The Annals of Applied Probability
2024, Vol. 34, No. 4, 3943–3994
https://doi.org/10.1214/24-AAP2056

UNIVERSALITY OF APPROXIMATE MESSAGE PASSING ALGORITHMS AND TENSOR NETWORKS

BY TIANHAO WANG[a], XINYI ZHONG[b] AND ZHOU FAN[c]

Department of Statistics and Data Science, Yale University, [a]*tianhao.wang@yale.edu,* [b]*xinyi.zhong@yale.edu,* [c]*zhou.fan@yale.edu*

Approximate message passing (AMP) algorithms provide a valuable tool for studying mean-field approximations and dynamics in a variety of applications. Although these algorithms are often first derived for matrices having independent Gaussian entries or satisfying rotational invariance in law, their state evolution characterizations are expected to hold over larger universality classes of random matrix ensembles.

We develop several new results on AMP universality. For AMP algorithms tailored to independent Gaussian entries, we show that their state evolutions hold over broadly defined generalized Wigner and white noise ensembles, including matrices with heavy-tailed entries and heterogeneous entrywise variances that may arise in data applications. For AMP algorithms tailored to rotational invariance in law, we show that their state evolutions hold over delocalized sign-and-permutation-invariant matrix ensembles that have a limit distribution over the diagonal, including sensing matrices composed of subsampled Hadamard or Fourier transforms and diagonal operators.

We establish these results via a simplified moment-method proof, reducing AMP universality to the study of products of random matrices and diagonal tensors along a tensor network. As a by-product of our analyses, we show that the aforementioned matrix ensembles satisfy a notion of asymptotic freeness with respect to such tensor networks, which parallels usual definitions of freeness for traces of matrix products.

CONTENTS

Received August 2022; revised October 2023.
MSC2020 subject classifications. 68W40.
Key words and phrases. Random matrices, free probability, dynamical mean-field theory.

1. Introduction. Approximate message passing (AMP) algorithms are a general family of iterative algorithms, driven by a random matrix $\mathbf{W}$, whose iterates admit a simple distributional characterization in the asymptotic limit of increasing dimensions. Their origins may be traced separately in the engineering, statistics, and probability literatures [11, 29, 45], where these algorithms have since provided an important tool for studying mean-field phenomena in many probabilistic models. Without seeking to be exhaustive, we mention here their applications to analyses of spin glass and perceptron models [12, 13, 25, 38, 39], recovery thresholds and asymptotic phenomena in high-dimensional statistical models [5, 15, 26, 28, 29, 49, 56, 58, 61, 62, 68, 76], and mean-field dynamics of other first-order optimization algorithms including discrete-time and continuous-time gradient descent [20, 21]. We refer readers to [40] for a recent review.

Asymptotic distributional characterizations of the AMP iterates, known as their *state evolutions*, are often first proved for orthogonally invariant matrices $\mathbf{W}$ using an inductive conditioning technique. For $\mathbf{W}$ with i.i.d. Gaussian entries, this method was developed in [7, 11] and has been extended to analyze AMP algorithms of increasing generality in [9, 41, 43, 58, 61]. For $\mathbf{W}$ satisfying rotational invariance in law, a similar technique has been applied to analyze various AMP algorithms in [37, 50, 51, 63, 67, 69–71], with a parallel line of work [17–19, 60] deriving related algorithms using nonrigorous methods of dynamic functional theory.

It is expected—and in some settings known—that the state evolution characterizations of AMP algorithms should extend beyond orthogonally invariant matrices, to describe also the limit distributions of iterates when applied to broader universality classes of random matrix ensembles. For example, it was shown in [6] that AMP algorithms designed for i.i.d. Gaussian matrices and having polynomial nonlinearities admit state evolutions that are universal across matrices with sub-Gaussian entries of common variance. In [22], universality over a similar matrix class for AMP with Lipschitz nonlinearities was proven using a different Gaussian interpolation method, and extended to spectrally initialized algorithms for spiked matrix models. Moving beyond matrices with independent entries, in [32] it was shown that the state evolution of a linear AMP algorithm for phase retrieval holds universally for sub-sampled Hadamard matrices. Recently, results of [33, 34]—fruit of parallel research efforts—showed universality for AMP algorithms having divergence-free nonlinearities over a broad model of semi-random matrices with randomly signed rows/columns and delocalized entries. The latter work [34] also applied these results to establish universality classes of matrices for more general first-order iterative algorithms, including proximal gradient methods and general versions of AMP. We discuss the relation of these results to our work in more detail at the conclusion of the following section.

1.1. *Contributions*. Our current work has the two-fold goal of extending the scope of some of these universality results of [6, 22, 33], and of presenting a more direct and elementary proof for AMP universality. We summarize our contributions as follows:

1. For AMP algorithms designed for i.i.d. Gaussian matrices, we show that their state evolutions hold more broadly over generalized Wigner and white noise ensembles, with entries having potentially heteroskedastic variances and higher moments growing rapidly with the dimension n. This includes standardized adjacency matrices of sparse random graphs down to sparsity levels of $(\log n)/n$, as well as data matrices arising in contexts of count-valued and missing observations after applying practical row and column normalization schemes. We discuss two motivating applications in Examples 2.24 and 2.25 of Section 2.4. In the random matrix theory literature, global spectral laws and spectral CLTs for related ensembles were studied in [2], and universality of local spectral statistics in [35, 36].

2. For AMP algorithms designed for rotationally invariant matrix ensembles, we show that their state evolutions hold over universality classes of "generalized invariant matrices" that satisfy only invariances of permutation and sign and whose generated algebra over the diagonal, in the sense of [4], consists of matrices with delocalized entries and common normalized trace. Importantly, this includes matrices composed of subsampled Hadamard or discrete Fourier transforms and diagonal operators, which admit fast matrix-vector multiplication for signal processing applications. We discuss a specific application to universality of the compressed sensing phase transition for AMP [6, 27] in Example 2.26 of Section 2.4. Related models of permutation-and-sign-invariant matrices have been studied in the context of asymptotic liberation in [1].

3. We introduce a simplified two-step proof of AMP universality, in the first step reducing universality to the study of products of $\mathbf{W}$ with diagonal tensors along a tensor network, and in the second step establishing universality of the values of these matrix-tensor products. The second step admits a simple combinatorial analysis for all of the preceding matrix ensembles. Our argument for the first step is general and holds irrespective of the specific matrix ensemble. We propose this two-step proof framework in part to enable easier extensions of AMP universality to other random matrix models (e.g., having sufficiently weak or short-range correlation across entries) as this need arises in applications.

4. For symmetric matrices $\mathbf{W} \in \mathbb{R}^{n\times n}$, our definition of a tensor network is a natural generalization of expressions of the form

$$\frac{1}{n}\mathbf{u}^\top \mathbf{W}\mathbf{T}_1\mathbf{W}\mathbf{T}_2\cdots\mathbf{T}_k\mathbf{W}\mathbf{v}$$

for deterministic vectors $\mathbf{u}$, $\mathbf{v}$ and diagonal matrices $\mathbf{T}_1, \ldots, \mathbf{T}_k$ to expressions involving higher-order diagonal tensors. As a by-product of our analyses, we show for both the preceding classes of generalized Wigner and generalized invariant matrices $\mathbf{W}$ that they satisfy a notion of asymptotic freeness with respect to such tensor networks, namely, that if all diagonal tensors have asymptotically vanishing normalized trace, then evaluations of expressions of this form are also 0 in the asymptotic limit. This is parallel to notions of asymptotic freeness [74], usually defined with respect to normalized traces of matrix products, in settings of products with higher-order tensors. Our analysis of tensor networks has also similarities to the analysis of graph observables in the theory of traffic freeness developed in [52].

Our proofs use a moment-method and polynomial approximation strategy, similar to [6]. In heuristic derivations of AMP algorithms from belief propagation for matrices in the Gaussian universality class, the Onsager correction terms arise from the removal of single-step-backtracking messages. The arguments of [6] showed a corresponding equivalence between such AMP algorithms and a tensorial unfolding of AMP using nonbacktracking paths. To our knowledge, the correction terms in the algorithms of [37] for rotationally-invariant ensembles do not have a similar combinatorial interpretation, motivating us to analyze a simpler tensorial unfolding without nonbacktracking structure. Our results for the Gaussian universality class may be obtained via either approach; we take the opportunity to present unified proofs for both the Gaussian and non-Gaussian universality classes using the same unfolding, and to simplify the polynomial approximation arguments of [6] using more recent state evolution results of [37] for AMP with non-Lipschitz functions. We remark that, as in the AMP universality analysis of [22] which developed a different continuous interpolation argument, our method of proof applies also to more general first-order iterative algorithms of the form studied in [21] that are characterizable by an asymptotic state evolution.

Our analyses for generalized invariant ensembles (Definitions 2.6 and 2.20) are complementary to those of the recent works [33, 34], which studied an important family of Vector-AMP style methods that have divergence-free nonlinearities [17, 51, 67, 69]. As discussed in

[34], the universality classes for these algorithms are broader than that of the more general AMP algorithms we study here, for example, containing matrices with differing spectral distributions having common second moment. [33, 34] prove universality of these algorithms for semi-random sign-invariant matrices and i.i.d. side information vectors, by developing a Hermite-polynomial unfolding of the AMP iterations and leveraging the vanishing of certain terms in this unfolding due to the divergence-free form. The latter work [34] extends this result to also derive certain spectral and strongly semi-random universality classes for first-order algorithms that do not have this divergence-free structure. Our methods here establish universality over a class of matrices that has similarities to, and is partially inspired by, these latter classes studied in [34] (cf. Proposition 2.7(b)). We obtain these results via an alternative analysis of a simpler tensorial unfolding in the standard monomial basis. As we discuss in Remark 2.15, our proofs also establish the existence and universality of the limit empirical distribution of iterates for first-order methods applied to matrices beyond the orthogonally invariant universality class, suggesting the possible development of new iterative algorithms with characterizable state evolutions for such matrices.

1.2. *Notation.* We denote entries of $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{W} \in \mathbb{R}^{n\times n}$ as $x[i]$ and $W[i, j]$. For vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ and a random vector $(X_1, \dots, X_k)$, we write

$$(\mathbf{x}_1, \dots, \mathbf{x}_k) \overset{W_p}{\to} (X_1, \dots, X_k) \quad \text{as } n \to \infty$$

for the Wasserstein-p convergence of the empirical distribution of rows of $(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^{n\times k}$ to the joint law of $(X_1, \dots, X_k)$. This means, for any continuous function $f : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$|f(x_1, \dots, x_k)| \le C\big(1 + \|(x_1, \dots, x_k)\|_2^p\big) \quad \text{for a constant } C > 0, \tag{1.1}$$

we have as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} f\big(x_1[i], \dots, x_k[i]\big) \to \mathbb{E}\big[f(X_1, \dots, X_k)\big]. \tag{1.2}$$

We write

$$(\mathbf{x}_1, \dots, \mathbf{x}_k) \overset{W}{\to} (X_1, \dots, X_k)$$

to mean that the above Wasserstein-p convergence holds for every order $p \ge 1$.

For a function $f : \mathbb{R}^k \to \mathbb{R}$ and vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$, we denote by $f(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^n$ the evaluation of $f(\cdot)$ on each row of $(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^{n\times k}$. We write $\langle\cdot\rangle$ for the empirical average of the coordinates of a vector, and introduce the shorthand $\mathbf{x}_{1:k} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ and $X_{1:k} = (X_1, \dots, X_k)$. Thus (1.2) may be expressed as $\langle f(\mathbf{x}_{1:k})\rangle \to \mathbb{E}[f(X_{1:k})]$.

For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \odot \mathbf{y} \in \mathbb{R}^n$ is their entrywise product. $\mathrm{diag}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ or $\mathrm{diag}(\mathbf{x}) \in \mathbb{R}^{n\times\cdots\times n}$ denotes the diagonal matrix or tensor with $\mathbf{x}$ along the main diagonal, that is, $\mathrm{diag}(\mathbf{x})[i, \dots, i] = \mathbf{x}[i]$ and $\mathrm{diag}(\mathbf{x})$ has all other entries equal to 0. For $\mathbf{x} \in \mathbb{R}^{\min(m,n)}$, we write also $\mathrm{diag}(\mathbf{x}) \in \mathbb{R}^{m\times n}$ for the rectangular diagonal matrix where each (i, i) entry is $x[i]$; we will indicate the dimensions if needed to disambiguate these notation. $\|\mathbf{W}\|_{\mathrm{op}}$ is the $\ell_2 \to \ell_2$ operator norm of the matrix $\mathbf{W}$. We denote $[n] = \{1, \dots, n\}$, and reserve Roman letters $i, j, \dots$ for indices in $[n]$ and Greek letters $\alpha, \beta, \dots$ for indices in $[m]$.

2. Main results.

2.1. *Universality of AMP algorithms for symmetric matrices.* Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a symmetric random matrix. Consider an initialization $\mathbf{u}_1 \in \mathbb{R}^n$ and auxiliary "side information"

vectors $\mathbf{f}_1, \dots, \mathbf{f}_k \in \mathbb{R}^n$, independent of $\mathbf{W}$. In applications, such side information vectors may play the role of the external field in spin glass models, the true signal vector in spiked matrix models, or the signal and residual error vectors in regression models. We refer to [7, 61, 62] for several examples. Let $u_2, u_3, u_4, \dots$ be a sequence of nonlinear functions, where $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$. We study a general form for an AMP algorithm with separable nonlinearities that computes, for $t = 1, 2, 3, \dots$

$$\mathbf{z}_t = \mathbf{W}\mathbf{u}_t - \sum_{s=1}^{t} b_{ts}\mathbf{u}_s, \tag{2.1a}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{z}_1, \dots, \mathbf{z}_t, \mathbf{f}_1, \dots, \mathbf{f}_k), \tag{2.1b}$$

where $\{b_{ts}\}_{s \le t}$ are deterministic scalar "Onsager correction" coefficients. We will characterize the iterates of this algorithm in the large system limit as $n \to \infty$, for fixed $k \ge 0$.

We assume throughout the following conditions for $(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k)$.

ASSUMPTION 2.1. Almost surely as $n \to \infty$,

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \overset{W}{\to} (U_1, F_1, \dots, F_k) \tag{2.2}$$

for a joint limit law $(U_1, F_1, \dots, F_k)$ having finite moments of all orders, where $\mathbb{E}[U_1^2] > 0$. Furthermore, multivariate polynomials are dense in the real L^2-space of functions $f : \mathbb{R}^{k+1} \to \mathbb{R}$ with inner-product

$$(f, g) \mapsto \mathbb{E}\big[f(U_1, F_1, \dots, F_k)g(U_1, F_1, \dots, F_k)\big].$$

REMARK 2.2. The convergence (2.2) holds, for example, if rows of $(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \in \mathbb{R}^{n \times (k+1)}$ are i.i.d. and equal in law to $(U_1, F_1, \dots, F_k)$. The density of polynomials holds if $\|(U_1, F_1, \dots, F_k)\|_2$ has finite moment generating function in a neighborhood of 0; see [66], Section 14.1 and Corollary 14.24.

In an AMP algorithm, the coefficients $\{b_{ts}\}$ of (2.1) are defined so that the iterates $\{\mathbf{z}_t\}$ are described by a simple *state evolution* in the asymptotic limit as $n \to \infty$. For $\mathbf{W} \sim \mathrm{GOE}(n)$ (cf. Definition 2.3), this may be done as follows: Set $\mathbf{\Sigma}_1 = \mathbb{E}[U_1^2] \in \mathbb{R}^{1 \times 1}$. Inductively, having defined $\mathbf{\Sigma}_t \in \mathbb{R}^{t \times t}$, let $Z_{1:t} \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ be independent of $(U_1, F_{1:k})$, set $U_{s+1} = u_{s+1}(Z_{1:s}, F_{1:k})$ for each $s = 1, \dots, t$, and define

$$\mathbf{\Sigma}_{t+1} = \big(\mathbb{E}[U_r U_s]\big)_{r,s=1}^{t+1} \in \mathbb{R}^{(t+1)\times(t+1)}. \tag{2.3}$$

Let $b_{tt} = 0$, and for each $s < t$, define the coefficient b_{ts} as

$$b_{ts} = \mathbb{E}\big[\partial_s u_t(Z_{1:t-1}, F_{1:k})\big], \tag{2.4}$$

where $\partial_s u_t$ is the partial derivative of $u_t(\cdot)$ in its $s^t h$ argument. We will call (2.3) and (2.4) the *GOE prescriptions* for $\mathbf{\Sigma}_t$ and b_{ts}. Results of [7, 43] (see also [57], Proposition 2.1, for this form) then imply that, for any Lipschitz functions $u_t(\cdot)$, the iterates of (2.1) satisfy the state evolution, almost surely as $n \to \infty$ for any fixed $t \ge 1$,

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W}{\to} (U_1, F_1, \dots, F_k, Z_1, \dots, Z_t).$$

We note that a variant of this algorithm may instead use the empirical average $b_{ts} = \langle \partial_s u_t(\mathbf{z}_{1:t-1}, \mathbf{f}_{1:k}) \rangle$, for which the same state evolution continues to hold (cf. Remark 2.9).

In [37], building upon work of [60], an extension of this result was proven for a larger class of orthogonally invariant matrices and nonlinear functions: We say that $\mathbf{W}$ is *orthogonally invariant* if it has spectral decomposition $\mathbf{W} = \mathbf{ODO}^\top$ where $\mathbf{O} \sim \mathrm{Haar}(\mathbb{O}(n))$ is Haar-distributed on the orthogonal group and independent of $\mathbf{D} = \mathrm{diag}(\mathbf{d})$. Suppose that $\mathbf{d} \overset{W}{\to} D$ as

$n \to \infty$, where D represents the limit spectral law of $\mathbf{W}$. Set $\mathbf{\Sigma}_1 = \mathrm{Var}[D] \cdot \mathbb{E}[U_1^2] \in \mathbb{R}^{1\times 1}$. Having defined $\mathbf{\Sigma}_t \in \mathbb{R}^{t\times t}$, let $Z_{1:t} \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ be independent of $(U_1, F_{1:k})$, let $U_{s+1} = u_{s+1}(Z_{1:s}, F_{1:k})$ for each $s = 1, \dots, t$, and define

$$\mathbf{\Sigma}_{t+1} = \mathbf{\Sigma}_{t+1}\big(\{\mathbb{E}[U_r U_s]\}_{1\le r,s\le t+1}, \{\mathbb{E}\big[\partial_r u_{s+1}(Z_{1:s}, F_{1:k})\big]\}_{1\le r\le s\le t}\big) \in \mathbb{R}^{(t+1)\times(t+1)} \tag{2.5}$$

for a continuous function $\mathbf{\Sigma}_{t+1}(\cdot)$ whose form depends only on the law of D. For each $s \le t$ and a continuous function $b_{ts}(\cdot)$ whose form also depends only on the law of D, define

$$b_{ts} = b_{ts}\big(\{\mathbb{E}[U_q U_r]\}_{1\le q,r\le t}, \{\mathbb{E}\big[\partial_q u_{r+1}(Z_{1:r}, F_{1:k})\big]\}_{1\le q\le r<t}\big). \tag{2.6}$$

We will call (2.5) and (2.6) the *orthogonally invariant prescriptions* for $\mathbf{\Sigma}_t$ and b_{ts}. We refer to [37], Section 4, for their precise functional forms, which will not be important for our current work. When $\mathbf{W} \sim \mathrm{GOE}(n)$ and D has Wigner's semicircle law on $[-2, 2]$, these reduce to the previous GOE prescriptions of (2.3) and (2.4). It was shown in [37] that for weakly differentiable functions $u_t(\cdot)$ whose derivatives have at most polynomial growth, the iterates of (2.1) again satisfy the state evolution, almost surely as $n \to \infty$ for any fixed $t \ge 1$,

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W}{\to} (U_1, F_1, \dots, F_k, Z_1, \dots, Z_t).$$

Our main results are universality statements that extend the state evolution characterizations of these AMP algorithms to more general random matrix ensembles. Corresponding to $\mathbf{W} \sim \mathrm{GOE}(n)$, we study the following universality class of generalized Wigner matrices, having possibly heteroskedastic entrywise variances and heavy-tailed entries.

DEFINITION 2.3. $\mathbf{W} \in \mathbb{R}^{n\times n}$ is a *generalized Wigner matrix* with (deterministic) variance profile $\mathbf{S} \in \mathbb{R}^{n\times n}$ if

(a) $\mathbf{W}$ is symmetric, and entries on and above the diagonal $(W[i, j] : 1 \le i \le j \le n)$ are independent.

(b) Each $W[i, j]$ has mean 0, variance $n^{-1}S[i, j]$, and higher moments satisfying, for each integer $p \ge 3$,

$$\lim_{n\to\infty} n \cdot \max_{i,j=1}^{n} \mathbb{E}\big[|W[i, j]|^p\big] = 0.$$

(c) For a constant $C > 0$ independent of n,

$$\max_{i,j=1}^{n} S[i, j] \le C \quad \text{and} \quad \lim_{n\to\infty} \max_{i=1}^{n} \left|\frac{1}{n}\sum_{j=1}^{n} S[i, j] - 1\right| = 0.$$

We write $\mathbf{W} \sim \mathrm{GOE}(n)$ for the special case where $W[i, j] \sim \mathcal{N}(0, 1/n)$ and $S[i, j] = 1$ for all $i < j$, and $W[i, i] \sim \mathcal{N}(0, 2/n)$ and $S[i, i] = 2$ for all i.

The moment assumption in condition (b) weakens a uniform sub-Gaussianity condition for $\sqrt{n}W[i, j]$ that is assumed in the previous AMP universality results of [6, 22] and that would require instead $\mathbb{E}[|W[i, j]|^p] \lesssim n^{-p/2}$ for all $p \ge 3$. This condition (b) is weak enough to encompass centered and normalized adjacency matrices of sparse random graphs with slowly growing average vertex degree. Condition (c) allows general patterns of entrywise variances whose rows and columns have approximately the same sum, where we also require in (2.7) of Theorem 2.4 below that these rows and columns are "asymptotically unaligned" with the initialization and side information vectors $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k$. We discuss two applications in Examples 2.24 and 2.25 of Section 2.4.

The following theorem shows that the state evolution of AMP algorithms for GOE random matrices remains valid for matrices $\mathbf{W}$ in this generalized Wigner universality class.

THEOREM 2.4. *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a generalized Wigner matrix with variance profile $\mathbf{S}$, and let $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k$ be independent of $\mathbf{W}$ and satisfy Assumption 2.1. Suppose that*

1. *Each function $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuous, satisfies the polynomial growth condition (1.1) for some order $p \geq 1$, and is Lipschitz in its first t arguments.*
2. *$\|\mathbf{W}\|_{\mathrm{op}} < C$ for a constant $C > 0$ almost surely for all large n.*
3. *Let $\mathbf{s}_i$ be the $i^t h$ row of $\mathbf{S}$. For any fixed polynomial function $q : \mathbb{R}^{k+1} \to \mathbb{R}$, almost surely as $n \to \infty$,*

$$\max_{i=1}^{n} \big|\langle q(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \odot \mathbf{s}_i \rangle - \langle q(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \rangle \cdot \langle \mathbf{s}_i \rangle \big| \to 0. \tag{2.7}$$

Let $\{b_{ts}\}$ and $\{\mathbf{\Sigma}_t\}$ be defined by the GOE prescriptions (2.3) and (2.4), where each matrix $\mathbf{\Sigma}_t$ is nonsingular. Then for any fixed $t \geq 1$, almost surely as $n \to \infty$, the iterates of (2.1) satisfy

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W_2}{\to} (U_1, F_1, \dots, F_k, Z_1, \dots, Z_t),$$

where $(Z_1, \dots, Z_t) \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ is independent of $(U_1, F_1, \dots, F_k)$, that is, this limit has the same joint law as described by the AMP state evolution for $\mathbf{W} \sim \mathrm{GOE}(n)$.

Next, corresponding to orthogonal invariance, we study universality classes of matrices that are permutation-and-sign-invariant in law and that have limit distributions over the diagonal, in the following sense inspired by [4]: Let $\Delta : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be the diagonal map that preserves only the entries on the diagonal, that is,

$$\Delta(\mathbf{M}) = \mathrm{diag}\big(M[1,1], \dots, M[n,n]\big) \in \mathbb{R}^{n\times n}.$$

Let $\Delta\langle \mathbf{x} \rangle$ denote the set of all words in $\mathbf{x}$ and $\Delta(\cdot)$, for example,

$$\mathbf{xx}, \qquad \mathbf{x}\Delta(\mathbf{xx})\mathbf{x}, \qquad \Delta\big(\mathbf{xx}\Delta(\mathbf{x})\big)\mathbf{x}, \qquad \mathbf{xxx}\Delta\big(\Delta(\mathbf{x})\big)\Delta(\mathbf{xx}).$$

We refer to $\Delta\langle \mathbf{x} \rangle$ as the set of *diagonal monomials* in $\mathbf{x}$. For $p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle$ and $\mathbf{M} \in \mathbb{R}^{n\times n}$, we write $p(\mathbf{M}) \in \mathbb{R}^{n\times n}$ for its evaluation at $\mathbf{x} = \mathbf{M}$.

DEFINITION 2.5. The *distribution over the diagonal* of $\mathbf{M}$ is the mapping[1]

$$p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle \mapsto \frac{1}{n} \operatorname{Tr} p(\mathbf{M}).$$

Matrices $\mathbf{M} \in \mathbb{R}^{n\times n}$ *converge in diagonal distribution a.s.* if $\lim_{n\to\infty} \frac{1}{n} \operatorname{Tr} p(\mathbf{M})$ exists almost surely (and is finite) for every fixed $p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle$. The *limit diagonal distribution* of $\mathbf{M}$, which we will refer to as $\mathcal{D}_{\mathrm{diag}}$, is then the mapping

$$p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle \mapsto \lim_{n\to\infty} \frac{1}{n} \operatorname{Tr} p(\mathbf{M}).$$

We remark that $\mathcal{D}_{\mathrm{diag}}$ specifies the limit of $\frac{1}{n} \operatorname{Tr} \mathbf{M}^v$ for each fixed integer $v \geq 1$, and hence also the limit spectral distribution of $\mathbf{M}$ when this distribution has compact support.

We call $\mathbf{\Pi} = \mathbf{P}\mathbf{\Xi} \in \mathbb{R}^{n\times n}$ a *uniformly random signed permutation matrix* if $\mathbf{\Xi} = \mathrm{diag}(\xi[1], \dots, \xi[n]) \in \mathbb{R}^{n\times n}$ where each diagonal entry $\xi[i]$ is independently chosen from $\{+1, -1\}$ with equal probability, and $\mathbf{P} \in \mathbb{R}^{n\times n}$ is a uniformly random permutation matrix independent of $\mathbf{\Xi}$. Note that for any symmetric matrix $\mathbf{M}$ and signed permutation matrix $\mathbf{\Pi}$, we have $\Delta(\mathbf{\Pi}\mathbf{M}\mathbf{\Pi}^\top) = \mathbf{\Pi}\Delta(\mathbf{M})\mathbf{\Pi}^\top$, so also $p(\mathbf{\Pi}\mathbf{M}\mathbf{\Pi}^\top) = \mathbf{\Pi}p(\mathbf{M})\mathbf{\Pi}^\top$ for every diagonal monomial $p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle$. In particular, $\mathbf{M}$ and $\mathbf{\Pi}\mathbf{M}\mathbf{\Pi}^\top$ have the same distributions over the diagonal. The following then defines our universality class.

[1] We define the distribution over the diagonal by the values of $\frac{1}{n} \operatorname{Tr} p(\mathbf{M}) \in \mathbb{R}$ rather than $\Delta(p(\mathbf{M})) \in \mathbb{R}^{n\times n}$ as might be more standard in operator-valued free probability.

DEFINITION 2.6. $\mathbf{W} = \mathbf{\Pi M \Pi}^\top \in \mathbb{R}^{n\times n}$ is a *symmetric generalized invariant matrix*[2] with limit diagonal distribution $\mathcal{D}_{\text{diag}}$ if, as $n \to \infty$,

(a) $\mathbf{M} \in \mathbb{R}^{n\times n}$ converges in diagonal distribution a.s. to a limit $\mathcal{D}_{\text{diag}}$.
(b) For any $\varepsilon > 0$ and any fixed $p(\mathbf{x}) \in \Delta\langle \mathbf{x}\rangle$, almost surely for all large n,

$$\max_{i\neq j} \big| p(\mathbf{M})[i,j] \big| < n^{-1/2+\varepsilon}.$$

(c) $\mathbf{\Pi} \in \mathbb{R}^{n\times n}$ is a uniformly random signed permutation, independent of $\mathbf{M}$.

Our result on AMP universality will pertain specifically to such matrices $\mathbf{W}$ whose limit diagonal distribution $\mathcal{D}_{\text{diag}}$ coincides with that of an orthogonally invariant matrix. In this setting, the next proposition clarifies that $\mathcal{D}_{\text{diag}}$ is determined uniquely by the limit spectral law of $\mathbf{W}$, and it also provides simpler conditions inspired by the spectral universality class in [34] that imply Definition 2.6. We have stated Definition 2.6 for more general limits $\mathcal{D}_{\text{diag}}$ because, as discussed in Remark 2.15 to follow, we will in fact prove a general lemma showing the existence and universality of the limit empirical distribution of iterates for first-order iterative algorithms applied to any such matrix $\mathbf{W}$, even if $\mathcal{D}_{\text{diag}}$ does not correspond to an orthogonally-invariant model.

PROPOSITION 2.7. *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues $\mathbf{d} \in \mathbb{R}^n$ satisfying $\mathbf{d} \overset{W}{\to} D$ almost surely as $n \to \infty$, where D has finite moments of all orders.*

(a) *If $\mathbf{W}$ is orthogonally invariant, then $\mathbf{W}$ is a symmetric generalized invariant matrix in the sense of Definition 2.6, and its limit diagonal distribution $\mathcal{D}_{\text{diag}}$ is determined uniquely by the law of D.*
(b) *Suppose that either*:

1. *$\mathbf{W} = \mathbf{ODO}^\top$ where $\mathbf{D} = \text{diag}(\mathbf{d})$ and $\mathbf{O} = \mathbf{\Pi}_V \mathbf{H} \mathbf{\Pi}_E$, such that $\mathbf{\Pi}_V, \mathbf{\Pi}_E \in \mathbb{R}^{n\times n}$ are uniformly random signed permutations independent of each other and of $(\mathbf{D}, \mathbf{H})$, and $\mathbf{H}$ is an orthogonal matrix with entries satisfying*

$$\max_{i,j\in[n]} \big| H[i,j] \big| < n^{-1/2+\varepsilon} \tag{2.8}$$

for any fixed $\varepsilon > 0$, almost surely for all large n.

2. *$\mathbf{W} = \mathbf{\Pi M \Pi}^\top$ such that $\mathbf{\Pi}$ is a uniformly random signed permutation independent of $\mathbf{M}$ (which has eigenvalues $\mathbf{d}$), and for each fixed integer $\nu \geq 1$, the matrix $\mathbf{M}^\nu$ satisfies*

$$\max_{i=1}^{n} \left| M^\nu[i,i] - \frac{1}{n} \text{Tr}\, \mathbf{M}^\nu \right| < n^{-1/2+\varepsilon}, \qquad \max_{i\neq j} \big| M^\nu[i,j] \big| < n^{-1/2+\varepsilon} \tag{2.9}$$

for any fixed $\varepsilon > 0$, almost surely for all large n.

Then $\mathbf{W}$ is a generalized invariant matrix in the sense of Definition 2.6, and its limit diagonal distribution $\mathcal{D}_{\text{diag}}$ coincides with that of the orthogonally invariant matrix in part (a).

We prove Proposition 2.7 in Appendix B. Important examples for applications are when $\mathbf{W}$ is a composition of permutations, deterministic Hadamard/Fourier matrices, and diagonal operators. We discuss one such application to compressed sensing in Example 2.26 of Section 2.4.

