Pure and Applied Functional Analysis
Volume 5, Number 6, 2020, 1395–1424

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

ON RANDOM MATRICES ARISING IN DEEP NEURAL NETWORKS: GAUSSIAN CASE

LEONID PASTUR

ABSTRACT. The paper deals with distribution of singular values of product of random matrices arising in the analysis of deep neural networks. The matrices resemble the product analogs of the sample covariance matrices, however, an important difference is that the population covariance matrices, which are assumed to be non-random in the standard setting of statistics and random matrix theory, are now random, moreover, are certain functions of random data matrices. The problem has been considered in recent work [21] by using the techniques of free probability theory. Since, however, free probability theory deals with population matrices which are independent of the data matrices, its applicability in this case requires an additional justification. We present this justification by using a version of the standard techniques of random matrix theory under the assumption that the entries of data matrices are independent Gaussian random variables. In the subsequent paper [18] we extend our results to the case where the entries of data matrices are just independent identically distributed random variables with several finite moments. This, in particular, extends the property of the so-called macroscopic universality on the considered random matrices.

1. Introduction

Deep neural networks with multiple hidden layers have achieved remarkable performance in a wide variety of domains, see e.g. [2–4, 9, 25, 27] for reviews. Among numerous research directions of the field those using random matrices of large size are of considerable amount and interest. They treat random untrained networks (allowing for the study their initialization and learning dynamics, the information propagation through generic deep random neural networks, etc.), the expressivity and the geometry of neural networks, the analysis of the Bayesian approach, etc., see e.g. [7, 10–12, 21–24, 26] and references therein.

Consider an untrained, feed-forward, fully connected neural network with L layers of width n_l for the lth layer and pointwise nonlinearities φ. Let

$$x^0 = \{x^0_{j_0}\}_{j_0=1}^{n_0} \in \mathbb{R}^{n_0} \tag{1.1}$$

be the input to the network, and $x^L = \{x^L_{j_L}\}_{j_L=1}^{n_L} \in \mathbb{R}^{n_L}$ be its output. The components of the activations x^l in the lth layer and the post-affine transformations

2010 *Mathematics Subject Classification.* Primary 15B52; Secondary 92B20.
Key words and phrases. random matrices, deep neural networks.

y^l of x^l are $\{x^l_{j_l}\}_{j_l=1}^{n_l}$ and $\{y^l_{j_l}\}_{j_l=1}^{n_l}$ respectively and are related as follows

$$y^l = W^l x^{l-1} + b^l,\ x^l_{j_l} = \varphi(y^l_{j_l}),\ j_l = 1, ..., n_l,\ l = 1, ..., L, \tag{1.2}$$

where

$$W^l = \{W^l_{j_l j_{l-1}}\}_{j_l, j_{l-1}=1}^{n_l, n_{l-1}},\ l = 1, ..., L \tag{1.3}$$

are $n_l \times n_{l-1}$ rectangular weight matrices,

$$b^l = \{b^l_{j_l}\}_{j_l=1}^{n_l},\ l = 1, 2, ..., L \tag{1.4}$$

are n_l-component bias vectors and $\varphi : \mathbb{R} \to \mathbb{R}$ is the component-wise nonlinearity.

Assume that the biases components $\{b^l_{j_l}\}_{j_l=1}^{n_l}$ are the Gaussian random variables such that:

$$\mathbf{E}\{b^l_{j_l}\} = 0,\ \mathbf{E}\{b^{l_1}_{j_{l_1}} b^{l_2}_{j_{l_2}}\} = \sigma_b^2 \delta_{l_1 l_2} \delta_{j_{l_1} j_{l_2}}. \tag{1.5}$$

As for the weight matrices $W^l,\ l = 1, 2, ..., L$, it is assumed that

$$\begin{aligned} W^l &= n_{l-1}^{-1/2} X^l = n_{l-1}^{-1/2} \{X^l_{j_l j_{l-1}}\}_{j_l, j_{l-1}=1}^{n_l, n_{l-1}}, \\ \mathbf{E}\{X^l_{j_l j_{l-1}}\} &= 0,\ \mathbf{E}\{X^{l_1}_{j_{l_1} j_{l_1-1}} X^{l_2}_{j_{l_2} j_{l_2-1}}\} = \delta_{l_1 l_2} \delta_{j_{l_1} j_{l_1-1}} \delta_{j_{l_2} j_{l_2-1}}, \end{aligned} \tag{1.6}$$

the matrices $X^l,\ l = 1, 2, ..., L$ are independent and identically distributed and for every l we view X^l as the upper left rectangular block of the semi-infinite random matrix

$$\{X^l_{j_l j_{l-1}}\}_{j_l, j_{l-1}=1}^{\infty, \infty} \tag{1.7}$$

with the standard Gaussian entries.

