Electron. J. Probab. **26** (2021), article no. 150, 1–37.
ISSN: 1083-6489 https://doi.org/10.1214/21-EJP699

Eigenvalue distribution of some nonlinear models of random matrices*

Lucas Benigni† Sandrine Péché‡

Abstract

This paper is concerned with the asymptotic empirical eigenvalue distribution of some non linear random matrix ensemble. More precisely we consider $M = \frac{1}{m}YY^*$ with $Y = f(WX)$ where W and X are random rectangular matrices with i.i.d. centered entries. The function f is applied pointwise and can be seen as an activation function in (random) neural networks. We compute the asymptotic empirical distribution of this ensemble in the case where W and X have sub-Gaussian tails and f is real analytic. This extends a result of [32] where the case of Gaussian matrices W and X is considered. We also investigate the same questions in the multi-layer case, regarding neural network and machine learning applications.

Keywords: random matrices; machine learning; neural networks.
MSC2020 subject classifications: 15B52; 62M45.
Submitted to EJP on April 13, 2021, final version accepted on September 3, 2021.
Supersedes arXiv:1904.03090v3.

1 Introduction

In this article, we are interested in the asymptotic spectral properties of some Gram matrices whose definition comes from machine learning. Machine learning has shined through a large list of successful applications over the past five years or so (see for instance applications in image or speech recognition [25, 23] or translation [37]) but is also used now in video, style transfer, dialogues, games and countless other topics. The interested reader can go to [35] for an overview of the subject. However, a complete theoretical and mathematical understanding of learning is still missing. The main difficulty comes from the complexity of studying non-convex functions of a very large number of parameters [9, 31]. We also refer to [10] for a comprehensive exposition of the problem.

An artificial neural network can be modeled as follows: some input column vector $x \in \mathbb{R}^{n_0}$ goes through a multistage architecture of alternated layers with both linear and

*Research was accomplished while Sandrine Péché was supported by the Institut Universitaire de France.
†University of Chicago, United States of America. E-mail: lbenigni@uchicago.edu
‡Université de Paris, France. E-mail: peche@lpsm.paris

non linear functionals: let $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \dots, L$ be some given *activation functions* and $W_i, i = 1 \dots L$ be $n_i \times n_{i-1}$ matrices. The output vector after layer L is s_L where

$$s_1 = g_1(W_1 x), \quad s_i = g_i(W_i s_{i-1}), i = 2, \dots, L. \tag{1.1}$$

The functions g_i are here applied componentwise. The matrices W_i are the (synaptic) weights in the layer i and the activation function g_i models the impact of the neurons in the architecture. There are different possible choices for the activation functions: some notable examples are $g(x) = \max(0, x)$ (known as the ReLU activation function for Rectified Linear Unit) or the sigmoid function $g(x) = (1 + e^{-x})^{-1}$. The parameter L is called the depth of the neural network.

In supervised machine learning, one is given a $n_0 \times m$ matrix dataset X coinjointly with a target dataset Z of size $d \times m$. Here m is the sample size and the parameters to be learned are the weight matrices. A possible idea to understand better such large complex systems is to approximate the elements of the system by random variables: the weights are random. This is the place where random matrix theory can bring its techniques in principle. The aim of the so-called training phase (with X and Z) is to determine a function h so that, given a *new* photo x, the output of the function $h(x)$ yields an acceptable approximation of the target (true) object. The error when performing such an approximation is measured through a loss function. In the context of Feed Forward Neural Networks as in (1.1), when the input vector is high dimensional and the sample size is comparably large, one of the commonly used learning method is ridge regression (and h is linear in $Y = g_1(WX)$). More precisely, in the one layer case ($L = 1$) the loss function is

$$B \in \mathbb{R}^{d \times n_1} \mapsto \mathcal{L}(B) := \frac{1}{2dm} ||Z - B^*(g_1(W_1 X))||_F^2 + \gamma ||B||_F^2,$$

where γ is the penalizing parameter. The optimal matrix B can then be proved to be proportional to YQZ^* where

$$Q = \left(\frac{1}{m} Y^* Y + \gamma I \right)^{-1}. \tag{1.2}$$

As a consequence, the performance of such a learning procedure can be measured thanks to the asymptotic spectral properties of the matrix $\frac{1}{m} Y^* Y$. Indeed, for the one layer case, the expected training loss can be proved to be related to its asymptotic e.e.d. (and Stieltjes transform). It is given by $\mathbb{E}(\mathcal{L}(B)) = -\frac{\gamma^2}{m} \frac{\partial}{\partial \gamma} \mathbb{E}(\mathrm{Tr}\, Q)$, where Q is given by (1.2) and Tr denotes the unnormalized trace. The case where $L = 1$, which is a particular model of interest in this paper, is rather known as extreme learning. One may also refer to [27] for more information on the development of deep learning ($L > 1$).

Random matrix theory has already been proved to be useful in machine learning. In [19] for instance, neural networks with random Gaussian weights have been studied for practical interest while eigenvalues of non-Hermitian matrices were used to understand neural networks in [34]. See also [38] who study echo state networks used to model nonlinear dynamical systems. In [5, 12], a random matrix approach has been used to study spectral clustering by considering the random Gram matrix WW^*. They compute the asymptotic deterministic empirical distribution (e.e.d.) of this matrix which allows the analysis of the spectral clustering algorithm in large dimensions. The e.e.d. of nonlinear random matrix models of the form $f(XX^*)$ have also been studied in [14] and [8] with different variance scalings for the entries of X. We also mention [15] where the question is further studied including the behavior of extreme eigenvalues, with a view to statistical estimation of the covariance of the population.

We are here interested in random neural networks where both the number of samples m and the number of parameters n_0 are large. We consider rectangular matrices of size $n_0 \times m$ in the regime where n_0/m goes to some constant ϕ as the dimension grows to infinity. The study of such matrix models for random neural networks was first accomplished in [29, 32], where they consider

$$M = \frac{1}{m} Y^* Y \in \mathbb{R}^{n_1 \times n_1} \quad \text{with} \quad Y_{ij} = f\left(\frac{1}{\sqrt{n_0}}(WX)_{ij}\right) \quad \text{for} \quad 1 \leqslant i \leqslant n_1, \quad 1 \leqslant j \leqslant m.$$

In the above equation f is a nonlinear activation function, W is the $n_1 \times n_0$ matrix corresponding to the weights and X the $n_0 \times m$ matrix of the data. There are several possibilities to incorporate randomness in this model. In [29], the authors consider random weights with *deterministic data* X. The weights are given by functions of Gaussian random variables and the asymptotic eigenvalue distribution of M is studied thanks to concentration inequalities in the case where the function f is Lipschitz continuous. They prove that the eigenvalue distribution corresponds to that of a (usual) sample covariance matrix $\frac{1}{m} T^* X^* X T$ with population covariance $T^* T = \overline{M}$ as studied in [36]. However, there is a difference from a usual sample covariance matrix ensemble, which is the non universality of the eigenvalue distribution (as $\overline{M}$ depends on the distribution of W beyond its first two moments). The authors [29] use this equation to study the effect of the fourth moments of the distribution for the efficiency of the neural networks. The general approach based on concentration arguments that they develop is detailed in the recent preprint [28].

The model we are interested in has been introduced by [32]. Another approach, based on entropy and an information theory approach, has been obtained for the same model in [18]. [32] consider the case where both matrices W and X are random and independent with normalized Gaussian entries. Interestingly, they derive a fixed point equation for the Stieltjes transform of the asymptotic e.e.d., which is a *quartic equation* (recalled in Theorem 2.3 below). Before discussing our result, one may note that the quartic equation specializes in some special cases of the parameters to the Marčenko-Pastur equation for the Stieltjes transform: $zm(z)^2 + \left(\left(1 - \frac{\psi}{\phi}\right) z - 1\right) m(z) + \frac{\psi}{\phi} = 0$. Thus there exists a class of functions such that the nonlinear matrix model has the same limiting e.e.d. as that of Wishart matrices. The equation also becomes cubic when the function f is linear and corresponds to the product Wishart matrix. The limiting e.e.d. of such matrices, known as the Fuss–Catalan or Raney distribution, has been computed in [33, 13, 17, 3]. For the general case, [30] (see Theorem 1.4) shows that μ is actually the limiting e.e.d. of an information plus noise sample covariance matrix.

We refer the reader to Sections 4 and 5 of [32] for a more detailed discussion on machine learning applications of such a result. Regarding potential applications, the question of multiple layers is of particular interest. In particular [32] use this equation to facilitate the choice of activation function, a problem which has a crucial impact on the training procedure. In [22], the choice of activation function was studied for random neural networks after going through a large number of layers. One may first note that for linear models, the multilayer case corresponds to the maybe simpler setting of products of random matrices. One refers the reader to [26, 11, 1, 2, 33] for products of complex Ginibre matrices and to [21] where a large product of large random matrices is considered. In the non linear setting, [32] conjecture that the Marčenko-Pastur is invariant through multiple layers for some appropriate activation functions f, which could speed up training through the network. This is also a question we are interested in.

The scope of this paper is first theoretical: one aims to study the asymptotic e.e.d. of Gram matrices $f(WX)f(WX)^*$ where f is applied entrywise and to extend the result

established by [32] to non-Gaussian matrices. In particular, the question of universality of the limiting e.e.d. is of interest here as initial weights can be chosen to be non-Gaussian (a typical example is the uniform distribution as in [20]). In this setting, it has to be compared to the result of [14] where some kernel matrices are investigated (with another universal limiting empirical eigenvalue distribution). Our result may be of use for applications as it provides an easy way to compare different possible activation functions for a certain class of distribution for both weights and data. We also investigate the multilayer case $Y^{(\ell)} = f(W^{(\ell-1)}Y^{(\ell-1)})$ for $\ell = 1 \dots L$ with L fixed and study again the asymptotic empirical eigenvalue distribution for a class of activation functions. This is actually a toy model for the multi-layer model (1.1), due to the fact that the matrices $W^{(\ell-1)}$ are here independent. This gives one step to understand the multilayer random neural networks and also confirms the conjecture made in [32].

2 Model and results

Consider a random matrix $X \in \mathbb{R}^{n_0 \times m}$ with $i.i.d.$ elements with distribution ν_1. Let also $W \in \mathbb{R}^{n_1 \times n_0}$ be a random matrix with $i.i.d.$ entries with distribution ν_2. W is called the weight matrix. Both distributions are centered and we denote the variance of each distribution by

$$\mathbb{E}\left[X_{ij}^2\right] = \sigma_x^2 \quad \text{and} \quad \mathbb{E}\left[W_{ij}^2\right] = \sigma_w^2. \tag{2.1}$$

We also need the following assumption on the tails of W and X: there exist constants ϑ_w, $\vartheta_x > 0$ and $\alpha > 1$ such that for any $t > 0$ we have

$$\mathbb{P}\left(|W_{11}| > t\right) \leqslant e^{-\vartheta_w t^\alpha} \quad \text{and} \quad \mathbb{P}\left(|X_{11}| > t\right) \leqslant e^{-\vartheta_x t^\alpha}. \tag{2.2}$$

Note that the above implies that there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n_0}}\sum_{k=1}^{n_0} W_{1k}X_{k1}\right| > t\right) \leqslant Ce^{-t^2/2}. \tag{2.3}$$

We now consider a smooth function $f : \mathbb{R} \to \mathbb{R}$ scaled so that it has zero Gaussian mean in the sense that

$$\int f(\sigma_w\sigma_x x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x = 0. \tag{2.4}$$

This assumption has no impact on the asymptotic e.e.d. as this amounts to consider a rank one modification of the model if needed. However this greatly simplifies the exposition of the proof.

As an additional assumption, we also suppose that there exist positive constants C_f and c_f and $A_0 > 0$ such that for any $A \geqslant A_0$ and any $n \in \mathbb{N}$ we have,

$$\sup_{x\in[-A,A]} |f^{(n)}(x)| \leqslant C_f A^{c_f n}. \tag{2.5}$$

Remark 2.1. (2.5) guarantees that the function is real analytic which may be seen as a strong restriction. However, commonly used activation functions fall within the scope of this paper such as the sigmoid function $f(x) = (1 + e^{-x})^{-1}$, $f(x) = \tanh x$ or the softplus function $f(x) = \beta^{-1}\log(1 + e^{\beta x})$, i.e. a smooth variant of the ReLU. Extensions to more general (non analytic) functions f is the object of current research.

We consider the following random matrix,

$$M = \frac{1}{m}YY^* \in \mathbb{R}^{n_1 \times n_1} \quad \text{with} \quad Y = f\left(\frac{WX}{\sqrt{n_0}}\right) \tag{2.6}$$

where f is applied entrywise. We suppose that the dimensions of both the columns and the rows of each matrix grow together in the following sense: there exist positive constants ϕ and ψ such that

$$\frac{n_0}{m} \xrightarrow[m\to\infty]{} \phi, \quad \frac{n_0}{n_1} \xrightarrow[m\to\infty]{} \psi$$

Denote by $(\lambda_1, \dots, \lambda_{n_1})$ the eigenvalues of M given by (2.6) and define its e.e.d. by

$$\mu_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{\lambda_i}. \tag{2.7}$$

Theorem 2.2. *There exists a deterministic compactly supported measure μ such that we have*

$$\mu_{n_1}^{(f)} \xrightarrow[n_1\to\infty]{} \mu \quad \textit{weakly almost surely.}$$

Similarly we denote by $(\tilde{\lambda}_1, \dots, \tilde{\lambda}_{n_1}, 0 \dots, 0)$ the eigenvalues of $\frac{1}{m} Y^* Y$ (note that $m - n_1$ such eigenvalues are necessarily null). We set $\tilde{\mu}_m$ its e.e.d. and by $\tilde{\mu}$ its limit.

The moments of the asymptotic empirical eigenvalue distribution depend on the two following parameters of the function f: we set

$$\theta_1(f) = \int f^2(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x \quad \text{and} \quad \theta_2(f) = \left(\sigma_w \sigma_x \int f'(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x \right)^2. \tag{2.8}$$

We also define the following Stieltjes transforms: let $z \in \mathbb{C} \setminus \mathbb{R}$, we set

$$G(z) := \int \frac{\mathrm{d}\mu(x)}{x - z}, \quad \tilde{G}(z) := \int \frac{\mathrm{d}\tilde{\mu}(x)}{x - z} \quad \text{and } H(z) := \frac{\psi - 1}{\psi} - \frac{z}{\psi} G(z).$$

Theorem 2.3. *The measure μ satisfies the following fixed point equation for its Stieljes transform G:*

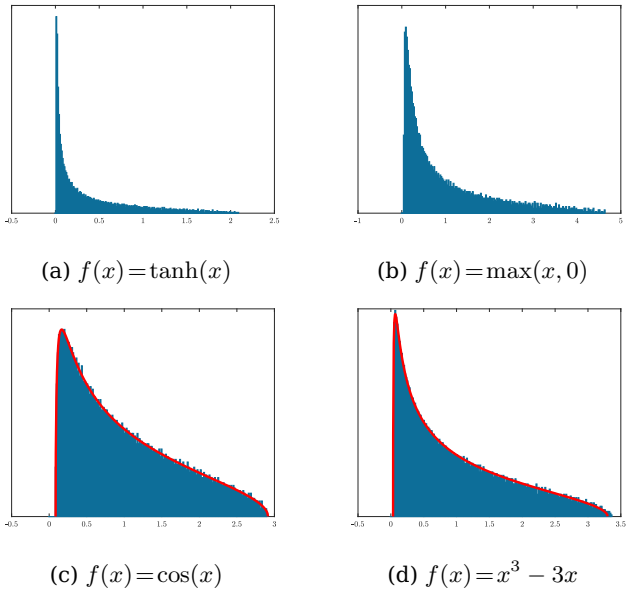
$$\frac{H(z)}{z} = \frac{1}{z} + \frac{G(z)\tilde{G}(z)(\theta_1(f) - \theta_2(f))}{\psi} + \frac{G(z)\tilde{G}(z)\theta_2(f)}{\psi - zG(z)\tilde{G}(z)\theta_2(f)},$$

with $\theta_1(f)$ and $\theta_2(f)$ are defined in (2.8)*.*

Remark 2.4. Theorem 2.3 is a universality statement: the asymptotic e.e.d. derived in [32] is universal for some class of distributions and activation functions. We observe first that all the dependence on f lies in the two parameters $\theta_1(f)$ and $\theta_2(f)$. The class of activation function that we consider is restrictive as we need strong regularity. However, these assumptions are needed in the proof to consider less assumptions on the random variables since we do not need strong concentration bounds. Our proof is based on a method of moments to recover the self-consistent equation for the Stieltjes transform. We defer the study of the asymptotic behavior of the largest eigenvalue as in [15] to another article.

