Analysis of one-hidden-layer neural networks via the resolvent method

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we investigate the asymptotic spectral density of the random feature matrix $M = YY^*$ with $Y = f(WX)$ generated by a single-hidden-layer neural network, where W and X are random rectangular matrices with i.i.d. centred entries and f is a non-linear smooth function which is applied entry-wise. We prove that the Stieltjes transform of the limiting spectral distribution approximately satisfies a quartic self-consistent equation, which is exactly the equation obtained by Pennington and Worah [21] and Benigni and Péché [6] with the moment method. We extend the previous results to the case of additive bias $Y = f(WX + B)$ with B being an independent rank-one Gaussian random matrix, closer modelling the neural network infrastructures encountered in practice. Our key finding is that in the case of additive bias it is impossible to choose an activation function leaving the layer-to-layer singular value distribution invariant, in sharp contrast to the bias-free case where a simple integral constraint is sufficient for this type of invariance. Compared to previous works, we follow the *resolvent method* from random matrix theory, rather than the more commonly used moment method. We find that the former is more robust and less combinatorial and expect that it will apply also for models where the combinatorics of the former become intractable.

1 Introduction

Machine learning has seen many successful achievements in recent years. Applications in face identification, object and speech recognition, translation, email spam filtering, navigation, medical diagnosis, etc. have proved the enormous potential of machine learning for day-to-day live [16, 11]. Deep neural networks have turned out to be a particularly powerful machine learning method, and understanding the theoretical underpinning of their success has received tremendous attention in mathematics, physics and computer science.

A fully-connected, feed-forward neural network with L hidden layers of dimensions $n_1, \dots, n_L$ can be modelled as follows:

$$f_\theta(\boldsymbol{x}) = \beta^* f(W^{(L)} f(W^{(L-1)} f(\dots f(W^{(1)}\boldsymbol{x})\dots))) \in \mathbb{R}^d,$$

where $\boldsymbol{x} \in \mathbb{R}^{n_0}$ denotes the input data vector and $f\colon \mathbb{R} \to \mathbb{R}$ is a non-linear activation function which is applied entry-wise. We denote the parameters of the network by $\theta \coloneqq (W^{(1)}, \dots, W^{(L)}, \beta)$, where $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ for $1 \le l \le L$ and $\beta \in \mathbb{R}^{n_L \times d}$ are the matrices of the weights. In the classical setting of supervised learning, we are given a training set of (say) m samples of input feature vectors $\boldsymbol{x}_i \in \mathbb{R}^{n_0}$ with associated target vectors $\boldsymbol{z}_i \in \mathbb{R}^d$. For example, $\boldsymbol{x}_i$ may encode the pixels of a photograph of an animal and the target $\boldsymbol{z}_i$ may label the species of the animal in the image. Roughly speaking, the goal of supervised learning is to learn the mapping between the feature and the target vectors based on a given training set in order to predict the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

output of new unlabelled data. Let $X = (\boldsymbol{x}_1 \dots \boldsymbol{x}_m) \in \mathbb{R}^{n_0 \times m}$ be the matrix of the data and let $Z = (\boldsymbol{z}_1 \dots \boldsymbol{z}_m) \in \mathbb{R}^{d \times m}$ be the target matrix. Then, the aim of the network is to find optimal parameters θ such that $f_\theta(X) = (f_\theta(\boldsymbol{x}_1), \dots, f_\theta(\boldsymbol{x}_m)) \in \mathbb{R}^{d \times m}$ approximates the target Z optimally. During the training phase, weights are adjusted in order to minimize the empirical risk $\mathcal{R}(\theta) = \mathbf{E}\,\mathcal{L}(f_\theta(X), Z)$, where $\mathcal{L}(\cdot,\cdot)$ is a given loss function, usually involving some penalty for large weights θ in order to avoid *over-fitting*. Stochastic gradient descent (SGD) and its variants with *back-propagation* are the most commonly used algorithms for training multilayer networks by iteratively updating the parameters into the direction of the negative of the gradient of the empirical risk. For a much more complete survey, we refer the reader to [11].

In the present paper, we will focus on a single-hidden-layer neural network of the form $f_\theta(X) = \beta^* Y$ with $Y = f(WX)$. This model was first studied by Louart, Liao, and Couillet [19] for the case where the data matrix X is deterministic and W is a matrix of random weights (in particular, the weights are given by functions of standard Gaussian random variables), and by Pennington and Worah [21] for the case where X and W are independent random matrices with both centred Gaussian entries. In both papers, the matrix $\beta \in \mathbb{R}^{n_1 \times d}$ is the only parameter to be learned and is chosen as the unique minimizer of the ridge-regularized least squares loss function

$$\mathcal{L}(f_\theta(X), Z) = \frac{1}{2dm}\|Z - \beta^* Y\|_F^2 + \gamma\|\beta\|_F^2,$$

where $\gamma > 0$ is the *learning rate*. The unique minimizing weight matrix $\hat{\beta}$ is then equal to $\hat{\beta} = YG(-\gamma)Z^*$, where

$$G(-\gamma) = \left(\frac{1}{m}Y^*Y + \gamma \mathbf{1}_m\right)^{-1}$$

is the resolvent of $\frac{1}{m}Y^*Y$. As proved in [19, 21], the expected training loss E_{train} is related to $-\gamma\frac{\partial}{\partial\gamma}G(-\gamma)$, and thus also to the Stieltjes transform of the limiting spectral measure of $\frac{1}{m}Y^*Y$. Here the Stieltjes transform m_μ of a probability measure μ on $\mathbb{R}$ is defined as $m_\mu(z) := \int_{\mathbb{R}}(x-z)^{-1}\,\mathrm{d}\mu(x)$ for $z \in \mathbb{C}$ such that $\Im z \geq 0$, and for μ_{n_1} being the empirical probability measure of the n_1 eigenvalues of $\frac{1}{m}Y^*Y$ is related to the resolvent via $m_{\mu_{n_1}}(z) = \frac{1}{n_1}\operatorname{Tr} G(z)$. The performance of one-hidden-layer neural networks depends on the asymptotic spectral properties of the matrix $\frac{1}{m}Y^*Y$. Pennington and Worah [21] investigated the limiting spectral measure of the random matrix $M = \frac{1}{m}YY^*$ and derived the quartic self-consistent equation

$$\begin{aligned} 1 + zm = \theta_1(f)m\left(1 - \frac{\phi}{\psi}(1+zm)\right) - \frac{\theta_2(f)}{\psi}m(1+zm)\left(1 - \frac{\phi}{\psi}(1+zm)\right) \\ + \frac{\theta_2(f)(\theta_1(f) - \theta_2(f))}{\psi}m^2\left(1 - \frac{\phi}{\psi}(1+zm)\right)^2, \end{aligned} \tag{1}$$

which is approximately satisfied, $m(z) \approx g(z)$, by the Stieltjes transform $g(z) := \frac{1}{n_1}\operatorname{Tr} G(z)$ in case of Gaussian W, X. It is notable that the asymptotic spectrum of Y^*Y for large dimensions such that $n_0/m \to \phi \in (0,\infty)$ and $n_0/n_1 \to \psi \in (0,\infty)$ depends on the non-linear function f only through two integral parameters $\theta_1(f), \theta_2(f)$ defined below in (5). Benigni and Péché [6] then extended this model to random matrices W and X with general i.i.d. centred entries, and obtained the same self-consistent equation (1). We mention that (1) may be reduced, for some special cases, to the quadratic equation that is satisfied by the Stieltjes transform $m_{\mu_{MP}}$ of the Marchenko-Pastur distribution μ_{MP} [20]. This means that for some activation functions, the non-linear random matrix model has the same limiting spectral distribution as that of sample covariance matrices XX^* (in other cases, the equation can simplify to the cubic equation approximately satisfied by product Wishart matrices [7, 10]). This can be generalised to multilayer networks: [21] found experimentally that the Marchenko-Pastur distribution is invariant through multiple layers for activation functions with $\theta_2(f) = 0$, and this conjecture was proved in [6] for the general case of bounded activation functions. Moreover, [19] performed a spectral analysis on the Gram matrix model with general training data and proved that, in the large dimensional regime, the resolvent of Y^*Y has a similar behaviour as that observed in sample covariance matrix models. This was extended in [17] by considering Gaussian mixture of data. We also refer to the recent paper [18]. In the context of multilayer feedforward neural networks, Fan and Wang [9] analysed the eigenvalue distribution of the Gram matrix model, where the weights are at random and the input vectors are assumed to be approximately pairwise orthogonal.

In particular, they showed that the limiting spectral distribution converges to a deterministic limit and, at each intermediate layer, this limit corresponds to the Marchenko-Pastur map of a linear transformation of that of the previous layer.

In recent years, there have been some progress in the asymptotic analysis of the eigenvalue distribution of another Gram matrix, the so-called Neural Tangent Kernel (NTK). Consider a multilayer neural network and denote by $J = \nabla_\theta f_\theta(X)$ the Jacobian matrix of the network outputs with respect to the weights θ. Then, the NTK is the Gram matrix of J, defined by $K^{\mathrm{NTK}} = J^* J$. It was shown in [14] that the NTK at random initialization converges, in the infinite-width limit, to a deterministic kernel and it remains constant during the whole training time of the network. Subsequently, [22] analysed the spectrum of the sample covariance matrix JJ^* in a single-hidden-layer neural network, and provided an exact asymptotic characterization of the spectral distribution of JJ^* with random Gaussian weights and data. Recently, [9] proved that the limiting spectral measure of the NTK converges to a deterministic measure, which may be described by recursive fixed-point equations that extend the Marchenko-Pastur distribution.

The present paper is structured as follows: In the first part of this paper, we consider the non-linear random matrix model studied in [6]. We employ the resolvent method, and we compute the self-consistent equation that is approximately satisfied by the Stieltjes transform of the limiting spectral distribution. This is a quartic equation and is the same as that found in [6]. However, our method is quite different from [6, 21]. Indeed, in both [6] and [21], the authors approximated general non-linear functions by polynomial ones and then computed the asymptotics of high moments $\mathbf{E}\,\mathrm{Tr}(Y^k)$ with $Y = f(WX)$ to obtain the limiting measure via its moments. Conversely, we approach matters in a more robust and less combinatorial fashion by considering Y as a random matrix with correlated entries and then we directly derive a self-consistent equation for its resolvent. In particular, we prove that the random matrix Y has cycle correlations, in the sense that the joint cumulant does not vanish when the random variables Y_{ij}'s are joined by a cycle graph. We find that the variance of Y_{ij} is given by the parameter $\theta_1(f)$, whereas for $k > 1$ the cumulants $\kappa(Y_{i_1 i_2}, Y^*_{i_2 i_3}, \ldots, Y^*_{i_{2k} i_1})$ are powers of $\theta_2(f)$. We note that in the random matrix literature matrices with general decaying correlations have been studied previously, see e.g. [3, 8, 1]. However, the cycle correlations of Y considered in the present paper are much stronger compared to these previous results. The second part of this paper concerns the additive bias case which is a more realistic model for machine learning applications. More precisely, we consider the random feature matrix $Y = f(WX + B)$, where B is a rectangular rank-one Gaussian random matrix, and derive a characterization of the Stieltjes transform of the limiting spectral density. We chose B to be rank-one since for the most commonly used neural network architectures the added bias is chosen equal for each sample. [2] studied the bias case for deterministic data and i.i.d. Gaussian random weights, and computed the exact training error of a ridge-regularized noisy autoenconder in the high-dimensional regime. Interestingly we find that in the case of additive bias it is impossible to choose an activation function f giving rise to an invariant distribution throughout multiple layers, unlike in the bias-free case where $\theta_2(f) = 0$ yields the invariant Marchenko-Pastur distribution.

2 Model and main results

We consider a random data matrix $X \in \mathbb{R}^{n_0 \times m}$ with i.i.d. random variables X_{ij} with distribution ν_1 and a random weight matrix $W \in \mathbb{R}^{n_1 \times n_0}$ with i.i.d. weights W_{ij} with distribution ν_2. We assume that both distributions are centred with variance $\mathbf{E}X_{ij}^2 = \sigma_x^2$ and $\mathbf{E}W_{ij}^2 = \sigma_w^2$. Since for $1 \le i \le n_1$ and $1 \le j \le m$ we have

$$\left(\frac{WX}{\sqrt{n_0}}\right)_{ij} = \frac{1}{\sqrt{n_0}} \sum_{k=1}^{n_0} W_{ik} X_{kj},$$

we note that in light of the central limit theorem the entries of the matrix $\frac{WX}{\sqrt{n_0}}$ are approximately $\mathcal{N}(0, \sigma_w^2 \sigma_x^2)$-normally distributed random variables. Therefore, for any $t > 0$, we have the large deviation estimate

$$\mathbf{P}\left(\max_{i,j} \left|\frac{(WX)_{ij}}{\sqrt{n_0}}\right| > t\right) \lesssim n_0^2\, e^{-t^2/2\sigma_w^2\sigma_x^2}.$$

Let $f\colon \mathbb{R} \to \mathbb{R}$ be a C^∞ function with zero mean with respect to the Gaussian density of standard deviation $\sigma_w \sigma_x$, i.e.