Likewise, for every l we view b^l in (1.4) as the first n_l components of the semi-infinite vector

$$\{b^l_{j_l}\}_{j_l=1}^{\infty} \tag{1.8}$$

whose components are Gaussian random variables normalized by (1.5) with $n_l = \infty,\ l = 1, 2, ..., L$.

As a result of this form of weights and biases of the lth layer they are for all $n_l = 1, 2, ...$ defined on the same infinite-dimensional product probability space Ω^l generated by (1.7) – (1.8). Let also

$$\Omega_l = \Omega^l \times \Omega^{l-1} \times ... \times \Omega^1,\ l = 1, ..., L \tag{1.9}$$

be the infinite-dimensional probability space on which the recurrence (1.2) is defined for a given L (the number of layers). This will allow us to formulate our results on the large size asymptotics of the eigenvalue distribution of matrices (1.12) as those valid with probability 1 in Ω_L.

Note that matrices $W^l(W^l)^T$ of (1.3) and (1.6) are known in statistics as the Wishart matrices [14].

Consider the input-output Jacobian

$$J^L_{\mathbf{n}_L} := \left\{ \frac{\partial x^L_{j_L}}{\partial x^0_{j_0}} \right\}_{j_0, j_L=1}^{n_0, n_L} = \prod_{l=1}^{L} n_{l-1}^{-1/2} D^l X^l,\ \mathbf{n}_L = (n_1, ..., n_L) \tag{1.10}$$

i.e., a $n_L \times n_0$ random matrix, where

$$D^l = \{D^l_{j_l}\delta_{j_l k_l}\}_{j_l,k_l=1}^{n_l},\ D^l_{j_l} = \varphi'\Big(n_{l-1}^{-1/2}\sum_{j_{l-1}=1}^{n_{l-1}} X^l_{j_l j_{l-1}} x^{l-1}_{j_{l-1}} + b^l_{j_l}\Big) \tag{1.11}$$

are diagonal random matrices.

We are interested in the spectrum of singular values of $J^L_{\mathbf{n}_L}$, i.e., the square roots of eigenvalues of

$$M^L_{\mathbf{n}_L} := J^L_{\mathbf{n}_L}(J^L_{\mathbf{n}_L})^T \tag{1.12}$$

for networks with the above random weights and biases and for large $\{n_l\}_{l=1}^L$, i.e., for deep networks with wide layers, see [10, 12, 20–22, 26] for motivations and settings. More precisely, we will study in this paper the asymptotic case determined by the simultaneous limits

$$\lim_{N_l\to\infty}\frac{n_{l-1}}{n_l} = c_l \in (0,\infty),\ n_l \to \infty,\ l = 1, ..., L \tag{1.13}$$

denoted below as

$$\lim_{\mathbf{n}_L\to\infty}. \tag{1.14}$$

Denote $\{\lambda^L_t\}_{t=1}^{n_L}$ the eigenvalues of the $n_L \times n_L$ random matrix $M^L_{\mathbf{n}_L}$ and define its Normalized Counting Measure (NCM) as

$$\nu_{M^L_{\mathbf{n}_L}} := n_L^{-1}\sum_{t=1}^{N_L}\delta_{\lambda^L_t}. \tag{1.15}$$

We will deal with the leading term of $\nu_{M^L_{\mathbf{n}_L}}$ in the asymptotic regime (1.13) – (1.14), i.e., with the limit

$$\nu_{M^L} := \lim_{\mathbf{n}_L\to\infty}\nu_{M^L_{\mathbf{n}_L}} \tag{1.16}$$

if any. Note that since $\nu_{M^L_{\mathbf{n}_L}}$ is random, the meaning of the limit has to be stipulated.