[2]More formally, these definitions of generalized Wigner and generalized invariant matrices are describing sequences of matrices $\mathbf{W} \in \mathbb{R}^{n\times n}$ of increasing dimensions $n \to \infty$, rather than a single matrix. We will choose not make this terminological distinction in our work.

The following is our main theorem on AMP universality in this context, showing that the state evolution of AMP algorithms for orthogonally invariant matrices holds universally over the class of generalized invariant matrices with matching limit diagonal distribution.

THEOREM 2.8. *Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ be a symmetric generalized invariant matrix whose limit diagonal distribution $\mathcal{D}_{\mathrm{diag}}$ coincides with that of an orthogonally invariant matrix $\mathbf{G}$. Let $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k$ be independent of $\mathbf{W}$ and satisfy Assumption 2.1. Suppose that*

1. *Each function $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuous, satisfies the polynomial growth condition (1.1) for some order $p \geq 1$, and is Lipschitz in its first t arguments.*
2. *$\|\mathbf{W}\|_{\mathrm{op}} < C$ for a constant $C > 0$ almost surely for all large n.*

Let $\{b_{ts}\}$ and $\{\mathbf{\Sigma}_t\}$ be defined by the orthogonally invariant prescriptions (2.5) and (2.6) for the limit spectral distribution D specified by $\mathcal{D}_{\mathrm{diag}}$. Suppose that $\mathrm{Var}[D] > 0$ and each matrix $\mathbf{\Sigma}_t$ is nonsingular. Then for any fixed $t \geq 1$, almost surely as $n \to \infty$, the iterates of (2.1) satisfy

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W_2}{\to} (U_1, F_1, \dots, F_k, Z_1, \dots, Z_t),$$

where $(Z_1, \dots, Z_t) \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ is independent of $(U_1, F_1, \dots, F_k)$, that is, this limit has the same joint law as described by the AMP state evolution for $\mathbf{G}$.

REMARK 2.9. Theorems 2.4 and 2.8 hold equally for AMP algorithms where, in the prescriptions (2.4) and (2.6) for b_{ts}, the quantities $\mathbb{E}[\partial_r u_{s+1}(Z_{1:s}, F_{1:k})]$ and $\mathbb{E}[U_r U_s]$ are replaced by the empirical averages

$$\big\langle \partial_r u_{s+1}(\mathbf{z}_{1:s}, \mathbf{f}_{1:k}) \big\rangle = \frac{1}{n}\sum_{i=1}^{n} \partial_r u_{s+1}\big(z_{1:s}[i], f_{1:k}[i]\big), \qquad \langle \mathbf{u}_r \odot \mathbf{u}_s \rangle = \frac{1}{n}\sum_{i=1}^{n} u_r[i] u_s[i].$$

For example, such an AMP algorithm for GOE matrices $\mathbf{W}$ and nonlinearities $u_{t+1}(z_{1:t}, f_{1:k}) = u_{t+1}(z_t)$ consists of the iterations

$$\mathbf{z}_t = \mathbf{W}\mathbf{u}_t - \big\langle u_t'(\mathbf{z}_{t-1}) \big\rangle \mathbf{u}_{t-1}, \qquad \mathbf{u}_{t+1} = u_{t+1}(\mathbf{z}_t).$$

To see this, note that b_{11} depends only on $\mathbb{E}[U_1^2]$, so these prescriptions for b_{11} asymptotically coincide by Assumption 2.1. Then the state evolution holds for $\mathbf{z}_1$. Inductively, validity of the state evolution for $\mathbf{z}_{1:t}$ ensures that, almost surely as $n \to \infty$,

$$\big\langle \partial_r u_{s+1}(\mathbf{z}_{1:s}, \mathbf{f}_{1:k}) \big\rangle \to \mathbb{E}\big[\partial_r u_{s+1}(Z_{1:s}, F_{1:k})\big] \quad \text{for all } r \leq s \leq t,$$
$$\langle \mathbf{u}_r \odot \mathbf{u}_s \rangle \to \mathbb{E}[U_r U_s] \quad \text{for all } r, s \leq t+1,$$

where the first statement follows from Wasserstein-2 convergence of $(\mathbf{z}_{1:s}, \mathbf{f}_{1:k})$ and Stein's lemma (cf. [37], Proposition E.5). Then the presciptions of (2.4) and (2.6) for $\{b_{t+1,s}\}_{s\leq t+1}$ asymptotically coincide with their empirical versions defined by $\langle \partial_r u_{s+1}(\mathbf{z}_{1:s}, \mathbf{f}_{1:k}) \rangle$ and $\langle \mathbf{u}_r \odot \mathbf{u}_s \rangle$, which in turn implies validity of the state evolution for $\mathbf{z}_{1:(t+1)}$.

REMARK 2.10. Theorems 2.4 and 2.8 show universality of AMP algorithms with an initialization $\mathbf{u}_1$ that is independent of $\mathbf{W}$. For spiked matrix models with a low-rank signal component, alternative AMP algorithms with spectral initializations have been studied for example, in [56, 58, 76]. Universality for such algorithms may be shown using the preceding results, by approximating the spectral initialization with a large number of linear AMP iterations starting from an initialization $\mathbf{u}_1$ that is, independent of $\mathbf{W}$ but correlated with the true signal; we refer to [22], Section 8, and [76], Section A.2, for examples of this type of argument.

Since we allow the nonlinearities $u_{t+1}(\cdot)$ to be functions of all preceding iterates $\mathbf{z}_1, \dots, \mathbf{z}_t$, universality of AMP with matrix-valued iterates in $\mathbb{R}^{n\times J}$ for a fixed dimension $J \geq 1$ may also be deduced from the preceding results, by simulating each iteration of any such algorithm using J iterations of an algorithm with iterates in $\mathbb{R}^n$. We leave the further study of these extensions to future work, as the need arises in applications.

2.2. *Tensor networks and strategy of proof.* We describe here our high-level strategy of proof for Theorems 2.4 and 2.8. The full proofs of these results are contained in Section 3.

DEFINITION 2.11. A *diagonal tensor network* $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$ in k variables is an undirected tree graph with vertices $\mathcal{V}$ and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, each of whose vertices $v \in \mathcal{V}$ is labeled by a polynomial function $q_v : \mathbb{R}^k \to \mathbb{R}$. The *value* of T on a symmetric matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ and vectors $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ is

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \frac{1}{n} \sum_{\mathbf{i}\in[n]^{\mathcal{V}}} q_{\mathbf{i}|T} \cdot W_{\mathbf{i}|T},$$

where, for each index tuple $\mathbf{i} = (i_v : v \in \mathcal{V}) \in [n]^{\mathcal{V}}$, we set

$$q_{\mathbf{i}|T} = \prod_{v\in\mathcal{V}} q_v(x_1[i_v], \dots, x_k[i_v]), \qquad W_{\mathbf{i}|T} = \prod_{(u,v)\in\mathcal{E}} W[i_u, i_v].$$

This value may be understood as:

1. Associating to each vertex $v \in \mathcal{V}$ a diagonal tensor $\mathbf{T}_v = \mathrm{diag}(q_v(\mathbf{x}_1, \dots, \mathbf{x}_k)) \in \mathbb{R}^{n\times\cdots\times n}$, where the order of this tensor equals the degree of v in the tree.[3]
2. Associating to each edge the symmetric matrix $\mathbf{W}$.
3. Iteratively contracting all tensor-matrix-tensor products represented by the edges of the tree.

For example, if $\mathcal{V} = [w+1]$ and T is the line graph $1-2-\cdots-w-(w+1)$, then $\mathbf{T}_1, \mathbf{T}_{w+1} \in \mathbb{R}^n$ are vectors, $\mathbf{T}_v \in \mathbb{R}^{n\times n}$ is a diagonal matrix for each vertex $v \in \{2, \dots, w\}$, and the value (in usual matrix-vector product notation) is

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \frac{1}{n}\mathbf{T}_1^\top \mathbf{W}\mathbf{T}_2\mathbf{W}\cdots\mathbf{W}\mathbf{T}_w\mathbf{W}\mathbf{T}_{w+1}.$$

When each tensor $\mathbf{T}_v$ has all 1's along the main diagonal, this definition is an example of the graph sum used to show asymptotic freeness of Wigner and diagonal matrices in [53, 54], and it is also a specific case of a "graph monomial" in the notion of traffic freeness in [52].

We will show in Lemma 3.10 that for any AMP algorithm (or more generally, any first-order iterative algorithm of the form (2.1)) with polynomial nonlinearities $u_2, u_3, u_4, \dots$, and for any polynomial test function $p(\cdot)$, the coordinate average

$$\langle p(\mathbf{u}_{1:t}, \mathbf{z}_{1:t}, \mathbf{f}_{1:k})\rangle = \frac{1}{n}\sum_{i=1}^n p\big(u_{1:t}[i], z_{1:t}[i], f_{1:k}[i]\big)$$

of $p(\cdot)$ evaluated on the AMP iterates and side information vectors is a linear combination of values of different tensor networks on $\mathbf{W}$ and $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k$. Then, leveraging state evolution results of [37] to perform an inductive polynomial approximation argument, the proof reduces the universality of AMP for Lipschitz nonlinearities to the universality of these tensor network values. This reduction is encapsulated in the following lemma.

[3]We remind readers our notation that $q_v(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^n$ indicates the application of $q_v : \mathbb{R}^k \to \mathbb{R}$ row-wise to $(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^{n\times k}$, and $\mathbf{T}_v$ is then a diagonal tensor with $q_v(\mathbf{x}_1, \dots, \mathbf{x}_k) \in \mathbb{R}^n$ along its main diagonal.

LEMMA 2.12. *Let* $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k \in \mathbb{R}^n$ *satisfy Assumption* 2.1. *Let* $\mathbf{W}, \mathbf{G} \in \mathbb{R}^{n\times n}$ *be symmetric random matrices independent of* $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k$ *such that*

1. $\mathbf{G} = \mathbf{ODO}^\top$ *is an orthogonally invariant matrix, where* $\mathbf{D} = \operatorname{diag}(\mathbf{d})$ *and* $\mathbf{d} \overset{W}{\to} D$ *for a limit law* D *with compact support and* $\operatorname{Var}[D] > 0$.
2. $\|\mathbf{W}\|_{\mathrm{op}} < C$ *for a constant* $C > 0$, *almost surely for all large* n.
3. *For every diagonal tensor network* T *in* $k+1$ *variables, almost surely as* $n \to \infty$,

$$\operatorname{val}_T(\mathbf{W}; \mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) - \operatorname{val}_T(\mathbf{G}; \mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \to 0.$$

Let $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ *be continuous functions which satisfy the polynomial growth condition* (1.1) *for some order* $p \geq 1$, *and are Lipschitz in their first* t *arguments. Let* $\{b_{ts}\}$ *and* $\{\mathbf{\Sigma}_t\}$ *be defined by the orthogonally invariant prescriptions* (2.5) *and* (2.6) *for the limit law* D, *where each* $\mathbf{\Sigma}_t$ *is nonsingular. Then the iterates* (2.1) *applied to* $\mathbf{W}$ *satisfy, almost surely as* $n \to \infty$ *for any fixed* $t \geq 1$,

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W_2}{\to} (U_1, F_1, \dots, F_k, Z_1, \dots, Z_t),$$

where this limit has the same joint law as described by the AMP state evolution for $\mathbf{G}$.

This lemma applies also in the special case of $\mathbf{G} \sim \mathrm{GOE}(n)$, where the definitions of $\{b_{ts}\}$ and $\{\mathbf{\Sigma}_t\}$ reduce to the GOE prescriptions of (2.3) and (2.4). The lemma does not assume any particular matrix model for $\mathbf{W}$, and thus may be used as a tool to establish AMP universality for matrix models beyond the ones we consider in this work.

Theorems 2.4 and 2.8 then follow from the next two lemmas, which verify the universality of tensor network values for the classes of generalized Wigner matrices and symmetric generalized invariant matrices.

LEMMA 2.13. *Let* $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ *be* (*random or deterministic*) *vectors and let* $(X_1, \dots, X_k)$ *have finite moments of all orders, such that almost surely as* $n \to \infty$,

$$(\mathbf{x}_1, \dots, \mathbf{x}_k) \overset{W}{\to} (X_1, \dots, X_k). \tag{2.10}$$

Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ *be a generalized Wigner matrix, independent of* $\mathbf{x}_1, \dots, \mathbf{x}_k$, *with variance profile matrix* $\mathbf{S}$. *Let* $\mathbf{s}_i$ *be the* $i^t h$ *row of* $\mathbf{S}$, *and suppose for each fixed polynomial function* $q : \mathbb{R}^k \to \mathbb{R}$ *that*

$$\max_{i=1}^{n} \left| \langle q(\mathbf{x}_1, \dots, \mathbf{x}_k) \odot \mathbf{s}_i \rangle - \langle q(\mathbf{x}_1, \dots, \mathbf{x}_k) \rangle \cdot \langle \mathbf{s}_i \rangle \right| \to 0. \tag{2.11}$$

Then for any diagonal tensor network T *in* k *variables, there is a deterministic value* $\text{lim-val}_T(X_1, \dots, X_k)$ *depending only on* T *and the joint law of* $(X_1, \dots, X_k)$ *such that almost surely,*

$$\lim_{n\to\infty} \operatorname{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \text{lim-val}_T(X_1, \dots, X_k).$$

In particular, this limit value is the same for $\mathbf{W}$ *as for* $\mathbf{G} \sim \mathrm{GOE}(n)$.

LEMMA 2.14. *Let* $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ *be* (*random or deterministic*) *vectors and let* $(X_1, \dots, X_k)$ *have finite moments of all orders, such that almost surely as* $n \to \infty$,

$$(\mathbf{x}_1, \dots, \mathbf{x}_k) \overset{W}{\to} (X_1, \dots, X_k). \tag{2.12}$$

Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ *be a symmetric generalized invariant matrix, independent of* $\mathbf{x}_1, \dots, \mathbf{x}_k$, *with limit diagonal distribution* $\mathcal{D}_{\mathrm{diag}}$. *Then for any diagonal tensor network* T *in* k *variables,*

there is a deterministic limit value $\lim\text{-val}_T(X_1, \dots, X_k, \mathcal{D}_{\text{diag}})$ *depending only on* T, *the joint law of* $(X_1, \dots, X_k)$, *and* $\mathcal{D}_{\text{diag}}$ *such that almost surely,*

$$\lim_{n\to\infty} \text{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \lim\text{-val}_T(X_1, \dots, X_k, \mathcal{D}_{\text{diag}}).$$

In particular, if there exists an orthogonally invariant matrix $\mathbf{G}$ *having the same limit diagonal distribution* $\mathcal{D}_{\text{diag}}$, *then this limit value is the same for* $\mathbf{W}$ *as for* $\mathbf{G}$.

REMARK 2.15. Lemma 2.14 applies to any class of symmetric generalized invariant matrices satisfying Definition 2.6, where the limit diagonal distribution $\mathcal{D}_{\text{diag}}$ does not necessarily coincide with that of an orthogonally invariant model.

This has the following implication: Consider any first-order iterative algorithm having the structure (2.1), where b_{ts} are arbitrary fixed constants and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ are polynomial functions applied entrywise. Then for any polynomial test function $p(\cdot)$, the value

$$\frac{1}{n}\sum_{i=1}^{n} p\big(u_{1:t}[i], z_{1:t}[i], f_{1:k}[i]\big)$$

is a linear combination of tensor network values (cf. Lemma 3.10) and hence has a universal limit as $n \to \infty$. Under mild moment assumptions, this implies that there exists a limit law for the empirical distribution of each iterate $\mathbf{u}_t$ and $\mathbf{z}_t$, and this law is universal across such matrices having the same limit diagonal distribution $\mathcal{D}_{\text{diag}}$.

When $\mathcal{D}_{\text{diag}}$ is not described by an orthogonally invariant model, we believe it may be an interesting open question to develop such an algorithm that has a more succinct state-evolution characterization of its iterates in terms of this limit diagonal law.

The proofs of Lemmas 2.13 and 2.14 result in forms for the limit tensor network values that are, in general, combinatorially complex. However, a by-product of the proofs is that these forms reduce to 0 when all diagonal tensors of the tensor network have vanishing normalized trace. This may be viewed as a version of asymptotic freeness for tensor networks, and we state the result here for independent interest.

PROPOSITION 2.16.

(a) *In the setting of Lemma 2.13, let* T *be a diagonal tensor network such that, for every vertex of* v *of* T, *almost surely*

$$\lim_{n\to\infty} \langle q_v(\mathbf{x}_1, \dots, \mathbf{x}_k)\rangle = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} q_v\big(x_1[i], \dots, x_k[i]\big) = 0. \tag{2.13}$$

Then $\lim\text{-val}_T(X_1, \dots, X_k) = 0$.

(b) *In the setting of Lemma 2.14, suppose* T *is a diagonal tensor network for which* (2.13) *holds almost surely for every vertex* v. *Suppose also that there exists an orthogonally invariant matrix having the same limit diagonal distribution* $\mathcal{D}_{\text{diag}}$ *as* $\mathbf{W}$, *and* $\lim_{n\to\infty} \frac{1}{n}\operatorname{Tr}\mathbf{W} = 0$. *Then* $\lim\text{-val}_T(X_1, \dots, X_k, \mathcal{D}_{\text{diag}}) = 0$.

2.3. *Universality of AMP algorithms for rectangular matrices.* Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be a rectangular random matrix. Consider an initialization $\mathbf{u}_1 \in \mathbb{R}^m$ and vectors of side information $\mathbf{f}_1, \dots, \mathbf{f}_k \in \mathbb{R}^m$ and $\mathbf{g}_1, \dots, \mathbf{g}_\ell \in \mathbb{R}^n$, all independent of $\mathbf{W}$. Let $v_1, v_2, v_3, \dots$ and $u_2, u_3, u_4, \dots$ be two sequences of nonlinear functions where $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} :$

$\mathbb{R}^{t+k} \to \mathbb{R}$. We study an AMP algorithm that computes, for $t = 1, 2, 3, \ldots$

$$\mathbf{z}_t = \mathbf{W}^\top \mathbf{u}_t - \sum_{s=1}^{t-1} b_{ts} \mathbf{v}_s, \tag{2.14a}$$

$$\mathbf{v}_t = v_t(\mathbf{z}_1, \ldots, \mathbf{z}_t, \mathbf{g}_1, \ldots, \mathbf{g}_\ell), \tag{2.14b}$$

$$\mathbf{y}_t = \mathbf{W}\mathbf{v}_t - \sum_{s=1}^{t} a_{ts} \mathbf{u}_s, \tag{2.14c}$$

$$\mathbf{u}_{t+1} = u_{t+1}(\mathbf{y}_1, \ldots, \mathbf{y}_t, \mathbf{f}_1, \ldots, \mathbf{f}_k), \tag{2.14d}$$

where $\{b_{ts}\}_{s<t}$ and $\{a_{ts}\}_{s\le t}$ are deterministic "Onsager correction" coefficients. We will characterize the iterates of this algorithm in the limit as $m, n \to \infty$ proportionally with $m/n \to \gamma \in (0, \infty)$, for fixed $k, \ell \ge 0$. For Gaussian and bi-orthogonally invariant matrices $\mathbf{W}$ (see the definition after Definition 2.20), we review the forms for these correction coefficients and the corresponding state evolutions in (D.1)–(D.2) and (D.4)–(D.5) of Appendix D (see the Supplementary Material [75]).

We assume the following condition for $(\mathbf{u}_1, \mathbf{f}_1, \ldots, \mathbf{f}_k)$ and $(\mathbf{g}_1, \ldots, \mathbf{g}_\ell)$, which is analogous to Assumption 2.1.

ASSUMPTION 2.17. Almost surely as $m, n \to \infty$,

$$(\mathbf{u}_1, \mathbf{f}_1, \ldots, \mathbf{f}_k) \xrightarrow{W} (U_1, F_1, \ldots, F_k) \quad \text{and} \quad (\mathbf{g}_1, \ldots, \mathbf{g}_\ell) \xrightarrow{W} (G_1, \ldots, G_\ell)$$

for joint limit laws $(U_1, F_1, \ldots, F_k)$ and $(G_1, \ldots, G_\ell)$ having finite moments of all orders, where $\mathbb{E}[U_1^2] > 0$. Multivariate polynomials are dense in the real L^2-spaces of functions $f : \mathbb{R}^{k+1} \to \mathbb{R}$ and $g : \mathbb{R}^\ell \to \mathbb{R}$ with the inner-products

$$(f, \tilde{f}) \mapsto \mathbb{E}\big[f(U_1, F_1, \ldots, F_k)\tilde{f}(U_1, F_1, \ldots, F_k)\big],$$
$$(g, \tilde{g}) \mapsto \mathbb{E}\big[g(G_1, \ldots, G_\ell)\tilde{g}(G_1, \ldots, G_\ell)\big].$$

Our main results show that the state evolution characterizations of AMP algorithms for Gaussian and orthogonally invariant matrices are universal across the following matrix ensembles, analogous to Definitions 2.3 and 2.6 in the symmetric setting.

DEFINITION 2.18. $\mathbf{W} \in \mathbb{R}^{m\times n}$ is a *generalized white noise matrix* with (deterministic) variance profile $\mathbf{S} \in \mathbb{R}^{m\times n}$ if

(a) All entries $W[\alpha, i]$ are independent.

(b) Each entry $W[\alpha, i]$ has mean 0, variance $n^{-1}S[\alpha, i]$, and higher moments satisfying, for each integer $p \ge 3$,

$$\lim_{m,n\to\infty} n \cdot \max_{\alpha=1}^{m} \max_{i=1}^{n} \mathbb{E}\big[|W[\alpha, i]|^p\big] = 0.$$

(c) For a constant $C > 0$ independent of m, n,

$$\max_{\alpha=1}^{m} \max_{i=1}^{n} S[\alpha, i] \le C, \qquad \lim_{m,n\to\infty} \max_{\alpha=1}^{m} \left| \frac{1}{n} \sum_{i=1}^{n} S[\alpha, i] - 1 \right| = 0,$$

$$\lim_{m,n\to\infty} \max_{i=1}^{n} \left| \frac{1}{m} \sum_{\alpha=1}^{m} S[\alpha, i] - 1 \right| = 0.$$

We call $\mathbf{W}$ a *Gaussian white noise matrix* in the special case where $W[\alpha, i] \sim \mathcal{N}(0, 1/n)$ and $S[\alpha, i] = 1$ for all $(\alpha, i) \in [m] \times [n]$.

Next, we introduce a notion of diagonal distribution for rectangular matrices, analogous to Definition 2.5. Recall the diagonal map $\Delta(\cdot)$, and let $\Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle$ be the set of all words in $\mathbf{x}$, $\mathbf{I}_m$, $\mathbf{I}_n$ and $\Delta(\cdot)$, for example,

$$\mathbf{x}\mathbf{I}_m, \qquad \mathbf{x}\Delta(\mathbf{I}_n\mathbf{x})\mathbf{I}_m, \qquad \Delta\big(\mathbf{x}\mathbf{x}\Delta(\mathbf{I}_n)\big)\mathbf{x}, \qquad \mathbf{I}_m\mathbf{x}\mathbf{I}_n\Delta\big(\Delta(\mathbf{x})\big)\Delta(\mathbf{x}\mathbf{I}_m).$$

For $p(\mathbf{x}) \in \Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle$ and $\mathbf{M} \in \mathbb{R}^{m\times n}$, we write $p(\widetilde{\mathbf{M}}) \in \mathbb{R}^{(m+n)\times(m+n)}$ for its evaluation at $\mathbf{x} = \widetilde{\mathbf{M}}$,

$$\mathbf{I}_m = \begin{pmatrix} \mathrm{Id}_m & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}, \qquad \mathbf{I}_n = \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id}_n \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n)},$$

where we define the symmetric embedding

$$\widetilde{\mathbf{M}} = \begin{pmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\top & 0 \end{pmatrix} \in \mathbb{R}^{(m+n)\times(m+n)} \tag{2.15}$$

and the identity matrices $\mathrm{Id}_m \in \mathbb{R}^{m\times m}$ and $\mathrm{Id}_n \in \mathbb{R}^{n\times n}$.

DEFINITION 2.19. The distribution over the diagonal of a rectangular matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ is the mapping

$$p(\mathbf{x}) \in \Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle \mapsto \frac{1}{m+n} \operatorname{Tr} p(\widetilde{\mathbf{M}}).$$

Matrices $\mathbf{M} \in \mathbb{R}^{m\times n}$ converge in diagonal distribution a.s. if $\lim_{m,n\to\infty} \frac{1}{m+n} \operatorname{Tr} p(\widetilde{\mathbf{M}})$ exists almost surely (and is finite) for every fixed $p(\mathbf{x}) \in \Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle$, as $m, n \to \infty$ with $m/n \to \gamma \in (0, \infty)$. The limit diagonal distribution of $\mathbf{M}$, which we will refer to as $\mathcal{D}_{\mathrm{diag}}$, is then the mapping

$$p(\mathbf{x}) \in \Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle \mapsto \lim_{m,n\to\infty} \frac{1}{m+n} \operatorname{Tr} p(\widetilde{\mathbf{M}}).$$

Note that $\mathcal{D}_{\mathrm{diag}}$ and γ specify the limit of $\frac{1}{m}\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top)^\nu$ for each fixed integer $\nu \geq 1$, and hence also the limit singular value distribution of $\mathbf{M}$ when this distribution has compact support. Note also that, similar to the symmetric setting, $\mathbf{M}$ and $\mathbf{\Pi}_U\mathbf{M}\mathbf{\Pi}_V^\top$ must have the same limit diagonal distribution $\mathcal{D}_{\mathrm{diag}}$ for any signed permutation matrices $\mathbf{\Pi}_U \in \mathbb{R}^{m\times m}$ and $\mathbf{\Pi}_V \in \mathbb{R}^{n\times n}$.

DEFINITION 2.20. $\mathbf{W} = \mathbf{\Pi}_U\mathbf{M}\mathbf{\Pi}_V^\top \in \mathbb{R}^{m\times n}$ is a *rectangular generalized invariant matrix* with limit diagonal distribution $\mathcal{D}_{\mathrm{diag}}$ if, as $m, n \to \infty$ with $m/n \to \gamma \in (0, \infty)$,

(a) $\mathbf{M}$ converges in diagonal distribution a.s. to a limit $\mathcal{D}_{\mathrm{diag}}$.
(b) For any $\varepsilon > 0$ and any fixed $p(\mathbf{x}) \in \Delta\langle \mathbf{x}, \mathbf{I}_m, \mathbf{I}_n \rangle$, almost surely for all large m, n,

$$\max_{i\neq j} \big|p(\widetilde{\mathbf{M}})[i, j]\big| < n^{-1/2+\varepsilon},$$

where $\widetilde{\mathbf{M}}$ is the symmetric embedding (2.15).
(c) $\mathbf{\Pi}_U \in \mathbb{R}^{m\times m}$ and $\mathbf{\Pi}_V \in \mathbb{R}^{n\times n}$ are uniformly random signed permutations independent of each other and of $\mathbf{M}$.

We call $\mathbf{W} \in \mathbb{R}^{m\times n}$ *bi-orthogonally invariant* if it has singular value decomposition $\mathbf{W} = \mathbf{ODQ}^\top$ where $\mathbf{O} \sim \mathrm{Haar}(\mathbb{O}(m))$ and $\mathbf{Q} \sim \mathrm{Haar}(\mathbb{O}(n))$ are Haar-distributed on the orthogonal groups independently of each other and of $\mathbf{D} = \mathrm{diag}(\mathbf{d}) \in \mathbb{R}^{m\times n}$. We verify in Proposition D.1 of Appendix D that such bi-orthogonally invariant matrices satisfy Definition 2.20, where $\mathcal{D}_{\mathrm{diag}}$ is determined uniquely by $\gamma = \lim_{m,n\to\infty} m/n$ and the limit singular value distribution of $\mathbf{D}$.

The following theorems show that the state evolution of AMP algorithms for Gaussian white noise matrices holds universally for generalized white noise matrices as in Definition 2.18, and the state evolution for bi-orthogonally invariant matrices holds universally for rectangular generalized invariant matrices as in Definition 2.20.

THEOREM 2.21. *Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be a generalized white noise matrix with variance profile matrix $\mathbf{S}$, and let $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{g}_1, \dots, \mathbf{g}_\ell$ be independent of $\mathbf{W}$ and satisfy Assumption 2.17. Suppose that*

1. *Each function $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuous, satisfies the polynomial growth condition (1.1) for some order $p \geq 1$, and is Lipschitz in its first t arguments.*
2. *$\|\mathbf{W}\|_{\mathrm{op}} < C$ for a constant $C > 0$ almost surely for all large m, n.*
3. *Let $\mathbf{s}_\alpha$ be the $\alpha^t h$ row of $\mathbf{S}$ and $\mathbf{s}^i$ be the $i^t h$ column of $\mathbf{S}$. For any fixed polynomial functions $p : \mathbb{R}^{k+1} \to \mathbb{R}$ and $q : \mathbb{R}^\ell \to \mathbb{R}$, almost surely as $m, n \to \infty$,*

$$\begin{aligned} \max_{\alpha=1}^{m} \big|\langle p(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k) \odot \mathbf{s}_\alpha \rangle - \langle p(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k)\rangle \cdot \langle \mathbf{s}_\alpha \rangle\big| &\to 0, \\ \max_{i=1}^{n} \big|\langle q(\mathbf{g}_1, \dots, \mathbf{g}_\ell) \odot \mathbf{s}^i \rangle - \langle q(\mathbf{g}_1, \dots, \mathbf{g}_\ell)\rangle \cdot \langle \mathbf{s}^i \rangle\big| &\to 0. \end{aligned} \tag{2.16}$$

Let $\{a_{ts}\}, \{b_{ts}\}, \{\mathbf{\Omega}_t\}, \{\mathbf{\Sigma}_t\}$ be defined by the white noise prescriptions (D.1) and (D.2), where each matrix $\mathbf{\Omega}_t$ and $\mathbf{\Sigma}_t$ is nonsingular. Then for any fixed $t \geq 1$, almost surely as $m, n \to \infty$ with $m/n \to \gamma \in (0, \infty)$, the iterates of (2.14) satisfy

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{y}_1, \dots, \mathbf{y}_t) \overset{W_2}{\to} (U_1, F_1, \dots, F_k, Y_1, \dots, Y_t),$$

$$(\mathbf{g}_1, \dots, \mathbf{g}_\ell, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W_2}{\to} (G_1, \dots, G_\ell, Z_1, \dots, Z_t),$$

where $(Z_1, \dots, Z_t) \sim \mathcal{N}(0, \mathbf{\Omega}_t)$ and $(Y_1, \dots, Y_t) \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ are independent of $(U_1, F_1, \dots, F_k)$ and $(G_1, \dots, G_\ell)$, that is, these limits have the same joint laws as described by the AMP state evolution for a Gaussian white noise matrix $\mathbf{W}$.

THEOREM 2.22. *Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ be a rectangular generalized invariant matrix whose limit diagonal distribution $\mathcal{D}_{\mathrm{diag}}$ coincides with that of a bi-orthogonally invariant matrix $\mathbf{G}$. Let $\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{g}_1, \dots, \mathbf{g}_\ell$ be independent of $\mathbf{W}$ and satisfy Assumption 2.17. Suppose that*

1. *Each function $v_t : \mathbb{R}^{t+\ell} \to \mathbb{R}$ and $u_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ is continuous, satisfies the polynomial growth condition (1.1) for some order $p \geq 1$, and is Lipschitz in its first t arguments.*
2. *$\|\mathbf{W}\|_{\mathrm{op}} < C$ for a constant $C > 0$ almost surely for all large m, n.*

Let $\{a_{ts}\}, \{b_{ts}\}, \{\mathbf{\Omega}_t\}, \{\mathbf{\Sigma}_t\}$ be defined by the bi-orthogonally invariant prescriptions (D.4) and (D.5) for the limit singular value distribution D specified by $\mathcal{D}_{\mathrm{diag}}$ and γ. Suppose that $\mathbb{E}[D^2] > 0$ and each $\mathbf{\Omega}_t$ and $\mathbf{\Sigma}_t$ is nonsingular. Then for any fixed $t \geq 1$, almost surely as $m, n \to \infty$ with $m/n \to \gamma \in (0, \infty)$, the iterates of (2.14) satisfy

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \mathbf{y}_1, \dots, \mathbf{y}_t) \overset{W_2}{\to} (U_1, F_1, \dots, F_k, Y_1, \dots, Y_t),$$

$$(\mathbf{g}_1, \dots, \mathbf{g}_\ell, \mathbf{z}_1, \dots, \mathbf{z}_t) \overset{W_2}{\to} (G_1, \dots, G_\ell, Z_1, \dots, Z_t),$$

where $(Z_1, \dots, Z_t) \sim \mathcal{N}(0, \mathbf{\Omega}_t)$ *and* $(Y_1, \dots, Y_t) \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ *are independent of* $(U_1, F_1, \dots, F_k)$ *and* $(G_1, \dots, G_\ell)$, *that is, these limits have the same joint laws as described by the AMP state evolution for* $\mathbf{G}$.

