsymbol{W}_{\beta'}(\boldsymbol{\mathcal{A}})\| \le (1 + \varepsilon\sqrt{n})^{O(1)}\,.$$

As a result,

$$\|\boldsymbol{A}\boldsymbol{w}_\alpha(\boldsymbol{A})\|_2 = \|\boldsymbol{A}\boldsymbol{W}_\beta(\boldsymbol{A})\boldsymbol{A}\boldsymbol{w}_{\beta_1}(\boldsymbol{A})\|_2 \le \|\boldsymbol{A}\| \cdot \|\boldsymbol{W}_\beta(\boldsymbol{A})\| \cdot \|\boldsymbol{A}\boldsymbol{w}_{\beta_1}(\boldsymbol{A})\|_2 \le (1 + \varepsilon\sqrt{n})^{O(1)}\,,$$

using again induction on $\|\boldsymbol{A}\boldsymbol{w}_{\beta_1}(\boldsymbol{A})\|_2$. This concludes the induction. □

### 5.7 Putting everything together: Proof of Theorem 5.3

*Proof of Theorem 5.3.* The first part follows from Proposition 5.19, and the second part follows from Proposition 5.27. For the third part, suppose that $\boldsymbol{H}$ satisfies Eqs. (12) to (14) with $\varepsilon^{(n)} = n^{-\frac{1}{2}+o(1)}$. Summarizing, we know that:

1. For all $\sigma \in \mathcal{C}$, $\frac{1}{n} w_\sigma(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} m_\sigma \in \mathbb{R}$ by assumption.
2. For all $\alpha \in \mathcal{E} \setminus \mathcal{C}$, $\frac{1}{n} z_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$ by the first part.
3. For all $\alpha \in \mathcal{A} \setminus \mathcal{E}$, $\frac{1}{n} w_\alpha(\boldsymbol{A}) \underset{n\to\infty}{\longrightarrow} 0$ by the second part.

By Lemma 3.14, the traffic distribution of $\boldsymbol{A}$ then exists and is uniquely determined by $\{m_\sigma : \sigma \in \mathcal{C}\}$, completing the proof. □

## 6 From Diagrams to Asymptotic GFOM Dynamics

The traffic distribution captures the limiting behavior of all scalar-valued, permutation-invariant polynomials. In this section, we show how to leverage this information to derive the limiting empirical laws of vector-valued, permutation-invariant polynomials. Our main application is a description of the limiting dynamics of GFOM.

We will mostly work under the assumption that the input matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfy the strong cactus property, which we recall is the statement that $\frac{1}{n}\mathbb{E}_{\boldsymbol{A}}\, z_\alpha(\boldsymbol{A}) \to 0$ as $n \to \infty$ for all non-cactus $\alpha$ (i.e., all $\alpha \in \mathcal{A} \setminus \mathcal{C}$, a statement about *scalar* graph polynomials). In Section 6.3.2 we will briefly suspend this assumption to discuss punctured matrices, so as to connect to the setting of Section 5.

We tackle two tasks in this section:

1. First, we study the joint asymptotic limit of the empirical distributions of the vector diagrams $\boldsymbol{z}_\alpha(\boldsymbol{A})$ over $\alpha \in \mathcal{A}_1$. Assuming the strong cactus property, we show that only the small subset of *treelike* $\alpha \in \mathcal{T}_1$ are asymptotically nonzero in the $z$-basis, in a sense to be made precise below. We then show that the asymptotic algebra of the treelike diagrams is isomorphic to a *Wick algebra*, an algebra defined by a family of Gaussian random variables. This will give a precise version of Theorem 1.12.

2. Second, we work with the asymptotic limit of treelike diagrams to identify a generalized Onsager correction, derive a treelike Approximate Message Passing algorithm, and prove its state evolution over arbitrary input matrices having the strong cactus property and a limiting diagonal distribution. This will give a precise version of Theorem 1.13.

### 6.1 Asymptotic limit of the vector diagrams

In this section, given a family $(X_i)_{i\in I}$ and $J \subseteq I$, we will write as a shortcut $X_J = (X_j)_{j\in J}$.

Recall that $\mathcal{C}_1$ denotes the set of rooted cactuses and $\mathcal{T}_1 \subseteq \mathcal{A}_1$ denotes the set of rooted trees with hanging cactuses. We call the diagrams in $\mathcal{T}_1$ *treelike*, and we call *Gaussian trees* the subset of diagrams $\mathcal{G}_1 \subseteq \mathcal{T}_1$ such that the root has degree exactly 1 after removing hanging cactuses.

**Definition 6.1** (Type)**.** *For each $\tau \in \mathcal{T}_1$, let $\mathrm{type}(\tau) \in \mathbb{N}^{\mathcal{G}_1 \cup \mathcal{C}_1}$, where $\mathrm{type}(\tau)_\alpha$ count the number of copies of $\alpha \in \mathcal{G}_1 \cup \mathcal{C}_1$ attached to the root of $\tau$, with the additional convention that $\mathrm{type}(\tau)_\alpha = 0$ for all $\alpha \in \mathcal{G}_1$ that has cactuses hanging at the root.*

The following theorem identifies the limiting distribution of $\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A})$ under the strong cactus property. We refer the reader to Appendix C for the definition of convergence in distribution for random elements indexed by countably infinite index sets.

**Theorem 6.2.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies Eq. (4), has the strong cactus property, and a limiting diagonal distribution. Then,*

$$\mathrm{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A})) \xrightarrow{(\mathrm{d})} Z^\infty_{\mathcal{A}_1}\,,$$

*where $Z^\infty_{\mathcal{A}_1} \in \mathbb{R}^{\mathcal{A}_1}$ is a random variable satisfying the following properties:*

1. *$Z^\infty_\alpha = 0$ for all non-treelike $\alpha$.*
2. *Conditioned on $Z^\infty_{\mathcal{C}_1}$, $Z^\infty_{\mathcal{G}_1}$ is a centered Gaussian process with covariance $\boldsymbol{\Sigma}^\infty$ from Eq. (31).*
3. *Let* He *denote the Wick product (Definition 2.9). Then for every $\tau \in \mathcal{T}_1$,*

$$Z^\infty_\tau = \mathrm{He}_{\mathrm{type}(\tau)}(Z^\infty_{\mathcal{G}_1}\,;\boldsymbol{\Sigma}^\infty)\cdot\prod_{\sigma\in\mathcal{C}_1}(Z^\infty_\sigma)^{\mathrm{type}(\tau)_\sigma}\,.$$

Theorem 6.2 shows how the limiting algebra $Z^\infty_{\mathcal{A}_1}$ of permutation-invariant, vector-valued polynomials in $\boldsymbol{A}$ can be derived from $Z^\infty_{\mathcal{C}_1}$. Although we have not specified the description of the law of $Z^\infty_{\mathcal{C}_1}$, it is fully determined by the limiting diagonal distribution of $\boldsymbol{A}$. For example, when $\boldsymbol{A}$ further satisfies the *factorizing* strong cactus property, $Z^\infty_{\mathcal{C}_1}$ is deterministic:

**Proposition 6.3.** *If $\boldsymbol{A}$ satisfies the factorizing strong cactus property and Eq. (4), then the conclusion of Theorem 6.2 holds with the additional property that for every $\sigma \in \mathcal{C}_1$,*

$$Z^\infty_\sigma = \prod_{\rho\in\mathrm{cyc}(\sigma)}\left(\lim_{n\to\infty}\frac{1}{n}\,\mathbb{E}\,z_\rho(\boldsymbol{A})\right)\,.$$

*Proof.* Let $\sigma \in \mathcal{C}_1$. The first moment of $\operatorname{samp}(\boldsymbol{z}_\sigma(\boldsymbol{A}))$ is

$$\mathbb{E}\operatorname{samp}(\boldsymbol{z}_\sigma(\boldsymbol{A})) = \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} \boldsymbol{z}_\sigma(\boldsymbol{A})[i] = \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_{\sigma_0}(\boldsymbol{A})\,, \tag{23}$$

where $\sigma_0$ is the unrooted version of $\sigma$. As $n \to \infty$, Eq. (23) converges to the deterministic constant

$$\kappa_{\sigma_0} := \lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_{\sigma_0}(\boldsymbol{A}) = \prod_{\rho \in \operatorname{cyc}(\sigma_0)} \left( \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\, z_\rho(\boldsymbol{A}) \right)$$

by the factorizing cactus property.

We now switch to the second moment,

$$\mathbb{E}\operatorname{samp}(\boldsymbol{z}_\sigma(\boldsymbol{A}))^2 = \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} \sum_{i=1}^{n} \boldsymbol{z}_\sigma(\boldsymbol{A})[i]^2\,.$$

Expand the scalar polynomial $\sum_{i=1}^n \boldsymbol{z}_\sigma(\boldsymbol{A})[i]^2$ in the $z$-basis. The support of that expansion is the set of diagrams that can be obtained by grafting two copies of $\sigma$ at the root and merging pairs of vertices across the two different copies. By the strong cactus property, it suffices to find which cactuses can be obtained in this way. By Lemma D.1, the only cactus that can occur in this way has no merging, and it contributes $\kappa_{\sigma_0}^2$ by the factorizing cactus property. Thus,

$$\mathbb{E}\operatorname{samp}(z_\sigma(\boldsymbol{A}))^2 = \kappa_{\sigma_0}^2 + o(1) = (\mathbb{E}\operatorname{samp}(z_\sigma(\boldsymbol{A}))^2 + o(1)\,.$$

We showed that $\operatorname{samp}(z_\sigma(\boldsymbol{A}))$ converges to the desired deterministic quantity in expectation, and furthermore that its variance converges to 0. This implies that it converges to the constant in distribution. By unicity of the limit in distribution, $Z_\sigma^\infty$ equals that constant almost surely. □

However, if we drop the factorizing cactus property assumption, the variables $Z_{\mathcal{C}_1}^\infty$ may no longer be deterministic. For example, this can be the case when $\boldsymbol{A}$ is a block-structured matrix as in Section 4.3:

**Example 6.4.** *Let $\boldsymbol{A}_1^{(n)}$ and $\boldsymbol{A}_2^{(n)}$ be two $n \times n$ matrices satisfying the assumptions of Theorem 6.2. Define the $2n \times 2n$ matrix,*

$$\boldsymbol{A}^{(2n)} = \begin{bmatrix} \boldsymbol{A}_1^{(n)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_2^{(n)} \end{bmatrix}$$

*From the block-diagonal structure, for any $\alpha \in \mathcal{A}_1$,*

$$\boldsymbol{z}_\alpha(\boldsymbol{A})[i] = \begin{cases} \boldsymbol{z}_\alpha(\boldsymbol{A}_1)[i] & \text{if } i \in [n] \\ \boldsymbol{z}_\alpha(\boldsymbol{A}_2)[i-n] & \text{if } i \in [2n] \setminus [n] \end{cases}$$

*Hence, the law of $Z_{\mathcal{C}_1}^\infty(\boldsymbol{A})$ is a uniform mixture of the law of $Z_{\mathcal{C}_1}^\infty(\boldsymbol{A}_1)$ and that of $Z_{\mathcal{C}_1}^\infty(\boldsymbol{A}_2)$.*

We will prove a generalization of Example 6.4 later; see Lemma 6.31.

In Example 6.4, the randomness of $Z_{\mathcal{C}_1}^\infty$ may be viewed as coming solely from the $\operatorname{samp}(\cdot)$ operator, but this is not always the case. For instance, our model also captures orthogonally invariant distributions that do not satisfy the traffic concentration property:

**Example 6.5.** *Let $(\lambda_n)_{n\geq 1}$ be an exchangeable sequence of random variables in $[-1,1]$ and consider*

$$\boldsymbol{A}^{(n)} = (\boldsymbol{Q}^{(n)})^\top \operatorname{diag}(\lambda_1, \dots, \lambda_n)\boldsymbol{Q}^{(n)}\,,$$

*for Haar-distributed matrices $\boldsymbol{Q}^{(n)} \in O(n)$, independent from $(\lambda_n)$. By de Finetti's theorem, there exists a latent random probability measure $\mu$ almost surely supported on $[-1,1]$ such that conditionally on $\mu$, $\lambda_1, \lambda_2, \dots$ are i.i.d. with common law $\mu$. By Theorem 4.2, $\boldsymbol{A}^{(n)}$ satisfies the strong cactus property conditionally on $\mu$, so it also satisfies the strong cactus property unconditionally.*

*Applying Theorem 6.2 and Proposition 6.3, we get that conditionally on $\mu$, $\operatorname{samp}(z_{\mathcal{C}_1}(\boldsymbol{A}))$ converges in distribution to*

$$\left(\prod_{\rho\in\operatorname{cyc}(\sigma)} \kappa_{|\rho|}(\mu)\right)_{\sigma\in\mathcal{C}_1}, \tag{24}$$

*where $(\kappa_q(\mu))_{q\geq 1}$ are the free cumulants of $\mu$. Therefore, unconditionally, $\operatorname{samp}(z_{\mathcal{C}_1}(\boldsymbol{A}))$ converges in distribution to the* random *quantity Eq. (24).*

Note that Examples 6.4 and 6.5 do not contradict Proposition 6.3 because in these examples, $\boldsymbol{A}^{(n)}$ typically does not satisfy the factorizing cactus property.

### 6.1.1 Non-treelike diagrams are asymptotically negligible

The remainder of Section 6.1 is dedicated to the proof of Theorem 6.2. In the whole proof, we drop the dependence of $z_\alpha$ and $w_\alpha$ on $\boldsymbol{A}$ to lighten notation. We start by proving that non-treelike diagrams are negligible.

**Lemma 6.6.** *Suppose that $\boldsymbol{A}$ satisfies the strong cactus property. Then for each non-treelike $\alpha$,*

$$\operatorname{samp}(\boldsymbol{z}_\alpha) \xrightarrow{L^2} 0\,.$$

*Proof.* By definition, we have

$$\mathbb{E}\operatorname{samp}(\boldsymbol{z}_\alpha)^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left(\boldsymbol{z}_\alpha^2\right)[i]\right]$$

By Lemma D.3, we can expand $\boldsymbol{z}_\alpha^2$ in the $\boldsymbol{z}$-basis to obtain, for some constant coefficients $c_\beta$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^n \sum_{\beta\in\mathcal{A}_1\setminus\mathcal{T}_1} c_\beta\, \mathbb{E}\,\boldsymbol{z}_\beta[i] \\ &= \frac{1}{n}\sum_{\beta\in\mathcal{A}_0\setminus\mathcal{T}_0} c'_\beta\, \mathbb{E}\, z_\beta \end{aligned}$$

for some other constant coefficients $c'_\beta$. Since no diagram in $\mathcal{A}_0 \setminus \mathcal{T}_0$ is a cactus, by the strong cactus property, we get $\mathbb{E}\operatorname{samp}(\boldsymbol{z}_\alpha)^2 \underset{n\to\infty}{\longrightarrow} 0$, as desired. □

### 6.1.2 Asymptotic limit of the treelike diagrams

Next, we analyze the treelike diagrams. All results in Section 6.1.2 are purely combinatorial, meaning that they hold for arbitrary $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$.

The covariance of treelike diagrams is defined in terms of *homeomorphic matchings* between them. We start by defining this new concept.

**Definition 6.7** (Core)**.** *Let $\tau \in \mathcal{T}_1$. Define* $\mathrm{core}(\tau)$ *to be the rooted tree obtained from $\tau$ by*

1. *Removing all hanging cactuses.*
2. *Removing all non-root degree-2 vertices and the two edges they are incident with, and adding back a new edge between their two neighbors.*

Note that the vertex set $V(\mathrm{core}(\tau))$ may be identified with a subset of $V(\tau)$, even though the second rule may lead to edges being present in $\mathrm{core}(\tau)$ that do not exist in $\tau$.

**Definition 6.8** (Homeomorphic matchings)**.** *Let $\tau_1, \tau_2 \in \mathcal{T}_1$. We say that a partial matching $P \subseteq V(\tau_1) \times V(\tau_2)$ of $\tau_1$ and $\tau_2$ is* homeomorphic *if*

1. $(\mathrm{root}(\tau_1), \mathrm{root}(\tau_2)) \in P$.
2. *Restricted to $V(\mathrm{core}(\tau_1)) \times V(\mathrm{core}(\tau_2))$, $P$ is a rooted graph isomorphism between* $\mathrm{core}(\tau_1)$ *and* $\mathrm{core}(\tau_2)$.
3. *Let $\{u, u'\} \in E(\mathrm{core}(\tau_1))$, let $(u = u_1, \dots, u_k = u')$ be the path between $u$ and $u'$ in $\tau_1$. Let $v = P(u)$, $v' = P(u')$, and $(v = v_1, \dots, v_\ell = v')$ be the path between $v$ and $v'$ in $\tau_2$. Then there is no matching edge between $\{u_1, \dots, u_k, v_1, \dots, v_\ell\}$ and its complement. Moreover, for all $(u_i, v_j) \in P$ and $(u_{i'}, v_{j'}) \in P$, we have $i \le i' \iff j \le j'$ (the matching restricted to the vertices in the paths is non-crossing).*
4. *No inner vertices from the hanging cactuses are matched.*

*We denote by $H(\tau_1, \tau_2)$ the set of homeomorphic matchings between $\tau_1$ and $\tau_2$.*

This definition is motivated by the following lemma stating that, when computing the covariance of two treelike diagrams, the matchings giving rise to cactuses are precisely the homeomorphic ones.

**Lemma 6.9.** *Let $\tau_1, \tau_2 \in \mathcal{T}_1$ and $\tau = \tau_1 \sqcup \tau_2$. For any matching $P \subseteq V(\tau_1) \times V(\tau_2)$ such that $(\mathrm{root}(\tau_1), \mathrm{root}(\tau_2)) \in P$, we have $\tau_P \in \mathcal{C}_1$ if and only if $P \in H(\tau_1, \tau_2)$.*

In particular, if $\tau_1, \tau_2 \in \mathcal{C}_1$, only the matching $P = \{(\mathrm{root}(\tau_1), \mathrm{root}(\tau_2))\}$ creates a cactus $\tau_P$. We are now ready to describe the algebra of treelike diagrams:

**Lemma 6.10.** *For all $\gamma_1, \dots, \gamma_\ell \in \mathcal{G}_1 \cup \mathcal{C}_1$,*

$$\prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} - \sum_{M \in \mathcal{M}(\ell)} \sum_{\substack{P_{uv} \in H(\gamma_u, \gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \bigoplus_{u \notin M} \gamma_u} \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}), \tag{25}$$

*where $\oplus$ denotes the grafting at the root.*

The proofs of Lemmas 6.9 and 6.10 are deferred to Appendix D.1. Note that the error in Lemma 6.10 is measured in terms of non-treelike diagrams.

By inverting Eq. (25), we can formulate the algebra of treelike diagrams in the language of *Wick products* (Definition 2.9).

**Corollary 6.11.** *For all* $\tau \in \mathcal{T}_1$,

$$\boldsymbol{z}_\tau - \mathrm{He}_{\mathrm{type}(\tau)}(\boldsymbol{z}_{\mathcal{G}_1}; \boldsymbol{\Sigma}) \prod_{\sigma \in \mathcal{C}_1} (Z_\sigma^\infty)^{\mathrm{type}(\tau)_\sigma} \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}), \tag{26}$$

*where for all* $\gamma, \gamma' \in \mathcal{G}_1$, *we defined the "finite-n" covariance matrix*

$$\boldsymbol{\Sigma}[\gamma, \gamma'] := \sum_{P \in H(\gamma, \gamma')} \boldsymbol{z}_{\gamma_P}. \tag{27}$$

*Proof.* We proceed by induction on the number of vertices of $\tau$. First, Eq. (26) trivially holds if $\tau$ has one vertex, which proves the base case. Now, suppose that $\tau = \gamma_1 \oplus \ldots \oplus \gamma_\ell$ is the grafting at the root of $\gamma_1, \ldots, \gamma_\ell \in \mathcal{G}_1 \cup \mathcal{C}_1$. By Lemma 6.10,

$$\boldsymbol{z}_\tau + \sum_{\substack{M \in \mathcal{M}(\ell) \\ M \neq \varnothing}} \sum_{\substack{P_{uv} \in H(\gamma_u, \gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \bigoplus_{u \notin M} \gamma_u} - \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}).$$

Applying the induction hypothesis and using additivity of types, we have:

$$\boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \bigoplus_{u \notin M} \gamma_u} - \prod_{uv \in M} \boldsymbol{z}_{\gamma_{P_{uv}}} \prod_{\substack{u \notin M \\ \gamma_u \in \mathcal{C}_1}} \boldsymbol{z}_{\gamma_u} \cdot \mathrm{He}_{\sum_{\substack{u \notin M \\ \gamma_u \in \mathcal{G}_1}} \mathrm{type}(\gamma_u)}(\boldsymbol{z}_{\mathcal{G}_1}; \boldsymbol{\Sigma}) \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}).$$

Since cactuses are not matched by homeomorphic matchings by definition, the product over cactuses $\boldsymbol{z}_{\gamma_u}$ is over all $u$ such that $\gamma_u \in \mathcal{C}_1$, which is independent of $M$ and can be factorized out. Therefore, in the rest of the proof we assume that $\gamma_i \in \mathcal{G}_1$ for all $i \in [\ell]$. Using Claim D.4, we obtain

$$z_\tau + \sum_{\substack{M \in \mathcal{M}(\ell) \\ M \neq \varnothing}} \prod_{uv \in M} \underbrace{\sum_{P \in H(\gamma_u, \gamma_v)} \boldsymbol{z}_{\gamma_P}}_{\boldsymbol{\Sigma}[\gamma_u, \gamma_v]} \cdot \mathrm{He}_{\sum_{u \notin M} \mathrm{type}(\gamma_u)}(\boldsymbol{z}_{\mathcal{G}_1}; \boldsymbol{\Sigma}) - \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}). \tag{28}$$

By the recursive formula of the Wick products (Corollary 2.11),

$$\sum_{\substack{M \in \mathcal{M}(\ell) \\ M \neq \varnothing}} \prod_{uv \in M} \boldsymbol{\Sigma}[\gamma_u, \gamma_v] \cdot \mathrm{He}_{\sum_{u \notin M} \mathrm{type}(\gamma_j)}(\boldsymbol{z}_{\mathcal{G}_1}; \boldsymbol{\Sigma}) + \mathrm{He}_{\mathrm{type}(\tau)}(\boldsymbol{z}_{\mathcal{G}_1}; \boldsymbol{\Sigma}) = \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j}. \tag{29}$$

Combining Eqs. (28) and (29) concludes the proof. □

Finally, if we reduce Lemma 6.10 modulo the larger class of non-cactus diagrams (which are the negligible diagrams *in expectation* under the strong cactus property), we deduce that the joint moments of the diagrams in $\mathcal{G}_1$ have an asymptotically Gaussian structure.

**Corollary 6.12.** *For all $\gamma_1, \dots, \gamma_\ell \in \mathcal{G}_1$ and $\sigma_1, \dots, \sigma_k \in \mathcal{C}_1$,*

$$\prod_{i=1}^{k} \boldsymbol{z}_{\sigma_i} \left[ \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} - \sum_{M \in \mathcal{M}_{\mathrm{perf}}(\ell)} \prod_{xy \in M} \boldsymbol{\Sigma}[\gamma_x, \gamma_y] \right] \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{C}_1}),$$

*where $\boldsymbol{\Sigma}$ is defined in Eq. (27).*

*Proof.* Every non-treelike term in Eq. (25) is a fortiori not a cactus. Also, the only cactuses in the subtracted term occur when $M$ is a perfect matching. In other words,

$$\prod_{i=1}^{k} \boldsymbol{z}_{\sigma_i} \left[ \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} - \sum_{M \in \mathcal{M}_{\mathrm{perf}}(\ell)} \sum_{\substack{P_{uv} \in H(\gamma_u, \gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{uv}}} \right] \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{C}_1}).$$

Therefore, by Lemma D.3, we deduce

$$\boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{uv}}} - \prod_{uv \in M} \boldsymbol{z}_{\gamma_{P_{uv}}} \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{C}_1}),$$

and the desired statement follows. □

### 6.1.3 Proof of Theorem 6.2

**Claim 6.13.** *Suppose that the traffic distribution of $\boldsymbol{A}$ exists. Then, for any $\alpha_1, \dots, \alpha_k \in \mathcal{A}_1$, the sequence $\mathbb{E} \operatorname{samp}(\boldsymbol{z}_{\alpha_1} \cdots \boldsymbol{z}_{\alpha_k})$ converges as $n \to \infty$.*

*Proof.* This is straightforward, as

$$\mathbb{E} \operatorname{samp}(\boldsymbol{z}_{\alpha_1} \cdots \boldsymbol{z}_{\alpha_k}) = \frac{1}{n} \mathbb{E} \sum_{i=1}^{n} \boldsymbol{z}_{\alpha_1}[i] \cdots \boldsymbol{z}_{\alpha_k}[i],$$

and the inner polynomial is a scalar polynomial of $\boldsymbol{A}$ that can be expanded in the $z$-basis of scalar diagrams as a linear combination of various quotients of the scalar diagram formed by forgetting the identity of the root in $\alpha_1 \oplus \cdots \oplus \alpha_k$. □

Claim 6.13 implies in particular that the sequence $\operatorname{samp}(\boldsymbol{z}_{\mathcal{A}_1})$ is tight. In the rest of the proof, we show that the limit in distribution actually exists and characterize it. The following lemma is a direct consequence of the fundamental theorem of graph polynomials.

**Lemma 6.14.** *If $\|\boldsymbol{A}\| \le O(1)$, then for each $\alpha \in \mathcal{E}_1$, there exists $C_\alpha > 0$ such that $|\operatorname{samp}(\boldsymbol{z}_\alpha)| \le C_\alpha$.*

*Proof.* By Lemma 3.9 and Lemma 3.13, we can expand for some coefficients $c_\beta = c_\beta(\alpha) \in \mathbb{R}$,

$$\boldsymbol{z}_\alpha = \sum_{\beta \in \mathcal{E}_1} c_\beta \boldsymbol{w}_\beta .$$

By Theorem 5.7, it holds for every $\beta \in \mathcal{E}_1$ that $\|\boldsymbol{w}_\beta\|_\infty \le \|\boldsymbol{A}\|^{|E(\beta)|}$, which is at most $O_\alpha(1)$ by assumption. The lemma follows by the triangle inequality. □

**Lemma 6.15.** *Suppose that the traffic distribution of $\boldsymbol{A}$ exists and that Eq. (4) holds. Then,* $\operatorname{samp}(\boldsymbol{z}_{\mathcal{C}_1})$ *converges in distribution to some stochastic process* $Z^{\infty}_{\mathcal{C}_1}$.