The problem was considered in [21] (see also [10,20]) in the case where all b^l and X^l, $l = 1, 2, ..., L$ in (1.5) – (1.6) are Gaussian and have the same size n and $n \times n$ respectively, i.e.,

$$n := n_0 = ... = n_L. \tag{1.17}$$

We will write in this case n instead of n_l, $l = 0, ..., L$. In [21] compact formulas for the limit

$$\overline{\nu}_{M^L} := \lim_{n\to\infty}\overline{\nu}_{M^L_n},\ \overline{\nu}_{M^L_n} := \mathbf{E}\{\nu_{M^L_n}\} \tag{1.18}$$

and its Stieltjes transform

$$f_{M^L}(z) = \int_{\infty}^{\infty}\frac{\overline{\nu}_{M^L}(d\lambda)}{\lambda - z},\ \Im z \neq 0 \tag{1.19}$$

were proposed. The formula for $\overline{\nu}_{M^L}$ is given in (2.4) below. To write the formula for f_{M^L} it is convenient to use the moment generating function

$$m_{M^L}(z) = \sum_{k=1}^{\infty} m_k z^k, \ m_k = \int_{\infty}^{\infty} \lambda^k \overline{\nu}_{M^L}(d\lambda), \tag{1.20}$$

related to f_{M^L} as

$$m_{M^L}(z) = -1 - z^{-1} f_{M^L}(z^{-1}). \tag{1.21}$$

Let

$$K_n^l := (D_n^l)^2 = \{(D_{j_l}^l)^2\}_{j_l=1}^n \tag{1.22}$$

be the square of the $n \times n$ random diagonal matrix (1.11) with $n_l = n$, denoted D_n^l to make explicit its dependence on n of (1.17), and let m_{K^l} be the moment generating function of the $n \to \infty$ limit $\overline{\nu}_{K^l}$ of the expectation of the NCM of K_n^l. Then we have according to formulas (14) and (16) in [21] in the case where $\overline{\nu}_{K^l}$, hence m_{K^l}, do not depend on l (see Remark 2.2 (i))

$$\begin{aligned} m_{M^L}(z) &= m_K(z^{1/L}\Psi_L(m_{M^L}(z))), \\ \Psi_L(z) &= (1+z)^{1/L} z^{1-1/L}. \end{aligned} \tag{1.23}$$

i.e., f_{M^L} of (1.19)) satisfies a certain functional equation, the standard situation in random matrix theory and its applications, see [17] for general results and [8, 15] for results on the products of random matrices. Note that our notation is different from that of [21]: our $f_{M^L}(z)$ of (1.19) is $-G_X(z)$ of (7) in [21] and our $m_{M^L}(z)$ of (1.20) is $M_X(1/z)$ of (9) in [21].

The derivation of (1.23) and the corresponding formula for the limiting mean NCM $\overline{\nu}_{M^L}$ in [21] was based on the claimed there asymptotic freeness of diagonal matrices $D_{n_l}^l = \{D_{j_l}^l\}_{j_l=1}^{n_l}$, $l = 1, 2..., L$ of (1.11) and Gaussian matrices $X_{n_l}^l$, $l = 1, 2, ..., L$ of (1.3) – (1.6) (see, e.g. [5, 13, 19] for the definitions and properties of asymptotic freeness). This leads directly to (1.23) in view of the multiplicative property of the moment generating functions (1.20) and the so-called S-transforms of $\overline{\nu}_{K^l}$ and of ν_{MP}, the mean limiting NCM's of $K_{n_l}^l$ and of $n^{-1}X_{n_l}^l(X_{n_l}^l)^T$ in the regime (1.13), see Remark 2.2 (ii) and Corollary 3.7.

There is, however, a delicate point in the proof of the above results in [21], since, to the best of our knowledge, the asymptotic freeness has been established so far for the Gaussian random matrices $X_{n_l}^l$ of (1.6) and deterministic (more generally, random but $X_{n_l}^l$-independent) diagonal matrices, see e.g. [5, 13, 19] and also [8, 15] treating the product matrices of form (1.12) with X_n^l-independent diagonal matrices. On the other hand, the diagonal matrices D_n^l in (1.11) depend explicitly on (X_n^l, b_n^l) of (1.3) – (1.4) and, implicitly, via x^{l-1}, on the all "preceding" $(X_n^{l'}, b_n^{l'})$, $l' = l-1, ..., 1$. Thus, the proof of validity of (1.23) requires an additional reasoning. The goal of this paper is to provide this reasoning, thereby justifying the basic formula (1.23) and the corresponding formulas for the mean limiting NCM $\overline{\nu}_{M^L}$ of (1.18), see formula (13) of [21] and formula (2.7) below. Moreover, we prove that the formula (1.16) is valid not only in the mean (see (1.18)), but also with probability 1 in Ω_L of (1.9) (recall that the measures in the r.h.s. of (1.16) are random) and that the limiting measure ν_{M^L} in the l.h.s. of (1.16) coincides with $\overline{\nu}_{M^L}$ of (1.18), i.e., ν_{M^L} is nonrandom.