The model given by (2.6) consists in passing the input data through one layer of a neural network as we apply the function f a single time. However, we could reinsert the output data through the network again, thus multiplying layers. It was conjectured in [32] that for activation functions such that $\theta_2(f) = 0$ the limiting e.e.d. is invariant and given by the Marčenko–Pastur distribution at each layer. We prove this statement in Theorem 2.5 below. We denote by L the number of layers and consider, for $p \in [\![0, L-1]\!]$ a family of independent matrices $W^{(p)} \in \mathbb{R}^{n_{p+1} \times n_p}$ where $(n_p)_p$ is a family of growing sequences of integers such that there exists ϕ and $(\psi_p)_p$ with

$$\frac{n_0}{m} \xrightarrow[m\to\infty]{} \phi \quad \frac{n_p}{n_{p+1}} \xrightarrow[m\to\infty]{} \psi_p.$$



(a) $f(x)=\tanh(x)$ (b) $f(x)=\max(x,0)$

(c) $f(x)=\cos(x)$ (d) $f(x)=x^3-3x$

Figure 1: Eigenvalues of M for different activation functions. Note that every function displayed here is actually scaled so that $\theta_1(f)=1$ and centered so that there is no very large eigenvalue. For the two bottom figures, we have $\theta_2(f)=0$ and the Marcenko–Pastur of shape parameter ϕ/ψ density is plotted in red.

We suppose that all the matrix entries $(W_{ij}^{(p)})_{ij}, 1\le i\le n_{p+1}, 1\le j\le n_p$, $p=0,\dots,L-1$ are $i.i.d$ with variance σ_w^2. Consider also $X\in\mathbb{R}^{n_0\times m}$ with $i.i.d$ entries of variance σ_x^2 and define the sequence of random matrices

$$Y^{(p+1)}=f\left(\frac{\sigma_x}{\sqrt{\theta_1(f)}}\frac{W^{(p)}Y^{(p)}}{\sqrt{n_p}}\right)\in\mathbb{R}^{n_{p+1}\times m}\quad\text{with}\quad Y^{(0)}=X. \tag{2.9}$$

The scaling is here chosen to normalize the variance of the entries of $Y^{(p)}$ at every layer. This normalization is known (adding centering) as batch normalization and is proved to improve the training speed [24]. The centering (2.4) is only important here. Now, one can define

$$M^{(L)}=\frac{1}{m}Y^{(L)}Y^{(L)*}\quad\text{and}\quad\mu_{n_L}^{(L)}=\frac{1}{n_L}\sum_{i=1}^{n_L}\delta_{\lambda_i^{(L)}},$$

where $(\lambda_k^{(L)})$ are the eigenvalues of $M^{(L)}$. We then prove the following theorem under the additional assumption that the function f is bounded.

Theorem 2.5. *Let L be a given integer. Suppose that f is a bounded analytic function such that* (2.4) *and* (2.5) *hold. In the case where $\theta_2(f)=0$, then the asymptotic e.e.d. $\mu_{n_L}^{(L)}$ is given almost surely by the Marčenko-Pastur distribution of shape parameter $\frac{\phi}{\psi_0\psi_1\cdots\psi_{L-1}}$.*

In particular the above result is consistent at any layer with the dimensions of the matrix $\tilde{M^{(L')}}=\frac{1}{n_L}{Y^{(L')}}^*Y^{(L')}, 1\le L'\le L$. This suggests that the limiting e.e.d. shall be the Marcenko-Pastur distribution for any L, (in particular when $L\to\infty$). Unfortunately our result does not encompass the case of a number of layers also growing to infinity.

Remark 2.6. The model we consider for several layers can be thought as a theoretical toy model since in practicality, weights would be updated along the neural network (using gradient descent for instance). Thus the independence assumption on the weights

shall not hold true. However, we are not able so far to handle the case where the entries of the the weight matrices are correlated.

The next section is dedicated to proving Theorem 2.2 for polynomial activation functions using the moment method. Our choice is motivated by the fact that [32] introduce a family of graphs to describe the asymptotic e.e.d. of Gaussian non linear random matrices, which we want to understand in greater generality. Thus the main part of the article has some combinatorial aspects and we believe that this point of view can give some insights in the study of these matrix models. Such a combinatorial method to study nonlinear random matrix models was previously used in [15] in the context of kernel matrices. In Section 4, we generalize the result to other functions by using a polynomial approximation. Finally, in Section 5 we first give a combinatorial description of the multilayer case for polynomials and then prove Theorem 2.5.

3 Limiting e.e.d. when f is a polynomial

The point of this section is to compute the moments of the empirical eigenvalue distribution of the matrix M when the activation is a polynomial. The following statement gives the expected moment of the distribution in this case using a graph enumeration. Before stating the result, we need the following definition.

Definition 3.1. *Let $q \geq 1$ be a given integer. A coincidence graph is a connected graph built up from the simple (bipartite) cycle of vertices labeled $i_1, j_1, i_2, \ldots, i_q, j_q$ (in order) by identifying some i-indices respectively and j-indices respectively. Such a graph is admissible if the formed cycles are joined to another by at most a common vertex and each edge belongs to a unique cycle.*

Remark 3.2. In the following, the edges and vertices of such an admissible graph are colored red.

Remark 3.3. An admissible graph has $2q$ edges. It can also be seen as a tree of cycles (simply replacing cycles by edges) also called a cactus graph. These graphs appear also in random matrix theory in the so-called theory of traffics when expanding injective traces (see [6] e.g.).

The basic admissible graph is given by the simple cycle (left figure on Figure 2) whose associated tree is a simple edge. The two right figures show a tree and one admissible graph that is associated to the tree: note that the points i_1 and j_1 where cycles are glued to each other are not determined by the tree, neither the lengths of the cycles.

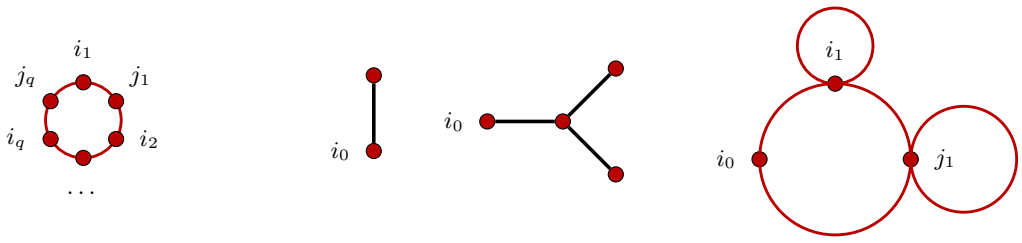


Figure 2: Two admissible graphs and their associated trees.

Definition 3.4. *$\mathcal{A}(q, I_i, I_j, b)$ is the number of admissible graphs with $2q$ edges, I_i i-identifications, I_j j-identifications and with exactly b cycles of size 2.*

We can now state the following Theorem. Let θ_1 and θ_2 be defined as in (2.8).

Theorem 3.5. *Let $f = \sum_{k=1}^{K} \frac{a_k}{k!}(x^k - k!!\mathbb{1}_{k \text{ even}})$ be a polynomial such that* (2.4) *holds.*

The degree of f, K, can grow with n_1 but we suppose that

$$K = \mathcal{O}\left(\frac{\log n_1}{\log\log n_1}\right). \tag{3.1}$$

Let $\mu_{n_1}^{(f)}$ be defined in (2.7) *and its expected moments $\overline{m}_q := \mathbb{E}\left[\langle \mu_{n_1}^{(f)}, x^q\rangle\right]$ We then have the following asymptotics*

$$\overline{m}_q = \sum_{I_i,I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \mathcal{A}(q,I_i,I_j,b)\theta_1(f)^b\theta_2(f)^{q-b}\psi^{I_i+1-q}\phi^{I_j}\left(1+o(1)\right). \tag{3.2}$$

Note that in this theorem we allow the degree K of the polynomial to grow with n_1 as in (3.1) but the theorem holds true for any fixed integer q (independent of n). It is possible to improve the assumption (3.1) in the sense that K could grow faster with n_1. However, this bound is enough for the polynomial approximation we need later (using a Taylor approximation of the function f). The proof of the above Theorem relies on combinatorial arguments we now develop.

3.1 Proof of Theorem 3.5 when f is a monomial of odd degree:

We first consider the case where $f(x) = \frac{x^k}{k!}$ for an odd integer k. We first assume that the entries of W and X are bounded in the following sense: there exists $A > 0$ such that

$$\max_{ij} |W_{ij}| + |X_{ij}| \leqslant A \quad \text{almost surely.}$$

3.1.1 Basic definitions

For this activation function, the entries of $Y = f(WX/\sqrt{n_0})$ are of the form

$$Y_{ij} = \frac{1}{k!}\left(\frac{WX}{\sqrt{n_0}}\right)^k_{ij} = \frac{1}{n_0^{k/2}k!}\left(\sum_{\ell=1}^{n_0} W_{ik}X_{kj}\right)^k = \frac{1}{n_0^{k/2}k!}\sum_{\ell_1,\dots\ell_k=1}^{n_0}\prod_{p=1}^{k} W_{i\ell_p}X_{\ell_p j}. \tag{3.3}$$

We want to study the normalized tracial moments of the matrix M. Thus we want to consider, for a positive integer q,

$$\begin{aligned}\frac{1}{n_1}\mathbb{E}\left[\operatorname{Tr} M^q\right] &= \frac{1}{n_1 m^q}\mathbb{E}\left[\operatorname{Tr}\left(YY^*\right)^q\right] \\ &= \frac{1}{n_1 m^q}\mathbb{E}\sum_{i_1,\dots,i_q=1}^{n_1}\sum_{j_1,\dots,j_q=1}^{m} Y_{i_1j_1}Y_{i_2j_1}Y_{i_2j_2}Y_{i_3j_2}\dots Y_{i_qj_q}Y_{i_1j_q}.\end{aligned} \tag{3.4}$$

We first encode each of the summand in (3.4) as a coincidence graph (not necessarily admissible) by simply marking the coinciding indices in the summand. Then injecting (3.3) in the previous equation we obtain the following expansion

$$\begin{aligned}&\frac{1}{n_1}\mathbb{E}\left[\operatorname{Tr} M^q\right] \\ &= \frac{1}{n_1 m^q n_0^{kq}(k!)^{2q}}\mathbb{E}\sum_{i_1,\dots,i_q}^{n_1}\sum_{j_1,\dots,j_q}^{m}\sum_{\substack{\ell_1^1,\dots\ell_k^1\\ \cdots\\ \ell_1^{2q}\dots\ell_k^{2q}}}^{n_0}\prod_{p=1}^{k} W_{i_1\ell_p^1}X_{\ell_p^1 j_1}\prod_{p=1}^{k} W_{i_2\ell_p^2}X_{\ell_p^2 j_1}\cdots\prod_{p=1}^{k} W_{i_1\ell_p^{2q}}X_{\ell_p^{2q}j_q}\end{aligned} \tag{3.5}$$

To take the l-indices into account, we now add to the red graph $2kq$ blue vertices. We can represent the vertices in a graph such as in Figure 3a. We call a red edge a *niche*. Each *niche* is decorated by k *blue* vertices from which leave blue edges corresponding to a term $W_{i\ell}X_{\ell j}$ in (3.5). Since the $W_{i\ell}$ and $X_{\ell j}$ are centered and independent, each such entry has to arise at least twice in the summand in equation (3.5). Thus, to compute the spectral moment, one needs to match the blue edges so that each entry arises with multiplicity at least 2. The matching of ℓ indices in (3.5) corresponds to a matching of the *blue* vertices. Then, the main contribution shall come from those summands maximizing the number of pairwise distinct indices.

3.1.2 The simplest admissible graph: a cycle of length $2q$

In this subsection, we assume that the i and j indices are pairwise distinct and consider the associated contribution to the spectral moment, which we denote by $\mathbf{E_q(k)}$. We show the following Lemma:

Lemma 3.6. *One has that*

$$\mathbf{E_q(k)} = \begin{cases} \theta_2^q(f)\psi^{1-q} + \mathcal{O}\left(\theta_2^q(f)\frac{q+k}{n_0}\right) & \text{if } q > 1 \\ \theta_1(f) + \mathcal{O}\left(\frac{k^2(2k-2)!!}{n_0(k!)^2}\right) & \text{if } q = 1. \end{cases}$$

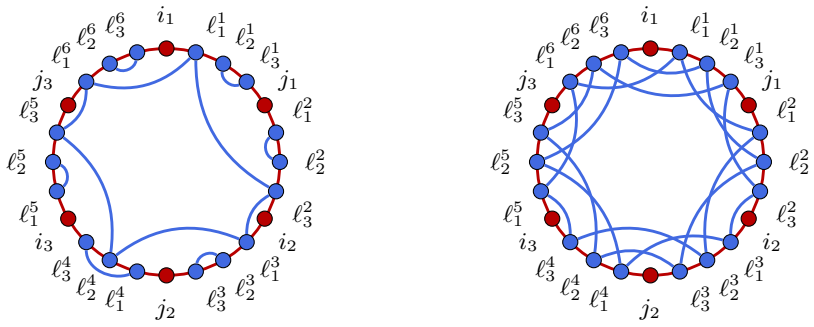
Proof of Lemma 3.6: Because the i- and j-indices are pairwise distinct, the associated red graph is the simple cycle of length $2q$. Thus we can really encode the products in the summand as in the left case of Figure 3. Since each matrix entry has to arise twice, say for instance W_{i_1,ℓ_1^1}, it needs to occur at least an other time in the product. There are then two different ways it can happen:

- (i) There exists $p \in \{2,\dots,k\}$ such that $\ell_p^1 = \ell_1^1$.
- (ii) There exists $p \in \{1,\dots,k\}$ such that $\ell_p^{2q} = \ell_1^1$. Applying the same reasoning for $X_{\ell_1^1,j_1}$, there exists $p' \in \{1,\dots,k\}$ such that $\ell_{p'}^2 = \ell_1^1$.

The same reasoning applies for each niche. Now, in order to maximize the number of pairwise distinct indices, one has to perform the most perfect matchings inside each niche. Note that, as k is odd, case (ii) necessarily occurs.

We first consider the contribution of those decorated graphs maximizing the number of pairwise distinct indices.

The case where $q > 1$: In this case, there is a blue cycle of size $2q$ as in Figure 3a. Thus we can construct the decorated graphs maximizing the number of pairwise distinct



(a) Leading order graph for $k = q = 3$ (b) Lower order graph for $k = q = 3$

Figure 3: The contribution of the simple cycle

indices in the following way: One chooses an index ℓ_p in each niche which is in the only blue cycle of the graph and then the remaining blue edges are perfectly matched inside niches. The corresponding contribution from the basic cycle to the moment is, as every entry exactly occurs twice in the products, using (2.1),

$$\mathbf{E_q}(\mathbf{k}) = \frac{((\sigma_w\sigma_x)^k k(k-1)!!)^{2q} n_0}{n_1 m^q n_0^{kq} (k!)^{2q}} \frac{m!}{(m-q)!} \frac{n_1!}{(n_1-q)!} \frac{n_0!}{(n_0-(k-1)q)!}$$

To obtain this formula, note that we choose the i-labels over n_1 possible indices and the j-labels over m indices. Now, we also choose the ℓ-labels over n_0: the one for the blue cycle and those vertices corresponding to matched edges. Finally, we have to fix the *blue* vertices belonging to the blue cycle: there are k^{2q} possible choices. The number of perfect matchings on the rest of the vertices in each niche is then equal to $((k-1)!!)^{2q}$. We then obtain that

$$\mathbf{E_q}(\mathbf{k}) = \left(\frac{(\sigma_w\sigma_x)^k k(k-1)!!}{k!}\right)^{2q} \psi^{1-q} + \mathcal{O}\left(\left(\frac{(\sigma_w\sigma_x)^k k(k-1)!!}{k!}\right)^{2q} \frac{q+k}{n_0}\right). \tag{3.6}$$

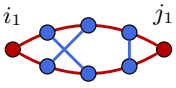
Note that, by (2.8), one has that $\theta_2(f) = \left(\frac{(\sigma_w\sigma_x)^k k(k-1)!!}{k!}\right)^2$ and we can write

$$\mathbf{E_q}(\mathbf{k}) = \theta_2^q(f)\psi^{1-q} + \mathcal{O}\left(\theta_2^q(f)\frac{q+k}{n_0}\right).$$

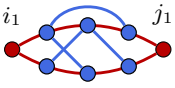
Case where $q = 1$. The case where $k = 1$ is slightly different. Indeed we can do any perfect matching between the $2k$ *blue* vertices. The graph can bee seen in Figure 4a. Thus, the contribution of the moments in this case is the following

$$\mathbf{E_1}(\mathbf{k}) = \frac{(\sigma_w\sigma_x)^{2k}(2k)!!}{(k!)^2} + \mathcal{O}\left(\frac{k^2(2k-2)!!}{n_0(k!)^2}\right) = \theta_1(f) + \mathcal{O}\left(\frac{k^2(2k-2)!!}{n_0(k!)^2}\right),$$

where the error comes from performing a matching which is not a perfect one. We now



(a) Contribution in the case where $q = 1$



(b) Subleading term in the case $q = 1$.