REMARK 2.23. As in Remark 2.9, Theorems 2.21 and 2.22 hold equally for AMP algorithms where, in the prescriptions (D.2) and (D.5) for a_{ts} and b_{ts}, the quantities $\mathbb{E}[\partial_r v_s(Z_{1:s}, G_{1:\ell})]$, $\mathbb{E}[\partial_r u_{s+1}(Y_{1:s}, F_{1:k})]$, $\mathbb{E}[U_r U_s]$, and $\mathbb{E}[V_r V_s]$ are replaced by the empirical averages

$$\langle \partial_r v_s(\mathbf{z}_{1:s}, \mathbf{g}_{1:\ell}) \rangle, \qquad \langle \partial_r u_{s+1}(\mathbf{y}_{1:s}, \mathbf{f}_{1:k}) \rangle, \qquad \langle \mathbf{u}_r \odot \mathbf{u}_s \rangle, \qquad \langle \mathbf{v}_r \odot \mathbf{v}_s \rangle.$$

For example, such an AMP algorithm for Gaussian white noise matrices $\mathbf{W}$ and nonlinearities $v_t(z_{1:t}, g_{1:\ell}) = v(z_t)$ and $u_{t+1}(y_{1:t}, f_{1:k}) = u(y_t)$ consists of the iterations

$$\mathbf{z}_t = \mathbf{W}^\top \mathbf{u}_t - \gamma \langle u'(\mathbf{y}_{t-1}) \rangle \mathbf{v}_{t-1}, \qquad \mathbf{v}_t = v(\mathbf{z}_t),$$
$$\mathbf{y}_t = \mathbf{W}\mathbf{v}_t - \langle v'(\mathbf{z}_t) \rangle \mathbf{u}_t, \qquad \mathbf{u}_{t+1} = u(\mathbf{y}_t).$$

The proofs of Theorems 2.21 and 2.22 are similar to those of Theorems 2.4 and 2.8 for symmetric matrices, and we defer them to Appendix D.

2.4. *Applications*.

EXAMPLE 2.24. AMP algorithms for the Gaussian universality class may be heuristically derived by approximating belief propagation on dense graphical models [30, 45]. Our assumptions in Theorems 2.4 and 2.21 are sufficiently weak to show that their state evolutions remain valid in sparse random graphs down to sparsity levels of $(\log n)/n$.

As a concrete example, consider the symmetric stochastic block model where G is an undirected graph over n vertices, divided into two communities $\mathcal{V}_+$ and $\mathcal{V}_-$ of equal sizes $n/2$. For two n-dependent probabilities $p_n > q_n$, each pair of vertices (i, j) in G (including self-loops, for simplicity of discussion) is independently connected with probability

$$\mathbb{P}[i \text{ is connected to } j] = \begin{cases} p_n & \text{if } i, j \in \mathcal{V}_+ \text{ or } i, j \in \mathcal{V}_-, \\ q_n & \text{if } i \in \mathcal{V}_+ \text{ and } j \in \mathcal{V}_- \text{ or if } i \in \mathcal{V}_- \text{ and } j \in \mathcal{V}_+. \end{cases}$$

Let $\mathbf{A} \in \{0, 1\}^{n \times n}$ be the adjacency matrix of G, and let $\bar{p}_n = (p_n + q_n)/2$ be the mean connectivity. Then the centered and normalized adjacency matrix takes the form

$$\frac{\mathbf{A} - \bar{p}_n}{\sqrt{n \bar{p}_n (1 - \bar{p}_n)}} = \frac{\sqrt{\lambda_n}}{n} \mathbf{f}\mathbf{f}^\top + \mathbf{W}, \tag{2.17}$$

where $\lambda_n = n(p_n - q_n)^2/[4\bar{p}_n(1 - \bar{p}_n)]$ is a parameter representing the signal-to-noise ratio of the model, $\mathbf{f} \in \{+1, -1\}^n$ is the binary indicator vector representing the membership of the vertices, and $\mathbf{W}$ is a symmetric noise matrix with independent entries. It may checked for each $(i, j) \in [n] \times [n]$ that

$$\mathbb{E}[W[i, j]] = 0, \qquad \mathbb{E}[W[i, j]^2] \in \left\{ \frac{p_n(1 - p_n)}{n \bar{p}_n (1 - \bar{p}_n)}, \frac{q_n(1 - q_n)}{n \bar{p}_n (1 - \bar{p}_n)} \right\},$$
$$|W[i, j]| \le \frac{1}{\sqrt{n \bar{p}_n (1 - \bar{p}_n)}}.$$

In the asymptotic regime where $n\bar{p}_n(1 - \bar{p}_n) \to \infty$ and $\lambda_n \to \lambda$ a positive constant, we have that $S[i, j] := n \cdot \mathbb{E}[W[i, j]^2] \to 1$ uniformly over $(i, j) \in [n] \times [n]$, so that $\mathbf{W}$ is a generalized Wigner matrix in the sense of Definition 2.3 (with variance profile $\mathbf{S}$ approximately constant in every entry). Furthermore, under a slightly stronger assumption

$$n \bar{p}_n (1 - \bar{p}_n) \ge c \log n \tag{2.18}$$

for any constant $c > 0$, [8], Theorem 2.7 and eq. (2.4), implies that $\|\mathbf{W}\|_{\mathrm{op}} < C$ almost surely for all large n. This encompasses the stochastic block model in regimes with sparsity $\bar{p}_n \gtrsim (\log n)/n$.[4]

It was shown in [24] that the mutual information between G and $\mathbf{f}$ has an asymptotic limit depending only on the limit signal-to-noise ratio λ, which is nontrivial when $\lambda > 1$. This was proven by interpolating between the model (2.17) and a "$\mathbb{Z}_2$-synchronization" model where $\mathbf{W} \sim \mathrm{GOE}(n)$, and applying an AMP analysis in the latter model. Our result of Theorem 2.4 implies that, under the additional condition (2.18), this AMP analysis may instead be directly applied to the model (2.17), bypassing interpolation to the GOE.

EXAMPLE 2.25. Let $\mathbf{Y} \in \mathbb{R}^{m \times n}$ be a signal-plus-noise data matrix modeled as

$$\mathbf{Y} = \mathbf{X} + \mathbf{E},$$

where $\mathbf{X} = \mathbb{E}[\mathbf{Y}] = \sum_{j=1}^{k} \mathbf{f}_j \mathbf{g}_j^\top \in \mathbb{R}^{m \times n}$ is a low-rank signal matrix, and $\mathbf{E} = \mathbf{Y} - \mathbf{X}$ is a mean-zero matrix of residual noise. We assume that $\mathbf{E}$ has independent entries, although in many applications involving count observations or missing data, these entries may have a heteroskedastic variance profile $\mathbf{V}$ where

$$V[\alpha, i] := \mathrm{Var}\big[E[\alpha, i]\big].$$

Such models where the variance $V[\alpha, i]$ is a quadratic function $a + bX[\alpha, i] + cX[\alpha, i]^2$ of the mean were discussed recently in [46], including Poisson and negative-binomial models for $\mathbf{Y}$ in the context of single-cell RNA sequencing applications [42, 65, 72]. Such models encompass also simple models of missing data, where $\mathbf{Y}$ is a partial observation of an underlying low-rank signal matrix $\widetilde{\mathbf{X}}$ so that

$$Y[\alpha, i] = \begin{cases} \widetilde{X}[\alpha, i] & \text{with probability } p, \\ 0 & \text{with probability } 1 - p, \end{cases}$$

independently for each entry. Then $X[\alpha, i] = p \cdot \widetilde{X}[\alpha, i]$ and $V[\alpha, i] = p(1-p) \cdot \widetilde{X}[\alpha, i]^2$ are the corresponding means and variances.

When the entries $V[\alpha, i]$ are heteroskedastic, the singular value spectrum of $\mathbf{E}$ does not generally conform to the Marcenko–Pastur law. However, row and column normalization is typically applied in practice prior to data analysis, with [46] suggesting the following normalization scheme: Determine via Sinkhorn iteration two diagonal matrices $\mathbf{D}_1 \in \mathbb{R}^{m \times m}$ and $\mathbf{D}_2 \in \mathbb{R}^{n \times n}$ for which $\mathbf{S} = \mathbf{D}_1 \mathbf{V} \mathbf{D}_2$ has all rows summing to n and all columns summing to m, and use these to standardize $\mathbf{Y}$ into the biwhitened matrix

$$\widetilde{\mathbf{Y}} = \frac{1}{\sqrt{n}} \cdot \mathbf{D}_1^{1/2} \mathbf{Y} \mathbf{D}_2^{1/2} = \frac{1}{\sqrt{n}} \cdot \mathbf{D}_1^{1/2} \mathbf{X} \mathbf{D}_2^{1/2} + \mathbf{W}, \qquad \mathbf{W} = \frac{1}{\sqrt{n}} \cdot \mathbf{D}_1^{1/2} \mathbf{E} \mathbf{D}_2^{1/2}.$$

[46] proved that such biwhitened count matrices have singular value spectra asymptotically described by the Marcenko–Pastur law, and showed a remarkable empirical agreement with the Marcenko–Pastur law for matrices arising in several domains of application, from single-cell biology to topic modeling of text.

In this standardized model $\widetilde{\mathbf{Y}}$, the error matrix $\mathbf{W}$ now has variance profile $\mathbf{S} = \mathbf{D}_1 \mathbf{V} \mathbf{D}_2$ which satisfies by construction $\frac{1}{n} \sum_{i=1}^{n} S[\alpha, i] = \frac{1}{m} \sum_{\alpha=1}^{m} S[\alpha, i] = 1$, and Theorem 2.21 describes conditions under which state evolution holds for Gaussian AMP algorithms applied

[4] Our universality result for AMP with polynomial nonlinearities does not require the operator norm bound $\|\mathbf{W}\|_{\mathrm{op}} < C$ and hence holds for any sparsity $\bar{p}_n \gg 1/n$, cf. Remark 3.12. We believe that the operator norm requirement in Condition 2 of Theorem 2.4 may be an artifact of our polynomial approximation proof.

In contrast, we do not expect AMP universality to hold for random graph models with sparsity $\bar{p}_n \asymp 1/n$, where the belief propagation recursions on such graphs may not admit asymptotic Gaussian approximations.

to this matrix $\mathbf{W}$. We note that to analyze AMP applied instead to $\widetilde{\mathbf{Y}}$, the condition (2.16) represents a potentially strong restriction on the relation between the variance profile matrix $\mathbf{S}$ and the low-rank mean signal. A modified analysis of AMP may be needed in settings where this restriction does not hold, and we leave this as a direction to explore in future work.

EXAMPLE 2.26. Much of the early development of AMP algorithms was motivated by compressed sensing applications of reconstructing sparse signals from linear measurements. Consider a model of m measurements

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\varepsilon} \in \mathbb{R}^m,$$

where $\mathbf{x} \in \mathbb{R}^n$ is the underlying signal, $\mathbf{W} \in \mathbb{R}^{m\times n}$ is a random sensing matrix, and $\boldsymbol{\varepsilon}$ is measurement noise. For i.i.d. Gaussian sensing matrices $\mathbf{W}$, pioneering work of [29, 30] proposed an AMP algorithm for reconstructing $\mathbf{x}$, where the nonlinearities are soft-thresholding functions tailored to the sparsity of $\mathbf{x}$. Analysis of the dynamics of this algorithm leads to a derivation of a sparsity-undersampling phase transition curve that matches a phase transition for ℓ_1-based reconstruction in this model [6, 7, 29, 31].

Extensive numerical experiments performed in [27, 55] suggested that this phase transition curve is universal across broad classes of non-Gaussian sensing matrices. Theorem 2.8 provides an extension of the AMP universality shown in [6] for this application, broadening the universality class to matrices composed of subsampled Fourier or Hadamard transforms and diagonal operators. Importantly, matrix-vector multiplication operations for such matrices may be computed in $O(n \log n)$ time without explicitly storing the matrices in memory, allowing applications of AMP at much larger scales than would be possible with i.i.d. sensing designs.

As an example, consider

$$\mathbf{W} = (\boldsymbol{\Pi}_U \mathbf{H} \boldsymbol{\Pi}_E)\mathbf{D}(\boldsymbol{\Pi}_V \mathbf{K} \boldsymbol{\Pi}_F)^\top \in \mathbb{R}^{m\times n}, \tag{2.19}$$

where $\mathbf{D} \in \mathbb{R}^{m\times n}$ is diagonal with its diagonal entries sampled i.i.d. from a Marcenko–Pastur law; $\mathbf{H}, \mathbf{K} \in \mathbb{R}^{n\times n}$ are orthogonal matrices representing deterministic Hadamard or discrete Fourier transforms; and $\boldsymbol{\Pi}_U$, $\boldsymbol{\Pi}_E$, $\boldsymbol{\Pi}_V$, $\boldsymbol{\Pi}_F$ are independent random signed permutations. We verify in Proposition D.1(b1) that this class of matrices satisfies Definition 2.20. If the signal vector $\mathbf{x}$, residual error $\boldsymbol{\varepsilon}$, and initialization $\mathbf{x}_1$ are each comprised of i.i.d. entries, then the random permutations in $\boldsymbol{\Pi}_U$, $\boldsymbol{\Pi}_V$ may be further absorbed into $\mathbf{x}_1$, $\mathbf{x}$, $\boldsymbol{\varepsilon}$. Thus the AMP iterates are equal in law to those of AMP applied with a simpler sensing matrix

$$\widetilde{\mathbf{W}} = \boldsymbol{\Xi}_U \mathbf{H} \widetilde{\mathbf{D}} \widetilde{\mathbf{K}}^\top \boldsymbol{\Xi}_V,$$

where $\boldsymbol{\Xi}_U$, $\boldsymbol{\Xi}_V$ are diagonal matrices of i.i.d. $\{+1, -1\}$ signs, $\widetilde{\mathbf{K}}^\top \in \mathbb{R}^{m\times n}$ is a random subsampling of m rows of $\mathbf{K}^\top$, and $\widetilde{\mathbf{D}} \in \mathbb{R}^{m\times m}$ is a diagonal matrix whose diagonal entries are given by those of $\mathbf{D}$ also multiplied by i.i.d. $\{+1, -1\}$ signs.

Theorem 2.22 implies that AMP applied to the above matrix $\mathbf{W}$ admits the same state evolution as when applied to an i.i.d. Gaussian sensing matrix $\mathbf{G}$. This universality extends beyond the Gaussian setting, to sensing matrices (2.19) where the diagonal entries of $\mathbf{D}$ are sampled from an arbitrary compactly supported singular value distribution. Theorem 2.22 then shows that the state evolution characterizations for the more general AMP algorithms of [37]—derived originally for bi-orthogonally invariant ensembles—are valid in such settings. For this compressed sensing application, we note that the resulting AMP algorithms are similar to the convolutional AMP algorithms developed and studied recently in [70, 71].

3. Proofs for symmetric matrices.

3.1. *Universality for generalized Wigner matrices.* In this section, we prove Lemma 2.13 on the universality of the tensor network value for generalized Wigner matrices.

Fix a tensor network $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$. Let $\mathcal{P}$ be the set of all partitions of $\mathcal{V}$. For each index tuple $\mathbf{i} \in [n]^{\mathcal{V}}$, define its induced partition $\pi(\mathbf{i}) \in \mathcal{P}$ such that vertices $u, v \in \mathcal{V}$ belong to the same block of $\pi(\mathbf{i})$ if and only if $i_u = i_v$. Then we can decompose the value of T as

$$\operatorname{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \frac{1}{n} \sum_{\pi\in\mathcal{P}} \sum_{\mathbf{i}\in[n]^{\mathcal{V}}:\pi(\mathbf{i})=\pi} q_{\mathbf{i}|T} \cdot W_{\mathbf{i}|T}. \tag{3.1}$$

DEFINITION 3.1. Let $(\mathcal{V}, \mathcal{E})$ be an undirected graph. For any partition π of $\mathcal{V}$, the *image of* $(\mathcal{V}, \mathcal{E})$ *under* π is the undirected multi-graph $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ that is the image of $(\mathcal{V}, \mathcal{E})$ under the graph homomorphism sending each vertex $u \in \mathcal{V}$ to the block of π containing u.

That is, the vertices $\mathcal{K}_\pi \equiv \pi$ of G_π are the blocks of π, and G_π has the same number of edges $|\mathcal{F}_\pi|$ (counting multiplicity and self-loops) as $|\mathcal{E}|$. For each edge $(u, v) \in \mathcal{E}$, there is a corresponding edge $(U, V) \in \mathcal{F}_\pi$ where $U, V \in \pi$ are the blocks for which $u \in U$ and $v \in V$.

For each $\pi \in \mathcal{P}$, let $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ be the image of $(\mathcal{V}, \mathcal{E})$ under π. For each block $U \in \mathcal{K}_\pi$, define the polynomial $Q_U = \prod_{u\in U} q_u$, and for each unique (undirected) edge (U, V) of G_π, let $e(U, V)$ be the number of times it appears in $\mathcal{F}_\pi$. Then, identifying the sum over $\{\mathbf{i} : \pi(\mathbf{i}) = \pi\}$ as a sum over one distinct index in $[n]$ for each block $U \in \mathcal{K}_\pi$, we have

$$\sum_{\mathbf{i}\in[n]^{\mathcal{V}}:\pi(\mathbf{i})=\pi} q_{\mathbf{i}|T} \cdot W_{\mathbf{i}|T} = \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot W^e_{\mathbf{i}|G_\pi},$$

where $\sum^*$ denotes the restriction of the summation to index tuples $\mathbf{i} = (i_U : U \in \mathcal{K}_\pi) \in [n]^{\mathcal{K}_\pi}$ having all indices distinct, and

$$Q_{\mathbf{i}|G_\pi} = \prod_{U\in\mathcal{K}_\pi} Q_U(x_1[i_U], \dots, x_k[i_U]), \qquad W^e_{\mathbf{i}|G_\pi} = \prod_{\text{unique edges } (U,V) \text{ of } G_\pi} W[i_U, i_V]^{e(U,V)}.$$

Applying this to (3.1), we obtain

$$\operatorname{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) = \frac{1}{n} \sum_{\pi\in\mathcal{P}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot W^e_{\mathbf{i}|G_\pi}. \tag{3.2}$$

We will compute the expectation of (3.2), and see that the only nonvanishing contributions in the limit $n \to \infty$ arise from partitions π where G_π is itself a tree and $e(U, V) = 2$ for each unique edge (U, V) of G_π. These nonvanishing terms may be related to the values of a reduced tensor network associated to G_π, evaluated on the matrix $\mathbf{S}/n$ in place of $\mathbf{W}$.

In anticipation of this computation, we first show the following lemma which establishes universality of the value of any tensor network evaluated on $\mathbf{S}/n$.

LEMMA 3.2. *Under the assumptions of Lemma 2.13, for any diagonal tensor network* $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$ *in* k *variables,*

$$\lim_{n\to\infty} \operatorname{val}_T(\mathbf{S}/n; \mathbf{x}_1, \dots, \mathbf{x}_k) = \prod_{v\in\mathcal{V}} \mathbb{E}[q_v(X_1, \dots, X_k)].$$

PROOF. Observe first that for any diagonal tensor network $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$, we have

$$\frac{1}{n} \sum_{\mathbf{i}\in[n]^{\mathcal{V}}} \prod_{v\in\mathcal{V}} |q_v(x_{1:k})[i_v]| \cdot \prod_{(u,v)\in\mathcal{E}} \frac{1}{n} \le C \tag{3.3}$$

for a constant $C := C(T) > 0$, almost surely for all large n. Indeed, since T is a tree, we have $\frac{1}{n}\prod_{(u,v)\in\mathcal{E}} \frac{1}{n} = n^{-|\mathcal{V}|}$ in the above. Each function $|q_v|$ is continuous and satisfies the polynomial growth condition (1.1), so (3.3) follows from the assumption (2.10).

Now note that since T is a tree, we can order its vertices as $1, 2, \dots, |\mathcal{V}|$ such that removing one vertex at a time in this order, the remaining graph is always still a tree. Denote the remaining tensor network after removing vertices $1, \dots, h-1$ by $T_h = (\mathcal{V}_h, \mathcal{E}_h, \{q_v\}_{v \geq h})$. The vertex h has only one neighbor in T_h, which we denote by $u_h \in \{h+1, \dots, |\mathcal{V}|\}$. Then

$$\begin{aligned}
&\operatorname{val}_T(\mathbf{S}/n; \mathbf{x}_{1:k}) \\
&= \frac{1}{n} \sum_{\mathbf{i} \in [n]^{\mathcal{V}}} \prod_{v \in \mathcal{V}} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}} \frac{S[i_u, i_v]}{n} \\
&= \frac{1}{n} \sum_{i_2, \dots, i_{|\mathcal{V}|}=1}^{n} \Bigg(\prod_{v \in \mathcal{V}_2} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}_2} \frac{S[i_u, i_v]}{n} \Bigg) \cdot \Bigg(\sum_{i_1=1}^{n} q_1(x_{1:k}[i_1]) \frac{S[i_1, i_{u_1}]}{n} \Bigg) \\
&= \frac{1}{n} \sum_{i_2, \dots, i_{|\mathcal{V}|}=1}^{n} \Bigg(\prod_{v \in \mathcal{V}_2} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}_2} \frac{S[i_u, i_v]}{n} \Bigg) \cdot \big(\langle q_1(\mathbf{x}_{1:k}) \rangle \cdot \langle \mathbf{s}_{i_{u_1}} \rangle + \delta(i_{u_1}) \big) \\
&= \frac{1}{n} \sum_{i_2, \dots, i_{|\mathcal{V}|}=1}^{n} \Bigg(\prod_{v \in \mathcal{V}_2} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}_2} \frac{S[i_u, i_v]}{n} \Bigg) \cdot \big(\mathbb{E}[q_1(X_{1:k})] + \delta'(i_{u_1}) \big).
\end{aligned}$$

Here, $\delta(i), \delta'(i)$ denote errors that satisfy $\lim_{n \to \infty} \max_{i \in [n]} |\delta(i)|, |\delta'(i)| = 0$, as follows from (2.11), the conditions of Definition 2.3(c), and (2.10). Note that

$$\left| \frac{1}{n} \sum_{i_2, \dots, i_{|\mathcal{V}|}=1}^{n} \Bigg(\prod_{v \in \mathcal{V}_2} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}_2} \frac{S[i_u, i_v]}{n} \Bigg) \cdot \delta'(i_{u_1}) \right| \to 0$$

as $n \to \infty$ by the condition $|S[i_u, i_v]| \leq C$ of Definition 2.3(c), the bound (3.3) applied to the network T_2 with vertex 1 removed, and the convergence $\max_{i \in [n]} |\delta'(i)| \to 0$. Thus

$$\lim_{n \to \infty} \operatorname{val}_T(\mathbf{S}/n; \mathbf{x}_{1:k}) = \lim_{n \to \infty} \frac{1}{n} \sum_{i_2, \dots, i_{|\mathcal{V}|}=1}^{n} \Bigg(\prod_{v \in \mathcal{V}_2} q_v(x_{1:k}[i_v]) \prod_{(u,v) \in \mathcal{E}_2} \frac{S[i_u, i_v]}{n} \Bigg) \cdot \mathbb{E}[q_1(X_{1:k})].$$

Repeating the above procedure by removing vertices $1, 2, \dots, |\mathcal{V}| - 1$ sequentially, we are left with the single vertex $|\mathcal{V}|$ and no edges, and

$$\lim_{n \to \infty} \operatorname{val}_T(\mathbf{S}/n; \mathbf{x}_{1:k}) = \lim_{n \to \infty} \frac{1}{n} \sum_{i_{|\mathcal{V}|}=1}^{n} q_{|\mathcal{V}|}(x_{1:k}[i_{|\mathcal{V}|}]) \cdot \prod_{v=1}^{|\mathcal{V}|-1} \mathbb{E}[q_v(X_{1:k})] = \prod_{v=1}^{|\mathcal{V}|} \mathbb{E}[q_v(X_{1:k})].$$

□

The next lemma relates summations over distinct indices to ones without the distinctness requirement.

LEMMA 3.3. *Let $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ and $(X_1, \dots, X_k)$ be such that*

$$(\mathbf{x}_1, \dots, \mathbf{x}_k) \xrightarrow{W} (X_1, \dots, X_k).$$

For a finite index set $\mathcal{S}$, let $(q_s : s \in \mathcal{S})$ be $|\mathcal{S}|$ continuous functions satisfying the polynomial growth condition (1.1) for some order $p \geq 1$. Then

$$\begin{aligned}
\lim_{n \to \infty} \frac{1}{n^{|\mathcal{S}|}} \sideset{}{^*}\sum_{\mathbf{i} \in [n]^{\mathcal{S}}} \prod_{s \in \mathcal{S}} q_s(x_1[i_s], \dots, x_k[i_s]) &= \lim_{n \to \infty} \frac{1}{n^{|\mathcal{S}|}} \sum_{\mathbf{i} \in [n]^{\mathcal{S}}} \prod_{s \in \mathcal{S}} q_s(x_1[i_s], \dots, x_k[i_s]) \\
&= \prod_{s \in \mathcal{S}} \mathbb{E}[q_s(X_1, \dots, X_k)].
\end{aligned}$$

PROOF. Let $\mathcal{P}$ be the set of partitions of $\mathcal{S}$, and let $\pi(\mathbf{i}) \in \mathcal{P}$ be the partition induced by $\mathbf{i} \in [n]^{\mathcal{S}}$. Let $0_{\mathcal{P}}$ the partition having $|\mathcal{S}|$ singleton blocks, corresponding to $\mathbf{i}$ having all indices distinct. Then for the first equality, it suffices to show that

$$\Delta := \frac{1}{n^{|\mathcal{S}|}} \sum_{\pi \in \mathcal{P}: \pi \neq 0_{\mathcal{P}}} \sum_{\mathbf{i} \in [n]^{\mathcal{S}}: \pi(\mathbf{i}) = \pi} \prod_{s \in \mathcal{S}} |q_s(x_{1:k}[i_s])| \tag{3.4}$$

vanishes as $n \to \infty$.

For any $\pi \in \mathcal{P}$ and block $R \in \pi$, define $Q_R = \prod_{u \in R} q_u$, and let $|\pi|$ be the number of blocks of π. Then, identifying the sum over $\{\mathbf{i} : \pi(\mathbf{i}) = \pi\}$ with the sum over one distinct index in $[n]$ for each block of π,

$$\frac{1}{n^{|\pi|}} \sum_{\mathbf{i} \in [n]^{\mathcal{S}}: \pi(\mathbf{i}) = \pi} \prod_{s \in \mathcal{S}} |q_s(x_{1:k}[i_s])| = \frac{1}{n^{|\pi|}} \sum_{\mathbf{i} \in [n]^{\pi}}^{*} \prod_{R \in \pi} |Q_R(x_{1:k}[i_R])|.$$

As an upper bound, adding back the excluded index tuples $\mathbf{i} \in [n]^{\pi}$ where some indices coincide,

$$\frac{1}{n^{|\pi|}} \sum_{\mathbf{i} \in [n]^{\mathcal{S}}: \pi(\mathbf{i}) = \pi} \prod_{s \in \mathcal{S}} |q_s(x_{1:k}[i_s])| \leq \prod_{R \in \pi} \left(\frac{1}{n} \sum_{i=1}^{n} |Q_R(x_{1:k}[i])| \right).$$

Since $\mathbf{x}_{1:k} \xrightarrow{W} X_{1:k}$ and $|Q_R|$ is a continuous function satisfying the polynomial growth condition (1.1), this upper bound is at most a constant $C(\pi)$ for all large n. For any $\pi \neq 0_{\mathcal{P}}$, we have $|\pi| \leq |\mathcal{S}| - 1$. As the number of partitions $\pi \in \mathcal{P}$ is independent of n, applying these observations to (3.4) shows $\Delta \leq C/n$ for a constant $C > 0$ and all large n, and hence $\Delta \to 0$ as desired.

The second equality of the lemma follows from the given condition $\mathbf{x}_{1:k} \xrightarrow{W} X_{1:k}$, hence

$$\frac{1}{n^{|\mathcal{S}|}} \sum_{\mathbf{i} \in [n]^{\mathcal{S}}} \prod_{s \in \mathcal{S}} q_s(x_{1:k}[i_s]) = \prod_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} q_s(x_{1:k}[i]) \to \prod_{s \in \mathcal{S}} \mathbb{E}[q_s(X_{1:k})]. \qquad \square$$

We now show that the limit tensor network value is universal in expectation over $\mathbf{W}$.

LEMMA 3.4. *Let $\mathbb{E}$ denote the expectation over $\mathbf{W}$, conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Then Lemma 2.13 holds for $\mathbb{E}[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \ldots, \mathbf{x}_k)]$ in place of $\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$.*

PROOF. Recall the decomposition of $\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \ldots, \mathbf{x}_k)$ in (3.2), where $\mathcal{P}$ is the set of partitions of the vertices $\mathcal{V}$ of T. Taking expectation on both sides yields

$$\mathbb{E}\big[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k})\big] = \frac{1}{n} \sum_{\pi \in \mathcal{P}} \sum_{\mathbf{i} \in [n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot \mathbb{E}\big[W^e_{\mathbf{i}|G_\pi}\big]. \tag{3.5}$$

First note that, since the indices of $\mathbf{i}$ are distinct and all entries of $\mathbf{W}$ have mean 0, $\mathbb{E}[W^e_{\mathbf{i}|G_\pi}]$ is nonzero only if each unique edge of $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ appears at least twice. Let $|\mathcal{K}_\pi|$ and $|\mathcal{F}_\pi|_*$ be the number of vertices and number of *unique* (undirected) edges of G_π. The graph G_π must be connected since the original tree T was connected, so $|\mathcal{K}_\pi| \leq |\mathcal{F}_\pi|_* + 1$. Then any G_π where each unique edge appears at least twice has $|\mathcal{K}_\pi| \leq |\mathcal{F}_\pi|_* + 1 \leq |\mathcal{E}|/2 + 1$, where $|\mathcal{E}|$ is the number of edges of the original tree T.