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

We consider the random feature model generated by a single-hidden-layer neural network,

$$M = \frac{1}{m} Y Y^* \in \mathbb{R}^{n_1 \times n_1} \quad \text{with } \; Y = f\left(\frac{WX}{\sqrt{n_0}}\right), \tag{3}$$

where the activation function f is applied entry-wise. Let $\chi\colon \mathbb{R} \to \mathbb{R}$ be a smooth cut-off function that is equal to one for $|x| \leq 1$ and zero for $|x| \geq 2$. We then replace f by $f(\cdot)\chi(\log^{-1}(n_0)\,\cdot)$. In particular, we now have that f is smooth with compact support. Moreover, for any $l > 0$ and n_0 large enough, with probability $1 - n_0^{-l}$, the singular values of Y remain the same.

We are interested in the eigenvalue density of the random matrix M in the infinite size limit. So, we assume that the dimensions of both the columns and the rows of each matrix are large and grow at the same speed, i.e. we introduce some positive constants ϕ and ψ such that

$$\frac{n_0}{m} \longrightarrow \phi \;\text{ and }\; \frac{n_0}{n_1} \longrightarrow \psi \quad \text{as } \; n_0, n_1, m \to \infty. \tag{4}$$

We denote by $(\lambda_1, \ldots, \lambda_{n_1})$ the eigenvalues of M and define its empirical spectral distribution by $\mu_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{\lambda_i}$. Then, as n_1 grows large, the empirical distribution of eigenvalues converges in distribution to some deterministic limiting density.

Theorem 2.1. *There exists a deterministic measure $\mu = \mu_{\phi,\psi}(\theta_1, \theta_2)$ such that almost surely weakly*

$$\mu_{n_1} \longrightarrow \mu \quad \textit{as} \;\; n_1 \to \infty.$$

We notice that if $m < n_1$, then $\operatorname{rank}(M) = \min(n_1, m) = m$ and M has $n_1 - m$ zero eigenvalues. In this case, since $\phi/\psi > 1$, there exists an atom at 0 with mass $\mu_{n_1}(0) = 1 - \psi/\phi > 0$, and we have

$$\mu_{n_1} = \frac{n_1 - m}{n_1} \delta_0 + \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_{\lambda_i}.$$

Conversely, if $n_1 < m$, the matrix M has full rank and it is invertible. Since the nonzero eigenvalues of YY^* and Y^*Y are the same, the limiting measure μ of Theorem 2.1 turns out to be

$$\mu = \left(1 - \frac{\psi}{\phi}\right)_+ \delta_0 + \tilde{\mu},$$

where $(\cdot)_+ = \max(0, \cdot)$, and $\tilde{\mu}$ is the limiting spectral measure of $\frac{1}{m} Y^* Y$. The measure $\tilde{\mu}$ is compactly supported and absolutely continuous w.r.t. the Lebesgue measure.

We will prove that the deterministic measure μ of Theorem 2.1 is characterized through a quartic self-consistent equation for its Stieltjes transform $g(z) = \frac{1}{n_1} \operatorname{Tr} G(z)$, where

$$G(z) = (M - z\mathbf{1}_{n_1})^{-1} \in \mathbb{C}^{n_1 \times n_1}$$

is the resolvent of the random matrix M and the spectral parameter z lies in the upper half plane $\mathbb{H} = \{z \in \mathbb{C} \,|\, \Im z \geq 0\}$. We set

$$\theta_1(f) := \int_{\mathbb{R}} f^2(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x \quad \text{and} \quad \theta_2(f) := \left(\sigma_w \sigma_x \int_{\mathbb{R}} f'(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x\right)^2. \tag{5}$$

Then, the following theorem characterizes g as the solution to a quartic equation which depends only on the two parameters $\theta_1(f)$ and $\theta_2(f)$.

Theorem 2.2. *For some $\delta, \epsilon > 0$ and any $z \in \mathbb{H}$ with $\Im z > n_1^{-\frac{1}{4}+\epsilon}$, the measure μ is characterized through the following self-consistent equation for its Stieltjes transform g,*

$$\left| 1 + zg - \left(\theta_1 - \frac{\theta_2}{\psi}(1 + zg)\right) g \left(1 - \frac{\phi}{\psi}(1 + zg)\right) - \frac{\theta_2(\theta_1 - \theta_2)}{\psi} g^2 \left(1 - \frac{\phi}{\psi}(1 + zg)\right)^2 \right| \leq n_1^{-\delta} \tag{6}$$

almost surely.

Remark 2.3.

(i) We obtain an analogous result for complex feature and weight matrices W, X, c.f. Remark F.1.

(ii) The limiting spectral measure itself can be recovered from the Stieltjes transform $g(z)$ via the Stieltjes inversion formula,

$$\mu(\lambda) = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \Im g(\lambda + i\epsilon)$$

Therefore, Theorem 2.1 follows from Theorem 2.2.

(iii) Note that the quartic self-consistent equation (6) *may not have a unique solution such that $\Im g(z) > 0$.*

Remark 2.4. *It should be noted that Theorem 2.1 and Theorem 2.2 were proven in [21, 6] under different assumptions and with a different method. The result in [21] was obtained for i.i.d. Gaussian features and weights, whereas [6] extends the result to the case where both the inputs and the random weights have sub-Gaussian tails but are not necessarily Gaussian.*

Observing equation (6), we note that if $\theta_2(f) = 0$, then the limiting measure μ is exactly the Marchenko-Pastur μ_{MP} distribution with parameter ϕ/ψ. Indeed, in this case, $g(z)$ approximately satisfies the quadratic equation

$$1 + \left(z + \theta_1(f) \left(\frac{\phi}{\psi} - 1 \right) \right) g(z) + \theta_1(f) \frac{\phi}{\psi} z g(z)^2 \approx 0, \tag{7}$$

which corresponds to the self-consistent equation satisfied by the Stieltjes transform of μ_{MP} [20]. As discussed in the introduction, this consideration is relevant when studying multilayer networks. Pennington and Worah [21] conjectured that the asymptotic spectral distribution is invariant through multiple layers only for $\theta_2(f) = 0$ and is given by the Marchenko-Pastur distribution at each layer. Benigni and Péché [6] then proved this conjecture for bounded activation functions satisfying $\theta_2(f) = 0$. Moreover, if $\theta_1(f) = \theta_2(f)$, then equation (6) becomes cubic. In particular, the equality $\theta_1(f) = \theta_2(f)$ holds if and only if f is a linear function (for more details, we refer to the supplementary material in [21]). In this case, $M = \frac{1}{m} YY^*$ with $Y = WX$, and thus the limiting measure μ corresponds to the limiting spectral distribution of a product Wishart matrix. The spectral density for matrices of this type has been computed in [7, 10].

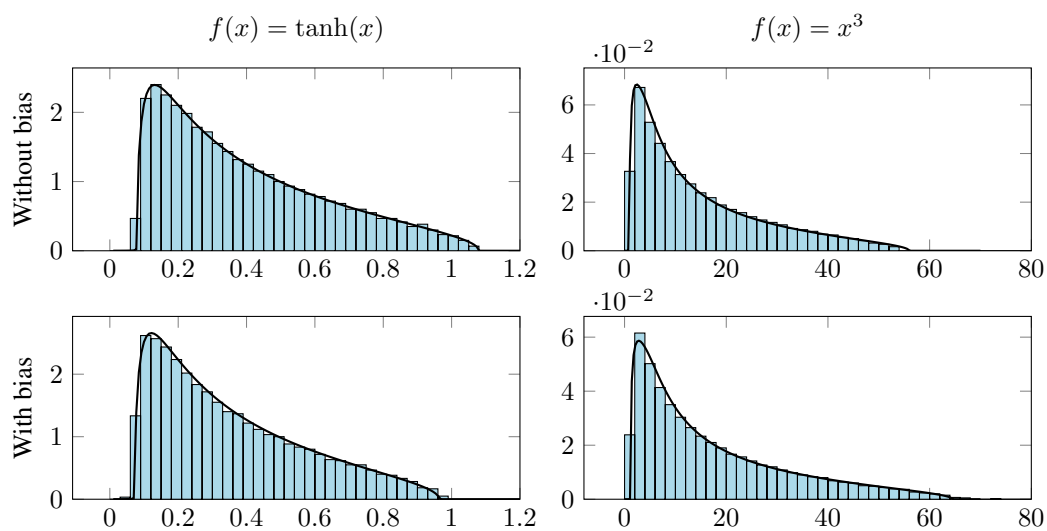


Figure 1: We present the eigenvalue histogram of the covariance matrix YY^* for a single random realisation together with the theoretical limit from Theorems 2.1 and 2.5 for the functions $f(x) = \tanh(x)$ and $f(x) = x^3$ with and without additive bias. We note that the presence of an additive bias can both increase or decrease the largest singular value. The numerical experiments were conducted for the parameters $n_1 = 3000$, $\phi = \sigma_x = \sigma_w = 1$, $\psi = 5$ (left) or $\psi = 2$ (right), and $\sigma_b = 0$ (top) or $\sigma_b = 0.25$ (bottom).

2.1 Additive bias case

The previous model can be generalised by adding random biases. In neural networks, the bias is an additional parameter that allows the model to better fit the given data. In this case, for each input data $x \in \mathbb{R}^{n_0}$, a bias vector $b \in \mathbb{R}^{n_1}$ is added to the vector $Wx \in \mathbb{R}^{n_1}$. We then apply a non-linear function $f\colon \mathbb{R} \to \mathbb{R}$ in an element-wise fashion to its vector arguments $Wx + b$ in order to obtain n_1 random features $f(Wx + b) \in \mathbb{R}^{n_1}$.

We consider a random bias matrix $B \in \mathbb{R}^{n_1 \times m}$ of i.i.d. Gaussian random variables $B_{ij} = B_i$ with zero mean and variance $\mathbf{E}B_i^2 = \sigma_b^2$. Note that the random matrix B has rank 1. Let $X \in \mathbb{R}^{n_0 \times m}$ and $W \in \mathbb{R}^{n_1 \times n_0}$ be random matrices with i.i.d. entries, defined as before. Moreover, let $f\colon \mathbb{R} \to \mathbb{R}$ be a C^∞ function satisfying

$$\int_{\mathbb{R}} f\left(\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}\, x\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x = 0. \tag{8}$$

Just as before, without loss of generality, upon replacing f by $f(\cdot)\chi(\log^{-1}(n_0)\,\cdot)$, we may assume that f is a smooth function with compact support. We then define the random matrix M by

$$M = \frac{1}{m} YY^* \in \mathbb{R}^{n_1 \times n_1} \quad \text{with } Y = f\left(\frac{WX}{\sqrt{n_0}} + B\right), \tag{9}$$

where f is applied entry-wise. We introduce the parameter

$$\widetilde{\sigma} \coloneqq \sqrt{\frac{\sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2 + 2\sigma_b^2)}{\sigma_w^2\sigma_x^2 + \sigma_b^2}},$$

and we define the following integral parameters:

$$\begin{aligned}
\theta_1(f) &\coloneqq \int_{\mathbb{R}} f^2\left(\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}x\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x, \\
\theta_{1,b}(f) &\coloneqq \frac{1}{2\pi\widetilde{\sigma}\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}} \int_{\mathbb{R}^2} f(x_1)f(x_2) \exp\left(-\frac{x_1^2 + x_2^2}{2\widetilde{\sigma}^2} + \frac{\sigma_b^2 x_1 x_2}{\widetilde{\sigma}^2(\sigma_w^2\sigma_x^2 + \sigma_b^2)}\right) \mathrm{d}\boldsymbol{x}, \\
\theta_2(f) &\coloneqq \frac{\sigma_w\sigma_x}{2\pi\widetilde{\sigma}\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}} \int_{\mathbb{R}^2} f'(x_1)f'(x_2) \exp\left(-\frac{x_1^2 + x_2^2}{2\widetilde{\sigma}^2} + \frac{\sigma_b^2 x_1 x_2}{\widetilde{\sigma}^2(\sigma_w^2\sigma_x^2 + \sigma_b^2)}\right) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{10}$$

We can now state the analogue of Theorem 2.2 in the additive bias case. In particular, the following theorem shows that the normalized trace of the resolvent of M approximately satisfies the self-consistent equation (6) with parameters given by (10).

Theorem 2.5. *The Stieltjes transform g satisfies* (6) *with parameters given by* (10)*, where $\theta_1(f)$ is replaced by $\theta_1(f) - \theta_{1,b}(f)$. Moreover, there exists a single outlier eigenvalue $\lambda_{\max} = n_1\theta_{1,b}(1 + \mathcal{O}(n_1^{-1/2}))$ of M that is separated from the support of the rest of the spectrum.*

We remark that the parameters $\theta_{1,b}(f), \theta_2(f)$ can be alternatively expressed as infinite series, directly demonstrating that for $\sigma_b \neq 0$ and non-trivial f both coefficients are strictly positive, $\theta_{1,b}(f), \theta_2(f) > 0$. For notational implicitly, we introduce the Hermite inner product

$$\langle f, g\rangle_{\mathrm{He}} \coloneqq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x)g(x)e^{-x^2/2}\,\mathrm{d}x.$$

Remark 2.6. *We have*

$$\begin{aligned}
\theta_{1,b}(f) &= \frac{\widetilde{\sigma}}{\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}} \sum_{k\geq 0} \frac{1}{k!}\Big(\frac{\sigma_b^2}{\sigma_w^2\sigma_x^2 + \sigma_b^2}\Big)^k \langle x^k, f(\widetilde{\sigma}\cdot)\rangle_{\mathrm{He}}^2 \\
\theta_2(f) &= \frac{\sigma_w^2\sigma_x^2\widetilde{\sigma}}{\sqrt{\sigma_w^2\sigma_x^2 + \sigma_b^2}} \sum_{k\geq 0} \frac{1}{k!}\Big(\frac{\sigma_b^2}{\sigma_w^2\sigma_x^2 + \sigma_b^2}\Big)^k \langle x^k, f'(\widetilde{\sigma}\cdot)\rangle_{\mathrm{He}}^2
\end{aligned} \tag{11}$$

and therefore $\theta_{1,b}(f) = 0$, $\sigma_b \neq 0$ implies that $f(\widetilde{\sigma}\cdot)$ is orthogonal to Hermite polynomials of any order, and consequently $f \equiv 0$. Similarly, $\theta_2(f) = 0$, $\sigma_b \neq 0$ implies that $f \equiv \mathrm{const}$.