consider the contribution of other matchings. We will show that $\mathbf{E_q}(\mathbf{k})$ is indeed the typical contribution from the basic cycle, that is all other matchings lead to a negligible contribution with respect to $\mathbf{E_q}(\mathbf{k})$. There are four different phenomena that can give a (lower order) contribution. First, there may be more than one cycle linking every niche as in Figure 3b. Also, in at least one niche there could be more identifications between ℓ-indices, which raises moments of entries of W and X. There could be an identification between the index of the cycle and an index from a perfect matching inside a niche. Finally, there could also exist identifications between two distinct niches; note we can only get higher moments in the case where the two niches are adjacent. While these four behaviors can happen simultaneously, we see the contribution separately since it would induce an even smaller order if counted together.

a) There is more than one cycle between niches. We call $E_q^{(1)}$ the contribution to the moments of such decorated graphs. Suppose there are c cycles. Note that

necessarily c is odd since k is odd and entries are centered, then we can write, if we suppose that indices ℓ not in cycles are being perfectly matched,

$$E_q^{(1)} = \frac{(k^c(k-c)!!)^{2q}}{n_1 m^q n_o^{kq}(k!)^{2q}} \sum_{\substack{i_1,\dots,i_q \\ \text{pairwise} \\ \text{distinct}}}^{n_1} \sum_{\substack{j_1,\dots,j_q \\ \text{pairwise} \\ \text{distinct}}}^{m} \sum_{\ell_0,\dots,\ell_c}^{n_0} \sum_{\substack{\ell_1^1,\dots\ell_{\frac{k-c}{2}}^1 \\ \dots \\ \ell_1^{2q}\dots\ell_{\frac{k-c}{2}}^{2q}}}^{n_0} (\sigma_w\sigma_x)^{2kq}$$

$$= \frac{((\sigma_w\sigma_x)^k k^c(k-c)!!)^{2q}}{n_1 m^q n_0^{kq-c}(k!)^{2q}} \frac{m!}{(m-q)!} \frac{n_1!}{(n_1-q)!} \frac{n_0!}{(n_0-(k-c)q)!}.$$

In order to understand the very first term, note that one has to select in each niche c *blue* vertices to create the cycles and then do a perfect matching for the rest of the vertices. Thus one has that

$$E_q^{(1)} = \frac{((\sigma_w\sigma_x)^k k(k-c)!!)^{2q}\psi^{1-q}}{n_0^{(c-1)(q-1)}(k!)^{2q}} (1+o(1)). \tag{3.7}$$

Thus this is of smaller order than (3.6) when the number of cycles is strictly greater than 1 as in Figure 3b for instance. Indeed, one obtains that

$$\frac{E_q^{(1)}}{\mathbf{E_q(k)}} = \mathcal{O}\left(\frac{1}{n_0^{(c-1)(q-1)}}\left(\frac{(k-c)!!}{(k-1)!!}\right)^{2q}\right).$$

b) The matching in each niche is not a perfect matching- apart from the vertex in the cycle. If the matching is more complicated than a perfect matching, the associated moments could be of higher order than the variance. Consider a matching inside a niche which is not a perfect one: there exists then an identification between $a_1,\dots,a_b$ entries such that $a_1+\dots+a_b = k-1$ and such that at least one of the a_i's is greater than 2. For ease we suppose that $a_1 = \dots = a_{b_1} = 2$ and $a_{b_1+1},\dots,a_b > 2$ for some $b_1 \in [\![1, b-1]\!]$. See e.g. Figure 5.

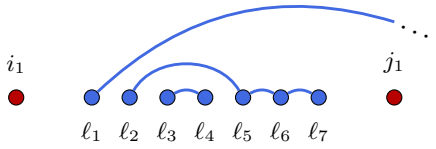


Figure 5: Niche where the induced graph is not a perfect matching which raises a fourth moment in the case where $k = 7$.

We call $E_q^{(2)}$ the contribution of all such matchings where a single niche breaks the perfect matching condition. Then we obtain that:

$$\frac{E_q^{(2)}}{\mathbf{E_q(k)}} = \sum_{b=1}^{\frac{k-1}{2}-1} \sum_{b_1=1}^{b-1} \sum_{\substack{a_{b_1+1}\dots a_b>2 \\ \sum a_j = k-1-2b_1}} \frac{(k-1)!}{(k-1)!!\prod_{i=1}^b a_i! b!} \frac{n_0!}{(n_0-(1+b+\frac{k-1}{2}(2q-1)))!} \times$$

$$\times \frac{(n_0-(1+(k-1)q)!}{n_0!} \frac{\prod_{b_1+1}^b \mathbb{E}|W_{11}|^{a_p}\mathbb{E}|X_{11}|^{a_p}}{(\sigma_w\sigma_x)^{2(\frac{k-1}{2}-b_1)}}.$$

The first term in the summand counts the number of matchings of the $k-1$ remaining *blue* vertices (after the choice of the cycle) into b classes of size $a_1 \dots a_b$. We can bound it in the following way

$$\frac{(k-1)!}{(k-1)!! \prod_{i=1}^{b} a_i! b!} \leqslant \frac{2^{\frac{k-1}{2}-b_1}(\frac{k-1}{2})!}{\prod_{i=b_1+1}^{b} a_i!} \leqslant \left(\frac{k-1}{2}\right)^{\frac{k-1}{2}-b} \frac{2^{\frac{k-1}{2}-b_1}}{\prod_{i=b_1+1}^{b} a_i! b!} \leqslant (k-1)^{(\frac{k-1}{2}-b)}.$$

In the first inequality we use the fact that $a_1 = \dots = a_{b_1} = 2$ and the definition of the double factorial. Then we expand the factorial and in the last inequality we use the fact that $a_i \geqslant 3$ for $i > b_1$. Now, for the second term, we compare the number of possible choices for ℓ indices, yielding that

$$\frac{(n_0 - (1+(k-1)q)!}{(n_0 - (1+b+\frac{k-1}{2}(2q-1)))!} \leqslant \frac{1}{n_0^{\frac{k-1}{2}-b}} e^{-\frac{C(kq)^2}{N}}.$$

Finally, the last term in the summand corresponds to the different possible moments, as only variances intervene in the leading contribution, while higher moments can appear inside the niche $\{i_1, j_1\}$. We use the fact that

$$\frac{\prod_{b_1+1}^{b} \mathbb{E}|W_{11}|^{a_p} \mathbb{E}|X_{11}|^{a_p}}{(\sigma_w \sigma_x)^{2(\frac{k-1}{2}-b_1)}} \leqslant \frac{A^{2\sum_{i \geqslant b_1+1} a_i}}{\prod_{i \geqslant b_1+1} \sigma_w^{a_i} \sigma_x^{a_i}} = \left(\frac{A^4}{\sigma_w^2 \sigma_x^2}\right)^{\frac{k-1}{2}-b_1}. \tag{3.8}$$

Now we need to bound the combinatorial factor coming from the sums:

$$\sum_{b_1=1}^{b-1} \sum_{\substack{a_{b_1+1},\dots,a_b \geqslant 3 \\ \sum a_j = k-1-2b_1}} \leqslant \sum_{b_1=1}^{b-1} \binom{k-1-3b-b_1+b-b_1-1}{b-b_1-1}$$

$$\leqslant \sum_{b_1=1}^{b-1} (k-1)^{k-1-3b+b_1} \leqslant (k-1)^{2(\frac{k-1}{2}-b)},$$

where we use in the first inequality that $\sum_j (a_j - 3) = k-1-2b_1-3(b-b_1)$. Finally, putting all these contributions together, we obtain the following comparison between $E_q^{(2)}$ and $\mathbf{E_q}(\mathbf{k})$,

$$\frac{E_q^{(2)}}{\mathbf{E_q}(\mathbf{k})} \leqslant \sum_{b=1}^{\frac{k-1}{2}-1} \left(\frac{CA^4}{\sigma_w^2 \sigma_x^2} \frac{(k-1)^3}{n_0}\right)^{3(\frac{k-1}{2}-b)} = \mathcal{O}\left(\frac{Ck^3}{n_0}\right). \tag{3.9}$$

Note that $k^3 = o(n_0)$. Here we suppose that in all other niches a perfect matching and a single cycle is used to match the blue vertices. The other cases are just negligible.

c) There are identifications between matchings from different niches If these niches are not adjacent, then such matchings would not increase the moments of the entries of W or X. On the contrary, matchings between adjacent niches may result into moments of higher order than the variance. We can then perform the same analysis as the previous one where we replace $k-1$ (the remaining indices after the choice of the cycle in one niche) to $2k-2$ corresponding to the number of vertices of two adjacent niches. This yields a contribution in the order of (3.9) with respect to $\mathbf{E_q}(\mathbf{k})$.

d) There are identifications between the cycle and perfect matchings inside niches. Suppose that these identifications happen in d niches, and for $p \in \{1, \dots, d\}$,

we identify the index from the cycle with $2b_p$ *blue* vertices from the niche. Indeed if the number of identifications was odd, in order to obtain a non-vanishing term, we would need to either create another cycle or perform more identifications inside the niches. Thus, we obtain the following upper bound

$$\frac{E_q^{(3)}}{\mathbf{E_q(k)}} = \sum_{d=1}^{2q} \sum_{b_1,\dots,b_d=1}^{\frac{k-1}{2}} \binom{2q}{d} \left[\prod_{p=1}^{d} \binom{k-1}{b_p}\right] \prod_{i=1}^{d} \mathbb{E}|W_{11}|^{2+2b_p} \mathbb{E}|X_{11}|^{2+2b_p} \times \\ \times \frac{((k-1)!!)^{2q-d} \prod_{p=1}^{d} ((k-2b_p-1)!!)}{n_0^{\sum_{p=1}^{d} b_p} ((k-1)!!)^{2q} (\sigma_w \sigma_x)^{2d+\sum_{p=1}^{d} 2b_p}}.$$

This comes from the choices of the niches, the identifications we make in each niche, and the perfect matchings we perform in the other niches. Finally, we suppose that we perform perfect matchings in the rest of the d niches. Then, we can use the bounds

$$\prod_{p=1}^{d} \frac{1}{b_p!} \leqslant 1, \quad \prod_{i=1}^{d} \mathbb{E}|W_{11}|^{2+2b_p} \mathbb{E}|X_{11}|^{2+2b_p} \leqslant A^{4d+4\sum_{i=1}^{d} b_i} \tag{3.10}$$

and

$$\frac{(k-1)!!^{2q-d} \prod_{p=1}^{d} (k-1-2b_p)!!}{(k-1)!!^{2q}} \leqslant 1.$$

From the above we obtain that

$$\frac{E_q^{(3)}}{\mathbf{E_q(k)}} \leqslant \sum_{d=1}^{2q} \binom{2q}{d} \left(\frac{A^4}{\sigma_w^2 \sigma_x^2}\right)^d \sum_{b_1,\dots,b_p=1}^{\frac{k-1}{2}} \left(\frac{A^4(k-1)}{2\sigma_w^2 \sigma_x^2 n_0}\right)^{\sum_{i=1}^{p} b_i} \\ = \sum_{d=1}^{2q} \binom{2q}{d} \left(\frac{A^4}{\sigma_w \sigma_x} \sum_{b=1}^{\frac{k-1}{2}} \left(\frac{A^4(k-1)}{2\sigma_w^2 \sigma_x^2 n_0}\right)^b\right)^d.$$

Now since $k \ll n_0$, we obtain that $\frac{E_q^{(3)}}{\mathbf{E_q(k)}} = \mathcal{O}\left(\frac{Ck}{n_0}\right)$. This finishes the proof of Lemma 3.6. □

3.1.3 Contribution of general admissible graphs

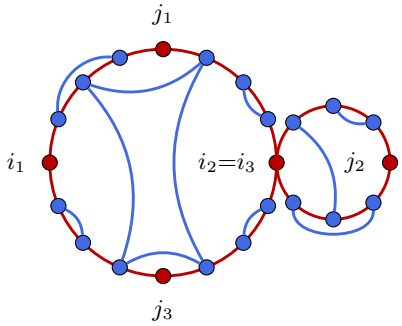
Lemma 3.7. *The total contribution from admissible graphs to the spectral moment is*

$$E_q(k) = \sum_{I_i,I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \mathcal{A}(q, I_i, I_j, b) \theta_1(f)^b \theta_2(f)^{q-b} \psi^{I_i+1-q} \phi^{I_j} (1 + o(1)). \tag{3.11}$$

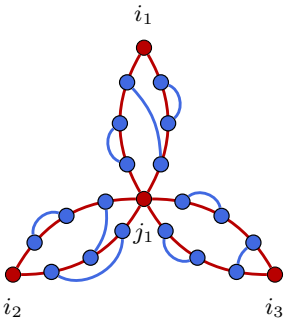
Remark 3.8. Lemma 3.7 is almost the statement of Theorem 3.5.

Proof of Lemma 3.7: We now suppose that there are I_i identifications between the vertices indexed by i labels and I_j identifications between the vertices indexed by j labels. Note that by our definition, such a graph is admissible if and only if it consists of $I_i + I_j + 1$ cycles. See for example Figures 6a and 6b. As seen earlier in the case of a simple cycle, the case of a cycle of size 2 has to be considered separately. Thus we denote by b the number of cycles of size 2.

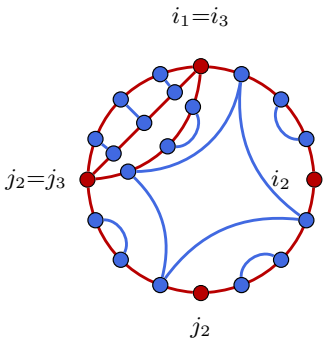
We can do a similar analysis in the case of general admissible graphs because we can realize *blue* identifications inside each red cycle as they are well defined. Thus, recalling



(a) Admissible graph with the i-identification $i_2 = i_3$ for $k = q = 3$.



(b) Admissible graph with the j-identification $j_1 = j_2 = j_3$ for $k = q = 3$.



(c) Non-admissible graph for $k = q = 3$.

Figure 6: Examples of admissible and non-admissible graphs

$\mathcal{A}(q, I_i, I_j, b)$ from Definition 3.4, we can write the contribution from all admissible graphs as

$$E'_q(k) = \frac{1+o(1)}{n_1 m^q n_0^{kq}} \sum_{I_i, I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \frac{n_1!}{(n_1-q+I_i)!} \frac{m!}{(m-q+I_j)!} \times \\ \times \mathcal{A}(k, q, I_i, I_j, b) \theta_1^b(f) n_0^{kb} \theta_2^{q-b}(f) n_0^{(k-1)(q-b)+I_i+I_j+1-b}.$$

Thus we obtain (3.11) provided we show that the error terms are negligible.

Note that the same error terms arise as in cases a), b), c) or d) for each red cycle: their contribution is then negligible as before as soon as matchings are still performed inside each cycle.

Another possible contribution may come from cross-cycle *blue* identifications: we now show this contribution is subleading. Consider the first case where such a cross-cycle identification arises around an i-identification or a j-identification: see e.g. Figure 7. These *blue* edges match entries of W to get a non-vanishing moments. However, in order to match the corresponding X entries, some new identifications are needed. This either implies that inside a niche, the matching is not a perfect matching. In this case the total final contribution is in the order of (3.9). Either this implies that two *blue* cycles going through two cycles bear the same vertex. Thus the total contribution of such cases is in the order of $n_0^{-1} E_q$ due to the fact that one loses a possible choice for the index of a *blue*

cycle. There may also exist cross-cycle *blue* identifications which do not arise around an $i-$ or $j-$ identification. It is not difficult to check in this case that the total contribution is again at most of the order of $n_0^{-1}E_q$. This finishes the proof of Lemma 3.7. □

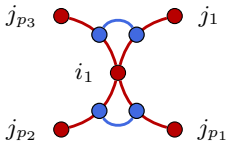


Figure 7: Subleading *blue* identifications around an i-identification

3.1.4 Contribution from non-admissible graphs

Now we estimate the contribution of non-admissible graphs which we denote $E_q^{(\mathrm{NA})}$. Our aim is to show the following Lemma.

Lemma 3.9. *One has that*

$$\frac{E_q^{(\mathrm{NA})}}{E_q(k)} = \mathcal{O}\left(\frac{q^5(1+q^{2k})}{n_0}\right). \tag{3.12}$$

Once Lemma 3.9 is proved, this finishes the proof of Theorem 3.5 in the case where f is an odd monomial.

Proof of Lemma 3.9: Let us first come back to admissible graphs. Starting from the origin i_1 of an admissible graph G, there is a single way to run through the different cycles and return to the origin. Note that all the cycles are oriented e.g. counter clockwise. They are called the *fundamental* cycles: they correspond to the cycles where we perform a matching on the *blue* vertices (otherwise the contribution is negligible). An admissible graph G can then be (partially) encoded into a rooted tree $T = (V, E)$ as follows. A fundamental cycle is represented by an edge and adjacent cycles yield adjacent edges. The tree inherits the orientation of the graph.

A non-admissible graph is a multigraph $G = (V, E_1, E_2)$ where E_1 denotes the set of single edges and E_2 is the set of multiple edges. There are first multiple ways to determine the fundamental cycles. Thus, we have to count the number of non-admissible graphs labeled by their fundamental cycles (see Figure 8 for an illustration). There may be also multiple ways to go through the whole graph: we explain later how this can be counted thanks to the associated admissible graph.