Furthermore, we claim that the contribution from partitions π where $|\mathcal{K}_\pi| \le |\mathcal{E}|/2$ is negligible. To see this, we apply Definition 2.3(b) to get

$$\begin{aligned}|\mathbb{E}[W^e_{\mathbf{i}|G_\pi}]| &= \prod_{(U,V):e(U,V)=2} \mathbb{E}[W[i_U,i_V]^2] \prod_{(U,V):e(U,V)>2} |\mathbb{E}[W[i_U,i_V]^{e(U,V)}]| \\ &\le \prod_{(U,V):e(U,V)=2} \frac{C}{n} \prod_{(U,V):e(U,V)>2} \frac{o(1)}{n}.\end{aligned}$$

If there is an edge (U,V) of G_π with $e(U,V) > 2$, then this shows $|\mathbb{E}[W^e_{\mathbf{i}|G_\pi}]| \le o(1)/n^{|\mathcal{F}_\pi|_*} \le o(1)/n^{|\mathcal{K}_\pi|-1}$. If, conversely, every edge in G_π appears exactly twice, then by assumption $|\mathcal{F}_\pi|_* = |\mathcal{E}|/2 \ge |\mathcal{K}_\pi|$, so this shows $|\mathbb{E}[W^e_{\mathbf{i}|G_\pi}]| \le (C/n)^{|\mathcal{F}_\pi|_*} \le o(1)/n^{|\mathcal{K}_\pi|-1}$ also. Therefore,

$$\left|\frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot \mathbb{E}[W^e_{\mathbf{i}|G_\pi}]\right| \le \frac{o(1)}{n^{|\mathcal{K}_\pi|}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} |Q_{\mathbf{i}|G_\pi}|.$$

As an upper bound, adding back the excluded tuples $\mathbf{i} \in [n]^{\mathcal{K}_\pi}$ where not all indices are distinct, we have

$$\frac{1}{n^{|\mathcal{K}_\pi|}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} |Q_{\mathbf{i}|G_\pi}| \le \prod_{U\in\mathcal{K}_\pi} \frac{1}{n}\sum_{i=1}^{n} |Q_U(x_{1:k}[i])|. \tag{3.6}$$

By (2.10), this upper bound is at most a constant $C(\pi)$ for all large n, so

$$\left|\frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot \mathbb{E}[W^e_{\mathbf{i}|G_\pi}]\right| \to 0$$

as claimed.

Thus the only nonvanishing contributions to (3.5) come from partitions π where $|\mathcal{K}_\pi| = |\mathcal{F}_\pi|_* + 1 = |\mathcal{E}|/2 + 1$. Then each unique edge of G_π appears exactly twice, and these edges form a tree. In this case, we have

$$\frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot \mathbb{E}[W^e_{\mathbf{i}|G_\pi}] = \frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} Q_{\mathbf{i}|G_\pi} \cdot \prod_{\text{unique edges } (U,V) \text{ of } G_\pi} \frac{S[i_U,i_V]}{n}. \tag{3.7}$$

Let $\mathcal{I}_*$ be the set of tuples $\mathbf{i} \in [n]^{\mathcal{K}_\pi}$ where all indices are distinct. Then, applying $|S[i_U,i_V]| \le C$ from Definition 2.3(c), for a constant $C' = C(\pi) > 0$,

$$\left|\frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}\setminus\mathcal{I}_*} Q_{\mathbf{i}|G_\pi} \prod_{\text{unique edges } (U,V) \text{ of } G_\pi} \frac{S[i_U,i_V]}{n}\right| \le \frac{C'}{n^{1+|\mathcal{F}_\pi|_*}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}\setminus\mathcal{I}_*} |Q_{\mathbf{i}|G_\pi}|.$$

Since $|\mathcal{K}_\pi| = |\mathcal{F}_\pi|_* + 1$, the first equality of Lemma 3.3 shows that this vanishes as $n \to \infty$. Then the right side of (3.7) has the same limit as

$$\frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}} Q_{\mathbf{i}|G_\pi} \cdot \prod_{\text{unique edges } (U,V) \text{ of } G_\pi} \frac{S[i_U,i_V]}{n}.$$

This is the limit value of the tensor network $T_\pi = (\mathcal{K}_\pi, \text{unique edges of } \mathcal{F}_\pi, \{Q_U\}_{U\in\mathcal{K}_\pi})$ applied to $\mathbf{S}/n$, which by Lemma 3.2 equals $\prod_{U\in\mathcal{K}_\pi} \mathbb{E}[Q_U(X_{1:k})]$. Applying this back to (3.7) and (3.5),

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})] &= \sum_{\pi\in\mathcal{P}:|\mathcal{K}_\pi|=|\mathcal{F}_\pi|_*+1=|\mathcal{E}|/2+1} \prod_{U\in\mathcal{K}_\pi} \mathbb{E}[Q_U(X_{1:k})] \\ &=: \text{lim-val}_T(X_1,\dots,X_k).\end{aligned} \tag{3.8}$$

This limit depends only on T and the joint law of $X_1,\dots,X_k$, concluding the proof. □

We make a brief interlude to show here the asymptotic freeness result of Proposition 2.16(a).

PROOF OF PROPOSITION 2.16(A). From the preceding proof, only partitions π where $|\mathcal{K}_\pi| = |\mathcal{E}|/2 + 1$ contribute to $\text{lim-val}_T(X_1, \dots, X_k)$. For every such partition, since T has $|\mathcal{E}| + 1$ vertices, this implies that some vertex of $\mathcal{K}_\pi$, that is, some block U of π, contains only a single vertex v of T. For this block U, we have

$$\mathbb{E}\big[Q_U(X_{1:k})\big] = \mathbb{E}\big[q_v(X_{1:k})\big] = 0$$

by the condition (2.13). Therefore, every summand in (3.8) vanishes, implying as desired $\text{lim-val}_T(X_1, \dots, X_k) = 0$. □

To complete the proof of Lemma 2.13, it remains to establish the concentration of the value of any tensor network around its mean as $n \to \infty$.

LEMMA 3.5. *Let $\mathbb{E}$ denote the expectation over $\mathbf{W}$, conditional on $\mathbf{x}_1, \dots, \mathbf{x}_k$. Under the setting of Lemma 2.13, almost surely as $n \to \infty$,*

$$\text{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) - \mathbb{E}\big[\text{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k)\big] \to 0.$$

PROOF. We write $\text{val}(\mathbf{W}) = \text{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k)$. We will bound the fourth moment of $\text{val}(\mathbf{W}) - \mathbb{E}[\text{val}(\mathbf{W})]$ and apply the Borel–Cantelli lemma. (Note that $\text{val}(\mathbf{W}) - \mathbb{E}[\text{val}(\mathbf{W})]$ typically fluctuates on the order of $1/\sqrt{n}$, so that bounding the variance would not suffice to show almost-sure convergence.)

First, we expand

$$\begin{aligned}
&\mathbb{E}\big[(\text{val}(\mathbf{W}) - \mathbb{E}\big[\text{val}(\mathbf{W})\big])^4\big] \\
&\quad = \mathbb{E}\big[\text{val}(\mathbf{W})^4\big] - 4\mathbb{E}\big[\text{val}(\mathbf{W})^3\big]\mathbb{E}\big[\text{val}(\mathbf{W})\big] \\
&\qquad + 6\mathbb{E}\big[\text{val}(\mathbf{W})^2\big]\mathbb{E}\big[\text{val}(\mathbf{W})\big]^2 - 3\mathbb{E}\big[\text{val}(\mathbf{W})\big]^4.
\end{aligned} \tag{3.9}$$

We introduce four independent copies of the matrix $\mathbf{W}$ as $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}^{(3)}$, $\mathbf{W}^{(4)}$, define four index tuples $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \mathbf{i}_4 \in [n]^{\mathcal{V}}$, and write as shorthand

$$q_{\mathbf{i}_{1:4}} = q_{\mathbf{i}_1|T} \cdot q_{\mathbf{i}_2|T} \cdot q_{\mathbf{i}_3|T} \cdot q_{\mathbf{i}_4|T}, \qquad W^{(a_1,a_2,a_3,a_4)}_{\mathbf{i}_{1:4}} = W^{(a_1)}_{\mathbf{i}_1|T} \cdot W^{(a_2)}_{\mathbf{i}_2|T} \cdot W^{(a_3)}_{\mathbf{i}_3|T} \cdot W^{(a_4)}_{\mathbf{i}_4|T},$$

where each $W^{(a)}_{\mathbf{i}|T}$ is defined by the copy $\mathbf{W}^{(a)}$. Then

$$\begin{aligned}
\mathbb{E}\big[\text{val}(\mathbf{W})^4\big] &= \frac{1}{n^4} \sum_{\mathbf{i}_1,\dots,\mathbf{i}_4 \in [n]^{\mathcal{V}}} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W^{(1,1,1,1)}_{\mathbf{i}_{1:4}}\big], \\
\mathbb{E}\big[\text{val}(\mathbf{W})^3\big]\mathbb{E}\big[\text{val}(\mathbf{W})\big] &= \frac{1}{n^4} \sum_{\mathbf{i}_1,\dots,\mathbf{i}_4 \in [n]^{\mathcal{V}}} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W^{(1,1,1,2)}_{\mathbf{i}_{1:4}}\big], \\
\mathbb{E}\big[\text{val}(\mathbf{W})^2\big]\mathbb{E}\big[\text{val}(\mathbf{W})\big]^2 &= \frac{1}{n^4} \sum_{\mathbf{i}_1,\dots,\mathbf{i}_4 \in [n]^{\mathcal{V}}} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W^{(1,1,2,3)}_{\mathbf{i}_{1:4}}\big], \\
\mathbb{E}\big[\text{val}(\mathbf{W})\big]^4 &= \frac{1}{n^4} \sum_{\mathbf{i}_1,\dots,\mathbf{i}_4 \in [n]^{\mathcal{V}}} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W^{(1,2,3,4)}_{\mathbf{i}_{1:4}}\big].
\end{aligned} \tag{3.10}$$

Corresponding to each index tuple $\mathbf{i}_{1:4}$, consider a multi-graph $G(\mathbf{i}_{1:4})$ whose vertices are the unique index values in $\mathbf{i}_{1:4}$, with one edge $(i_{a,u}, i_{a,v})$ for every combination of

$a = 1, 2, 3, 4$ and edge $(u, v) \in \mathcal{E}$, counting multiplicity. (One may visualize $G(\mathbf{i}_{1:4})$ as a multi-graph whose vertices are a subset of $[n]$, and having edges of 4 colors corresponding to $a = 1, 2, 3, 4$.) Then the edges of $G(\mathbf{i}_{1:4})$ corresponding to each single index $a = 1, 2, 3, 4$ must belong to a single connected component, so the number of connected components in $G(\mathbf{i}_{1:4})$ can be either 1, 2, 3, or 4. Let us partition the index tuples $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \mathbf{i}_4 \in [n]^{\mathcal{V}}$ into the three sets

$$\begin{aligned}\mathcal{I}_2 &= \{\mathbf{i}_{1:4} : G(\mathbf{i}_{1:4}) \text{ has 1 or 2 connected components}\},\\ \mathcal{I}_3 &= \{\mathbf{i}_{1:4} : G(\mathbf{i}_{1:4}) \text{ has 3 connected components}\},\\ \mathcal{I}_4 &= \{\mathbf{i}_{1:4} : G(\mathbf{i}_{1:4}) \text{ has 4 connected components}\},\end{aligned}$$

and define correspondingly for $j = 2, 3, 4$,

$$\begin{aligned}A_j = \frac{1}{n^4} \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_j} q_{\mathbf{i}_{1:4}} \big(&\mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,1)}\big] - 4 \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,2)}\big]\\ &+ 6 \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,2,3)}\big] - 3 \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big]\big).\end{aligned} \tag{3.11}$$

Then by (3.9) and (3.10), we have $\mathbb{E}[(\mathrm{val}(\mathbf{W}) - \mathbb{E}[\mathrm{val}(\mathbf{W})])^4] = A_2 + A_3 + A_4$. Below, we will show that $A_3 = A_4 = 0$ and $A_2 = O(1/n^2)$ as $n \to \infty$.

For A_4, observe that since $G(\mathbf{i}_{1:4})$ has 4 connected components, the tuples $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \mathbf{i}_4$ have no common indices. Then due to the independence between entries of $\mathbf{W}$, we have

$$\begin{aligned}\mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(a_1,a_2,a_3,a_4)}\big] &= \mathbb{E}\big[W_{\mathbf{i}_1}^{(a_1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_2}^{(a_2)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_3}^{(a_3)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_4}^{(a_4)}\big]\\ &= \mathbb{E}\big[W_{\mathbf{i}_1}^{(1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_2}^{(1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_3}^{(1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_4}^{(1)}\big] = \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,1)}\big]\end{aligned}$$

for any a_1, a_2, a_3, a_4. Applying this to A_4 defined in (3.11), we get $A_4 = 0$.

Next, for A_3, we write $\mathbf{i}_j \parallel \mathbf{i}_{j'}$ if $\mathbf{i}_j$ and $\mathbf{i}_{j'}$ share at least one index. Note that for any $\mathbf{i}_{1:4} \in \mathcal{I}_3$, there is a unique pair $\mathbf{i}_j, \mathbf{i}_{j'}$ such that $\mathbf{i}_j \parallel \mathbf{i}_{j'}$, so

$$\mathcal{I}_3 = \bigsqcup_{1 \le j < j' \le 4} \{\mathbf{i}_{1:4} \in \mathcal{I}_3 : \mathbf{i}_j \parallel \mathbf{i}_{j'}\}.$$

We will repeatedly apply this six-fold decomposition of $\mathcal{I}_3$ and the independence between the different copies of $\mathbf{W}$. If $\mathbf{i}_3 \parallel \mathbf{i}_4$, we have

$$\mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,1)}\big] = \mathbb{E}\big[W_{\mathbf{i}_1}^{(1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_2}^{(1)}\big] \cdot \mathbb{E}\big[W_{\mathbf{i}_3}^{(1)} W_{\mathbf{i}_4}^{(1)}\big] = \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,3)}\big]$$

which together with permutation symmetry between the labels 1, 2, 3, and 4 further implies that

$$\begin{aligned}\sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,1)}\big] &= 6 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_3 \parallel \mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,1)}\big]\\ &= 6 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_3 \parallel \mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,3)}\big].\end{aligned} \tag{3.12}$$

Similarly, considering the two cases $\mathbf{i}_3 \parallel \mathbf{i}_4$ and $\mathbf{i}_2 \parallel \mathbf{i}_3$ and their symmetric equivalents,

$$\begin{aligned}&\sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,1,2)}\big]\\ &= 3 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_3 \parallel \mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] + 3 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_2 \parallel \mathbf{i}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,2,3)}\big]\\ &= 3 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_3 \parallel \mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] + 3 \sum_{\mathbf{i}_{1:4} \in \mathcal{I}_3, \mathbf{i}_3 \parallel \mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,3)}\big].\end{aligned} \tag{3.13}$$

Considering the three cases $\mathbf{i}_3 \parallel \mathbf{i}_4$, $\mathbf{i}_2 \parallel \mathbf{i}_4$, $\mathbf{i}_1 \parallel \mathbf{i}_2$ and their symmetric equivalents,

$$\begin{aligned}
&\sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,2,3)}\big] \\
&\quad= \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_3\parallel\mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] + 4 \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_2\parallel\mathbf{i}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] \\
&\qquad+ \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_1\parallel\mathbf{i}_2} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,1,2,3)}\big] \\
&\quad= 5 \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_3\parallel\mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] + \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_3\parallel\mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,3)}\big].
\end{aligned} \tag{3.14}$$

Finally, by symmetry,

$$\sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big] = 6 \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_3,\mathbf{i}_3\parallel\mathbf{i}_4} q_{\mathbf{i}_{1:4}} \cdot \mathbb{E}\big[W_{\mathbf{i}_{1:4}}^{(1,2,3,4)}\big]. \tag{3.15}$$

Collecting (3.12), (3.13), (3.14), and (3.15) and applying them to A_3 defined in (3.11), we get $A_3 = 0$.

Finally, we bound A_2. Let $|\mathcal{V}_{G(\mathbf{i}_{1:4})}|$ and $|\mathcal{E}_{G(\mathbf{i}_{1:4})}|_*$ be the number of vertices and number of unique (undirected) edges of $G(\mathbf{i}_{1:4})$. Since $G(\mathbf{i}_{1:4})$ has at most 2 connected components, we have $|\mathcal{V}_{G(\mathbf{i}_{1:4})}| \le |\mathcal{E}_{G(\mathbf{i}_{1:4})}|_* + 2$. By Definition 2.3(b), for a constant $C > 0$ and any a_1, a_2, a_3, a_4, we have $|\mathbb{E}[W_{\mathbf{i}_{1:4}}^{(a_1,a_2,a_3,a_4)}]| \le C/n^{|\mathcal{E}_{G(\mathbf{i}_{1:4})}|_*} \le C/n^{|\mathcal{V}_{G(\mathbf{i}_{1:4})}|-2}$. Therefore,

$$\frac{1}{n^4}\Big|\sum_{\mathbf{i}_{1:4}\in\mathcal{I}_2} q_{\mathbf{i}_{1:4}} \mathbb{E}\big[\mathbf{W}_{\mathbf{i}_{1:4}}^{(a_1,a_2,a_3,a_4)}\big]\Big| \le \frac{1}{n^4}\sum_{v=1}^{4|\mathcal{V}|}\frac{C}{n^{v-2}} \sum_{\mathbf{i}_{1:4}\in\mathcal{I}_2:|\mathcal{V}_{G(\mathbf{i}_{1:4})}|=v} |q_{\mathbf{i}_{1:4}}|.$$

Stratifying the inner summation over $\{\mathbf{i}_{1:4} \in \mathcal{I}_2 : |\mathcal{V}_{G(\mathbf{i}_{1:4})}| = v\}$ by its induced partition $\pi(\mathbf{i}_{1:4})$ of the $4|\mathcal{V}|$ total indices (having exactly v blocks), and applying the same argument as in (3.6), this inner summation may be bounded as $\sum_{\mathbf{i}_{1:4}\in\mathcal{I}_2:|\mathcal{V}_{G(\mathbf{i}_{1:4})}|=v} |q_{\mathbf{i}_{1:4}}| \le Cn^v$ for a constant $C > 0$. Applying this bound for each term of A_2, we obtain $|A_2| \le C/n^2$.

Combining the analyses of A_2, A_3, and A_4, we get $\mathbb{E}[(\mathrm{val}(\mathbf{W}) - \mathbb{E}[\mathrm{val}(\mathbf{W})])^4] \le C/n^2$. Then by Markov's inequality, for any $\varepsilon > 0$, $\mathbb{P}[|\mathrm{val}(\mathbf{W}) - \mathbb{E}[\mathrm{val}(\mathbf{W})]| > \varepsilon] \le C/(\varepsilon^4 n^2)$. This bound is summable over all $n \ge 1$, so almost-sure convergence follows by the Borel–Cantelli lemma. □

Combining Lemmas 3.4 and 3.5 concludes the proof of Lemma 2.13.

3.2. *Universality for symmetric generalized invariant matrices.* In this section, we prove Lemma 2.14 on the universality of tensor network values for generalized invariant matrices.

Fix the tensor network $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$. Expanding the product $\mathbf{W} = \mathbf{\Pi}\mathbf{M}\mathbf{\Pi}^\top$, the tensor network value is given by

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{V}}}\sum_{\mathbf{j},\mathbf{l}\in[n]^{\mathcal{E}}}\prod_{v\in\mathcal{V}} q_v(x_{1:k}[i_v]) \prod_{e=(u,v)\in\mathcal{E}} \Pi[i_u, j_e]M[j_e, l_e]\Pi[i_v, l_e].$$

The matrix $\mathbf{\Pi}$ may be written as $\mathbf{\Pi} = \mathbf{\Xi}\mathbf{P}$ where $\mathbf{\Xi}$ is a random sign matrix and $\mathbf{P}$ is a random permutation matrix independent of $\mathbf{\Xi}$. Let σ denote the permutation of $[n]$ for which $P[i, \sigma(i)] = 1$ for all $i \in [n]$. Then $\Pi[i_u, j_e]$ is nonzero if and only if $j_e = \sigma(i_u)$. Therefore, the tensor network value is equivalently expressed as

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{V}}}\prod_{v\in\mathcal{V}} q_v(x_{1:k}[i_v]) \prod_{e=(u,v)\in\mathcal{E}} \Xi[i_u]\cdot\Xi[i_v]\cdot M\big[\sigma(i_u), \sigma(i_v)\big]. \tag{3.16}$$

Let $\mathcal{P}$ be the set of partitions of $\mathcal{V}$. For each $\pi \in \mathcal{P}$, let $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ be the image of $(\mathcal{V}, \mathcal{E})$ under π, in the sense of Definition 3.1. For each $\mathbf{i} \in [n]^{\mathcal{V}}$, let $\pi(\mathbf{i}) \in \mathcal{P}$ be the partition induced by $\mathbf{i}$. Stratifying the summation over $\mathbf{i} \in [n]^{\mathcal{V}}$ by its induced partition $\pi(\mathbf{i})$,

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \sum_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{\mathbf{i} \in [n]^{\mathcal{V}} : \pi(\mathbf{i}) = \pi} \prod_{v \in \mathcal{V}} q_v\big(x_{1:k}[i_v]\big) \prod_{e=(u,v) \in \mathcal{E}} \Xi[i_u] \cdot \Xi[i_v] \cdot M\big[\sigma(i_u), \sigma(i_v)\big].$$

Then, defining $Q_R = \prod_{u \in R} q_u$ for the blocks $R \in \mathcal{K}_\pi \equiv \pi$, and identifying the sum over $\{\mathbf{i} \in [n]^{\mathcal{V}} : \pi(\mathbf{i}) = \pi\}$ as a sum over one distinct index for each block $R \in \mathcal{K}_\pi$, we have

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \sum_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{\mathbf{i} \in [n]^{\mathcal{K}_\pi}}^{*} \prod_{R \in \mathcal{K}_\pi} Q_R\big(x_{1:k}[i_R]\big) \prod_{(R,S) \in \mathcal{F}_\pi} \Xi[i_R] \cdot \Xi[i_S] \cdot M\big[\sigma(i_R), \sigma(i_S)\big],$$

where we recall the notation that $\sum^{*}$ restricts the summation to index tuples $\mathbf{i}$ having all indices distinct. For every $R \in \mathcal{K}_\pi$, let $\deg_{\mathrm{ext}}(R)$ be its *external degree* in G_π, that is, the total number of edges of $\mathcal{F}_\pi$ containing R (counting multiplicity) that are not self-loops. Then for every $R \in \mathcal{K}_\pi$, the number of times the factor $\Xi[i_R]$ appears in the above product is exactly $\deg_{\mathrm{ext}}(R)$ plus twice the number of self-loops on R. Since $\Xi[i_R]^2 = 1$, this implies

$$\begin{aligned} &\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) \\ &\quad = \sum_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{\mathbf{i} \in [n]^{\mathcal{K}_\pi}}^{*} \prod_{R \in \mathcal{K}_\pi} Q_R\big(x_{1:k}[i_R]\big) \Xi[i_R]^{\deg_{\mathrm{ext}}(R)} \prod_{(R,S) \in \mathcal{F}_\pi} M\big[\sigma(i_R), \sigma(i_S)\big]. \end{aligned} \tag{3.17}$$

LEMMA 3.6. *Let $\mathbb{E}$ be the expectation over $\mathbf{\Pi}$ conditional on $\mathbf{M}$ and $\mathbf{x}_1, \dots, \mathbf{x}_k$. Then Lemma 2.14 holds for $\mathbb{E}[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k)]$ in place of $\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k)$.*

PROOF. Taking expectations in (3.17) with respect to the independent signs $\mathbf{\Xi}$ and permutation σ, observe that

- If $R \in \mathcal{K}_\pi$ is such that $\deg_{\mathrm{ext}}(R)$ is odd, then $\mathbb{E}[\Xi[i_R]^{\deg_{\mathrm{ext}}(R)}] = 0$. Thus by independence of the diagonal entries of $\mathbf{\Xi}$ and distinctness of the indices of $\mathbf{i}$,

$$\mathbb{E}\Big[\prod_{R \in \mathcal{K}_\pi} \Xi[i_R]^{\deg_{\mathrm{ext}}(R)}\Big] = \mathbf{1}\big\{\deg_{\mathrm{ext}}(R) \text{ is even for all } R \in \mathcal{K}_\pi\big\}. \tag{3.18}$$

- Since σ is a uniformly random permutation on $[n]$, for any fixed tuple $\mathbf{i} \in [n]^{\mathcal{K}_\pi}$ with all entries distinct,

$$\mathbb{E}\Big[\prod_{(R,S) \in \mathcal{F}_\pi} M\big[\sigma(i_R), \sigma(i_S)\big]\Big] = \frac{(n - |\mathcal{K}_\pi|)!}{n!} \sum_{\mathbf{j} \in [n]^{\mathcal{K}_\pi}}^{*} \prod_{(R,S) \in \mathcal{F}_\pi} M[j_R, j_S], \tag{3.19}$$

  where $n!/(n - |\mathcal{K}_\pi|)!$ counts the total number of tuples $\mathbf{j} \in [n]^{\mathcal{K}_\pi}$ having distinct entries, and the right side represents a uniform average over such tuples $\mathbf{j}$.

Let us call a partition $\pi \in \mathcal{P}$ *even* if every vertex of $\mathcal{K}_\pi$ has even external degree. Then applying the above observations to take the expectation in (3.17), we obtain

$$\mathbb{E}\big[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k})\big] = \sum_{\text{even } \pi \in \mathcal{P}} B_n(\pi) \cdot Q_n(\pi) \cdot M_n(\pi), \tag{3.20}$$

where we set

$$B_n(\pi) = n^{|\mathcal{K}_\pi|} \cdot \frac{(n - |\mathcal{K}_\pi|)!}{n!}, \tag{3.21}$$

$$Q_n(\pi) = \frac{1}{n^{|\mathcal{K}_\pi|}} \sum_{\mathbf{i} \in [n]^{\mathcal{K}_\pi}}^{*} \prod_{R \in \mathcal{K}_\pi} Q_R\big(x_{1:k}[i_R]\big), \tag{3.22}$$

$$M_n(\pi) = \frac{1}{n} \sum_{\mathbf{j} \in [n]^{\mathcal{K}_\pi}}^{*} \prod_{(R,S) \in \mathcal{F}_\pi} M[j_R, j_S]. \tag{3.23}$$

It is clear that $\lim_{n\to\infty} B_n(\pi) = 1$ for every fixed π. For $Q_n(\pi)$, we may apply Lemma 3.3 with the identifications $\mathcal{S} \leftrightarrow \mathcal{K}_\pi$ and $\{q_s : s \in \mathcal{S}\} \leftrightarrow \{Q_R : R \in \mathcal{K}_\pi\}$. Then

$$\lim_{n\to\infty} Q_n(\pi) = \prod_{R \in \mathcal{K}_\pi} \mathbb{E}\big[Q_R(X_{1:k})\big].$$

For $M_n(\pi)$, since the original tensor network is connected, the graph $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ corresponding to each partition π must also be connected. Consequently, applying Lemma 3.7 below (with $p_e(\mathbf{M}) = \mathbf{M}$ for every edge e), there exists a deterministic limit value $M(G_\pi, \mathcal{D}_{\text{diag}})$ depending only on G_π and $\mathcal{D}_{\text{diag}}$ such that, almost surely, $\lim_{n\to\infty} M_n(\pi) = M(G_\pi, \mathcal{D}_{\text{diag}})$. Applying these statements to every π in (3.20), we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\big[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k})\big] &= \sum_{\text{even } \pi \in \mathcal{P}} \Big(\prod_{R \in \mathcal{K}_\pi} \mathbb{E}\big[Q_R(X_{1:k})\big] \Big) M(G_\pi, \mathcal{D}_{\text{diag}}) \\ &=: \text{lim-val}_T(X_{1:k}, \mathcal{D}_{\text{diag}}).\end{aligned}$$

This limit value depends only on T, the joint law of $(X_1, \dots, X_k)$, and the limit diagonal distribution $\mathcal{D}_{\text{diag}}$, and does not depend on the specific matrix $\mathbf{M}$, concluding the proof. □

LEMMA 3.7. *Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a deterministic symmetric matrix with limit diagonal distribution $\mathcal{D}_{\text{diag}}$, satisfying the following condition: For any fixed $\varepsilon > 0$, any diagonal monomial $p(\mathbf{x}) \in \Delta\langle \mathbf{x} \rangle$, and all large n,*

$$\max_{i \neq j} \big| p(\mathbf{M})[i, j] \big| < n^{-1/2+\varepsilon}.$$

Let $G = (\mathcal{K}, \mathcal{F})$ be a connected multi-graph such that the external degree $\deg_{\text{ext}}(R)$ is even for every vertex $R \in \mathcal{K}$. For every edge $e \in \mathcal{F}$, let $p_e(\mathbf{x})$ be a diagonal monomial labeling this edge. Then there exists a value $M(G, \mathcal{D}_{\text{diag}})$ depending only on G and $\mathcal{D}_{\text{diag}}$ such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\mathbf{j} \in [n]^{\mathcal{K}}}^{*} \prod_{e=(R,R') \in \mathcal{F}} p_e(\mathbf{M})[j_R, j_{R'}] = M(G, \mathcal{D}_{\text{diag}}).$$

PROOF. For convenience, we denote

$$M_n(G) := \frac{1}{n} \sum_{\mathbf{j} \in [n]^{\mathcal{K}}}^{*} \prod_{e=(R,R') \in \mathcal{F}} p_e(\mathbf{M})[j_R, j_{R'}].$$

We proceed by induction over the number of vertices $|\mathcal{K}|$. For the base case $|\mathcal{K}| = 1$, all edges of $\mathcal{F}$ must be self-loops, and we have

$$M_n(G) = \frac{1}{n} \sum_{j=1}^{n} \prod_{e \in \mathcal{F}} p_e(\mathbf{M})[j, j] = \frac{1}{n} \operatorname{Tr} \prod_{e \in \mathcal{F}} \Delta\big(p_e(\mathbf{M})\big).$$

Here $\prod_{e\in\mathcal{F}}\Delta(p_e(\mathbf{x}))$ is a diagonal monomial. Then, since $\mathbf{M}$ has a limit diagonal distribution, the above quantity admits a limit value as $n\to\infty$.

Next, supposing that the result is true for every multi-graph $G=(\mathcal{K},\mathcal{F})$ with $|\mathcal{K}|\le K$, we prove the result for $|\mathcal{K}|=K+1$. Define $\mathcal{K}_*:=\{R\in\mathcal{K}:\deg_{\text{ext}}(R)=2\}$.

First, consider the case where $|\mathcal{K}_*|=0$. Then

- Since every $\deg_{\text{ext}}(R)$ is even, we must have $\deg_{\text{ext}}(R)\ge 4$ for all $R\in\mathcal{K}$. Therefore, denoting by $\mathcal{F}_{\text{ext}}\subseteq\mathcal{F}$ those edges that are not self-loops, we have $4|\mathcal{K}|\le 2|\mathcal{F}_{\text{ext}}|$.
- We may assume without loss of generality that each vertex $R\in\mathcal{K}$ has exactly one self-loop: For R without a self-loop, we may add the self-loop $e=(R,R)$ with the identity label $p_e(\mathbf{M})=\text{Id}$. For R with multiple self-loops $\{e\in\mathcal{F}:e=(R,R)\}$, we may replace these by a single self-loop $e'=(R,R)$ having label $p_{e'}(\mathbf{M})=\prod_{e\in\mathcal{F}:e=(R,R)}\Delta(p_e(\mathbf{M}))$. These operations do not change the value of $M_n(G)$.