Note that a possible version of the proof of the above assertions could be carried out by extending the corresponding proofs in free probability (see, e.g. [13, 19]) to the case where the diagonal matrices are given by (1.11). We will prefer, however, another approach based on the standard techniques of random matrix theory, see e.g. [17]. There the main technical tools are some differentiation formulas (see, e.g. (3.50)), providing certain identities for expectations of essential spectral characteristics, and bounds (Poincaré, martingale) for the variance of these characterizes, guaranteing the vanishing of their fluctuations in the large size (layer width) limit, thereby allowing for the conversion of the obtained identities into functional equations for the characteristics in question, the Stieltjes transform of the limiting NCM in particular. This, however, has to be complemented (in fact, preceded) by a certain assertion (see Lemma 3.3) justifying the asymptotic replacement of random $X^l_{n_l}$-dependent matrices $D^l_{n_l}$ in (1.10) – (1.11) by certain random but $X^l_{n_l}$-independent matrices (see (3.19) – (3.20)) and allowing us not only to substantiate the results of [21], but also to extend them to the case of i.i.d. but not necessarily Gaussian $(X^l_{n_l}, b^l_{n_l})$, $l = 1, ..., L$ [18].

The paper is organized as follows. In the next section we prove the validity of (1.16) with probability 1 in Ω_L of (1.9), formula (1.23) and the corresponding formula for $\nu_{M^L} = \overline{\nu}_{M^L}$ of [21]. The proofs are based on a natural inductive procedure allowing for the passage from the lth to the $(l+1)$th layer. In turn, the induction procedure is based on a formula relating the limiting (in the layer width) Stieltjes transforms of the NCM's of two subsequent layers. The formula is more or less standard both in its form and its derivation in the case where the matrices D^l_n in (1.10) are deterministic or random but independent of $(X^{l'}_n, b^{l'}_n)$, $l' = l, l-1, ..., 1$, see e.g. [6, 17] and references therein.The case of dependent D^l_n as in (1.11) is treated in Section 3.

It follows from the results of the section that the coincidence of the limiting eigenvalue distribution of matrices of two indicated cases is due to the form of dependence of D^l_n on $(X^{l'}_n, b^{l'}_n)$, $l' = l, l-1, ..., 1$ given by (1.11), which is, so to say, "slow varying" and does not contribute to the leading term (the limit (1.16)) of the corresponding eigenvalue distribution.

2. Main Result and its Proof.

As was already mention ed in Introduction, our goal is to present a more complete proof of the results of work [21] by using random matrix theory. Thus, to formulate our results, we need several facts of the theory.

Consider for every positive integer n: (i) the $n \times n$ random matrix X_n with entries which are independent standard (mean zero and variance 1) Gaussian random variables; (ii) positive definite (and independent of X_n) matrices K_n and R_n that may be also random but independent of X_n and such that their Normalized Counting Measures ν_{K_n} and ν_{R_n} (see (1.15)) converge weakly (with probability 1 if random) as $n \to \infty$ to non-random measures ν_{K} and ν_{R}. Set $\mathsf{M}_n = n^{-1}\mathsf{R}_n^{1/2} X^T \mathsf{K}_n X_n \mathsf{R}_n^{1/2}$. According to random matrix theory (see, e.g. Lemma 3.5 below, [6] and references therein), in this case the Normalized Counting Measures $\nu_{\mathbf{M}_n}$ of M_n converge weakly with probability 1 as $n \to \infty$ to a non-random measure $\nu_{\mathbf{M}}$ which is uniquely determined by the limiting measures ν_{K} and ν_{R} via a certain analytical procedure (see, e.g. formulas (1.19) and (3.10) – (3.12) below). We can write down this fact as

$$\nu_{\mathsf{M}} = \nu_{\mathsf{K}} \diamond \nu_{\mathsf{R}} \tag{2.1}$$

and say that the procedure defines a binary operation in the set of non-negative measures with the total mass 1 and a support belonging to the positive semiaxis (see more details in Lemma 3.5 and Corollary 3.7). The main result of [21] and of this paper is that the limiting Normalized Counting Measure (1.16) of random matrices (1.12), where K_n is given by (1.11) and (1.22) and depends on Gaussian matrices X^l's of (1.6), can be found as the "product" with respect the operation (2.1) of L measures ν_{K^l}, $l = 1, ..., L$ which are indicated in Theorem 2.1 and are the limiting Normalized Counting Measures of special random matrices that do not depend on X^l's of (1.6), see (3.20 and the subsequent text.