Our aim is to obtain all the non admissible graphs from the set of admissible ones by adding $i-$ and j- identifications. Consider an admissible graph G_0 with $\mathcal{C}$ fundamental cycles, we can then choose two cycles and *glue* them together in the sense of identifying one vertex of one cycle to one of the other. This adds one identification to the initial graph and encodes one non-admissible graph with identified fundamental cycles (by G_0). Now, one can repeat this identification process either for the same two cycles or for some others. The number of possible ways to choose two cycles which are then identified at r pairs of vertices is then at most:

$$\binom{\mathcal{C}}{2}(2q)^r \leqslant (Cq)^{r+2}, \tag{3.13}$$

for some constant C, since we need to choose two edges and then two vertices. And the number of possible ways to go through the whole graph is then multiplied by a factor at

most $(2r)!$. Indeed one needs to discover the fundamental cycles in the order they are initially numbered on the admissible graph. The moments of time where one can make a choice when running through the graph correspond to the vertices of degree greater than 2. And a vertex v of degree $2r_i > 2$ induces at most $r_i!$ possible ways to leave v after discovering the fundamental cycles starting from v.
However, one loses a power of n_1 (or m) for each additional identification as one loses a possible choice of index. Denote by $E_q^{(\mathrm{NA},r)}$ the contribution of non admissible graphs with no cross matchings (apart from the cycle inside each red cycle). Then one has that

$$E_q^{(\mathrm{NA},r)} = \sum_r \left(\frac{q^5}{n_0}\right)^r E_q(k) \ll E_q(k), \tag{3.14}$$

as $q^5 \ll n_0$. Hereabove r denotes the total number of additionnal identifications.

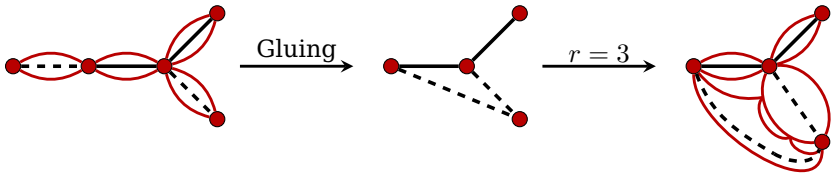


Figure 8: The left picture is an admissible graph with its encoding tree. The two dashed lines correspond to the two edges we glue together. The second graph correspond to the glued tree where there is now a cycle. Now the last step consists in choosing the number of identifications: here we have three total identifications between the two cycles.

Now, once the fundamental cycles are identified, cross identifications between *blue* edges from distinct niches (or fundamental cycles) are subleading unless in the following case: there are multiple cycles of length 2. Consider a cycle of length 2, with multiplicity p. First the number of ways to run through the graph is modifed as follows. Consider a vertex v to which are attached p_1 cycles and a cycle of length 2 with multiplicity p. The number of possible ways to leave v is at most $(p+p_1)!/p!$. This does not exceed $(2p_1)!$ if $p \leq p_1$ or $2^{2p}p^p$ if $p > p_1$. Then pk *blue* vertices have to be matched. While the leading order is given by performing a perfect matching between these vertices such as in Figure 9, we can do any kind of matching and use the similar analysis we did for (3.9). Suppose that we have an identification between $a_1, \dots, a_b$ entries such that $a_1 + \cdots + a_b = pk$. For ease we suppose that $a_1 = \cdots = a_{b_1} = 2$ and $a_{b_1+1}, \dots, a_b > 2$ for some $b_1 \in [\![1, b-1]\!]$, then we can compare their contribution to that of the admissible graph (used to encode it) by

$$\sum_{p=2}^{q} \sum_{b=1}^{\frac{m(e)k}{2}} \frac{p^p}{n_0^{p-1}} \sum_{b_1=1}^{b} \sum_{\substack{a_{b_1+1},\dots,a_b>2 \\ \sum a_i = pk-2b_i}} \frac{(pk)!}{((2k)!!)^{p/2} b! \prod_{i=b_1+1}^{b} a_i!} \frac{n_0^b \prod_{i=b_1+1}^{b} \mathbb{E}|W_{11}|^{a_i} \mathbb{E}|X_{11}|^{a_i}}{n_0^{kp/2} (\sigma_w^2 \sigma_x^2)^{kp/2-b}} \tag{3.15}$$

The factor of n_0^{1-p} comes from the additional identifications between i's and j's in order to obtain a multiple edge. For instance in Figure 9 there are less identifications in the admissible graph than in the corresponding non-admissible graph. The first term in the summand compares the number of possible matchings of the pk edges to that of a perfect matching in every single cycle. There exists a constant $C > 0$ such that

$$\frac{(pk)!}{((2k)!!)^{p/2} b! \prod_{i=b_1+1}^{b} a_i!} \leqslant (Cp)^{kp}.$$

The second term now comes from the number of ℓ indices chosen and the ratio of moments and we bound it in the same way as in (3.9),

$$\frac{\prod_{b_1+1}^{b}\mathbb{E}|W_{11}|^{a_p}\mathbb{E}|X_{11}|^{a_p}}{n_0^{kp/2-b}(\sigma_w\sigma_x)^{2(kp/2-b)}} \leqslant \frac{A^{2\sum_{i\geqslant b_1+1}a_i}}{n_0^{kp/2-b}\prod_{i\geqslant b_1+1}\sigma_w^{a_i}\sigma_x^{a_i}} = \left(\frac{A^4}{n_0\sigma_w^2\sigma_x^2}\right)^{\frac{kp}{2}-b}.$$

Also in the same way as in (3.9), we can bound the combinatorial factor coming from the sums as

$$\sum_{b_1=1}^{b}\sum_{\substack{a_{b_1+1},\dots,a_b>2\\ \sum a_i=pk-2b_i}} \leqslant (pk)^{2(\frac{pk}{2}-b)}.$$

Finally, putting all the contribution together we have

$$n_0\sum_{p=2}^{q}\left(\frac{Cp^{k+1}}{n_0}\right)^p\sum_{b=1}^{pk/2}\left(\frac{A^4p^2k^2}{n_0\sigma_w^2\sigma_x^2}\right)^{\frac{kp}{2}-b} = \mathcal{O}\left(\frac{q^{2k+1}}{n_0}\right), \tag{3.16}$$

where we used the fact that the leading order comes from the case where $b=\frac{kp}{2}$. Actually (3.16) can be improved to $O(\frac{q^{kj}}{n_0^{j-1}})$ for any integer $1<j<q$.

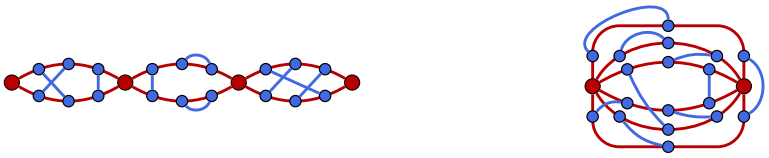


Figure 9: Different behavior between an admissible graph and a multiple edge.

Thus, combining (3.14), (3.13) and (3.16) finishes the proof of Lemma 3.9. □

3.2 Proof of Theorem 3.5 when f is a monomial of even degree:

In the case of an even monomial we center the function f, to do so we substract a constant given by the corresponding expectation. We then consider centered monomial of the form

$$f(x)=\frac{x^k-k!!}{k!},\ k=2p \quad \text{so that} \quad \theta_1(f)=\frac{(\sigma_w\sigma_x)^{2k}}{(k!)^2}\left((2k)!!-(k!!)^2\right) \quad \text{and} \quad \theta_2(f)=0.$$

Here, the fact that $\theta_2(f)$ vanishes means that all admissible graphs which have at least one cycle of size greater than 2 are subleading so that we see admissible graphs consisting only in cycles of size 2 such as Figure 6b for instance. Note that we have seen earlier that we can write

$$\mathbb{E}\left[\frac{1}{k!}\left(\frac{(WX)_{ij}}{\sqrt{n_0}}\right)^k\right]=\frac{1}{n_0^{k/2}k!}\mathbb{E}\sum_{\ell_1,\dots,\ell_k=1}^{n_0}\prod_{p=1}^{k}W_{i\ell_p}X_{\ell_pj}=\frac{k!!}{k!}(\sigma_w\sigma_x)^k\left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right).$$

Thus, by developing the tracial moments of M we obtain the following formula,

$$\frac{1}{n_1}\mathbb{E}\left[\operatorname{Tr} M^q\right] = \left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right)\frac{1}{n_1 m^q}\mathbb{E}\sum_{i_1,\dots,i_q}^{n_1}\sum_{j_1,\dots,j_q}^{m}\left[\frac{1}{n_0^{kq}(k!)^{2q}}\sum_{\substack{\ell_1^1,\dots\ell_k^1\\ \cdots\\ \ell_1^{2q}\dots\ell_k^{2q}}}^{n_0}\prod_{p=1}^{k} W_{i_1\ell_p^1}X_{\ell_p^1 j_1}\times\right.$$
$$\left.\times\prod_{p=1}^{k} W_{i_2\ell_p^2}X_{\ell_p^2 j_1}\cdots\prod_{p=1}^{k} W_{i_1\ell_p^{2q}}X_{\ell_p^{2q}j_q} - c_0^{2q}\right]. \quad (3.17)$$

Now it is not difficult to check that

$$c_0^{2q} = \left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right)\mathbb{E}\sum_{\ell_1^1,\dots\ell_k^1\dots\ell_1^{2q}\dots\ell_k^{2q}\,**}^{n_0}\prod_{p=1}^{k} W_{i_1\ell_p^1}X_{\ell_p^1 j_1}\prod_{p=1}^{k} W_{i_2\ell_p^2}X_{\ell_p^2 j_1}\cdots\prod_{p=1}^{k} W_{i_1\ell_p^{2q}}X_{\ell_p^{2q}j_q},$$

where the $**$ means that the ℓ-indices are matched according to a perfect matching inside each niche. Thus the centering by c_0 corresponds to the contribution of the admissible graphs where *blue* vertices make a perfect matching inside each niche.

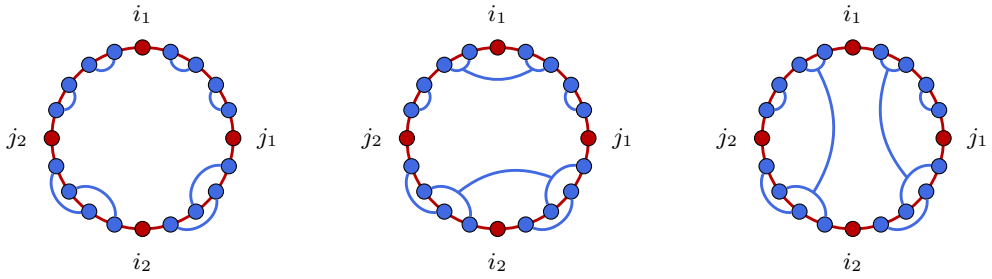


Figure 10: The left figure corresponds to the leading order before centering while the two others illustrate leading order graphs after centering. The center figure involves $\mathbb{E}W_{11}^4$ while the right one involves $\mathbb{E}X_{11}^4$.

After centering, the typical graphs may be those which have additional identifications between niches which have a common *red* vertex as in Figure 10. We first consider the contribution of one *red* cycle to the moments and then deduce the contribution of all admissible graphs. One can see that to maximize the number of possible choices of *blue* indices, we can first perform a perfect matching into each niche as before the centering, then we can choose either the i-vertices or the j-vertices and add identifications around the corresponding niches. This prevents having a perfect matching inside any niche (which is forbidden by the centering) but still the gives the maximal number of *blue* indices. With such a matching, moments of order 4 arise in the contribution and we obtain:

$$\begin{aligned}
&E_{q,1}(k)\\
&= \frac{1}{n_1(k!)^{2q}}\psi^{-q}\left(\frac{k}{2}(k!!)\right)^{2q}\left(\sigma_w^{2q(k-1)}\sigma_x^{2kq}(\mathbb{E}W_{11}^4)^q + \sigma_w^{2qk}\sigma_x^{2q(k-1)}(\mathbb{E}X_{11}^4)^q\right) + o\left(\frac{\theta_3(f)}{n_1}\right)\\
&= \frac{1}{n_1}\left[\theta_3(f)\left(\frac{\mathbb{E}W_{11}^4}{\sigma_w^4}+\frac{\mathbb{E}X_{11}^4}{\sigma_x^4}\right)\right]^q\psi^{-q} + o\left(\frac{\theta_3(f)}{n_1}\right)
\end{aligned}$$

where we defined

$$\theta_3(f) = \left(\frac{(\sigma_w \sigma_x)^2}{2} \int f''(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x \right)^2 .$$

Note that the contribution is of order n_1^{-1} and thus is negligible compared with the contribution from odd polynomials. For the number of distinct indices we obtain $n_1^{(k-1)q}$. We could try to instead create cycles between niches as for the odd polynomial case, but one can see that we would need to create two cycles instead of one and would obtain $(k-2)q+2$ distinct indices which is of lower order. Now if we only create one cycle, we need to perform at least identifications between three vertices in each niche since we would have an odd number of *blue* vertices left and the number of distinct indices becomes at most $(k-4)q+2q+1$ which is also of lower order than Figure 10. If, instead of identifying between different niches we would identify *blue* vertices inside the same niche we can only obtain at most $(k-4)q+2q$ distinct indices which is of lower order than Figure 10.

Now, in the same way, the case of a *simple cycle* (i.e. with length 2) is slightly different due to the centering. Indeed, at least one (thus two) vertices has to be connected to some other niche. Note also that any perfect matching where the two niches are connected is of the same order, thus we obtain for the leading order

$$E_1(k) = \frac{(\sigma_w \sigma_x)^{2k}}{(k!)^2} \left((2k)!! - (k!!)^2 \right) + \mathcal{O}\left(\frac{(2k-2)!! k^2}{(k!)^2 n_0} \right) = \theta_1(f) + \mathcal{O}\left(\frac{(2k-2)!! k^2}{(k!)^2 n_0} \right).$$

The above formula is self explanatory.

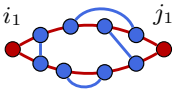


Figure 11: Contribution in the case $q = 1$ for an even monomial.

For the general case of admissible graphs with possible identifications, we use the fact that the contribution is just a product over the different cycles. For simplicity, we suppose that we have $\mathbb{E}W_{11}^4 \sigma_x^4 = \mathbb{E}X_{11}^4 \sigma_w^4$. Since the contribution of the cycles of length greater than 2 are $O(n_1^{-1})$, the terms involving the 4-th moments of the entries of X and W (which are not n_1-dependent) are subleading. Thus, this condition does not impact the overall order of the contribution but gives a simpler formula.

The leading order of a q-moment, corresponding to the total contribution of admissible graphs with $2q$ edges can be written as

$$E_q(k) = \frac{1+o(1)}{n_1 m^q n_0^{kq}} \sum_{I_i, I_j = 0}^{q} \sum_{b=0}^{I_i+I_j+1} \frac{n_1!}{(n-q+I_i)!} \frac{m!}{(m-q+I_j)!} \times$$
$$\times \mathcal{A}(q,k,I_i,I_j,b) \theta_1(f)^b n_0^{kb} 2^{I_i+I_j+1-b} \theta_3(f)^{q-b} \left(\frac{\mathbb{E}W_{11}^4}{\sigma_w^4} \right)^{q-b} n_0^{\frac{k-1}{2}(2q-2b)}$$

which gives asymptotically,

$$\begin{aligned}
E_q(k) &= (1+o(1)) \sum_{I_i,I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \left(\frac{2}{n_0}\right)^{(I_i+I_j+1)-b} \mathcal{A}(q,I_i,I_j,b)\theta_1(f)^b \left[\theta_3(f)\frac{\mathbb{E}W_{11}^4}{\sigma_w^4}\right]^{q-b} \phi^{I_j}\psi^{I_i+1-q} \\
&= (1+o(1)) \sum_{\substack{I_i,I_j=0 \\ I_i+I_j+1=q}} \mathcal{A}(q,I_i,I_j,I_i+I_j+1)\theta_1(f)^{I_i+I_j+1}\phi^{I_j}\psi^{I_i+1-q} \\
&= (1+o(1)) \sum_{I_i,I_j=0}^{q} \sum_{b=0}^{I_i+I_j+1} \mathcal{A}(q,I_i,I_j,b)\theta_1(f)^b\theta_2(f)^{q-b}\psi^{I_i+1-q}\phi^{I_j}
\end{aligned}$$

where we used in the last equality the fact that $\theta_2(f)=0$ in order to prove the expression (3.2). Note again that we did not give here all the errors since we have computed them in the previous subsection, the case of even monomials can be done similarly. Thus we can see that only the graphs which correspond to a tree of simple cycles contribute to the moments.

We can lead the analysis of the contribution from non-admissible graphs as in the previous section, as the *non admissible structure* only concerns the *red* graph while the (centered) polynomial involves only the matching on *blue* vertices. We leave the detail to the reader.