*Proof.* First, assume that $\sup_{n\geq 1} \|\boldsymbol{A}^{(n)}\| \leq K$ holds almost surely, for some universal constant $K > 0$. All the moments of $\operatorname{samp}(\boldsymbol{z}_{\mathcal{C}_1})$ converge by Claim 6.13. Since cactuses are 2-edge-connected, by Lemma 6.14, all random variables $\operatorname{samp}(\boldsymbol{z}_\alpha)$ for $\alpha \in \mathcal{C}_1$ are uniformly bounded in $n$. Hence, the moments satisfy the growth condition Eq. (62), so that $\operatorname{samp}(\boldsymbol{z}_{\mathcal{C}_1})$ converges in distribution by Theorem C.2. Finally, if we assume Eq. (4) rather than uniform boundedness, the result can be deduced from the latter case using Lemma C.3. □

*Proof of Theorem 6.2.* In the rest of the proof, we assume that the assumptions of Theorem 6.2 are satisfied. We start by analyzing convergence of the subtracted term from Corollary 6.12. By convergence in distribution of the cactuses (Lemma 6.15) and the continuous mapping theorem, we have for any $\gamma_1, \ldots, \gamma_\ell \in \mathcal{G}_1$ and $\sigma_1, \ldots, \sigma_k \in \mathcal{C}_1$,

$$\operatorname{samp}\left(\prod_{i=1}^{k} \boldsymbol{z}_{\sigma_i} \sum_{M\in\mathcal{M}_{\mathrm{perf}}(\ell)} \prod_{xy\in M} \boldsymbol{\Sigma}[\gamma_x, \gamma_y]\right) \xrightarrow{(\mathrm{d})} \prod_{i=1}^{k} Z^{\infty}_{\sigma_i} \sum_{M\in\mathcal{M}_{\mathrm{perf}}(k)} \prod_{xy\in M} \boldsymbol{\Sigma}^{\infty}[\gamma_x, \gamma_y]\,, \tag{30}$$

where we defined, for any $\gamma_1, \gamma_2 \in \mathcal{G}_1$, the "limiting" covariance matrix

$$\boldsymbol{\Sigma}^{\infty}[\gamma_1, \gamma_2] := \sum_{P\in H(\gamma_1,\gamma_2)} Z^{\infty}_{\gamma_P}\,. \tag{31}$$

Since all joint moments converge by Claim 6.13, the sequence of random variables on the left-hand side of Eq. (30) is uniformly integrable. So we also get convergence of the mean,

$$\mathbb{E}\operatorname{samp}\left(\prod_{i=1}^{k} \boldsymbol{z}_{\sigma_i} \sum_{M\in\mathcal{M}_{\mathrm{perf}}(\ell)} \prod_{xy\in M} \boldsymbol{\Sigma}[\gamma_x, \gamma_y]\right) \xrightarrow[n\to\infty]{} \mathbb{E}\left[\prod_{i=1}^{k} Z^{\infty}_{\sigma_i} \sum_{M\in\mathcal{M}_{\mathrm{perf}}(k)} \prod_{xy\in M} \boldsymbol{\Sigma}^{\infty}[\gamma_x, \gamma_y]\right].$$

Combining with Corollary 6.12 and the strong cactus property,

$$\mathbb{E}\operatorname{samp}\left(\prod_{i=1}^{k} \boldsymbol{z}_{\sigma_i} \prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j}\right) \xrightarrow[n\to\infty]{} \mathbb{E}\left[\prod_{i=1}^{k} Z^{\infty}_{\sigma_i} \sum_{M\in\mathcal{M}_{\mathrm{perf}}(\ell)} \prod_{xy\in M} \boldsymbol{\Sigma}^{\infty}[\gamma_x, \gamma_y]\right]. \tag{32}$$

The right-hand side of Eq. (32) coincides with the moments of $Z^{\infty}_{\mathcal{G}_1\cup\mathcal{C}_1}$. Recall that the law of $Z^{\infty}_{\mathcal{G}_1}$ satisfies that after sampling $Z^{\infty}_{\mathcal{C}_1}$ from its marginal (which is bounded almost surely by Lemma 6.14), then $Z^{\infty}_{\mathcal{G}_1}$ conditioned on $Z^{\infty}_{\mathcal{C}_1}$ is a Gaussian process with covariance kernel given by Eq. (31). This object satisfies the moment growth condition Eq. (62). So Theorem C.2 applies and we obtain convergence in distribution of $\operatorname{samp}(\boldsymbol{z}_{\mathcal{G}_1\cup\mathcal{C}_1})$ to $Z^{\infty}_{\mathcal{G}_1\cup\mathcal{C}_1}$.

By Lemma 6.6, the non-treelike diagrams converge in $L^2$ to 0, so by Slutsky's lemma, we obtain joint convergence in distribution, except for the remaining treelike, non-Gaussian trees. By Corollary 6.11, these are continuous images of cactuses and non-treelike diagrams, so by the continuous mapping theorem, all diagrams converge jointly in distribution to $Z^{\infty}_{\mathcal{A}_1}$. □

## 6.2 The treelike AMP algorithm

Now we turn to studying the dynamics of GFOM operations.

**Definition 6.16** (Asymptotic state)**.** *Let $(\boldsymbol{x}_i)_{i\in\mathcal{I}}$ be a family of random vectors, $\boldsymbol{x}_i \in \mathbb{R}^n$. We say that a stochastic process $(X_i)_{i\in\mathcal{I}}$ is the asymptotic state of $(\boldsymbol{x}_i)_{i\in\mathcal{I}}$ if, for any $k \geq 1$, $i_1, \dots, i_k \in \mathcal{I}$, and any bounded continuous or polynomial function $\varphi : \mathbb{R}^k \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\, \varphi(\boldsymbol{x}_{i_1}[j], \dots, \boldsymbol{x}_{i_k}[j]) = \mathbb{E}\, \varphi(X_{i_1}, \dots, X_{i_k})\,. \tag{33}$$

Definition 6.16 requires in particular for $(\boldsymbol{x}_i)_{i\in\mathcal{I}}$ to converge in distribution to $(X_i)_{i\in\mathcal{I}}$. As with convergence in distribution in general, this suffers from the caveat that the law of the limit in distribution of $(X_i)_{i\in\mathcal{I}}$ is unique, but the probability space on which the limit $(X_i)_{i\in\mathcal{I}}$ is realized is not. Thus when we speak of "the asymptotic state" we refer to a specific law, not a specific collection of random variables. Nonetheless, the sampling procedure in Theorem 6.2 suggests a natural way to sample an asymptotic state of the iterates of a pGFOM, since, provided we know how to sample from $Z^\infty_{\mathcal{C}_1}$ (which we must address on a case-by-case basis), the other $Z^\infty_\alpha$ are conditionally Gaussian or deterministic functions thereof.

Translating the limiting variables $Z^\infty_\alpha$ from Theorem 6.2 to a construction of an asymptotic state, we find:

**Lemma 6.17.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies the assumptions of Theorem 6.2. Let*

$$\boldsymbol{x} = \sum_{\alpha\in\mathcal{A}_1} c_\alpha \boldsymbol{z}_\alpha(\boldsymbol{A}) \tag{34}$$

*for some finitely supported coefficients $(c_\alpha)_{\alpha\in\mathcal{A}_1}$. Then,*

$$X := \sum_{\alpha\in\mathcal{A}_1} c_\alpha Z^\infty_\alpha \tag{35}$$

*is the asymptotic state of $\boldsymbol{x}$. Moreover, if $\boldsymbol{x}_t$ is of the form Eq. (34) for any $t \geq 1$ and $X_t$ is correspondingly defined as in Eq. (35), then $(X_t)_{t\geq 1}$ is the asymptotic state of $(\boldsymbol{x}_t)_{t\geq 1}$.*

We emphasize that the index set $t \geq 1$ is independent of $n$, and so our results hold for all fixed iterates $t$ independent of $n$, in the limit $n \to \infty$.

*Proof.* The statement for bounded continuous test functions $\varphi$ follows from Theorem 6.2 and the continuous mapping theorem. For polynomial $\varphi$, we proceed by a truncation argument. Let $S_n := \mathrm{samp}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)$ and $S := (X_1, \dots, X_t)$. Fix a cutoff $K > 0$ and consider any bounded continuous function $\varphi_K$ such that $|\varphi_K| \leq |\varphi|$, $\varphi_K(s) = \varphi(s)$ for all $\|s\|_2 \leq K$ and $\varphi_K(s) = 0$ for all $\|s\|_2 > 2K$ (standard approximations show that such a function exists). First, $|\mathbb{E}\, \varphi_K(S_n) - \mathbb{E}\, \varphi_K(S)|$

converges to 0 as $n \to \infty$ by the bounded continuous case. Next,

$$\begin{aligned}
|\mathbb{E}\,\varphi(S_n) - \mathbb{E}\,\varphi_K(S_n)| &\le \mathbb{E}\left[|\varphi(S_n)|\mathbf{1}_{\|S_n\|_2 > K}\right] && \text{(Definition of the truncated function)} \\
&\le (\mathbb{E}\,\varphi(S_n)^2)^{\frac{1}{2}} \Pr(\|S_n\|_2 > K)^{\frac{1}{2}} && \text{(Cauchy-Schwarz inequality)} \\
&\le (\mathbb{E}\,\varphi(S_n)^2)^{\frac{1}{2}} \frac{(\mathbb{E}\,\|S_n\|_2^2)^{\frac{1}{2}}}{K} && \text{(Markov inequality)}
\end{aligned}$$

Note that these quantities are respectively equal to

$$\mathbb{E}\,\varphi(S_n)^2 = \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{x}_1[i], \dots, \boldsymbol{x}_t[i])^2 \quad \text{and} \quad \mathbb{E}\,\|S_n\|_2^2 = \sum_{s=1}^{t} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_s[i]^2\,,$$

which both converge as $n \to \infty$ by existence of the traffic distribution, and in particular are bounded uniformly in $n$. Hence, there exists $C > 0$ independent of $n$ and $K$ such that

$$\limsup_{n\to\infty} |\mathbb{E}\,\varphi(S_n) - \mathbb{E}\,\varphi_K(S_n)| \le \frac{C}{K}\,,$$

and the same holds for $\mathbb{E}\,\varphi(S) - \mathbb{E}\,\varphi_K(S)$ by the same argument, using the fact that all moments exist in the space generated by $Z^{\infty}_{\mathcal{A}_1}$. We obtain $\limsup_{n\to\infty} |\mathbb{E}\,\varphi(S_n) - \mathbb{E}\,\varphi_K(S_n)| \le 2C/K$, and the claim follows from taking the limit $K \to \infty$. □

By Claim 1.5, the iterates of a pGFOM are of the form Eq. (34), so they have an asymptotic state. By definition, these asymptotic states determine the limiting distribution of any (bounded continuous or polynomial) observable. Motivated by this, we introduce a family of approximate message passing algorithms whose asymptotic states are conditionally Gaussian.

**Theorem 6.18** (Treelike AMP)**.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies the assumptions of Theorem 6.2. Let $f_t : \mathbb{R} \to \mathbb{R}$ be polynomial functions.*[12] *Define:*

$$\begin{aligned}
\boldsymbol{x}_0 &= \mathbf{1}\,, \qquad \boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{f}_{t-1} - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot \boldsymbol{f}_s\,, \\
\boldsymbol{b}_{s,t}[i] &:= \sum_{\substack{i_s,\dots,i_{t-1}\in[n] \text{ distinct} \\ i_s = i}} \left(\prod_{r=s+1}^{t-1} \boldsymbol{A}[i_{r-1}, i_r]\boldsymbol{f}_r'[i_r]\right) \boldsymbol{A}[i_{t-1}, i_s]\,, \qquad (36) \\
\boldsymbol{f}_t &:= f_t(\boldsymbol{x}_t)\,, \qquad \boldsymbol{f}_t' := f_t'(\boldsymbol{x}_t)\,, \qquad \boldsymbol{f}_0 = \mathbf{1}\,.
\end{aligned}$$

*Then $\boldsymbol{x}_t \in \operatorname{span}(\boldsymbol{z}_{\mathcal{G}_1 \cup (\mathcal{A}_1 \setminus \mathcal{T}_1)}(\boldsymbol{A}))$. Therefore, the asymptotic state of $(\boldsymbol{x}_t)_{t\ge 1}$ defined in Eq. (35) is a centered Gaussian process conditionally on $Z^{\infty}_{\mathcal{C}_1}$.*

To prove Theorem 6.18, motivated by the results in Section 6.1, we introduce the following handy notations:

**Definition 6.19** (Equality modulo non-treelike diagrams)**.** *For $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1})$, we write $\boldsymbol{x} \overset{\infty}{=} \boldsymbol{y}$ if $\boldsymbol{x} - \boldsymbol{y} \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1})$. We denote by $\operatorname{cactus}(\boldsymbol{x})$ the projection of $\boldsymbol{x}$ onto the span of the cactus diagrams $\mathcal{C}_1$, and by $\operatorname{gaussian}(\boldsymbol{x})$ the projection of $\boldsymbol{x}$ onto the span of the Gaussian diagrams $\mathcal{G}_1$.*

[12] For ease of exposition $f_t$ is assumed to be "memoryless", meaning that it only takes the most recent $\boldsymbol{x}_t$ as input.

The iterates of the treelike AMP algorithm Eq. (36) are engineered to asymptotically generate a self-avoiding walk. That is, whenever the algorithm performs a matrix multiplication operation, the Onsager correction terms in Eq. (36) (the subtracted terms involving $\boldsymbol{b}_{s,t}$) are chosen to subtract off the terms in the resulting diagram expansion which (1) are treelike and (2) revisit an existing vertex in any diagram.

**Example 6.20** (Self-avoiding walk)**.** *For intuition, consider the case of Theorem 6.18 where $f_t(x) = x$. Let $\pi_t$ be the $t$-path diagram and $\rho_t$ the $t$-cycle diagram. We can expand exactly:*

$$\boldsymbol{A}\boldsymbol{z}_{\pi_t} = \boldsymbol{z}_{\pi_{t+1}} + \sum_{s=0}^{t} \boldsymbol{z}_{\rho_{s+1} \oplus \pi_{t-s}} \,.$$

*For each term on the right-hand side, we have the approximate factorization (by Lemma 6.10) $\boldsymbol{z}_{\rho_{s+1}\oplus\pi_{t-s}} \overset{\infty}{=} \boldsymbol{z}_{\rho_{s+1}} \cdot \boldsymbol{z}_{\pi_{t-s}}$, which holds up to non-treelike terms. Then, we define a self-avoiding version of power iteration by:*

$$\boldsymbol{x}_0 = \boldsymbol{1}, \qquad \boldsymbol{x}_{t+1} = \boldsymbol{A}\boldsymbol{x}_t - \sum_{s=0}^{t} \boldsymbol{z}_{\rho_{s+1}} \cdot \boldsymbol{x}_{t-s} \,.$$

*By construction, we have $\boldsymbol{x}_t \overset{\infty}{=} \boldsymbol{z}_{\pi_t}$ and therefore, assuming the conditions of Theorem 6.2, the asymptotic state $X_t$ of $\boldsymbol{x}_t$ is Gaussian.*

To analyze a general iteration in the proof of Theorem 6.18, we separate the diagram expansion of $\boldsymbol{f}_t$ into its linear and nonlinear parts:

$$\boldsymbol{f}_t = \underbrace{\sum_{\gamma \in \mathcal{G}_1} c_\gamma \boldsymbol{z}_\gamma(\boldsymbol{A})}_{=:\boldsymbol{f}_t^1} + \underbrace{\sum_{\tau \in \mathcal{T}_1 \setminus \mathcal{G}_1} c_\tau \boldsymbol{z}_\tau(\boldsymbol{A})}_{=:\boldsymbol{f}_t^{\neq 1}} + \sum_{\alpha \in \mathcal{A}_1 \setminus \mathcal{T}_1} c_\alpha \boldsymbol{z}_\alpha(\boldsymbol{A}) \,.$$

We call $\boldsymbol{f}_t^1 = \mathrm{gaussian}(\boldsymbol{f}_t)$ the "linear part" since it should be thought of as the degree-1 part of the Hermite expansion of $\boldsymbol{f}_t$ with respect to the Gaussian vectors $\boldsymbol{z}_{\mathcal{G}_1}(\boldsymbol{A})$, while $\boldsymbol{f}_t^{\neq 1}$ equals all of the other components of the Hermite expansion. More precisely, when $\boldsymbol{f}_t$ is of the form $f_t(\boldsymbol{x}_t)$ for some Gaussian vector $\boldsymbol{x}_t$, which is the situation for AMP, then $\boldsymbol{f}_t^1$ has the following simple form.

**Lemma 6.21.** *Let $\boldsymbol{x} \in \mathrm{span}(\boldsymbol{z}_{\mathcal{G}_1})$ and let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial. Then,*

$$\mathrm{gaussian}(f(\boldsymbol{x})) \overset{\infty}{=} \mathrm{cactus}(f'(\boldsymbol{x})) \cdot \boldsymbol{x} \,.$$

*Proof.* Suppose that $f(x) = x^\ell$ for some integer $\ell \geq 0$ (the general case follows by linearity). By Lemma 6.10, the product of diagrams in $\mathcal{G}_1$ yields a diagram in $\mathcal{G}_1$ only when every diagram except one is matched. Formally, write $\boldsymbol{x} = \sum_{\gamma \in \mathcal{G}_1} c_\gamma \boldsymbol{z}_\gamma$, then by Lemma 6.10,

$$\boldsymbol{f}^1 \overset{\infty}{=} \ell \sum_{\gamma_1, \ldots, \gamma_\ell \in \mathcal{G}_1} c_{\gamma_1} \ldots c_{\gamma_\ell} \sum_{M \in \mathcal{M}_{\mathrm{perf}}(\ell-1)} \sum_{\substack{P_{uv} \in H(\gamma_u, \gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \gamma_\ell} \,. \tag{37}$$

Viewing $\bigoplus_{uv \in M} \gamma_{P_{u,v}}$ as a fixed cactus, by Lemma 6.10, every term on the right-hand side satisfies

$$\boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \gamma_\ell} \overset{\infty}{=} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}}} \cdot \boldsymbol{z}_{\gamma_\ell} \,. \tag{38}$$

Applying Lemma 6.10 one last time, it remains to observe that the cactus part of $f'(\boldsymbol{x}) = \ell \boldsymbol{x}^{\ell-1}$ is

$$\operatorname{cactus}(f'(\boldsymbol{x})) \overset{\infty}{=} \ell \sum_{\gamma_1,\dots,\gamma_{\ell-1} \in \mathcal{G}_1} c_{\gamma_1} \dots c_{\gamma_{\ell-1}} \sum_{M \in \mathcal{M}_{\text{perf}}(\ell-1)} \sum_{\substack{P_{uv} \in H(\gamma_u,\gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}}} . \tag{39}$$

Combining Eqs. (37) to (39) yields the desired claim. □

The next key Lemma 6.23 derives an explicit asymptotic formula for the AMP iterates: $\boldsymbol{x}_t$ is generated by taking a self-avoiding walk from each nonlinear term $\boldsymbol{f}_s^{\neq 1}$.

**Definition 6.22.** *Let $\boldsymbol{c}_t = \operatorname{cactus}(f_t'(\boldsymbol{x}_t))$. Define the self-avoiding walk matrix $\boldsymbol{B}_{s,t}$ generated by the iteration between time $s$ and $t$ to be:*

$$\boldsymbol{B}_{s,t}[i,j] := \sum_{\substack{i_s,\dots,i_t \in [n] \text{ distinct} \\ i_s = j,\, i_t = i}} \boldsymbol{A}[i_s, i_{s+1}] \cdots \boldsymbol{A}[i_{t-1}, i_t] \cdot \boldsymbol{c}_{s+1}[i_{s+1}] \cdots \boldsymbol{c}_{t-1}[i_{t-1}] .$$

Recalling Definition 5.4, $\boldsymbol{B}_{s,t}$ is a linear combination of open cactus matrices in the $z$-basis (up to non-treelike terms which arise from intersections involving the $\boldsymbol{c}_{s+i}$). For example, $\boldsymbol{B}_{t-1,t}$ equals $\boldsymbol{A}$ with the diagonal elements set to zero. We note the analogy between $\boldsymbol{b}_{s,t}$ and $\boldsymbol{B}_{s,t}$, which contain similar self-avoiding walks that return to the start and do not return to the start, respectively.

**Lemma 6.23.** *Define $\boldsymbol{x}_t, \boldsymbol{f}_t$ by Eq. (36) and let $\boldsymbol{c}_t = \operatorname{cactus}(f_t'(\boldsymbol{x}_t))$. Then for $t \geq 1$:*

$$\boldsymbol{x}_t \overset{\infty}{=} \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t} \boldsymbol{f}_s^{\neq 1} \quad \text{and} \quad \boldsymbol{f}_t \overset{\infty}{=} \boldsymbol{c}_t \cdot \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t} \boldsymbol{f}_s^{\neq 1} + \boldsymbol{f}_t^{\neq 1} .$$

*Proof.* First, note that for a fixed $t$, the second equation follows from the first:

$$\begin{aligned}
\boldsymbol{f}_t &\overset{\infty}{=} \boldsymbol{f}_t^1 + \boldsymbol{f}_t^{\neq 1} \\
&\overset{\infty}{=} \boldsymbol{c}_t \cdot \boldsymbol{x}_t + \boldsymbol{f}_t^{\neq 1} && \text{(Lemma 6.21)} \\
&\overset{\infty}{=} \boldsymbol{c}_t \cdot \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t} \boldsymbol{f}_s^{\neq 1} + \boldsymbol{f}_t^{\neq 1} && \text{(first equation and Lemma D.3)}
\end{aligned}$$

To establish the equations, we use induction on $t$. In the base case $t = 1$ we have $\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{f}_0 - \boldsymbol{b}_{0,1} \cdot \boldsymbol{f}_0 = \boldsymbol{B}_{0,1}\boldsymbol{f}_0$ as needed. Now, assume that the formulas hold for $0, \dots, t$. Denote by $\boldsymbol{C}_t$ the diagonal matrix with entries $\boldsymbol{c}_t$. The equation for $\boldsymbol{f}_t$ implies:

$$\boldsymbol{A}\boldsymbol{f}_t \overset{\infty}{=} \sum_{s=0}^{t-1} \boldsymbol{A}\boldsymbol{C}_t\boldsymbol{B}_{s,t}\boldsymbol{f}_s^{\neq 1} + \boldsymbol{A}\boldsymbol{f}_t^{\neq 1} \tag{40}$$

If we expand the matrix product $\boldsymbol{A}\boldsymbol{C}_t\boldsymbol{B}_{s,t}$ we can partition the sum based on whether the matrix

$\boldsymbol{A}$ revisits a vertex already on the walk:

$$(\boldsymbol{A}\boldsymbol{C}_t\boldsymbol{B}_{s,t})[i,j] = \sum_{k=1}^{n} \boldsymbol{A}[i,k]\boldsymbol{c}_t[k] \sum_{\substack{i_s,\dots,i_t\in[n] \text{ distinct} \\ i_s=j,\, i_t=k}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_{t-1},i_t]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_{t-1}[i_{t-1}]$$

$$= \sum_{\substack{i_s,\dots,i_{t+1}\in[n] \text{ distinct} \\ i_s=j,\, i_{t+1}=i}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_t,i_{t+1}]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_t[i_t] \tag{41}$$

$$+ \sum_{r=s}^{t} \sum_{\substack{i_s,\dots,i_t\in[n] \text{ distinct} \\ i_s=j,\, i_r=i}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_{t-1},i_t]\cdot\boldsymbol{A}[i_r,i_t]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_t[i_t]\,. \tag{42}$$

The first term Eq. (41) is self-avoiding and equals $\boldsymbol{B}_{s,t+1}[i,j]$. In the second term Eq. (42), the term $r=s$ is diagrammatically a cycle and is equal to $\boldsymbol{b}_{s,t+1}[i]$ when $i=j$, and 0 otherwise:

**Claim 6.24.** *We have:*

$$\boldsymbol{b}_{s,t} \overset{\infty}{=} \left( \sum_{\substack{i_s,\dots,i_{t-1}\in[n] \text{ distinct} \\ i_s=i}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_{t-2},i_{t-1}]\boldsymbol{A}[i_{t-1},i_s]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_{t-1}[i_{t-1}] \right)_{i\in[n]}.$$

*Proof.* The only difference between this formula and the definition of $\boldsymbol{b}_{s,t}$ is that the vectors $\boldsymbol{f}'_t$ at the internal vertices of the cycle have been replaced by $\boldsymbol{c}_t$. This holds up to non-treelike terms since placing a non-cactus diagram at any internal vertex of the cycle will create only non-treelike diagrams. □

The remaining terms in Eq. (42) are a cycle and a path joined together at vertex $r$:

**Claim 6.25.** *Let $r\in\{s+1,\dots,t\}$. For $i\in[n]$, let*

$$\boldsymbol{u}[i] = \sum_{\substack{i_s,\dots,i_t\in[n] \text{ distinct} \\ i_r=i}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_{t-1},i_t]\cdot\boldsymbol{A}[i_r,i_t]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_t[i_t]\boldsymbol{f}_s^{\neq 1}[i_s]\,.$$

*Then $\boldsymbol{u}\overset{\infty}{=}\boldsymbol{b}_{r,t+1}\cdot\boldsymbol{C}_r\boldsymbol{B}_{s,r}\boldsymbol{f}_s^{\neq 1}$.*

*Proof.* By expanding definitions, we can conveniently interpret

$$\boldsymbol{b}_{r,t+1}\cdot\boldsymbol{C}_r\boldsymbol{B}_{s,r}\boldsymbol{f}_s^{\neq 1}[i] = \sum_{\substack{i_s,\dots,i_t\in[n] \\ i_s,\dots,i_r \text{ distinct} \\ i_r,\dots,i_t \text{ distinct} \\ i_r=i}} \boldsymbol{A}[i_s,i_{s+1}]\cdots\boldsymbol{A}[i_{t-1},i_t]\cdot\boldsymbol{A}[i_r,i_t]\cdot\boldsymbol{c}_{s+1}[i_{s+1}]\cdots\boldsymbol{c}_t[i_t]\boldsymbol{f}_s^{\neq 1}[i_s]\,.$$

Since the diagram induced on $\{i_r,\dots,i_t\}$ is a cycle, any intersection between the vertices $\{i_s,\dots,i_r\}$ and $\{i_r,\dots,i_t\}$ would create a non-treelike diagram. □

Plugging Claim 6.25 in to Eq. (40), we have:

$$\begin{aligned}\boldsymbol{A}\boldsymbol{f}_t &\overset{\infty}{=} \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t+1}\boldsymbol{f}_s^{\neq 1} + \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t+1}\cdot \boldsymbol{f}_s^{\neq 1} + \sum_{s=0}^{t-1}\sum_{r=s+1}^{t} \boldsymbol{b}_{r,t+1}\cdot(\boldsymbol{C}_r\boldsymbol{B}_{s,r}\boldsymbol{f}_s^{\neq 1}) + \underbrace{\boldsymbol{A}\boldsymbol{f}_t^{\neq 1}}_{=\boldsymbol{B}_{t,t+1}\boldsymbol{f}_t^{\neq 1}+\boldsymbol{b}_{t,t+1}\cdot\boldsymbol{f}_t^{\neq 1}} \\ &= \sum_{s=0}^{t} \boldsymbol{B}_{s,t+1}\boldsymbol{f}_s^{\neq 1} + \sum_{r=0}^{t} \boldsymbol{b}_{r,t+1}\cdot\left(\boldsymbol{C}_r\sum_{s=0}^{r-1}\boldsymbol{B}_{s,r}\boldsymbol{f}_s^{\neq 1} + \boldsymbol{f}_r^{\neq 1}\right) \\ &\overset{\infty}{=} \sum_{s=0}^{t} \boldsymbol{B}_{s,t+1}\boldsymbol{f}_s^{\neq 1} + \sum_{r=0}^{t} \boldsymbol{b}_{r,t+1}\cdot\boldsymbol{f}_r\end{aligned}$$

The last equality is the inductive formula for $\boldsymbol{f}_r$. The Onsager correction subtracts off the second sum, leaving only the desired first sum for $\boldsymbol{x}_{t+1}$. □

*Proof of Theorem 6.18.* We prove the following purely combinatorial claim about Eq. (36): $\boldsymbol{x}_t$ is in the span of non-treelike diagrams and *Gaussian* treelike diagrams. By Lemma 6.17, this will imply that conditioned on $Z_{\mathcal{C}_1}^{\infty}$, the asymptotic state of $(\boldsymbol{x}_t)_{t\geq 1}$ is a centered Gaussian process, as desired.