2.2 Multiple layers

In [21] it was observed empirically that in the bias-free case activation functions with $\theta_2(f) = 0$ have the remarkable property that for multiple layers

$$Y^{(l+1)} := f(W^{(l)}Y^{(l)}), \qquad Y^{(0)} := X \tag{12}$$

the singular value distributions of $Y^{(1)}, Y^{(2)}, \ldots$ all asymptotically agree (up to scaling) with the probability distribution $\mu(\theta_1, \theta_2) = \mu(\theta_1, 0)$ from Theorem 2.1. This observation is very natural from our point of view since we find that $Y^{(1)}$ is approximately an i.i.d. random matrix if $\theta_2(f) = 0$, c.f. Proposition 3.2 below.

An interesting corollary of our Theorem 2.5 is that a similar invariance property *cannot* be present for the case of additive bias

$$Y^{(l+1)} := f(W^{(l)}Y^{(l)} + B^{(l)}), \qquad Y^{(0)} := X \tag{13}$$

since for $\sigma_b \neq 0$ we have $\theta_{1,b}(f), \theta_2(f) > 0$ for all activation functions f, and therefore already the random matrix $Y^{(1)}$ necessarily has leading order correlations, c.f. Proposition 3.3 below. In Fig. 2 we test this observation experimentally and choose the activation function $f(x) = c_1|x| - c_2$ with c_1, c_2 such that (2) is satisfied and $\theta_1(f) = 1$. Exactly as in [21, Fig. 1][1], we find that irrespective of the network depth the eigenvalues of the covariance matrix $Y^{(l)}(Y^{(l)})^*$ converge to their theoretical limit from Theorem 2.1. However, in the case of an additive bias no such convergence can be observed.

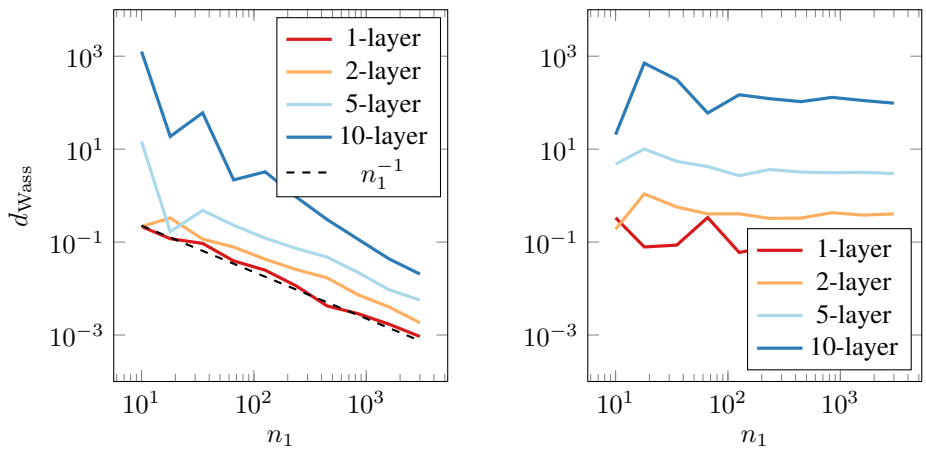


Figure 2: For randomly generated neural networks of varying depth and width, we compute the Wasserstein distance d_{Wass} between the empirical eigenvalue density of the covariance matrix $Y^{(l)}(Y^{(l)})^*$ to the distribution μ from Theorem 2.1 for the activation function $f(x) = c_1|x| - c_2$. In the bias-free case (left), the Wasserstein distance decays as the inverse of the network width, while in the case of an additive bias (right) no convergence can be observed. The numerical experiments were conducted for the parameters $\phi = \sigma_x = \sigma_w = 1$, $\psi = 2$ and $\sigma_b = 0$ (left) or $\sigma_b = 0.5$ (right).

The spectrum of the covariance matrix $Y^{(l)}(Y^{(l)})^*$ reflects the distortion of input data through the network and highly skewed distributions indicate poor conditioning which may impede learning performance [21]. *Batch normalization* seeks to remedy the distortion by normalising by the trace of the covariance matrix $Y^{(l)}(Y^{(l)})^t$ in each layer. In [21] it was suggested that choosing activation functions with $\theta_2(f) = 0$, i.e. functions which naturally leave the singular value distribution invariant, may serve as an alternative method of tuning networks for fast optimisation. Our result indicates that in the case of additive bias this alternative is not present. However, batch normalization seems to help stabilising the singular value distribution also in the additive bias case, c.f. Fig. 3.

[1] In the notation of [21], $f = f_1$.

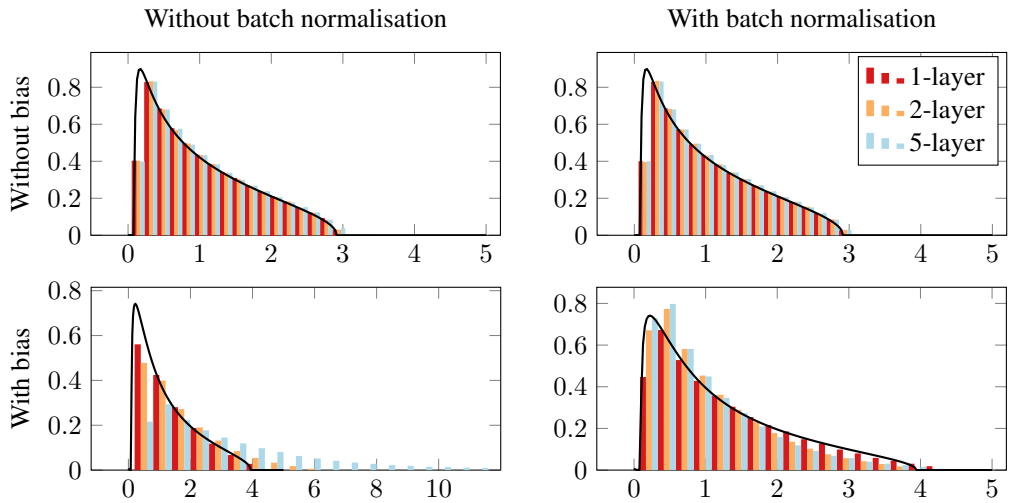


Figure 3: We present the eigenvalue distribution of neural networks of varying depth and in the presence/absence of both bias and batch normalization for the activation function $f(x) = c_1|x| - c_2$. In the bias-free case, batch normalisation has no effect on the spectral stability, and throughout the network the theoretical distribution from Theorem 2.1 matches the actual eigenvalue distribution of the covariance matrix $Y^{(l)}(Y^{(l)})^*$ well. In the case of an additive bias, the single-layer spectral density matches the theoretical limit from Theorem 2.5 to high accuracy. However, for multiple layers the spectral density diverges without additional batch normalization. Batch normalization alleviates the divergence, however the actual eigenvalue distribution deviates from the theoretical limit from Theorem 2.5. The numerical experiments were conducted for the parameters $n_1 = 3000$, $\phi = \sigma_x = \sigma_w = 1$, $\psi = 2$ and $\sigma_b = 0$ (top) or $\sigma_b = 0.5$ (bottom).

3 Outline of proof of Theorems 2.2 and 2.5

The proof of both Theorem 2.2 and 2.5 can be broken into two distinct parts. The first step is to show that $Y = f\left(\frac{WX}{\sqrt{n_0}}\right) \in \mathbb{R}^{n_1 \times m}$ can be viewed as a correlated random matrix with *cycle correlations*, c.f. Propositions 3.2 and 3.3 below. The second step is to prove the global law for the random matrix $M = \frac{1}{m}YY^*$ with the cycle correlations. In the following, we will sketch the derivation of the self-consistent equation. A more detailed proof is provided in the supplementary material.

The key idea is to use a multivariate cumulant expansion formula. Cumulants of a random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ can be defined in a combinatorial way by

$$\kappa(X_1, \dots, X_n) = \sum_{\pi} (-1)^{|\pi|-1}(|\pi|-1)! \prod_{B \in \pi} \mathbf{E}\left(\prod_{i \in B} X_i\right), \tag{14}$$

where the sum runs over all partitions π of the set $[n] = \{1, \dots, n\}$, the product runs over the blocks B of the partition π, and $|\pi|$ is the number of blocks in the partition. The following expansion is commonly referred to as a cumulant expansion and generalises the Gaussian integration by parts. In the context of random matrix theory, the usefulness of this expansion was first observed in [15] and later revived in [12, 13]. A proof of the following lemma is provided in Appendix C for completeness.

Lemma 3.1 (Cumulant expansion)**.** *If $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random vector with finite moments of all orders, then*

$$\mathbf{E}X_1 f(\boldsymbol{X}) = \sum_{l \geq 1} \sum_{i_1, \dots, i_l} \frac{\kappa(X_1, X_{i_1}, \dots, X_{i_l})}{l!} \mathbf{E}\partial_{i_1} \cdots \partial_{i_l} f(\boldsymbol{X}),$$

where $f \colon \mathbb{R}^n \to \mathbb{R}$ is smooth.

We start with the defining identity of the resolvent, $\mathbf{1}_{n_1} + zG = MG$, and we compute its average trace:

$$1 + zg = \frac{1}{n_1} \operatorname{Tr} \frac{YY^*G}{m} = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{m} Y_{ij} \left(\frac{Y^*G}{m} \right)_{ji}. \tag{15}$$

Since the random variable $(Y^*G)_{ji}$ can be seen as a function of Y_{ij}, we can take the expectation on both sides of (15) and apply Lemma 3.1:

$$1 + z\,\mathbf{E}g = \frac{1}{n_1} \sum_{k \geq 1} \sum_{i_1, \ldots, i_{2k}} \frac{\kappa(Y_{i_1 i_2}, Y_{i_3 i_4}, \ldots, Y_{i_{2k-1} i_{2k}})}{(k-1)!} \mathbf{E} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} \left(\frac{Y^*G}{m} \right)_{i_2 i_1}. \tag{16}$$

The main goal now is to show that Y can be viewed as a random matrix with cycle correlations given as in the Propositions 3.2 and 3.3 below: Prop. 3.2 refers to the bias-free case and Prop. 3.3 to the additive bias case. We postpone the proof of both propositions to Subsections A.2 and B.2, resp.

Proposition 3.2 (Correlation structure without bias)**.** *It holds that*

$$\kappa(Y_{ij}) = 0 \quad \textit{and} \quad \kappa(Y_{ij}, Y^*_{ji}) \approx \theta_1(f), \tag{17}$$

for $1 \leq i \leq n_1$ and $1 \leq j \leq m$. Moreover, for $k > 1$ and for $i_1, \ldots, i_{2k}$ distinct,

$$\kappa(Y_{i_1 i_2}, \ldots, Y^*_{i_{2k} i_1}) \approx \theta_2(f)^k n_0^{1-k}. \tag{18}$$

Proposition 3.3 (Correlation structure with bias)**.** *It holds that*

$$\kappa(Y_{i_1 i_2}) = 0, \quad \kappa(Y_{i_1 i_2}, Y^*_{i_2 i_1}) \approx \theta_1(f) \quad \textit{and} \quad \kappa(Y_{i_1 i_2}, Y^*_{i_3 i_1}) \approx \theta_{1,b}(f) \;\; (i_3 \neq i_2), \tag{19}$$

for $1 \leq i_1 \leq n_1$ and $1 \leq i_2, i_3 \leq m$. Moreover, for $k > 1$ and for $i_1, \ldots, i_{2k}$ distinct,

$$\kappa(Y_{i_1 i_2}, \ldots, Y^*_{i_{2k} i_1}) \approx \theta_2(f)^k n_0^{1-k}. \tag{20}$$

Applying Propositions 3.2 and 3.3 to (16), computing the partial derivatives and doing some bookkeeping, we get the desired equation (6) as $n_0, n_1, m \to \infty$. To complete the proofs of Theorems 2.2 and 2.5, one has to show the concentration of g around $\mathbf{E}g$, as stated in the following lemma.

Lemma 3.4. *For the random matrix $M = \frac{1}{m} YY^*$ and a complex number $z \in \mathbb{H}$ such that $\Im z > n_1^{-\frac{1}{4}+\epsilon}$, for some $\epsilon > 0$, it holds that*

$$\mathbf{E}_W |g(z) - \mathbf{E}_W g(z)|^4 \lesssim \frac{1}{n_1^2 (\Im z)^4} \tag{21}$$

with high probability in X, and analogously

$$\mathbf{E}_X |g(z) - \mathbf{E}_X g(z)|^4 \lesssim \frac{1}{n_1^2 (\Im z)^4} \tag{22}$$

with high probability in W, where $g(z) = \frac{1}{n_1} \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1}$ denotes the normalized trace of the resolvent of M, and $\mathbf{E}_X$ (resp. $\mathbf{E}_W$) the expectation in the X-space (resp. W-space).