3.3 Proof of Theorem 3.5 when f is a polynomial:

We now suppose that we can write

$$f(x) = \sum_{k=1}^{K} a_k f_k(x) \quad \text{with} \quad f_k(x) = \frac{x^k - k!!\mathbb{1}_{k \text{ even}}}{k!} \quad \text{and} \quad \sup_{k\in[\![1,K]\!]} |a_k| \leqslant C^k \quad \text{for some } C.$$

In particular, the parameters are in this case

$$\theta_1(f) = \sum_{\substack{k_1,k_2=1 \\ k_1+k_2=:k_0 \text{ even}}}^{K} \frac{a_{k_1}a_{k_2}(\sigma_w\sigma_x)^{k_0}\,(k_0!! - k_1!!k_2!!\mathbb{1}_{k_1 \text{ even}})}{k_1!k_2!},$$

$$\theta_2(f) = \left(\sum_{\substack{k=1 \\ k \text{ odd}}}^{K} \frac{a_k(\sigma_w\sigma_x)^k k(k-1)!!}{k!}\right)^2.$$

Note that for any polynomial, by expanding the moment as in (3.5), we have to compute the following quantity, for any $k_1,\dots,k_{2q}$ integers,

$$\begin{aligned}
\frac{1}{n_1}\mathbb{E}\left[\operatorname{Tr} M^q\right] = &\sum_{k_1,\dots,k_{2q}=1}^{K} \frac{a_{k_1}\dots a_{k_{2q}}}{n_1 m^q \prod_{i=1}^{2q} k_i!} \times \\
&\times \mathbb{E} \sum_{i_1,\dots,i_q}^{n_1} \sum_{j_1,\dots,j_q}^{m} \sum_{\substack{\ell_1^1,\dots\ell_k^1 \\ \dots \\ \ell_1^{2q}\dots\ell_k^{2q}}}^{n_0} f_{k_1}\left(\frac{WX}{\sqrt{n_0}}\right)_{i_1j_1} f_{k_2}\left(\frac{WX}{\sqrt{n_0}}\right)_{i_2j_1} \dots f_{k_{2q}}\left(\frac{WX}{\sqrt{n_0}}\right)_{i_1j_q}
\end{aligned} \tag{3.18}$$

To compute the leading term of this moment, first note that the centering creates disparity between even and odd monomials. Indeed let $q>1$, if we consider one *red* cycle of length $2q$, there are now $2q$ niches of different sizes, namely $k_1,\dots,k_{2q}$. We first bound these moments in order to see that, in each cycle, the niches with an even number

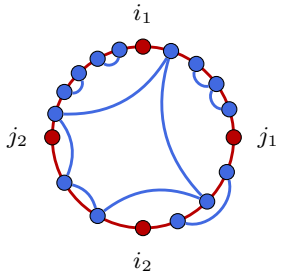


Figure 12: Admissible graph in the case of a polynomial with $(k_1, k_2, k_3, k_4) = (4, 3, 2, 5)$.

of vertices are subleading so that the dominant term in the asymptotic expansion of the moment corresponds to admissible graphs with only odd niches when expanding the polynomial. The behavior in a fundamental cycle can be understood as follows: there has to be at least one cycle connecting each niche for the odd or the centered even niches. Now, in each odd niche of length k_i, the leading term corresponds to a perfect matching of the $k_i - 1$ remaining vertices from (3.9). The number of pairwise distinct l indices in the niche is then $(k_i - 1)/2$, apart from the cycle. However, in the even niches, since there is already a cycle, there remains an odd number of vertices to be matched. The leading order is to disgard 2 vertices and then to perform a perfect matching of the $k_i - 2$ remaining vertices. The remaining vertices are matched to a blue cycle or to an existing matching. Then, the number of distinct l indices inside one niche is at most $(k_i - 2)/2$ (apart from cycles). Denote the number of choices of indices for *red* and *blue* vertices for a configuration of niches $k_1, \dots, k_{2q}$ by $C(k_1, \dots, k_{2q})$. Then we obtain

$$\frac{n_0^{-\sum_{i=1}^{2q} \frac{k_i}{2}}}{n_1 m^q} C(k_1, \dots, k_{2q}) = \frac{n_0^{-\sum_{i=1}^{2q} \frac{k_i}{2}}}{n_1 m^q} n_1^q m^q n_0^{1+\sum_{k_i \text{ odd}} \frac{k_i-1}{2} + \sum_{k_i \text{ even}} \frac{k_i-2}{2}} (1 + o(1))$$
$$= \frac{\psi^{1-q}}{n_0^{\frac{\#k_i \text{ even}}{2}}} (1 + o(1)).$$

This contribution can be understood in the following way: apart from the normalization, we have to choose the q i-indices, the q j-indices, the ℓ-indices. Thus, if we consider the contribution of cycles of size $q > 1$ for the polynomial $P = \sum \frac{a_k}{k!}(X^k - k!!\mathbb{1}_{k \text{ even}})$, we get the following asymptotic expansion for the moments

(3.18)

$$= \frac{1 + \mathcal{O}\left(\frac{1}{\sqrt{n_0}}\right)}{n_1 m^q} \sum_{\substack{k_1, \dots, k_{2q} \\ k_i \text{ odd}}} \left[\prod_{i=1}^{2q} \frac{a_{k_i}}{k_i!}\right] \frac{n_1^q m^q n_0^{1+\sum_{i, k_i \text{ odd}} \frac{k_i-1}{2}}}{n_0^{\sum_{i, k_i \text{ odd}} \frac{k_i}{2}}} \prod_{i, k_i \text{ odd}} (\sigma_w \sigma_x)^{k_i} k_i (k_i - 1)!!$$

$$= \psi^{1-q} \left(\sum_{k \text{ odd}} a_k (\sigma_w \sigma_x)^k k(k-1)!!\right)^{2q} \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n_0}}\right)\right) = \psi^{1-q} \theta_2^q(f) + \mathcal{O}\left(\frac{\theta_2^q(f)}{\sqrt{n_0}}\right).$$

As we now explain, in the case of a cycle consisting of two edges decorated by k_1 and k_2 *blue* vertices, there are three different possibilities: $i)$ if k_1 and k_2 are odd: the contribution to the moment is $(\sigma_x \sigma_w)^{k_1+k_2}(k_1 + k_2)!!$.; $ii)$ if k_1 and k_2 are even: the contribution is $(\sigma_w \sigma_x)^{k_1+k_2}((k_1 + k_2)!! - k_1!!k_2!!)$; iii) while if k_1 is even and k_2 is odd: the leading term in the asymptotic expansion is of order $n_0^{-1/2}$ due to the discrepancy. Thus,

the 1-moment for a polynomial f is

$$\sum_{\substack{k_1,k_2=1\\ k_1+k_2 \text{ even}}}^{K} \left(\frac{a_{k_1}a_{k_2}}{k_1!k_2!}(\sigma_w\sigma_x)^{k_1+k_2}\left((k_1+k_2)!! - k_1!!k_2!!\mathbb{1}_{k_1 \text{ even}}\right) + \mathcal{O}\left(\frac{(k_1+k_2)(k_1+k_2-1)!!}{\sqrt{n_0}k_1!k_2!}\right)\right) = \theta_1(f) + \mathcal{O}\left(\frac{K}{\sqrt{n_0}}\right)$$

where we used the fact that for any k_1 and k_2, $(k_1+k_2)!!/(k_1!k_2!)$ is bounded. While these analysis work in the case of a single cycle, we can do the same generalization to any (non) admissible graphs as before. Thus we get the following q-moment in the case of a polynomial

$$m_q := \frac{1}{n_1}\mathbb{E}\left[\mathrm{Tr}M^q\right] = (1+o(1))\sum_{I_i,I_j=0}^{q}\sum_{b=0}^{I_i+I_j+1}\mathcal{A}(q,I_i,I_j,b)\theta_1^b(f)\theta_2^{q-b}(f)\psi^{I_i+1-q}\phi^{I_j}.$$

This finishes the proof of Theorem 3.5 when f is a polynomial.

3.4 Convergence of moments in probability

In the previous subsection, we have proved convergence of the expected moments of the empirical eigenvalue distribution. We turn to the proof of the convergence in probability of these moments.

Lemma 3.10. *Let $f(x) = \sum_k^K a_k x^k$ be a polynomial activation function and consider the associated matrix M with enpirical eigenvalue distribution μ_{n_1}. Denote by m_q th moments $m_q = \frac{1}{n_1}\sum_{i=1}^{n_1}\lambda_i^q = \frac{1}{n_1}\mathrm{Tr}\,M^q$ and $\overline{m}_q = \mathbb{E}\left[m_q\right]$ we then have, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(|m_q - \overline{m}_q| > \varepsilon\right) \xrightarrow[n_1\to\infty]{} 0. \tag{3.19}$$

In addition there exists a constant C such that

$$\mathrm{Var}\, m_q = \mathcal{O}\left(\frac{(q^2K^2+q^4)C^q}{n_1^2}\right)$$

Proof. We can write the variance of the moments in the following way

$$\mathrm{Var}\, m_q = \mathbb{E}\left[\left(\frac{1}{n_1}\mathrm{Tr}\,M^q\right)^2\right] - \overline{m}_q^2 = \frac{1}{n_1^2}\sum_{\mathcal{G}_1,\,\mathcal{G}_2}\sum_{\boldsymbol{\ell_1},\boldsymbol{\ell_2}}\mathbb{E}\left[M_{\mathcal{G}_1}(\boldsymbol{\ell_1})M_{\mathcal{G}_2}(\boldsymbol{\ell_2})\right] - \mathbb{E}\left[M_{\mathcal{G}_1}(\boldsymbol{\ell_1})\right]\mathbb{E}\left[M_{\mathcal{G}_2}(\boldsymbol{\ell_2})\right]$$

with $\mathcal{G}_p = (G_p, \mathbf{i_p}, \mathbf{j_p})$ are labeled graphs with the i-labels and j-labels given respectively by $\mathbf{i_p}$, $\mathbf{j_p}$. For a given labeled graph $\mathcal{G} = (G, \mathbf{i}, \mathbf{j})$ and a matching $\boldsymbol{\ell}$, the notation $M_{\mathcal{G}}(\boldsymbol{\ell})$ corresponds to the following product after expansion

$$M_{\mathcal{G}}(\boldsymbol{\ell}) = \sum_{k_1,\dots,k_{2q}=1}^{K}\frac{a_{k_1}\dots a_{k_{2q}}}{m^q n_0^{\sum k_i/2}}\prod_{p=1}^{k_1}W_{i_1\ell_p^1}X_{\ell_p^1 j_1}\prod_{p=1}^{k_2}W_{i_2\ell_p^2}X_{\ell_p^2 j_1}\cdots\prod_{p=1}^{k_{2q}}W_{i_1\ell_p^{2q}}X_{\ell_p^{2q}j_q}.$$

Now, note that the shape of the graph and the possible expansion of the polynomial f does not depend on n_0, n_1 or m. By independence, the two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have to share an edge otherwise the contribution to the variance is null. In particular, the concatenated graph $\mathcal{G}$ cannot be admissible. Thus we only need to consider graphs $\mathcal{G}_1$,

$\mathcal{G}_2$ which share a common edge: either a red one or some $X_{\ell j}$ or $W_{i\ell}$ for some i, j, and ℓ. In other words the concatinated graph $\mathcal{G}$ cannot be admissible. We here assume for ease that $\mathcal{G}_1$ and $\mathcal{G}_2$ have $2q$ edges. The case where the number of edges is different in each cycle can be similarly handled.
To simplify the exposition of the argument further, we suppose that $\mathcal{G}_1$ and $\mathcal{G}_2$ are both a cycle and f is an odd monomial x^k. Note that the generalization comes from the fact that admissible graphs are a tree of cycles and non-admissible graphs yield a lower order contribution from (3.12). If we suppose that the coincidence between the two graphs comes from an i-label and a ℓ-label, in other words an entry $W_{i\ell}$, we have different possibilities that we now develop.

The first case consists in taking the two *red* cycles and attaching them at a fixed vertex i_0. We then perform a cross-cycle identification as in Figure 7 in order to match two entries $W_{i_0\ell_0}$ together from $\mathcal{G}_1$ and $\mathcal{G}_2$. Once these W entries are matched, note that the corresponding X entries have not been matched yet. We then need to identify this l-vertex with another vertex from an adjacent niche (and then creating a *blue* cycle going over the whole *red* cycle) or to another vertex in the same niche. Finally, it can be seen as simply performing the dominant matching into each graph, identifying two i indices and then identifying two *blue* edges from niches adjacent to i. Finally we can compute the contribution of these graphs in the covariance as

$$\sum_{\boldsymbol{\ell_1},\boldsymbol{\ell_2}} \mathrm{Cov}^{(1)}(M_{\mathcal{G}_1}(\boldsymbol{\ell_1}), M_{\mathcal{G}_2}(\boldsymbol{\ell_2})) = \mathcal{O}\left(q^2k^2\psi^{1-2q}\theta_2(f)^{2q}\left(\frac{\mathbb{E}W_{11}^4}{\sigma_w^4}-1\right)\right).$$

Indeed, in each graph we perform the typical matching corresponding to a *blue* cycle going over every niche and perfect matchings between the remaining indices in each niche. Now the fact that we identify two $W_{i_0\ell_0}$ entries create a moment of order 4 when we compute $\mathbb{E}[M_{\mathcal{G}_1}M_{\mathcal{G}_2}]$. We then have to count the number of possible choices for indices: we have n_1^{2q-1} choices for the i indices as we identify two from $\mathcal{G}_1$ and $\mathcal{G}_2$, m^{2q} for the j indices, $n_0^{2+4q(k-1)/2-1}$ choices for the ℓ indices (2 cycles, $4q$ niches and an identification between the two graphs). Taking into account the normalization $m^{-2q}n_0^{-2kq}$, this yields a factor ψ^{1-2q} asymptotically. In the same way, for general polynomial and admissible graphs, for such an identification we would obtain that

$$\frac{1}{n_1^2}\sum_{\mathcal{G}_1,\mathcal{G}_2}\sum_{\boldsymbol{\ell_1},\boldsymbol{\ell_2}} \mathrm{Cov}^{(1)}(M_{\mathcal{G}_1}(\boldsymbol{\ell_1}), M_{\mathcal{G}_1}(\boldsymbol{\ell_1})) \\ = \mathcal{O}\left(\frac{q^2k^2}{n_1^2\psi}\left(1-\frac{\phi}{\psi}\right)\mathfrak{m}_q^2\left(\frac{\mathbb{E}W_{11}^4}{\sigma_w^4}-1\right)\right) = \mathcal{O}\left(\frac{q^2k^2C^q}{n_1^2}\right),$$

for some $C>0$. Indeed, we get the q^2k^2 from the choices for the edge we want to identify between the two graphs, the constant factor in ϕ and ψ consists in the choice of choosing a $\{i,\ell\}$ edge or a $\{j,\ell\}$ edge. Then the previous computation in the case of a cycle can be generalized to all graphs as the construction only involves one cycle in each graph. For the second equality we use the fact that $\mathfrak{m}_q \le C^q$ as proved in the next subsection.

The second case consists in identifying a pair of *red* vertices in each graph. Such a pair is chosen in *one* fundmaental cycle in both $\mathcal{G}_1$ and $\mathcal{G}_2$. Then we identify the pair from one graph to the other pair. This allows the existence of edges belonging to the two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. The whole graph $\mathcal{G}$ created by this construction is non admissible as we have two identifications and two fundamental cycles. We thus need to choose the fundamental cycles in $\mathcal{G}$. The fundamental cycles we choose for this *red* graph are given by the cycles between the two vertices with edges belonging to both graphs in each cycle.

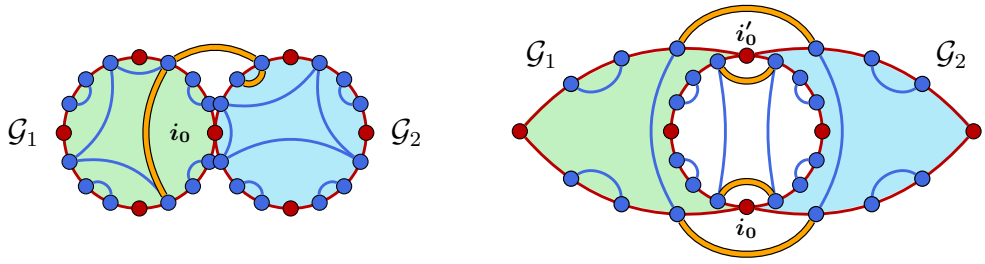
Since we need to choose a pair of vertices in each graph we have q^4 choices. In each fundamental cycles, we perform the typical *blue* matching and we have an edge between a niche from $\mathcal{G}_1$ and a niche from $\mathcal{G}_2$ (corresponding to the cycle going over every niche for instance). Thus we have a common W or X entry between the two graphs and the contribution to the covariance does not vanish. Considering the q^4 choices for the *red* vertices, we can see that we have

$$\frac{1}{n_1^2}\sum_{\mathcal{G}_1,\mathcal{G}_2}\sum_{\boldsymbol{\ell_1},\boldsymbol{\ell_2}}\mathrm{Cov}^{(1)}(M_{\mathcal{G}_1}(\boldsymbol{\ell_1}),M_{\mathcal{G}_2}(\boldsymbol{\ell_2}))=\mathcal{O}\left(\frac{q^4C^q}{n_1^2}\right).$$

Regarding the number of possible choices for the vertices, the number of l-indices is unchanged while that of $i-$ or j-indices decreases of 2 if we compare to the computation of the expected moment. Finally, we obtain that

$$\mathrm{Var}\, m_q=\mathcal{O}\left(\frac{(q^2k^2+q^4)C^q}{n_1^2}\right).$$

Using Bienaymé-Chebyshev inequality, one easily deduces (3.19). □



(a) In this figure, the two highlighted cycles correspond to the graph $\mathcal{G}_1$ and $\mathcal{G}_2$ which are attached at i_0. We perform a typical *blue* matching in each graph and then add an identification between the two graphs. The highlighted orange edges correspond to the edges common to the two graphs, which yields a moment of order 4.