Note that the operation is just a version of the so-called multiplicative convolution of free probability theory [13, 19], having the above random matrices as a basic analytic model.

We will follow [21] and confine ourselves to the case (1.17) where all the weight matrices and bias vectors are of the same size n. The general case of different sizes is essentially the same (see, e.g. Remark 3.2 (iii)).

Theorem 2.1. *Let M_n^L be the random matrix (1.12) defined by (1.2) – (1.11) and (1.17), where the biases b^l and weights W^l are random Gaussian variables satisfying (1.5) – (1.6) and the input vector x^0 (1.1) (deterministic or random) is such that there exists a finite limit*

$$q^1 := \lim_{n\to\infty} q_n^1 > \sigma_b^2 > 0, \ \ q_n^1 = n^{-1} \sum_{j_0=1}^{n} (x_{j_0}^0)^2 + \sigma_b^2. \tag{2.2}$$

Assume also that the nonlinearity φ in (1.2) is a piecewise differentiable function such that φ' is not zero identically and denote

$$\sup_{t\in\mathbb{R}} |\varphi(t)| = \Phi_0 < \infty, \ \sup_{t\in\mathbb{R}} |\varphi'(t)| = \Phi_1 < \infty. \tag{2.3}$$

Then the Normalized Counting Measure (NCM) $\nu_{M_n^L}$ of M_n^L (see (1.15)) converges weakly with probability 1 in the probability space Ω_L of (1.9) to the non-random measure

$$\nu_{M^L} = \nu_{K^1} \diamond \nu_{K^2} ... \diamond \nu_{K^L} \diamond \delta_1, \tag{2.4}$$

where the operation "$\diamond$" is defined in (2.1) (see also Lemma 3.5 and Corollary 3.7), δ_1 is the unit measure concentrated at 1 and ν_{K^l}, $l = 1, ..., L$ is the probability distribution of the random variable $(\varphi'(\gamma\sqrt{q^l}))^2$ with the standard Gaussian random variable γ and q^l determined by the recurrence

$$q^l = \mathbf{E}\{\varphi^2(\gamma\sqrt{q^{l-1}})\}, \ l \geq 2, \tag{2.5}$$

with q^1 given by (2.2).

Remark 2.2. (i) If

$$q_L = \cdots = q_1, \tag{2.6}$$

then $\nu_K := \nu_{K^l}$, $l = 1, ..., L$, (2.4) becomes

$$\nu_{M^L} = \underbrace{\nu_K \diamond \nu_K \cdots \diamond \nu_K}_{L \text{ times}} \diamond \delta_1. \tag{2.7}$$

An important case of equalities (2.6) is where $q^1 = q^*$ and q^* is a fixed point of (2.5), see [12, 22, 26] for a detailed analysis of (2.5) and its role in the deep neural networks functioning.

(ii) Let us show now that Theorem 2.1 implies the results of [21]. It follows from the theorem, (2.21), and Corollary 3.7 that the functional inverse $z_{M^{l+1}}$ of the moment generating function $m_{M^{l+1}}$ (see (1.20) – (1.21)) of the limiting NCM $\nu_{M^{l+1}}$ of matrix M_n^{l+1} and that of M_n^l are related as in (3.87), i.e.,

$$z_{M^{l+1}}(m) = z_{K^{l+1}}(m) z_{M^l}(m) m^{-1}. \tag{2.8}$$

Passing from the moment generating functions to the S-transforms of free probability theory via the formula $S(m) = (1+m)m^{-1}z(m)$ and taking into account that the S-transform of the limiting NCM of the Wishart matrix $n^{-1}X_nX_n^T$ is $S_{MP} = (1+m)^{-1}$ (see [13]), we obtain from (2.8)

$$S_{M^{l+1}}(m) = S_{K^{l+1}}(m) S_{MP}(m) S_{M^l}(m). \tag{2.9}$$

Iterating this relation from $l = 1$ to $l = L-1$, we obtain formula (13) of [21]. The functional equation (1.23) arising in the case (2.6) of the l-independent parameters q_l of (2.5) is derived from (2.9) in [21].