The proof of this lemma relies on a standard argument (e.g. see the proof of the concentration inequality in [5, Subsection 3.3.2]) and is given in Appendix E.

4 Conclusion

In this paper, we analysed the singular value distribution of fully random neural networks and found that in the case of additive biases it is impossible to tune the activation function in a way leaving the limiting spectral distribution invariant through multiple layers. In addition, we showed that the resolvent method from random matrix theory also applies to the neural network analysis, despite the non-linearities and we expect that this robust method will prove to be useful in contexts where the conventionally used moment method becomes intractable.

Broader impact

Our result is a purely theoretical one for fully random features, weights and biases. Therefore, we do not expect our contribution to have ethical concerns or adverse future societal consequences.

References

[1] A. Adhikari and Z. Che, *Edge universality of correlated Gaussians*, Electron. J. Probab. **24**, Paper No. 44, 25 (2019), MR3949269.

[2] B. Adlam, J. Levinson, and J. Pennington, *A random matrix perspective on mixtures of nonlinearities for deep learning*, preprint (2019), arXiv:1912.00827.

[3] O. H. Ajanki, L. Erdős, and T. Krüger, *Stability of the matrix Dyson equation and random matrices with correlations*, Probab. Theory Related Fields **173**, 293–373 (2019), MR3916109.

[4] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Vol. 118, Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2010), pp. xiv+492, MR2760897.

[5] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, Second, Springer Series in Statistics (Springer, New York, 2010), pp. xvi+551, MR2567175.

[6] L. Benigni and S. Péché, *Eigenvalue distribution of nonlinear models of random matrices*, preprint (2019), arXiv:1904.03090.

[7] T. Dupic and I. P. Castillo, *Spectral density of products of wishart dilute random matrices. Part I: the dense case*, preprint (2014), arXiv:1401.7802.

[8] L. Erdős, T. Krüger, and D. Schröder, *Random matrices with slow correlation decay*, Forum Math. Sigma **7**, Paper No. e8, 89 (2019), MR3941370.

[9] Z. Fan and Z. Wang, "Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks", Advances in neural information processing systems, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (2020), pp. 7710–7721.

[10] P. J. Forrester, *Eigenvalue statistics for product complex Wishart matrices*, J. Phys. A **47**, 345202, 22 (2014), MR3251989.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, MA, 2016), pp. xxii+775, MR3617773.

[12] Y. He and A. Knowles, *Mesoscopic eigenvalue statistics of Wigner matrices*, Ann. Appl. Probab. **27**, 1510–1550 (2017), MR3678478.

[13] Y. He, A. Knowles, and R. Rosenthal, *Isotropic self-consistent equations for mean-field random matrices*, Probab. Theory Related Fields **171**, 203–249 (2018), MR3800833.

[14] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: convergence and generalization in neural networks", Proceedings of the 32nd international conference on neural information processing systems, NIPS'18 (2018), pp. 8580–8589.

[15] A. M. Khorunzhy, B. A. Khoruzhenko, and L. A. Pastur, *Asymptotic properties of large random matrices with independent entries*, J. Math. Phys. **37**, 5033–5060 (1996), MR1411619.

[16] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521**, 436–44 (2015), PMID26017442.

[17] Z. Liao and R. Couillet, "On the spectrum of random features maps of high dimensional data", Proceedings of the 35th international conference on machine learning, Vol. 80, edited by J. Dy and A. Krause, Proceedings of Machine Learning Research (2018), pp. 3063–3071.

[18] C. Louart and R. Couillet, *Concentration of measure and large random matrices with an application to sample covariance matrices*, preprint (2018), arXiv:1805.08295.

[19] C. Louart, Z. Liao, and R. Couillet, *A random matrix approach to neural networks*, Ann. Appl. Probab. **28**, 1190–1248 (2018), MR3784498.

[20] V. A. Marčenko and L. A. Pastur, *Distribution of eigenvalues in certain sets of random matrices*, Mat. Sb. (N.S.) **72 (114)**, 507–536 (1967), MR0208649.

[21] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning", Advances in neural information processing systems, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (2017).

[22] J. Pennington and P. Worah, "The spectrum of the fisher information matrix of a single-hidden-layer neural network", Advances in neural information processing systems, Vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (2018).

[23] T. P. Speed, *Cumulants and partition lattices*, Austral. J. Statist. **25**, 378–388 (1983), MR725217.

Checklist

1. For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

(b) Did you describe the limitations of your work? [Yes] See 4.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] See 4.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.

(b) Did you include complete proofs of all theoretical results? [Yes] See Section 3 and the supplementary material, i.e. Appendix A- E.

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We uploaded the code here `https://github.com/wirhabenzeit/nonlinearRMT`.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The figures indicate the parameters we used.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to sharp concentration (c.f. Lemma 3.4), the error can be controlled even theoretically.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] It does not apply since the small experiments can be run within minutes on a personal computer.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [No] We do not work with existing assets.

(b) Did you mention the license of the assets? [No] We do not work with existing assets.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included a URL. See 3a.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We do not use data from people.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We do not use data from people.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We did not work with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We did not work with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We did not work with human subjects.

A Proof of Theorem 2.2

A.1 Derivation of the self-consistent equation

We start from (16). Applying Proposition 3.2 to (16), we have

$$
\begin{aligned}
1 + z\,\mathbf{E}g &= \frac{1}{n_1 m} \sum_{k\geq 1} \sum_{i_1,\ldots,i_{2k}} \frac{\kappa(Y_{i_1 i_2}, Y_{i_3 i_4}, \ldots, Y_{i_{2k-1} i_{2k}})}{(k-1)!}\, \mathbf{E}\partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} (Y^* G)_{i_2 i_1} \\
&= \frac{1}{n_1 m} \sum_{k\geq 1} \sum_{i_1,\ldots,i_{2k}} \kappa(Y_{i_1 i_2}, Y^*_{i_2 i_3}, \ldots, Y^*_{i_{2k} i_1})\, \mathbf{E}\partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} (Y^* G)_{i_2 i_1} \\
&= \frac{1}{n_1 m} \sum_{i_1, i_2} \kappa(Y_{i_1 i_2}, Y^*_{i_2 i_1})\, \mathbf{E}\partial_{Y^*_{i_2 i_1}} (Y^* G)_{i_2 i_1} \\
&\quad + \frac{1}{n_1 m} \sum_{k\geq 2} \sum_{i_1,\ldots,i_{2k}} \kappa(Y_{i_1 i_2}, \ldots, Y^*_{i_{2k} i_1})\, \mathbf{E}\partial_{Y^*_{i_2 i_3}} \cdots \partial_{Y^*_{i_{2k} i_1}} (Y^* G)_{i_2 i_1} \\
&\approx \frac{\theta_1}{n_1 m} \sum_{i_1, i_2} \mathbf{E}\partial_{Y^*_{i_2 i_1}} (Y^* G)_{i_2 i_1} + \frac{1}{n_1 m} \sum_{k\geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_1,\ldots,i_{2k}} \mathbf{E}\partial_{Y^*_{i_2 i_3}} \cdots \partial_{Y^*_{i_{2k} i_1}} (Y^* G)_{i_2 i_1}\,,
\end{aligned}
\tag{23}
$$

where in the second step the factorial $(k-1)!$ disappears since there are exactly $(k-1)!$ ways to map the variables $Y_{i_3 i_4}, \ldots, Y_{i_{2k-1} i_{2k}}$ into $Y^*_{i_2 i_3}, Y_{i_3 i_4}, \ldots, Y^*_{i_{2k} i_1}$ with distinct $i_1, \ldots, i_{2k}$. From this point onwards, we will omit reference to $\mathbf{E}$ to simplify notation slightly.

We now need to compute the partial derivatives in (23). The proof of the following lemma is included in Appendix C.

Lemma A.1. *Let $G(z) = (M - z\mathbf{1}_{n_1})^{-1}$, $z \in \mathbb{H}$, be the resolvent of the random matrix $M = \frac{1}{m} Y Y^* \in \mathbb{R}^{n_1 \times n_1}$. Then, it holds that*

$$
\partial_{Y^*_{i_2 i_1}} (Y^* G)_{i_2 i_1} = G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right), \tag{24a}
$$

$$
\partial_{Y^*_{i_2 i_3}} \cdots \partial_{Y^*_{i_{2k} i_1}} (Y^* G)_{i_2 i_1} \approx -\partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} \left(\frac{GY}{m} \right)_{i_3 i_{2k}} G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right). \tag{24b}
$$

Thus, using Lemma A.1 in (23) we have

$$
\begin{aligned}
1 + zg &\approx \frac{\theta_1}{n_1 m} \sum_{i_1, i_2} G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right) \\
&\quad - \frac{1}{n_1 m} \sum_{k\geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_1,\ldots,i_{2k}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} \left(\frac{GY}{m} \right)_{i_3 i_{2k}} G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right) \\
&= \theta_1 g - \theta_1 \frac{n_1}{m} g \left\langle \frac{Y^* G Y}{m} \right\rangle \\
&\quad - \left(g - \frac{n_1}{m} g \left\langle \frac{Y^* G Y}{m} \right\rangle \right) \frac{1}{m} \sum_{k\geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_3,\ldots,i_{2k}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} (GY)_{i_3 i_{2k}}\,,
\end{aligned}
\tag{25}
$$

where $\left\langle \frac{Y^*GY}{m} \right\rangle := \frac{1}{n_1} \operatorname{Tr} \frac{Y^*GY}{m} = 1 + zg$ from (15). Again, we stress that the equalities are meant in expectation. Moreover, shifting the index in the above summation, we get

$$
\begin{aligned}
&\frac{1}{m} \sum_{k \geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_3, \ldots, i_{2k}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} (GY)_{i_3 i_{2k}} \\
&= \theta_2 \frac{n_1}{n_0} \frac{1}{m} \sum_{k \geq 1} \frac{\theta_2^k}{n_1 n_0^{k-1}} \sum_{i_3, \ldots, i_{2k+2}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k+1} i_{2k+2}}} (GY)_{i_3 i_{2k+2}} \\
&= \theta_2^2 \frac{n_1}{n_0} \frac{1}{n_1 m} \sum_{i_3, i_4} \partial_{Y_{i_3 i_4}} (GY)_{i_3 i_4} \\
&\quad + \theta_2 \frac{n_1}{n_0} \frac{1}{n_1 m} \sum_{k \geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_3, \ldots, i_{2k+2}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k+1} i_{2k+2}}} (GY)_{i_3 i_{2k+2}} \\
&\approx \theta_2^2 \frac{n_1}{n_0} \left(g - \frac{n_1}{m} g \left\langle \frac{Y^*GY}{m} \right\rangle \right) + \theta_2 \frac{n_1}{n_0} \left(1 + zg - \theta_1 g + \theta_1 \frac{n_1}{m} g \left\langle \frac{Y^*GY}{m} \right\rangle \right) \\
&= \theta_2 \frac{n_1}{n_0} (1 + zg) - \theta_2 (\theta_1 - \theta_2) \frac{n_1}{n_0} g \left(1 - \frac{n_1}{m} (1 + zg) \right),
\end{aligned}
$$

where in the third step we used (23). Finally, together with (25), we have

$$
\begin{aligned}
1 + zg \approx \theta_1 g \left(1 - \frac{n_1}{m} (1 + zg) \right) - \theta_2 \frac{n_1}{n_0} g (1 + zg) \left(1 - \frac{n_1}{m} (1 + zg) \right) \\
+ \theta_2 (\theta_1 - \theta_2) \frac{n_1}{n_0} g^2 \left(1 - \frac{n_1}{m} (1 + zg) \right)^2,
\end{aligned}
\tag{26}
$$

which corresponds to the desired equation (6) as $n_0, n_1, m \to \infty$. Thus, (26) combined with the concentration inequality given in Lemma 3.4 completes the proof of Theorem 2.2.

Proof of Theorem 2.2. We need to show the concentration w.r.t. $\mathbf{E}_{W,X} \equiv \mathbf{E}$. By the triangle and Jensen inequality we have

$$
\begin{aligned}
\mathbf{E}|g(z) - \mathbf{E}g(z)|^4 &\lesssim \mathbf{E}|g(z) - \mathbf{E}_W g(z)|^4 + \mathbf{E}_X |\mathbf{E}_W g(z) - \mathbf{E}g(z)|^4 \\
&\leq \mathbf{E}_X \left(\mathbf{E}_W |g(z) - \mathbf{E}_W g(z)|^4 \right) + \mathbf{E}_W \left(\mathbf{E}_X |g(z) - \mathbf{E}_X g(z)|^4 \right) \lesssim \frac{2}{n_1^2 (\Im z)^4}
\end{aligned}
$$

and thus the almost sure convergence follows from the Borel-Cantelli Lemma, completing the proof of Theorem 2.2 together with (26). □

A.2 Proof of Proposition 3.2

First, we observe that in light of the central limit theorem, the random variables

$$
\left(\frac{WX}{\sqrt{n_0}} \right)_{ij} = \frac{1}{\sqrt{n_0}} \sum_{k=1}^{n_0} W_{ik} X_{kj},
$$

with $1 \leq i \leq n_1$ and $1 \leq j \leq m$, are approximately $\mathcal{N}(0, \sigma_w^2 \sigma_x^2)$-normally distributed. Our next goal is to compute their cumulants. We start with the second cumulant.