To show the claim, we start from the conclusion of Lemma 6.23:

$$\boldsymbol{x}_t \overset{\infty}{=} \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t}\boldsymbol{f}_s^{\neq 1}\,.$$

The diagrams in $\boldsymbol{B}_{s,t}\boldsymbol{f}_s^{\neq 1}$ are obtained by: (1) choose a diagram from $\boldsymbol{f}_s^{\neq 1}$, (2) choose a cactus diagram from $\boldsymbol{c}_r$ at each internal vertex of $\boldsymbol{B}_{s,t}$ (i.e. each internal vertex along a path of length $t-s$), (3) multiply these diagrams together. Since none of the diagrams in $\boldsymbol{f}_s^{\neq 1}$ have degree 1 at the root by definition, the only treelike terms in the product are formed by grafting the diagrams together without intersections. In particular, the root is the endpoint of the path in $\boldsymbol{B}_{s,t}$ and has degree 1. This concludes the proof. □

### 6.2.1 Covariance structure of treelike AMP

While Theorem 6.18 shows that the treelike AMP iterates are asymptotically Gaussian, it does not identify their covariance. We calculate the covariance "combinatorially" by calculating the cactus diagrams appearing in the expansion of $\langle \boldsymbol{x}_s, \boldsymbol{x}_t\rangle$.

**Proposition 6.26.** *Let $\boldsymbol{x}_t$ follow the iteration Eq. (36). Then for any $s,t\geq 1$,*

$$\boldsymbol{x}_s\cdot\boldsymbol{x}_t - \sum_{s'=0}^{s-1}\sum_{t'=0}^{t-1}\boldsymbol{B}_{s'st't}(\boldsymbol{f}_{s'}\cdot\boldsymbol{f}_{t'}) \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1\setminus\mathcal{C}_1})\,, \tag{43}$$

*where for $0\leq s'\leq s, 0\leq t'\leq t$, we define the matrix $\boldsymbol{B}_{s'st't}\in\mathbb{R}^{n\times n}$ by*

$$\boldsymbol{B}_{s'st't}[i,j] := \sum_{\substack{i_{s'},\dots,i_s,j_{t'},\dots,j_t\in[n]\\ \text{distinct except}\\ i_{s'}=j_{t'}=j,\, i_s=j_t=i}} \prod_{r=s'}^{s-1}\boldsymbol{A}[i_r,i_{r+1}]\prod_{r=t'}^{t-1}\boldsymbol{A}[j_r,j_{r+1}]\prod_{r=s'+1}^{s-1}\boldsymbol{c}_r[i_r]\prod_{r=t'+1}^{t-1}\boldsymbol{c}_r[j_r]\,.$$

When we apply this lemma, we will average Eq. (43) over the coordinates $i\in[n]$ and over $\boldsymbol{A}$.

Since the error terms are all in the span of non-cactus diagrams, all error terms converge to 0 by the strong cactus property. On the other hand, the average of the term $\boldsymbol{x}_s \cdot \boldsymbol{x}_t$ converges to the covariance $\mathbb{E}[X_s X_t]$ which we want to calculate. The subtracted terms involving the $\boldsymbol{B}_{s't'st}$ matrices converge to limits depending on the asymptotic values of the cactuses $Z_{\mathcal{C}_1}^{\infty}$. For some settings (such as Proposition 6.3), the values $Z_{\mathcal{C}_1}^{\infty}$ are deterministic. For other settings, the values $Z_{\mathcal{C}_1}^{\infty}$ are random, and we will condition on them in order to obtain the conditional covariance.

*Proof.* Since $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ have degree exactly one at the root, in order to form a cactus in $\boldsymbol{x}_s \cdot \boldsymbol{x}_t$, the paths from the root of $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ in the expansion from Lemma 6.23 must meet at some point. This intersection cannot happen at a vertex from $\boldsymbol{f}_{s'}^{\neq 1}$ or $\boldsymbol{f}_{t'}^{\neq 1}$ (that would create edges in two cycles). Let $s' \in \{0, \dots, s-1\}$ and $t' \in \{0, \dots, t-1\}$ denote the integers such that the first intersection corresponds to the indices $i_{s'}$ (for $\boldsymbol{x}_s$) and $i_{t'}$ (for $\boldsymbol{x}_t$) in Definition 6.22. Then, we can decompose

$$\boldsymbol{x}_s \cdot \boldsymbol{x}_t - \sum_{s'=0}^{s-1} \sum_{t'=0}^{t-1} \boldsymbol{B}_{s'st't}((\boldsymbol{c}_{s'} \cdot \boldsymbol{x}_{s'} + \boldsymbol{f}_{s'}^{\neq 1}) \cdot (\boldsymbol{c}_{t'} \cdot \boldsymbol{x}_{t'} + \boldsymbol{f}_{t'}^{\neq 1})) \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{C}_1}),$$

and the conclusion follows from the equality (Lemma 6.21) $\boldsymbol{f}_s \overset{\infty}{=} \boldsymbol{c}_s \cdot \boldsymbol{x}_s + \boldsymbol{f}_s^{\neq 1}$. □

The cactus expansion of $\boldsymbol{B}_{s'st't}(\boldsymbol{f}_{s'} \cdot \boldsymbol{f}_{t'})$ can be obtained explicitly by combining a cycle of length $s - s' + t - t'$ along the edges of $\boldsymbol{B}_{s'st't}$, a cactus from $\boldsymbol{c}_r$ hanging at every vertex $r$ in the cycle, and a homeomorphic matching of the tree components of $\boldsymbol{f}_{s'}$ and $\boldsymbol{f}_{t'}$ (Definition 6.8).

## 6.3 Examples of state evolution

In this section, we specialize Theorem 6.18 to obtain a more explicit description of the state evolution of the treelike AMP algorithm for several concrete matrix models.

**Notation 6.27.** *For a vector $\boldsymbol{x} \in \mathbb{R}^n$, we will use the following notation for empirical averages:*

$$\langle \boldsymbol{x} \rangle := \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Technically, most algorithms in this section are not pGFOM since they calculate empirical averages. Assuming that the traffic distribution concentrates and the vector $\boldsymbol{x}$ lies in the diagram basis, then the empirical average $\langle \boldsymbol{x} \rangle$ concentrates, and we can replace $\langle \boldsymbol{x} \rangle$ by its limit $\mathbb{E}\, X$ without changing the asymptotic state of the algorithm. This is formally proven in Lemma D.10.

### 6.3.1 Orthogonally invariant random matrices

In the special case that $\boldsymbol{A}$ is drawn from an orthogonally invariant random matrix ensemble, the treelike AMP algorithm recovers the orthogonal AMP algorithm of Fan [Fan22], giving a new proof of this result.

**Theorem 6.28** (State evolution for orthogonally invariant matrices)**.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}_{\mathrm{sym}}^{n \times n}$ be an orthogonally invariant random matrix converging in tracial moments in $L^2$ to a probability measure with free cumulants $(\kappa_q)_{q \geq 1}$. Assume $\boldsymbol{A}$ satisfies Eq. (4). Let $f_t : \mathbb{R} \to \mathbb{R}$ be polynomial functions*

*and define the iteration*

$$\boldsymbol{x}_0 = \mathbf{1}, \qquad \boldsymbol{x}_t = \boldsymbol{A} f_{t-1}(\boldsymbol{x}_{t-1}) - \sum_{s=0}^{t-1} \kappa_{t-s} \left( \prod_{r=s+1}^{t-1} \langle f_r'(\boldsymbol{x}_r) \rangle \right) f_s(\boldsymbol{x}_s) \quad \forall t \geq 1 \,. \tag{44}$$

*Then, the asymptotic state of $(\boldsymbol{x}_t)_{t\geq 1}$ is a centered Gaussian process $(X_t)_{t\geq 1}$ with covariance*

$$\mathbb{E}\,[X_s X_t] = \sum_{s'=0}^{s-1} \sum_{t'=0}^{t-1} \kappa_{s-s'+t-t'} \left( \prod_{r=s'+1}^{s-1} \mathbb{E}\, f_r'(X_r) \right) \left( \prod_{r=t'+1}^{t-1} \mathbb{E}\, f_r'(X_r) \right) \mathbb{E}\,[f_{s'}(X_{s'}) f_{t'}(X_{t'})] \quad \forall s, t \geq 1 \,,$$

*with $X_0 := 1$.*

*Proof.* By Theorem 4.2, $\boldsymbol{A}$ satisfies the factorizing strong cactus property and its diagonal distribution exists, so the assumptions of Theorem 6.2 and Theorem 6.18 are satisfied. Therefore, the treelike AMP algorithm in Eq. (36) has Gaussian asymptotic state.

We now specialize the Onsager correction term in Eq. (36) to this model. The term $\boldsymbol{b}_{s,t}$ is represented by a cycle of length $t - s$, with $f_r'(\boldsymbol{x}_r)$ attached to the $r$th vertex of the cycle for each $s < r < t$. By Lemma D.3, we only need to look at treelike contributions in $\boldsymbol{b}_{s,t}$. Because of the base cycle, these are only cactuses, obtained by attaching cactuses from $(f_r'(\boldsymbol{x}_r))_{s<r<t}$ along the base cycle. By Proposition 6.3, $\boldsymbol{b}_{s,t}$ has constant asymptotic state equal to $\kappa_{t-s} \prod_{r=s+1}^{t-1} \mathbb{E}\, f_r'(X_r)$. The cactuses in $\boldsymbol{b}_{s,t}$ persist until the end of the algorithm, so that they will eventually contribute this value towards the asymptotic state. Hence it does not affect the asymptotic state to replace $\boldsymbol{b}_{s,t}$ immediately by its limiting constant value.

Moreover, by Lemma D.10 and Lemma B.7, we may replace $\mathbb{E}\, f_r'(X_r)$ by the empirical average $\langle f_r'(\boldsymbol{x}_r) \rangle$ to obtain Eq. (44) without affecting the asymptotic state. Now the asymptotic state $X_t$ of Eq. (44) matches that of Eq. (36), and we may apply Theorem 6.18 to deduce that $X_t$ is Gaussian.

To calculate the covariance $\mathbb{E}\,[X_s X_t]$, we average Proposition 6.26 over the coordinates $i \in [n]$ and take the limit $n \to \infty$. On the right side of Eq. (43), the cycle of $\boldsymbol{B}_{s'st't}$ contributes $\kappa_{s-s'+t-t'}$ and the hanging diagrams $f_r'(\boldsymbol{x}_r)$ inside $\boldsymbol{B}_{s'st't}$ contribute $\mathbb{E}\, f_r'(X_r)$ by the factorizing cactus property. The cactuses in $f_{s'}(\boldsymbol{x}_{s'}) \cdot f_{t'}(\boldsymbol{x}_{t'})$ contribute $\mathbb{E}\,[f_{s'}(X_{s'}) f_{t'}(X_{t'})]$, which establishes the desired recurrence. □

Note that this proof only uses the strong factorizing cactus property and the concentration of the traffic distribution, which explains why Theorem 6.28 also holds for non-orthogonally invariant matrix models such as Wigner matrices (Section 4.1).

### 6.3.2 Punctured random and deterministic matrices

The punctured matrices studied in Section 5 do not satisfy the strong cactus property, so we cannot directly apply Theorem 6.18 to derive an AMP iteration for them. However, a reduction allows us to derive the state evolution of punctured orthogonally invariant random matrices from that of their unpunctured counterparts. These matrices are central because, by Theorem 5.3, they provide an intermediate step in deriving the state evolution of sequences of punctured *deterministic* matrices satisfying Assumption 5.2.

Note that a GFOM run on a punctured matrix must be initialized with a random vector $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, rather than $\boldsymbol{x}_0 = \mathbf{1}$, to avoid triviality.

**Theorem 6.29** (State evolution for punctured matrices)**.** *Let $\boldsymbol{H} = \boldsymbol{H}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be a sequence of orthogonally invariant random matrices satisfying Eq. (4) and converging in tracial moments in $L^2$ to a probability measure with free cumulants $(\kappa_q)_{q\ge 1}$. Let $\boldsymbol{A}$ denote the puncturing of $\boldsymbol{H}$ (Definition 2.1). Let $f_t : \mathbb{R} \to \mathbb{R}$ be polynomial functions with $f_0(x) = x$, and consider the pGFOM:*

$$\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})\,, \quad \boldsymbol{x}_t = \boldsymbol{A} f_{t-1}(\boldsymbol{x}_{t-1}) - \sum_{s=0}^{t-1} \kappa_{t-s} \left( \prod_{r=s+1}^{t-1} \langle f_r'(\boldsymbol{x}_r)\rangle \right) (f_s(\boldsymbol{x}_s) - \langle f_s(\boldsymbol{x}_s)\rangle \mathbf{1}) \quad \forall t \ge 1\,.$$

*Then for any $t \ge 1$ and any polynomial $\varphi : \mathbb{R}^t \to \mathbb{R}$, we have*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{x}_0} \langle \varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)\rangle = \mathbb{E}\, \varphi(X_1, \dots, X_t)\,,$$

*where $(X_t)_{t\ge 1}$ is a centered Gaussian process with covariance given by*

$$\mathbb{E}\,[X_s X_t] = \sum_{s'=0}^{s-1} \sum_{t'=0}^{t-1} \kappa_{s-s'+t-t'} \left( \prod_{r=s'+1}^{s-1} \mathbb{E}\, f_r'(X_r) \right) \left( \prod_{r=t'+1}^{t-1} \mathbb{E}\, f_r'(X_r) \right) \mathbb{E}\left[ \overline{F_{s'}}\, \overline{F_{t'}} \right] \quad \forall s, t \ge 1\,,$$
$$\overline{F_0} := 1\,, \quad \overline{F_t} := f_t(X_t) - \mathbb{E}\, f_t(X_t) \quad \forall t \ge 1\,.$$

By Theorem 5.1, the conclusion of Theorem 6.29 also holds for any sequence of deterministic matrices satisfying the delocalization assumption Assumption 5.2 and having a limiting diagonal distribution that factorizes over cycles (that is, matches the diagonal distribution of some orthogonally invariant random matrix ensemble). In particular, the conclusion holds for the Walsh-Hadamard matrices and the Discrete Cosine and Sine Transform matrices, for which the $\kappa_q$ are the free cumulants of the ROM (Eq. (9)).

The proof of Theorem 6.29 proceeds by reducing to the following iteration on the original, non-punctured matrix, initialized at the all-ones vector:

$$\boldsymbol{u}_0 = \mathbf{1}\,, \qquad \boldsymbol{u}_t = \boldsymbol{H}\overline{\boldsymbol{f}_{t-1}} - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot \overline{\boldsymbol{f}_s} \quad \forall t \ge 1\,,$$
$$\text{where } \boldsymbol{b}_{s,t}[i] := \sum_{\substack{i_s,\dots,i_{t-1}\in[n] \text{ distinct} \\ i_s = i}} \left( \prod_{r=s+1}^{t-1} \boldsymbol{H}[i_{r-1}, i_r] \boldsymbol{f}_r'[i_r] \right) \boldsymbol{H}[i_{t-1}, i_s] \quad \forall t > s \ge 0\,, \tag{45}$$
$$\overline{\boldsymbol{f}_0} := \boldsymbol{u}_0\,, \quad \overline{\boldsymbol{f}_t} := \boldsymbol{\Pi} f_t(\boldsymbol{u}_t) \quad \forall t \ge 1\,.$$

**Lemma 6.30.** *For any $t \ge 1$ and any polynomial $\varphi : \mathbb{R}^t \to \mathbb{R}$,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{x}_0} \langle \varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)\rangle = \lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{H}} \langle \varphi(\boldsymbol{u}_1, \dots, \boldsymbol{u}_t)\rangle\,.$$

The proof of Lemma 6.30 is deferred to Appendix D.3.

*Proof of Theorem 6.29.* We apply Theorem 6.28 to $\boldsymbol{u}_t$ after replacing iteratively each occurrence of $\boldsymbol{\Pi} f_t(\boldsymbol{u}_t)$ by $f_t(\boldsymbol{u}_t) - \mathbb{E}\,[f_t(U_t)] \cdot \mathbf{1}$ (where $U_t$ is the asymptotic state of $\boldsymbol{u}_t$ as predicted by Theorem 6.28).

By Lemma D.10, this transformation does not change the asymptotic state of $\boldsymbol{u}_t$. The state evolution formula for polynomial test functions then transfers to $\boldsymbol{x}_t$ by Lemma 6.30. □

### 6.3.3 Block-structured random matrices

Our final example is the class of block-structured matrices whose blocks satisfy the factorizing strong cactus property, which we introduced in Section 4.3. As anticipated in Example 6.4, these matrices do not themselves satisfy the factorizing strong cactus property. Therefore, we start by describing the random limit $Z^{\infty}_{\mathcal{C}_1}$.

**Lemma 6.31.** *Let $q \in \mathbb{N}$. For $r, c \in [q]$, let $\boldsymbol{A}_{r,c} = \boldsymbol{A}^{(n)}_{r,c} \in \mathbb{R}^{\frac{n}{q}\times\frac{n}{q}}_{\mathrm{sym}}$ be a sequence of symmetric random matrices such that $\boldsymbol{A}_{r,c} = \boldsymbol{A}_{c,r}$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be the block matrix with blocks $(\boldsymbol{A}_{r,c})_{r,c\in[q]}$. Assume that each $\boldsymbol{A}_{r,c}$ satisfies the factorizing strong cactus property, and $(\boldsymbol{A}_{r,c})_{1\le r\le c\le q}$ are asymptotically traffic independent. Let $(\kappa^{\{r,c\}}_{\ell})_{\ell\ge 1}$ be the limiting free cumulants of $\boldsymbol{A}_{r,c}$. Then,*

$$(\mathrm{block}(i), \boldsymbol{z}_{\mathcal{C}_1}(\boldsymbol{A})[i]) \xrightarrow{(\mathrm{d})} (R, Z^{\infty}_{\mathcal{C}_1}(R))\,, \qquad i \sim \mathrm{Unif}([n])\,, \quad R \sim \mathrm{Unif}([q])\,,$$

*where the (deterministic) sequence $Z^{\infty}_{\mathcal{C}_1}(r)$ for $r \in [q]$ is defined recursively by:*

*(i) For the singleton cactus, $Z^{\infty}_{\mathrm{singleton}}(r) := 1$.*

*(ii) Suppose $\sigma \in \mathcal{C}_1$ is rooted at a vertex $u_1$ of degree 2. Let $(u_1, \dots, u_\ell)$ be the cycle incident to the root. Let $\sigma_2, \dots, \sigma_\ell \in \mathcal{C}_1$ be the rooted cactuses attached to the vertices of the cycle. Then*

$$Z^{\infty}_{\sigma}(r) := \begin{cases} \displaystyle\sum_{c\in[q]} \left[ \kappa^{\{r,c\}}_{\ell} \prod_{\substack{k=2\\ k \text{ odd}}}^{\ell} Z^{\infty}_{\sigma_k}(r) \prod_{\substack{k=2\\ k \text{ even}}}^{\ell} Z^{\infty}_{\sigma_k}(c) \right] & \text{if } \ell \text{ is even} \\ \displaystyle \kappa^{\{r,r\}}_{\ell} \prod_{k=2}^{\ell} Z^{\infty}_{\sigma_k}(r) & \text{if } \ell \text{ is odd} \end{cases}$$

*(iii) If $\sigma \in \mathcal{C}_1$ decomposes as $\sigma = \bigoplus_{k=1}^{\ell} \sigma_k$, then $Z^{\infty}_{\sigma}(r) := \prod_{k=1}^{\ell} Z^{\infty}_{\sigma_k}(r)$.*

*In particular, the law of the limit $Z^{\infty}_{\mathcal{C}_1}$ is $\mathrm{Unif}(\{Z^{\infty}_{\mathcal{C}_1}(r) : r \in [q]\})$.*

The proof is deferred to Appendix D.4. By unicity of the limit in distribution, Lemma 6.31, together with Theorem 6.2, determines the law of $Z^{\infty}_{\mathcal{A}_1}$. The joint convergence with $\mathrm{block}(i)$ clarifies the source of the randomness of $Z^{\infty}_{\mathcal{C}_1}$: it arises because of the random choice of the block an entry belongs to in the $\mathrm{samp}(\cdot)$ operation.

Using Lemma 6.31, we can specialize the treelike AMP iteration and its state evolution to concrete block-structured models. We start with block GOE matrices (Definition 4.4). A family of AMP iterations for such matrices was derived in [Ran11, JM13]. As we will discuss below, these iterations have the same asymptotic state as treelike AMP.

**Theorem 6.32** (State evolution for the block GOE model)**.** *Let $\boldsymbol{A} \sim \mathsf{BlockGOE}(n, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}_{\ge 0}$ is a symmetric matrix. Given polynomial functions $f_t : \mathbb{R} \to \mathbb{R}$, let*

$$\boldsymbol{x}_0 = \mathbf{1}\,, \quad \boldsymbol{x}_1 = \boldsymbol{A} f_0(\boldsymbol{x}_0)\,, \quad \boldsymbol{x}_t = \boldsymbol{A} f_{t-1}(\boldsymbol{x}_{t-1}) - (\boldsymbol{A}^{\odot 2} f'_{t-1}(\boldsymbol{x}_{t-1})) \cdot f_{t-2}(\boldsymbol{x}_{t-2}) \quad \forall t \ge 2\,. \tag{46}$$

*Then the asymptotic state of $(\boldsymbol{x}_t)_{t\geq 1}$ is a mixture $\frac{1}{q}\sum_{r\in[q]}\mu_r$ where $\mu_r$ denotes the law of a centered Gaussian process $(X_t)_{t\geq 1}$ with covariance kernel $\boldsymbol{\Gamma}_r$ defined recursively by*

$$\boldsymbol{\Gamma}_r[s,t] = \sum_{c\in[q]} \left[\boldsymbol{\Sigma}[r,c] \mathop{\mathbb{E}}_{(X_T)_{T\geq 1}\sim\mu_c} [f_{s-1}(X_{s-1})f_{t-1}(X_{t-1})]\right] \quad \forall r\in[q]\,, \quad \forall s,t\geq 1\,,$$

*with $X_0 := 1$.*

*Proof.* By the discussion after Theorem 4.7, $\boldsymbol{A}$ has a traffic distribution and satisfies the strong cactus property[13], so it satisfies the assumption of Theorem 6.18. We consider the treelike AMP iteration $\boldsymbol{x}_t$ in Eq. (36) applied to $\boldsymbol{A}$. We show that this iteration has the same asymptotic state as Eq. (46) by simplifying the Onsager correction term.

The free cumulants of the GOE are 0 except for $\kappa_2$, so by Lemma 6.31, the only asymptotically non-negligible cactuses are those such that every cycle is a 2-cycle. For any $s < t-2$, $\boldsymbol{b}_{s,t}$ contains an injective cycle of length larger than 2 that cannot be destroyed by later operations. For $s = t-2$, we have:

$$\boldsymbol{b}_{t-2,t}[i] = \sum_{\substack{j=1\\ j\neq i}}^{n} \boldsymbol{A}[i,j]^2 \boldsymbol{f}'_{t-1}[j] = (\boldsymbol{A}^{\odot 2}\boldsymbol{f}'_{t-1})[i] - \boldsymbol{A}[i,i]^2\boldsymbol{f}'_{t-1}[i]\,.$$

Both $\boldsymbol{A}[i,i]^2\boldsymbol{f}'_{t-1}[i]$ and $\boldsymbol{b}_{t-1,t}$ contain a self-loop that also cannot be destroyed by later operations. In conclusion, the treelike AMP algorithm from Eq. (36) and the iteration in Eq. (46) are equal up to negligible diagrams. By Theorem 6.18, the asymptotic state $(X_t)_{t\geq 1}$ of $(\boldsymbol{x}_t)_{t\geq 1}$ in Eq. (46) exists and is Gaussian conditionally on $Z^{\infty}_{\mathcal{C}_1}$, and so, in the construction from Lemma 6.31, it is Gaussian conditionally on the random variable $R$.

Next, we specialize the covariance formula given by Proposition 6.26. Since only cactuses of 2-cycles are nonzero in the traffic distribution of $\boldsymbol{A}$ (this may be induced from Lemma 6.31), only the term for $s' = s-1$ and $t' = t-1$ is non-negligible in the expansion of $\frac{1}{n}\mathbb{E}\,\boldsymbol{x}_s\cdot\boldsymbol{x}_t$ given by Proposition 6.26. The expansion into cactuses of that term is obtained by grafting together a 2-cycle at the root, and cactuses of 2-cycles from $\boldsymbol{f}_{s-1}$ and $\boldsymbol{f}_{t-1}$ at the child of the root. Applying the recursive formula for $Z^{\infty}_{\mathcal{C}_1}(r)$ in Lemma 6.31, we obtain:

$$\mathbb{E}\,[X_sX_t \mid R=r] = \sum_{c\in[q]} \boldsymbol{\Sigma}[r,c]\,\mathbb{E}\,[f_{s-1}(X_{s-1})f_{t-1}(X_{t-1}) \mid R=c] \qquad \forall r\in[q]\,.$$

Thus, we have shown that, conditionally on $R\sim\mathrm{Unif}([q])$, $(X_t)_{t\geq 1}$ is a Gaussian process with the required covariance. The result follows by taking $\mu_r$ to be the law of $(X_t)_{t\geq 1}$ conditionally on $R=r$. □

To illustrate the modularity of our approach, we also study a different block-structured matrix model whose blocks are not all GOE.

**Theorem 6.33** (State evolution for the community model)**.** *Let $\boldsymbol{M}\in\mathbb{R}^{\frac{n}{q}\times\frac{n}{q}}_{\mathrm{sym}}$ be an orthogonally invariant random matrix converging in tracial moments to a probability measure with free cumulants $(\kappa_q)_{q\geq 1}$ such that $\kappa_2 = \frac{1}{q}$. Let $\boldsymbol{A}$ be the random symmetric $n\times n$ matrix with blocks $(\boldsymbol{A}_{r,c})_{r,c\in[q]}$*

[13] One can also verify that it satisfies Eq. (4). But note that this was only needed in the proof of Theorem 6.2 to ensure the existence of the limit $Z^{\infty}_{\mathcal{C}_1}$, which we established directly in Lemma 6.31.