(iii) In the subsequent work [18] we consider a more general case of not necessarily Gaussian random variables, i.e., where the entries of independent random matrices X^l, $l = 1, 2, ...$ in (1.10) – (1.11) are i.i.d. random variables satisfying (1.6) and certain moment conditions and the component of independent vectors b^l, $l = 1, 2, ...$ are i.i.d. random variables satisfying (1.5). It is shown that in this, more general case, the conclusion of the theorem is still valid, however the measure ν_{K^l}, $l = 1, 2, ...$ is now the probability distribution of $(\varphi'(\gamma\sqrt{(q^{l-1} - \sigma_b^2)} + b_1^l))^2$, where γ is again the standard Gaussian random variable and (2.5) is replaced by

$$q^l = \int \varphi^2\Big(\gamma\sqrt{q^{l-1} - \sigma_b^2} + b\Big)\Gamma(d\gamma)F(db), \ l \geq 2, \tag{2.10}$$

where $\Gamma(d\gamma) = (2\pi)^{1/2}e^{-\gamma^2/2}d\gamma$, F is the probability law of b_1^l and q^1 is again given by (2.2).

(iv) If the input vector (1.1) are random, then it is assumed that they are defined on the same probability space Ω_{x^0} for all n_0 and the limit q^1 exists with probability 1 in Ω_{x^0}. An example of this situation is where $\{x_{j_0}^l\}_{j^0=1}^{n_0}$ are the first n_0 components of an ergodic sequence $\{x_{j_0}^l\}_{j^0=1}^{\infty}$ (e.g. a sequence of i.i.d. random variables) with finite second moment. Here q_1 in (2.2) exists with probability 1 on Ω_{x^0} and even is non-random just by ergodic theorem (the strong Law of Large Numbers in the case of i.i.d sequence) and the theorem is valid with probability 1 in $\Omega_l \times \Omega_{x^0}$.

We present now the proof of Theorem 2.1.

Proof. We prove the theorem by induction in L. We have from (1.2) – (1.12) and (1.17) with $L = 1$ the following $n \times n$ matrix

$$M_n^1 = J_n^1 (J_n^1)^T = n^{-1} D_n^1 X_n^1 (X_n^1)^T D_n^1. \tag{2.11}$$

It is convenient to pass from M_n^1 to the $n \times n$ matrix

$$\mathcal{M}_n^1 = (J_n^1)^T J_n^1 = n^{-1} (X_n^1)^T K_n^1 X_n^1, \ K_n^1 = (D_n^1)^2 \tag{2.12}$$

which has the same spectrum, hence the same Normalized Counting Measure as M_n^1. The matrix $\mathcal{M}_n^1$ is a particular case with $S_n = \mathbf{1}_n$ of matrix (3.1) treated in Theorem 3.1 below. Since the NCM of the unit matrix 1_n is the Dirac measure δ_1, conditions (3.2) – (3.3) of the theorem are evident. Condition (3.9) of the theorem is just (2.2). It follows then from Corollary 3.7 that the assertion of our theorem, i.e., formula (2.4) with q^1 of (2.2) is valid for $L = 1$.

Consider now the case $L = 2$ of (1.2) – (1.12) and (1.17):

$$M_n^2 = n^{-1} D_n^2 X_n^2 M_n^1 (X_n^2)^T D_n^2. \tag{2.13}$$

Since M_n^1 is positive definite, we have

$$M_n^1 = (S_n^1)^2 \tag{2.14}$$

with a positive definite S_n^1, hence

$$M_n^2 = n^{-1} D_n^2 X_n^2 (S_n^1)^2 (X_n^2)^T D_n^2 \tag{2.15}$$

and the corresponding $\mathcal{M}_n^2$ is

$$\mathcal{M}_n^2 = n^{-1} S_n^1 (X_n^2)^T K_n^2 X_n^2 S_n^1, \ K_n^2 = (D_n^2)^2. \tag{2.16}$$

We observe that $\mathcal{M}_n^2$ is a particular case of matrix (3.1) of Theorem 3.1 with $M_n^1 = (S_n^1)^2$ as $R_n = (S_n)^2$, X_n^2 as X_n, K_n^2 as K_n, $\{x_{j_1}^1\}_{j_1=1}^n$ as $\{x_{\alpha n}\}_{\alpha=1}^n$, $\Omega_1 = \Omega^1$ of (1.9) as Ω_{Rx} and Ω^2 of (1.9) as Ω_{Xb}, i.e., the case of the random but $\{X_n^2, b_n^2\}$ -independent R_n and $\{x_{\alpha n}\}_{\alpha=1}^n$ in (3.1) as described in Remark 3.2 (i). Let us check that conditions (3.2) – (3.3) and (3.9) of Theorem 3.1 are satisfied for $\mathcal{M}_n^2$ of (2.16) with probability 1 in the probability space $\Omega_1 = \Omega^1$ generated by $\{X_n^1, b_n^1\}$ for all n and independent of the space Ω^2 generated by $\{X_n^2, b_n^2\}$ for all n.