We denote by e_R the unique self-loop on each vertex $R\in\mathcal{K}$. Then it follows that

$$\begin{aligned}
\big|M_n(G)\big| &\le \frac{1}{n}\sum_{\mathbf{j}\in[n]^{\mathcal{K}}}^{*}\prod_{e=(R,R')\in\mathcal{F}}\big|p_e(\mathbf{M})[j_R,j_{R'}]\big| \\
&\le \frac{1}{n}\sum_{\mathbf{j}\in[n]^{\mathcal{K}}}\prod_{R\in\mathcal{K}}\big|p_{e_R}(\mathbf{M})[j_R,j_R]\big|\cdot\prod_{e\in\mathcal{F}_{\text{ext}}}\max_{i\ne j}\big|p_e(\mathbf{M})[i,j]\big| \\
&\le \frac{1}{n}\cdot n^{(-1/2+\varepsilon)|\mathcal{F}_{\text{ext}}|}\cdot\sum_{\mathbf{j}\in[n]^{\mathcal{K}}}\prod_{R\in\mathcal{K}}\big|p_{e_R}(\mathbf{M})[j_R,j_R]\big| \\
&\le n^{-1+\varepsilon|\mathcal{F}_{\text{ext}}|}\prod_{R\in\mathcal{K}}\left(\frac{1}{n}\sum_{j=1}^{n}\big|p_{e_R}(\mathbf{M})[j,j]\big|\right).
\end{aligned}$$

Here, the second inequality uses the constraint that indices j_R, $j_{R'}$ are distinct if $R\ne R'$, the third inequality holds for any fixed $\varepsilon>0$ and all large n by the given assumption on $\mathbf{M}$, and the last inequality applies $n^{-|\mathcal{F}_{\text{ext}}|/2}\le n^{-|\mathcal{K}|}$ as follows from the above bound $4|\mathcal{K}|\le 2|\mathcal{F}_{\text{ext}}|$. By Cauchy–Schwarz, we have

$$\left(\frac{1}{n}\sum_{j=1}^{n}\big|p_{e_R}(\mathbf{M})[j,j]\big|\right)^2\le\frac{1}{n}\sum_{j=1}^{n}p_{e_R}(\mathbf{M})[j,j]^2=\frac{1}{n}\operatorname{Tr}\Delta\big(p_{e_R}(\mathbf{M})\big)\Delta\big(p_{e_R}(\mathbf{M})\big),$$

where $\Delta(p_{e_R}(\mathbf{x}))\Delta(p_{e_R}(\mathbf{x}))$ is a diagonal monomial. Then this quantity has a limit value as $n\to\infty$, for each $R\in\mathcal{K}$. Choosing $\varepsilon<1/|\mathcal{F}_{\text{ext}}|$, we conclude that $M_n(G)\to 0$.

Next, consider the case where $|\mathcal{K}_*|>0$. We pick an arbitrary vertex $R_*\in\mathcal{K}_*$, and let $R_1,R_2\in\mathcal{K}$ be its two neighbors (where $R_1\ne R_*$ and $R_2\ne R_*$, but possibly $R_1=R_2$). Denote $e_1=(R_*,R_1)$, $e_2=(R_*,R_2)$, and assume without loss of generality as above that R_* has a unique self-loop $e_*=(R_*,R_*)$. Then

$$\begin{aligned}
M_n(G)&=\frac{1}{n}\sum_{\mathbf{j}\in[n]^{\mathcal{K}\setminus R_*}}^{*}\prod_{e=(R,R')\in\mathcal{F}\setminus\{e_1,e_2\}}p_e(\mathbf{M})[j_R,j_{R'}] \\
&\quad\times\sum_{\substack{j_{R_*}=1\\ j_{R_*}\notin\{j_S:S\in\mathcal{K}\setminus R_*\}}}^{n}p_{e_1}(\mathbf{M})[j_{R_*},j_{R_1}]p_{e_2}(\mathbf{M})[j_{R_*},j_{R_2}]p_{e_*}(\mathbf{M})[j_{R_*},j_{R_*}] \\
&:=\mathrm{I}-\sum_{S\in\mathcal{K}\setminus R_*}\mathrm{II}(S),
\end{aligned}$$

where we set

$$\mathrm{I} = \frac{1}{n} \sum_{\mathbf{j} \in [n]^{\mathcal{K} \setminus R_*}}^{*} \Bigg(\prod_{e=(R,R') \in \mathcal{F} \setminus \{e_1, e_2\}} p_e(\mathbf{M})[j_R, j_{R'}] \Bigg) \cdot \big(p_{e_1} \Delta(p_{e_*}) p_{e_2}\big)(\mathbf{M})[j_{R_1}, j_{R_2}],$$

$$\mathrm{II}(S) = \frac{1}{n} \sum_{\mathbf{j} \in [n]^{\mathcal{K} \setminus R_*}}^{*} \Bigg(\prod_{e=(R,R') \in \mathcal{F} \setminus \{e_1, e_2\}} p_e(\mathbf{M})[j_R, j_{R'}] \Bigg) \times p_{e_1}(\mathbf{M})[j_S, j_{R_1}] p_{e_2}(\mathbf{M})[j_S, j_{R_2}] p_{e_*}(\mathbf{M})[j_S, j_S].$$

Here I corresponds to the full summation over $j_{R_*} \in [n]$ without restriction, and each term $-\mathrm{II}(S)$ removes the contribution from the case $j_{R_*} = j_S$.

The term I is exactly equal to $M_n(G')$ for a multi-graph G' obtained from G by removing vertex R_* and the edges e_1, e_2, adding a new edge between R_1 and R_2 with label $p_{e_1}(\mathbf{x}) \Delta (p_{e_*}(\mathbf{x})) p_{e_2}(\mathbf{x})$. This graph G' is connected and has one fewer vertex than G. Each remaining vertex in G' has the same external degree as in G if $R_1 \neq R_2$, and if $R_1 = R_2$ then the external degree of $R_1 = R_2$ is reduced by 2. In both cases, all external degrees in G' remain even. Then applying the inductive hypothesis to G', $\lim_{n \to \infty} \mathrm{I}$ exists and depends only on $(G', \mathcal{D}_{\mathrm{diag}})$.

Each term $\mathrm{II}(S)$ is exactly equal to $M_n(G')$ for a multi-graph G' that merges the vertices S and R_* of G into a single vertex S_* in G', and preserves all edges and their labels. The new vertex S_* in G' has external degree equal to $\deg_{\mathrm{ext}}(S) + \deg_{\mathrm{ext}}(R_*) - 2|\{e \in \mathcal{F} : e = (S, R_*)\}|$, which is even. It is clear that G' remains connected, and the external degrees of all other vertices of G' remain the same as in G. Then applying the inductive hypothesis to G', also $\lim_{n \to \infty} \mathrm{II}(S)$ exists and depends only on $(G', \mathcal{D}_{\mathrm{diag}})$, completing the induction. □

REMARK 3.8. In the language of [52], our proof of Lemma 3.6 shows that if $\mathbf{W}$ is invariant in law under conjugation by permutations, then the expected tensor network value has a limit if $\mathbf{W}$ converges in traffic distribution, and this value is universal across matrices having the same limiting traffic distribution. Our arguments of Lemmas 3.6 and 3.7 further establish that if $\mathbf{W}$ is also invariant under conjugation by random signs and satisfies the additional delocalization conditions of Definition 2.6, then it has a limit traffic distribution that is uniquely determined by its limit diagonal law.

We provide in Appendix C an alternative computation of $\text{lim-val}_T(X_1, \dots, X_k, \mathcal{D}_{\mathrm{diag}})$ for the special case where $\mathbf{W}$ is orthogonally invariant in law, using the orthogonal Weingarten calculus [23]. We establish the asymptotic freeness statement of Proposition 2.16(b) also in Appendix C via this computation.

Finally, we conclude the proof of Lemma 2.14 by showing concentration of the tensor network value.

LEMMA 3.9. *Let $\mathbb{E}$ be the expectation over $\mathbf{\Pi}$ conditional on $\mathbf{M}$ and $\mathbf{x}_1, \dots, \mathbf{x}_k$. Under the setting of Lemma 2.14, almost surely as $n \to \infty$,*

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k) - \mathbb{E}\big[\mathrm{val}_T(\mathbf{W}; \mathbf{x}_1, \dots, \mathbf{x}_k)\big] \to 0.$$

PROOF. Let us write as shorthand $\mathrm{val}(\mathbf{W}) = \mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k})$. By Jensen's inequality,

$$\mathbb{E}\big[\big(\mathrm{val}(\mathbf{W}) - \mathbb{E}\,\mathrm{val}(\mathbf{W})\big)^4\big] \le \mathbb{E}\big[\big(\mathrm{val}(\mathbf{W}) - \mathrm{val}(\bar{\mathbf{W}})\big)^4\big],$$

where $\bar{\mathbf{W}} = \bar{\mathbf{\Pi}} \mathbf{M} \bar{\mathbf{\Pi}}^\top$, $\bar{\mathbf{\Pi}}$ is an independent copy of $\mathbf{\Pi}$, and the expectation on the right side is over $(\mathbf{\Pi}, \bar{\mathbf{\Pi}})$. We proceed to bound this expectation.

Recall the tensor network $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$. Let $(\mathcal{V}^{(1)}, \mathcal{E}^{(1)}), \dots, (\mathcal{V}^{(4)}, \mathcal{E}^{(4)})$ denote four copies of the tree $(\mathcal{V}, \mathcal{E})$. For any subset $A \subseteq \{1,2,3,4\}$, let

$$(\mathcal{V}_A, \mathcal{E}_A) \cong \bigsqcup_{a\in A} (\mathcal{V}^{(a)}, \mathcal{E}^{(a)})$$

denote the graph that is the disjoint union of those copies corresponding to $a \in A$, that is, $(\mathcal{V}_A, \mathcal{E}_A)$ has $|A|$ connected components, each a copy of $(\mathcal{V}, \mathcal{E})$. We label each vertex $v \in \mathcal{V}^{(a)} \subseteq \mathcal{V}_A$ with the same label q_v as in the original tensor network T. We write $\bar{A} = \{1,2,3,4\} \setminus A$ as the complement of A, and $\bar{\boldsymbol{\Xi}}$ and $\bar{\sigma}$ for the random sign matrix and random permutation corresponding to $\bar{\boldsymbol{\Pi}}$. Then we have, similar to (3.16),

$$\begin{aligned}
&\big(\mathrm{val}(\mathbf{W}) - \mathrm{val}(\bar{\mathbf{W}})\big)^4 \\
&\quad= \sum_{A\subseteq\{1,2,3,4\}} (-1)^{|A|} \prod_{a\in A} \mathrm{val}(\mathbf{W}) \prod_{a\in\bar{A}} \mathrm{val}(\bar{\mathbf{W}}) \\
&\quad= \sum_{A\subseteq\{1,2,3,4\}} (-1)^{|A|} \frac{1}{n^4} \sum_{\mathbf{i}\in[n]^{\mathcal{V}_A}} \prod_{v\in\mathcal{V}_A} q_v\big(x_{1:k}[i_v]\big) \\
&\qquad\times \prod_{(u,v)\in\mathcal{E}_A} \Xi[i_u]\cdot\Xi[i_v]\cdot M\big[\sigma(i_u),\sigma(i_v)\big] \\
&\qquad\times \sum_{\mathbf{j}\in[n]^{\mathcal{V}_{\bar{A}}}} \prod_{v\in\mathcal{V}_{\bar{A}}} q_v\big(x_{1:k}[j_v]\big) \prod_{(u,v)\in\mathcal{E}_{\bar{A}}} \bar{\Xi}[j_u]\cdot\bar{\Xi}[j_v]\cdot M\big[\bar{\sigma}(j_u),\bar{\sigma}(j_v)\big].
\end{aligned} \tag{3.24}$$

Let $\mathcal{P}_A$ be the set of partitions of $\mathcal{V}_A$, and denote by $\pi(\mathbf{i}) \in \mathcal{P}_A$ the partition induced by $\mathbf{i} \in [n]^{\mathcal{V}_A}$. For each $\pi \in \mathcal{P}_A$, let $G_\pi = (\mathcal{K}_\pi, \mathcal{F}_\pi)$ be the image of $(\mathcal{V}_A, \mathcal{E}_A)$ under π, in the sense of Definition 3.1. Note that here, G_π is not necessarily connected but can consist of up to $|A| \le 4$ connected components. Define $B_n(\pi)$ and $Q_n(\pi)$ exactly as in (3.21)–(3.22), let $\mathcal{C}(\pi)$ denote the set of connected components of G_π, and define

$$M_n(\pi) = \frac{1}{n^{|\mathcal{C}(\pi)|}} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_\pi}}^{*} \prod_{(R,S)\in\mathcal{F}_\pi} M[j_R, j_S]. \tag{3.25}$$

This coincides with our previous definition of (3.23) when $\mathcal{C}(\pi) = 1$. Define similarly $B_n(\bar{\pi})$, $Q_n(\bar{\pi})$, $M_n(\bar{\pi})$ via the graph $G_{\bar{\pi}} = (\mathcal{K}_{\bar{\pi}}, \mathcal{F}_{\bar{\pi}})$ that is, the image of $(\mathcal{V}_{\bar{A}}, \mathcal{E}_{\bar{A}})$ under $\bar{\pi} \in \mathcal{P}_{\bar{A}}$. Then, stratifying the sums over $\mathbf{i}$ and $\mathbf{j}$ by $\pi(\mathbf{i}) \in \mathcal{P}_A$ and $\pi(\mathbf{j}) \in \mathcal{P}_{\bar{A}}$, and taking the expectation in (3.24) over $(\boldsymbol{\Pi}, \bar{\boldsymbol{\Pi}})$ using (3.18)–(3.19), we get analogously to (3.20)

$$\begin{aligned}
&\mathbb{E}\big[\big(\mathrm{val}(\mathbf{W}) - \mathrm{val}(\bar{\mathbf{W}})\big)^4\big] \\
&\quad= \sum_{A\subseteq\{1,2,3,4\}} (-1)^{|A|} \sum_{\substack{\text{even } \pi\in\mathcal{P}_A \\ \text{even } \bar{\pi}\in\mathcal{P}_{\bar{A}}}} \frac{n^{|\mathcal{C}(\pi)|+|\mathcal{C}(\bar{\pi})|}}{n^4} \\
&\qquad\times B_n(\pi)B_n(\bar{\pi})\cdot Q_n(\pi)Q_n(\bar{\pi})\cdot M_n(\pi)M_n(\bar{\pi}).
\end{aligned} \tag{3.26}$$

For $\pi \in \mathcal{P}_A$ and $\bar{\pi} \in \mathcal{P}_{\bar{A}}$, we define $\tau = \pi \oplus \bar{\pi} \in \mathcal{P}_{\{1,2,3,4\}}$ as the combined partition of all vertices in $\mathcal{V}_{\{1,2,3,4\}}$ given by taking the blocks of both π and $\bar{\pi}$. We write $G_\tau = (\mathcal{K}_\tau, \mathcal{F}_\tau)$ as the image of $(\mathcal{V}_{\{1,2,3,4\}}, \mathcal{E}_{\{1,2,3,4\}})$ under τ; this is the disjoint union of G_π and $G_{\bar{\pi}}$, so in particular

$$|\mathcal{K}_\tau| = |\mathcal{K}_\pi| + |\mathcal{K}_{\bar{\pi}}|, \qquad |\mathcal{C}(\tau)| = |\mathcal{C}(\pi)| + |\mathcal{C}(\bar{\pi})|.$$

We now proceed to approximate $B_n(\pi)B_n(\bar{\pi})$, $Q_n(\pi)Q_n(\bar{\pi})$, and $M_n(\pi)M_n(\bar{\pi})$ by quantities that depend only on τ, and not on the individual partitions $\pi, \bar{\pi}$. We write $O(n^{-\nu})$ for any error of magnitude at most C/n^{ν} for a constant $C := C(\pi, \bar{\pi}) > 0$ and all large n.

For B_n, observe from the definition (3.21) that

$$\begin{aligned} B_n(\tau) &= \frac{n}{n} \cdot \frac{n}{n-1} \cdot \frac{n}{n-2} \cdot \ldots \cdot \frac{n}{n-|\mathcal{K}_\tau|+1} \\ &= 1 + \frac{\sum_{k=0}^{|\mathcal{K}_\tau|-1} k}{n} + O(n^{-2}) = 1 + n^{-1}\binom{|\mathcal{K}_\tau|}{2} + O(n^{-2}). \end{aligned}$$

Similarly,

$$B_n(\pi)B_n(\bar{\pi}) = 1 + n^{-1}\left(\binom{|\mathcal{K}_\pi|}{2} + \binom{|\mathcal{K}_{\bar{\pi}}|}{2}\right) + O(n^{-2}).$$

In particular,

$$B_n(\pi)B_n(\bar{\pi}) = B_n(\tau) + O(n^{-1}) = 1 + O(n^{-1}). \tag{3.27}$$

In the case where $G_\tau = (\mathcal{K}_\tau, \mathcal{F}_\tau)$ has 4 connected components, that is, each block of both π and $\bar{\pi}$ is contained within a single copy $\mathcal{V}^{(a)}$ of $\mathcal{V}$, let us write $G_\tau(a) = (\mathcal{K}_\tau(a), \mathcal{F}_\tau(a))$ for the component corresponding to the partition of $\mathcal{V}^{(a)}$. Given any $\pi \in \mathcal{P}_A$, $\bar{\pi} \in \mathcal{P}_{\bar{A}}$, and $\tau = \pi \oplus \bar{\pi}$, we then have

$$\binom{|\mathcal{K}_\tau|}{2} = \binom{|\mathcal{K}_\pi|}{2} + \binom{|\mathcal{K}_{\bar{\pi}}|}{2} + \sum_{a\in A, b\notin A} |\mathcal{K}_\tau(a)| \cdot |\mathcal{K}_\tau(b)|$$

because to choose two elements of $\mathcal{K}_\tau$, we may choose them both from $\mathcal{K}_\pi$, both from $\mathcal{K}_{\bar{\pi}}$, or one from $\mathcal{K}_\tau(a) \subseteq \mathcal{K}_\pi$ and the other from $\mathcal{K}_\tau(b) \subseteq \mathcal{K}_{\bar{\pi}}$ for some $a \in A$, $b \in \bar{A}$. This gives a refinement of (3.27),

$$B_n(\pi)B_n(\bar{\pi}) = B_n(\tau) + \sum_{a\in A, b\notin A} B_n(\tau, a, b) + O(n^{-2}), \tag{3.28}$$

where we define $B_n(\tau, a, b) = -n^{-1}|\mathcal{K}_\tau(a)| \cdot |\mathcal{K}_\tau(b)|$. Here $B_n(\tau, a, b) = O(n^{-1})$.

For Q_n, observe from the definition (3.22) that the distinction between $Q_n(\pi)Q_n(\bar{\pi})$ and $Q_n(\tau)$ is that the former does not restrict indices of summation corresponding to π to be distinct from those corresponding to $\bar{\pi}$, that is,

$$\begin{aligned} Q_n(\pi)Q_n(\bar{\pi}) = Q_n(\tau) + \frac{1}{n^{|\mathcal{K}_\tau|}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_{\bar{\pi}}}}^{*} \mathbf{1}\{\text{there is at least 1 pair of coinciding} \\ \text{indices between } \mathbf{i} \text{ and } \mathbf{j}\} \times \prod_{R\in\mathcal{K}_\pi} Q_R(x_{1:k}[i_R]) \prod_{R'\in\mathcal{K}_{\bar{\pi}}} Q_{R'}(x_{1:k}[j_{R'}]). \end{aligned}$$

By Lemma 3.3, the quantities $Q_n(\pi)$, $Q_n(\bar{\pi})$, and $Q_n(\tau)$ all have deterministic limit values as $n \to \infty$. Furthermore, by a simple inclusion-exclusion argument together with Lemma 3.3, in the above double summation the contribution from pairs $(\mathbf{i}, \mathbf{j})$ coinciding in exactly k pairs of indices is of size $O(n^{-k})$. Then in particular,

$$Q_n(\pi)Q_n(\bar{\pi}) = Q_n(\tau) + O(n^{-1}) = O(1). \tag{3.29}$$

In the case where G_τ has 4 connected components, let us write more explicitly

$$\begin{aligned} &Q_n(\pi)Q_n(\bar{\pi}) \\ &\quad = Q_n(\tau) + \frac{1}{n^{|\mathcal{K}_\tau|}} \sum_{\mathbf{i}\in[n]^{\mathcal{K}_\pi}}^{*} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_{\bar{\pi}}}}^{*} \mathbf{1}\{\text{there is exactly 1 pair of coinciding} \\ &\qquad \text{indices between } \mathbf{i} \text{ and } \mathbf{j}\} \times \prod_{R\in\mathcal{K}_\pi} Q_R\big(x_{1:k}[i_R]\big) \prod_{R'\in\mathcal{K}_{\bar{\pi}}} Q_{R'}\big(x_{1:k}[j_{R'}]\big) + O\big(n^{-2}\big). \end{aligned}$$

We may choose the coinciding index pair by choosing 1 vertex $R \in \mathcal{K}_\tau(a) \subseteq \mathcal{K}_\pi$ for some $a \in A$, and 1 vertex $R' \in \mathcal{K}_\tau(b) \subseteq \mathcal{K}_{\bar{\pi}}$ for some $b \in \bar{A}$. Now viewing $R \in \pi$ and $R' \in \bar{\pi}$ as disjoint blocks of vertices of $\mathcal{V}$, note that if $S = R \cup R' \subseteq \mathcal{V}$ is the block obtained upon merging R, R', then by definition $Q_S = Q_R \cdot Q_{R'}$. Thus, the above is equivalent to

$$Q_n(\pi)Q_n(\bar{\pi}) = Q_n(\tau) + \sum_{a\in A, b\notin A} Q_n(\tau, a, b) + O\big(n^{-2}\big), \tag{3.30}$$

where we define

$$\begin{aligned} Q_n(\tau, a, b) = \frac{1}{n^{|\mathcal{K}_\tau|}} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_\tau}} &\mathbf{1}\{\mathbf{j} \text{ has } |\mathcal{K}_\tau| - 1 \text{ distinct indices, and 1 index from} \\ &\mathcal{K}_\tau(a) \text{ coincides with 1 index from } \mathcal{K}_\tau(b)\} \times \prod_{R\in\mathcal{K}_\tau} Q_R\big(x_{1:k}[j_R]\big). \end{aligned}$$

By the preceding arguments, $Q_n(\tau, a, b) = O(n^{-1})$.

For M_n, consider any $A \subseteq \{1, 2, 3, 4\}$ and $\pi \in \mathcal{P}_A$. Recall that $\mathcal{C}(\pi)$ is the set of connected components of G_π. Each component in $\mathcal{C}(\pi)$ takes the form $G_\sigma = (\mathcal{K}_\sigma, \mathcal{F}_\sigma)$ where σ is a partition that contains a subset of the blocks of π. Let us write $\sum^{**}_{\mathbf{j}\in[n]^{\mathcal{K}_\pi}}$ for the summation over tuples $\mathbf{j}$ such that indices corresponding to each component $\mathcal{K}_\sigma \subseteq \mathcal{K}_\pi$ are distinct, but they are not necessarily distinct across different components. Recalling (3.25), define

$$M_n^{**}(\pi) := \prod_{G_\sigma\in\mathcal{C}(\pi)} M_n(\sigma) = \frac{1}{n^{|\mathcal{C}(\pi)|}} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_\pi}}^{**} \prod_{(R,S)\in\mathcal{F}_\pi} M[j_R, j_S],$$

where $M_n^{**}(\pi)$ is now a multiplicative function over connected components of π. Since each G_σ is connected, Lemma 3.7(a) implies the existence of the limit

$$M_n^{**}(\pi) \to \prod_{G_\sigma\in\mathcal{C}(\pi)} M(G_\sigma, \mathcal{D}_{\text{diag}}). \tag{3.31}$$

Comparing the definitions of $M_n^{**}(\pi)$ and $M_n(\pi)$, we have $M_n^{**}(\pi) = M_n(\pi)$ if G_π has a single connected component $|\mathcal{C}(\pi)| = 1$, and more generally

$$\begin{aligned} M_n(\pi) = M_n^{**}(\pi) - \frac{1}{n^{|\mathcal{C}(\pi)|}} \sum_{\mathbf{j}\in[n]^{\mathcal{K}_\pi}}^{**} &\mathbf{1}\{\text{some indices of } \mathbf{j} \text{ for different connected} \\ &\text{components of } G_\pi \text{ coincide}\} \times \prod_{(R,S)\in\mathcal{F}_\pi} M[j_R, j_S]. \end{aligned}$$

For each $\mathbf{j}$ where this summand is nonzero, define $\pi'(\mathbf{j}) \in \mathcal{P}_A$ as the partition that merges those blocks of π where the corresponding indices of $\mathbf{j}$ coincide. Let $\mathcal{P}(\pi)$ be the set of all

possible such partitions $\pi'(\mathbf{j})$. (If $|\mathcal{C}(\pi)| = 1$, then $\mathcal{P}(\pi) = \varnothing$.) Then, stratifying the summation over $\mathbf{j}$ by $\pi'(\mathbf{j}) \in \mathcal{P}(\pi)$, letting $G_{\pi'} = (\mathcal{K}_{\pi'}, \mathcal{F}_{\pi'})$ be the image of $(\mathcal{V}_A, \mathcal{E}_A)$ under π', and identifying the sum over $\{\mathbf{j} : \pi'(\mathbf{j}) = \pi'\}$ as a sum over one distinct index for each $R \in \mathcal{K}_{\pi'}$,

$$\begin{aligned} M_n(\pi) &= M_n^{**}(\pi) - \frac{1}{n^{|\mathcal{C}(\pi)|}} \sum_{\pi' \in \mathcal{P}(\pi)} \sum_{\mathbf{j} \in [n]^{\mathcal{K}_{\pi'}}}^{*} \prod_{(R,S) \in \mathcal{F}_{\pi'}} M[j_R, j_S] \\ &= M_n^{**}(\pi) - \sum_{\pi' \in \mathcal{P}(\pi)} \frac{1}{n^{|\mathcal{C}(\pi)| - |\mathcal{C}(\pi')|}} M_n(\pi'). \end{aligned} \tag{3.32}$$

For any $\pi' \in \mathcal{P}(\pi)$, its number of connected components satisfies $|\mathcal{C}(\pi')| \le |\mathcal{C}(\pi)| - 1$. In particular, if $|\mathcal{C}(\pi)| = 2$, then $|\mathcal{C}(\pi')| = 1$ for all $\pi' \in \mathcal{P}(\pi)$, so $M_n(\pi') = M_n^{**}(\pi')$ on the right side of (3.32). If $|\mathcal{C}(\pi)| \ge 3$, then we may apply this identity (3.32) recursively to further approximate $M_n(\pi')$ on the right side of (3.32) by $M_n^{**}(\pi')$, until only instances of M_n^{**} and no instances of M_n remain. Applying (3.31) to each instance of M_n^{**} in this final expression, this shows that

$$M_n(\pi) = M_n^{**}(\pi) + O(n^{-1}) = O(1).$$

Applying this for $\pi \in \mathcal{P}_A$, $\bar{\pi} \in \mathcal{P}_{\bar{A}}$, and $\tau = \pi \oplus \bar{\pi}$, and recalling that M_n^{**} is multiplicative across connected components so that $M_n^{**}(\tau) = M_n^{**}(\pi) M_n^{**}(\bar{\pi})$, this yields

$$M_n(\pi) M_n(\bar{\pi}) = M_n(\tau) + O(n^{-1}) = O(1). \tag{3.33}$$

When G_τ has 4 connected components, let us derive a more explicit expression for this $O(n^{-1})$ error. Applying (3.32) and the above arguments to τ, we have

$$M_n(\tau) = M_n^{**}(\tau) - \frac{1}{n} \sum_{\tau' \in \mathcal{P}(\tau) : |\mathcal{C}(\tau')| = |\mathcal{C}(\tau)| - 1} M_n^{**}(\tau') + O(n^{-2}).$$

If $\tau' \in \mathcal{P}(\tau)$ and $|\mathcal{C}(\tau')| = |\mathcal{C}(\tau)| - 1$, then τ' is obtained by picking exactly two connected components of G_τ, say $G_\tau(a) = (\mathcal{K}_\tau(a), \mathcal{F}_\tau(a))$ and $G_\tau(b) = (\mathcal{K}_\tau(b), \mathcal{F}_\tau(b))$, and merging one or more pairs of blocks $R \in \mathcal{K}_\tau(a)$ with $R' \in \mathcal{K}_\tau(b)$. We write the set of such partitions $\tau' \in \mathcal{P}(\tau)$ corresponding to the two fixed indices $a \ne b$ as $\mathcal{P}(\tau, a, b)$. Then

$$M_n(\tau) = M_n^{**}(\tau) - \frac{1}{n} \sum_{1 \le a < b \le 4} \sum_{\tau' \in \mathcal{P}(\tau, a, b)} M_n^{**}(\tau') + O(n^{-2}).$$

Similarly

$$M_n(\pi) = M_n^{**}(\pi) - \frac{1}{n} \sum_{\substack{a<b \\ a,b \in A}} \sum_{\pi' \in \mathcal{P}(\pi, a, b)} M_n^{**}(\pi') + O(n^{-2}),$$

$$M_n(\bar{\pi}) = M_n^{**}(\bar{\pi}) - \frac{1}{n} \sum_{\substack{a<b \\ a,b \in \bar{A}}} \sum_{\bar{\pi}' \in \mathcal{P}(\bar{\pi}, a, b)} M_n^{**}(\bar{\pi}') + O(n^{-2}).$$

Taking the product of these two expressions and applying multiplicativity of M_n^{**}, we deduce

$$M_n(\pi) M_n(\bar{\pi}) = M_n(\tau) + \sum_{a \in A, b \in \bar{A}} M_n(\tau, a, b) + O(n^{-2}), \tag{3.34}$$

where we define $M_n(\tau, a, b) = \frac{1}{n} \sum_{\tau' \in \mathcal{P}(\tau, a, b)} M_n^{**}(\tau')$. Here again, $M_n(\tau, a, b) = O(n^{-1})$.

Equipped with these approximations, we now bound (3.26). Given $\tau \in \mathcal{P}_{\{1,2,3,4\}}$, let $\mathcal{A}(\tau)$ be the set of $A \subseteq \{1, 2, 3, 4\}$ for which $\tau = \pi \oplus \bar{\pi}$ for some $\pi \in \mathcal{P}_A$ and $\bar{\pi} \in \mathcal{P}_{\bar{A}}$, that is,

$A \in \mathcal{A}(\tau)$ if and only if each connected component of G_τ corresponds to vertices belonging entirely to $\mathcal{V}_A$ or entirely to $\mathcal{V}_{\bar{A}}$. Note that given $\tau = \pi \oplus \bar{\pi}$ and $A \in \mathcal{A}(\tau)$, this uniquely determines $\pi \in \mathcal{P}_A$ and $\bar{\pi} \in \mathcal{P}_{\bar{A}}$. Then, stratifying the summation in (3.26) by the number of connected components $|\mathcal{C}(\tau)| = |\mathcal{C}(\pi)| + |\mathcal{C}(\bar{\pi})|$, we have

$$\mathbb{E}\big[\big(\mathrm{val}(\mathbf{W}) - \mathrm{val}(\bar{\mathbf{W}})\big)^4\big] = E_1 + E_2 + E_3 + E_4,$$

where

$$E_j = \frac{1}{n^{4-j}} \sum_{\substack{\text{even } \tau \in \mathcal{P}_{\{1,2,3,4\}} \\ |\mathcal{C}(\tau)| = j}} \sum_{A \in \mathcal{A}(\tau)} (-1)^{|A|} B_n(\pi) B_n(\bar{\pi}) \cdot Q_n(\pi) Q_n(\bar{\pi}) \cdot M_n(\pi) M_n(\bar{\pi}).$$

Here, π and $\bar{\pi}$ on the right side denote those partitions that are uniquely determined by $\tau = \pi \oplus \bar{\pi}$ and $A \in \mathcal{A}(\tau)$; we omit their dependence on (τ, A) for brevity.

Applying the simple the bound $B_n(\pi) B_n(\bar{\pi}) Q_n(\pi) Q_n(\bar{\pi}) M_n(\pi) M_n(\bar{\pi}) = O(1)$ from (3.27), (3.29), and (3.33), we get $E_1 = O(n^{-3})$ and $E_2 = O(n^{-2})$.