(b) In this figure, the two highlighted cycles correspond to the graph $\mathcal{G}_1$ and $\mathcal{G}_2$ which are attached at two vertices i_0 and i_0'. The graph is non-admissible and we choose the fundamental cycles so that neither $\mathcal{G}_1$ or $\mathcal{G}_2$ are fundamental cycles. The typical matching in the chosen cycles create common edges between the two graphs highlighted in orange on the figure.

3.5 From bounded to sub-Gaussian random variables

We have computed the limiting expected moments in the case of bounded random variables. However, note that while high moments of W or X can appear in the error terms, as in (3.8), (3.10) and (3.15), one may use for such sub-Gaussian random variables the following bound

$$\mathbb{E}\left[|X_{11}|^k\right]\leqslant C^k k^{k/\alpha},\quad \mathbb{E}\left[|W_{11}|^k\right]\leqslant C^k k^{k/\alpha},$$

for some constant C. Thus one may simply replace in all the error terms A by $k^{1/\alpha}$. Since k is of order $\frac{\log n_1}{\log\log n_1}$ all the errors are still $o(1)$.

3.6 Weak convergence of the empirical spectral measure

In this section we briefly finish the proof of Theorems 2.2 and 2.3 for a polynomial activation function. The fact that the sequence of moments

$$\mathfrak{m}_q:=\sum_{I_i,I_j=0}^{q}\sum_{b=0}^{I_i+I_j+1}\mathcal{A}(q,I_i,I_j,b)\theta_1(f)^b\theta_2(f)^{q-b}\psi^{I_i+1-q}\phi^{I_j}\tag{3.20}$$

uniquely defines a probability measure μ so that $\int x^q \mathrm{d}\mu(x) = \mathfrak{m}_q$ follows from Carleman's condition. Indeed, denote by $\Theta(q)$ the number of unlabeled cactus graphs with q vertices. It has been shown in [16] that, regardless of the number of identifications or simple cycles, there exists numerical constants $\delta > 0$ and $\xi > 1$ such that $\Theta(q) \sim \frac{3\delta}{4\sqrt{\pi}} \frac{\xi^{q+3/2}}{q^5}$. Thus there exists a constant C such that $\mathfrak{m}_q \leqslant C^q$. This can also been used to show that the measure has compact support.

3.7 Derivation of the self-consistent equation for the Stieltjes transform

Consider the Stieltjes transform of the limiting empirical eigenvalue distribution of M,

$$G(z) = \int \frac{\mathrm{d}\mu(x)}{x - z}.$$

One can also write it as the following generating function of moments, since the following equality makes sense at least on a neighborhood of infinity,

$$-G(z) = \frac{1}{z} + \sum_{q=1}^{\infty} \frac{\mathfrak{m}_q}{z^{q+1}}.$$

Using that

$$\mathfrak{m}_q = \psi^{1-q} \sum_{I_i, I_j = 0}^{q} \sum_{b=0}^{I_i + I_j + 1} \mathcal{A}(q, I_i, I_j, b) \theta_1^b(f) \theta_2^{q-b}(f) \psi^{I_i} \phi^{I_j},$$

one can write the Stieltjes transform as

$$-G(z) = \frac{1 - \psi}{z} + \frac{\psi}{z} H(z)$$

with

$$H(z) = \sum_{q=0}^{\infty} \frac{1}{(\psi z)^q} \sum_{I_i, I_j = 0}^{q} \sum_{b=0}^{I_i + I_j + 1} \mathcal{A}(q, I_i, I_j, b) \theta_1^b(f) \theta_2^{q-b}(f) \psi^{I_i} \phi^{I_j}.$$

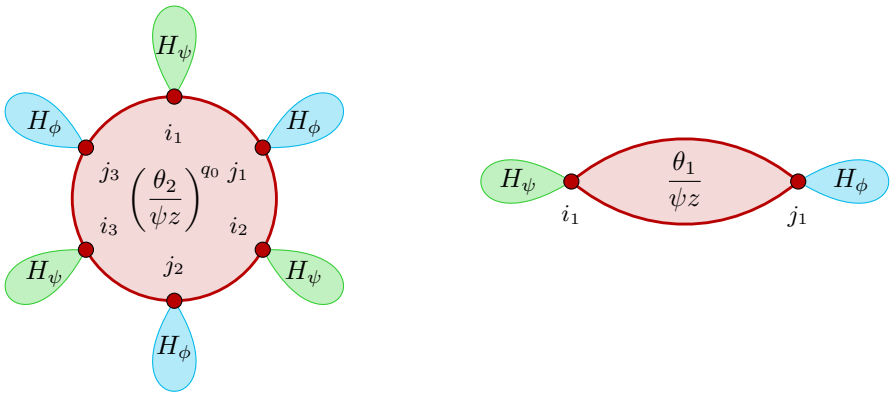
Fix a vertex v and denote q_0 the length of one of the fundamental cycles containing v. Suppose first that we have $q_0 > 1$, this cycle contains $2q_0$ edges with q_0 vertices labeled with i and q_0 vertices labeled with j. On each vertex labeled with i, either a graph is attached and we have a i-identification on this vertex, or nothing is attached. Thus, considering the formula above, we have that the contributions for identifications for each vertex is

$$H_\psi(z) := 1 - \psi + \psi H(z) \quad \text{for } i\text{-labels and} \quad H_\phi(z) := 1 - \phi + \phi H(z) \quad \text{for } j\text{-labels.}$$

Also, one can see in the leading order of the moment that a cycle of length q_0 give a contribution of $\left(\frac{\theta_2(f)}{\psi z}\right)^{q_0}$. Now, if the cycle is of length 1, in the same way, there is a single i-labeled vertex and a single j-labeled vertex which can give a contribution of H_ψ and H_ϕ but the contribution of a simple cycle is not given in terms of $\theta_2(f)$ but by $\frac{\theta_1(f)}{\psi z}$. This is illustrated in Figure 14. Thus, we have the following recursion relation for H,

$$\begin{aligned} H(z) = 1 + \frac{H_\phi(z) H_\psi(z) \theta_1}{\psi z} + \sum_{q_0 = 2}^{\infty} \left(\frac{H_\phi(z) H_\psi(z) \theta_2}{\psi z} \right)^{q_0} \\ = 1 + \frac{H_\phi(z) H_\psi(z) (\theta_1 - \theta_2)}{\psi z} + \frac{H_\phi(z) H_\psi(z) \theta_2}{\psi z - H_\phi(z) H_\psi(z) \theta_2}. \end{aligned}$$

Note that we obtain the final equation from Theorem 2.3 by noting that $H_\psi(z) = -zG(z)$ and $H_\phi(z) = -z\tilde{G}(z)$.



(a) Case of a large cycle ($q_0 = 3$)

(b) Case of a simple cycle.

Figure 14: Illustration of the recursion for the derivation of the self-consistent equation.

4 Proof of Theorem 2.2 for general activation function

In this section, we now allow the activation function to belong to a wider class, thus proving Theorem 2.2. For ease, we assume that $\sigma_w = \sigma_x = 1$, which can be achieved by scaling.

Proof of Theorem 2.2. We begin by defining the following polynomial which approximates f up to a constant, for $x \in \mathbb{R}$ we define

$$P_k(x) := \sum_{j=1}^{k} f^{(j)}(0)\frac{x^j - j!!}{j!} = \sum_{j=0}^{k} f^{(j)}(0)\frac{x^j}{j!} - a_n \quad \text{with} \quad a_n = \sum_{j=0}^{k} f^{(j)}(0)\frac{j!!}{j!} \tag{4.1}$$

with the convention that $j!! = 0$ for j odd and $0!! = 1$. This choice ensures that the polynomial is centered with respect to the Gaussian distribution. Thus, using Taylor's theorem, we obtain the following approximation for any $A > 0$

$$\sup_{x\in[-A,A]} |(f(x) - a_{k-1})(x) - P_{k-1}(x)| \leqslant C_f \frac{A^{(1+c_f)k}}{k!}. \tag{4.2}$$

Now, we compare the Hermitized version of the matrix M (up to finite rank modification), and define

$$\begin{aligned} Y^{(a_k)} &= f\left(\frac{WX}{\sqrt{n_0}}\right) - a_k, \quad Y_k = P_k\left(\frac{WX}{\sqrt{n_0}}\right), \\ \mathcal{E} &= \frac{1}{\sqrt{m}} \begin{pmatrix} 0 & Y^{(a_{k-1})} - Y_k \\ \left(Y^{(a_{k-1})} - Y_k\right)^* & 0 \end{pmatrix}. \end{aligned} \tag{4.3}$$

We want to control the spectral radius of the $(m + n_1) \times (m + n_1)$ symmetric matrix $\mathcal{E}$. Now consider the event, for $\delta_1 \in (0, \frac{1}{2})$,

$$\mathcal{A}_{n_1}(\delta_1) = \bigcap_{1\leqslant i\leqslant n_1} \bigcap_{1\leqslant j\leqslant m} \left\{ \left| \left(\frac{WX}{\sqrt{n_0}}\right)_{ij} \right| \leqslant (\log n_1)^{1/2+\delta_1} \right\}. \tag{4.4}$$

On this event, we have, considering the approximation (4.2),

$$\rho(\mathcal{E}) \leqslant C_f \sqrt{m} \frac{(\log n_1)^{k(1/2+\delta_1)(1+c_f)}}{k!}.$$

We then choose

$$k \geqslant c_0 \frac{\log n_1}{\log\log n_1} \quad \text{with} \quad c_0 > \frac{1}{2(1-(1+c_f)(\frac{1}{2}+\delta_1))}. \tag{4.5}$$

We obtain, by using Stirling formula, that there exists a $\delta_2 > 0$ such that for any $\varepsilon > 0$ we have

$$\rho(\mathcal{E}) = \mathcal{O}\left(\frac{n_1^{\varepsilon}}{n_1^{\delta_2}}\right).$$

By taking ε small enough we then see that, on the event $\mathcal{A}_{n_1}(\delta_1)$ and with k as in (4.5), $\rho(\mathcal{E}) \to 0$ as $n_1 \to \infty$. It remains to see that the event $\mathcal{A}_{n_1}(\delta_1)$ occurs with high probability which comes from the assumption on the entries W_{ij} and X_{ij}. Indeed,

$$\mathbb{P}\left(\mathcal{A}_{n_1}(\delta_1)^c\right) = \mathbb{P}\left(\exists\, i,j \text{ such that } \left|\left(\frac{WX}{\sqrt{n_0}}\right)_{ij}\right| > (\log n_1)^{1/2+\delta_1}\right) \leqslant C n_1 m e^{-\frac{(\log n_1)^{1+2\delta_1}}{2}} \tag{4.6}$$

which goes to zero faster than any polynomial in n_1. Now we know the limiting e.e.d. of the matrix M_{P_k} constructed with the centered polynomial P_k as activation function. The above argument yields it is the same for M_{f-a_k} constructed with $f - a_k$ instead. Now $Y^{(a_k)}$ is just a rank one deformation of Y and by the rank inequalities (see [4] for instance), M and M_{f-a_k} have the same limiting e.e.d.. This finishes the proof of Theorem 2.2. □

5 Propagation of eigenvalue distribution through multiple layers

In this section, we study the eigenvalue distribution of a nonlinear matrix model when the data passes through several layers of the neural network. The case of a single layer has been considered in Theorems 2.2 and 2.3 where we describe the asymptotic e.e.d. in the one layer case. It has been conjectured in [32] that the limiting e.e.d. is stable through the layers in the case where $\theta_2(f) = 0$. We give here a positive answer to this conjecture (with the appropriate normalization).

We first develop the combinatorial arguments for an odd monomial of the form

$$f(x) = \frac{x^k}{k!}. \tag{5.1}$$

for several layers. It can be shown as in Subsection 3.2 that the even monomial are subleading. Thus the leading order for moments is given by the contribution of odd monomial only. From now on, we assume (5.1) holds true. We can write the entries of the two layers data matrix $Y^{(2)}$ as

$$Y^{(2)}_{ij} = \frac{1}{k!}\left(\frac{\sigma_x}{\sqrt{\theta_1(f)}}\frac{W^{(1)}Y^{(1)}}{\sqrt{n_1}}\right)^k = \frac{\sigma_x^k}{n_1^{k/2}k!\theta_1(f)^{k/2}} \sum_{\ell_1,\dots,\ell_k=1}^{n_1} \prod_{p=1}^{k} W^{(1)}_{i\ell_p} Y^{(1)}_{\ell_p j}. \tag{5.2}$$

Then, developing the expected moment of the e.e.d. and using (5.2), we obtain the following

$$\begin{aligned}
&\frac{1}{n_2}\mathbb{E}\left[\mathrm{Tr}\left(M^{(2)}\right)^q\right] = \\
&= \frac{\sigma_x^{2kq}}{n_2 m^q n_1^{kq}(k!)^{2q}\theta_1(f)^{kq}}\mathbb{E}\sum_{i_1,\dots,i_q}^{n_2}\sum_{j_1,\dots,j_q}^{m}\sum_{\substack{\ell_1^1,\dots\ell_k^1\\ \cdots\\ \ell_1^{2q}\dots\ell_k^{2q}}}^{n_1}\prod_{p=1}^{k} W^{(1)}_{i_1\ell_p^1} Y^{(1)}_{\ell_p^1 j_1}\cdots\prod_{p=1}^{k} W^{(1)}_{i_1\ell_p^{2q}} Y^{(1)}_{\ell_p^{2q} j_q}.
\end{aligned} \tag{5.3}$$

We call the terms contributing in a non negligible way *typical*. Now, we can give a graphical representation of these terms as in the previous sections.We will see that the contributing graphs are actually the same admissible graphs from Definition 3.1. However, there are less constraints in the choices of the *blue* edges. Indeed, the entries of the matrix $Y^{(1)}$ are not independent: we do not need each entry to be matched with at least another. This constraint however holds for the entries of the matrix $W^{(1)}$.

5.1 The simpler case of the simple cycle

In this subsubsection, we explain the combinatorics in the case where the i-labels and j-labels are pairwise distinct. We first perform a matching on the entries of $W^{(1)}$. This matching on the $W^{(1)}$ entries induces one on the entries of $Y^{(1)}$. This matching thus induces another graph between j-labeled and ℓ-labeled vertices. The i-labeled vertices do not appear in the graph (as they correspond to entries of $W^{(1)}$). This graph can be constructed from the initial graph by seeing which niches are connected by a *blue* edge. Figure 15 explains this construction: ℓ_2 links the same niche adjacent to j_2 while ℓ_1 links the niches adjacent to j_1 and j_2.

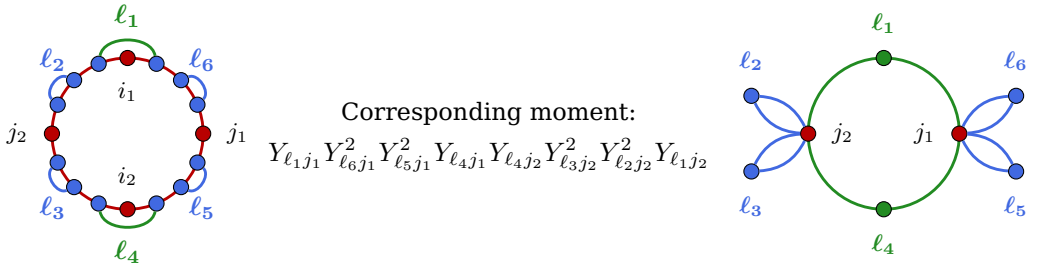


Figure 15: Graph obtained after a *blue* matching in the initial graph. The *green* edges, corresponding to bridges between niches, induce a cycle in the final graph. The remaining edges coming from matched pairs inside a niche create simple cycles attached to j labeled indices.