*given by* $\boldsymbol{A}_{1,1} = \boldsymbol{M}$ *and for all* $1 \le r \le c \le q$ *with* $(r,c) \neq (1,1)$, $\boldsymbol{A}_{r,c}$ *are i.i.d.* $\frac{n}{q} \times \frac{n}{q}$ *GOE matrices with entries of variance* $\frac{1}{n}$ *(and we set* $\boldsymbol{A}_{r,c} = \boldsymbol{A}_{c,r}$*).*

*Let* $\boldsymbol{x}_t$ *be the treelike AMP iteration Eq. (36) run on* $\boldsymbol{A}$ *with arbitrary polynomial nonlinearities. Then the asymptotic state* $(X_t)_{t\ge1}$ *of* $(\boldsymbol{x}_t)_{t\ge1}$ *is the mixture* $(1-\frac{1}{q})\mu_0 + \frac{1}{q}\mu_1$*, where* $\mu_i$ *is the law of a centered Gaussian process* $(X_t)_{t\ge1}$ *with covariance kernel* $\boldsymbol{\Gamma}_i$ *defined recursively by, for all* $s,t \ge 1$*:*

$$\begin{aligned}
&\boldsymbol{\Gamma}_0[s,t] = \mathbb{E}\left[F_{s-1}F_{t-1}\right] \\
&\boldsymbol{\Gamma}_1[s,t] = \mathbb{E}\left[F_{s-1}F_{t-1}\right] + \sum_{\substack{s'=0}}^{s-1}\sum_{\substack{t'=0 \\ (s',t')\neq(s-1,t-1)}}^{t-1} \kappa_{s-s'+t-t'} \left(\prod_{r=s'+1}^{s-1} \mathop{\mathbb{E}}_{\mu_1} F'_r\right)\left(\prod_{r=t'+1}^{t-1} \mathop{\mathbb{E}}_{\mu_1} F'_r\right) \mathop{\mathbb{E}}_{\mu_1}\left[F_{s'}F_{t'}\right] , \\
&F_t := f_t(X_t), \qquad F'_t := f'_t(X_t), \qquad X_0 = 1 ,
\end{aligned}$$

*where* $\mathbb{E}_{\mu_1}$ *denotes expectation with respect to* $(X_t)_{t\ge1} \sim \mu_1$*.*

*Proof.* The assumptions of Lemma 6.31 and Theorem 6.33 are satisfied. All blocks except the one in position $(1,1)$ have the same free cumulants (the GOE free cumulants, normalized so that $\kappa_2 = \frac{1}{q}$). Therefore, in the construction of Lemma 6.31, we have $Z^\infty_{\mathcal{C}_1}(r) = Z^\infty_{\mathcal{C}_1}(s)$ for all $r,s > 1$. Let $\mu_0$ (resp. $\mu_1$) be the law of the asymptotic state $(X_t)_{t\ge1}$ of the treelike AMP iteration $(\boldsymbol{x}_t)_{t\ge1}$ conditioned on $R > 1$ (resp. $R = 1$). By Theorem 6.18, both $\mu_0$ and $\mu_1$ are the laws of centered Gaussian processes. It remains to specialize the formula of Proposition 6.26 for their covariance to the present setting.

Conditionally on $R > 1$ (that is, outside the community), only 2-cycles at the root contribute to $Z^\infty_{\mathcal{C}_1}$. Thus, by combining the strong cactus property, Proposition 6.26, and Lemma 6.31, we obtain

$$\begin{aligned}
\mathbb{E}\left[X_sX_t \mid R>1\right] &= \frac{1}{q}\left(\mathbb{E}\left[f_{s-1}(X_s)f_{t-1}(X_t) \mid R=1\right] + (q-1)\,\mathbb{E}\left[f_{s-1}(X_s)f_{t-1}(X_t) \mid R>1\right]\right) \\
&= \mathbb{E}\left[f_{s-1}(X_s)f_{t-1}(X_t)\right] .
\end{aligned}$$

Conditionally on $R = 1$ (that is, inside the community), we also obtain a contribution of $\mathbb{E}\left[f_{s-1}(X_s)f_{t-1}(X_t)\right]$ from the term $s' = s-1$ and $t' = t-1$ in Proposition 6.26 (again using the normalization $\kappa_2 = \frac{1}{q}$ inside the community). For all of the remaining terms $s', t'$, when $s - s' + t - t'$ is an even integer larger than 2, we obtain a contribution only from $c = 1$ in Lemma 6.31, namely

$$\kappa_{s-s'+t-t'}\,\mathbb{E}\left[F_{s'}F_{t'} \mid R=1\right] \prod_{r=s'+1}^{s-1} \mathbb{E}\left[F'_r \mid R=1\right] \prod_{r=t'+1}^{t-1} \mathbb{E}\left[F'_r \mid R=1\right] .$$

When $s - s' + t - t'$ is odd, Lemma 6.31 yields exactly the same expression as the even case. Altogether, we obtain the recursion

$$\begin{aligned}
&\mathbb{E}\left[X_sX_t \mid R=1\right] = \mathbb{E}\left[F_{s-1}F_{t-1}\right] + \\
&\sum_{\substack{s'=0}}^{s-1}\sum_{\substack{t'=0 \\ (s',t')\neq(s-1,t-1)}}^{t-1} \kappa_{s-s'+t-t'}\,\mathbb{E}\left[F_{s'}F_{t'} \mid R=1\right] \prod_{r=s'+1}^{s-1} \mathbb{E}\left[F'_r \mid R=1\right] \prod_{r=t'+1}^{t-1} \mathbb{E}\left[F'_r \mid R=1\right] .
\end{aligned}$$

These are the desired covariance formulas for $\mu_0$ and $\mu_1$, and the mixing weights of the events $(R=1)$ and $(R>1)$ are indeed $\frac{1}{q}$ and $1-\frac{1}{q}$, respectively. □

### 6.3.4 Further extensions

There are several possible technical extensions of the methods we have developed here, whose full development is left for future work.

First, Lemma 6.31 applies to general orthogonally invariant distributions within the blocks, not just the GOE. In principle, one can then derive a corresponding state evolution formula mechanically for non-identically distributed orthogonally invariant blocks with arbitrary free cumulants, although the resulting expression is quite complicated.

Second, for technical reasons, we assumed that the blocks are square and symmetric, so that we could work with undirected graphs. The results of [Mal20, CDM24] extend to general matrices, and our techniques should also extend to the setting of varying block sizes and asymmetric matrices, leading to non-uniform mixtures in the recursion for the covariance kernel.

One caveat of the treelike AMP algorithm is that the Onsager correction term in Eq. (36) is not obviously efficient to compute in practice.[14] On the other hand, the vectors $\boldsymbol{b}_{s,t}$ have asymptotically constant entries in many settings, so that the Onsager correction can be replaced by a simpler asymptotically equivalent term, like in Theorems 6.28 and 6.29. This should also hold for block-structured models, as in the generalized AMP algorithm of Javanmard and Montanari [JM13]. For example, Eq. (46) is expected to be asymptotically equivalent to:

$$\boldsymbol{x}_0 = \mathbf{1}\,, \qquad \boldsymbol{x}_t = \boldsymbol{A} f_{t-1}(\boldsymbol{x}_{t-1}) - \boldsymbol{b}_{t-2,t} \cdot f_{t-2}(\boldsymbol{x}_{t-2})\,,$$
$$\boldsymbol{b}_{t-2,t}[i] = \sum_{c=1}^{q} \boldsymbol{\Sigma}[\mathrm{block}(i), c] \langle f'(x_{t-1}) \cdot \mathbf{1}_{\mathrm{block}=c} \rangle\,,$$

where $\mathbf{1}_{\mathrm{block}=c} \in \{0,1\}^n$ indicates the entries in block $c \in [q]$. The treelike AMP algorithm for Theorem 6.33 is expected to be asymptotically equivalent to:

$$\boldsymbol{x}_0 = \mathbf{1}\,, \qquad \boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{f}_{t-1} - \langle \boldsymbol{f}'_{t-1} \rangle \boldsymbol{f}_{t-2} - \sum_{\substack{s=0 \\ s \neq t-2}}^{t-1} \kappa_{t-s} \left( \prod_{r=s+1}^{t-1} \langle \boldsymbol{f}'_r \cdot \mathbf{1}_{\mathrm{block}=1} \rangle \right) \boldsymbol{f}_s \cdot \mathbf{1}_{\mathrm{block}=1}\,,$$

$$\boldsymbol{f}_t := f_t(\boldsymbol{x}_t)\,, \qquad \boldsymbol{f}'_t := f'_t(\boldsymbol{x}_t)\,,$$

where $\mathbf{1}_{\mathrm{block}=1} \in \{0,1\}^n$ indicates the entries in the first block. Because these expressions involve the blockwise indicators $\mathbf{1}_{\mathrm{block}=c}$, they could be represented and analyzed using an extended diagram basis in which certain indices are constrained to lie in a prescribed block. We leave the full development of this extension to future work.

A final open question is to characterize traffic distributions satisfying the (not necessarily factorizing) strong cactus property. Sequences of block matrices with orthogonally invariant blocks provide one general construction of matrices with the strong cactus property. If a sequence of matrices has the strong cactus property, must its traffic distribution arise as the limit (in an appropriate sense) of traffic distributions of block matrices with orthogonally invariant blocks (allowing the number of blocks to tend to infinity)?

[14] The $\boldsymbol{b}_{s,t}$ can be approximated with high probability to negligible error for all $0 \le s < t \le T$ in time $2^{O(T)}\,\mathrm{poly}(n)$ using the color coding technique [AYZ95, DSS20], but the exponential dependence on $T$ makes this algorithm impractical to implement for large $T$.

# References


[ABKZ20] Alia Abbara, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. On the universality of noiseless linear estimation with respect to the measurement matrix. *Journal of Physics A: Mathematical and Theoretical*, 53(16):164001, 2020. 6, 12

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2010. 24

[AYZ95] Noga Alon, Raphael Yuster, and Uri Zwick. Color-coding. *Journal of the ACM*, 42:844–856, 1995. 65

[Ban10] Teodor Banica. The orthogonal Weingarten formula in compact form. *Letters in Mathematical Physics*, 91(2):105–118, 2010. 76

[BHX25] Zhigang Bao, Qiyang Han, and Xiaocong Xu. A leave-one-out approach to approximate message passing. *Annals of Applied Probability*, 35(4):2716–2766, 2025. 12

[Bil95] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, third edition, 1995. 86

[BIPZ78] Edouard Brézin, Claude Itzykson, Giorgio Parisi, and Jean-Bernard Zuber. Planar diagrams. *Communications in Mathematical Physics*, 59:35–51, 1978. 70

[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. 3, 11, 88

[BM08] Adrian Bondy and U.S.R. Murty. *Graph Theory*. Springer, 2008. 33

[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. 1, 11

[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. 11

[BS10] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010. 32

[BSK15] Jean Barbier, Christophe Schulke, and Florent Krzakala. Approximate message-passing with spatially coupled structured operators, with applications to compressed sensing and sparse superposition codes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015. 3, 6, 11

[CDM24] Guillaume Cébron, Antoine Dahlqvist, and Camille Male. Traffic distributions and independence II: Universal constructions for traffic spaces. *Documenta Mathematica*, 29(1):39–114, 2024. 7, 25, 65, 75, 78, 84

[CGH$^+$17] Jordan S. Cotler, Guy Gur-Ari, Masanori Hanada, Joseph Polchinski, Phil Saad, Stephen H. Shenker, Douglas Stanford, Alexandre Streicher, and Masaki Tezuka. Black holes and random matrices. *Journal of High Energy Physics*, 2017(5):1–54, 2017. 70

[CHS24] Nicola Muca Cirone, Jad Hamdan, and Cristopher Salvi. Genus expansion for non-linear random matrix ensembles with applications to neural networks. *arXiv preprint arXiv:2407.08459*, 2024. 1

[CL21] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021. 11

[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. 3

[CMW20] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory (COLT 2020)*, pages 1078–1141. PMLR, 2020. 1, 2, 11

[ÇO19] Burak Çakmak and Manfred Opper. Memory-free dynamics for the Thouless-Anderson-Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices. *Physical Review E*, 99(6):062140, 2019. 6, 12

[CŚ06] Benoît Collins and Piotr Śniady. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics*, 264(3):773–795, 2006. 76

[DFGZJ95] Philippe Di Francesco, Paul Ginsparg, and Jean Zinn-Justin. 2D gravity and random matrices. *Physics Reports*, 254(1-2):1–133, 1995. 70

[DJM13] David L. Donoho, Adel Javanmard, and Andrea Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, 2013. 11

[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. 3, 6, 11, 12, 29

[DMM09] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009. 1, 3, 11

[DMR14] Yash Deshpande, Andrea Montanari, and Emile Richard. Cone-constrained principal component analysis. In *Advances in Neural Information Processing Systems (NIPS 2014)*, pages 2717–2725, 2014. 2

[DSL24] Rishabh Dudeja, Subhabrata Sen, and Yue M. Lu. Spectral universality in regularized linear regression with nearly deterministic sensing matrices. *IEEE Transactions on Information Theory*, 2024. 11

[DSS20] Jingqiu Ding, Hopkins Samuel, and David Steurer. Estimating Rank-One Spikes from Heavy-Tailed Noise via Self-Avoiding Walks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 5576–5586, 2020. 65

[EM03] Nicholas M. Ercolani and Kenneth McLaughlin. Asymptotics of the partition function for random matrices via Riemann-Hilbert techniques and applications to graphical enumeration. *International Mathematics Research Notices*, 2003:755–820, 2003. 75

[EMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *Annals of Probability*, 49(6):2922–2960, 2021. 2

[Eti24] Pavel Etingof. Mathematical ideas and notions of quantum field theory. *arXiv preprint arXiv:2409.03117*, 2024. 72

[Fan22] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *Annals of Statistics*, 50(1):197–224, 2022. 3, 10, 12, 59

[FVRS22] Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A Unifying Tutorial on Approximate Message Passing. *Foundations and Trends in Machine Learning*, 15(4):335–536, 2022. 1, 3, 11

[GHN26] Mohammed-Younes Gueddari, Walid Hachem, and Jamal Najim. Approximate Message Passing for General Non-Symmetric Random Matrices. *Journal of Theoretical Probability*, 39, 2026. 12

[GP13] Stavros Garoufalidis and Ionel Popescu. Analyticity of the planar limit of a matrix model. *Annales Henri Poincaré*, 14:499–565, 2013. 75

[GTM+24] Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous Dynamical Mean-Field Theory for Stochastic Gradient Descent Methods. *SIAM Journal on Mathematics of Data Science*, 6:400–427, 2024. 1

[Hol81] Ian Holyer. The NP-completeness of some edge-partition problems. *SIAM Journal on Computing*, 10(4):713–717, 1981. 75

[IS24] Misha Ivkov and Tselil Schramm. Semidefinite programs simulate approximate message passing robustly. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC 2024)*, pages 348–357. ACM, 2024. 11

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

[JM12] Adel Javanmard and Andrea Montanari. Subsampling at information theoretically optimal rates. In *Proceedings of the International Symposium on Information Theory (ISIT 2012)*, pages 2431–2435. IEEE, 2012. 6

[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. 3, 10, 11, 25, 62, 65

[JP25] Chris Jones and Lucas Pesenti. Fourier analysis of iterative algorithms. In *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, pages 102–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2025. 11, 24, 26, 92

[KMW24] Dmitriy Kunisky, Cristopher Moore, and Alexander S. Wein. Tensor cumulants for statistical inference on invariant distributions. In *65th Annual Symposium on Foundations of Computer Science (FOCS 2024)*, pages 1007–1026, 2024. 80

[LW22] Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*, 2022. 11

[LWF25] Max Lovig, Tianhao Wang, and Zhou Fan. On universality of non-separable approximate message passing algorithms. *arXiv preprint arXiv:2506.23010*, 2025. 3

[Mal20] Camille Male. *Traffic distributions and independence: permutation invariant random matrices and the three notions of independence*, volume 267:1300. American mathematical society, 2020. 5, 7, 23, 24, 26, 27, 65

[MFC+19] Antoine Maillard, Laura Foini, Alejandro Lage Castellanos, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. High-temperature expansions and message passing algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(11):113301, 2019. 7, 12, 19

[MM09] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. 3

[Mon12] Andrea Montanari. Graphical models concepts in compressed sensing. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 394–438. Cambridge University Press, 2012. 1, 11

[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. 2

[MP17] Junjie Ma and Li Ping. Orthogonal AMP. *IEEE Access*, 5:2020–2033, 2017. 12

[MPR94a] Enzo Marinari, Giorgio Parisi, and Felix Ritort. Replica field theory for deterministic models: I. binary sequences with low autocorrelation. *Journal of Physics A: Mathematical and General*, 27(23):7615, 1994. 6

[MPR94b] Enzo Marinari, Giorgio Parisi, and Felix Ritort. Replica field theory for deterministic models. II. A non-random spin glass with glassy behaviour. *Journal of Physics A: Mathematical and General*, 27(23):7647, 1994. 6

[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. 3

[MR15] Andrea Montanari and Emile Richard. Non-negative principal component analysis: message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3):1458–1484, 2015. 2

[MS12] James A. Mingo and Roland Speicher. Sharp bounds for sums associated to graphs of matrices. *Journal of Functional Analysis*, 262(5):2272–2288, 2012. 32

[MW24] Andrea Montanari and Yuchen Wu. Statistically optimal first-order algorithms: a proof via orthogonalization. *Information and Inference: A Journal of the IMA*, 13, 2024. 1

[MW25] Andrea Montanari and Alexander S. Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *Probability Theory and Related Fields*, 191:181–233, 2025. 2, 11

[NS06] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006. 77

[OÇW16] Manfred Opper, Burak Çakmak, and Ole Winther. A theory of solving TAP equations for Ising models with general invariant random matrices. *Journal of Physics A: Mathematical and Theoretical*, 49(11):114002, 2016. 12

[Pes26] Lucas Pesenti. *Algorithms beyond the union bound: polynomial optimization and discrepancy theory*. PhD thesis, Bocconi University, 2026. 2

[Pet82] L.C. Petersen. On the relation between the multidimensional moment problem and the one-dimensional moment problem. *Mathematica Scandinavica*, 51:361–366, 1982. 86

[PP95] Giorgio Parisi and Marc Potters. Mean-field equations for spin models with orthogonal interaction matrices. *Journal of Physics A: Mathematical and General*, 28(18):5267, 1995. 6, 7

[Ran11] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symposium on Information Theory (ISIT 2011)*, pages 2168–2172. IEEE, 2011. 1, 10, 62

[RFSK16] Sundeep Rangan, Alyson K. Fletcher, Philip Schniter, and Ulugbek S. Kamilov. Inference for generalized linear models via alternating directions and Bethe free energy minimization. *IEEE Transactions on Information Theory*, 63(1):676–697, 2016. 1

[Rob39] Herbert Ellis Robbins. A theorem on graphs, with an application to a problem of traffic control. *The American Mathematical Monthly*, 46(5):281–283, 1939. 33

[Rot64] Gian-Carlo Rota. On the foundations of combinatorial theory: I. Theory of Möbius functions. *Probability Theory and Related Fields*, 2(4):340–368, 1964. 18

[RSF19] Sundeep Rangan, Philip Schniter, and Alyson K. Fletcher. Vector approximate message passing. *IEEE Transactions on Information Theory*, 65(10):6664–6684, 2019. 10, 12

[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

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

[Sch20] Philip Schniter. A simple derivation of AMP and its state evolution via first-order cancellation. *IEEE Transactions on Signal Processing*, 68:4283–4292, 2020. 6, 12

[SSS19] Phil Saad, Stephen H. Shenker, and Douglas Stanford. JT gravity as a matrix integral. *arXiv preprint arXiv:1903.11115*, 2019. 70

[Tak19] Keigo Takeuchi. Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Transactions on Information Theory*, 66(1):368–386, 2019. 12

[tH74] Gerard 't Hooft. A planar diagram theory for strong interactions. *Nuclear Physics B*, 72(3):461–473, 1974. 70, 73

[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. 3

[WZF22] Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *Annals of Applied Probability*, 34:3943–3994, 2022. 3, 6, 7, 10, 11, 12, 28, 29

[ZK16] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. 3

[ZWF24] Xinyi Zhong, Tianhao Wang, and Zhou Fan. Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *Information and Inference: A Journal of the IMA*, 13(3):iaae024, 2024. 3, 29

# A Traffic Distributions via Feynman Diagrams

One of our motivations is to connect graph polynomials with the celebrated Feynman diagram technique from physics. In quantum field theory, Feynman diagram expansion is used to reduce matrix integrals into graphical calculations. We show in this section that this method can (heuristically) derive the traffic distribution of orthogonally invariant distributions (Theorem 4.2).

The matrix model that we consider in this section is specified by a *potential function* $V : \mathbb{R} \to \mathbb{R}$, and has partition function

$$Z := \int_{\mathcal{M}} \mathrm{d}\boldsymbol{A}\ e^{-\frac{n}{2}\operatorname{Tr} V(\boldsymbol{A})}, \tag{47}$$

where $\mathcal{M} := \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is the space of symmetric $n\times n$ matrices. Equivalently, this is the partition function of the random matrix $\boldsymbol{A} \in \mathcal{M}$ sampled from the probability measure $\mu_V(\boldsymbol{A}) \propto \exp(-\frac{n}{2}\operatorname{Tr} V(\boldsymbol{A}))$, which is a special case of an orthogonally invariant distribution (Section 4.2).

In physics, matrix integrals such as Eq. (47) are viewed as a 0-dimensional theory: the variable is a matrix, and the partition function is a finite-dimensional integral rather than a functional integral over fields on space-time. The large-$n$ expansion of such integrals is organized into diagrammatic contributions indexed by *Feynman diagrams*.

1. In the limit $n \to \infty$, only planar diagrams contribute at leading order, an observation going back to foundational work of 't Hooft [tH74, BIPZ78]. Related planarity phenomena also appear in mathematics, for example in the connections between large random matrices and non-crossing pairings.

2. In special scaling limits of the potential with $n$, the Feynman diagram expansion can be interpreted in terms of physical theories such as 2D gravity and certain string-theoretic models [DFGZJ95, CGH$^+$17, SSS19].

The combinatorial approach in this paper fits naturally into this perspective. First, our results are formulated in the large-$n$ limit, and the dominant combinatorial objects in that limit are planar, as in the 't Hooft limit. Second, we show that our $w$- and $z$-polynomials are planar dual to the Feynman diagrams traditionally used in physics. Third, while the Feynman diagram method is based on perturbative expansion around the GOE potential $V(x) = x^2/2$, our rigorous results Theorems 4.2 and 6.2 still remain valid beyond the radius of convergence for perturbative methods.

We present in this section the traditional approach for computing Eq. (47) based on Feynman diagrams. The argument is "combinatorially rigorous" (true at the level of generating functions), but not sufficient to rigorously derive the probabilistic conclusions.

## A.1 Calculation of the free energy

For now, we restrict to the case where the potential in Eq. (47) is $V(\boldsymbol{A}) = \frac{1}{2}\boldsymbol{A}^2 + \frac{g}{4}\boldsymbol{A}^4$, where the *coupling constant* $g$ measures the strength of the quartic interaction in the model. Such potentials appear in string theory, statistical physics (the $\lambda\phi^4$ theory), and the theory of integrable systems. The quartic term $\frac{g}{4}\operatorname{Tr}(\boldsymbol{A}^4)$ can be viewed as a correction term to the $\mathsf{GOE}$ model, for which $Z_{\mathsf{GOE}} = \int_{\mathcal{M}} \mathrm{d}\boldsymbol{A} \exp(-n\operatorname{Tr}(\boldsymbol{A}^2)/4)$.

The idea of the Feynman diagram technique is to perturbatively expand this correction term, reducing to a problem on Gaussian variables. We illustrate this by computing the free energy of the

quartic model, namely the quantity $\ln Z$ (this example can be found in physics textbooks). For an observable quantity $\mathcal{O}$, we write $\langle \mathcal{O} \rangle := \mathbb{E}_{\boldsymbol{A} \sim \mu_V}[\mathcal{O}]$, and $\langle \mathcal{O} \rangle_{\mathsf{GOE}} := \mathbb{E}_{\boldsymbol{A} \sim \mathsf{GOE}}[\mathcal{O}]$. We have

$$\begin{aligned} Z &= \int_{\mathcal{M}} \mathrm{d}\boldsymbol{A} \exp\Big(-\frac{n}{4}\operatorname{Tr}\big(\boldsymbol{A}^2\big) - \frac{gn}{8}\operatorname{Tr}\big(\boldsymbol{A}^4\big)\Big) \\ &= Z_{\mathsf{GOE}} \cdot \Big\langle \exp\Big(-\frac{gn}{8}\operatorname{Tr}\big(\boldsymbol{A}^4\big)\Big)\Big\rangle_{\mathsf{GOE}} . \end{aligned}$$

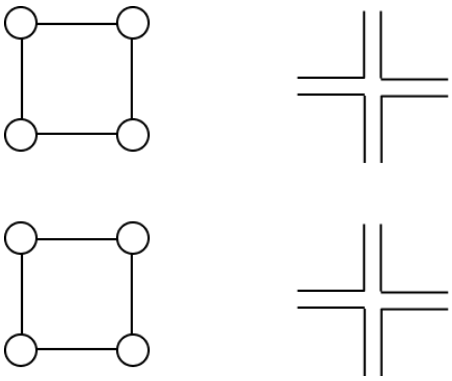
A simple calculation shows that $Z_{\mathsf{GOE}} = 2^{\frac{n}{2}} \left(\frac{2\pi}{n}\right)^{\frac{n(n+1)}{4}}$. We Taylor expand the remaining part and integrate term-by-term:

$$\Big\langle \exp\Big(-\frac{gn}{8}\operatorname{Tr}\big(\boldsymbol{A}^4\big)\Big)\Big\rangle_{\mathsf{GOE}} = \sum_{s=0}^{\infty} \frac{1}{s!}\Big(-\frac{gn}{8}\Big)^s \Big\langle \operatorname{Tr}\big(\boldsymbol{A}^4\big)^s \Big\rangle_{\mathsf{GOE}} . \tag{48}$$

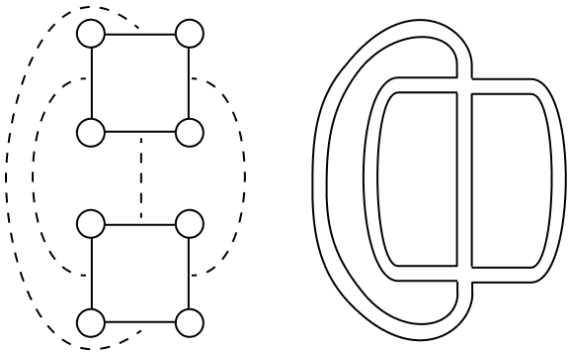
The quantities $\langle \operatorname{Tr}(\boldsymbol{A}^4)^s \rangle_{\mathsf{GOE}}$ on the right-hand side are expectations over Gaussian random variables, and can be computed by Wick's lemma (Lemma 2.8) to be a sum over all *Wick contractions* between the variables (in graph-theoretic terms, a sum over all perfect matchings). The *propagator* for a single contraction with a $\mathsf{GOE}$ matrix is the covariance of the Gaussians,

$$\langle \boldsymbol{A}[i,j]\boldsymbol{A}[k,\ell]\rangle_{\mathsf{GOE}} = \frac{1}{n}\delta_{ik}\delta_{j\ell} + \frac{1}{n}\delta_{i\ell}\delta_{jk} , \quad \text{where } \delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} . \tag{49}$$

A *Feynman diagram* represents a combinatorial type of Wick contractions. In the graphical notations of this paper, we would visualize each $\operatorname{Tr}(\boldsymbol{A}^4)$ as a square, with Wick contraction having the effect of gluing together edges of the squares. The 't Hooft double line notation, which is more common in physics, represents each $\operatorname{Tr}(\boldsymbol{A}^4)$ as a vertex with four incident double edges. These representations are dual to each other (in the sense of planar duality); see Fig. 5 for comparison.