For E_3, applying the approximation $B_n(\pi) B_n(\bar{\pi}) = B_n(\tau) + O(n^{-1})$, $Q_n(\pi) Q_n(\bar{\pi}) = Q_n(\tau) + O(n^{-1})$, and $M_n(\pi) M_n(\bar{\pi}) = M_n(\tau) + O(n^{-1})$ from (3.27), (3.29), and (3.33), we have

$$E_3 = \frac{1}{n} \sum_{\text{even } \tau \in \mathcal{P}_{\{1,2,3,4\}} : |\mathcal{C}(\tau)| = 3} B_n(\tau) Q_n(\tau) M_n(\tau) \sum_{A \in \mathcal{A}(\tau)} (-1)^{|A|} + O\big(n^{-2}\big).$$

Importantly, the leading term $B_n(\tau) Q_n(\tau) M_n(\tau)$ does not depend on A, so we have factored it outside of the sum over A, and the lower order terms all contribute to the $O(n^{-2})$ error. For any τ where $|\mathcal{C}(\tau)| = 3$, we have $\sum_{A \in \mathcal{A}(\tau)} (-1)^{|A|} = 0$: For example, if the 3 connected components of G_τ correspond to vertices in $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_{\{3,4\}}$, then $\mathcal{A}(\tau) = \{\varnothing, \{1\}, \{2\}, \{1, 2\}, \{3, 4\}, \{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$. Thus, we get $E_3 = O(n^{-2})$.

Finally, for E_4, we apply the finer approximations (3.28), (3.30), and (3.34) which hold when $|\mathcal{C}(\tau)| = 4$. In this case $\mathcal{A}(\tau)$ consists of all subsets of $\{1, 2, 3, 4\}$, so

$$\begin{aligned} E_4 = \sum_{\text{even } \tau \in \mathcal{P}_{\{1,2,3,4\}} : |\mathcal{C}(\tau)| = 4} \Bigg(& B_n(\tau) Q_n(\tau) M_n(\tau) \sum_{A \subseteq \{1,2,3,4\}} (-1)^{|A|} \\ & + \sum_{a \neq b \in \{1,2,3,4\}} \big[B_n(\tau) Q_n(\tau) M_n(\tau, a, b) + B_n(\tau) Q_n(\tau, a, b) M_n(\tau) \\ & + B_n(\tau, a, b) Q_n(\tau) M_n(\tau) \big] \sum_{\substack{A \subseteq \{1,2,3,4\} \\ a \in A, b \in \bar{A}}} (-1)^{|A|} \Bigg) + O\big(n^{-2}\big). \end{aligned}$$

Importantly, we have exchanged the order of summations over A and over $(a \in A, b \in \bar{A})$, and used that each term $B_n(\tau, a, b)$, $Q_n(\tau, a, b)$, $M_n(\tau, a, b)$ does not depend on the assignment of the remaining indices $\{1, 2, 3, 4\} \setminus \{a, b\}$ to A and $\bar{A}$. Then, applying $\sum_{A \subseteq \{1,2,3,4\}} (-1)^{|A|} = 0$ and also $\sum_{A \subseteq \{1,2,3,4\} : a \in A, b \in \bar{A}} (-1)^{|A|} = 0$ for each fixed pair a, b, we get $E_4 = O(n^{-2})$.

Combining the above, we have $\mathbb{E}[(\mathrm{val}(\mathbf{W}) - \mathbb{E}\,\mathrm{val}(\mathbf{W}))^4] \le C/n^2$ for a constant $C > 0$ and all large n. Then Lemma 3.9 follows from Markov's inequality and the Borel–Cantelli lemma. □

Combining Lemmas 3.6 and 3.9 completes the proof of Lemma 2.14.

3.3. *Universality of AMP via polynomial approximation.* We now prove Lemma 2.12, showing that the universality of AMP for Lipschitz nonlinearities[5] can be obtained from universality of tensor network values by polynomial approximation.

For the given AMP algorithm with Lipschitz nonlinearities $u_{t+1}(\cdot)$, we approximate it by an auxiliary AMP algorithm with polynomial nonlinearities $\tilde{u}_{t+1}(\cdot)$, where each $\tilde{u}_{t+1}$ is an L_2-approximation for u_{t+1} with respect to the state evolution of its arguments. A similar method of approximation was recently used in [33]. Combining this approximation, the validity of state evolution for polynomial AMP applied to $\mathbf{G}$, and the universality of tensor network values for $\mathbf{G}$ and $\mathbf{W}$, we show that iterates of the Lipschitz and polynomial AMP algorithms applied to $\mathbf{W}$ are close in (normalized) ℓ_2 distance. This will imply the desired W_2-convergence of the AMP iterates (2.1) to their state evolution.

We construct the auxiliary AMP algorithm as follows: Fix any $\varepsilon > 0$. For the same initialization $\tilde{\mathbf{u}}_1 = \mathbf{u}_1$ and vectors of side information $\mathbf{f}_1, \dots, \mathbf{f}_k$ as in the given Lipschitz AMP algorithm (2.1), define the iterates for $t = 1, 2, 3, \dots$

$$\begin{aligned} \tilde{\mathbf{z}}_t &= \mathbf{W}\tilde{\mathbf{u}}_t - \sum_{s=1}^{t} \tilde{b}_{ts}\tilde{\mathbf{u}}_s, \\ \tilde{\mathbf{u}}_{t+1} &= \tilde{u}_{t+1}(\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_t, \mathbf{f}_1, \dots, \mathbf{f}_k) \end{aligned} \tag{3.35}$$

such that

1. Each coefficient $\tilde{b}_{ts}$ is defined by $\tilde{u}_2, \tilde{u}_3, \dots, \tilde{u}_t$ and the orthogonally invariant prescription (2.6).
2. Let $\widetilde{\mathbf{\Sigma}}_t$ be the orthogonally invariant prescription (2.5), and let $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$ be the state evolution where $\widetilde{Z}_{1:t} \sim \mathcal{N}(0, \widetilde{\mathbf{\Sigma}}_t)$ is independent of $(U_1, F_{1:k})$. Then each polynomial $\tilde{u}_{t+1}(\cdot)$ is chosen to satisfy

$$\mathbb{E}\big[\big(\tilde{u}_{t+1}(\widetilde{Z}_{1:t}, F_{1:k}) - u_{t+1}(\widetilde{Z}_{1:t}, F_{1:k})\big)^2\big] < \varepsilon. \tag{3.36}$$

3. For any fixed arguments $z_{1:(t-1)}$ and $f_{1:k}$, the function $z_t \mapsto \tilde{u}_{t+1}(z_{1:t}, f_{1:k})$ has nonlinear dependence in z_t.

We write the iterates as $\tilde{\mathbf{z}}_t(\mathbf{W})$, $\tilde{\mathbf{u}}_t(\mathbf{W})$ if we want to make explicit that the algorithm is evaluated on the matrix $\mathbf{W}$.

The choice of $\tilde{u}_{t+1}$ in condition (2) above is possible by the polynomial density condition in Assumption 2.1, and by Lemma A.1 which ensures that the same density condition holds for $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$. If condition (3) does not hold for this polynomial $\tilde{u}_{t+1}$, then it must hold upon adding to $\tilde{u}_{t+1}$ a small multiple of z_t^2. The conditions of [37], Assumption 4.2, are verified by Assumption 2.1, the condition $\mathrm{Var}[D] > 0$ given in Lemma 2.12, and the above condition (3). Then [37], Theorem 4.3, ensures, almost surely as $n \to \infty$,

$$\big(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \tilde{\mathbf{z}}_1(\mathbf{G}), \dots, \tilde{\mathbf{z}}_t(\mathbf{G})\big) \overset{W}{\to} (U_1, F_1, \dots, F_k, \widetilde{Z}_1, \dots, \widetilde{Z}_t). \tag{3.37}$$

LEMMA 3.10. *Fix any $t \geq 1$. Let $(\tilde{\mathbf{u}}_1, \tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_t, \tilde{\mathbf{u}}_t)$ be the iterates of any algorithm of the form (3.35), where $\{\tilde{b}_{ts}\}$ are scalar constants and $\tilde{u}_{t+1} : \mathbb{R}^{t+k} \to \mathbb{R}$ are polynomial functions applied row-wise. Then for any polynomial $p : \mathbb{R}^{2t+k} \to \mathbb{R}$ and for some finite set $\mathcal{F}$ of diagonal tensor networks in $k+1$ variables,*

$$\langle p(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_t, \tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_t, \mathbf{f}_1, \dots, \mathbf{f}_k)\rangle = \sum_{T \in \mathcal{F}} \mathrm{val}_T(\mathbf{W}; \mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k).$$

[5] By this we mean that each nonlinearity $u_{t+1}(\cdot)$ is Lipschitz in its first t arguments $z_1, \dots, z_t$.

PROOF. First note that

$$\langle p(\tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k})\rangle = \mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}),$$

where T is a tensor network with only one vertex v whose associated polynomial is $q_v = p$.

We claim that given any tensor network $T = (\mathcal{V}, \mathcal{E}, \{q_v\}_{v\in\mathcal{V}})$ in the variables $(\tilde{u}_{1:t}, \tilde{z}_{1:t}, f_{1:k})$, we can decompose

$$\mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) = \sum_{T'\in\mathcal{F}} \mathrm{val}_{T'}(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k}), \tag{3.38}$$

where $\mathcal{F}$ is a finite set of tensor networks in the variables $(\tilde{u}_{1:t}, \tilde{z}_{1:(t-1)}, f_{1:k})$. To show this, recall that

$$\mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) = \frac{1}{n}\sum_{\mathbf{i}\in[n]^{\mathcal{V}}} \prod_{v\in\mathcal{V}} q_v\big(\tilde{u}_{1:t}[i_v], \tilde{z}_{1:t}[i_v], f_{1:k}[i_v]\big) W_{\mathbf{i}|T}.$$

Applying $\tilde{\mathbf{z}}_t = \mathbf{W}\tilde{\mathbf{u}}_t - \sum_{s=1}^t \tilde{b}_{ts}\tilde{\mathbf{u}}_s$ and expanding each q_v in terms of $(\tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k})$ and $\mathbf{W}\tilde{\mathbf{u}}_t$, we have

$$\begin{aligned} &q_v\big(\tilde{u}_{1:t}[i_v], \tilde{z}_{1:t}[i_v], f_{1:k}[i_v]\big) \\ &\quad= \sum_{\theta=0}^{\Theta_v} q_{v,\theta}\big(\tilde{u}_{1:t}[i_v], \tilde{z}_{1:(t-1)}[i_v], f_{1:k}[i_v]\big) \cdot \left(\sum_{j=1}^n W[i_v, j]\tilde{u}_t[j]\right)^{\theta}, \end{aligned}$$

where Θ_v is the maximum degree of q_v in $\tilde{z}_t$, and $q_{v,0}, q_{v,1}, \ldots, q_{v,\Theta_v}$ are polynomials that depend on q_v and $\{\tilde{b}_{ts}\}$. Therefore

$$\begin{aligned} \mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) = \frac{1}{n} &\sum_{\boldsymbol{\theta}\in\prod_{v\in\mathcal{V}}\{0,\ldots,\Theta_v\}} \sum_{\mathbf{i}\in[n]^{\mathcal{V}}} \prod_{v\in\mathcal{V}} q_{v,\theta_v}\big(\tilde{u}_{1:t}[i_v], \tilde{z}_{1:(t-1)}[i_v], f_{1:k}[i_v]\big) \\ &\cdot \left(\sum_{j=1}^n W[i_v, j]\tilde{u}_t[j]\right)^{\theta_v} \prod_{(u,v)\in\mathcal{E}} W[i_u, i_v]. \end{aligned}$$

For each $\boldsymbol{\theta} \in \prod_{v\in\mathcal{V}}\{0, \ldots, \Theta_v\}$, we define a new tensor network $T_{\boldsymbol{\theta}}$ from T as follows: (1) for each $v \in \mathcal{V}$, replace the associated polynomial q_v by q_{v,θ_v}; (2) for each $v \in \mathcal{V}$, connect v with θ_v new vertices, where the associated polynomial for each new vertex is $q(\tilde{u}_{1:t}, \tilde{z}_{1:(t-1)}, f_{1:k}) = \tilde{u}_t$. Then the above is precisely

$$\mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) = \sum_{\boldsymbol{\theta}\in\prod_{v\in\mathcal{V}}\{0,\ldots,\Theta_v\}} \mathrm{val}_{T_{\boldsymbol{\theta}}}(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k})$$

which shows the claim (3.38).

We next claim that for any tensor network T in the variables $(\tilde{u}_{1:t}, \tilde{z}_{1:(t-1)}, f_{1:k})$, we have

$$\mathrm{val}_T(\mathbf{W}; \tilde{\mathbf{u}}_{1:t}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k}) = \mathrm{val}_{T'}(\mathbf{W}; \tilde{\mathbf{u}}_{1:(t-1)}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k}) \tag{3.39}$$

for a tensor network T' in the variables $(\tilde{u}_{1:(t-1)}, \tilde{z}_{1:(t-1)}, f_{1:k})$. This holds because $\tilde{\mathbf{u}}_t = \tilde{u}_t(\tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k})$ is itself a polynomial of $(\tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k})$, so for each vertex v of T, we may write

$$q_v\big(\tilde{\mathbf{u}}_{1:(t-1)}, \tilde{u}_t(\tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k}), \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k}\big) = \tilde{q}_v(\tilde{\mathbf{u}}_{1:(t-1)}, \tilde{\mathbf{z}}_{1:(t-1)}, \mathbf{f}_{1:k})$$

for some polynomial $\tilde{q}_v$. Then we can define T' by replacing each polynomial q_v with $\tilde{q}_v$ and preserving all other structures of T.

Having shown the reductions (3.38) and (3.39), the proof is completed by recursively applying these reductions for $t, t-1, t-2, \ldots, 1$. □

Combining the above lemma, the state evolution (3.37) for the polynomial AMP algorithm applied to $\mathbf{G}$, and the given condition in Lemma 2.12 that tensor network values have the same limit for $\mathbf{G}$ and $\mathbf{W}$, we obtain the following state evolution guarantee for the polynomial AMP algorithm applied to $\mathbf{W}$.

LEMMA 3.11. *In the setting of Lemma 2.12, for any fixed $t \geq 1$, almost surely as $n \to \infty$*

$$(\mathbf{u}_1, \mathbf{f}_1, \dots, \mathbf{f}_k, \tilde{\mathbf{z}}_1(\mathbf{W}), \dots, \tilde{\mathbf{z}}_t(\mathbf{W})) \overset{W}{\to} (U_1, F_1, \dots, F_k, \widetilde{Z}_1, \dots, \widetilde{Z}_t).$$

PROOF. By Lemma 3.10, for any polynomial $p : \mathbb{R}^{t+k+1} \to \mathbb{R}$, we have

$$\langle p(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}(\mathbf{W}))\rangle = \sum_{T \in \mathcal{F}} \mathrm{val}_T(\mathbf{W}; \mathbf{u}_1, \mathbf{f}_{1:k}),$$

where $\mathcal{F}$ is a finite set of diagonal tensor networks, and the same decomposition holds for $\mathbf{G}$ in place of $\mathbf{W}$. Then by the condition given in Lemma 2.12 and the state evolution (3.37), almost surely

$$\lim_{n\to\infty} \langle p(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}(\mathbf{W}))\rangle = \lim_{n\to\infty} \langle p(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}(\mathbf{G}))\rangle = \mathbb{E}[p(U_1, F_{1:k}, \widetilde{Z}_{1:t})]. \tag{3.40}$$

In particular, this shows that on an event $\mathcal{E}$ having probability 1, all mixed moments of the empirical distribution of rows of $(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}(\mathbf{W}))$ converge to those of $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$. Lemma A.1 implies that the joint law of $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$ is uniquely determined by its mixed moments, so on this event $\mathcal{E}$, the empirical distribution of rows converges weakly to $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$ (cf. [10], Theorem 30.2, which extends to the multivariate setting by the same proof). On this event $\mathcal{E}$, also

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \|(u_1[i], f_{1:k}[i], \tilde{z}_{1:t}(\mathbf{W})[i])\|_2^{2d} = \mathbb{E}[\|(U_1, F_{1:k}, \widetilde{Z}_{1:t})\|_2^{2d}]$$

for each integer $d \geq 1$, which shows (cf. [73], Definition 6.8 and Theorem 6.9) that

$$(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}(\mathbf{W})) \overset{W}{\to} (U_1, F_{1:k}, \widetilde{Z}_{1:t}). \qquad \square$$

REMARK 3.12. Lemmas 3.11, 2.13, and 2.14 already imply universality of the state evolution for polynomial AMP algorithms, without requiring the assumption $\|\mathbf{W}\|_{\mathrm{op}} < C$.

We now proceed with an inductive comparison of the given Lipschitz AMP algorithm (2.1) and the polynomial AMP algorithm (3.35), both applied to $\mathbf{W}$. For each $t \geq 1$, let $(U_1, F_{1:k}, Z_{1:t})$ describe the state evolution of the given Lipschitz AMP algorithm (2.1), where $Z_{1:t} \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ and $\mathbf{\Sigma}_t$ is nonsingular by assumption in Lemma 2.12. Let $(U_1, F_{1:k}, \widetilde{Z}_{1:t})$ describe the state evolution of the auxiliary AMP algorithm (3.35) where $\widetilde{Z}_{1:t} \sim \mathcal{N}(0, \widetilde{\mathbf{\Sigma}}_t)$. We write as shorthand

$$U_{s+1} = u_{s+1}(Z_{1:s}, F_{1:k}), \qquad \nabla U_{s+1} = (\partial_1 u_{s+1}, \dots, \partial_s u_{s+1})(Z_{1:s}, F_{1:k}),$$
$$\widetilde{U}_{s+1} = \tilde{u}_{s+1}(\widetilde{Z}_{1:s}, F_{1:k}), \qquad \nabla \widetilde{U}_{s+1} = (\partial_1 \tilde{u}_{s+1}, \dots, \partial_s \tilde{u}_{s+1})(\widetilde{Z}_{1:s}, F_{1:k}),$$

where the gradients are with respect to the first s arguments.

All subsequent constants may depend on the Lipschitz nonlinearities $u_2, u_3, u_4, \dots$, the corresponding Onsager coefficients $\{b_{ts}\}$ and state evolution covariances $\{\mathbf{\Sigma}_t\}$, and joint laws of $(U_1, F_{1:k}, Z_{1:t})$, which we treat as fixed throughout this argument.

LEMMA 3.13. *Fix $t \geq 1$. Suppose (3.36) holds for $\varepsilon > 0$ and every polynomial $\tilde{u}_2, \ldots, \tilde{u}_{t+1}$. Suppose also $\|\mathbf{\Sigma}_t - \widetilde{\mathbf{\Sigma}}_t\|_{\mathrm{op}} < \delta$ for $\delta > 0$. Then for any sufficiently small δ, ε, we have*

$$\max_{s=1}^{t} \big\|\mathbb{E}[\nabla U_{s+1}] - \mathbb{E}[\nabla \widetilde{U}_{s+1}]\big\|_2 < \iota(\delta,\varepsilon), \qquad \max_{r,s=1}^{t+1} \big|\mathbb{E}[U_r U_s] - \mathbb{E}[\widetilde{U}_r \widetilde{U}_s]\big| < \iota(\delta,\varepsilon)$$

for a constant $\iota(\delta,\varepsilon) > 0$ satisfying $\iota(\delta,\varepsilon) \to 0$ as $(\delta,\varepsilon) \to (0,0)$.

PROOF. We write $\iota(\delta,\varepsilon)$ for any positive constant satisfying $\iota(\delta,\varepsilon) \to 0$ as $(\delta,\varepsilon) \to (0,0)$ and changing from instance to instance. Since $\|\mathbf{\Sigma}_t - \widetilde{\mathbf{\Sigma}}_t\|_{\mathrm{op}} < \delta$, $\mathbf{\Sigma}_t$ is invertible, and $\mathbf{\Sigma}_s$ is the upper-left submatrix of $\mathbf{\Sigma}_t$ for $s \leq t$, for sufficiently small $\delta > 0$ and each $s = 1, \ldots, t$ we have

$$\big\|\mathbf{\Sigma}_s^{-1} - \widetilde{\mathbf{\Sigma}}_s^{-1}\big\|_{\mathrm{op}} < \iota(\delta,\varepsilon), \qquad \mathbb{E}\big[\|Z_{1:s} - \widetilde{Z}_{1:s}\|_2^2\big] < \iota(\delta,\varepsilon) \tag{3.41}$$

for a coupling of $Z_{1:s}$ and $\widetilde{Z}_{1:s}$.

We introduce the additional abbreviations for the intermediate quantities

$$\overline{U}_{s+1} = u_{s+1}(\widetilde{Z}_{1:s}, F_{1:k}), \qquad \nabla \overline{U}_{s+1} = (\partial_1 u_{s+1}, \ldots, \partial_s u_{s+1})(\widetilde{Z}_{1:s}, F_{1:k}).$$

Then for any $s \in [t]$,

$$\begin{aligned} &\big\|\mathbb{E}[\nabla U_{s+1}] - \mathbb{E}[\nabla \widetilde{U}_{s+1}]\big\|_2 \\ &\quad \leq \big\|\mathbb{E}[\nabla U_{s+1}] - \mathbb{E}[\nabla \overline{U}_{s+1}]\big\|_2 + \big\|\mathbb{E}[\nabla \overline{U}_{s+1}] - \mathbb{E}[\nabla \widetilde{U}_{s+1}]\big\|_2. \end{aligned} \tag{3.42}$$

Applying Stein's lemma, $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, and Cauchy–Schwarz, the first term of (3.42) is bounded as

$$\begin{aligned} &\big\|\mathbb{E}[\nabla U_{s+1}] - \mathbb{E}[\nabla \overline{U}_{s+1}]\big\|_2^2 \\ &\quad = \big\|\mathbb{E}\big[U_{s+1} \cdot \mathbf{\Sigma}_s^{-1} Z_{1:s}^\top\big] - \mathbb{E}\big[\overline{U}_{s+1} \cdot \widetilde{\mathbf{\Sigma}}_s^{-1} \widetilde{Z}_{1:s}^\top\big]\big\|_2^2 \\ &\quad \leq 3\mathbb{E}\big[(U_{s+1} - \overline{U}_{s+1})^2\big] \cdot \mathbb{E}\big[\|\mathbf{\Sigma}_s^{-1} Z_{1:s}^\top\|_2^2\big] + 3\mathbb{E}\big[\overline{U}_{s+1}^2\big] \cdot \mathbb{E}\big[\|(\mathbf{\Sigma}_s^{-1} - \widetilde{\mathbf{\Sigma}}_s^{-1}) Z_{1:s}^\top\|_2^2\big] \\ &\qquad + 3\mathbb{E}\big[\overline{U}_{s+1}^2\big] \cdot \mathbb{E}\big[\|\widetilde{\mathbf{\Sigma}}_s^{-1}(Z_{1:s}^\top - \widetilde{Z}_{1:s}^\top)\|_2^2\big]. \end{aligned}$$

The latter two terms are at most $\iota(\delta,\varepsilon)$ by (3.41), and for the first term we have

$$\mathbb{E}\big[(U_{s+1} - \overline{U}_{s+1})^2\big] \leq L_s^2 \cdot \mathbb{E}\big[\|Z_{1:s} - \widetilde{Z}_{1:s}\|_2^2\big] < \iota(\delta,\varepsilon), \tag{3.43}$$

where L_s is the Lipschitz constant of $u_{s+1}(\cdot)$. The second term of (3.42) is bounded similarly as

$$\begin{aligned} \big\|\mathbb{E}[\nabla \overline{U}_{s+1}] - \mathbb{E}[\nabla \widetilde{U}_{s+1}]\big\|_2^2 &= \big\|\mathbb{E}\big[\overline{U}_{s+1} \cdot \widetilde{\mathbf{\Sigma}}_s^{-1} \widetilde{Z}_{1:s}^\top\big] - \mathbb{E}\big[\widetilde{U}_{s+1} \cdot \widetilde{\mathbf{\Sigma}}_s^{-1} \widetilde{Z}_{1:s}^\top\big]\big\|_2^2 \\ &\leq \mathbb{E}\big[(\overline{U}_{s+1} - \widetilde{U}_{s+1})^2\big] \cdot \mathbb{E}\big[\|\widetilde{\mathbf{\Sigma}}_s^{-1} \widetilde{Z}_{1:s}^\top\|_2^2\big], \end{aligned}$$

where by (3.36) we have

$$\mathbb{E}\big[(\overline{U}_{s+1} - \widetilde{U}_{s+1})^2\big] = \mathbb{E}\big[\big(u_{s+1}(\widetilde{Z}_{1:s}, F_{1:K}) - \tilde{u}_{s+1}(\widetilde{Z}_{1:s}, F_{1:K})\big)^2\big] < \varepsilon. \tag{3.44}$$

Combining these bounds and applying them to (3.42) yields the first claim of the lemma, $\|\mathbb{E}[\nabla U_{s+1}] - \mathbb{E}[\nabla \widetilde{U}_{s+1}]\|_2 < \iota(\delta,\varepsilon)$. For the second claim, for any $r, s \in [t+1]$, we have

$$\big|\mathbb{E}[U_{r+1}U_{s+1}] - \mathbb{E}[\widetilde{U}_{r+1}\widetilde{U}_{s+1}]\big| \leq \big|\mathbb{E}\big[(U_{r+1} - \widetilde{U}_{r+1})U_{s+1}\big]\big| + \big|\mathbb{E}\big[\widetilde{U}_{r+1}(U_{s+1} - \widetilde{U}_{s+1})\big]\big|.$$

Applying again Cauchy–Schwarz and the bounds (3.43) and (3.44) yields the second claim $|\mathbb{E}[U_{r+1}U_{s+1}] - \mathbb{E}[\widetilde{U}_{r+1}\widetilde{U}_{s+1}]| < \iota(\delta,\varepsilon)$. □

LEMMA 3.14. *Fix $t \geq 1$. Suppose (3.36) holds for $\varepsilon > 0$ and every polynomial $\tilde{u}_2, \ldots, \tilde{u}_{t+1}$. Then for any sufficiently small ε, almost surely for all large n, we have*

$$\max_{s=1}^{t} \frac{1}{\sqrt{n}} \|\mathbf{z}_s(\mathbf{W}) - \tilde{\mathbf{z}}_s(\mathbf{W})\|_2 < \iota(\varepsilon), \qquad \max_{s=1}^{t} \frac{1}{\sqrt{n}} \|\mathbf{z}_s(\mathbf{W})\|_2 < C, \qquad \|\boldsymbol{\Sigma}_t - \widetilde{\boldsymbol{\Sigma}}_t\|_{\mathrm{op}} < \iota(\varepsilon)$$

for constants $C > 0$ and $\iota(\varepsilon) > 0$ satisfying $\iota(\varepsilon) \to 0$ as $\varepsilon \to 0$.

PROOF. We write $\mathbf{z}_t$, $\tilde{\mathbf{z}}_t$, $\mathbf{u}_t$, $\tilde{\mathbf{u}}_t$ for the iterates of the Lipschitz and polynomial AMP algorithms applied to $\mathbf{W}$. We prove the extended claim that there are constants $\iota(\varepsilon) > 0$ and $C > 0$, satisfying $\iota(\varepsilon) \to 0$ as $\varepsilon \to 0$, for which almost surely for all large n,

(a) $\max_{s=1}^{t} |b_{ts} - \tilde{b}_{ts}| < \iota(\varepsilon)$;
(b) $\max_{s=1}^{t} \frac{1}{\sqrt{n}} \|\mathbf{z}_s - \tilde{\mathbf{z}}_s\|_2 < \iota(\varepsilon)$ and $\max_{s=1}^{t} \frac{1}{\sqrt{n}} \|\mathbf{z}_s\|_2 < C$;
(c) $\|\boldsymbol{\Sigma}_t - \widetilde{\boldsymbol{\Sigma}}_t\|_{\mathrm{op}} < \iota(\varepsilon)$;
(d) $\max_{s=0}^{t} \frac{1}{\sqrt{n}} \|\mathbf{u}_{s+1} - \tilde{\mathbf{u}}_{s+1}\|_2 < \iota(\varepsilon)$ and $\max_{s=0}^{t} \frac{1}{\sqrt{n}} \|\mathbf{u}_{s+1}\|_2 < C$.

We induct on t. For the base case $t = 0$, statements (a–c) are vacuous, and (d) holds by the equality of initializations $\tilde{\mathbf{u}}_1 = \mathbf{u}_1$ and by the convergence of $\mathbf{u}_1$ in Assumption 2.1.

Consider any $t \geq 1$, and suppose inductively that claims (a–d) all hold for $t - 1$. Denoting the constants in this inductive claim for $t - 1$ as $\iota_{t-1}(\varepsilon)$ and C_{t-1}, let us write $\iota_t(\varepsilon)$ and C_t for any positive constants depending on $\iota_{t-1}(\varepsilon)$ and C_{t-1}, satisfying $\iota_t(\varepsilon) \to 0$ as $\varepsilon \to 0$ and $\iota_{t-1}(\varepsilon) \to 0$, and changing from instance to instance.

For (a), by the prescription (2.6), each b_{ts} is a continuous function of $\mathbb{E}[\nabla U_{s+1}]$ for $s \leq t - 1$ and $\mathbb{E}[U_r U_s]$ for $1 \leq r \leq s \leq t$. Then $\max_{s=1}^{t} |b_{ts} - \tilde{b}_{ts}| < \iota_t(\varepsilon)$ by statement (c) of the inductive hypothesis and Lemma 3.13.

For (b), by the definition of $\mathbf{z}_t$ and $\tilde{\mathbf{z}}_t$, we have

$$\frac{\|\mathbf{z}_t - \tilde{\mathbf{z}}_t\|_2}{\sqrt{n}} \leq \frac{\|\mathbf{W}(\mathbf{u}_t - \tilde{\mathbf{u}}_t)\|_2}{\sqrt{n}} + \sum_{s=1}^{t} |b_{ts}| \cdot \frac{\|\mathbf{u}_s - \tilde{\mathbf{u}}_s\|_2}{\sqrt{n}} + \sum_{s=1}^{t} |b_{ts} - \tilde{b}_{ts}| \cdot \frac{\|\tilde{\mathbf{u}}_s\|_2}{\sqrt{n}}.$$

By the assumption that $\|\mathbf{W}\|_{\mathrm{op}} \leq C$ almost surely for all large n, by claim (d) of the inductive hypothesis, and by claim (a) already shown, this is at most $\iota_t(\varepsilon)$. Also,

$$\frac{\|\mathbf{z}_t\|_2}{\sqrt{n}} \leq \frac{\|\mathbf{W}\mathbf{u}_t\|_2}{\sqrt{n}} + \sum_{s=1}^{t} |b_{ts}| \cdot \frac{\|\mathbf{u}_s\|_2}{\sqrt{n}}$$

which similarly is at most C_t.

For (c), by the prescription (2.5), the matrix $\boldsymbol{\Sigma}_t$ is (as in the proof of (a) above) a continuous function of $\mathbb{E}[\nabla U_{s+1}]$ for $s \leq t - 1$ and $\mathbb{E}[U_r U_s]$ for $1 \leq r \leq s \leq t$. Then $\|\boldsymbol{\Sigma}_t - \widetilde{\boldsymbol{\Sigma}}_t\|_{\mathrm{op}} < \iota_t(\varepsilon)$ again by statement (c) of the inductive hypothesis and Lemma 3.13.