We start with general observations. The largest number of possible distinct ℓ indices is kq, which is obtained as follows: One matches at least two indices from different adjacent niches of an i-label index and perform a perfect matching between the $2k-2$ remaining indices. Such a matching gives kq different ℓ indices and matches every $W^{(1)}$ entry with another. This is illustrated in the leftmost graph in Figure 15. Note that this type of matching gives kq distinct ℓ indices but is actually not necessarily typical (see Figure 16 for an illustration) and is not the sole typical configuration.
As in Figure 15, we see that the matching on the initial graph induces another admissible graph. Note that it does not consist in one cycle but in a cycle (in green on the figure) where $k-1$ cycles of length 2 are attached to each j-labeled vertex. Also one has to note that it is possible to perform identifications between the *blue* edges and obtain a graph contributing in a non negligible way to the asymptotic expansion (see Figure 17 for an illustration). This behavior is explained in the second step when we develop the entries of $Y^{(1)}$. Let us briefly indicate, as in Figure 16, a *blue* matching on the initial

cycle which maximizes the number of distinct indices may give rise to a non-admissible induced graph. This comes from the fact that too many edges link two distinct niches.

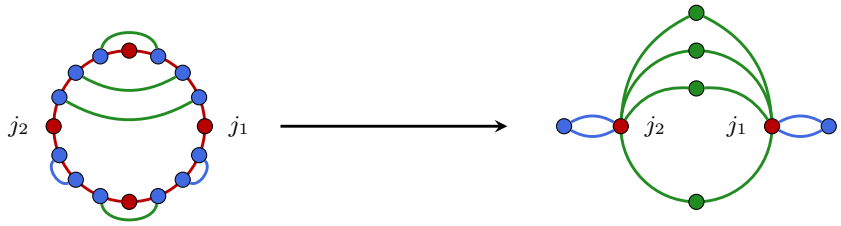


Figure 16: Non-admissible graph obtained after a *blue* matching which induces a maximum number of distinct indices in the initial cycle. We can see that several *green* bridges between the same niches create a non-admissible graph and is thus subleading via the analysis from the previous section.

The main tool to understand the combinatorial arguments for the multilayer case is the following Lemma. It states that the leading order is actually given by the matchings as in Figure 15.

Lemma 5.1. *Consider a cycle of length $q > 2$, then the typical matchings on the* blue *vertices consist in the following:*

i) Two niches adjacent to the same i-labeled vertex are linked by a single edge called a bridge.

ii) Remaining edges inside a niche are matched according to a perfect matching.

iii) We can add identifications between bridges *only.*

If the cycle is of length 2 then we perform a perfect matching between the $2k$ blue *vertices in the cycle.*

Proof. The proof is based on the construction of the second graph and the fact that the typical graphs are admissible. We first show that any other matching gives a non-admissible second layer graph. Firstly, more than one *bridge* between two distinct niches breaks the tree structure and thus yields a non admissible graph. The same reasoning holds for possible identifications between bridges and a matched pair inside a niche. If we identify two matched pairs inside a niche, we can see via the construction of the graph that it creates double edges and we would obtain an entry of $Y^{(1)}$ to the power of 4. However, note that in the initial cycle of size q, we can add identifications between the q *bridges* and still keep the second graph admissible. This behavior is illustrated in Figure 17 where we perform identifications between bridges and still obtain an admissible graph.

We now need to show that the contribution of the matchings leading to a non admissible graph is subleading. As in Subsection 3.1.4, we have additional identifications between the vertices and we need to choose the fundamental cycles as well as the way one runs through the graph. Suppose we have I_ℓ identifications between the ℓ vertices. Then if the graph was admissible we would have $I_\ell + q(k-1) + 1$ fundamental cycles in the induced graph on (j, ℓ) vertices. Thus, if the graph is non-admissible, we have at most $I_\ell + q(k-1)$ fundamental cycles.

Let $m \leqslant I_\ell + q(k-1)$ be the number of fundamental cycles of the induced graph. We denote by $\mathscr{C}_1, \dots, \mathscr{C}_b, \mathscr{C}_{b+1}, \dots, \mathscr{C}_m$ its cycles such that if $\ell(\mathscr{C}_i)$ denotes the length of the cycle $\mathscr{C}_i$ we have: $\ell(\mathscr{C}_1) = \dots = \ell(\mathscr{C}_b) = 2$ and $\ell(\mathscr{C}_{b+1}), \dots, \ell(\mathscr{C}_m) > 2$. One then has that $\sum_{i=1}^m \ell(\mathscr{C}_i) = 2kq$. Now, the contribution of such graphs (initial and induced), taking into

account the normalization and the number of ways to run through the graph, is at most

$$\frac{(1+o(1))C^{I_l-m+I_j}}{n_1 m^q n_o^{kq+k^2q}} n_1^q m^q n_0^{kq-I_\ell} n_0^{kb} n_0^{m-b+\frac{k-1}{2}\sum_{i=b+1}^m \ell(\mathscr{C}_i)} = \mathcal{O}\left(\left(\frac{C}{n_0}\right)^{(I_\ell+(k-1)q+1)-m}\right),$$

for some constant C. Indeed one has to choose the i- and j-labels of vertices in the initial cycle, the ℓ vertices in the initial graph with the constraint that there are I_ℓ identifications. Then, in the induced graph, there are at most k indices in each cycle of length 2 and $1+(k-1)/2\ell(\mathscr{C}_i)$ indices in the cycle $\mathscr{C}_i$ for $i > b$. Thus the contribution is negligible due the constraint that $m \leqslant I_\ell + q(k-1)$.

Some (negligible) contribution depending on k comes from the possible multiple cycles of length 2 attached together as in Figure 9. Here the error is slightly bigger since the induced graph has $2kq$ edges instead of simply $2q$. Fix a vertex j_0, if we match together $2p$ ℓ-indices together in the niche adjacent to j_0, using (3.16), the corresponding error is given by $\mathcal{O}(n_0(k(2p)^k/n_0)^p)$. However, up to $2k$ indices can be matched together so that the contribution of non admissible graphs in this case is given by

$$\sum_{p=2}^{k} n_0 \left(\frac{k(2p)^k}{n_0}\right)^p = o(1) \quad \text{for} \quad k \leqslant \frac{\log n}{\log\log n}.$$

It actually decays faster than any polynomial for such k. This finishes the proof of the Lemma. □

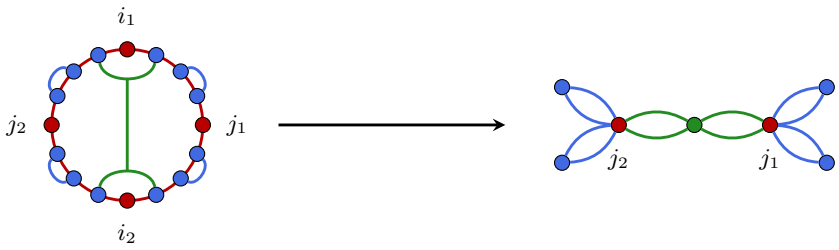


Figure 17: Admissible graph after a matching with an identification between two *bridges*. While two ℓ vertices are identified, the matching is of leading order as one more cycle is in the induced graph.

Lemma 5.1 has been proved for the two layers case. It can readily be extended to the case of $L \geq 2$ layers: the number of possible distinct l-indices is multiplied by k at each layer and the final graph after performing matchings has to be admissible so that it contributes in the limit. The proof is similar to that of Lemma 5.1. The detail is left to the reader.

5.2 Invariance of the distribution in the case when $\theta_2(f)$ vanishes.

In light of the previous combinatorial arguments, it is interesting to consider the special case where $\theta_2(f) = 0$. Indeed, for the one layer case, by Theorem 2.2, the limiting e.e.d. is the Marčenko-Pastur distribution with shape $\frac{\phi}{\psi}$, denoted by $\mu_{\phi/\psi}$, as proved also by the following lemma.

Lemma 5.2. *Let q be a positive integer we have the following equality*

$$\sum_{\substack{I_j,I_i=0\\ I_i+I_j+1=q}}^{q-1} \mathcal{A}(q,I_i,I_j,q)\psi^{1-q+I_i}\phi^{I_j}\theta_1^q(f) = \theta_1^q(f)\sum_{k=0}^{q-1}\left(\frac{\phi}{\psi}\right)^k \frac{1}{k+1}\binom{q}{k}\binom{q-1}{k} = \theta_1^q(f)\langle x^q, \mu_{\phi/\psi}\rangle.$$

Proof. Firstly, we can slightly rewrite the left hand side as

$$\sum_{\substack{I_j,I_i=0\\ I_i+I_j+1=q}}^{q-1} \mathcal{A}(q,I_i,I_j,q)\psi^{1-q+I_i}\phi^{I_j}\theta_1^q(f) = \theta_1^q(f)\sum_{k=0}^{q-1}\left(\frac{\phi}{\psi}\right)^k \mathcal{A}(q,q-k-1,k,q).$$

Now there only remains to see that

$$\mathcal{A}(q,q-k-1,k,q) = \frac{1}{k+1}\binom{q}{k}\binom{q-1}{k}. \tag{5.4}$$

This fact comes from another representation of admissible graphs. Consider admissible graphs with $2q$ edges, q cycles of length 2, k j-identifications and $q-k-1$ i-identifications. Thus we can count this as double trees, in the sense that one of every two vertices are i-labeled and the others are j-labeled, with the appropriate number of each type of vertex ($q-k$ j-labeled vertices and $k+1$ i-labeled vertices). This number is known as a Narayana number [7] and given by (5.4). □

This fact then means that if we consider a function f such that $\theta_2(f)=0$, the e.e.d. (up to a change in variance and shape) is "stable" after going through one layer of the network. Indeed, if one considers the matrix $\frac{1}{m\sigma_x^2}XX^*$, the asymptotic e.e.d. is given by μ_ϕ the Marčenko-Pastur distribution with shape parameter ϕ. Now, after a layer of the network, we see that for $\frac{1}{m\theta_1(f)}YY^*$ it is given by $\mu_{\phi/\psi}$.

We now consider the case of an arbitrary but fixed number of layers, mostly interested in the case where $\theta_2(f)=0$. Let $Y^{(L+1)}$ be as in (2.9), and consider the matrices

$$M^{(L+1)} = \frac{1}{m\theta_1(f)}Y^{(L+1)}Y^{(L+1)^*}. \tag{5.5}$$

Theorem 5.3. *Let L be a given integer. Let $f=\sum_{k=1}^K \frac{a_k}{k!}(x^k - k!!\mathbb{1}_{k \text{ even}})$ be a polynomial such that* (2.4) *holds and $\theta_2(f)=0$. The degree of f, K, can grow with n_1 but suppose that $K \leqslant \frac{1}{L-1}\frac{\log n_1}{\log\log n_1}$. Denote the e.e.d. of $M^{(L)}$ constructed as in* (5.5) *by $\mu_{n_L}^{(L)} = \frac{1}{n_L}\sum_{i=1}^{n_L}\delta_{\lambda_i^{(\ell)}}$ and its expected moments by $\overline{m}_q^{(L)} := \mathbb{E}\left[\langle \mu_{n_L}^{(L)}, x^q\rangle\right]$, then*

$$\overline{m}_q^{(L)} = \left(\sum_{k=1}^{q-1}\left(\frac{\phi}{\prod_{i=0}^{L-1}\psi_i}\right)^k \frac{1}{k+1}\binom{q}{k}\binom{q-1}{k}\right)(1+o(1)). \tag{5.6}$$

Proof. We again first develop the arguments in the case of a monomial of odd degree $f(x)=x^k$ since the case of an even monomial is completely similar (we only consider graphs with simple cycles).We study and count the admissible graphs along each layer. It is enough to identify in the asymptotic expansion of the moment those terms where no θ_2 arises. Thus one can consider only admissible graphs made of cycles of length 2. For the error terms one has to consider also admissible graphs with longer cycles but where the matching in each niche does not yield an occurence of θ_2.

We begin with the case where $q = 1$ for two layers. Then the cycle has length 2 as in Figure 4a. The dominant term in the asymptotic expansion consists in performing a perfect matching between all edges from Lemma 5.1. The contribution coming from this first construction (in Lemma 5.1) is given by

$$\frac{\sigma_x^{2k}}{n_2 m n_1^k} n_2 m n_1^k \left(\sigma_w^{2k}(2k)!!\right) = \theta_1(f)(1+o(1)).$$

This follows from the choices for the i index, the j index and the ℓ indices. Now, this construction on the initial graph induces a second graph as in Figure 18. This induced graph is an admissible graph where all j's are identified to a single vertex and k cycles of length 2 are attached to it (corresponding to the k *blue* edges in the initial cycle). We use the same reasoning as before and develop the entries $Y^{(1)}$ as a product of entries of $W^{(0)}$ and X. Since the graph is admissible, the dominant term in the asymptotic expansion corresponds to performing a perfect matching in all cycles of length 2 as in Section 3 (illustrated in Figure 18). Thus this adds a contribution of

$$\frac{1}{n_0^{k^2}\theta_1(f)^k} n_0^{k^2} \left(\sigma_w^{2k}\sigma_x^{2k}(2k)!!\right)^k = 1 + o(1).$$

Here, the normalization in $n_0^{-k^2}$ comes from the fact there are $2k$ entries with a normalization of $n_0^{-k/2}$. We then have to choose $n_0^{k^2}$ indices in the second graph. Finally, we obtain for the final contribution for a cycle of length 2 that $E_1(f) = \theta_1(f)\,(1+o(1))$.

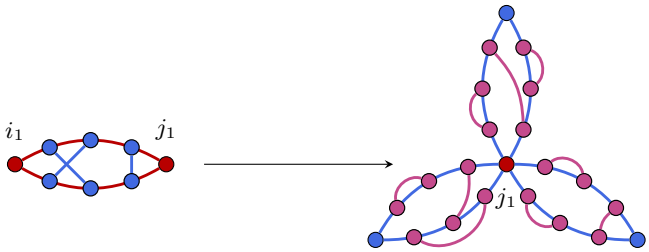


Figure 18: Construction/matching on the second layer graphs from a matching on the initial graph. The first graph gives a combinatorial factor of $(2k)!!$ while the second graph gives a factor of $(2k)!!^k$.

For the general case we saw that the first step of the procedure (by the construction explained before) yields a forest of *star admissible graph* where each graph is given by a certain number of cycles of length 2 attached to a unique j-labeled vertex. Consider now a connected component (of the induced forest) which corresponds to a unique j vertex. The number of cycles of length 2 attached to j is then k times the total number of cycles adjacent to j in the previous steps (since we have k *blue* edges in each cycle of length 2). From this first process we then get the following contribution for this first two steps

$$\frac{\sigma_x^{2kq+2k^2q}(1+o(1))}{n_L m^q \theta_1(f)^{q+kq+k^2q}} \sum_{\substack{I_i,I_j \\ I_i+I_j+1=q}} \mathcal{A}(q,I_j,I_j,q) n_L^{q-I_i} \frac{m^{q-I_j}}{n_{L-1}^{kq}} n_{L-1}^{kq} \left(\sigma_w^{2k}(2k)!!\right)^q \times$$

$$\times \frac{1}{n_{L-2}^{k^2q}} n_{L-2}^{k^2q} \left(\sigma_w^{2k}(2k)!!\right)^{kq} = (1+o(1)) \sum_{k=0}^{q-1} \mathcal{A}(q,q-k-1,k,q) \left(\frac{n_L}{m}\right)^k.$$

Let us explain the above formula: there are $n_L^{q-I_i}$ choices needed to label the i-labeled vertices and m^{q-I_j} for the j-labeled vertices. For the powers of n_{L-1} we take into

account the normalization and the corresponding number of ℓ indices to choose. Finally in each cycle of length 2 we perform a perfect matching between the two niches: there are q cycles of length 2 in the initial graph and kq such cycles in the forest obtained. See Figure 19 for an illustration.

Now, we can perform one more step of the procedure, we now have a forest of these *star admissible graphs* where each graph has only one j vertex. To the j vertex are now attached k times more cycles than in the previous step. Thus, for the 3 step procedure, the total number of cycles of length 2 in the forest is given by $k^3 q$. We can perform this for each layer the data goes through as the only parameter to be changed is the number of cycles of length 2 attached to each j vertex.

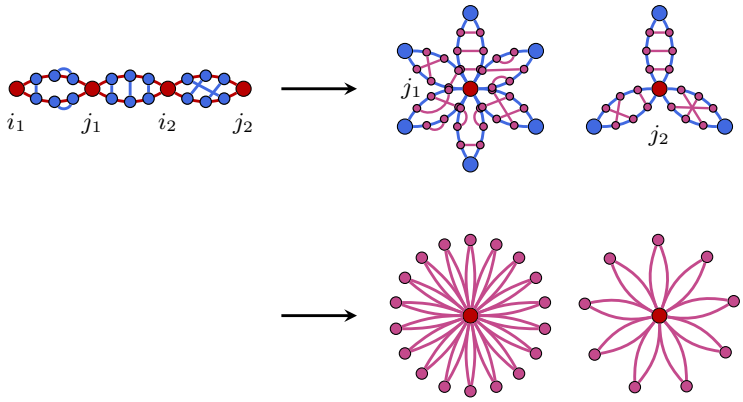


Figure 19: Effect on going through several layers for admissible graphs with only cycles of length 2. The first step consists of separating each j-labeled vertex into his own graph where it is attached to cycles of length 2. At each layer after the first one, we multiply by k the number of cycles attached.