(a) $\operatorname{Tr}(\boldsymbol{A}^4)^2$ represented as two squares in our notation, compared to the 't Hooft double line notation.



(b) One of the Wick contractions appearing in $\langle \operatorname{Tr}(\boldsymbol{A}^4)^2 \rangle_{\mathsf{GOE}}$. The edges of the squares are glued together according to the matching to make a "pillow".

Figure 5: Our Feynman diagram notation vs. the 't Hooft double line notation.

The delta functions in the propagator enforce that the vertices of the squares have a consistent index $i$ when the edges of the squares are glued together. Note that the propagator in Eq. (49) for

the GOE model allows $\boldsymbol{A}[i,j], \boldsymbol{A}[k,\ell]$ to be glued in either orientation (in contrast to the Gaussian Unitary Ensemble which would only have one term). Therefore, we define a Feynman diagram for the GOE to be an *oriented* perfect matching between the edges of the squares. For each Feynman diagram $\gamma$, the contribution of $\gamma$ to Eq. (48) is:

(i) a factor $n$ per vertex of $\gamma$, since each vertex holds an index from $[n]$ which is summed over in $\mathrm{Tr}(\boldsymbol{A}^4)$.

(ii) a factor $\frac{1}{n}$ per paired edge of $\gamma$ from the propagator, Eq. (49).

(iii) a factor $-\frac{gn}{8}$ per square face of $\gamma$ from Eq. (48). There is also an overall factor of $\frac{1}{|F(\gamma)|!}$ where $|F(\gamma)|$ equals the number of square faces in $\gamma$.

For example, the $s = 1$ term in Eq. (48) is

$$\left(-\frac{gn}{8}\right) \cdot \langle \mathrm{Tr}\left(\boldsymbol{A}^4\right) \rangle_{\mathsf{GOE}} = \left(-\frac{g}{8}\right) \cdot \left(2 \cdot n^2 + 5 \cdot n + 5\right) .$$

The Feynman diagrams are enumerated in Fig. 6.

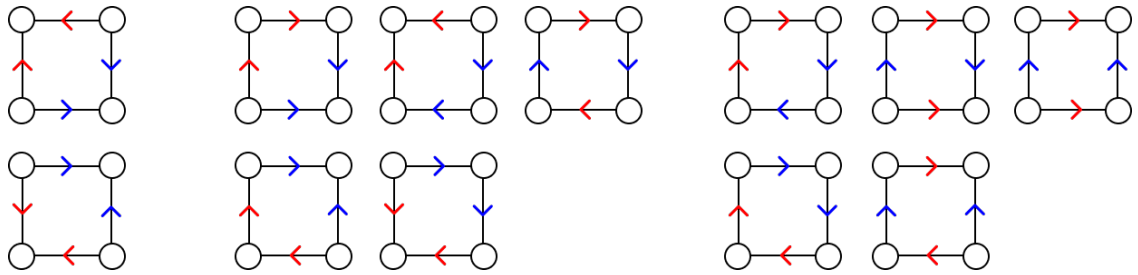


Figure 6: The 12 Feynman diagrams associated to $\langle \mathrm{Tr}(\boldsymbol{A}^4) \rangle_{\mathsf{GOE}}$. The two red edges are matched in the orientation specified by the arrows, and similarly for the blue edges. Gluing either of the left two diagrams results in a "taco". Gluing the remaining diagrams results in degenerate polyhedra. After gluing, the "tacos" have 3 vertices, the middle diagrams have 2 vertices, and the right diagrams have 1 vertex, respectively.

For a given Feynman diagram $\gamma$, the total factor of $n$ is $|V(\gamma)| - |E(\gamma)| + |F(\gamma)| =: \chi(\gamma)$ which is the *Euler characteristic* of the polyhedron $\gamma$. In total, we obtain a Feynman diagram expansion for the partition function,

$$Z = Z_{\mathsf{GOE}} \sum_{\gamma \in \Gamma} \frac{1}{|F(\gamma)|!} \left(-\frac{g}{8}\right)^{|F(\gamma)|} n^{\chi(\gamma)} , \tag{50}$$

where $\Gamma$ is the set of Feynman diagrams, the set of polyhedra built from square faces. Formally, $\Gamma = \sqcup_{s \geq 0} \Gamma_s$, where $\Gamma_s$ is the set of oriented perfect matchings between the edges of $s$ squares.

Taking the logarithm has the effect of restricting the summation to *connected* Feynman diagrams; this is the linked cluster theorem in quantum field theory [Eti24, Section 3.5]. We obtain:

$$\ln\left(\frac{Z}{Z_{\mathsf{GOE}}}\right) = \sum_{\gamma \in \Gamma_c} \frac{1}{|F(\gamma)|!} \left(-\frac{g}{8}\right)^{|F(\gamma)|} n^{\chi(\gamma)} , \tag{51}$$

where $\Gamma_c \subseteq \Gamma$ are connected Feynman diagrams.

### A.1.1 Asymptotic limit $n \to \infty$

As $n \to \infty$, Eq. (51) significantly simplifies because only the *planar* diagrams survive, i.e. polyhedra $\gamma$ with "no holes," which have the maximum possible Euler characteristic among connected graphs ($\chi(\gamma) = 2$). This foundational observation goes back to 't Hooft [tH74].[15] We obtain, at first order,

$$\frac{1}{n^2} \ln\left(\frac{Z}{Z_{\mathsf{GOE}}}\right) = \sum_{\substack{\gamma \in \Gamma_c \\ \text{planar}}} \frac{1}{|F(\gamma)|!} \left(-\frac{g}{8}\right)^{|F(\gamma)|} + O(n^{-2}) \,. \tag{52}$$

In summary, the Feynman diagram method shows that the non-Gaussian component of the matrix model can be replaced by a generating function for graphs/surfaces which, in the $n \to \infty$ limit, restricts to a generating function for planar graphs/surfaces with genus 0. This restriction leads to significant simplifications in diagrammatic calculations, in the same way as our cactus property and treelike property in the rest of the paper.

## A.2 Calculation of general observables: Argument for Theorem 4.2

We now assume that the potential $V(\boldsymbol{A})$ has the general form $V(\boldsymbol{A}) = \frac{1}{2}\boldsymbol{A}^2 + \sum_{j \geq 3} c_j \boldsymbol{A}^j$ (arbitrary coefficients on $\boldsymbol{A}$ and $\boldsymbol{A}^2$ can be handled by centering and rescaling, respectively). We compute the traffic distribution of $\boldsymbol{A}$, which consists of all $S_n$-invariant observables of $\boldsymbol{A}$. The $z$-polynomials are a basis for these observables where, for each multigraph $\alpha$,

$$\frac{1}{n}\langle z_\alpha(\boldsymbol{A})\rangle = \frac{1}{n} \sum_{i : V(\alpha) \hookrightarrow [n]} \left\langle \prod_{\{u,v\} \in E(\alpha)} \boldsymbol{A}[i(u), i(v)] \right\rangle .$$

Separating out the Gaussian part of the action from the higher-order interactions:

$$\frac{1}{n}\langle z_\alpha(\boldsymbol{A})\rangle = \frac{1}{n} \cdot \frac{\left\langle z_\alpha(\boldsymbol{A}) \exp\left(-\sum_{j \geq 3} c_j n \operatorname{Tr}(\boldsymbol{A}^j)\right)\right\rangle_{\mathsf{GOE}}}{\left\langle \exp\left(-\sum_{j \geq 3} c_j n \operatorname{Tr}(\boldsymbol{A}^j)\right)\right\rangle_{\mathsf{GOE}}} .$$

The dual Feynman diagrams are built from polygons with $j \geq 3$ sides, each of which comes with a factor of $-c_j$, generalizing the situation from the previous section. A small generalization of the argument shows that the denominator is

$$\left\langle \exp\left(-\sum_{j \geq 3} c_j n \operatorname{Tr}\left(\boldsymbol{A}^j\right)\right)\right\rangle_{\mathsf{GOE}} = \sum_{\gamma \in \Gamma} \left(\prod_{j \geq 3} \frac{(-c_j)^{|F_j(\gamma)|}}{|F_j(\gamma)|!}\right) n^{\chi(\gamma)} , \tag{53}$$

where $|F_j(\gamma)|$ denotes the number of $j$-sided faces in $\gamma$.

The numerator can also be calculated diagrammatically. The Wick contractions go between a collection of polygons as well as the additional edges $\boldsymbol{A}[i,j]$ in $z_\alpha(\boldsymbol{A})$. Let $\Gamma(\alpha)$ be the set of

[15] 't Hooft studies unitarily invariant matrix models instead of orthogonally invariant ones. He takes a further step by sending $g \to 0$ at the rate $\Theta(1/\sqrt{n})$, i.e., fixing $\lambda = g^2 n$ to be constant. His claim is that $\lambda$ is the only parameter characterizing the physical properties of observables in the large-$n$ limit, and by taking $\lambda \to \infty$ one gains some intuition on the physical phenomena of strongly interacting particles. The limit $g \to 0$ is less interesting for us, since the traffic distribution (hence also the spectrum) is asymptotically the same as the GUE whenever $g = o(1)$.

Feynman diagrams, visualized as polyhedra built on a set of "boundary" edges $\alpha$. Then

$$\left\langle z_\alpha(\boldsymbol{A}) \exp\left(-\sum_{j\geq 3} c_j n \operatorname{Tr}\left(\boldsymbol{A}^j\right)\right)\right\rangle_{\mathsf{GOE}} = \sum_{\gamma\in\Gamma(\alpha)} \left(\prod_{j\geq 3} \frac{(-c_j)^{|F_j(\gamma)|}}{|F_j(\gamma)|!}\right) n^{\chi(\gamma)} \cdot (1 - O(n^{-1}))\,. \tag{54}$$

Note $\alpha$ is considered a boundary and does not count towards the faces $F_j(\gamma)$.

To enforce that the labels of $z_\alpha(\boldsymbol{A})$ are injective, we remove from $\Gamma(\alpha)$ any matching which causes two vertices of $\alpha$ to have the same label. The factor $1 - O(n^{-1})$ arises because each vertex is summed over $n - O(1)$ indices to maintain injectivity, instead of precisely $n$ which we had previously.

We obtain the final result by dividing Eq. (54) by Eq. (53). This has the effect of restricting to the set of connected Feynman diagrams $\Gamma_c(\alpha) \subseteq \Gamma(\alpha)$ by an alternate version of the linked cluster theorem. The final Feynman diagram formula is:

$$\frac{1}{n}\langle z_\alpha(\boldsymbol{A})\rangle = \sum_{\gamma\in\Gamma_c(\alpha)} \left(\prod_{j\geq 3} \frac{(-c_j)^{|F_j(\gamma)|}}{|F_j(\gamma)|!}\right) \cdot n^{\chi(\gamma)-1} \cdot (1 - O(n^{-1}))\,. \tag{55}$$

**Remark A.1.** *An alternative approach to the calculation would be to first symmetrize $z_\alpha(\boldsymbol{A})$ over $O(n)$ which is the symmetry group of the matrix model (and is larger than $S_n$), then to plug in the values of the $O(n)$-invariant observables (the trace polynomials). We find it simpler to Taylor expand the action directly.*

### A.2.1 Asymptotic limit $n \to \infty$

In the asymptotic limit $n \to \infty$, the only diagrams in Eq. (55) with constant-order magnitude are those such that $\alpha$ is a cactus graph, and $\gamma$ consists of polyhedra with genus 0 attached to each cycle of the cactus, which has $\chi(\gamma) = 1$. We prove this combinatorially in the forthcoming Lemma A.2.

The large-$n$ combinatorial summation factors over the cycles of the cactus, since the genus-0 polyhedra on each cycle can be chosen independently. We obtain

$$\frac{1}{n}\langle z_\alpha(\boldsymbol{A})\rangle = \begin{cases} \displaystyle\prod_{\sigma\in\text{cycles}(\alpha)} \frac{1}{n}\langle z_\sigma(\boldsymbol{A})\rangle + O(n^{-1}) & \text{if } \alpha \text{ is a cactus} \\ O(n^{-1}) & \text{otherwise} \end{cases}$$

The limiting value $\frac{1}{n}\langle z_\sigma(\boldsymbol{A})\rangle$ of the $q$-cycle diagram $\sigma$ is equal to $\kappa_q + O(n^{-1})$ by the moment/free cumulant relation Eq. (8). Thus, Eq. (55) recovers Theorem 4.2.

**Lemma A.2.** *Let $\alpha \in \mathcal{A}$ be a connected multigraph, and let $\gamma \in \Gamma_c(\alpha)$. Then $\chi(\gamma) = 1$ if and only if $\alpha$ is a cactus and $\gamma$ consists of genus-0 polyhedra attached to each cycle of $\alpha$.*

*Proof.* The only $\alpha$ for which $\Gamma_c(\alpha)$ is nonzero are the Eulerian $\alpha$, since a polyhedron $\gamma \in \Gamma_c(\alpha)$ with boundary $\alpha$ must have a boundary which is a union of cycles. Therefore, it remains to argue about Eulerian graphs $\alpha$.

For Eulerian $\alpha$, the $\gamma \in \Gamma_c(\alpha)$ which maximize the quantity $\chi(\gamma) = |V(\gamma)| - |E(\gamma)| + |F(\gamma)|$ are given by decomposing $\alpha$ into the maximum number of simple cycles, then attaching a genus

0 polyhedron to each cycle. This achieves $\chi(\gamma) = |V(\alpha)| - |E(\alpha)| + C$ where $C$ is the number of cycles.[16]

We argue that:

$$|V(\alpha)| - |E(\alpha)| + C \leq 1 \tag{56}$$

for all Eulerian graphs $\alpha$ and this is achieved if and only if $\alpha$ is a cactus. Fix a maximum cycle partition of $\alpha$. The $C$ cycles are edge-disjoint so we can remove one edge from each one while maintaining that the graph is connected. Let $\alpha'$ be the resulting graph. Then $|V(\alpha')| - |E(\alpha')| = |V(\alpha)| - |E(\alpha)| + C$. Since $\alpha'$ is still connected we have $|V(\alpha')| - |E(\alpha')| \leq 1$. The final inequality is an equality if and only if $\alpha'$ is a tree and hence $\alpha$ is a cactus. This proves Eq. (56) and completes the lemma. □

### A.3 Mathematical comments on the Feynman diagram method

The Feynman diagram method is not mathematically rigorous, with (in our opinion) the main obstruction being that intermediate summations such as Eqs. (50), (51) and (54) are divergent. The Euler characteristic grows with the number of disconnected polyhedra, but the method proceeds anyway to divide out the disconnected polyhedra, which ultimately yields a convergent summation in Eq. (52) (for sufficiently small values of the coupling constant $g \geq 0$).

The Feynman diagram method is a perturbative expansion because it holds for sufficiently small perturbations of the GOE density, up to the radius of convergence of the Feynman diagram summations [EM03, GP13]. On the other hand, Theorem 4.2 holds beyond the radius of convergence of the Feynman diagram expansion in Eq. (55), so it would be impossible to prove the theorem using a perturbative expansion alone.

## B Traffic Distributions via Weingarten Calculus

We now present different tools and calculations for the traffic distributions of orthogonally invariant matrices based on the *Weingarten formula* for the moments of entries of Haar-random orthogonal matrices. These essentially follow the ideas of similar calculations by [CDM24], but use the version of the Weingarten formula for the orthogonal group, which we review below.

### B.1 Weingarten formula for orthogonal matrices

For $\boldsymbol{i} = (i_1, \ldots, i_k)$ and a perfect matching $\alpha \in \mathcal{M}_{\text{perf}}([k])$, define

$$\delta_\alpha(\boldsymbol{i}) = \begin{cases} 1 & \text{if } i_u = i_v \text{ for all } \{u, v\} \in \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

The Weingarten calculus expresses the moments of the Haar measure on $O(n)$ in terms of a certain "Weingarten function" $W_n(\alpha, \beta)$ on pairs of matchings.

[16] Note that the computational problem of, given an Eulerian graph $\alpha$, compute a partition of $E(\alpha)$ into the maximum number of cycles, is NP-hard [Hol81].

**Lemma B.1** (Weingarten formula)**.** *Let $\boldsymbol{Q} \sim O(n)$ be a Haar-random orthogonal matrix. There exists a function $W_n : \mathcal{M}_{\mathrm{perf}}([k])^2 \to \mathbb{R}$ such that*

$$\mathop{\mathbb{E}}_{\boldsymbol{Q}\sim O(n)} [\boldsymbol{Q}[i_1, j_1] \cdots \boldsymbol{Q}[i_k, j_k]] = \sum_{\alpha,\beta\in\mathcal{M}_{\mathrm{perf}}([k])} W_n(\alpha, \beta)\delta_\alpha(\boldsymbol{i})\delta_\beta(\boldsymbol{j}).$$

See [CŚ06, Ban10] for an explicit definition of $W_n(\alpha, \beta)$. We will only be interested in asymptotics for $k$ constant and $n \to \infty$, for which the approximations below will suffice.

When $k$ is odd, $\mathcal{M}_{\mathrm{perf}}([k]) = \varnothing$, so the right-hand side above is zero, and indeed the left-hand side is easily seen to be zero without invoking the Weingarten formula, because $\boldsymbol{Q}$ has the same law as $-\boldsymbol{Q}$. So, the only interesting case is $k$ even. In that case, we give $\mathcal{M}_{\mathrm{perf}}([k])$ the structure of a metric space, where $\Delta(\alpha, \beta)$ is defined as the minimum number of *swap operations* needed to reach $\beta$ from $\alpha$ (a swap replaces pairs $\{a, b\}$, $\{c, d\}$ with pairs $\{a, c\}$, $\{b, d\}$). It is easy to check that $\Delta$ is a metric (indeed, it is the distance on a certain graph structure defined on $\mathcal{M}_{\mathrm{perf}}([k])$). Further, write $\mathrm{cyc}(\alpha, \beta)$ for the set of even cycles formed by the disjoint union of $\alpha$ and $\beta$. Then, it is easy to show the alternative characterization

$$\Delta(\alpha, \beta) = \frac{k}{2} - |\mathrm{cyc}(\alpha, \beta)|\,.$$

As a sanity check, $|\mathrm{cyc}(\alpha, \beta)| \le \frac{k}{2}$ with equality achieved if and only if $\alpha = \beta$, which is precisely the case $\Delta(\alpha, \beta) = 0$.

For $\alpha, \beta \in \mathcal{M}_{\mathrm{perf}}([k])$, let $\mathcal{P}(\alpha, \beta)$ be the set of geodesic paths from $p$ to $q$ in $\mathcal{M}_{\mathrm{perf}}([k])$, i.e., of sequences $\alpha = \gamma_0, \gamma_1, \ldots, \gamma_t = \beta$ with $\gamma_i \ne \gamma_{i+1}$ for all $i = 0, \ldots, t-1$ and with $\sum_{i=0}^{t-1} \Delta(\gamma_i, \gamma_{i+1}) = \Delta(\alpha, \beta)$. For such a path $P = (\gamma_0, \ldots, \gamma_t)$, write $|P| := t$. Then, we define

$$\mu(\alpha, \beta) = \sum_{P\in\mathcal{P}(\alpha,\beta)} (-1)^{|P|}$$

This may be viewed as a Möbius function of the partially ordered set whose chains are geodesics from a given "base" matching $p$ to each other matching. An explicit formula from [CŚ06] is

$$\mu(\alpha, \beta) = \prod_{C\in\mathrm{cyc}(\alpha,\beta)} (-1)^{\frac{|C|}{2}-1}\mathrm{Cat}\left(\frac{|C|}{2} - 1\right) \tag{57}$$

where $\mathrm{Cat}(\cdot)$ are the Catalan numbers. The key asymptotic for the Weingarten function for our purposes is then the following:

**Proposition B.2** ([CŚ06])**.** *For a fixed $k$ and $\alpha, \beta \in \mathcal{M}_{\mathrm{perf}}([k])$, as $n \to \infty$ we have*

$$W_n(\alpha, \beta) = n^{-k+\mathrm{cyc}(\alpha,\beta)}\left(\mu(\alpha, \beta) + O(n^{-1})\right)\,.$$

Note that the maximum possible scaling of this quantity is $n^{-k/2}$, which corresponds to the fact that with high probability the entries of $\boldsymbol{Q}$ are all roughly of order $n^{-1/2}$.

## B.2 Möbius inversion on non-crossing partitions

Recall that $\mathrm{NC}(k)$ is the partially ordered set of non-crossing partitions, i.e., those whose parts do not cross when drawn as a partition of vertices of the $k$-cycle. We review some standard properties of this partially ordered set; see, e.g., [NS06] for a standard reference.

Each non-crossing partition $\pi \in \mathrm{NC}(k)$ has a natural dual partition, called the *Kreweras complement* and denoted $K(\pi)$. On the cycle graph $C_k$, this may be viewed as the maximal non-crossing partition of the midpoints of the *edges* of $C_k$ that does not cross the boundaries of $\pi$. Alternatively, one may view both partitions as placed on a single cycle graph of twice the size, $C_{2k}$, on alternating sets of vertices. We show this viewpoint with an example in Fig. 7. The map $K : \mathrm{NC}(k) \to \mathrm{NC}(k)$ is easily checked to be an involution.

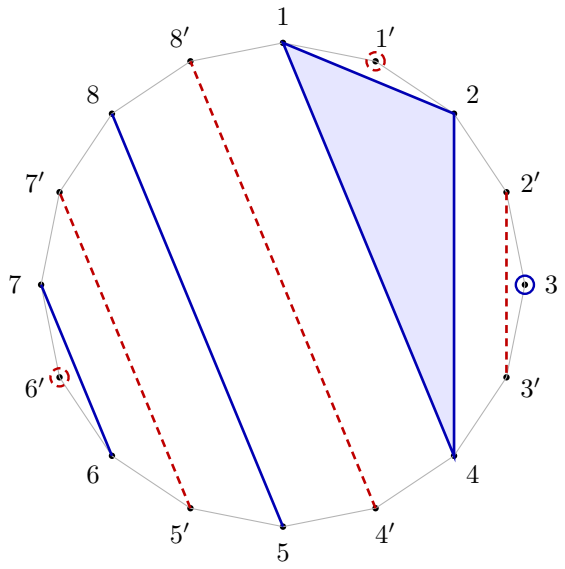



Figure 7: An illustration of the Kreweras complement operation on non-crossing partitions. The parts of a partition $\pi \in \mathrm{NC}(8)$ are drawn in blue, and the parts of the Kreweras complement $K(\pi) \in \mathrm{NC}(8)$ in red.

We give $\mathrm{NC}(k)$ the usual partial ordering of refinement of partitions, written $\pi \preceq \rho$, using that a refinement of a non-crossing partition remains non-crossing. This partial ordering has a minimal element $\underline{0} \in \mathrm{NC}(k)$, the partition where every block is a singleton, and a maximal element $\underline{1} \in \mathrm{NC}(k)$, the partition with just one block. The Kreweras complement is an *anti-isomorphism* of this ordering: it is a bijection that reverses the ordering, i.e. $K(\pi) \preceq K(\rho)$ if and only if $\pi \succeq \rho$. In particular, $K(\underline{0}) = \underline{1}$ and $K(\underline{1}) = \underline{0}$.

The Möbius function for the $\mathrm{NC}(k)$ poset gives values $\mu(\pi, \rho)$ for each pair $\pi \preceq \rho$. The Kreweras complement interacts with the Möbius function in the following way that will be crucial for our purposes:

$$\mu(\underline{0}, \pi) = \mu(K(\pi), \underline{1})\,. \tag{58}$$

Further, evaluations of the Möbius function as on the left-hand side may be expanded as products over the blocks of $\pi$, and the factors turn out to be the same as the combinatorial quantities appearing in Eq. (57); there is a combinatorial explanation for this coincidence but we will just need

to use that this indeed occurs:

$$\mu(\underline{0}, \pi) = \prod_{A \in \pi} (-1)^{|A|-1} \mathrm{Cat}(|A| - 1)\,.$$

Note that, applying Möbius inversion to Eq. (7), we obtain an explicit formula for the free cumulants in terms of the moments, as mentioned earlier in the main text: if $m_k$ are the moments of a probability measure, then the free cumulants $\kappa_k$ are

$$\kappa_k = \sum_{\pi \in \mathrm{NC}(k)} \mu(\pi, \underline{1}_k) \prod_{A \in \pi} m_{|A|}\,. \tag{59}$$

### B.3 Tracial moments concentration

The result of [CDM24] also assumes the following formula for the joint moments of the various trace powers of a matrix, that we also use in our proof. We show that it follows from our assumptions.

**Lemma B.3** (Tracial moments concentration)**.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$ be random matrices that converge in tracial moments in $L^2$ to some $\mu$. Then for any cycle diagrams $\rho_1, \dots, \rho_k$,*

$$\lim_{n \to \infty} \mathbb{E} \left[ \prod_{j=1}^{k} \frac{1}{n} w_{\rho_j}(\boldsymbol{A}) \right] = \prod_{j=1}^{k} \lim_{n \to \infty} \mathbb{E} \frac{1}{n} w_{\rho_j}(\boldsymbol{A})\,.$$

*Proof.* Let us write $T_q = T_q^{(n)} := \frac{1}{n} \operatorname{Tr} \boldsymbol{A}^q$. For any finite multiset of integers $\mathcal{Q}$, we can expand

$$\mathbb{E} \left[ \prod_{q \in \mathcal{Q}} T_q \right] - \prod_{q \in \mathcal{Q}} \mathbb{E}\, T_q = \sum_{\varnothing \neq \mathcal{Q}' \subseteq \mathcal{Q}} \mathbb{E} \left[ \prod_{q \in \mathcal{Q}'} (T_q - \mathbb{E}\, T_q) \right] \prod_{q \in \mathcal{Q} \setminus \mathcal{Q}'} \mathbb{E}\, T_q\,.$$

Our goal is now to show that each term in the sum over $\mathcal{Q}'$ converges to 0 as $n \to \infty$. Fix $\mathcal{Q}' \subseteq \mathcal{Q}$ such that $\mathcal{Q}' \neq \varnothing$, and select an arbitrary element $q_0 \in \mathcal{Q}'$. By Cauchy-Schwarz, we have

$$\left( \mathbb{E} \left[ \prod_{q \in \mathcal{Q}'} (T_q - \mathbb{E}\, T_q) \right] \right)^2 \leq \mathbb{E}(T_{q_0} - \mathbb{E}\, T_{q_0})^2 \cdot \mathbb{E} \left[ \prod_{q \in \mathcal{Q}' \setminus \{q_0\}} (T_q - \mathbb{E}\, T_q)^2 \right]\,. \tag{60}$$

We know that $\mathbb{E}(T_{q_0} - \mathbb{E}\, T_{q_0})^2$ converges to 0 as $n \to \infty$ by the $L^2$ tracial moments convergence assumption. For the remaining product of expectations from Eq. (60), we apply the bound $T_q^2 \leq T_{2p}^{q/p}$ for all $q \leq p$ to get: for all $\mathcal{Q}'' \subseteq \mathcal{Q}'$,

$$\prod_{q \in \mathcal{Q}''} T_q^2 \leq T_2^{\sum_{q \in \mathcal{Q}''} q}\,.$$

Therefore, all terms in the expansion of Eq. (60) can be bounded by products of terms of the form $\mathbb{E}\, T_q$ for $q \in \mathbb{N}$. These are all bounded as $n \to \infty$, since convergence in $L^2$ also implies convergence in expectation. Together, we deduce

$$\lim_{n \to \infty} \mathbb{E} \left[ \prod_{q \in \mathcal{Q}} T_q \right] = \prod_{q \in \mathcal{Q}} \lim_{n \to \infty} \mathbb{E}\, T_q\,,$$

which is equivalent to the desired statement. □

**Remark B.4.** *This property is a statement about concentration of the tracial moments. For an example where it does not hold, one can take* $\boldsymbol{A}^{(n)} = a\boldsymbol{I}_n$ *for* $a \sim \mathrm{Unif}(\{\pm 1\})$*, in which case* $\lim_{n\to\infty} \mathbb{E}\frac{1}{n}\operatorname{Tr}(\boldsymbol{A}) = \lim_{n\to\infty} \mathbb{E}\frac{1}{n}\operatorname{Tr}(\boldsymbol{A}^3) = 0$*, while* $\lim_{n\to\infty} \mathbb{E}[\frac{1}{n}\operatorname{Tr}(\boldsymbol{A}) \cdot \frac{1}{n}\operatorname{Tr}(\boldsymbol{A}^3)] = 1$.