Lemma A.2. *The cumulant of* $\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}$ *and* $\frac{(WX)_{i_3 i_4}}{\sqrt{n_0}}$ *is nonzero only if* $i_1 = i_3$ *and* $i_2 = i_4$*, and in this case it holds that*

$$
\kappa \left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \frac{(WX)^*_{i_2 i_1}}{\sqrt{n_0}} \right) = \sigma_w^2 \sigma_x^2.
$$

Proof. Since the random variable $\frac{(WX)_{ij}}{\sqrt{n_0}}$ has zero mean, we have

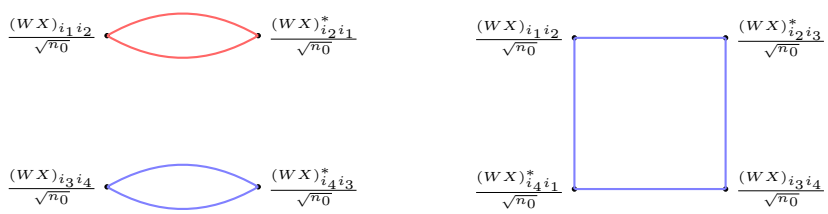
$$\begin{aligned}\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}},\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}\right) &= \frac{1}{n_0}\mathbf{E}(WX)_{i_1i_2}(WX)_{i_3i_4}\\ &= \frac{1}{n_0}\sum_{k_1,k_2=1}^{n_0}\mathbf{E}W_{i_1k_1}X_{k_1i_2}W_{i_3k_2}X_{k_2i_4}\\ &= \frac{1}{n_0}\sum_{k_1=1}^{n_0}\delta_{i_1i_3}\delta_{i_2i_4}\,\mathbf{E}W_{i_1k_1}^2X_{k_1i_2}^2 = \delta_{i_1i_3}\delta_{i_2i_4}\sigma_w^2\sigma_x^2.\end{aligned}$$

Thus, the second cumulant is nonzero if $i_1 = i_3$ and $i_2 = i_4$, and in this case it is exactly the variance of the random variable $\frac{(WX)_{ij}}{\sqrt{n_0}}$. □

We now consider four random entries, and we compute

$$\frac{1}{n_0^2}\mathbf{E}(WX)_{i_1i_2}(WX)_{i_3i_4}(WX)_{i_5i_6}(WX)_{i_7i_8}.$$

We can encode the product of the four random variables as a graph. We then obtain a non-vanishing expectation if we either pair the four random variables or loop them. We can represent the first case with a disconnected graph of two distinct cycles of length two, while the second can be described by a cycle graph of length four. See Fig. 4 for an illustration of both ways.



(a) Two cycle graphs of length two. (b) A cycle graph of length four.

Figure 4: The two possible ways to encode the product of four random elements as a graph such that the expectation does not vanish.

In the first case, the cumulant vanishes. Indeed, we have

$$\begin{aligned}&\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}},\frac{(WX)^*_{i_2i_1}}{\sqrt{n_0}},\frac{(WX)_{i_3i_4}}{\sqrt{n_0}},\frac{(WX)^*_{i_4i_3}}{\sqrt{n_0}}\right)\\ &= \frac{1}{n_0^2}\mathbf{E}\,(WX)^2_{i_1i_2}(WX)^2_{i_3i_4} - \frac{1}{n_0^2}\mathbf{E}(WX)^2_{i_1i_2}\,\mathbf{E}(WX)^2_{i_3i_4} \approx 0.\end{aligned}$$

On the other hand, when considering case (b), the cumulant is given by

$$\begin{aligned}&\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}},\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}},\frac{(WX)_{i_3i_4}}{\sqrt{n_0}},\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right)\\ &= \frac{1}{n_0^2}\mathbf{E}(WX)_{i_1i_2}(WX)^*_{i_2i_3}(WX)_{i_3i_4}(WX)^*_{i_4i_1}\\ &= \frac{1}{n_0^2}\sum_{k_1,k_2,k_3,k_4=1}^{n_0}\mathbf{E}W_{i_1k_1}X_{k_1i_2}W_{i_3k_2}X_{k_2i_2}W_{i_3k_3}X_{k_3i_4}W_{i_1k_4}X_{k_4i_4}\\ &= \frac{1}{n_0^2}\sum_{k_1=1}^{n_0}\mathbf{E}W_{i_1k_1}^2X_{k_1i_2}^2W_{i_3k_1}^2X_{k_1i_4}^2 = \frac{\left(\sigma_w^2\sigma_x^2\right)^2}{n_0} + \mathcal{O}(n_0^{-2}).\end{aligned}$$

As a consequence, in the case of four random entries, the only contribution comes from

$$\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}},\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}},\frac{(WX)_{i_3i_4}}{\sqrt{n_0}},\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right).$$

This result can be generalised:

Lemma A.3. *The nonzero cumulant of $2k$ entries of the random matrix $\frac{WX}{\sqrt{n_0}}$ is given by*

$$\kappa\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \ldots, \frac{(WX)^*_{i_{2k} i_1}}{\sqrt{n_0}}\right) = \frac{\left(\sigma_w^2 \sigma_x^2\right)^k}{n_0^{k-1}} + \mathcal{O}(n_0^{-k})$$

for $i_1, \ldots, i_{2k}$ distinct.

Proof. As illustrated for the case with four random variables, to have a nonzero cumulant, we can encode the $2k$ random variables as a cycle graph of length $2k$. Then, the only contribution comes from

$$\kappa\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \ldots, \frac{(WX)^*_{i_{2k} i_1}}{\sqrt{n_0}}\right) = \frac{1}{n_0^k}\,\mathbf{E}(WX)_{i_1 i_2} \cdots (WX)^*_{i_{2k} i_1} = \frac{\left(\sigma_w^2 \sigma_x^2\right)^k}{n_0^{k-1}} + \mathcal{O}(n_0^{-k}),$$

which completes the proof. □

We now use the cumulant asymptotics in order to asymptotically compute the characteristic function of the random variables $\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \ldots, \frac{(WX)^*_{i_{2k} i_1}}{\sqrt{n_0}}$.

Lemma A.4. *The characteristic function of $\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \ldots, \frac{(WX)^*_{i_{2k} i_1}}{\sqrt{n_0}}$, for $k \geq 2$, is given by*

$$\begin{aligned}
&\mathbf{E}\exp\left(it_1 \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + \cdots + it_{2k} \frac{(WX)^*_{i_{2k} i_1}}{\sqrt{n_0}}\right) \\
&= \left(1 + \frac{\left(\sigma_w^2 \sigma_x^2\right)^k}{n_0^{k-1}} \prod_{i=1}^{2k} t_i + \mathcal{O}(n_0^{-k})\right) \exp\left(-\frac{\sigma_w^2 \sigma_x^2}{2} \sum_{i=1}^{2k} t_i^2\right)
\end{aligned}$$

for $t_1, \ldots, t_{2k} \in \mathbb{R}$ such that $\prod_{i=1}^{2k} t_i < n_0^{k-1}$.

Proof. The key idea is to write the log-characteristic function using cumulants and compute them. The computation of the cumulants is given in Appendix D. We then obtain the characteristic function by taking the exponential. □

Finally, we compute the cumulants of the entries of the random matrix Y. Since the activation function f is applied component-wise, it follows from the previous results that the only contribution comes from $\kappa(Y_{i_1 i_2}, Y^*_{i_2 i_3}, \ldots, Y^*_{i_{2k} i_1})$ for $k \geq 1$ and $i_1, \ldots, i_{2k}$ distinct, thus proving that Y has cycle correlations.

Proof of Proposition 3.2. Since f is applied entry-wise, a straightforward computation yields

$$\kappa(Y_{ij}) = \mathbf{E}Y_{ij} \approx \int_{\mathbb{R}} f(x) \frac{e^{-x^2/2\sigma_w^2\sigma_x^2}}{\sigma_w \sigma_x \sqrt{2\pi}} \mathrm{d}x = \int_{\mathbb{R}} f(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x = 0,$$

and

$$\kappa(Y_{ij}, Y^*_{ji}) \approx \int_{\mathbb{R}} f^2(\sigma_w \sigma_x x) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x = \theta_1(f),$$

which proves (17). Proving (18) is more challenging. Let $k > 1$. Then, since f is a smooth function with compact support, we have that f is in C^l for some integer $l > 1 + \frac{2k^2}{k-1}$. Using the Fourier inversion theorem, it follows that

$$\begin{aligned}
f(x_1) &= \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(t_1)\, e^{it_1 x_1} \mathrm{d}t_1 \\
&= \frac{1}{2\pi} \int_{|t_1| \leq n_0^{\frac{k-1}{2k}}} \hat{f}(t_1)\, e^{it_1 x_1} \mathrm{d}t_1 + \frac{1}{2\pi} \int_{|t_1| > n_0^{\frac{k-1}{2k}}} \hat{f}(t_1)\, e^{it_1 x_1} \mathrm{d}t_1 \\
&= \frac{1}{2\pi} \int_{|t_1| \leq n_0^{\frac{k-1}{2k}}} \hat{f}(t_1)\, e^{it_1 x_1} \mathrm{d}t_1 + \mathcal{O}\left((n_0^{\frac{k-1}{2k}})^{1-l}\right),
\end{aligned}$$

where we used $|\hat{f}(t_1)| \le \frac{c}{(1+|t_1|)^l}$, for some positive constant c. For notational simplicity we work in the case $k=2$, but the argument when $k>2$ is the same. We compute

$$\begin{aligned}\mathbf{E}Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1} &= \frac{1}{(2\pi)^4}\int_{\mathbb{R}^4}\int_{\mathbb{R}^4} f(x_1)f(x_2)f(x_3)f(x_4)\,\varphi(\boldsymbol{t})\,e^{-i\boldsymbol{t}\cdot\boldsymbol{x}}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{(2\pi)^4}\int_{\forall i,\,|t_i|\le n_0^{\frac14}} \hat{f}(t_1)\hat{f}(t_2)\hat{f}(t_3)\hat{f}(t_4)\,\varphi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} + \mathcal{O}(n_0^{-2}),\end{aligned}$$

where here $l>9$ and $\varphi(\boldsymbol{t})$ denotes the characteristic function of $\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}$, $\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}$, $\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}$, and $\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}$. Since $t_1t_2t_3t_4<n_0$, we can use Lemma A.4 and we obtain

$$\begin{aligned}&\mathbf{E}Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1}\\ &= \frac{1}{(2\pi)^4}\int \hat{f}(t_1)\hat{f}(t_2)\hat{f}(t_3)\hat{f}(t_4)\left(1+\frac{(\sigma_w^2\sigma_x^2)^2}{n_0}\prod_{i=1}^4 t_i\right)e^{-\frac{\sigma_w^2\sigma_x^2}{2}\sum_{i=1}^4 t_i^2}\mathrm{d}\boldsymbol{t} + \mathcal{O}(n_0^{-2})\\ &= \left(\frac{1}{2\pi}\int \hat{f}(t)e^{-\sigma_w^2\sigma_x^2\,t^2/2}\,\mathrm{d}t\right)^4 + \frac{1}{n_0}\left(\sigma_w\sigma_x\frac{1}{2\pi}\int \widehat{f}'(t)e^{-\sigma_w^2\sigma_x^2\,t^2/2}\,\mathrm{d}t\right)^4 + \mathcal{O}(n_0^{-2}),\end{aligned} \tag{27}$$

where in the last step we used the property $\widehat{f}'(t) = it\hat{f}(t)$. By the Fourier inversion theorem, the first integral in (27) is zero by assumption:

$$\begin{aligned}\frac{1}{2\pi}\int \hat{f}(t)e^{-\sigma_w^2\sigma_x^2\,t^2/2}\,\mathrm{d}t &= \frac{1}{2\pi}\int f(x)\left(\frac{2\pi}{\sigma_w^2\sigma_x^2}\right)^{1/2} e^{-x^2/2\sigma_w^2\sigma_x^2}\,\mathrm{d}x\\ &= \frac{1}{\sqrt{2\pi}\sigma_w\sigma_x}\int f(x)e^{-x^2/2\sigma_w^2\sigma_x^2}\,\mathrm{d}x = 0.\end{aligned}$$

On the other hand, the second integral in (27) is equal to

$$\begin{aligned}\frac{\sigma_w\sigma_x}{2\pi}\int \widehat{f}'(t)e^{-\sigma_w^2\sigma_x^2\,t^2/2}\,\mathrm{d}t &= \frac{1}{\sqrt{2\pi}}\int f'(x)e^{-x^2/2\sigma_w^2\sigma_x^2}\,\mathrm{d}x\\ &= \sigma_w\sigma_x\int f'(\sigma_w\sigma_x x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x = \theta_2(f)^{1/2}.\end{aligned}$$

Thus, we obtain

$$\kappa(Y_{i_1i_2}, Y^*_{i_2i_3}, Y_{i_3i_4}, Y^*_{i_4i_1}) \approx \mathbf{E}Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1} = \theta_2(f)^2 n_0^{-1} + \mathcal{O}(n_0^{-2}),$$

which proves (18). □

B Proof of Theorem 2.5

B.1 Derivation of the self-consistent equation

We proceed as in Subsection A.1. We know from (15) that

$$\frac{1}{m}\sum_{i=1}^m\left(\frac{Y^*GY}{m}\right)_{ii} = \frac{n_1}{m}\left\langle\frac{YY^*G}{m}\right\rangle = \frac{n_1}{m}(1+zg). \tag{28}$$

We further claim the following.