We will need here an important fact on the operator norm of $n \times n$ random matrices with independent standard Gaussian entries. Namely, if X_n is such $n \times n$ matrix, then we have with probability 1

$$\lim_{n \to \infty} n^{-1/2} ||X_n|| = 2, \tag{2.17}$$

thus, with the same probability

$$||X_n|| \leq C n^{1/2}, \ C > 2 \tag{2.18}$$

if n is large enough.

For the Gaussian matrices relation (2.17) has already been known in the Wigner's school of the early 1960th, see [17]. It follows in this case from the orthogonal polynomial representation of the density of the NCM of $n^{-1} X_n X_n^T$ and the asymptotic formula for the corresponding orthogonal polynomials. For the modern form of (2.17) and (2.18), in particular their validity for i.i.d matrix entries with mean zero and finite fourth moment, see [1, 28] and references therein.

We will also need the bound

$$||K_n^1|| \le (\Phi_1)^2, \tag{2.19}$$

following from (1.11), (1.22) and (2.3) and valid everywhere in Ω_1 of (1.9).

Now, by using (2.12), (2.18), (2.19) and the inequality

$$|\mathrm{Tr}AB| \le ||A||\mathrm{Tr}B, \tag{2.20}$$

valid for any matrix A and a positive definite matrix B, we obtain with probability 1 in Ω_1 and for sufficiently large n

$$n^{-1}\mathrm{Tr}(M_n^1)^2 = n^{-3}\mathrm{Tr}(K_n^1 X_n^1 (X_n^1)^T)^2 \le (C\Phi_1)^4.$$

We conclude that M_n^1, which plays here the role of R_n of Theorem 3.1 and Remark 3.2 (i) according to (2.14), satisfies condition (3.2) with $r_2 = (C\Phi_1)^4$ and with probability 1 in our case, i.e., on a certain $\Omega_{11} \subset \Omega_1$, $\mathbf{P}(\Omega_{11}) = 1$.

Next, it follows from the above proof of the theorem for $L = 1$, i.e., in fact, from Theorem 3.1, that there exists $\Omega_{12} \subset \Omega_1$, $\mathbf{P}(\Omega_{12}) = 1$ on which the NCM $\nu_{M_n^1}$ converges weakly to a non-random limit ν_{M^1}, hence condition (3.3) is also satisfied with probability 1, i.e., on Ω_{12}.

At last, according to Lemma 3.11 (i) and (2.2), there exists $\Omega_{13} \subset \Omega_1$, $\mathbf{P}(\Omega_{13}) = 1$ on which there exists

$$\lim_{n\to\infty} n^{-1} \sum_{j_1=1}^{n} (x_{j_1}^1)^2 + \sigma_b^2 = q^2 > \sigma_b^2,$$

i.e., condition (3.9) is also satisfied.

Hence, we can apply Theorem 3.1 on the subspace $\overline{\Omega}_1 = \Omega_{11} \cap \Omega_{12} \cap \Omega_{13} \subset \Omega_1$, $\mathbf{P}(\overline{\Omega}_1) = 1$ where all the conditions of the theorem are valid, i.e., $\overline{\Omega}_1$ plays the role of Ω_{Rx} of Remark 3.2 (i). Thus the theorem implies that for any $\omega_1 \in \overline{\Omega_1}$ there exists subspace $\overline{\Omega^2}(\omega_1)$ of the space Ω^2 generated by $\{X_n^2, B_n^2\}$ for all n and such that $\mathbf{P}(\overline{\Omega^2}(\omega_1)) = 1$ and formulas (2.4) – (2.5) are valid for $L = 2$. It follows then from the Fubini theorem that the same is true on a certain $\overline{\Omega}_2 \subset \Omega_2$, $\mathbf{P}(\overline{\Omega}_2) = 1$ where Ω_2 is defined by (1.9) with $L = 2$.