We will further show below that analogous formulas hold for joint moments of elements of the $w$- and $z$-bases of polynomials, not just the cycle diagrams.

## B.4 Traffic distribution of orthogonally invariant matrices

We now prove Theorem 4.2 by computing the traffic distribution of an orthogonally invariant matrix $\boldsymbol{A}$, which we recall consists of the limits of expressions of the form $\frac{1}{n}\mathbb{E}z_\alpha(\boldsymbol{A})$ for $\alpha \in \mathcal{A}_0$.

First, for a graph $\alpha = (V(\alpha), E(\alpha))$, define $\mathrm{HE} = \mathrm{HE}(\alpha)$ to be the set of *half-edges* in $\alpha$, a set of size $|\mathrm{HE}| = 2|E|$ which may be identified with pairs $(v, \{v, w\})$ for each choice of $v \in V$ and $\{v, w\} \in E(\alpha)$. Then, to $\alpha$ itself is associated a distinguished perfect matching $\widetilde{\alpha} \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE})$, which matches each pair $(v, \{v, w\})$ and $(w, \{v, w\})$ of half-edges that correspond to the same edge of $\alpha$ (this is the perfect matching that would realize $\alpha$ under the configuration model).

We say that a matching $\beta \in \mathcal{M}(\mathrm{HE})$ is $\alpha$-*local* if all of its matches are between half-edges of the form $(v, e_1)$, $(v, e_2)$, i.e., between pairs of half-edges associated to the same *vertex* (rather than the same *edge*) for $\widetilde{\alpha}$. Let $\mathrm{Loc}(\alpha) \subseteq \mathcal{M}_{\mathrm{perf}}(\mathrm{HE}(\alpha))$ be the set of all $\alpha$-local matchings. Note that $\mathrm{Loc}(\alpha) \neq \varnothing$ if and only if $\alpha$ is Eulerian, i.e., if every vertex has even degree.

At the heart of the matter is the distance between $\widetilde{\alpha}$ and the set $\mathrm{Loc}(\alpha)$, which is minimized precisely by the cactus graphs $\alpha \in \mathcal{C}$:

**Proposition B.5.** *For any graph* $\alpha$*, not necessarily connected, all of whose connected components are Eulerian, we have*

$$\Delta(\widetilde{\alpha}, \mathrm{Loc}(\alpha)) = \min_{\beta \in \mathrm{Loc}(\alpha)} \Delta(\widetilde{\alpha}, \beta) \geq |V(\alpha)| - |\mathrm{conn}(\alpha)|\,,$$

*with equality if and only if every connected component of* $\alpha$ *is a cactus. Further, in that case, there is a unique* $\beta \in \mathrm{Loc}(\alpha)$ *achieving equality, which is the (unique) such* $\beta$ *that matches pairs of half-edges belonging to the same cycle in* $\alpha$*.*

*Proof.* It suffices to consider $\alpha$ connected; the general case follows by considering each connected component separately.

We may rewrite

$$\Delta(\widetilde{\alpha}, \mathrm{Loc}(\alpha)) = \min_{\beta \in \mathrm{Loc}(\alpha)} \Delta(\widetilde{\alpha}, \beta) = |E| - \max_{\beta \in \mathrm{Loc}(\alpha)} |\mathrm{cyc}(\widetilde{\alpha}, \beta)|$$

and therefore it suffices to show that, for all $\alpha$-local matchings of half-edges $\beta$, we have

$$|\mathrm{cyc}(\widetilde{\alpha}, \beta)| \overset{(?)}{\leq} |E| - |V| + 1\,.$$

The set of cycles in the disjoint union of $\widetilde{\alpha}$ and an $\alpha$-local $\beta$ is equivalently the number of cycles in a cycle cover of $\alpha$ (i.e., a partition of its edges into cycles).

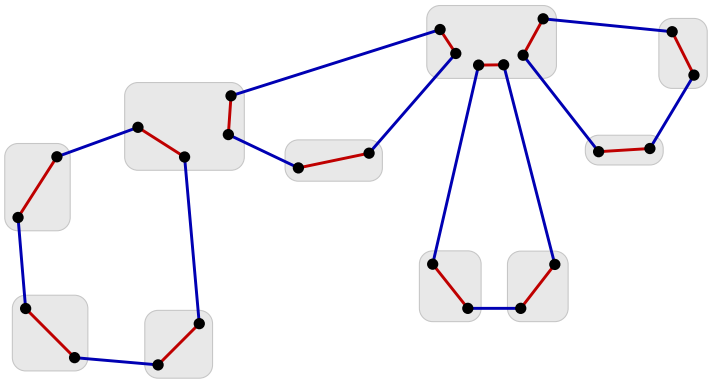


Figure 8: An illustration of the matchings involved in Proposition B.5 and the arguments afterwards. Gray regions represent vertices of a cactus graph $\alpha$, three triangles joined at a vertex with one 4-cycle attached to one of those triangles at a different vertex. This graph has $|V| = 10$ and $|E| = 13$. Black dots in those regions represent half-edges of $\alpha$. In blue, we draw the matching $\widetilde{\alpha}$ realizing the graph $\alpha$; if the gray regions are each contracted to a point and only blue edges are retained, then the resulting graph is the cactus $\alpha$. In red, we draw the unique $\alpha$-local matching $\beta$ that maximizes $|\mathrm{cyc}(\widetilde{\alpha}, \beta)| = |E| - |V| + 1 = 4$; its being $\alpha$-local corresponds to making matches only within the gray regions.

The bound is tight for cycles. Suppose $C_1, \dots, C_k$ is a cycle cover of some connected multigraph $\alpha$. Since $\alpha$ is connected, it is possible to order the $C_i$ such that $C_{i+1}$ has a vertex in common with the union of $C_1, \dots, C_i$ for each $i = 1, \dots, k-1$. Adding each successive $C_i$ then increases $|E| - |V| + 1$ by at least 1, so the bound follows. If the bound is tight, then in the above ordering $C_{i+1}$ must have exactly one vertex in common with the union of $C_1, \dots, C_i$, and thus is a cactus. In that case, there is only one cycle cover, and thus the minimizer $\beta$ is unique and must be as specified in the statement. □

We now proceed to Theorem 4.2 by calculating the traffic distribution of a sequence of orthogonally invariant random matrices $\boldsymbol{A} = \boldsymbol{A}^{(n)}$. We could view this as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ for a Haar-distributed orthogonal $\boldsymbol{Q}$ and some random diagonal $\boldsymbol{D}$, but actually this will not be necessary. Instead, let us take a perspective similar to the calculations in, for instance, [KMW24], which we believe is useful in general. Our idea will be to average the $z_\alpha(\boldsymbol{A})$ over a random rotation $\boldsymbol{Q}$ drawn independently of $\boldsymbol{A}$. Regardless of the structure of $\boldsymbol{A}$, this defines another family of polynomials:

$$\bar{z}_\alpha(\boldsymbol{A}) := \mathop{\mathbb{E}}_{\boldsymbol{Q}} z_\alpha(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top) \,.$$

If $\boldsymbol{Q}$ is drawn from Haar measure, then the $\bar{z}_\alpha$ will be orthogonally invariant polynomials, a greater symmetry than permutation invariance of $z_\alpha$. In particular, since the invariants of matrices under the $O(n)$ action are generated by traces of matrix powers, $\bar{z}_\alpha$ will be a polynomial in these.

*Proof of Theorem 4.2.* Let $\boldsymbol{Q} \in O(n)$ be Haar-distributed and independent of $\boldsymbol{A}$, and let $\alpha = (V, E)$ be a graph. As above, write $\mathrm{HE} = \mathrm{HE}(\alpha)$ for the set of half-edges. We start by directly expanding

the averaged polynomial $\bar{z}$ introduced above:

$$\begin{aligned}
&\bar{z}_\alpha(\boldsymbol{A}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Q}} z_\alpha(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Q}} \sum_{i:V\hookrightarrow[n]} \prod_{\{v,w\}\in E} (\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top)[i(v), i(w)] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Q}} \sum_{i:V\hookrightarrow[n]} \prod_{\{v,w\}\in E} \left( \sum_{j_1,j_2=1}^{n} \boldsymbol{Q}[i(v), j_1]\boldsymbol{A}[j_1, j_2]\boldsymbol{Q}[i(w), j_2] \right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{Q}} \sum_{\substack{i:V\hookrightarrow[n] \\ j:\mathrm{HE}\to[n]}} \prod_{\{v,w\}\in E} \boldsymbol{Q}[i(v), j(v,\{v,w\})]\boldsymbol{Q}[i(w), j(w,\{v,w\})]\boldsymbol{A}[j(v,\{v,w\}), j(w,\{v,w\})] \\
&= \sum_{\substack{i:V\hookrightarrow[n] \\ j:\mathrm{HE}\to[n]}} \left( \mathop{\mathbb{E}}_{\boldsymbol{Q}} \prod_{\{v,w\}\in E} \boldsymbol{Q}[i(v), j(v,\{v,w\})]\boldsymbol{Q}[i(w), j(w,\{v,w\})] \right) \\
&\qquad\qquad \cdot \prod_{\{v,w\}\in E} \boldsymbol{A}[j(v,\{v,w\}), j(w,\{v,w\})]
\end{aligned}$$

Here, we may use the Weingarten calculus, viewing the matchings involved as matchings of half-edges, provided that we view $i : V \to [n]$ as extended to $i' : \mathrm{HE} \to [n]$ by $i'(v, e) := i(v)$, i.e., labelling a half-edge by the vertex involved. This gives:

$$\begin{aligned}
&= \sum_{\substack{i:V\hookrightarrow[n] \\ j:\mathrm{HE}\to[n]}} \sum_{\beta,\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} W(\beta,\gamma)\delta_\beta(i')\delta_\gamma(j) \prod_{\{v,w\}\in E} \boldsymbol{A}[j(v,\{v,w\}), j(w,\{v,w\})] \\
&= \sum_{j:\mathrm{HE}\to[n]} \sum_{\beta,\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \left( \sum_{i:V\hookrightarrow[n]} \delta_\beta(i') \right) W(\beta,\gamma)\delta_\gamma(j) \prod_{\{v,w\}\in E} \boldsymbol{A}[j(v,\{v,w\}), j(w,\{v,w\})]
\end{aligned}$$

The summation over $i$ is zero unless $\beta \in \mathrm{Loc}(\alpha)$, and in that case each choice of $i$ contributes 1, for a total of $n^{|V|}(1 + O(n^{-1}))$. So, we have

$$\begin{aligned}
&= \left(1 + O\left(\frac{1}{n}\right)\right) n^{|V|} \sum_{\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \left( \sum_{\beta\in\mathrm{Loc}(\alpha)} W(\beta,\gamma) \right) \\
&\qquad\qquad \sum_{j:\ \mathrm{HE}\to[n]} \delta_\gamma(j) \prod_{\{v,w\}\in E} \boldsymbol{A}[j(v,\{v,w\}), j(w,\{v,w\})]
\end{aligned}$$

The remaining summation may be grouped into summations over the cycles in the disjoint union of $\gamma$ and $\widetilde{\alpha}$, which gives

$$= \left(1 + O\left(\frac{1}{n}\right)\right) n^{|V|} \sum_{\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \left( \sum_{\beta\in\mathrm{Loc}(\alpha)} W(\beta,\gamma) \right) \prod_{C\in\mathrm{cyc}(\widetilde{\alpha},\gamma)} \mathrm{Tr}\left(\boldsymbol{A}^{\frac{|C|}{2}}\right)$$

Now, we may use the asymptotic formula in Proposition B.2 and normalize the traces to get

$$
\begin{aligned}
&= \left(1 + O\left(\frac{1}{n}\right)\right) n^{|V|} \sum_{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \\
&\qquad \left( \sum_{\beta \in \mathrm{Loc}(\alpha)} \left(1 + O\left(\frac{1}{n}\right)\right) n^{-|\mathrm{HE}| + \mathrm{cyc}(\beta,\gamma)} \mu(\beta,\gamma) \right) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} \mathrm{Tr}\left(\boldsymbol{A}^{\frac{|C|}{2}}\right) \\
&= \left(1 + O\left(\frac{1}{n}\right)\right) n^{|V|} \sum_{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \\
&\qquad \left( \sum_{\beta \in \mathrm{Loc}(\alpha)} \left(1 + O\left(\frac{1}{n}\right)\right) n^{-|\mathrm{HE}| + \mathrm{cyc}(\beta,\gamma) + \mathrm{cyc}(\widetilde{\alpha},\gamma)} \mu(\beta,\gamma) \right) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} \frac{1}{n} \mathrm{Tr}\left(\boldsymbol{A}^{\frac{|C|}{2}}\right) \\
&= \left(1 + O\left(\frac{1}{n}\right)\right) n^{|V|} \sum_{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \\
&\qquad \left( \sum_{\beta \in \mathrm{Loc}(\alpha)} \left(1 + O\left(\frac{1}{n}\right)\right) n^{-\Delta(\beta,\gamma) - \Delta(\widetilde{\alpha},\gamma)} \mu(\beta,\gamma) \right) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} \frac{1}{n} \mathrm{Tr}\left(\boldsymbol{A}^{\frac{|C|}{2}}\right).
\end{aligned}
$$

Let us pause to notice that we have achieved our initial goal, expressing the orthogonally invariant polynomial $\bar{z}_\alpha(\boldsymbol{A})$ as a polynomial in traces of powers of $\boldsymbol{A}$. We now use that, if $\boldsymbol{A}$ was orthogonally invariant to begin with, then

$$\mathop{\mathbb{E}}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A}) = \mathop{\mathbb{E}}_{\boldsymbol{A}} \bar{z}_\alpha(\boldsymbol{A})$$

and continue to determine the right-hand side as $n \to \infty$.

By the triangle inequality, we have

$$\Delta(\beta,\gamma) + \Delta(\widetilde{\alpha},\gamma) \geq \Delta(\beta,\widetilde{\alpha}) \geq \Delta(\widetilde{\alpha}, \mathrm{Loc}(\alpha)) \geq |V| - 1\,.$$

Therefore, under our assumptions, all terms are negligible as $n \to \infty$ except for those where equality is achieved throughout above.

By Proposition B.5, we then find that if $\alpha$ is not a cactus then

$$\lim_{n\to\infty} \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A}) = 0\,.$$

So, suppose that $\alpha$ is a cactus. Then, using the factorization property (Lemma B.3), we have in the limit that

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{A}} z_\alpha(\boldsymbol{A}) = \sum_{\substack{\beta \in \mathrm{Loc}(\alpha) \\ \Delta(\beta,\widetilde{\alpha}) = |V|-1}} \sum_{\substack{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE}) \\ \Delta(\beta,\gamma) + \Delta(\gamma,\widetilde{\alpha}) = \Delta(\beta,\widetilde{\alpha})}} \mu(\beta,\gamma) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} m_{|C|/2}$$

where $m_k$ are the spectral moments. Letting $\eta$ be the $\alpha$-local matching of half-edges belonging to the same cycle around each vertex, by the uniqueness clause of Proposition B.5 we further have that only the term $\beta = \eta$ contributes, giving

$$= \sum_{\substack{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE}) \\ \Delta(\eta,\gamma)+\Delta(\eta,\widetilde{\alpha})=\Delta(\eta,\widetilde{\alpha})}} \mu(\eta,\gamma) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} m_{|C|/2}$$

Suppose there are $k$ cycles in $\alpha$. Then, $|\mathrm{cyc}(\eta,\widetilde{\alpha})| = k$, and, rewriting the condition on $\gamma$ in terms of cycle counts and using the explicit formula for the Möbius function from Eq. (57), we have

$$= \sum_{\substack{\gamma \in \mathcal{M}_{\mathrm{perf}}(\mathrm{HE}) \\ |\mathrm{cyc}(\eta,\gamma)|+|\mathrm{cyc}(\gamma,\widetilde{\alpha})|=|E|+k}} \prod_{C \in \mathrm{cyc}(\eta,\gamma)} (-1)^{\frac{|C|}{2}-1} \mathrm{Cat}\left(\frac{|C|}{2}-1\right) \prod_{C \in \mathrm{cyc}(\widetilde{\alpha},\gamma)} m_{|C|/2}$$

Now, we use that all $\gamma$ appearing in the sum must only match half-edges belonging to the same cycle. Since $\eta$ and $\widetilde{\alpha}$ both have this property also, the various sets of cycles above all form partitions of the cycles in $\alpha$. Thus, the entire sum factorizes over the cycles of $\alpha$. Further, those $\gamma$ that are in the sum have both the partitions of $\mathrm{cyc}(\eta,\gamma)$ and $\mathrm{cyc}(\gamma,\widetilde{\alpha})$ corresponding to non-crossing partitions of each cycle of $\alpha$, and these two non-crossing partitions are Kreweras complements of one another. Putting together all these combinatorial observations, we find:

$$= \prod_{C \in \mathrm{cyc}(\alpha)} \left( \sum_{\pi \in \mathrm{NC}(|C|)} \prod_{A \in K(\pi)} (-1)^{|A|-1} \mathrm{Cat}(|A|-1) \cdot \prod_{B \in \pi} m_{|B|} \right).$$

Now we use Eq. (58) and Eq. (59) to complete the proof:

$$\begin{aligned}
&= \prod_{C \in \mathrm{cyc}(\alpha)} \left( \sum_{\pi \in \mathrm{NC}(|C|)} \mu(\underline{0}_{|C|}, K(\pi)) \cdot \prod_{B \in \pi} m_{|B|} \right) \\
&= \prod_{C \in \mathrm{cyc}(\alpha)} \left( \sum_{\pi \in \mathrm{NC}(|C|)} \mu(\pi, \underline{1}_{|C|}) \cdot \prod_{B \in \pi} m_{|B|} \right) \\
&= \prod_{C \in \mathrm{cyc}(\alpha)} \kappa_{|C|}\,,
\end{aligned}$$

where we have at last identified the free cumulants, completing the calculation. □

We also note that, by exactly the same argument but using the disconnected case of Proposition B.5, we may equally well calculate suitably normalized limits of the values of *disconnected* diagrams in the $z$-basis, which factorize over their connected components:

**Proposition B.6.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be a sequence of orthogonally invariant random matrices that converge in tracial moments in $L^2$ to a probability measure $\mu$. Let $\mathcal{D}$ denote their limiting traffic distribution, which exists by Theorem 4.2 and is given by the explicit formula stated there. Then, for all $k \geq 1$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$,*

$$\lim_{n\to\infty} \frac{1}{n^k} \mathbb{E}\, z_{\alpha_1 \sqcup \cdots \sqcup \alpha_k}(\boldsymbol{A}) = \prod_{i=1}^{k} \mathcal{D}(\alpha_i)\,.$$

## B.5 Concentration of traffic observables

As a corollary, we may also conclude that the traffic distribution is concentrated in the sense of Definition 3.15. This also extends [CDM24, Theorem 4.7] to orthogonally invariant distributions.

**Lemma B.7.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ be orthogonally invariant random matrices that converge in tracial moments in $L^2$ to a probability measure $\mu$. Then the traffic distribution concentrates for $\boldsymbol{A}$ (in the sense of Definition 3.15).*

*Proof.* Let $k \geq 2$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$. Then, by Lemma 3.17, it suffices to show the concentration property in the $z$-basis, namely that:

$$\lim_{n\to\infty} \mathbb{E}\left[\prod_{i=1}^{k} \frac{1}{n} z_{\alpha_i}(\boldsymbol{A})\right] \overset{(?)}{=} \lim_{n\to\infty} \prod_{i=1}^{k} \mathbb{E}\,\frac{1}{n} z_{\alpha_i}(\boldsymbol{A})\,.$$

Note that, upon expanding the summations in the $z$-basis polynomials, we have

$$z_{\alpha_1}(\boldsymbol{A}) \cdots z_{\alpha_k}(\boldsymbol{A}) = z_{\alpha_1 \sqcup \cdots \sqcup \alpha_k}(\boldsymbol{A}) + z_{\beta_1}(\boldsymbol{A}) + \cdots + z_{\beta_M}(\boldsymbol{A})\,,$$

where $\alpha_1 \sqcup \cdots \sqcup \alpha_k$ is the disjoint union, while the $\beta_i$ are various graphs formed by identifying subsets of the vertices of this disjoint union according to different non-trivial partitions of the vertices, provided that no two vertices of the same $\alpha_j$ are identified. In particular, all $\beta_i$ have at most $k-1$ connected components. Therefore, by Proposition B.6, we have

$$\lim_{n\to\infty} \frac{1}{n^k} z_{\beta_i}(\boldsymbol{A}) = 0$$

for all $i \in [M]$. Thus,

$$\lim_{n\to\infty} \frac{1}{n^k}\,\mathbb{E}\left[\prod_{i=1}^{k} z_{\alpha_i}(\boldsymbol{A})\right] = \lim_{n\to\infty} \frac{1}{n^k}\,\mathbb{E}[z_{\alpha_1 \sqcup \cdots \sqcup \alpha_k}(\boldsymbol{A})]\,,$$

and the result then follows by Proposition B.6. □

## B.6 Traffic distribution of punctured orthogonally invariant matrices

Since the r-ROM plays an important role in our main results, let us sketch how similar calculations can give an explicit combinatorial description of its traffic distribution, and indeed that of the puncturing of any orthogonally invariant random matrices. Recall that in the main text we relied entirely on the implicit description of this traffic distribution via Lemma 3.14. The closed form we give below is completely explicit, but, being in terms of a rather complicated summation over matchings, seems less useful than the implicit one.

We follow the notation from the proof in the previous section. Additionally, for a graph $\alpha$ and a matching $\beta$ of the half-edges of $\alpha$, we write $\mathrm{loc}(\beta)$ for the set of edges of $\beta$ that go between half-edges of the same vertex of $\alpha$, and $\mathrm{nonloc}(\beta)$ for the set of edges of $\beta$ that go between half-edges of different vertices of $\alpha$. Recall also that $\widetilde{\alpha}$ is the matching of half-edges of $\alpha$ corresponding to the edges actually in the graph $\alpha$.

**Theorem B.8.** *Let $\boldsymbol{A} = \boldsymbol{A}^{(n)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ be a sequence of orthogonally invariant random matrices that converges in tracial moments in $L^2$ to a probability measure $\mu$. Write $m_k$ for the $k$th moment of $\mu$ and $\boldsymbol{\Pi} = \boldsymbol{\Pi}^{(n)} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Then, for all $\alpha \in \mathcal{A}$,*

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{\boldsymbol{A}} z_\alpha(\boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{\Pi}) = \sum_{\substack{\beta\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE}(\alpha))\\ \alpha\sqcup\mathrm{nonloc}(\beta)\ \textit{is a cactus}}} (-1)^{|\mathrm{nonloc}(\beta)|} \sum_{\substack{\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE}(\alpha))\\ \Delta(\beta,\gamma)+\Delta(\gamma,\widetilde{\alpha})=\Delta(\beta,\widetilde{\alpha})}} \mu(\beta,\gamma) \prod_{C\in\mathrm{cyc}(\widetilde{\alpha},\gamma)} m_{|C|}\,.$$

*Proof.* Following the same calculations as in the proof of Theorem 4.2 above but now applied to $\boldsymbol{\Pi}\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top\boldsymbol{\Pi}$, we instead find:

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_{\boldsymbol{Q},\boldsymbol{A}} z_\alpha(\boldsymbol{\Pi}\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^\top\boldsymbol{\Pi})\\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{A}} \sum_{\beta,\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} W_n(\beta,\gamma)\cdot \prod_{C\in\mathrm{cyc}(\widetilde{\alpha},\gamma)} \mathrm{Tr}\Big(\boldsymbol{A}^{|C|}\Big)\cdot z_{G(\beta)}(\boldsymbol{\Pi})
\end{aligned}$$

where $G(\beta)$ denotes the graph formed by "wiring together" the matching of half-edges $\beta$ (so that, for example, $G(\widetilde{\alpha}) = \alpha$). Note that here if we replaced $\boldsymbol{\Pi}$ by $\boldsymbol{I}$, we would get $z_{G(\beta)}(\boldsymbol{I}) = \mathbf{1}_{\beta\in\mathrm{Loc}(\alpha)} n^{|V|}(1 + O(n^{-1}))$, compatible with the previous calculation in the proof of Theorem 4.2, and indeed the above is true for an arbitrary symmetric matrix $\boldsymbol{\Pi}$, not only the particular projection we are concerned with. But, in our particular case, since $\boldsymbol{\Pi}$ is constant on the diagonal and on the off-diagonal, we have

$$= \frac{1}{n}\sum_{\beta,\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} W_n(\beta,\gamma)\cdot \mathbb{E}_{\boldsymbol{A}} \prod_{C\in\mathrm{cyc}(\widetilde{\alpha},\gamma)} \mathrm{Tr}\Big(\boldsymbol{A}^{|C|}\Big)\cdot n^{\underline{|V|}} \left(-\frac{1}{n}\right)^{|\mathrm{nonloc}(\beta)|} \left(1-\frac{1}{n}\right)^{|\mathrm{loc}(\beta)|}$$

and now by the same asymptotics as before,

$$= \sum_{\beta,\gamma\in\mathcal{M}_{\mathrm{perf}}(\mathrm{HE})} \left(1 + O\left(\frac{1}{n}\right)\right) n^{-\Delta(\widetilde{\alpha},\gamma)-\Delta(\beta,\gamma)-|\mathrm{nonloc}(\beta)|+|V|-1} \mu(\beta,\gamma)(-1)^{|\mathrm{nonloc}(\beta)|} \prod_{C\in\mathrm{cyc}(\widetilde{\alpha},\gamma)} m_{|C|}$$

We claim that, for any connected $\alpha$ realized by the matching $\widetilde{\alpha}$ of its half-edges, and any other matching $\beta$ of the half-edges of $\alpha$, we have

$$\Delta(\widetilde{\alpha},\beta) + |\mathrm{nonloc}(\beta)| \geq |V| - 1\,.$$

As before, this is equivalent to having

$$|\mathrm{cyc}(\widetilde{\alpha},\beta)| \leq |E| + |\mathrm{nonloc}(\beta)| - |V| + 1\,.$$

Consider an ancillary graph $\alpha'$ constructed by adding edges to $\alpha$ for each non-local match in $\beta$. This graph is still connected, by parity considerations it must be Eulerian, and it has a total of $|E| + |\mathrm{nonloc}(\beta)|$ edges. $|\mathrm{cyc}(\widetilde{\alpha},\beta)|$ is now the size of a cycle cover of $\alpha'$, and the claim then follows by the bounds from the proof of Proposition B.5 applied to $\alpha'$.