Lemma B.1. *It holds that*

$$\frac{1}{m}\sum_{i=1}^m\sum_{j=1}^{n_1}\left(\frac{Y^*GY}{m}\right)_{ij} = 1 + \mathcal{O}\left((\theta_{1,b}(f)\,n_1)^{-1}\right). \tag{29}$$

Together with (28), Lemma B.1 implies

$$\frac{1}{m}\sum_{i\ne j}\left(\frac{Y^*GY}{m}\right)_{ij} \approx 1 - \frac{n_1}{m}(1+zg). \tag{30}$$

Proof. Using the Woodbury matrix identity[2], we have

$$\frac{1}{m}\left(\frac{Y^*GY}{m}\right) = \frac{1}{m^2}Y^*\left(\frac{YY^*}{m}-z\right)^{-1}Y = \frac{1}{m}+\frac{z}{m}\left(\frac{Y^*Y}{m}-z\right)^{-1},$$

which implies

$$\sum_{i,j}\frac{1}{m}\left(\frac{Y^*GY}{m}\right)_{ij} = \sum_{i,j}\frac{1}{m}\delta_{ij}+\sum_{i,j}\frac{z}{m}\left(\frac{Y^*Y}{m}-z\right)^{-1}_{ij} = 1+\sum_{i,j}\frac{z}{m}\left(\frac{Y^*Y}{m}-z\right)^{-1}_{ij}.$$

So, we need to show that $\sum_{i,j}\frac{z}{m}\left(\frac{Y^*Y}{m}-z\right)^{-1}_{ij}$ is approximately zero. Let $e := \frac{1}{\sqrt{m}}[1\ \cdots\ 1]^T$ be a normalized vector in $\mathbb{R}^m$. We then write

$$\sum_{i,j}\frac{z}{m}\left(\frac{Y^*Y}{m}-z\right)^{-1}_{ij} = z\,\langle e, \left(\frac{Y^*Y}{m}-z\right)^{-1}e\rangle.$$

It turns out that e is approximately an eigenvector of $\frac{1}{m}Y^*Y$. Indeed, it holds that

$$\mathbf{E}\left(\frac{Y^*Y}{m}e\right)_i = \frac{1}{m\sqrt{m}}\sum_{j=1}^{m}\sum_{k=1}^{n_1}\mathbf{E}\,Y^*_{ik}Y_{kj} \approx m^{-1/2}n_1\,\theta_{1,b}(f) = (n_1\,\theta_{1,b}(f))e_i.$$

Moreover, the variance is approximately $\mathcal{O}(n_1/m)$, which means that the standard deviation is of order 1, while the expectation of order n_1. Thus, e is approximately an eigenvector of $\frac{1}{m}Y^*Y$ with eigenvalue $n_1\theta_{1,b}(f)$. Since $\theta_{1,b}(f)$ is nonzero by assumption, we have that e is approximately an eigenvector of the matrix $\left(\frac{Y^*Y}{m}-z\mathbf{1}_m\right)^{-1}$ with eigenvalue $(n_1\theta_{1,b}(f)-z)^{-1}$, from which the result follows:

$$\left|\langle e, \left(\frac{Y^*Y}{m}-z\right)^{-1}e\rangle\right| \approx \left|(n_1\,\theta_{1,b}(f)-z)^{-1}\right| \ll 1. \qquad \square$$

Given Lemma B.1 and Proposition 3.3, we can now prove the global law for the random matrix M with the cycle correlations.

Proof of Theorem 2.5. Applying Proposition 3.3 to (16) we have

$$\begin{aligned}
1+zg &= \frac{1}{n_1m}\sum_{i_1,i_2}\kappa(Y_{i_1i_2},Y^*_{i_2i_1})\,\partial_{Y^*_{i_2i_1}}(Y^*G)_{i_2i_1} + \frac{1}{n_1m}\sum_{\substack{i_1,i_2,i_3\\ i_2\neq i_3}}\kappa(Y_{i_1i_2},Y^*_{i_3i_1})\,\partial_{Y^*_{i_3i_1}}(Y^*G)_{i_2i_1}\\
&\quad+\frac{1}{n_1m}\sum_{k\geq 2}\sum_{i_1,\ldots,i_{2k}}\kappa(Y_{i_1i_2},\ldots,Y^*_{i_{2k}i_1})\,\partial_{Y^*_{i_2i_3}}\cdots\partial_{Y^*_{i_{2k}i_1}}(Y^*G)_{i_2i_1}\\
&\approx \frac{\theta_1(f)}{n_1m}\sum_{i_1,i_2}\partial_{Y^*_{i_2i_1}}(Y^*G)_{i_2i_1} + \frac{\theta_{1,b}(f)}{n_1m}\sum_{i_1}\sum_{i_2\neq i_3}\partial_{Y^*_{i_3i_1}}(Y^*G)_{i_2i_1}\\
&\quad+\frac{1}{n_1m}\sum_{k\geq 2}\frac{\theta_2^k(f)}{n_0^{k-1}}\sum_{i_1,\ldots,i_{2k}}\partial_{Y^*_{i_2i_3}}\cdots\partial_{Y^*_{i_{2k}i_1}}(Y^*G)_{i_2i_1},
\end{aligned} \tag{31}$$

where we omitted reference to $\mathbf{E}$ to simplify notation. Given Lemma A.1, we only need to compute $\partial_{Y^*_{i_3i_1}}(Y^*G)_{i_2i_1}$:

$$\partial_{Y^*_{i_3i_1}}(Y^*G)_{i_2i_1} = \sum_{j=1}^{n_1}\partial_{Y^*_{i_3i_1}}\left(Y^*_{i_2j}G_{ji_1}\right) \approx -G_{i_1i_1}\left(\frac{Y^*GY}{m}\right)_{i_2i_3},$$

[2]For $A\in\mathbb{R}^{n\times n}$, $C\in\mathbb{R}^{r\times r}$, $U\in\mathbb{R}^{n\times r}$ and $V\in\mathbb{R}^{r\times n}$ the *Woodbury matrix identity* is given by

$$(A+UCV)^{-1} = A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.$$

where we omitted the contribution of $\partial_{Y^*_{i_3 i_1}} Y^*_{i_2 j}$ since it is very small. Plugging the partial derivatives into (31), we get

$$
\begin{aligned}
1 + zg \approx{}& \frac{\theta_1(f)}{n_1 m} \sum_{i_1, i_2} G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right) - \frac{\theta_{1,b}(f)}{n_1 m} \sum_{i_1} \sum_{i_2 \neq i_3} G_{i_1 i_1} \left(\frac{Y^* G Y}{m} \right)_{i_2 i_3} \\
& - \frac{1}{n_1 m} \sum_{k \geq 2} \frac{\theta_2^k(f)}{n_0^{k-1}} \sum_{i_1, \ldots, i_{2k}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} \left(\frac{GY}{m} \right)_{i_3 i_{2k}} G_{i_1 i_1} \left(1 - \left(\frac{Y^* G Y}{m} \right)_{i_2 i_2} \right) \\
\approx{}& \theta_1(f) g \left(1 - \frac{n_1}{m}(1 + zg) \right) - \theta_{1,b}(f) g \left(1 - \frac{n_1}{m}(1 + zg) \right) \\
& - g \left(1 - \frac{n_1}{m}(1 + zg) \right) \sum_{k \geq 2} \frac{\theta_2^k}{n_0^{k-1}} \sum_{i_3, \ldots, i_{2k}} \partial_{Y_{i_3 i_4}} \cdots \partial_{Y_{i_{2k-1} i_{2k}}} \left(\frac{GY}{m} \right)_{i_3 i_{2k}},
\end{aligned}
$$

where in the second step we used (28) and (30). Finally, by shifting the index in the summation and doing some simple bookkeeping, we have

$$
\begin{aligned}
1 + zg \approx{}& (\theta_1 - \theta_{1,b}) g \left(1 - \frac{n_1}{m}(1 + zg) \right) - \theta_2 \frac{n_1}{n_0} g (1 + zg) \left(1 - \frac{n_1}{m}(1 + zg) \right) \\
& + \theta_2 (\theta_1 - \theta_{1,b} - \theta_2) \frac{n_1}{n_0} g^2 \left(1 - \frac{n_1}{m}(1 + zg) \right)^2,
\end{aligned}
$$

which corresponds to the self-consistent equation (6) as $n_0, n_1, m \to \infty$, where θ_1 is replaced by $\theta_1 - \theta_{1,b}$. In the same way as in the bias-free case, the concentration inequality of Lemma 3.4 can also be applied here, thereby concluding that g is approximately equal to its mean with high probability. The first claim of Theorem 2.5 then follows. The second claim follows easily from Lemma B.1. Since $n_1 \theta_{1,b}(f)$ is approximately an eigenvalue of the random matrix $\frac{1}{m} Y^* Y$, and since the nonzero eigenvalues of $Y^* Y$ are the same as the one of $Y Y^*$, we have that $\lambda_{\max} \approx n_1 \theta_{1,b}(f)$ is an eigenvalue of M located away from the rest of the spectrum (called *outlier*). This concludes the proof of Theorem 2.5. □

B.2 Proof of Proposition 3.3

In light of the central limit theorem, the random variables $\frac{(WX)_{ij}}{\sqrt{n_0}} + B_i$, for $1 \leq i \leq n_1$ and $1 \leq j \leq m$, are approximately normally distributed with zero mean and variance $\sigma_w^2 \sigma_x^2 + \sigma_b^2$. In contrast to the bias-free case, here we have two different nonzero second cumulants of the entries of the random matrix $\frac{WX}{\sqrt{n_0}} + B$, and therefore also of the Y_{ij}'s.

Proof of (19). The first identity follows in a straightforward manner by assumption (8):

$$
\kappa(Y_{ij}) = \mathbf{E} Y_{ij} = \int_{\mathbb{R}} f(x) \frac{e^{-x^2/2(\sigma_w^2 \sigma_x^2 + \sigma_b^2)}}{\sqrt{2\pi(\sigma_w^2 \sigma_x^2 + \sigma_b^2)}} \, \mathrm{d}x = 0.
$$

For the second cumulant, we first compute

$$
\begin{aligned}
\kappa \left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + B_{i_1}, \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} + B_{i_3} \right) &= \mathbf{E} \left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + B_{i_1} \right) \left(\frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} + B_{i_3} \right) \\
&= \frac{1}{n_0} \mathbf{E} (WX)_{i_1 i_2} (WX)_{i_3 i_4} + \mathbf{E} B_{i_1} B_{i_3} \\
&= \delta_{i_1 i_3} \delta_{i_2 i_4} \, \sigma_w^2 \sigma_x^2 + \delta_{i_1 i_3} \sigma_b^2.
\end{aligned}
$$

For $i_1 = i_3$ and $i_2 = i_4$, the cumulant $\kappa(Y_{i_1 i_2}, Y^*_{i_2 i_1})$ follows easily:

$$
\kappa(Y_{i_1 i_2}, Y^*_{i_2 i_1}) \approx \int_{\mathbb{R}} f^2(x) \frac{e^{-x^2/2(\sigma_w^2 \sigma_x^2 + \sigma_b^2)}}{\sqrt{2\pi(\sigma_w^2 \sigma_x^2 + \sigma_b^2)}} \, \mathrm{d}x = \theta_1(f).
$$

On the other hand, for $i_1 = i_3$ and $i_2 \neq i_4$, to compute the cumulant $\kappa(Y_{i_1 i_2}, Y^*_{i_4 i_1})$, we need the characteristic function of $\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + B_{i_1}$ and $\frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} + B_{i_1}$. Using the definition of multivariate

cumulants, we know that the log-characteristic function is given by

$$\sum_{l_1,l_2} \kappa\left(\left\{\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}\right\}^{l_1}, \left\{\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}+B_{i_1}\right\}^{l_2}\right)\frac{(it_1)^{l_1}(it_2)^{l_2}}{l_1!\,l_2!},$$

for $t_1, t_2 \in \mathbb{R}$. When $l_1 = l_2 = 1$, we have

$$\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}, \frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}+B_{i_1}\right) \approx \sigma_b^2,$$

and if $l_1 = 2$ and $l_2 = 0$, or vice versa, it holds that

$$\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}, \frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}\right) \approx \sigma_w^2\sigma_x^2+\sigma_b^2.$$

As all other cumulants vanish to leading order, the characteristic function turns out to be asymptotically equal to

$$\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}(t_1^2+t_2^2)-\sigma_b^2t_1t_2\right).$$

Now, we can compute the cumulant of $Y_{i_1i_2}$ and $Y^*_{i_4i_1}$:

$$\begin{aligned}\kappa(Y_{i_1i_2}, Y^*_{i_4i_1}) &\approx \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} f(x_1)f(x_2)e^{-i\boldsymbol{t}\cdot\boldsymbol{x}}\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}(t_1^2+t_2^2)-\sigma_b^2t_1t_2\right)\mathrm{d}\boldsymbol{t}\,\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\hat{f}(t_1)\hat{f}(t_2)\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}(t_1^2+t_2^2)-\sigma_b^2t_1t_2\right)\mathrm{d}t_1\,\mathrm{d}t_2,\end{aligned}$$

where in the second step we applied the Fourier inversion theorem. We denote the covariance matrix Σ by

$$\Sigma := \begin{pmatrix}\sigma_w^2\sigma_x^2+\sigma_b^2 & \sigma_b^2\\ \sigma_b^2 & \sigma_w^2\sigma_x^2+\sigma_b^2\end{pmatrix} \tag{32}$$

with determinant $\det(\Sigma) = \sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2+2\sigma_b^2)$ and inverse matrix

$$\Sigma^{-1} = \frac{1}{\det(\Sigma)}\begin{pmatrix}\sigma_w^2\sigma_x^2+\sigma_b^2 & -\sigma_b^2\\ -\sigma_b^2 & \sigma_w^2\sigma_x^2+\sigma_b^2\end{pmatrix}.$$

Again applying the Fourier inversion formula, we obtain

$$\begin{aligned}\kappa(Y_{i_1i_2}, Y^*_{i_4i_1}) &\approx \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\hat{f}(t_1)\hat{f}(t_2)e^{-\frac{1}{2}\langle\boldsymbol{t},\Sigma\boldsymbol{t}\rangle}\mathrm{d}\boldsymbol{t}\\ &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} f(x_1)f(x_2)\frac{2\pi}{\sqrt{\det(\Sigma)}}e^{-\frac{1}{2}\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{x}\\ &= \frac{1}{2\pi\sqrt{\sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2+2\sigma_b^2)}}\int_{\mathbb{R}^2} f(x_1)f(x_2)e^{-\frac{1}{2}\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{x} = \theta_{1,b}(f),\end{aligned}$$

where

$$e^{-\frac{1}{2}\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle} = \exp\left(-\frac{(\sigma_w^2\sigma_x^2+\sigma_b^2)(x_1^2+x_2^2)-2\sigma_b^2x_1x_2}{2\sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2+2\sigma_b^2)}\right).$$

This shows (19). □

To complete the proof of Proposition 3.3, it remains to compute the cumulant of $Y_{i_1i_2}$, $Y^*_{i_2i_3}$, …, $Y^*_{i_{2k}i_1}$ with $k > 1$ and $i_1, \dots, i_{2k}$ distinct.