In the whole, adding the layer L_0 multiplies the contribution with no θ_2 by a factor

$$\frac{1}{n_{L_0}^{k^{L-L_0}q}\theta_1^{k^{L-L_0}q}} n_{L_0}^{k^{L-L_0}q}\theta_1^{k^{L-L_0}q}(f).$$

Thus the whole contribution can be written in the following way:

$$(1+o(1))\sum_{k=0}^{q-1}\left(\frac{n_L}{m}\right)^k \mathcal{A}(q,q-k-1,k,q).$$

And we obtain the final result by using that $\frac{n_L}{m} \to \frac{\phi}{\psi_0\psi_1\ldots\psi_{L-1}}$ and $\mathcal{A}(q,q-k-1,k,q) = \frac{1}{k+1}\binom{r}{k}\binom{r-1}{k}$.

Now, in the statement of the theorem we do not explicit the leading contribution of admissible graphs with at least one cycle of length greater than 2. We only need now to get an estimate on the other possible errors and show that they are negligible. Using Subsection 3.6, the total number of cactus trees with q edges does not exceed $C\xi^q$ for some constant C. As we are interested in the case where θ_2 vanishes, it is enough to show that the error terms cannot grow faster than the Marcenko-Pastur moment. Actually using the arguments of Section 3, the whole analysis of errors remains true. The errors can only come from subleading matchings on the graph at each possible step. However, the main difference comes from the number of vertices at each step which is $k^{L_0}q$ instead of just kq. Note that it still only consists of a power of k which grows slower than any power of n_1. Again, the leading contribution of the errors comes from

possible multiple edges arising in the graph. Say that a given j vertex is first connected to r cycles of length 2 in the initial graph. At the step L_0, it is now connected to $k^{L_0-1}r$ cycles of length 2. Thus if at this stage we connect *blue* indices together, say p of them we obtain at the next step a multiple edge of multiplicity $2p$. We have a total of $2k^{L_0}r$ *blue* indices to match at this stage since we have $2k$ vertices per cycle of length 2. Thus, by comparing the contribution of such matchings with the typical matching we obtain, similarly to (3.16),

$$\sum_{p=2}^{k^{L_0}r} n_0 \left(\frac{Ckp^k}{n_0}\right)^p = o(1) \quad \text{for} \quad k \leqslant \frac{1}{L_0}\frac{\log n_1}{\log\log n_1}.$$

Now L_0 ranges from 1 to $L-1$ so that we obtain the needed bound if $k \leqslant \frac{1}{L-1}\frac{\log n_1}{\log\log n_1}$. □

We now finish the proof of Theorem 2.5.

Proof of Theorem 2.5. We have shown that for a polynomial of degree up to $\frac{1}{L-1}\frac{\log n_1}{\log\log(n_1)}$, the expected moments of the e.e.d. are those of the Marčenko-Pastur distribution with the appropriate shape parameter. We first see that the variance of the moments is of order k^L/n_1^2 in order to show convergence of the actual moments. The principle is similar to that of Lemma 3.10 as we count the corresponding graphs such that their covariance is non zero.

We can perform the same expansion as in Lemma 3.10 and see that we have for the first layer

$$\operatorname{Var} m_q^{(L)} = \frac{1}{n_1^2}\sum_{\mathcal{G}_1,\mathcal{G}_2}\sum_{\boldsymbol{\ell_1},\boldsymbol{\ell_2}} \mathbb{E}\left[M_{\mathcal{G}_1}^{(L)}(\boldsymbol{\ell_1})M_{\mathcal{G}_2}^{(L)}(\boldsymbol{\ell_2})\right] - \mathbb{E}\left[M_{\mathcal{G}_1}^{(L)}(\boldsymbol{\ell_1})\right]\mathbb{E}\left[M_{\mathcal{G}_2}^{(L)}(\boldsymbol{\ell_2})\right] \tag{5.7}$$

with

$$M_{\mathcal{G}}^{(L)}(\boldsymbol{\ell}) = \sum_{k_1,\dots,k_{2q}=1}^{K} \frac{a_{k_1}\dots a_{k_{2q}}}{m^q n_0^{\sum k_i/2}} \prod_{p=1}^{k_1} W_{i_1\ell_p^1}^{(L)} Y_{\ell_p^1 j_1}^{(L)} \prod_{p=1}^{k_2} W_{i_2\ell_p^2}^{(L)} Y_{\ell_p^2 j_1}^{(L)} \cdots \prod_{p=1}^{k_{2q}} W_{i_1\ell_p^{2q}}^{(L)} Y_{\ell_p^{2q} j_q}^{(L)}.$$

Now, in order to have a non vanishing contribution to the variance (5.7), we need to have additional identifications between the two graphs. Indeed, either at a given layer L_0 an entry of $W^{(L_0)}$ is matched between $\mathcal{G}_1$ and $\mathcal{G}_2$ or at the last layer there are identifications between the X entries. In the case where there are identifications of $Y^{(L_0)}$ entries we see, by expanding the expansion with respect to the entries of $W^{(L_0-1)}$, that this implies that there are further identifications in the layers beyond L_0. Since at each step we would lose an order $\mathcal{O}(q^2(k)^{2L_0})/n_0)$ (from the choice of which vertices to identify and the fact that we have one less choice for possible indices), we see that the leading order comes from identifying X entries in the two last layers.

Thus, since the main contribution to moments are still given by admissible graphs, a similar analysis can be done as in Lemma 3.10: we can, right at the first layer, identify i and j vertices to obtain an identification on the $W^{(L)}$ entries. Or one can choose two $W^{(L_0)}$ entries to be identified at a given layer L_0 (or X entries at the last layers $L_0 = 1$) and thus we obtain

$$\operatorname{Var} m_q^{(L)} = \mathcal{O}\left(\frac{q^4 + q^2\sum_{L_0=1}^{L} k^{2L_0} + \sum_{L_0=1}^{L} k^{4L_0}}{n_0^2} C^q\right) = \mathcal{O}\left(\frac{k^{4L+4}}{n_0^2}\right),$$

since q is fixed here.

Let us now extend the result to a bounded function f. As in Section 4, we consider a polynomial P_k such that, for some $A > 0$, $\sup_{x\in[-A,A]} |(f(x) - a_k) - P_k(x)| \leqslant C_f \frac{A^{(1+c_f)k}}{(n+1)!}$.

Now, we can consider $Y^{(L,a_k)}$ the matrix constructed as (2.9) with $f - a_k$ as an activation function and $Y^{(L,P_k)}$ the same matrix constructed with P_k. Note that we consider the same sampling of W and X for the construction of this model. We describe the case of $L = 2$ as we can recursively do the same reasoning for a higher number of layers, for simplicity we also forget the change of variance $\sigma_x/\sqrt{\theta_1(f)}$ at each layer. As we saw in Section 4, we simply need to bound

$$\frac{1}{\sqrt{m}} \max_{1\leqslant i\leqslant n_2} \sum_{j=1}^{m} \left| Y_{ij}^{(2,a_k)} - Y_{ij}^{(2,P_k)} \right| \\ = \frac{1}{\sqrt{m}} \max_{1\leqslant i\leqslant n_2} \sum_{j=1}^{m} \left| f\left(\frac{W^{(1)}Y^{(1,a_k)}}{\sqrt{n_1}} \right)_{ij} - a_k - P_k\left(\frac{W^{(1)}Y^{(1,P_k)}}{\sqrt{n_1}} \right)_{ij} \right|.$$

We split the right hand side into two parts and write

$$\left| Y_{ij}^{(2,a_k)} - Y_{ij}^{(2,P_k)} \right| \leqslant \left| f\left(\frac{W^{(1)}Y^{(1,a_k)}}{\sqrt{n_1}} \right)_{ij} - f\left(\frac{W^{(1)}Y^{(1,P_k)}}{\sqrt{n_1}} \right)_{ij} \right| \tag{5.8}$$

$$+ \left| f\left(\frac{W^{(1)}Y^{(1,P_k)}}{\sqrt{n_1}} \right)_{ij} - a_k - P_n\left(\frac{W^{(1)}Y^{(1,P_k)}}{\sqrt{n_1}} \right)_{ij} \right|. \tag{5.9}$$

For the first term on the right hand side of the previous equation, we bound it from the polynomial approximation. Indeed, we consider the following event

$$\mathcal{A}_1(\delta_1) = \bigcap_{i=1}^{n_1} \bigcap_{j=1}^{m} \left\{ \left| \left(\frac{W^{(0)}X}{\sqrt{n_0}} \right)_{ij} \right| \leqslant (\log n_1)^{1/2+\delta_1} \right\} \bigcap \left\{ \left| W_{ij}^{(1)} \right| \leqslant (\log n)^{1/\alpha+\delta_1} \right\}.$$

This event occurs with overwhelming probability for any $\delta_1 > 0$ in the sense that its probability decays faster than any polynomial. Now, on this event we can bound

$$\left| \left(\frac{W^{(1)}Y^{(1,a_k)}}{\sqrt{n_1}} \right)_{ij} - \left(\frac{W^{(1)}Y^{(1,P_n)}}{\sqrt{n_1}} \right)_{ij} \right| \leqslant C^n \sqrt{n_1} (\log n_1)^{1/\alpha+\delta_1} \frac{(\log n_1)^{(1/2+\delta_1)n}}{n!},$$

where we expand the entries and use the polynomial approximation. This also decays faster than any polynomial for $n = \mathcal{O}(\frac{\log n_1}{\log\log n_1})$. Finally, using the fact that f has a bounded derivative on the event $\mathcal{A}_2(\delta_2)$ defined in (5.10), the first term in (5.8) goes to zero providing that $\mathcal{A}_2$ occurs with high probability.

For the second term in (5.8), by the previous analysis and as in Section 4 we only need to prove that the following event occurs with probability tending to one:

$$\mathcal{A}_2(\delta_2) = \bigcap_{i=1}^{n_2} \bigcap_{j=1}^{m} \left\{ \frac{1}{\sqrt{n_1}} \sum_{\ell_1=1}^{n_1} W_{i\ell_1}^{(1)} P_n\left(\frac{1}{\sqrt{n_0}} \sum_{\ell_0=1}^{n_0} W_{\ell_1\ell_0}^{(0)} X_{\ell_0 j} \right) \leqslant (\log n_1)^{1/2+\delta_1} \right\}. \tag{5.10}$$

Since we suppose that f is bounded we know that on the event $\mathcal{A}_1(\delta_1)$ (which occurs with very high probability) we have that $\sup_{ij} |Y_{ij}^{(1,P_k)}| \leqslant C$. Besides, since $W_{i\ell_1}^{(1)}$ has zero expectation, has a sub-Gaussian tail and is independent of the entries of $W^{(0)}$ and X, the random variable $(W^{(1)}Y^{(1)})_{ij}$ is sub-Gaussian as well. So that we obtain that there exists a $C > 0$ such that

$$\mathbb{P}\left(\sum_{\ell_1=1}^{n_1} W_{i\ell_1}^{(1)} P_n\left(\frac{1}{\sqrt{n_0}} \sum_{\ell_0=1}^{n_0} W_{\ell_1\ell_0}^{(0)} X_{\ell_0 j} \right) > \sqrt{n_1} (\log n_1)^{1/2+\delta_1} \right) \leqslant C e^{-c(\log n_1)^{1+2\delta_1}}.$$

And finally $\mathbb{P}(\mathcal{A}_2(\delta_2)) \geqslant 1 - n_1^{-D}$ for any $D > 0$. □

References

[1] G. Akemann and Z. Burda. Universal microscopic correlation functions for products of independent ginibre matrices. *J. Phys. A Math. Theor.*, **45**, (2012), no. 46, 465201. MR2993423

[2] G. Akemann, J. Ipsen, and M. Kieburg. Products of rectangular random matrices: singular values and progressive scattering. *Phys. Rev. E*, **88**, (2013), 052118.

[3] N. Alekseev, F. Götze, and A. Tikhomirov. On the asymptotics of the distribution of singular numbers of a power of a random matrix. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, **408**, (2012), 9–42. MR3032206

[4] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, 2 edition, 2010. MR2567175

[5] F. Benaych-Georges and R. Couillet. Spectral analysis of the gram matrix of mixture models. *ESAIM Probab. Stat.*, **20**, (2016), 217–237. MR3528625

[6] G. Cébron, A. Dahlqvist, and C. Male. Universal constructions for spaces of traffics. arXiv:1601.00168.

[7] W. Y. C. Chen, S. H. F. Yan, and L. L. M. Yang. Identities from weighted motzkin paths. *Adv. in Appl. Math.*, **41**, (2008), no. 3, 329–334. MR2449594

[8] X. Cheng and A. Singer. The spectrum of random inner-product of kernel matrices. *Random matrix Theory and Applications*, **2**, (2013), no. 4, 1350010. MR3149440

[9] A. Choromanska, M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun. The loss surfaces of multilayer networks. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, 2015.

[10] C. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys*, **91**, (2019), no. 4, 045002.

[11] T. Claeys, A. B. J. Kuijlaars, and D. Wang. Correlation kernels for sums and products of random matrices. *Random Matrices: Theory and Applications*, **4**, (2015), no. 4, 1550017. MR3418846

[12] R. Couillet and F. Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electron. J. Stat.*, **10**, (2016), no. 1, 1393–1454. MR3507369

[13] T. Dupic and I. P. Castillo. Spectral density of products of Wishart dilute random matrices. part i: the dense case. arXiv:1401.7802

[14] N. El Karoui. The spectrum of kernel random matrices. *Ann. Statist.*, **38**, (2010), no. 1, 1–50. MR2589315

[15] Z. Fan and A. Montanari. The spectral norm of random inner-product kernel matrices. *Prob. Theory Rel. Fields*, **173**, (2019), no. 1-2, 27–85. MR3916104

[16] G. W. Ford and G. E. Uhlenbeck. Combinatorial problems in the theory of graphs. III. *Proc. Nat. Acad. Sci. U.S.A.*, **42**, (1956), 529–535. MR0080912

[17] P. Forrester and D. Liu. Raney distributions and random matrix theory. *Journ. Stat. Phys.*, **158**, (2015), no. 5, 1051–1082. MR3313617

[18] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborova. Entropy and mutual information in models of deep neural networks. *NeurIPS*, 2018. MR4063612

[19] R. Giryes, G. Sapiro, and A. M. Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, **64**, (2016), no. 13, 3444–3457. MR3515693

[20] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, PMLR **9**, (2010), 249–256.

[21] B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.*, **376**, (2020), no. 1, 287-322. MR4093863

[22] S. Hayou, A. Doucet, and J. Rousseau. On the selection of initialization and activation function for deep neural networks. arXiv:1805.08266

[23] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and Others. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, **29**, (2012), no. 6, 82–97.

[24] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR **37**, (2015), 448–456.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, **25**, (2012), 1097–1105.

[26] A. B. J. Kuijlaars and L. Zhang. Singular values of products of ginibre random matrices, multiple orthogonal polynomials and hard edge scaling limits. *Comm. Math. Phys.*, **332**, (2014), no. 2:759–781. MR3257662

[27] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, **521**, (2015), 436–444.

[28] C. Louart and R. Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. arXiv:1805.08295

[29] C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *Ann. Appl. Probab.*, **28**, (2018), no. 2, 1190–1248 MR3784498

[30] S. Péché. A note on the Pennington–Worah distribution. *Elec. Comm. Probab.*, **24**, (2019), no. 66, 7 pp. MR4029435

[31] J. Pennington and Y. Bahri. Geometry of neural network loss surfaces via random matrix theory. *Proceedings of the 34th International Conference on Machine Learning*, PMLR **70**, (2017), 2798–2806.

[32] J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. *Advances in Neural Information Processing Systems*, (2017), 2637–2646. MR4063603

[33] K. Penson and K. Zyczkowski. Product of Ginibre matrices: Fuss-Catalan and Raney distributions. *Phys. Rev. E*, **83**, (2011), no. 6, 061118.

[34] K. Rajan and L. F. Abbott. Eigenvalue spectra of random matrices for neural networks. *Phys. Rev. Lett.*, **97**, (2006), no. 18, 188104.

[35] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, **61**, (2015), 85 – 117.

[36] J. W. Silverstein and Z. D. Bai. On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.*, **54**, (1995), no.2, 175–192. MR1345534

[37] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, and Others. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144

[38] B. Zhang, D. J. Miller, and Y. Wang. Nonlinear system modeling with random matrices: echo state networks revisited. *IEEE trans. Neural Netw. Learn. Syst.*, **23**, (2012), no. 1, 175–182.

Acknowledgments. The authors would like to thank D. Schröder and Z. Fan for pointing out errors in a previous version of the article as well as anonymous referees for helpful suggestions on how to improve the present paper.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[2]EJMS: Electronic Journal Management System `http://www.vtex.lt/en/ejms.html`
[3]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[4]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[5]Project Euclid: `https://projecteuclid.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`