We also again have by the triangle inequality that

$$\Delta(\widetilde{\alpha},\gamma) + \Delta(\beta,\gamma) \geq \Delta(\widetilde{\alpha},\beta)\,.$$

Thus, all terms in the sum above are of at most constant order. Further, those of constant order are those where the exponent of $n$ is zero, which are those where the above bound is tight. By the characterization in Proposition B.5, this is precisely when $\alpha'$ as formed above is a cactus, and the stated result follows after rearranging. □

# C Convergence of Stochastic Processes

In Section 6, we deal with convergence in distribution of stochastic processes indexed by countably infinite set, intended as weak convergence in the product topology. Equivalently, this means that every finite-dimensional marginal converges in distribution.

**Definition C.1.** *Let $\mathcal{A}$ be a countable set. For random variables $(\boldsymbol{x}^{(n)})_{n\geq 1}$ and $\boldsymbol{x}^{\infty}$ taking values in $\mathbb{R}^{\mathcal{A}}$, we say that $\boldsymbol{x}^{(n)}$ converges in distribution to $\boldsymbol{x}^{\infty}$ and write*

$$\boldsymbol{x}^{(n)} \xrightarrow{(\mathrm{d})} \boldsymbol{x}^{\infty}$$

*if, for every $k \geq 1$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$, we have*

$$(x^{(n)}_{\alpha_1}, \ldots, x^{(n)}_{\alpha_k}) \xrightarrow{(\mathrm{d})} (X^{\infty}_{\alpha_1}, \ldots X^{\infty}_{\alpha_k}) \,.$$

To show convergence in distribution, we will use the method of moments [Bil95, Theorems 29.4, 30.1, 30.2]. The following theorem follows from Carleman's conditions on moment-determinacy of a distribution on $\mathbb{R}$, combined with [Pet82].

**Theorem C.2** (Method of moments)**.** *Let $(\boldsymbol{x}^{(n)})_{n\geq 1}$ be a sequence of stochastic processes indexed by a countable set $\mathcal{A}$. Assume that*

1. *All joint moments converge: for any $k \geq 1$ and $\alpha_1, \ldots, \alpha_k \in \mathcal{A}$, the limit of the joint moments*

$$\lim_{n\to\infty} \mathbb{E}\left[\prod_{i=1}^{k} x^{(n)}_{\alpha_i}\right] \tag{61}$$

   *exists.*

2. *All marginals are subexponential: for every $\alpha \in \mathcal{A}$, there exists $C_\alpha > 0$ such that for all $p \geq 1$,*

$$\lim_{n\to\infty} \mathbb{E}\left(x^{(n)}_{\alpha}\right)^{2p} \leq (C_\alpha p)^{2p} \,. \tag{62}$$

*Then $\boldsymbol{x}^{(n)}$ converges in distribution to the unique law on $\mathbb{R}^{\mathcal{A}}$ with moments given by Eq. (61).*

**Lemma C.3** (Truncation)**.** *Let $(x_n)_{n\geq 1}$ and $(y_n)_{n\geq 1}$ be sequences of random variables such that*

1. *For any $K > 0$, conditionally on $|x_n| \leq K$, $(y_n)_{n\geq 1}$ converges in distribution.*
2. *$(x_n)_{n\geq 1}$ is tight, i.e., $\sup_{n\geq 1} \Pr(|x_n| > K) \underset{K\to\infty}{\longrightarrow} 0$.*

*Then, $(y_n)_{n\geq 1}$ converges in distribution.*

*Proof.* First, we prove:

**Claim C.4.** *$(y_n)_{n\geq 1}$ is tight.*

*Proof.* For any $K, L > 0$, we have $\Pr(|y_n| > L) \leq \Pr(|y_n| > L \mid |x_n| \leq K) + \Pr(|x_n| > K)$. Pick $K$ large enough so that the second term is bounded by $\varepsilon$ uniformly in $n$. $(y_n)_{n\geq 1}$ is tight conditionally on $|x_n| \leq K$, so there exists $L > 0$ large enough so that the first term is also bounded by $\varepsilon$ uniformly in $n$. □

By Claim C.4 and Prokhorov's theorem, it remains to show that every subsequence of $(y_n)_{n\geq 1}$ that converges in distribution, converges to the same limit. Fix $f : \mathbb{R} \to \mathbb{R}$ to be a bounded continuous function and $\varepsilon > 0$. Then, by the law of total expectations, for any $n \geq 1$,

$$\begin{aligned}|\mathbb{E}\, f(y_n) - \mathbb{E}\left[f(y_n) \mid |x_n| \leq K\right]| &= \Pr(|x_n| > K)\left(\mathbb{E}\left[f(y_n) \mid |x_n| > K\right] - \mathbb{E}\left[f(y_n) \mid |x_n| \leq K\right]\right) \\ &\leq 2\|f\|_\infty \Pr(|x_n| > K) \\ &\leq \varepsilon\end{aligned}$$

by setting $K = K(\varepsilon)$ to be a large enough constant (with the second assumption). By the first assumption, there exists $N \geq 1$ such that for any $n, m \geq N$,

$$|\mathbb{E}\left[f(y_n) \mid |x_n| \leq K\right] - \mathbb{E}\left[f(y_m) \mid |x_m| \leq K\right]| \leq \varepsilon\,.$$

In turn, this implies $|\mathbb{E}\, f(y_n) - \mathbb{E}\, f(y_m)| \leq 3\varepsilon$ by the triangle inequality, so $(\mathbb{E}\, f(y_n))_{n\geq 1}$ is a Cauchy sequence, so it converges as $n \to \infty$. This implies that every weak subsequential limit of $(y_n)_{n\geq 1}$ converges to the same limit, which concludes the proof. □

## C.1 Connection with convergence of the empirical distribution

Let us also remark on certain details concerning modes of convergence that are important to the use and interpretation of Theorem 6.2.

Recall that we "stack" the $\boldsymbol{z}_\alpha(\boldsymbol{A})$ for $\alpha \in \mathcal{A}_1$ into a single vector with more complicated entries, $\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}) \in (\mathbb{R}^{\mathcal{A}_1})^n$. Using our notation from Section 1, we then sample a random coordinate of this vector, forming a further random countably infinite vector $\mathrm{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A})) \in \mathbb{R}^{\mathcal{A}_1}$. This contains the $i$th entry of each $\boldsymbol{z}_\alpha(\boldsymbol{A})$, for a single shared randomly chosen $i \sim \mathrm{Unif}([n])$. Define the infinite random vector $Z^\infty_{\mathcal{A}_1}$ similarly. Theorem 6.2 states that:

$$\mathrm{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}^{(n)})) \xrightarrow[n\to\infty]{\text{(d)}} Z^\infty_{\mathcal{A}_1}\,, \tag{63}$$

By the Cramér-Wold theorem, this is equivalent to: for any bounded continuous function $\varphi$ and any finitely supported vector of coefficients $c_\alpha$,

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{A}} \frac{1}{n}\sum_{i=1}^n \varphi\left(\sum_{\alpha\in\mathcal{A}_1} c_\alpha \boldsymbol{z}_\alpha(\boldsymbol{A})[i]\right) = \mathop{\mathbb{E}}_{\boldsymbol{A}} \varphi\left(\sum_{\alpha\in\mathcal{A}_1} c_\alpha Z^\infty_\alpha\right).$$

Alternatively, we may also make sense of this statement in terms of empirical distributions, which are just the laws of the random variables $\mathrm{samp}(\boldsymbol{x})$ discussed above.

**Definition C.5** (Empirical distribution)**.** *For $\boldsymbol{x} \in \mathbb{R}^n$, we write* $\mathrm{ed}(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}[i]}$ *for the empirical distribution of the entries of $\boldsymbol{x}$.*

Then, $\mathrm{ed}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}))$ is a random probability measure on the space $\mathbb{R}^{\mathcal{A}_1}$, and the random variable $\mathrm{samp}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}^{(n)}))$ is a single draw from this random probability measure. Its law is a *deterministic* probability measure on the space $\mathbb{R}^{\mathcal{A}_1}$, which is the expectation of the random measure $\mathrm{ed}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}))$ (if $\mu$ is a random measure, then its expectation takes values $(\mathbb{E}\mu)(A) = \mathbb{E}[\mu(A)]$). Thus, the above Eq. (63) is further equivalent to the weak convergence of probability measures

$$\mathbb{E}\,\mathrm{ed}(\boldsymbol{z}_{\mathcal{A}_1}(\boldsymbol{A}^{(n)})) \xrightarrow[n\to\infty]{\text{(w)}} \mathrm{Law}(Z^{\infty}_{\mathcal{A}_1})\,.$$

Again by the Cramér-Wold theorem, this is equivalent to, for any finitely supported coefficient vector of $c_\alpha$, having

$$\mathbb{E}\,\mathrm{ed}\left(\sum_{\alpha\in\mathcal{A}_1} c_\alpha \boldsymbol{z}_\alpha(\boldsymbol{A}^{(n)})\right) \xrightarrow[n\to\infty]{\text{(w)}} \mathrm{Law}\left(\sum_{\alpha\in\mathcal{A}_1} c_\alpha Z^{\infty}_\alpha\right).$$

In particular, since the output $\boldsymbol{x}_t$ of a GFOM can be viewed in the above way, we see that the empirical distributions of $\boldsymbol{x}_t = \boldsymbol{x}_t(\boldsymbol{A})$ are related to the asymptotic states $X^{\infty}_t$ by

$$\mathbb{E}\,\mathrm{ed}(\boldsymbol{x}_t(\boldsymbol{A}^{(n)})) \xrightarrow[n\to\infty]{\text{(w)}} \mathrm{Law}(X^{\infty}_t)\,.$$

Thus our results, interpreted in terms of convergence of the random empirical distributions of GFOM iterates, give convergence of the expectations of random measures. Often it is desirable to prove stronger modes of convergence in such situations, by proving that not only do we have

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{A}} \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{x}_t(\boldsymbol{A}^{(n)})[i]) = \mathbb{E}\varphi(X^{\infty}_t)\,,$$

but also that the random variable inside the expectation *concentrates* over the randomness in $\boldsymbol{A}$. We do not pursue this here, because it would require introducing additional assumptions on the matrices $\boldsymbol{A}$ involved, which may vary from application to application. As the example discussed in Remark B.4 shows, this kind of concentration does not follow automatically from the convergence in expectation that we show. An instructive example is the argument in [BLM15], which uses similar proof techniques to ours, but, to show that the above kind of convergence also happens in $L^2$ uses a trick involving the entrywise independence of the Wigner matrices they work with (see their Proposition 5).

In our much more general setting, it seems reasonable to ask instead for the convergence in the definition of the traffic distribution in Eq. (2) to happen in a stronger mode such as $L^2$. We leave the exploration of such conditions and the determination of which random matrix distributions they hold for to future work.

# D Omitted Proofs

## D.1 Combinatorial lemmas

We gather here lemmas involving only graph combinatorics.

**Lemma D.1.** *For all $\sigma, \sigma' \in \mathcal{C}_1$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$,*

$$\boldsymbol{z}_\sigma(\boldsymbol{A}) \cdot \boldsymbol{z}_{\sigma'}(\boldsymbol{A}) - \boldsymbol{z}_{\sigma\oplus\sigma'}(\boldsymbol{A}) \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1\setminus\mathcal{C}_1}(\boldsymbol{A}))\,,$$

*where $\sigma \oplus \sigma' \in \mathcal{C}_1$ is the grafting of $\sigma$ and $\sigma'$ at the root.*

*Proof.* In the $z$-basis expansion of $\boldsymbol{z}_\sigma(\boldsymbol{A}) \cdot \boldsymbol{z}_{\sigma'}(\boldsymbol{A})$, we sum over all possible partial matchings of the vertices of $\sigma$ and $\sigma'$. The empty matching contributes exactly $z_{\sigma\oplus\sigma'}(\boldsymbol{A})$. Any other matching that merges some vertices $u \in V(\sigma)$ and $v \in V(\sigma')$ creates 4 edge-disjoint paths between the root and the merged vertex. Merging additional vertices of $\sigma$ and $\sigma'$ can only increase the number of edge-disjoint paths, so the resulting graphs cannot be cactuses. □

**Lemma D.2.** *For all $\sigma \in \mathcal{C}_1$, $\alpha \in \mathcal{A}_1 \setminus \mathcal{C}_1$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$,*

$$\boldsymbol{z}_\sigma(\boldsymbol{A}) \cdot \boldsymbol{z}_\alpha(\boldsymbol{A}) \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1\setminus\mathcal{C}_1}(\boldsymbol{A}))\,.$$

*Proof.* The proof is similar to Lemma D.1. In this case, the graph corresponding to the empty matching is not a cactus because $\alpha$ is not. All other matchings create at least 3 edge-disjoint paths between the root and the merged vertex. □

**Lemma D.3.** *For each $\alpha \in \mathcal{A}_1 \setminus \mathcal{T}_1$ and $\beta \in \mathcal{A}_1$,*

$$\boldsymbol{z}_\alpha(\boldsymbol{A}) \cdot \boldsymbol{z}_\beta(\boldsymbol{A}) \in \mathrm{span}(\boldsymbol{z}_{\mathcal{A}_1\setminus\mathcal{T}_1})\,.$$

*Proof.* The non-treelike diagrams $\mathcal{A}_1 \setminus \mathcal{T}_1$ can be characterized as:

**Claim D.4.** *Let $\alpha \in \mathcal{A}_1$. Then $\alpha \in \mathcal{A}_1 \setminus \mathcal{T}_1$ if and only if one of the following holds:*

*(i) there exists a bridge edge which does not have a path to the root using only bridge edges,*

*(ii) or there exist a pair of vertices with three edge-disjoint paths between them.*

*Proof of Claim D.4.* It is clear that either structure forbids $\alpha$ from being treelike. Conversely, if there are at most two edge-disjoint paths between all pairs of vertices, then the bridge edges of $\alpha$ go between cactuses. Then condition *(i)* characterizes whether all bridge edges are connected to the root. □

Using the claim, if $\alpha$ has a structure of type *(ii)* then this is preserved in the product terms with any $\beta$. Suppose then that $\alpha$ has a structure of type *(i)* and call the bridge edge $e$. Note that both $\alpha, \beta$ are connected by definition of $\mathcal{A}_1$ . If no descendants of $e$ intersect with $\beta$, then the type *(i)* structure is preserved. Conversely, if any descendant of $e$ intersects with $\beta$, then we obtain a new path from the descendant to the root through $\beta$ which is disjoint from the other edges of $\alpha$. Edge $e$ has at least one ancestor which is not a bridge edge, hence there were already two edge-disjoint paths containing this ancestor. Together with the new path we obtain a structure of type *(ii)*. In all cases, the product terms remain in $\mathcal{A}_1 \setminus \mathcal{T}_1$. □

*Proof of Lemma 6.9.* First, if $P$ matches an internal vertex of a hanging cactus, then it creates three edge-disjoint paths from the root to that vertex. These paths cannot be eliminated by merging

other vertices, so $\tau_P$ cannot be a cactus. Therefore, we may assume without loss of generality that $\tau_1$ and $\tau_2$ contain no hanging cactuses.

It is straightforward to check that any homeomorphic matching yields a cactus. We focus on the converse. Specifically, suppose that we are given a matching $P$ between the vertices of $\tau_1$ and $\tau_2$ such that $\tau_P \in \mathcal{C}_1$. We prove $P \in H(\tau_1, \tau_2)$ by induction on $|V(\tau_1)| + |V(\tau_2)|$.

For the base case, suppose that $\tau_1$ or $\tau_2$ has only one vertex. Then $\tau_P$ can be a cactus only if both $\tau_1$ and $\tau_2$ consist of a single vertex.

For the inductive step, let $u_1^1, \ldots, u_k^1$ be the children of the root of $\tau_1$, and let $u_1^2, \ldots, u_\ell^2$ be the children of the root of $\tau_2$. A necessary condition for $\tau_P$ to be a cactus is that $k = \ell$ (and this is also necessary for $P$ to be a homeomorphic matching). Moreover, after reordering $u_1^2, \ldots, u_k^2$ if necessary, we may assume that for all $i \in [k]$, $u_i^1$ and $u_i^2$ lie on the same cycle in $\tau_P$, and that these form exactly $k$ distinct cycles in $\tau_P$ incident to the root.

For each $i \in [k]$ and $j \in \{1, 2\}$, let $S_i^j$ denote the non-root vertices of $\tau_j$ that are mapped under $P$ to the same cycle of $\tau_P$ as $u_i^1, u_i^2$.

**Claim D.5.** *For every $i \in [k]$, there is exactly one vertex $v_i^1 \in S_i^1$ and exactly one vertex $v_i^2 \in S_i^2$ that are mapped to the same vertex of $\tau_P$.*

*Proof.* Since $\tau_1$ and $\tau_2$ are acyclic, creating a cycle in $\tau_P$ requires identifying two other vertices than the root. Conversely, identifying more than one pair of vertices would create three edge-disjoint paths to the root in $\tau_P$, contradicting the fact that the latter is a cactus. □

**Claim D.6.** *For each $i \in [k]$ and $j \in \{1, 2\}$, every pair in $P$ incident to a vertex in the subtree rooted at $v_i^j$ has its other endpoint in the subtree rooted at $v_i^{3-j}$.*

*Proof.* Suppose for contradiction that there is a pair of $P$ between a vertex $w^1$ in the subtree rooted at $v_i^1$ and a vertex $w^2$ in the subtree rooted at $v_{i'}^2$ for some $i' \neq i$. Then in $\tau_P$ there are three edge-disjoint paths from the image of $v_i^1$ to the root: two lie on the cycle formed by $S_i^1 \cup S_i^2$, and the third is obtained by concatenating the path from $v_i^1$ to $w^1$ with the path from $w^2$ to the root. This contradicts the fact that $\tau_P$ is a cactus. □

By Claim D.6, we may apply the induction hypothesis for each $i \in [k]$ to the subtree of $\tau_1$ rooted at $v_i^1$ and the subtree of $\tau_2$ rooted at $v_i^2$. Thus, the restriction of $P$ to these subtrees is a homeomorphic matching. In particular, $v_i^1$ and $v_i^2$ have the same degree.

**Claim D.7.** *Let $i \in [k]$. Then $v_i^1$ and $v_i^2$ are either both in the core of their respective trees or both outside of it. Moreover, for each $j \in \{1, 2\}$, no vertex in $S_i^j \setminus \{v_i^j\}$ lies in the core of $\tau_j$.*

*Proof.* For the first part, since $v_i^1$ and $v_i^2$ have the same degree, they are either both in the core or both outside the core.

For the second part, suppose for contradiction that some $w \in S_i^j \setminus \{v_i^j\}$ lies in the core of $\tau_j$. Since $w$ has degree greater than 2, its image in the cactus $\tau_P$ is an articulation vertex. Let $\rho$ be a cycle of $\tau_P$ incident to $w$ that is distinct from the cycle induced by $S_i^1 \cup S_i^2$. Then the two neighbors of $w$ in $\rho$ are images of vertices of $\tau_j$. Since $\tau_j$ is acyclic, the cycle $\rho$ must contain a vertex $w'$ that is the image of a vertex of $\tau_{3-j}$. But then $\tau_P$ contains three edge-disjoint paths from $w$ to the root: two through the cycle induced by $S_i^1 \cup S_i^2$, and a third obtained by following $\rho$ from $w$ to $w'$ and then the path from $w'$ to the root. This contradicts the fact that $\tau_P$ is a cactus. □

Let $i \in [k]$. For $j \in \{1,2\}$, let $w_i^j$ be the first descendant of $u_i^j$ that lies in the core of $\tau_j$. By Claim D.7, there are only two cases:

1. Either $v_i^j = w_i^j$ for both $j \in \{1,2\}$. In this case, there are no non-core vertices to match on the path from $u_i^j$ to $v_i^j$, so the induced matching is empty (and hence trivially order-preserving).
2. Or $v_i^j \neq w_i^j$ for both $j \in \{1,2\}$. In this case, by induction, the matching between $v_i^j$ and $w_i^j$ is order-preserving. Matching $v_i^1$ to $v_i^2$ and adding the matching from $v_i^j$ to $w_i^j$ yields an order-preserving matching from $u_i^j$ to $w_i^j$.

By induction, the restriction of $P$ induces an isomorphism between the cores of $\tau_1$ and $\tau_2$ within each subtree rooted at $v_i^j$. Since there is no core vertex on the path from $u_i^j$ to $v_i^j$ by Claim D.7, these local isomorphisms extend to an isomorphism between the cores of $\tau_1$ and $\tau_2$ globally. This concludes the proof. □

*Proof of Lemma 6.10.* Given $\gamma_1, \dots, \gamma_\ell \in \mathcal{G}_1$, we can expand

$$\prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} = \sum_P \boldsymbol{z}_{\gamma_P},$$

where $P$ ranges over all partitions of $V(\gamma_1) \cup \ldots \cup V(\gamma_\ell)$ such that all roots are in the same block, but no two vertices of the same $\gamma_i$ are in the same block. Suppose that $\gamma_P$ is treelike.

**Claim D.8.** *Every internal vertex of a hanging cactus forms a singleton block.*

*Proof.* Suppose for contradiction that an internal vertex $u$ of a hanging cactus in $\gamma_1$ lies in the same block as some vertex $v$ of $\gamma_2$. Let $u'$ be the attachment vertex of the cycle containing $u$. In $\gamma_P$, there are three edge-disjoint paths between the images of $u$ and $u'$: two are inherited from $\gamma_1$, while the third is obtained by following the path in $\tau_2$ from $v$ to the root and then the path in $\tau_1$ from the root to $u'$. This contradicts Lemma D.3, since $\gamma_P$ is assumed to be treelike. □

By Claim D.8, we may temporarily delete the hanging cactuses from $\gamma_1, \dots, \gamma_\ell$ and then reattach them in $\gamma_P$; this does not affect whether $\gamma_P$ is treelike. Hence, we may assume without loss of generality that none of $\gamma_1, \dots, \gamma_\ell$ contains a hanging cactus.

**Claim D.9.** *Let $M$ be the graph on $[\ell]$ with an edge between $i, j \in [\ell]$ if there exist $u \in V(\gamma_i)$ and $v \in V(\gamma_j)$ that lie in the same block of $P$. Then $M$ is a matching.*

*Proof.* Suppose for contradiction that $M$ is not a matching. Then there exist non-root vertices $u \in V(\gamma_1)$, $v, v' \in V(\gamma_2)$, and $w \in V(\gamma_3)$ such that $u$ and $v$ (resp. $w$ and $v'$) lie in the same block of $P$. Let $v''$ be the lowest common ancestor of $v$ and $v'$ in $\gamma_2$. Since $\gamma_2 \in \mathcal{G}_1$, $v''$ is not the root of $\gamma_2$. In $\gamma_P$, there are three edge-disjoint paths from the image of $v''$ to the root: one is the inherited path from $v''$ to the root inside $\gamma_2$; the second follows the path in $\gamma_2$ from $v''$ to $v$ and then the path in $\gamma_1$ from $u$ to the root; and the third follows the path in $\gamma_2$ from $v''$ to $v'$ and then the path in $\gamma_3$ from $w$ to the root. This contradicts Lemma D.3, since $\gamma_P$ is treelike by assumption. □

By Claim D.9, it follows that

$$\prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} - \sum_{M \in \mathcal{M}(\ell)} \sum_{\substack{P_{u,v} \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \bigoplus_{u \notin M} \gamma_u} \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}),$$

where for each edge $uv \in M$, the sum over $P_{u,v}$ ranges over all partial matchings between $V(\gamma_u)$ and $V(\gamma_v)$ that fix the roots.

Finally, note that unless $P_{u,v}$ is empty, $\gamma_{P_{u,v}}$ cannot be a treelike diagram that is not a cactus. Indeed, no vertices in the hanging cactuses can be matched; otherwise it would create three edge-disjoint paths. Moreover, if we match two tree vertices, that would create two edge-disjoint paths to the root, and thus would force the diagram to be a cactus. Since the grafting of non-treelike diagrams is again non-treelike, the only treelike contributions arise when each factor $\gamma_{P_{u,v}}$ is a cactus. By Lemma 6.9, this forces $P_{u,v}$ to be a homeomorphic matching. Hence,

$$\prod_{j=1}^{\ell} \boldsymbol{z}_{\gamma_j} - \sum_{M \in \mathcal{M}(\ell)} \sum_{\substack{P_{uv} \in H(\gamma_u, \gamma_v) \\ \forall uv \in M}} \boldsymbol{z}_{\bigoplus_{uv \in M} \gamma_{P_{u,v}} \oplus \bigoplus_{u \notin M} \gamma_u} \in \operatorname{span}(\boldsymbol{z}_{\mathcal{A}_1 \setminus \mathcal{T}_1}),$$

as desired. □

## D.2 Handling empirical averages

To represent expressions involving empirical averages, we allow the coefficients in a diagram representation to be formal polynomials in the quantities $\{\langle \boldsymbol{z}_\alpha(\boldsymbol{A}) \rangle : \alpha \in \mathcal{A}_1\}$. Another approach would be to use disconnected diagrams, as in [JP25].