Proof of (20). For notational simplicity, we prove the statement for $k = 2$. First, we use the cumulant asymptotics in order to asymptotically compute the characteristic function. The cumulants have exactly the same values as in the bias-free case (see Appendix D), except for

$$\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}, \frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1}\right) \approx \sigma_w^2\sigma_x^2+\sigma_b^2.$$

In addition to all these cumulants, we also have

$$\kappa\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}+B_{i_1},\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}+B_{i_1}\right)=\kappa\left(\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}+B_{i_3},\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}+B_{i_3}\right)\approx\sigma_b^2.$$

Therefore, the log-characteristic function is given by

$$\begin{aligned}&-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}\sum_{i=1}^4 t_i^2-\sigma_b^2(t_1t_4+t_2t_3)+\sum_{n\geq 1}\frac{(-1)^{n-1}}{n}\left(\frac{(\sigma_w^2\sigma_x^2)^2}{n_0}\prod_{i=1}^4 t_i+\mathcal{O}(n_0^{-2})\right)^n\\&=-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}\sum_{i=1}^4 t_i^2-\sigma_b^2(t_1t_4+t_2t_3)+\log\left(1+\frac{(\sigma_w^2\sigma_x^2)^2}{n_0}\prod_{i=1}^4 t_i+\mathcal{O}(n_0^{-2})\right),\end{aligned}$$

for $t_1,t_2,t_3,t_4\in\mathbb{R}$ such that $t_1t_2t_3t_4<n_0$. We obtain the characteristic function by taking the exponential of the above expression. By the same argument as in the proof of Proposition A.2, we have

$$\mathbf{E}Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1}=\frac{1}{(2\pi)^4}\int_{\forall i,\,|t_i|\leq n_0^{\frac14}}\prod_{i=1}^4\hat{f}(t_i)\,\varphi(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t}+\mathcal{O}(n_0^{-2}),$$

where $\varphi(\boldsymbol{t})$ denotes the characteristic function. Next, we compute

$$\begin{aligned}&\mathbf{E}\,Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1}\\&=\frac{1}{(2\pi)^4}\int\prod_{i=1}^4\hat{f}(t_i)\left(1+\frac{(\sigma_w^2\sigma_x^2)^2}{n_0}\prod_{i=1}^4 t_i\right)\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}\sum_{i=1}^4 t_i^2-\sigma_b^2(t_1t_4+t_2t_3)\right)\mathrm{d}\boldsymbol{t}\\&\quad+\mathcal{O}(n_0^{-2})\\&=\left(\frac{1}{(2\pi)^2}\int\hat{f}(t_1)\hat{f}(t_2)\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}(t_1^2+t_2^2)-\sigma_b^2t_1t_2\right)\mathrm{d}t_1\mathrm{d}t_2\right)^2\\&\quad+\frac{1}{n_0}\left(\frac{\sigma_w^2\sigma_x^2}{(2\pi)^2}\int\widehat{f}'(t_1)\widehat{f}'(t_2)\exp\left(-\frac{\sigma_w^2\sigma_x^2+\sigma_b^2}{2}(t_1^2+t_2^2)-\sigma_b^2t_1t_2\right)\mathrm{d}t_1\mathrm{d}t_2\right)^2+\mathcal{O}(n_0^{-2})\\&=\left(\frac{1}{2\pi\sqrt{\sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2+2\sigma_b^2)}}\int f(x_1)f(x_2)e^{-\frac12\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{x}\right)^2\\&\quad+\frac{1}{n_0}\left(\frac{\sigma_w^2\sigma_x^2}{2\pi\sqrt{\sigma_w^2\sigma_x^2(\sigma_w^2\sigma_x^2+2\sigma_b^2)}}\int f'(x_1)f'(x_2)e^{-\frac12\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{x}\right)^2+\mathcal{O}(n_0^{-2}),\end{aligned}$$

where Σ is the matrix defined by (32). It then follows that

$$\begin{aligned}\kappa(Y_{i_1i_2},Y^*_{i_2i_3},Y_{i_3i_4},Y^*_{i_4i_1})&=\mathbf{E}Y_{i_1i_2}Y^*_{i_2i_3}Y_{i_3i_4}Y^*_{i_4i_1}-\mathbf{E}Y_{i_1i_2}Y^*_{i_4i_1}\,\mathbf{E}\,Y^*_{i_2i_3}Y_{i_3i_4}\\&=\frac{1}{n_0}\left(\frac{\sigma_w\sigma_x}{2\pi\sqrt{\sigma_w^2\sigma_x^2+2\sigma_b^2}}\int f'(x_1)f'(x_2)e^{-\frac12\langle\boldsymbol{x},\Sigma^{-1}\boldsymbol{x}\rangle}\mathrm{d}\boldsymbol{x}\right)^2\\&\quad+\mathcal{O}(n_0^{-2})\\&=\theta_2(f)^2\,n_0^{-1}+\mathcal{O}(n_0^{-2}),\end{aligned}$$

as desired. The proof for $k>2$ is similar. □

C Proofs of auxiliary results

Proof of Lemma 3.1. By applying the Fourier inversion theorem, we have

$$\mathbf{E}X_1f(\boldsymbol{X})=\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\int_{\mathbb{R}^n}x_1f(\boldsymbol{x})e^{-i\boldsymbol{t}\cdot\boldsymbol{x}}\varphi_{\boldsymbol{X}}(\boldsymbol{t})\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{t},$$

where $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$ is the characteristic function of the n-dimensional random vector $\boldsymbol{X}$. It holds that $\int_{\mathbb{R}^n}(-ix_1)f(\boldsymbol{x})e^{-i\boldsymbol{t}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x} = \partial_{t_1}\hat{f}(\boldsymbol{t})$. Then, it follows that

$$\begin{aligned}
\mathbf{E}X_1 f(\boldsymbol{X}) &= \frac{i}{(2\pi)^n}\int_{\mathbb{R}^n}\Big(\partial_{t_1}\hat{f}(\boldsymbol{t})\Big)\varphi_{\boldsymbol{X}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \\
&= -\frac{i}{(2\pi)^n}\int_{\mathbb{R}^n}\hat{f}(\boldsymbol{t})\Big(\partial_{t_1}\varphi_{\boldsymbol{X}}(\boldsymbol{t})\Big)\mathrm{d}\boldsymbol{t} \\
&= -\frac{i}{(2\pi)^n}\int_{\mathbb{R}^n}\hat{f}(\boldsymbol{t})\Big(\partial_{t_1}e^{\log\varphi_{\boldsymbol{X}}(\boldsymbol{t})}\Big)\mathrm{d}\boldsymbol{t} \\
&= -\frac{i}{(2\pi)^n}\int_{\mathbb{R}^n}\hat{f}(\boldsymbol{t})\Big(\partial_{t_1}\log\varphi_{\boldsymbol{X}}(\boldsymbol{t})\Big)\varphi_{\boldsymbol{X}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t}.
\end{aligned}$$

Cumulants can also be defined in an analytical way as the coefficients of the $\log$-characteristic function

$$\log\mathbf{E}e^{i\boldsymbol{t}\cdot\boldsymbol{X}} = \sum_{\boldsymbol{l}}\kappa_{\boldsymbol{l}}\frac{(i\boldsymbol{t})^{\boldsymbol{l}}}{\boldsymbol{l}!}, \tag{33}$$

where $\sum_{\boldsymbol{l}}$ is the sum over all multi-indices $\boldsymbol{l} = (l_1,\dots,l_n)\in\mathbb{N}^n$. We note that $\kappa_{\boldsymbol{l}}(X_1,\dots,X_n) = \kappa(\{X_1\}^{l_1},\dots,\{X_n\}^{l_n})$ means that X_i appears l_i times. One can prove that this definition of cumulants is equivalent to the combinatorial one given by 14 (see [23] for a proof). Using definition (33) results in

$$\partial_{t_1}\log\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = i\sum_{\boldsymbol{l}}\kappa_{\boldsymbol{l}+\boldsymbol{e_1}}\frac{(i\boldsymbol{t})^{\boldsymbol{l}}}{\boldsymbol{l}!},$$

where $\boldsymbol{l}+\boldsymbol{e_1} = (l_1+1, l_2,\dots,l_n)$. Since $(i\boldsymbol{t})^{\boldsymbol{l}}\hat{f}(\boldsymbol{t}) = \widehat{f^{(\boldsymbol{l})}}(\boldsymbol{t})$, we finally obtain

$$\mathbf{E}X_1 f(\boldsymbol{X}) = \sum_{\boldsymbol{l}}\frac{\kappa_{\boldsymbol{l}+\boldsymbol{e_1}}}{\boldsymbol{l}!}\frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\widehat{f^{(\boldsymbol{l})}}(\boldsymbol{t})\varphi_{\boldsymbol{X}}(\boldsymbol{t})\mathrm{d}\boldsymbol{t} = \sum_{\boldsymbol{l}}\frac{\kappa_{\boldsymbol{l}+\boldsymbol{e_1}}}{\boldsymbol{l}!}\mathbf{E}\,f^{(\boldsymbol{l})}(\boldsymbol{X}),$$

where we again applied the Fourier inversion formula. □

Proof of Lemma A.1. Let $\Delta^{i,j}$ denote a $m\times n_1$ matrix such that $\Delta^{i,j}_{kl} = \mathbf{1}_{\{(i,j)=(k,l)\}}$. Then, applying the resolvent identity, we get

$$\frac{\partial G}{\partial Y^*_{ij}} = \lim_{\epsilon\to 0}\frac{\left(\frac{Y(Y^*+\epsilon\Delta^{i,j})}{m}-z\right)^{-1}-\left(\frac{YY^*}{m}-z\right)^{-1}}{\epsilon} = -\frac{GY\Delta^{i,j}G}{m}.$$

It follows that $\partial_{Y^*_{ij}}G_{ab} = -\left(\frac{GY}{m}\right)_{ai}G_{jb}$ for $1\le a,b\le n_1$, $1\le i\le m$, and $1\le j\le n_1$. Therefore, we have

$$\partial_{Y^*_{i_2i_1}}(Y^*G)_{i_2i_1} = \sum_{j=1}^{n_1}\partial_{Y^*_{i_2i_1}}\left(Y^*_{i_2j}G_{ji_1}\right) = G_{i_1i_1}\left(1-\left(\frac{Y^*GY}{m}\right)_{i_2i_2}\right),$$

which proves (3.6a). We now compute

$$\begin{aligned}
\sum_{j=1}^{n_1}\partial_{Y^*_{i_2i_3}}\partial_{Y^*_{i_{2k}i_1}}\left(Y^*_{i_2j}G_{ji_1}\right) &\approx -\sum_{j=1}^{n_1}\partial_{Y^*_{i_2i_3}}\left(Y^*_{i_2j}\left(\frac{GY}{m}\right)_{ji_{2k}}G_{i_1i_1}\right) \\
&\approx -\left(\frac{GY}{m}\right)_{i_3i_{2k}}G_{i_1i_1}+\left(\frac{Y^*GY}{m}\right)_{i_2i_2}\left(\frac{GY}{m}\right)_{i_3i_{2k}}G_{i_1i_1},
\end{aligned}$$

where the approximation in the first line comes from the fact that the contribution of $\partial_{Y^*_{i_{2k}i_1}}Y^*_{i_2j}$ is very small and can therefore be neglected. Since the off-diagonals of the resolvent of random matrices are small if $\Im z \gg n_1^{-1}$, the partial derivative $\partial_{Y^*_{i_2i_3}}G_{i_1i_1}$ can be omitted. This justifies the second approximation. So, we obtain

$$\partial_{Y^*_{i_2i_3}}\cdots\partial_{Y^*_{i_{2k}i_1}}(Y^*G)_{i_2i_1} \approx -\partial_{Y_{i_3i_4}}\cdots\partial_{Y_{i_{2k-1}i_{2k}}}\left(\frac{GY}{m}\right)_{i_3i_{2k}}G_{i_1i_1}\left(1-\left(\frac{Y^*GY}{m}\right)_{i_2i_2}\right),$$

which completes the proof of Lemma A.1. □

D Computation of the characteristic function

Proof of Lemma A.4. We prove Lemma A.4 for $k = 2$ in what follows. Let $t_1, t_2, t_3, t_4 \in \mathbb{R}$ be such that $\prod_{i=1}^4 t_i < n_0$. The log-characteristic function is given by

$$\begin{aligned}
&\log \mathbf{E} \exp\left(it_1 \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + it_2 \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} + it_3 \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} + it_4 \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right) \\
&= \sum_{l_1 + \cdots + l_4 \geq 1} \kappa\left(\left\{ \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right\}^{l_1}, \left\{ \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} \right\}^{l_2}, \left\{ \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} \right\}^{l_3}, \left\{ \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right\}^{l_4} \right) \\
&\qquad \cdot \frac{(it_1)^{l_1} (it_2)^{l_2} (it_3)^{l_3} (it_4)^{l_4}}{l_1! \, l_2! \, l_3! \, l_4!},
\end{aligned} \tag{34}$$

thus we first need to compute the cumulants and claim that only five terms contribute to leading order, $l_1 = 2, l_2 = l_3 = l_4 = 0$ (resp. $l_2 = 2, l_1 = l_3 = l_4 = 0$, resp. $l_3 = 2, l_1 = l_2 = l_4 = 0$, resp. $l_4 = 2, l_1 = l_2 = l_3 = 0$) and $l_1 = l_2 = l_3 = l_4 = 1$, as stated in the following lemma.