For (d), it follows from the definitions of $\mathbf{u}_{t+1}$ and $\tilde{\mathbf{u}}_{t+1}$ that

$$\begin{aligned}
\frac{\|\mathbf{u}_{t+1} - \tilde{\mathbf{u}}_{t+1}\|_2}{\sqrt{n}} &= \frac{\|u_{t+1}(\mathbf{z}_{1:t}, \mathbf{f}_{1:k}) - \tilde{u}_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k})\|_2}{\sqrt{n}} \\
&\leq \frac{\|u_{t+1}(\mathbf{z}_{1:t}, \mathbf{f}_{1:k}) - u_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k})\|_2}{\sqrt{n}} \\
&\quad + \frac{\|u_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) - \tilde{u}_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k})\|_2}{\sqrt{n}}.
\end{aligned}$$

The first term is at most $\iota_t(\varepsilon)$ by statement (b) already proved and the fact that $u_{t+1}(\cdot)$ is Lipschitz. For the second term, Lemma 3.11 shows $(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) \overset{W}{\to} (\widetilde{Z}_{1:t}, F_{1:k})$ almost surely

as $n \to \infty$, and the function $(u_{t+1}(\cdot) - \tilde{u}_{t+1}(\cdot))^2$ satisfies the polynomial growth condition (1.1) by the given conditions for $u_{t+1}(\cdot)$. Then

$$\begin{aligned}\lim_{n\to\infty} &\frac{\|u_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k}) - \tilde{u}_{t+1}(\tilde{\mathbf{z}}_{1:t}, \mathbf{f}_{1:k})\|_2^2}{n} \\ &= \mathbb{E}\big[\big(u_{t+1}(\widetilde{Z}_{1:t}, F_{1:k}) - \tilde{u}_{t+1}(\widetilde{Z}_{1:t}, F_{1:k})\big)^2\big] < \varepsilon,\end{aligned} \tag{3.45}$$

where the inequality is due to (3.36) in the construction of $\tilde{u}_{t+1}$. Thus $\|\mathbf{u}_{t+1} - \tilde{\mathbf{u}}_{t+1}\|_2/\sqrt{n} < \iota_t(\varepsilon)$. Similarly,

$$\frac{\|\mathbf{u}_{t+1}\|_2}{\sqrt{n}} \le \frac{\|u_{t+1}(\mathbf{z}_{1:t}, \mathbf{f}_{1:k}) - u_{t+1}(\mathbf{0}, \mathbf{f}_{1:k})\|_2}{\sqrt{n}} + \frac{\|u_{t+1}(\mathbf{0}, \mathbf{f}_{1:k})\|_2}{\sqrt{n}}.$$

The first term is at most C_t by statement (b) already proved and the fact that $u_{t+1}(\cdot)$ is Lipschitz. For the second term, we have $\lim_{n\to\infty} \frac{1}{n}\|u_{t+1}(\mathbf{0}, \mathbf{f}_{1:k})\|_2^2 = \mathbb{E}[u_{t+1}(0, F_{1:k})^2]$ which is also at most a constant C_t. This concludes the proof of (d) and completes the induction. □

Finally, we apply Lemma 3.14 to prove Lemma 2.12.

PROOF OF LEMMA 2.12. Let $\mathbf{u}_{1:t}$, $\tilde{\mathbf{u}}_{1:t}$, $\mathbf{z}_{1:t}$, $\tilde{\mathbf{z}}_{1:t}$ be the iterates of the Lipschitz AMP algorithm and polynomial AMP algorithm applied to $\mathbf{W}$. We write $\iota(\varepsilon)$ for a positive constant such that $\iota(\varepsilon) \to 0$ as $\varepsilon \to 0$, and changing from instance to instance.

To show W_2-convergence of $(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t})$, consider any function $g : \mathbb{R}^{t+k+1} \to \mathbb{R}$ satisfying $|g(\mathbf{x}) - g(\mathbf{y})| \le C(1 + \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)\|\mathbf{x} - \mathbf{y}\|_2$ for a constant $C > 0$. Then

$$\begin{aligned}&\big|\langle g(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t}) - g(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t})\rangle\big| \\ &\quad\le \frac{C}{n}\sum_{i=1}^n \big(1 + \big\|\big(u_1[i], f_{1:k}[i], z_{1:t}[i]\big)\big\|_2 + \big\|\big(u_1[i], f_{1:k}[i], \tilde{z}_{1:t}[i]\big)\big\|_2\big) \cdot \big\|z_{1:t}[i] - \tilde{z}_{1:t}[i]\big\|_2 \\ &\quad\le \frac{C}{n}\sqrt{3\sum_{i=1}^n \big(1 + \big\|\big(u_1[i], f_{1:k}[i], z_{1:t}[i]\big)\big\|_2^2 + \big\|\big(u_1[i], f_{1:k}[i], \tilde{z}_{1:t}[i]\big)\big\|_2^2\big)} \\ &\qquad\cdot \sqrt{\sum_{i=1}^n \big\|z_{1:t}[i] - \tilde{z}_{1:t}[i]\big\|_2^2} \\ &\quad= \frac{C}{n}\sqrt{3n + 6\|\mathbf{u}_1\|_2^2 + 6\sum_{j=1}^k \|\mathbf{f}_j\|_2^2 + 3\sum_{s=1}^t \big(\|\mathbf{z}_s\|_2^2 + \|\tilde{\mathbf{z}}_s\|_2^2\big)} \cdot \sqrt{\sum_{s=1}^t \|\mathbf{z}_s - \tilde{\mathbf{z}}_s\|_2^2}.\end{aligned}$$

This implies, by the statements for $\mathbf{z}_{1:t}$ in Lemma 3.14 and the convergence of $(\mathbf{u}_1, \mathbf{f}_1, \ldots, \mathbf{f}_k)$ in Assumption 2.1, that almost surely for all large n,

$$\big|\langle g(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t}) - g(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t})\rangle\big| < \iota(\varepsilon). \tag{3.46}$$

Since $(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t}) \xrightarrow{W} (U_1, F_{1:k}, \widetilde{Z}_{1:t})$ by Lemma 3.11, we have

$$\lim_{n\to\infty} \langle g(\mathbf{u}_1, \mathbf{f}_{1:k}, \tilde{\mathbf{z}}_{1:t})\rangle = \mathbb{E}\big[g(U_1, F_{1:k}, \widetilde{Z}_{1:t})\big]. \tag{3.47}$$

By the statement for $\mathbf{\Sigma}_t$ in Lemma 3.14, there is a coupling of $Z_{1:t}$ and $\widetilde{Z}_{1:t}$ such that $\mathbb{E}[\|Z_{1:t} - \widetilde{Z}_{1:t}\|_2^2] < \iota(\varepsilon)$. Then similarly

$$\begin{aligned}
&\big|\mathbb{E}\big[g(U_1, F_{1:k}, Z_{1:t})\big] - \mathbb{E}\big[g(U_1, F_{1:k}, \widetilde{Z}_{1:t})\big]\big| \\
&\quad \le C \cdot \mathbb{E}\big[\big(1 + \big\|(U_1, F_{1:k}, Z_{1:t})\big\|_2 + \big\|(U_1, F_{1:k}, \widetilde{Z}_{1:t})\big\|_2\big) \cdot \|Z_{1:t} - \widetilde{Z}_{1:t}\|_2\big] \\
&\quad \le C\sqrt{\mathbb{E}[3 + 6U_1^2 + 6\|F_{1:k}\|_2^2 + 3\|Z_{1:t}\|_2^2 + 3\|\widetilde{Z}_{1:t}\|_2^2}} \cdot \sqrt{\mathbb{E}\big[\|Z_{1:t} - \widetilde{Z}_{1:t}\|_2^2\big]} \\
&\quad < \iota(\varepsilon).
\end{aligned} \tag{3.48}$$

Combining (3.46), (3.47), and (3.48), we obtain for a (different) constant $\iota(\varepsilon) > 0$, almost surely for all large n, $|\langle g(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t})\rangle - \mathbb{E}[g(U_1, F_{1:k}, Z_{1:t})]| < \iota(\varepsilon)$. Since $\varepsilon > 0$ is arbitrary and $\iota(\varepsilon) \to 0$ as $\varepsilon \to 0$, we conclude that $\lim_{n\to\infty}\langle g(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t})\rangle = \mathbb{E}[g(U_1, F_{1:k}, Z_{1:t})]$. This holds for all bounded Lipschitz functions $g(\cdot)$ as well as for $g(U_1, F_{1:k}, Z_{1:t}) = \|(U_1, F_{1:k}, Z_{1:t})\|_2^2$, which implies $(\mathbf{u}_1, \mathbf{f}_{1:k}, \mathbf{z}_{1:t}) \overset{W_2}{\to} (U_1, F_{1:k}, Z_{1:t})$ (cf. [73], Definition 6.8 and Theorem 6.9). □

Combining Lemmas 2.12 and 2.13 for $\mathbf{G} \sim \mathrm{GOE}(n)$ concludes the proof of Theorem 2.4, and combining Lemmas 2.12 and 2.14 for an orthogonally invariant matrix $\mathbf{G}$ with limit spectral distribution D concludes the proof of Theorem 2.8.

4. Discussion. In this work, we have established universality of the state evolution for AMP algorithms applied to ensembles of matrices in both Gaussian and non-Gaussian universality classes, using an unfolding of polynomial AMP algorithms into linear combinations of matrix-tensor networks. Our analyses also reveal universality classes of matrices for which these tensor networks have common limiting values, but where a more succinct characterization of the limiting behavior of first-order iterative algorithms is currently unknown. We hope that our work may inspire the development of dynamical mean-field theory descriptions of such algorithms for these broader matrix ensembles.

Recently, motivated by statistical applications, a burgeoning line of work [16, 47, 48, 64] has studied nonasymptotic guarantees for AMP algorithms, in settings where the underlying structure (e.g., sparsity) and the nonlinearities applied may depend on the dimension n, and for a number of iterations of the algorithm that may also grow with the dimension n. The study of AMP universality in such settings falls outside the scope of our current analyses, and we believe this is an interesting direction for future work.

APPENDIX A: DENSITY OF POLYNOMIALS

LEMMA A.1. *Let μ_X and μ_Y be probability laws on $\mathbb{R}^m$ and $\mathbb{R}^n$ having finite moments of all orders, such that multivariate polynomials are dense in the real L^2-spaces $L^2(\mu_X)$ and $L^2(\mu_Y)$. Then multivariate polynomials are also dense in $L^2(\mu_X \times \mu_Y)$.*

PROOF. Consider any measurable $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$, and let χ_A, χ_B, $\chi_{A\times B}$ be the indicator functions of A, B, and $A \times B$. For any $\varepsilon > 0$, by the density conditions for $L^2(\mu_X)$ and $L^2(\mu_Y)$, we may take polynomials p_A, p_B such that $\|\chi_A - p_A\|_{L^2(\mu_X)} < \varepsilon/2$ and $\|\chi_B - p_B\|_{L^2(\mu_Y)} < \varepsilon/(2\|p_A\|_{L^2(\mu_X)})$. Then

$$\begin{aligned}
&\|\chi_{A\times B} - p_A p_B\|_{L^2(\mu_X\times\mu_Y)} \\
&\quad \le \|\chi_A - p_A\|_{L^2(\mu_X)}\|\chi_B\|_{L^2(\mu_Y)} + \|p_A\|_{L^2(\mu_X)}\|\chi_B - p_B\|_{L^2(\mu_Y)} < \varepsilon.
\end{aligned}$$

Taking $\varepsilon \to 0$ shows that polynomials are dense in the linear span of indicator functions $\{\chi_{A\times B} : \text{measurable } A \subseteq \mathbb{R}^m, B \subseteq \mathbb{R}^n\}$. This linear span is in turn dense in $L^2(\mu_X \times \mu_Y)$, showing the lemma. □

APPENDIX B: SUFFICIENT CONDITIONS FOR GENERALIZED INVARIANCE

In this appendix, we prove Proposition 2.7, providing examples of matrix models that satisfy the generalized invariance condition of Definition 2.6.

LEMMA B.1. *Let $\mathbf{M} \in \mathbb{R}^{n\times n}$ be a symmetric matrix having eigenvalues $\mathbf{d} \in \mathbb{R}^n$. Suppose $\mathbf{d} \overset{W}{\to} D$ almost surely as $n \to \infty$, where D has finite moments of all orders. Suppose $\mathbf{M}$ satisfies (2.9) almost surely for all large n. Then for any $p(\mathbf{x}) \in \Delta\langle \mathbf{x}\rangle$,*

(a) $\lim_{n\to\infty} \frac{1}{n} \operatorname{Tr} p(\mathbf{M})$ *exists almost surely, is finite, and depends only on the law of D.*
(b) *For any $\varepsilon > 0$ and all large n, we have*

$$\max_{i=1}^{n} \left| p(\mathbf{M})[i,i] - \frac{1}{n} \operatorname{Tr} p(\mathbf{M}) \right| < n^{-1/2+\varepsilon}, \qquad \max_{i\neq j} \left| p(\mathbf{M})[i,j] \right| < n^{-1/2+\varepsilon}.$$

PROOF. By the definition of diagonal monomials, every $p(\mathbf{x}) \in \Delta\langle \mathbf{x}\rangle$ is a word of the form

$$p(\mathbf{x}) = \mathbf{x}^{r_1} \Delta\big(p_1(\mathbf{x})\big) \mathbf{x}^{r_2} \Delta\big(p_2(\mathbf{x})\big) \cdots \mathbf{x}^{r_L} \Delta\big(p_L(\mathbf{x})\big) \mathbf{x}^{r_{L+1}}, \tag{B.1}$$

where each $p_\ell(\mathbf{x}) \in \Delta\langle \mathbf{x}\rangle$ and each $r_\ell \geq 0$. We define the *depth* of $p(\mathbf{x})$, denoted by $\delta(p)$, as $\delta(p) = 0$ if $L = 0$ (so that $p(\mathbf{x}) = \mathbf{x}^r$ for some $r \geq 0$), and $\delta(p) = 1 + \max_{\ell=1}^{L} \delta(p_\ell)$ if $L \geq 1$. Thus $\delta(p)$ is the maximum number of "nested" applications of $\Delta(\cdot)$. We induct on $\delta(p)$.

For the base case where $\delta(p) = 0$ and $p(\mathbf{x}) = \mathbf{x}^r$, we have

$$\frac{1}{n} \operatorname{Tr} p(\mathbf{M}) = \frac{1}{n} \operatorname{Tr}\big(\mathbf{M}^r\big) = \frac{1}{n} \sum_{i=1}^{n} d[i]^r \to \mathbb{E}\big[D^r\big]$$

almost surely, by the assumption $\mathbf{d} \overset{W}{\to} D$. Thus statement (a) holds, and statement (b) holds by the assumed condition (2.9).

Suppose inductively that the lemma is true for all $p(\mathbf{x})$ with $\delta(p) \leq K$, and consider $p(\mathbf{x})$ with $\delta(p) = K + 1$. Fix any $\varepsilon > 0$. By the definition of depth, every $p_\ell(\mathbf{x})$ in (B.1) satisfies $\delta(p_\ell) \leq K$. Then for every $\ell = 1, 2, \ldots, L$, by claim (b) of the induction hypothesis, we can decompose $p_\ell(\mathbf{M}) = \frac{1}{n} \operatorname{Tr} p_\ell(\mathbf{M}) \cdot \mathrm{Id} + \mathbf{E}_\ell$ where $\mathbf{E}_\ell$ satisfies $\max_{i,j\in[n]} |E_\ell[i,j]| < n^{-1/2+\varepsilon}$ almost surely for all large n. Fix any $i, j \in [n]$ and write $i_0 \equiv i$ and $i_{L+1} \equiv j$. Then, applying this decomposition to every $p_\ell(\mathbf{M})$, we obtain

$$\begin{aligned}
p(\mathbf{M})[i,j] &= \sum_{\mathbf{i}\in[n]^L} M^{r_1}[i_0,i_1] p_1(\mathbf{M})[i_1,i_1] M^{r_2}[i_1,i_2] \cdots p_L(\mathbf{M})[i_L,i_L] M^{r_{L+1}}[i_L,i_{L+1}] \\
&= \sum_{\mathbf{i}\in[n]^L} \prod_{\ell=1}^{L+1} M^{r_\ell}[i_{\ell-1},i_\ell] \prod_{\ell=1}^{L} \left(\frac{1}{n} \operatorname{Tr} p_\ell(\mathbf{M}) + E_\ell[i_\ell,i_\ell] \right) \\
&= \sum_{\mathcal{J}\subseteq[L]} \left(\prod_{\ell\in[L]\setminus\mathcal{J}} \frac{1}{n} \operatorname{Tr} p_\ell(\mathbf{M}) \right) \sum_{\mathbf{i}\in[n]^L} \prod_{\ell=1}^{L+1} M^{r_\ell}[i_{\ell-1},i_\ell] \prod_{\ell\in\mathcal{J}} E_\ell[i_\ell,i_\ell].
\end{aligned}$$

By the induction hypothesis, the limit

$$M_{\mathcal{J}} := \lim_{n\to\infty} \prod_{\ell\in[L]\setminus\mathcal{J}} \frac{1}{n} \operatorname{Tr} p_\ell(\mathbf{M}) \tag{B.2}$$

exists, is finite, and depends only on the law of D. We set $M_{\mathcal{J}} = 1$ if $\mathcal{J} = [L]$. Note that this convergence is uniform over pairs $i, j \in [n]$. Therefore, for an error $\xi_{\mathcal{J}} = o(1)$ independent

of i and j,

$$p(\mathbf{M})[i,j] = \sum_{\mathcal{J}\subseteq[L]} (M_{\mathcal{J}} + \xi_{\mathcal{J}}) \sum_{\mathbf{i}\in[n]^L} \prod_{\ell=1}^{L+1} M^{r_\ell}[i_{\ell-1}, i_\ell] \prod_{\ell\in\mathcal{J}} E_\ell[i_\ell, i_\ell].$$

We first sum over all indices $\{i_\ell : \ell \notin \mathcal{J}\}$: Write explicitly $\mathcal{J} = \{\ell_1, \dots, \ell_{|\mathcal{J}|}\}$ where $1 \le \ell_1 < \cdots < \ell_{|\mathcal{J}|} \le L$. Let $\ell_0 = 0$ and $\ell_{|\mathcal{J}|+1} = L+1$, and denote $R_\rho = r_{\ell_{\rho-1}+1} + \cdots + r_{\ell_\rho}$. Then this gives

$$p(\mathbf{M})[i,j] = \sum_{\mathcal{J}\subseteq[L]} (M_{\mathcal{J}} + \xi_{\mathcal{J}}) \sum_{\mathbf{i}\in[n]^{\mathcal{J}}} \prod_{\rho=1}^{|\mathcal{J}|+1} M^{R_\rho}[i_{\ell_{\rho-1}}, i_{\ell_\rho}] \prod_{\ell\in\mathcal{J}} E_\ell[i_\ell, i_\ell]. \tag{B.3}$$

We denote by $C > 0$ a constant depending only on $p(\mathbf{x})$, $\mathcal{J}$, and the law of D, and changing from instance to instance. By (2.9), we have $\max_{i\in[n]} |M^{R_\rho}[i,i]| < \frac{1}{n}\operatorname{Tr}\mathbf{M}^{R_\rho} + n^{-1/2+\varepsilon} < C$ and $\max_{i\neq j} |M^{R_\rho}[i,j]| < n^{-1/2+\varepsilon}$ for each $\rho \in [|\mathcal{J}|+1]$, almost surely for all large n. For any $\mathbf{i} \in [n]^{\mathcal{J}}$, define $\Psi(\mathbf{i}) = \{\rho \in [|\mathcal{J}|+1] : i_{\ell_{\rho-1}} \neq i_{\ell_\rho}\}$. Then this implies

$$\prod_{\rho=1}^{|\mathcal{J}|+1} \bigl|M^{R_\rho}[i_{\ell_{\rho-1}}, i_{\ell_\rho}]\bigr| \prod_{\ell\in\mathcal{J}} \bigl|E_\ell[i_\ell, i_\ell]\bigr| \le C n^{(-1/2+\varepsilon)(|\Psi(\mathbf{i})|+|\mathcal{J}|)}. \tag{B.4}$$

Moreover, if $\psi \ge 1$, then note that $|\{\mathbf{i} \in [n]^{\mathcal{J}} : |\Psi(\mathbf{i})| = \psi\}| \le C n^{\psi-1}$, because $i_{\ell_0} = i_0 = i$ and $i_{\ell_{|\mathcal{J}|+1}} = i_{L+1} = j$ are fixed, so there is freedom to choose $\psi - 1$ remaining index values. Combining this with (B.4), for any $\psi \ge 1$,

$$\begin{aligned}\sum_{\mathbf{i}\in[n]^{\mathcal{J}}:|\Psi(\mathbf{i})|=\psi} \prod_{\rho=1}^{|\mathcal{J}|+1} \bigl|M^{R_\rho}[i_{\ell_{\rho-1}}, i_{\ell_\rho}]\bigr| \prod_{\ell\in\mathcal{J}} \bigl|E_\ell[i_\ell, i_\ell]\bigr| &\le C n^{\psi-1} \cdot n^{(-1/2+\varepsilon)(\psi+|\mathcal{J}|)} \\ &\le C n^{-1/2+\varepsilon(2|\mathcal{J}|+1)},\end{aligned} \tag{B.5}$$

where the last inequality follows from the observation that we always have $|\Psi(\mathbf{i})| \le |\mathcal{J}| + 1$. For $\psi = 0$, we must have $i = j$ and $|\{\mathbf{i} \in [n]^{\mathcal{J}} : |\Psi(\mathbf{i})| = \psi\}| = 1$. Then by (B.4), this bound (B.5) still holds as long as $|\mathcal{J}| \ge 1$, that is, $\mathcal{J} \neq \varnothing$.

Applying (B.5) for all nonempty $\mathcal{J} \subseteq [L]$, it follows from (B.3) that

$$\bigl|p(\mathbf{M})[i,j] - \mathbf{1}\{i=j\}(M_\varnothing + \xi_\varnothing) M^R[i,i]\bigr| \le C n^{-1/2+\varepsilon(2L+1)}, \tag{B.6}$$

where we set $R = r_1 + \cdots + r_{L+1}$. The above bounds all hold uniformly over $i, j \in [n]$, and hence (B.6) holds simultaneously for all pairs $i, j \in [n]$, almost surely for all large n. Thus, combining with the condition (2.9) for $M^R[i,i]$, we conclude that both $\max_{i\neq j} |p(\mathbf{M})[i,j]|$ and $\max_{i=1}^n |p(\mathbf{M})[i,i] - (M_\varnothing + \xi_\varnothing)\cdot\frac{1}{n}\operatorname{Tr}\mathbf{M}^R|$ are at most $n^{-1/2+\varepsilon(2L+2)}$ for all large n. Then $\{p(\mathbf{M})[i,i] : i \in [n]\}$ are uniformly close to a value independent of $i \in [n]$, which implies also $\max_{i=1}^n |p(\mathbf{M})[i,i] - \frac{1}{n}\operatorname{Tr} p(\mathbf{M})| < 2n^{-1/2+\varepsilon(2L+2)}$. These statements hold for any $\varepsilon > 0$, showing the inductive claim (b) for $p(\mathbf{x})$. Moreover, averaging (B.6) over $i = j$ gives

$$\lim_{n\to\infty} \frac{1}{n}\operatorname{Tr} p(\mathbf{M}) = \lim_{n\to\infty} (M_\varnothing + \xi_\varnothing)\cdot\frac{1}{n}\operatorname{Tr} M^R = M_\varnothing \cdot \mathbb{E}\bigl[D^R\bigr],$$

and we recall from (B.2) that $M_\varnothing$ depends only on the law of D. This shows the inductive claim (a) for $p(\mathbf{x})$, completing the induction. □

PROOF OF PROPOSITION 2.7. Lemma B.1 implies that the matrix model in Proposition 2.7(b2) satisfies Definition 2.6, where the limit diagonal law $\mathcal{D}_{\text{diag}}$ is determined uniquely by the limit spectral distribution D. To complete the proof of Proposition 2.7, it suffices to

verify that the orthogonally invariant matrix model of part (a) and the model of part (b1) are both special cases of the model in part (b2).

If $\mathbf{W} = \mathbf{ODO}^\top$ is orthogonally invariant, that is, $\mathbf{O} \sim \mathrm{Haar}(\mathbb{O}(n))$ is independent of $\mathbf{D}$, then also $\mathbf{O} \overset{L}{=} \mathbf{\Pi}_V \mathbf{O} \mathbf{\Pi}_E$ where $\mathbf{\Pi}_V$, $\mathbf{\Pi}_E$ are uniformly random signed permutations independent of $\mathbf{O}$. The entries of $\mathbf{O}$ satisfy the delocalization condition (2.8) almost surely for all large n, as is implied by [44], Theorem 1. Thus $\mathbf{W}$ is a special case of the model in part (b1).

Now suppose $\mathbf{W}$ is any matrix satisfying the description of part (b1). Then $\mathbf{W}$ has the simpler form $\mathbf{W} = \mathbf{\Pi M \Pi}^\top$ where $\mathbf{\Pi} = \mathbf{\Pi}_V$,

$$\mathbf{M} = \mathbf{HPDP}^\top \mathbf{H}^\top,$$

and $\mathbf{P} = \mathbf{P}_E$ is the random permutation corresponding to $\mathbf{\Pi}_E = \mathbf{P}_E \mathbf{\Xi}_E$. Here, we have eliminated the diagonal sign matrices $\mathbf{\Xi}_E$ from the expression using $\mathbf{\Xi}_E \mathbf{D} \mathbf{\Xi}_E^\top = \mathbf{D}$. To show that $\mathbf{W}$ is an example of the model in part (b2), it remains to show that this matrix $\mathbf{M}$ satisfies the condition (2.9) almost surely for all large n.

Consider $\mathbf{M}^\nu$ for any fixed integer $\nu \geq 1$. Let $\mathbf{h}_i \in \mathbb{R}^n$ denote the $i^t h$ row of $\mathbf{H}$, and let σ be the permutation of $[n]$ for which $P[i, \sigma(i)] = 1$ for all $i \in [n]$. Then

$$M^\nu[i, j] = \big(\mathbf{HPD}^\nu \mathbf{P}^\top \mathbf{H}^\top\big)[i, j] = \sum_{k=1}^n h_i[k] D^\nu\big[\sigma(k), \sigma(k)\big] h_j[k]. \tag{B.7}$$

We condition on $(\mathbf{D}, \mathbf{H})$, and write $\mathbb{E}$ for the expectation over only the permutation σ. Then for each fixed $k \in [n]$, we have $\mathbb{E}[D^\nu[\sigma(k), \sigma(k)]] = n^{-1} \operatorname{Tr} \mathbf{D}^\nu$, so

$$\mathbb{E}\big[M^\nu[i, j]\big] = \frac{1}{n} \operatorname{Tr} \mathbf{D}^\nu \cdot \mathbf{h}_i^\top \mathbf{h}_j = \frac{1}{n} \operatorname{Tr} \mathbf{M}^\nu \cdot 1\{i = j\}.$$

We now show concentration of $M^\nu[i, j]$ around this expectation by computing its high moments: Consider first any fixed $i \neq j \in [n]$, and abbreviate $\tilde{h}[k] = h_i[k] h_j[k]$ and $\tilde{d}[k] = D^\nu[k, k]$. Then from (B.7), $M^\nu[i, j] = \sum_{k=1}^n \tilde{h}[k] \tilde{d}[\sigma(k)]$, so for any even integer $p \geq 2$,

$$\mathbb{E}\big[\big(M^\nu[i, j]\big)^p\big] = \sum_{\mathbf{k} \in [n]^p} \tilde{h}[k_1] \cdots \tilde{h}[k_p] \mathbb{E}\big[\tilde{d}\big[\sigma(k_1)\big] \cdots \tilde{d}\big[\sigma(k_p)\big]\big].$$

Let $\mathcal{P}$ be the lattice of partitions of $[p]$, endowed with the usual partial ordering by refinement. For each $\mathbf{k} \in [n]^p$, let $\pi(\mathbf{k}) \in \mathcal{P}$ be the partition induced by $\mathbf{k}$, that is, $i, j \in [p]$ belong to a common block of π if and only if $k_i = k_j$. Then

$$\begin{aligned} \mathbb{E}\big[\big(M^\nu[i, j]\big)^p\big] &= \sum_{\pi \in \mathcal{P}} \sum_{\mathbf{k} \in [n]^p : \pi(\mathbf{k}) = \pi} \tilde{h}[k_1] \cdots \tilde{h}[k_p] \mathbb{E}\big[\tilde{d}\big[\sigma(k_1)\big] \cdots \tilde{d}\big[\sigma(k_p)\big]\big] \\ &= \sum_{\pi \in \mathcal{P}} \sum_{\substack{\mathbf{k} \in [n]^p \\ \pi(\mathbf{k}) = \pi}} \tilde{h}[k_1] \cdots \tilde{h}[k_p] \cdot \frac{(n - |\pi|)!}{n!} \sum_{\substack{\mathbf{l} \in [n]^p \\ \pi(\mathbf{l}) = \pi}} \tilde{d}[l_1] \cdots \tilde{d}[l_p], \end{aligned} \tag{B.8}$$

the second equality using that the permutation σ is uniformly random, so the expectation over σ yields a uniform average over new choices for the $|\pi|$ distinct index values of $\mathbf{k}$.

Let $\mu(\pi, \pi')$ for $\pi \leq \pi'$ be the Möbius function over $\mathcal{P}$, satisfying the inversion relation (see, e.g., [59], eq. (10.10)) $\sum_{\tau \in \mathcal{P} : \pi \leq \tau \leq \pi'} \mu(\pi, \tau) = \mathbf{1}\{\pi = \pi'\}$. Then for any function f,

$$\sum_{\substack{\mathbf{k} \in [n]^p \\ \pi(\mathbf{k}) = \pi}} f(\mathbf{k}) = \sum_{\substack{\mathbf{k} \in [n]^p \\ \pi(\mathbf{k}) \geq \pi}} f(\mathbf{k}) \cdot \sum_{\substack{\tau \in \mathcal{P} \\ \pi \leq \tau \leq \pi(\mathbf{k})}} \mu(\pi, \tau) = \sum_{\substack{\tau \in \mathcal{P} \\ \tau \geq \pi}} \mu(\pi, \tau) \sum_{\substack{\mathbf{k} \in [n]^p \\ \pi(\mathbf{k}) \geq \tau}} f(\mathbf{k}). \tag{B.9}$$

Applying this to the term involving $\tilde{h}$ in (B.8),

$$\sum_{\mathbf{k} \in [n]^p : \pi(\mathbf{k}) = \pi} \tilde{h}[k_1] \ldots \tilde{h}[k_p] = \sum_{\tau \in \mathcal{P} : \tau \geq \pi} \mu(\pi, \tau) \prod_{R \in \tau} \sum_{k=1}^n \tilde{h}[k]^{|R|}.$$

Recalling $\tilde{h}[k] = h_i[k]h_j[k]$ where $i \neq j$, we have $\sum_{k=1}^n \tilde{h}[k] = \mathbf{h}_i^\top \mathbf{h}_j = 0$. Thus the summand for τ vanishes if τ has a singleton block. For all other partitions $\tau \in \mathcal{P}$, its number of blocks satisfies $|\tau| \le p/2$. Then applying $|\tilde{h}[k]| \le n^{2(-1/2+\varepsilon)}$ by the delocalization condition (2.8) for $\mathbf{H}$, for any fixed $\varepsilon > 0$ and all large n,

$$\left| \prod_{R \in \tau} \sum_{k=1}^n \tilde{h}[k]^{|R|} \right| \le n^{2p(-1/2+\varepsilon)} \cdot n^{|\tau|} \le n^{-p/2+2p\varepsilon}.$$

Thus $|\sum_{\mathbf{k} \in [n]^p : \pi(\mathbf{k}) = \pi} \tilde{h}[k_1] \dots \tilde{h}[k_p]| \le C n^{-p/2+2p\varepsilon}$ where, here and below, we denote by $C > 0$ a (π, D)-dependent constant that may change from instance to instance. By the assumption $\mathbf{d} \overset{W}{\to} D$ and Lemma 3.3 (applied with $\mathcal{S}$ being the blocks of π and $q_S(x) = \tilde{d}(x)^{|S|}$ for $S \in \pi$), also $n^{-|\pi|} |\sum_{\mathbf{l} \in [n]^p : \pi(\mathbf{l}) = \pi} \tilde{d}[l_1] \dots \tilde{d}[l_p]| \le C$. Applying these to (B.8), we obtain $\mathbb{E}[(M^\nu[i,j])^p] \le C n^{-p/2+2p\varepsilon}$, so $\mathbb{P}[|M^\nu[i,j]| > n^{-1/2+3\varepsilon}] \le C n^{-p\varepsilon}$ by Markov's inequality. Choosing even $p \ge 2$ sufficiently large and taking a union bound over all $i \neq j$, this shows that the second condition of (2.9) holds almost surely for all large n.