**Lemma D.10.** *Assume that $\boldsymbol{A} = \boldsymbol{A}^{(n)}$ satisfies the assumptions of Theorem 6.2, and furthermore, the traffic distribution concentrates for $\boldsymbol{A}$ (Definition 3.15). Let*

$$\boldsymbol{x} = \sum_{\alpha \in \mathcal{A}_1} c_\alpha \boldsymbol{z}_\alpha(\boldsymbol{A}) \tag{64}$$

*for some finitely supported coefficients $(c_\alpha)_{\alpha \in \mathcal{A}_1}$ which are polynomials $c_\alpha \in \mathbb{R}[\mathcal{V}]$ with $\mathcal{V} := \{\langle \boldsymbol{z}_\alpha(\boldsymbol{A}) \rangle : \alpha \in \mathcal{A}_1\}$. Then,*

$$X := \sum_{\alpha \in \mathcal{A}_1} c_\alpha(\mathbb{E}\, Z^\infty_{\mathcal{A}_1}) \cdot Z^\infty_\alpha \tag{65}$$

*is the asymptotic state of $\boldsymbol{x}$. Moreover, if $\boldsymbol{x}_t$ is of the form Eq. (64) for any $t \geq 1$ and $X_t$ is correspondingly defined as in Eq. (65), then $(X_t)_{t \geq 1}$ is the asymptotic state of $(\boldsymbol{x}_t)_{t \geq 1}$.*

*Proof.* For polynomial test functions, the convergence in Eq. (33) follows directly from the concentration of the traffic distribution. Moreover, Lemma 3.16 implies that $\frac{1}{n} z_\alpha(\boldsymbol{A})$ converges in $L^2$ to a deterministic limit for any $\alpha \in \mathcal{A}_1$. So we can combine Lemma 6.17 with Slutsky's lemma to obtain that Eq. (33) also holds for bounded continuous functions. □

## D.3 Proof of Lemma 6.30

In this section, we prove Lemma 6.30. We assume throughout that $\boldsymbol{H}$ satisfies the assumptions of Theorem 6.29. We will prove that $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$ have the same state evolution by relating them to the following intermediate iteration:

$$\boldsymbol{y}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{\Pi} f_{t-1}(\boldsymbol{y}_{t-1}) - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot (\boldsymbol{\Pi} f_s(\boldsymbol{y}_s)) \quad \forall t \geq 1, \tag{66}$$

where $\boldsymbol{b}_{s,t}$ is defined in Eq. (45). Unless specified otherwise, all expectations in this section are taken with respect to both $\boldsymbol{H}$ and $\boldsymbol{y}_0$.

Theorem 6.18 does not apply to $\boldsymbol{y}_t$ because of the Gaussian initialization $\boldsymbol{y}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ (instead of $\boldsymbol{y}_0 = \boldsymbol{1}$). To analyze this initialization, we extend the class of diagrams to *generalized diagrams*, that is, graphs $\alpha = (V(\alpha), E(\alpha))$ together with an additional label $p(v) \in \mathbb{N}$ assigned to each vertex. The $z$-polynomial associated with a graph $\alpha$ is

$$z_\alpha(\boldsymbol{A}, \boldsymbol{y}_0) := \sum_{i: V(\alpha) \hookrightarrow [n]} \prod_{\{u,v\} \in E(\alpha)} \boldsymbol{A}[i(u), i(v)] \prod_{v \in V(\alpha)} \boldsymbol{y}_0[i(v)]^{p(v)}.$$

The collection of generalized vector diagrams $\mathcal{A}_1(\boldsymbol{y}_0)$ is defined analogously. Definitions such as $\mathcal{T}_1$, $\mathcal{G}_1$, and $\overset{\circ\circ}{=}$ extend to generalized diagrams by simply ignoring the labels $p(v)$.

As in the proof of Theorem 6.28, one caveat is that $\boldsymbol{y}_t$ cannot be directly expanded as a linear combination of *connected* generalized vector diagrams, because the iteration involves the scalar quantity $\langle f_t(\boldsymbol{y}_t) \rangle$. We therefore proceed as in Lemma D.10, viewing the coefficients in the diagram expansion as formal polynomials in these variables whenever necessary.

Our first observation is that taking expectation over $\boldsymbol{y}_0$ in the $z$-basis turns (up to a scaling factor) a generalized diagram $\alpha$ into the same diagram $\alpha$ where the labels are ignored.

**Lemma D.11.** *For any generalized scalar diagram $\alpha$ (not necessarily connected) and any $\boldsymbol{H} \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{y}_0} z_\alpha(\boldsymbol{H}, \boldsymbol{y}_0) = \begin{cases} \left(\prod_{v \in V(\alpha)} (p(v) - 1)!!\right) z_\alpha(\boldsymbol{H}) & \text{if } p(v) \text{ is even for every } v \in V(\alpha) \\ 0 & \text{otherwise} \end{cases}$$

*Proof.* In the $z$-basis, all vertices are assigned distinct labels. Therefore, we may take the expectation over $\boldsymbol{y}_0$ separately at each vertex, since the coordinates of $\boldsymbol{y}_0$ are independent. For each vertex $v$, we have $\mathbb{E}_{Z \sim \mathcal{N}(0,1)} Z^{p(v)} = (p(v) - 1)!!$ if $p(v)$ is even or 0 if $p(v)$ is odd. □

Next, we describe structural properties of the labels $p(v)$ appearing in the diagram expansion of the iterates of the AMP iteration Eq. (66).

**Lemma D.12.** *We have $\boldsymbol{y}_t \overset{\circ\circ}{=} \sum_\tau c_\tau \boldsymbol{z}_\tau(\boldsymbol{H}, \boldsymbol{y}_0)$ and $f_t(\boldsymbol{y}_t) \overset{\circ\circ}{=} \sum_\tau c'_\tau \boldsymbol{z}_\tau(\boldsymbol{H}, \boldsymbol{y}_0)$, where $c_\tau$ and $c'_\tau$ are supported on (generalized) treelike diagrams $\tau$ such that, for all $v \in V(\tau)$:*

$$p(v) = \begin{cases} 1 & \text{if } v \text{ is a leaf vertex of } \tau \\ 0 \text{ or } 2 & \text{if } v \text{ is in a hanging cactus} \\ 0 & \text{otherwise} \end{cases}.$$

*Leaves of treelike diagrams are defined after removing hanging cactuses.*

*Proof.* First, the proof of Lemma 6.23 still goes through with the nonlinearities $g_t(y) = f_t(y) - \langle f_t(\boldsymbol{y}_t)\rangle$, after extending the coefficient field from $\mathbb{R}$ to the ring of formal polynomials in $\{\langle \boldsymbol{z}_\alpha(\boldsymbol{A})\rangle : \alpha \in \mathcal{A}_1\}$. Therefore, we obtain

$$\boldsymbol{y}_t \overset{\infty}{=} \sum_{s=0}^{t-1} \boldsymbol{B}_{s,t}(\boldsymbol{\Pi} f_s(\boldsymbol{y}_s))^{\neq 1}\,. \tag{67}$$

We now argue by induction on $t$. The base case is $f_0(\boldsymbol{y}_0) = \boldsymbol{y}_0$ which is the singleton with $p(v) = 1$.

Now, suppose that the claim holds for $\boldsymbol{y}_t$. The treelike diagrams appearing in $f_t(\boldsymbol{y}_t)$ are obtained by considering all possible products of treelike diagrams $\gamma_1, \dots, \gamma_\ell \in \mathcal{G}_1 \cup \mathcal{C}_1$ appearing in $\boldsymbol{y}_t$. By Lemma 6.10, each such product can be written as a sum over matchings among the $\gamma_i$, where each $\gamma_i$ is either paired into a cactus or does not intersect any other $\gamma_j$. In the second case, the values $p(v)$ within $\gamma_i$ are unchanged. In the first case, the values $p(v)$ at the leaves are updated from 1 to 2, while all other values $p(v)$ within $\gamma_i$ remain unchanged.

Moreover, no non-trivial intersection between $\boldsymbol{B}_{s,t}$ and $(\boldsymbol{\Pi} f_s(\boldsymbol{y}_s))^{\neq 1}$ can produce a treelike diagram. Hence, the decomposition of $\boldsymbol{y}_{t+1}$ given by Eq. (67), together with the induction hypothesis, shows that in every treelike diagram appearing in $\boldsymbol{y}_{t+1}$, the condition on $p(v)$ is inherited directly from the corresponding property of $f_s(\boldsymbol{y}_s)$. This completes the induction. □

**Lemma D.13.** *For any $t \geq 1$ and any polynomial $\varphi : \mathbb{R}^t \to \mathbb{R}$,*

$$\lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{x}_0} \langle \varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)\rangle = \lim_{n\to\infty} \mathop{\mathbb{E}}_{\boldsymbol{H},\boldsymbol{y}_0} \langle \varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\rangle\,.$$

*Proof of Lemma D.13.* An iteration involving $\boldsymbol{A}$ can be reduced to one involving $\boldsymbol{H}$ by expanding

$$\boldsymbol{A} f_t(\boldsymbol{y}_t) = \boldsymbol{H} f_t(\boldsymbol{y}_t) - \langle f_t(\boldsymbol{y}_t)\rangle \boldsymbol{H}\mathbf{1} - \langle \boldsymbol{H}\boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle \mathbf{1}\,. \tag{68}$$

Set $\delta_t := \langle \boldsymbol{H}\boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle$ and $m_t := \langle f_t(\boldsymbol{y}_t)\rangle$. We first compare $\boldsymbol{y}_t$ with the following modified iteration, which differs from $\boldsymbol{x}_t$ only in the formula for the Onsager correction term:

$$\tilde{\boldsymbol{y}}_0 = \boldsymbol{y}_0\,, \quad \tilde{\boldsymbol{y}}_t = \boldsymbol{A} f_{t-1}(\tilde{\boldsymbol{y}}_{t-1}) - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot (\boldsymbol{\Pi} f_s(\tilde{\boldsymbol{y}}_s))\,,$$

where $\boldsymbol{b}_{s,t}$ is defined in Eq. (45).

**Claim D.14.** *For any $t \in \mathbb{N}$, we have*

$$\tilde{\boldsymbol{y}}_t - \boldsymbol{y}_t = \sum_{\alpha} c_{t,\alpha}(\delta_0, \dots, \delta_{t-1}, m_0, \dots, m_{t-1}) \boldsymbol{z}_\alpha(\boldsymbol{H}, \boldsymbol{y}_0)\,, \tag{69}$$

*where the sum runs over finitely many generalized vector diagrams, and each $c_{t,\alpha}$ is a polynomial in $\delta_0, \dots, \delta_{t-1}, m_0, \dots, m_{t-1}$ that is divisible by $\delta_s$ for some $s \in \{0, \dots, t-1\}$.*

*Proof of Claim D.14.* We argue by induction on $t$. For $t = 0$, $\boldsymbol{y}_0 - \tilde{\boldsymbol{y}}_0 = 0$, establishing the base case. Let $t \geq 1$ and suppose that Eq. (69) holds for all $s < t$. First, one easily verifies from the

induction hypothesis that the same property Eq. (69) holds for $\boldsymbol{\Delta}_s := f_s(\tilde{\boldsymbol{y}}_s) - f_s(\boldsymbol{y}_s)$ for every $s < t$. By Eq. (68), we can then write

$$\boldsymbol{A} f_{t-1}(\tilde{\boldsymbol{y}}_{t-1}) - \boldsymbol{H}\boldsymbol{\Pi} f_{t-1}(\boldsymbol{y}_{t-1}) = \boldsymbol{H}\boldsymbol{\Pi}\boldsymbol{\Delta}_{t-1} - \delta_{t-1}\mathbf{1} - \langle \boldsymbol{H}\boldsymbol{\Pi}\boldsymbol{\Delta}_{t-1}\rangle \mathbf{1}\,,$$

and each of the three terms on the right-hand side satisfies a decomposition of the form Eq. (69) by the induction hypothesis. Finally, the correction terms differ by

$$\sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot (\boldsymbol{\Pi} f_s(\tilde{\boldsymbol{y}}_s)) - \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot (\boldsymbol{\Pi} f_s(\boldsymbol{y}_s)) = \sum_{s=0}^{t-1} \boldsymbol{b}_{s,t} \cdot \boldsymbol{\Pi}\boldsymbol{\Delta}_s\,,$$

which again satisfies the property Eq. (69) by the induction hypothesis. Combining these observations, we conclude that $\tilde{\boldsymbol{y}}_t - \boldsymbol{y}_t$ satisfies Eq. (69), completing the induction. □

Next, fix any polynomial $\varphi : \mathbb{R}^t \to \mathbb{R}$. By Claim D.14, we have

$$\langle \varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\rangle - \langle \varphi(\tilde{\boldsymbol{y}}_1, \dots, \tilde{\boldsymbol{y}}_t)\rangle = \sum_{\alpha} c_\alpha(\delta_0, \dots, \delta_{t-1}, m_0, \dots, m_{t-1}) \langle \boldsymbol{z}_\alpha(\boldsymbol{H}, \boldsymbol{y}_0)\rangle\,, \tag{70}$$

where the sum runs over finitely many generalized scalar diagrams, and each $c_\alpha$ is a polynomial in $\delta_0, \dots, \delta_{t-1}, m_0, \dots, m_{t-1}$ that is divisible by some $\delta_s$. In the remainder of the proof, we show that each term on the right-hand side of Eq. (70) converges to 0 in expectation. The reason is that each coefficient $c_\alpha$ contains a factor $\delta_s$, and these quantities converge to 0 in $L^2$:

**Claim D.15.** $\langle \boldsymbol{H}\boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle \xrightarrow{L^2} 0$ *for any* $t \geq 1$.

*Proof of Claim D.15.* The claim is equivalent to the statement that $\frac{1}{n^2}\mathbb{E}\langle \boldsymbol{H}\mathbf{1}, \boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle^2$ converges to 0. This quantity can be expanded as a linear combination of terms of the form

$$\frac{1}{n^2}\,\mathbb{E}\left[z_\alpha(\boldsymbol{H}, \boldsymbol{y}_0) z_\beta(\boldsymbol{H}, \boldsymbol{y}_0)\right]$$

where $\alpha, \beta \in \mathcal{A}$ both belong to the support of the expansion of $\langle \boldsymbol{H}\mathbf{1}, \boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle$. As in the proof of Lemma B.7,

$$\frac{1}{n^2}\,\mathbb{E}\left[z_\alpha(\boldsymbol{H}, \boldsymbol{y}_0) z_\beta(\boldsymbol{H}, \boldsymbol{y}_0)\right] = \frac{1}{n^2}\,\mathbb{E}\left[z_{\alpha \sqcup \beta}(\boldsymbol{H}, \boldsymbol{y}_0)\right] + o(1)\,.$$

Indeed, each identification of vertices across the two copies yields a connected diagram whose expectation, after normalization by $1/n^2$, converges to 0 by the existence of the traffic distribution. This holds for every realization of $\boldsymbol{y}_0$, and therefore also after taking expectation over $\boldsymbol{y}_0$.

Taking expectation over $\boldsymbol{y}_0$ and using Lemma D.11, each term either vanishes or becomes a constant multiple of $\mathbb{E}_{\boldsymbol{H}}\left[z_{\alpha \sqcup \beta}(\boldsymbol{H})\right]$, where $\alpha \sqcup \beta$ is viewed as an ordinary scalar diagram obtained by ignoring the labels $p(v)$. By Lemma B.7 and the strong cactus property, the only terms that contribute to the limit are those for which both $\alpha$ and $\beta$ are cactuses. Viewing $\boldsymbol{H}\mathbf{1}$ as a rooted tree with one edge, the cactuses in the $z$-basis expansion of $\langle \boldsymbol{H}\mathbf{1}, \boldsymbol{\Pi} f_t(\boldsymbol{y}_t)\rangle$ arise when the child of $\boldsymbol{H}\mathbf{1}$ is merged with a leaf of a diagram from $\boldsymbol{\Pi} f_t(\boldsymbol{y}_t)$. By Lemma D.12, such leaves satisfy $p(v) = 1$. Applying Lemma D.11 once again, we find that each of these cactus terms has expectation 0 over $\boldsymbol{y}_0$, which concludes the proof. □

After taking expectation over $\boldsymbol{H}$ and $\boldsymbol{y}_0$, any monomial appearing on the right-hand side

of Eq. (70) has the following form for some $p_i, q_i \in \mathbb{N}$:

$$\mathbb{E}\left[\delta_s \prod_{i=0}^{t-1} \delta_i^{p_i} m_i^{q_i} \langle z_\alpha(\boldsymbol{H}, \boldsymbol{z}_0)\rangle\right] \le \left(\mathbb{E}\,\delta_s^2\right)^{\frac{1}{2}} \cdot \left(\mathbb{E}\left[\prod_{i=0}^{t-1} \delta_i^{2p_i} m_i^{2q_i} \langle z_\alpha(\boldsymbol{H}, \boldsymbol{z}_0)\rangle^2\right]\right)^{\frac{1}{2}}, \tag{71}$$

where the inequality follows from Cauchy-Schwarz.

The first factor on the right-hand side of Eq. (71) converges to 0 as $n \to \infty$ by Claim D.15. The second factor can be expanded in the $z$-basis as a finite linear combination of products of generalized $z$-diagrams. Taking expectation over $\boldsymbol{y}_0$ and using Lemma D.11, each such term either vanishes or becomes a constant multiple of a product of ordinary scalar $z$-diagrams. By Lemma B.7, the normalized expectation of each of these terms has a finite limit as $n \to \infty$ and in particular, is uniformly bounded in $n$. Therefore, the second factor on the right-hand side of Eq. (71) is bounded, and hence the right-hand side of Eq. (70) converges to 0 in expectation. In summary, we have shown:

$$\lim_{n\to\infty} \mathbb{E}\langle\varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\rangle - \mathbb{E}\langle\varphi(\tilde{\boldsymbol{y}}_1, \dots, \tilde{\boldsymbol{y}}_t)\rangle = 0\,. \tag{72}$$

Finally, as in the proof of Theorem 6.28, we may use the traffic concentration property (Lemma B.7) to replace $\boldsymbol{b}_{s,t}$ by $\kappa_{t-s} \prod_{s<r<t} \langle f_r'(\tilde{\boldsymbol{y}}_r)\rangle$ in the iteration for $\tilde{\boldsymbol{y}}_t$ without affecting the asymptotic state. This yields

$$\lim_{n\to\infty} \mathbb{E}\langle\varphi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_t)\rangle - \mathbb{E}\langle\varphi(\tilde{\boldsymbol{y}}_1, \dots, \tilde{\boldsymbol{y}}_t)\rangle = 0\,.$$

Combining this with Eq. (72) completes the proof. □

*Proof of Lemma 6.30.* First, we can replace every occurrence of $\boldsymbol{\Pi} f_0(\boldsymbol{y}_0)$ in $\boldsymbol{y}_t$ by $f_0(\boldsymbol{y}_0)$ using the traffic concentration property, since $\langle f_0(\boldsymbol{y}_0)\rangle = \langle \boldsymbol{y}_0\rangle$, which converges to 0 as $n \to \infty$. After this update, the iterates $\boldsymbol{y}_t$ and $\boldsymbol{u}_t$ have the same generalized diagram expansion as functions of their initializations $\boldsymbol{y}_0$ and $\boldsymbol{u}_0$. Note that this expansion is formal in the variables $\langle \boldsymbol{z}_\alpha(\boldsymbol{A})\rangle$ for $\alpha \in \mathcal{A}_1(\boldsymbol{y}_0)$, because the puncturing operation introduces terms of the form $\langle f_t(\boldsymbol{y}_t)\rangle$.

First, by the strong cactus property and Lemma D.11, all non-cactus terms in the generalized diagram expansions of $\langle\varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\rangle$ and $\langle\varphi(\boldsymbol{u}_1, \dots, \boldsymbol{u}_t)\rangle$ converge to 0 in expectation. Second, using Lemma D.12 and extending the same argument one further step to $\varphi$, all cactus diagrams in the generalized diagram expansions of $\boldsymbol{y}_t$ and $\boldsymbol{u}_t$ satisfy $p(v) \in \{0, 2\}$, since they have no non-root leaves (the iterates for $t \ge 1$ have no singleton component). Therefore, by Lemma D.11, the expectations of the cactus terms remain unchanged as $n \to \infty$ if we replace $\boldsymbol{y}_0$ by $\boldsymbol{u}_0 = \mathbf{1}$.

Combining these facts with the traffic concentration property for $\boldsymbol{H}$ (Lemma B.7) shows that $\langle\varphi(\boldsymbol{y}_1, \dots, \boldsymbol{y}_t)\rangle - \langle\varphi(\boldsymbol{u}_1, \dots, \boldsymbol{u}_t)\rangle$ converges to 0 in expectation, as desired. □

## D.4 Proof of Lemma 6.31

In this section, we prove the auxiliary lemmas for block matrices.

**Definition D.16.** *Let $\alpha \in \mathcal{C}_1$ be a cactus diagram. For a coloring $\chi : V(\alpha) \to [q]$ of the vertices of $\alpha$ with $q$ colors, we say that $\chi$ is valid if for every cycle $\rho = (u_1, \dots, u_k, u_1) \in \mathrm{cyc}(\alpha)$, there exist $r, c \in [q]$ such that $\chi(u_i) = r$ when $i$ is even and $\chi(u_i) = c$ when $i$ is odd, with $r = c$ if $k$ is odd. We write $\chi(\rho) = \{r, c\}$ in this case.*

Our main diagrammatic calculation for block models is the following, which gives the traffic distribution on each block:

**Lemma D.17.** *Let $\boldsymbol{A}$ be as in the setting of Lemma 6.31. Then for all $\alpha \in \mathcal{A}_1$ and $r \in [q]$:*

$$\frac{q}{n} \sum_{\substack{i \in [n] \\ \mathrm{block}(i) = r}} \mathbb{E}\, \boldsymbol{z}_\sigma(\boldsymbol{A})[i] \xrightarrow[n\to\infty]{} \begin{cases} \displaystyle\sum_{\substack{\chi : V(\alpha) \to [q] \\ \chi \text{ valid} \\ \chi(\mathrm{root}) = r}} \prod_{\rho \in \mathrm{cyc}(\alpha)} \kappa_{|\rho|}^{\chi(\rho)} & \text{if } \alpha \in \mathcal{C}_1 \\ 0 & \text{if } \alpha \in \mathcal{A}_1 \setminus \mathcal{C}_1 \end{cases}$$

*Proof.* We partition the sum defining $\boldsymbol{z}_\alpha(\boldsymbol{A})$ according to the block of each vertex, as in the proof of Proposition 4.6:

$$\boldsymbol{z}_\alpha(\boldsymbol{A}) = \sum_{\chi : V(\alpha) \setminus \{\mathrm{root}\} \to [q]} \boldsymbol{z}_{\alpha_\chi}((\boldsymbol{A}_{r,c})_{r,c \in [q]})$$

where $\alpha_\chi$ is a diagram whose edges are colored by the matrices $\boldsymbol{A}_{r,c}$. For a fixed $r' \in [q]$, we get

$$\frac{q}{n} \sum_{\substack{i \in [n] \\ \mathrm{block}(i) = r'}} \mathbb{E}\, \boldsymbol{z}_\alpha(\boldsymbol{A})[i] = \sum_{\chi : V(\alpha) \setminus \{\mathrm{root}\} \to [q]} \frac{q}{n} \sum_{\substack{i \in [n] \\ \mathrm{block}(i) = r'}} \mathbb{E}\, \boldsymbol{z}_{\alpha_\chi}((\boldsymbol{A}_{r,c})_{r,c \in [q]})[i]\,.$$

By the definition of traffic independence (Definition 4.5), the limit as $n \to \infty$ exists for each term indexed by $\chi$ on the right-hand side. Hence, the limit of the left-hand side also exists. Arguing as in the proof of Proposition 4.6, we find that the limit is zero for all $\alpha \in \mathcal{A}_1 \setminus \mathcal{C}_1$.

For cactus diagrams $\alpha \in \mathcal{C}_1$, asymptotic traffic independence and the strong factorizing cactus property of the individual blocks imply that the only nonzero contributions arise when every cycle of $\alpha_\chi$ is monochromatic, in the sense that it involves only a single matrix $\boldsymbol{A}_{r,c}$. This happens if and only if $\chi$ is a valid coloring, in which case the corresponding term contributes asymptotically

$$\prod_{\rho \in \mathrm{cyc}(\sigma)} \kappa_{|\rho|}^{\chi(\rho)}\,,$$

as desired. □

*Proof of Lemma 6.31.* Let $\mathcal{L}_{\mathcal{C}_1}(r)$ denote the values from Lemma D.17:

$$\mathcal{L}_\sigma(r) := \sum_{\substack{\chi : V(\sigma) \to [q] \\ \chi \text{ valid} \\ \chi(\mathrm{root}) = r}} \prod_{\rho \in \mathrm{cyc}(\sigma)} \kappa_{|\rho|}^{\chi(\rho)}\,.$$

We first prove that all joint moments of $\boldsymbol{z}_{\mathcal{C}_1}(\boldsymbol{A})[i]$ conditioned on $\mathrm{block}(i) = r$ converge to the

moments of the deterministic sequence $Z^{\infty}_{\mathcal{C}_1}(r)$. For any $\sigma_1, \dots, \sigma_k \in \mathcal{C}_1$, we have

$$\begin{aligned}
&\frac{q}{n} \sum_{\substack{i \in [n] \\ \mathrm{block}(i) = r}} \mathbb{E}\left[\boldsymbol{z}_{\sigma_1}(\boldsymbol{A})[i] \cdots \boldsymbol{z}_{\sigma_k}(\boldsymbol{A})[i]\right] \\
&= \frac{q}{n} \sum_{\substack{i \in [n] \\ \mathrm{block}(i) = r}} \mathbb{E}\, \boldsymbol{z}_{\sigma_1 \oplus \dots \oplus \sigma_k}(\boldsymbol{A})[i] + o(1) && \text{(by Lemmas D.1 and D.17)} \\
&= \mathcal{L}_{\sigma_1 \oplus \dots \oplus \sigma_k}(r) + o(1) = \prod_{j=1}^{k} \mathcal{L}_{\sigma_j}(r) + o(1)\,. && \text{(by Lemma D.17)}
\end{aligned}$$

So it remains to prove that $\mathcal{L}_{\mathcal{C}_1}(r)$ satisfies the same recursion as $Z^{\infty}_{\mathcal{C}_1}(r)$. First, one readily checks that $\mathcal{L}_{\mathrm{singleton}}(r) = 1$, as in *(i)*. Next, suppose that $\sigma$ is rooted at a vertex of degree 2, and let $\ell$ and $\sigma_1, \dots, \sigma_{\ell-1}$ be as in *(ii)*. Then, by decomposing according to the value of cycle containing the root, we have

$$\mathcal{L}_{\sigma}(r) = \begin{cases} \displaystyle\sum_{c \in [q]} \left[ \kappa_{\ell}^{\{r,c\}} \prod_{\substack{k=2 \\ k \text{ odd}}}^{\ell} \mathcal{L}_{\sigma_k}(r) \prod_{\substack{k=2 \\ k \text{ even}}}^{\ell} \mathcal{L}_{\sigma_k}(c) \right] & \text{if } \ell \text{ is even} \\ \displaystyle \kappa_{\ell}^{\{r,r\}} \prod_{k=2}^{\ell} \mathcal{L}_{\sigma_k}(r) & \text{if } \ell \text{ is odd} \end{cases}$$

just like the recursion in *(ii)*. Similarly, *(iii)* follows from the fact that the definition of $\mathcal{L}_{\sigma}(r)$ factorizes over graftings at the root. Together, this shows that $\mathcal{L}_{\mathcal{C}_1}(r) = Z^{\infty}_{\mathcal{C}_1}(r)$.

Since the limit is deterministic, we have shown that conditionally on $\mathrm{block}(i) = r$, $\boldsymbol{z}_{\mathcal{C}_1}(\boldsymbol{A})[i]$ converges to $Z^{\infty}_{\mathcal{C}_1}(r)$ in $L^2$. Since $\mathrm{block}(i) \sim \mathrm{Unif}([q])$, it follows that $(\mathrm{block}(i), \boldsymbol{z}_{\mathcal{C}_1}(\boldsymbol{A})[i])$ converges in distribution to $(R, Z^{\infty}_{\mathcal{C}_1}(R))$, where $R \sim \mathrm{Unif}([q])$. □