Lemma D.1. *Let $l_1, l_2, l_3, l_4 \in \mathbb{N}$ such that $l_1 + l_2 + l_3 + l_4 \geq 1$. Then, it holds that*

$$\begin{aligned}
&\kappa\left(\left\{ \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right\}^{l_1}, \left\{ \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} \right\}^{l_2}, \left\{ \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} \right\}^{l_3}, \left\{ \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right\}^{l_4} \right) \\
&= \mathcal{O}(n_0^{-2}) + \begin{cases} \sigma_w^2 \sigma_x^2 & \text{if } l_i = 2, l_j = 0 \text{ for some } i \text{ and all } j \neq i \\ \frac{(\sigma_w^2 \sigma_x^2)^2}{n_0} & \text{if } l_1 = l_2 = l_3 = l_4 = 1 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}$$

where $\{X\}^n$ denotes n copies of the random variable X.

Using Lemma D.1 the log-characteristic function is given by

$$\begin{aligned}
&\log \mathbf{E} \exp\left(it_1 \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} + it_2 \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} + it_3 \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} + it_4 \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right) \\
&= -\frac{\sigma_w^2 \sigma_x^2}{2} (t_1^2 + t_2^2 + t_3^2 + t_4^2) + \frac{(\sigma_w^2 \sigma_x^2)^2}{n_0} t_1 t_2 t_3 t_4 + \mathcal{O}(n_0^{-2}) \\
&= -\frac{\sigma_w^2 \sigma_x^2}{2} (t_1^2 + t_2^2 + t_3^2 + t_4^2) + \log\left(1 + \frac{(\sigma_w^2 \sigma_x^2)^2}{n_0} t_1 t_2 t_3 t_4 + \mathcal{O}\left(\left(\frac{(\sigma_w^2 \sigma_x^2)}{n_0} t_1 t_2 t_3 t_4 \right)^2 \right) \right),
\end{aligned}$$

where in the last equality we used the assumption that $t_1 t_2 t_3 t_4 < n_0$. We get the characteristic function by taking the exponential of the above expression. This completes the proof of Lemma A.4 for $k = 2$. The proof for $k > 2$ follows by the same argument. □

Proof of Lemma D.1. We start with the case where (say) $l_1 = 2, l_2 = l_3 = l_4 = 0$, which gives

$$\kappa\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right) = \mathbf{E}\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right)^2 = \sigma_w^2 \sigma_x^2.$$

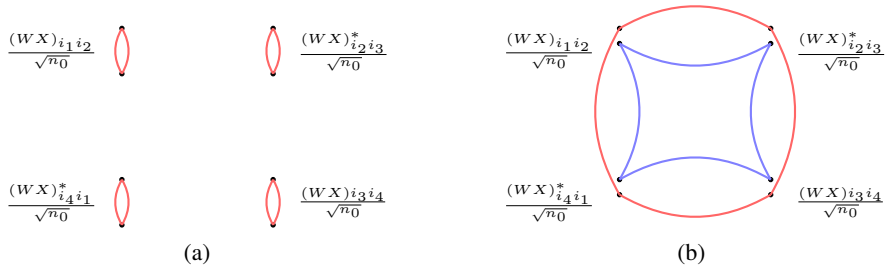
When considering $l_1 = l_2 = l_3 = l_4 = 1$, we already showed in Lemma A.3 that

$$\kappa\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}}, \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}}, \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}}, \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right) = \frac{(\sigma_w^2 \sigma_x^2)^2}{n_0} + \mathcal{O}(n_0^{-2}).$$

We now consider the case with $l_1 = l_2 = l_3 = l_4 = 2$ and compute

$$\begin{aligned}
&\mathbf{E}\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right)^2 \left(\frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} \right)^2 \left(\frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} \right)^2 \left(\frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right)^2 \\
&\approx \mathbf{E}\left(\frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \right)^2 \mathbf{E}\left(\frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} \right)^2 \mathbf{E}\left(\frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} \right)^2 \mathbf{E}\left(\frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right)^2 \\
&\quad + 8 \left(\mathbf{E} \frac{(WX)_{i_1 i_2}}{\sqrt{n_0}} \frac{(WX)^*_{i_2 i_3}}{\sqrt{n_0}} \frac{(WX)_{i_3 i_4}}{\sqrt{n_0}} \frac{(WX)^*_{i_4 i_1}}{\sqrt{n_0}} \right)^2
\end{aligned}$$

The expectation of the product of the square of the four random variables has then two main contributions which can be illustrated by coding the eight random variables as a graph: we can either couple the equal random variables, thus resulting in four cycles of length two, or we can obtain two cycles of the four different random variables. For the latter, there are eight different ways for drawing two distinct cycles of length four. See Fig. 5 for an illustration. We therefore obtain



(a) (b)

Figure 5: The two main contributions when computing the expectation of the product of the square of the four random variables. There is one way of coupling equal random variables (case (a)), and there are eight ways to draw two cycles of length four (case (b)).

$$
\begin{aligned}
&\kappa\left(\left\{\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}\right\}^2, \left\{\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}\right\}^2, \left\{\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}\right\}^2, \left\{\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right\}^2\right) \\
&\approx \mathbf{E}\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}\right)^2\left(\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}\right)^2\left(\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}\right)^2\left(\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right)^2 \\
&\quad - 8\left(\mathbf{E}\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right)^2 \\
&\quad - \mathbf{E}\left(\frac{(WX)_{i_1i_2}}{\sqrt{n_0}}\right)^2\mathbf{E}\left(\frac{(WX)^*_{i_2i_3}}{\sqrt{n_0}}\right)^2\mathbf{E}\left(\frac{(WX)_{i_3i_4}}{\sqrt{n_0}}\right)^2\mathbf{E}\left(\frac{(WX)^*_{i_4i_1}}{\sqrt{n_0}}\right)^2 = \mathcal{O}(n_0^{-3}).
\end{aligned}
$$

Similarly, it follows that all other cumulants with multiple occurrence of at least two of the random variables are negligible since their joint moment to leading order factorises into a product of lower order moments. □

E Concentration inequality

Proof of Lemma 3.4. Without loss of generality, it suffices to prove the statement w.r.t. $\mathbf{E}_X$ since by cyclicity the statement for $\mathbf{E}_W$ is analogous. We write $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)$ with $\boldsymbol{x}_k = (x_{1k}, \dots, x_{n_0k})'$, and similarly, $Y = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_m)$. We denote by $\mathcal{F}_k$, $1 \le k \le m$, the filtration generated by $\{\boldsymbol{x}_l,\, 1 \le l \le k\}$ and by $\mathbf{E}_k[\cdot] := \mathbf{E}_X[\cdot \mid \mathcal{F}_k]$ the conditional expectation w.r.t. $\mathcal{F}_k$. Now, we decompose $g(z) - \mathbf{E}_X g(z)$ as a sum of martingale differences

$$D_k := \mathbf{E}_k \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} - \mathbf{E}_{k-1} \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1}, \quad \text{for } k = 1, \dots, m.$$

By construction, we have $\mathbf{E}_m \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} = \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1}$ and $\mathbf{E}_0 \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} = \mathbf{E}_X \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1}$. It then follows that

$$g(z) - \mathbf{E}_X g(z) = \frac{1}{n_1}\sum_{k=1}^m \mathbf{E}_k \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} - \mathbf{E}_{k-1} \operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} = \frac{1}{n_1}\sum_{k=1}^m D_k.$$

Next, we define $M_k := M - \boldsymbol{y}_k\boldsymbol{y}_k^*$. We note that

$$\mathbf{E}_k \operatorname{Tr}(M_k - z\mathbf{1}_{n_1})^{-1} = \mathbf{E}_{k-1} \operatorname{Tr}(M_k - z\mathbf{1}_{n_1})^{-1},$$

since M_k is independent of $\boldsymbol{y}_k$ and therefore is also independent of $\boldsymbol{x}_k$. So, we have

$$D_k = (\mathbf{E}_k - \mathbf{E}_{k-1})[\operatorname{Tr}(M - z\mathbf{1}_{n_1})^{-1} - \operatorname{Tr}(M_k - z\mathbf{1}_{n_1})^{-1}].$$

Then, by the Shermann-Morrison formula, we have

$$\begin{aligned}\left|\mathrm{Tr}(M - z\mathbf{1}_{n_1})^{-1} - \mathrm{Tr}(M_k - z\mathbf{1}_{n_1})^{-1}\right| &= \left|\frac{\boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-2}\boldsymbol{y_k}}{1 + \boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-1}\boldsymbol{y_k}}\right| \\ &\leq \frac{|\boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-2}\boldsymbol{y_k}|}{\Im(\boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-1}\boldsymbol{y_k})} \\ &\leq \frac{1}{\Im z},\end{aligned}$$

where the last inequality follows from the resolvent identity:

$$\begin{aligned}|\boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-2}\boldsymbol{y_k}| &\leq \boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-1}(M_k - \bar{z}\mathbf{1}_{n_1})^{-1}\boldsymbol{y_k} \\ &= \frac{\boldsymbol{y}_k^*\left((M_k - z\mathbf{1}_{n_1})^{-1} - (M_k - \bar{z}\mathbf{1}_{n_1})^{-1}\right)\boldsymbol{y}_k}{2i\,\Im z} \\ &= \frac{\Im(\boldsymbol{y}_k^*(M_k - z\mathbf{1}_{n_1})^{-1}\boldsymbol{y_k})}{\Im z}.\end{aligned}$$

Thus, $|D_k| \leq 2(\Im z)^{-1}$, and so $g(z) - \mathbf{E}_X g(z)$ is a sum of bounded martingale differences. We can now apply the Burkholder's inequality which states that for $\{D_k, 1 \leq k \leq m\}$ being a complex-valued martingale difference sequence, for $p > 1$,

$$\mathbf{E}\left|\sum_{k=1}^{m} D_k\right|^p \leq C\,\mathbf{E}\left(\sum_{k=1}^{n} |D_k|^2\right)^{p/2},$$

where C is a positive constant depending on p. We refer to [5, Lemma 2.12] for a proof of this inequality. By choosing $p = 4$, we get

$$\begin{aligned}\mathbf{E}_X\,|g(z) - \mathbf{E}_X g(z)|^4 &= \frac{1}{n_1^4}\,\mathbf{E}_X\left|\sum_{k=1}^{m} D_k\right|^4 \\ &\leq \frac{1}{n_1^4} C\,\mathbf{E}_X\left(\sum_{k=1}^{m} |D_k|^2\right)^2 \\ &\leq \frac{16\,C\,m^2}{n_1^4\,(\Im z)^4} = \mathcal{O}(n_1^{-2}(\Im z)^{-4}),\end{aligned}$$

just as claimed. □

F Complex case

Remark F.1. *We can also consider matrices $X \in \mathbb{C}^{n_0 \times m}$ and $W \in \mathbb{C}^{n_1 \times n_0}$ of complex random entries with zero mean and variance $\mathbf{E}|X_{ij}|^2 = \sigma_x^2$ and $\mathbf{E}|W_{ij}|^2 = \sigma_w^2$. Let $M = \frac{1}{m}YY^*$ with $Y = f\left(\frac{WX}{\sqrt{n_0}}\right)$, and let $f\colon \mathbb{C} \to \mathbb{R}$ be a real-differentiable function satisfying $\int_{\mathbb{C}} f(\sigma_w\sigma_x z)\frac{e^{-|z|^2}}{\pi}\mathrm{d}^2 z = 0$. Set $\theta_1(f) = \int_{\mathbb{C}} |f(\sigma_w\sigma_x z)|^2\,\frac{e^{-|z|^2}}{\pi}\mathrm{d}^2 z$. Then, it can be proved that the normalized trace of the resolvent of M satisfies equation* (7).