The case $i = j$ is similar: Fix $i \in [n]$ and now abbreviate $\tilde{h}[k] = h_i[k]^2$ and $\tilde{d}[k] = D^\nu[k,k] - n^{-1} \operatorname{Tr} \mathbf{D}^\nu$. Then from (B.7), $M^\nu[i,i] - n^{-1} \operatorname{Tr} \mathbf{M}^\nu = \sum_k \tilde{h}[k]\tilde{d}[k]$, so we obtain analogously to (B.8)

$$\mathbb{E}\big[(M^\nu[i,i] - n^{-1} \operatorname{Tr} \mathbf{M}^\nu)^p\big] = \sum_{\substack{\mathbf{k} \in [n]^p \\ \pi(\mathbf{k}) = \pi}} \tilde{h}[k_1] \cdots \tilde{h}[k_p] \cdot \frac{(n - |\pi|)!}{n!} \sum_{\substack{\mathbf{l} \in [n]^p \\ \pi(\mathbf{l}) = \pi}} \tilde{d}[l_1] \cdots \tilde{d}[l_p].$$

Applying the Möbius inversion relation (B.9) now to the second summation over $\mathbf{l}$,

$$\sum_{\mathbf{l} \in [n]^p : \pi(\mathbf{l}) = \pi} \tilde{d}[l_1] \cdots \tilde{d}[l_p] = \sum_{\tau \in \mathcal{P} : \tau \ge \pi} \mu(\pi, \tau) \prod_{R \in \tau} \sum_{l=1}^n \tilde{d}[l]^{|R|}.$$

Using that $\sum_{k=1}^n \tilde{d}[k] = 0$, the summand for τ vanishes if τ has a singleton block. For all other partitions $\tau \in \mathcal{P}$, applying $\mathbf{d} \overset{W}{\to} D$, we obtain

$$\left| \prod_{R \in \tau} \sum_{l=1}^n \tilde{d}[l]^{|R|} \right| \le C n^{|\tau|} \le C n^{p/2}.$$

Then $|\sum_{\mathbf{l} \in [n]^p : \pi(\mathbf{l}) = \pi} \tilde{d}[l_1] \dots \tilde{d}[l_p]| \le C n^{p/2}$. From (2.8), we have also

$$n^{-|\pi|} \sum_{\mathbf{k} \in [n]^p : \pi(\mathbf{k}) = \pi} |\tilde{h}[k_1] \cdots \tilde{h}[k_p]| \le n^{-2p(1/2+\varepsilon)}.$$

Then $\mathbb{E}[(M^\nu[i,i] - n^{-1} \operatorname{Tr} \mathbf{M}^\nu)^p] \le C n^{-p/2+2p\varepsilon}$, so the first condition of (2.9) follows also by Markov's inequality and a union bound. This verifies that $\mathbf{W}$ satisfying part (b1) also satisfies part (b2), as desired. □

APPENDIX C: TENSOR NETWORK VALUE UNDER ORTHOGONAL INVARIANCE

In this Appendix, we derive a more explicit combinatorial form for the tensor network value of Lemma 2.14 when $\mathbf{W}$ is an orthogonally invariant matrix, using the orthogonal Weingarten calculus. We then prove the asymptotic freeness statement of Proposition 2.16(b).

Let T be a tensor network with $w+1$ vertices and w edges. Then there are $2w$ vertex-edge pairs (v, e) where edge e is incident to vertex v. We label these vertex-edge pairs arbitrarily as $1, 2, \dots, 2w$. Let $\mathcal{P}$ be the lattice of partitions of $[2w]$, endowed with the usual partial ordering by refinement. We define two distinguished partitions $\pi_V, \pi_E \in \mathcal{P}$, such that vertex-edge pairs $\rho, \tau \in [2w]$ belong to the same block of π_V if and only if they have the same

vertex v, and to the same block of π_E if and only if they have the same edge e. (Thus π_V has $w+1$ blocks, one for each vertex of T, and π_E is a pairing with w pairs, one for each edge of T.)

Define a metric over $\mathcal{P}$ by

$$d(\pi,\pi') = |\pi| + |\pi'| - 2|\pi \vee \pi'|, \tag{C.1}$$

where $\pi \vee \pi'$ is the join (i.e., least upper bound) of π and π'. This is shown in [3, 14] to be equivalent to the smallest number of merge and divide operations needed to transform π into π', where a merge operation combines any two blocks into one block, and a divide operation splits any one block into two blocks. From this characterization, it is immediate that $d(\cdot,\cdot)$ satisfies the triangle inequality $d(\pi,\pi') + d(\pi',\pi'') \geq d(\pi,\pi'')$. We call a path $\pi_0 \to \pi_1 \to \cdots \to \pi_k$ of partitions a *d-geodesic* if it is a shortest path from π_0 to π_k in the metric $d(\cdot,\cdot)$, that is, if

$$d(\pi_0,\pi_k) = d(\pi_0,\pi_1) + d(\pi_1,\pi_2) + \cdots + d(\pi_{k-1},\pi_k).$$

The main result of this Appendix is the following proposition.

PROPOSITION C.1. *In the setting of Lemma 2.14, suppose in addition that $\mathbf{W} = \mathbf{ODO}^\top$ is orthogonally invariant, where $\mathbf{D} = \operatorname{diag}(\mathbf{d})$ and $\mathbf{d} \overset{W}{\to} D$ almost surely as $n \to \infty$. For $\pi \geq \pi_V$ and $\pi' \geq \pi_E$, define*

$$q(\pi) = \prod_{S \in \pi} \mathbb{E}\Bigg[\prod_{\text{distinct vertices } v \text{ in vertex-edge pairs of } S} q_v(X_1,\dots,X_k) \Bigg], \tag{C.2}$$

$$D(\pi') = \prod_{S \in \pi'} \mathbb{E}\big[D^{\text{number of distinct edges in vertex-edge pairs of } S} \big]. \tag{C.3}$$

Then

$$\lim_{n\to\infty} \operatorname{val}_T(\mathbf{W}; \mathbf{x}_1,\dots,\mathbf{x}_k) = \sum_{j \geq 0} \sum_{\substack{\text{distinct pairings } \pi_0,\dots,\pi_j \text{ of } [2w] \\ \pi_V \to \pi_0 \to \cdots \to \pi_j \to \pi_E \text{ is a } d\text{-geodesic}}} (-1)^j q(\pi_V \vee \pi_0) D(\pi_E \vee \pi_j).$$

(*Here π_j is not required to be distinct from π_E.*)

To show this result, we apply the following statements derived from the orthogonal Weingarten calculus of [23] for mixed moments of entries of Haar-orthogonal random matrices.

LEMMA C.2. *Let $\mathbf{O} \sim \operatorname{Haar}(\mathbb{O}(n))$. Let $\mathbf{i} = (i_1,\dots,i_{2w})$ and $\mathbf{j} = (j_1,\dots,j_{2w})$ be any index tuples in $[n]^{2w}$. Then*

$$\mathbb{E}\Bigg[\prod_{p=1}^{2w} O[i_p, j_p] \Bigg] = \sum_{\substack{\text{pairings } \pi,\pi' \text{ of } [2w] \\ \pi \leq \pi(\mathbf{i}), \pi' \leq \pi(\mathbf{j})}} \operatorname{Wg}_n[\pi,\pi'], \tag{C.4}$$

where Wg_n is the orthogonal Weingarten function. For fixed w, as $n \to \infty$, this satisfies

$$\operatorname{Wg}_n[\pi,\pi'] = n^{-w-d(\pi,\pi')/2} \cdot \mu_{\mathrm{NC}}(\pi,\pi') + O\big(n^{-w-d(\pi,\pi')/2-1}\big), \tag{C.5}$$

where $d(\pi,\pi')$ is the metric (C.1), and $\mu_{\mathrm{NC}}(\pi,\pi')$ is the Möbius function on the noncrossing partition lattice, given by

$$\mu_{\mathrm{NC}}(\pi,\pi') = \sum_{k \geq 0} \sum_{\substack{\text{distinct pairings } \pi_0,\pi_1,\dots,\pi_k \text{ of } [2w] \\ \pi_0 \to \pi_1 \to \cdots \to \pi_k \text{ is a } d\text{-geodesic from } \pi_0 = \pi \text{ to } \pi_k = \pi'}} (-1)^k. \tag{C.6}$$

PROOF. We may identify pairings π, π' of $[2w]$ as permutations in the symmetric group S_{2w}, each a product of w disjoint transpositions corresponding to the w pairs. The cycle decomposition of their product $\pi\pi'$ in S_{2w} has exactly two cycles for each set of their join partition $\pi \vee \pi'$. Then, the metric $l(\pi, \pi') = |\pi\pi'|/2$ used in [23], Section 3, (where $|\cdot|$ is the Cayley distance to the identity permutation in S_{2w}, given by $2w$ minus the number of cycles) is equivalently

$$l(\pi, \pi') = \frac{2w - 2|\pi \vee \pi'|}{2} = \frac{d(\pi, \pi')}{2}, \tag{C.7}$$

where the right side is our metric $d(\cdot, \cdot)$ restricted to pairings. The statements (C.4) and (C.5) then follow from [23], Corollary 3.4 and Theorem 3.13. The form (C.6) for the Möbius function follows from comparing [23], Theorem 3.13, with [23], Lemma 3.12, noting that the leading-order terms of [23], Lemma 3.12, come from paths of pairings satisfying $\pi_i \neq \pi_{i+1}$ for each $i = 0, \dots, k-1$ and also $l(\pi_0, \pi_1) + \cdots + l(\pi_{k-1}, \pi_k) = l(\pi_0, \pi_k)$. Any such path must be a geodesic of $k+1$ unique pairings in the metric $l(\cdot, \cdot)$, and hence also in the metric $d(\cdot, \cdot)$ by the equivalence (C.7), and this shows (C.6). □

PROOF OF PROPOSITION C.1. Expanding the product $\mathbf{W} = \mathbf{ODO}^\top$, the tensor network value is given by

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \frac{1}{n} \sum_{\mathbf{i} \in [n]^{\mathcal{V}}} \sum_{\mathbf{j} \in [n]^{\mathcal{E}}} \prod_{v \in \mathcal{V}} q_v\big(x_{1:k}[i_v]\big) \prod_{e=(u,v) \in \mathcal{E}} O[i_u, j_e] D[j_e, j_e] O[i_v, j_e].$$

For each vertex v or edge e, let $\rho(v), \rho(e) \in [2w]$ be an arbitrary choice of vertex-edge pair containing this vertex or this edge. Then this is equivalently expressed as

$$\mathrm{val}_T(\mathbf{W}; \mathbf{x}_{1:k}) = \frac{1}{n} \sum_{\substack{\mathbf{i} \in [n]^{2w} \\ \pi(\mathbf{i}) \geq \pi_V}} \sum_{\substack{\mathbf{j} \in [n]^{2w} \\ \pi(\mathbf{j}) \geq \pi_E}} \prod_{v \in \mathcal{V}} q_v\big(x_{1:k}[i_{\rho(v)}]\big) \prod_{e \in \mathcal{E}} D[j_{\rho(e)}, j_{\rho(e)}] \prod_{\rho=1}^{2w} O[i_\rho, j_\rho]. \tag{C.8}$$

Note that by the constraints $\pi(\mathbf{i}) \geq \pi_V$ and $\pi(\mathbf{j}) \geq \pi_E$, this expression is the same for any choices of vertex-edge pairs $\rho(v), \rho(e) \in [2w]$.

Let $\mathbb{E}$ be the expectation over $\mathbf{O}$, conditional on $\mathbf{x}_1, \dots, \mathbf{x}_k$ and $\mathbf{D}$. By Lemma C.2, we have

$$\mathbb{E}\Bigg[\prod_{p=1}^{2w} O[i_p, j_p]\Bigg] = \sum_{\text{pairings } \pi, \pi' \text{ of } [2w]} \mathbf{1}_{\pi(\mathbf{i}) \geq \pi} \mathbf{1}_{\pi(\mathbf{j}) \geq \pi'} \cdot n^{-w - d(\pi, \pi')} \big(\mu_{\mathrm{NC}}(\pi, \pi') + o(1)\big).$$

Note that

$$\mathbf{1}_{\pi(\mathbf{i}) \geq \pi_V} \mathbf{1}_{\pi(\mathbf{i}) \geq \pi} = \mathbf{1}_{\pi(\mathbf{i}) \geq \pi_V \vee \pi}, \qquad \mathbf{1}_{\pi(\mathbf{j}) \geq \pi_E} \mathbf{1}_{\pi(\mathbf{j}) \geq \pi'} = \mathbf{1}_{\pi(\mathbf{j}) \geq \pi_E \vee \pi'}.$$

Identifying summations over $\mathbf{i}, \mathbf{j} \in [n]^{2w}$ with $\pi(\mathbf{i}) \geq \pi_V \vee \pi$ and $\pi(\mathbf{j}) \geq \pi_E \vee \pi'$ as a summation over one index in $[n]$ for each block of $\pi_V \vee \pi$ and $\pi_E \vee \pi'$, and applying the given conditions that $\mathbf{x}_{1:k} \xrightarrow{W} X_{1:k}$ and $\mathrm{diag}(\mathbf{D}) \xrightarrow{W} D$ almost surely, observe that

$$\frac{1}{n^{|\pi_V \vee \pi|}} \sum_{\mathbf{i} \in [n]^{2w}} \mathbf{1}_{\pi(\mathbf{i}) \geq \pi_V \vee \pi} \prod_{v \in \mathcal{V}} q_v\big(x_{1:k}[i_{\rho(v)}]\big) \to q(\pi_V \vee \pi),$$

$$\frac{1}{n^{|\pi_E \vee \pi|}} \sum_{\mathbf{j} \in [n]^{2w}} \mathbf{1}_{\pi(\mathbf{j}) \geq \pi_E \vee \pi'} \prod_{e \in \mathcal{E}} D[j_{\rho(e)}, j_{\rho(e)}] \to D\big(\pi_E \vee \pi'\big),$$

where $q(\cdot)$ and $D(\cdot)$ are as defined in (C.2) and (C.3). Then, taking the expectation over $\mathbf{O}$ in (C.8) and applying these observations,

$$\begin{aligned}\mathbb{E}\big[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})\big] = {} & \sum_{\text{pairings } \pi,\pi' \text{ of } [2w]} \frac{1}{n}\cdot n^{|\pi_V\vee\pi|}\cdot n^{|\pi_E\vee\pi'|}\cdot n^{-w-d(\pi,\pi')} \\ & \cdot\big(\mu_{\mathrm{NC}}(\pi,\pi')\cdot q(\pi_V\vee\pi)\cdot D\big(\pi_E\vee\pi'\big)+o(1)\big).\end{aligned} \tag{C.9}$$

Recall that $|\pi_V| = w+1$ and $|\pi| = |\pi'| = |\pi_E| = w$ as these are all pairings of $[2w]$. Then by definition of the metric $d(\cdot,\cdot)$,

$$|\pi_V\vee\pi| = \frac{2w+1-d(\pi_V,\pi)}{2}, \qquad |\pi_E\vee\pi| = \frac{2w-d(\pi_E,\pi)}{2}.$$

So the above value simplifies to

$$\sum_{\text{pairings } \pi,\pi' \text{ of } [2w]} n^{\frac{2w-1-d(\pi_V,\pi)-d(\pi,\pi')-d(\pi',\pi_E)}{2}}\big(\mu_{\mathrm{NC}}(\pi,\pi')q(\pi_V\vee\pi)D\big(\pi_E\vee\pi'\big)+o(1)\big).$$

Applying the triangle inequality for $d(\cdot,\cdot)$ and the identity $|\pi_V\vee\pi_E| = 1$ since T is a connected tree, we have

$$d(\pi_V,\pi)+d\big(\pi,\pi'\big)+d\big(\pi',\pi_E\big)\geq d(\pi_V,\pi_E) = (w+1)+w-2 = 2w-1,$$

and equality holds if and only if $\pi_V\to\pi\to\pi'\to\pi_E$ is a d-geodesic. Thus, we obtain the limit value

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{E}\big[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})\big] \\ &= \sum_{\substack{\text{pairings } \pi,\pi' \text{ of } [2w]\\ \pi_V\to\pi\to\pi'\to\pi_E \text{ is a } d\text{-geodesic}}} \mu_{\mathrm{NC}}\big(\pi,\pi'\big)q(\pi_V\vee\pi)D\big(\pi_E\vee\pi'\big).\end{aligned} \tag{C.10}$$

Here, π and π' are pairings of $[2w]$ that may coincide with each other and/or with π_E.

Finally, we apply (C.6) to express $\mu_{\mathrm{NC}}(\pi,\pi')$ also as a summation over geodesic paths of pairings from π to π', giving

$$\lim_{n\to\infty}\mathbb{E}\big[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})\big] = \sum_{j\geq 0}\sum_{\substack{\text{distinct pairings } \pi_0,\ldots,\pi_j \text{ of } [2w]\\ \pi_V\to\pi_0\to\cdots\to\pi_j\to\pi_E \text{ is a } d\text{-geodesic}}} (-1)^j q(\pi_V\vee\pi_0)D(\pi_E\vee\pi_j).$$

We have set $\pi_0 = \pi$ and $\pi_j = \pi'$, and the terms of the sum with $j=0$ correspond to $\pi=\pi'$. This shows that the stated form is the almost-sure limit of $\mathbb{E}[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})]$ where $\mathbb{E}$ is the expectation over $\mathbf{O}$. Comparing with the result of Lemma 2.14, we conclude that this must be $\mathrm{lim\text{-}val}_T(X_{1:k},\mathcal{D}_{\mathrm{diag}})$. □

PROOF OF PROPOSITION 2.16(b). By the universality established in Lemma 2.14, it suffices to check that the limit of $\mathbb{E}[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})]$ for orthogonally invariant matrices $\mathbf{W}$, as computed in the preceding Proposition C.1, equals 0 under the given conditions.

The given condition $\frac{1}{n}\operatorname{Tr}\mathbf{W}\to 0$ implies $\mathbb{E}[D]=0$. If $\pi_E\vee\pi'$ has any block containing only the two vertex-edge pairs for a single edge, then this implies $D(\pi_E\vee\pi')=0$ in (C.3). Otherwise, each block must correspond to at least two edges, so $|\pi_E\vee\pi'|\leq w/2$. Similarly, if $\pi_V\vee\pi$ is such that any block contains the vertex-edge pairs for only a single vertex, then the condition (2.13) implies $q(\pi)=0$ in (C.2). Otherwise,each block must correspond to at least two vertices, so $|\pi_V\vee\pi|\leq(w+1)/2$. Thus if $q(\pi_V\vee\pi)D(\pi_E\vee\pi')\neq 0$, then

$$\frac{1}{n}\cdot n^{|\pi_V\vee\pi|}\cdot n^{|\pi_E\vee\pi'|}\cdot n^{-w-d(\pi,\pi')}\leq n^{-1+(w+1)/2+w/2-w}\leq n^{-1/2}.$$

Applying this to (C.9), we get $\mathbb{E}[\mathrm{val}_T(\mathbf{W};\mathbf{x}_{1:k})]\to 0$ as desired. □

Acknowledgments. We would like to thank Mark Sellke for asking us an interesting question about asymptotic freeness for random tensors, and Zhigang Bao and Yuxin Chen for asking about AMP algorithms for heteroskedastic variances, which motivated parts of this work.

Funding. XZ was supported in part by funding from Two Sigma Investments, LP. ZF was supported in part by NSF Grants DMS-1916198 and DMS-2142476.

SUPPLEMENTARY MATERIAL

Supplementary Appendix (DOI: 10.1214/24-AAP2056SUPP; .pdf). The supplementary appendix contains additional details about AMP algorithms for rectangular matrices and the rectangular generalized invariant universality class of Definition 2.20, and proofs of Theorems 2.21 and 2.22 on universality of AMP for rectangular matrices.

REFERENCES

[1] ANDERSON, G. W. and FARRELL, B. (2014). Asymptotically liberating sequences of random unitary matrices. *Adv. Math.* **255** 381–413. MR3167487 https://doi.org/10.1016/j.aim.2013.12.026

[2] ANDERSON, G. W. and ZEITOUNI, O. (2006). A CLT for a band matrix model. *Probab. Theory Related Fields* **134** 283–338. MR2222385 https://doi.org/10.1007/s00440-004-0422-3

[3] ARABIE, P. and BOORMAN, S. A. (1973). Multidimensional scaling of measures of distance between partitions. *J. Math. Psych.* **10** 148–203. MR0321559 https://doi.org/10.1016/0022-2496(73)90012-6

[4] AU, B., CÉBRON, G., DAHLQVIST, A., GABRIEL, F. and MALE, C. (2021). Freeness over the diagonal for large random matrices. *Ann. Probab.* **49** 157–179. MR4203335 https://doi.org/10.1214/20-AOP1447

[5] BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Proceedings of the* 30*th International Conference on Neural Information Processing Systems. NIPS'*16 424–432. Curran Associates Inc., Red Hook, NY, USA.

[6] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[7] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[8] BENAYCH-GEORGES, F., BORDENAVE, C. and KNOWLES, A. (2020). Spectral radii of sparse random matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** 2141–2161. MR4116720 https://doi.org/10.1214/19-AIHP1033

[9] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[10] BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR1324786

[11] BOLTHAUSEN, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.* **325** 333–366. MR3147441 https://doi.org/10.1007/s00220-013-1862-3

[12] BOLTHAUSEN, E. (2019). A Morita type proof of the replica-symmetric formula for SK. In *Statistical Mechanics of Classical and Disordered Systems. Springer Proc. Math. Stat.* **293** 63–93. Springer, Cham. MR4015008 https://doi.org/10.1007/978-3-030-29077-1_4

[13] BOLTHAUSEN, E., NAKAJIMA, S., SUN, N. and XU, C. (2021). Gardner formula for Ising perceptron models at small densities. Preprint. Available at arXiv:2111.02855.

[14] BOORMAN, S. A. and OLIVIER, D. C. (1973). Metrics on spaces of finite trees. *J. Math. Psych.* **10** 26–59. MR0317975 https://doi.org/10.1016/0022-2496(73)90003-5

[15] BU, Z., KLUSOWSKI, J. M., RUSH, C. and SU, W. J. (2021). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Trans. Inf. Theory* **67** 506–537. MR4231969 https://doi.org/10.1109/TIT.2020.3025272

[16] CADEMARTORI, C. and RUSH, C. (2023). A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. Preprint. Available at arXiv:2302.00088.

[17] ÇAKMAK, B. and OPPER, M. (2019). Memory-free dynamics for the Thouless–Anderson–Palmer equations of Ising models with arbitrary rotation-invariant ensembles of random coupling matrices. *Phys. Rev. E* **99** 062140, 14 pp. MR3984544

[18] ÇAKMAK, B. and OPPER, M. (2020). A dynamical mean-field theory for learning in restricted Boltzmann machines. *J. Stat. Mech. Theory Exp.* **10** 103303, 32 pp. MR4197533 https://doi.org/10.1088/1742-5468/abb8c9

[19] CAKMAK, B., WINTHER, O. and FLEURY, B. H. (2014). S-AMP: Approximate message passing for general matrix ensembles. In 2014 *IEEE Information Theory Workshop* (*ITW* 2014) 192–196. IEEE.

[20] CELENTANO, M., CHENG, C. and MONTANARI, A. (2021). The high-dimensional asymptotics of first order methods with random data. Preprint. Available at arXiv:2112.07572.

[21] CELENTANO, M., MONTANARI, A. and WU, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory* 1078–1141. PMLR.

[22] CHEN, W.-K. and LAM, W.-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.* **26** Paper No. 36, 44 pp. MR4235487 https://doi.org/10.1214/21-EJP604

[23] COLLINS, B. and ŚNIADY, P. (2006). Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Comm. Math. Phys.* **264** 773–795. MR2217291 https://doi.org/10.1007/s00220-006-1554-3

[24] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

[25] DING, J. and SUN, N. (2019). Capacity lower bound for the Ising perceptron. In *STOC'*19*—Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 816–827. ACM, New York. MR4003386 https://doi.org/10.1145/3313276.3316383

[26] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[27] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388 https://doi.org/10.1098/rsta.2009.0152

[28] DONOHO, D. L., JAVANMARD, A. and MONTANARI, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inf. Theory* **59** 7434–7464. MR3124654 https://doi.org/10.1109/TIT.2013.2274513

[29] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* **106** 18914–18919.

[30] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2010). Message passing algorithms for compressed sensing: I. Motivation and construction. In 2010 *IEEE Information Theory Workshop on Information Theory* (*ITW* 2010, *Cairo*) 1–5. IEEE.

[31] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716 https://doi.org/10.1073/pnas.0502258102

[32] DUDEJA, R. and BAKHSHIZADEH, M. (2022). Universality of linearized message passing for phase retrieval with structured sensing matrices. *IEEE Trans. Inf. Theory* **68** 7545–7574. MR4524656

[33] DUDEJA, R., LU, Y. M. and SEN, S. (2022). Universality of Approximate message passing with semi-random matrices. Preprint. Available at arXiv:2204.04281.

[34] DUDEJA, R., SEN, S. and LU, Y. M. (2022). Spectral universality of regularized linear regression with nearly deterministic sensing matrices. Preprint. Available at arXiv:2208.02753.

[35] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Bulk universality for generalized Wigner matrices. *Probab. Theory Related Fields* **154** 341–407. MR2981427 https://doi.org/10.1007/s00440-011-0390-3

[36] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147 https://doi.org/10.1016/j.aim.2011.12.010

[37] FAN, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Ann. Statist.* **50** 197–224. MR4382014 https://doi.org/10.1214/21-aos2101

[38] FAN, Z., LI, Y. and SEN, S. (2022). TAP equations for orthogonally invariant spin glasses at high temperature. Preprint. Available at arXiv:2202.09325.

[39] FAN, Z. and WU, Y. (2021). The replica-symmetric free energy for Ising spin glasses with orthogonally invariant couplings. Preprint. Available at arXiv:2105.02797.

[40] FENG, O. Y., VENKATARAMANAN, R., RUSH, C. and SAMWORTH, R. J. (2022). A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15** 335–536.

[41] GERBELOT, C. and BERTHIER, R. (2023). Graph-based approximate message passing iterations. *Inf. Inference* **12** Paper No. iaad020, 67 pp. MR4644961 https://doi.org/10.1093/imaiai/iaad020

[42] HAFEMEISTER, C. and SATIJA, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. **20** 1–15.

[43] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference* **2** 115–144. MR3311445 https://doi.org/10.1093/imaiai/iat004

[44] JIANG, T. (2005). Maxima of entries of Haar distributed matrices. *Probab. Theory Related Fields* **131** 121–144. MR2105046 https://doi.org/10.1007/s00440-004-0376-5

[45] KABASHIMA, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *J. Phys. A* **36** 11111–11121. MR2025247 https://doi.org/10.1088/0305-4470/36/43/030

[46] LANDA, B., ZHANG, T. T. C. K. and KLUGER, Y. (2022). Biwhitening reveals the rank of a count matrix. *SIAM J. Math. Data Sci*. **4** 1420–1446. MR4522878 https://doi.org/10.1137/21M1456807

[47] LI, G., FAN, W. and WEI, Y. (2023). Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302930120, 7 pp. MR4637851

[48] LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. Preprint. Available at arXiv:2208.03313.

[49] LI, Y. and WEI, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. Preprint. Available at arXiv:2110.09502.

[50] LIU, L., HUANG, S. and KURKOSKI, B. M. (2021). Memory approximate message passing. In 2021 *IEEE International Symposium on Information Theory* (*ISIT*) 1379–1384. IEEE.

[51] MA, J. and PING, L. (2017). Orthogonal AMP. *IEEE Access* **5** 2020–2033.

[52] MALE, C. (2020). Traffic distributions and independence: Permutation invariant random matrices and the three notions of independence. *Mem. Amer. Math. Soc*. **267** v+88. MR4197072 https://doi.org/10.1090/memo/1300

[53] MINGO, J. A. and SPEICHER, R. (2012). Sharp bounds for sums associated to graphs of matrices. *J. Funct. Anal*. **262** 2272–2288. MR2876405 https://doi.org/10.1016/j.jfa.2011.12.010

[54] MINGO, J. A. and SPEICHER, R. (2017). *Free Probability and Random Matrices. Fields Institute Monographs* **35**. Springer, New York. MR3585560 https://doi.org/10.1007/978-1-4939-6942-5

[55] MONAJEMI, H., JAFARPOUR, S., GAVISH, M., COLLABORATION, S. C. . and DONOHO, D. L. (2013). Deterministic matrices matching the compressed sensing phase transitions of Gaussian random matrices. *Proc. Natl. Acad. Sci. USA* **110** 1181–1186. MR3037097 https://doi.org/10.1073/pnas.1219540110

[56] MONDELLI, M. and VENKATARAMANAN, R. (2021). PCA initialization for approximate message passing in rotationally invariant models. *Adv. Neural Inf. Process. Syst*. **34** 29616–29629.

[57] MONTANARI, A. (2019). Optimization of the Sherrington–Kirkpatrick Hamiltonian. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 1417–1433. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228234 https://doi.org/10.1109/FOCS.2019.00087

[58] MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist*. **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

[59] NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability. London Mathematical Society Lecture Note Series* **335**. Cambridge Univ. Press, Cambridge. MR2266879 https://doi.org/10.1017/CBO9780511735127

[60] OPPER, M., ÇAKMAK, B. and WINTHER, O. (2016). A theory of solving TAP equations for Ising models with general invariant random matrices. *J. Phys. A* **49** 114002, 24 pp. MR3462332 https://doi.org/10.1088/1751-8113/49/11/114002

[61] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In 2011 *IEEE International Symposium on Information Theory Proceedings* 2168–2172. IEEE.

[62] RANGAN, S. and FLETCHER, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In 2012 *IEEE International Symposium on Information Theory Proceedings* 1246–1250. IEEE.

[63] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2019). Vector approximate message passing. *IEEE Trans. Inf. Theory* **65** 6664–6684. MR4009222 https://doi.org/10.1109/TIT.2019.2916359

[64] RUSH, C. and VENKATARAMANAN, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Trans. Inf. Theory* **64** 7264–7286. MR3876443 https://doi.org/10.1109/TIT.2018.2816681

[65] SARKAR, A. and STEPHENS, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet*. **53** 770–777. https://doi.org/10.1038/s41588-021-00873-4

[66] SCHMÜDGEN, K. (2017). *The Moment Problem. Graduate Texts in Mathematics* **277**. Springer, Cham. MR3729411

[67] SCHNITER, P., RANGAN, S. and FLETCHER, A. K. (2016). Vector approximate message passing for the generalized linear model. In 2016 50*th Asilomar Conference on Signals*, *Systems and Computers* 1525–1529. IEEE.

[68] SUR, P., CHEN, Y. and CANDÈS, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Related Fields* **175** 487–558. MR4009715 https://doi.org/10.1007/s00440-018-00896-9

[69] TAKEUCHI, K. (2017). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 501–505. IEEE.

[70] TAKEUCHI, K. (2020). Convolutional approximate message-passing. *IEEE Signal Process. Lett.* **27** 416–420.

[71] TAKEUCHI, K. (2021). Bayes-optimal convolutional AMP. *IEEE Trans. Inf. Theory* **67** 4405–4428. MR4306276 https://doi.org/10.1109/TIT.2021.3077471

[72] TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20** 1–16.

[73] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[74] VOICULESCU, D. V., DYKEMA, K. J. and NICA, A. (1992). *Free Random Variables*: *A Noncommutative Probability Approach to Free Products with Applications to Random Matrices*, *Operator Algebras and Harmonic Analysis on Free Groups*. *CRM Monograph Series* **1**. Amer. Math. Soc., Providence, RI. MR1217253 https://doi.org/10.1090/crmm/001

[75] WANG, T., ZHONG, X. and FAN, Z. (2024). Supplement to "Universality of approximate message passing algorithms and tensor networks." https://doi.org/10.1214/24-AAP2056SUPP

[76] ZHONG, X., WANG, T. and FAN, Z. (2021). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. Preprint. Available at arXiv:2110.02318.