This proves the theorem for $L = 2$. The proof for $L = 3, 4, ...$ is analogous, since (cf. (2.15))

$$M_n^{l+1} = n^{-1} D_n^{l+1} X_n^{l+1} M_n^l (X_n^{l+1})^T D_n^{l+1},\ l \ge 2. \tag{2.21}$$

In particular, we have with probability 1 on Ω_{l-1} of (1.9) for M_n^{l-1} playing the role of R_n of Theorem 3.1 on the lth step of the inductive procedure (cf. (3.2))

$$n^{-1}\mathrm{Tr}(M^l)^2 \le (C\Phi_1)^{4l},\ l \ge 2.$$

□

3. Auxiliary Results.

Our main result, Theorems 2.1 on the limiting eigenvalue distribution of random matrices (1.12) for any L, is proved above by induction in the layer number l, see formulas (2.13), (2.16) and (2.21). To carry out the passage from the lth to the $(l+1)$th layer we need an expression for the limiting NCM $\nu_{\mathcal{M}^{l+1}}$ of the matrix $\mathcal{M}_n^{l+1}$ via that of $\mathcal{M}_n^l$ in the infinite width limit $n \to \infty$. The corresponding results, which could be of independent interest, as well as certain auxiliary results are proved in this section. In particular, a functional equation relating the Stieltjes transform of $\nu_{\mathcal{M}_n^{l+1}}$ and $\nu_{\mathcal{M}_n^l}$ in the limit $n \to \infty$ is obtained.

Theorem 3.1. *Consider for every positive integer n the $n \times n$ random matrix*

$$\mathcal{M}_n = n^{-1} S_n X_n^T K_n X_n S_n, \tag{3.1}$$

where:

(a) *S_n is a positive definite $n \times n$ matrix such that*

$$\sup_n n^{-1} \mathrm{Tr} R_n^2 = r_2 < \infty, \; R_n = S_n^2, \tag{3.2}$$

and

$$\lim_{n\to\infty} \nu_{R_n} = \nu_R, \; \nu_R(\mathbb{R}_+) = 1, \tag{3.3}$$

where ν_{R_n} is the Normalized Counting Measure of R_n, ν_R is a non-negative measure not concentrated at zero and $\lim_{n\to\infty}$ denotes here the weak convergence of probability measures;

(b) *X_n is the $n \times n$ random matrix*

$$X_n = \{X_{j\alpha}\}_{j,\alpha=1}^n, \; \mathbf{E}\{X_{j\alpha}\} = 0, \; \mathbf{E}\{X_{j_1\alpha_1} X_{j_2\alpha_2}\} = \delta_{j_1 j_2} \delta_{\alpha_1 \alpha_2}, \tag{3.4}$$

with the independent standard Gaussian entries (cf. (1.6)), b_n is the n-component random vector

$$b_n = \{b_j\}_{j=1}^n, \; \mathbf{E}\{b_j\} = 0, \; \mathbf{E}\{b_{j_1} b_{j_2}\} = \sigma_b^2 \delta_{j_1 j_2} \tag{3.5}$$

with the independent Gaussian components of zero mean and variance σ_b^2 (cf. (1.5)) and for all n matrix X_n and the vector b_n viewed as defined on the probability space

$$\Omega_{Xb} = \Omega_X \times \Omega_b, \tag{3.6}$$

where Ω_X and Ω_b are generated by (1.7) and (1.8);

(c) *K_n and D_n are the diagonal random matrices*

$$K_n = D_n^2, \; D_n = \{\delta_{jk} D_{jn}\}_{j,k=1}^n, \; D_{jn} = \varphi'\Big(n^{-1/2} \sum_{a=1}^n X_{j\alpha} x_{\alpha n} + b_j\Big), \tag{3.7}$$

where $\varphi : \mathbb{R} \to \mathbb{R}$ is a piecewise differentiable function, such that (cf. (2.3))

$$\sup_{x\in\mathbb{R}} |\varphi(x)| = \Phi_0 < \infty, \; \sup_{x\in\mathbb{R}} |\varphi'(x)| = \Phi_1 < \infty, \tag{3.8}$$

and $x_n = \{x_{\alpha n}\}_{\alpha=1}^n$ is a collection of real numbers such that there exists

$$q = \lim_{n\to\infty} q_n > \sigma_b^2 > 0, \; q_n = n^{-1} \sum_{\alpha=1}^n (x_{\alpha n})^2 + \sigma_b^2. \tag{3.9}$$

Then the Normalized Counting Measure (NCM) $\nu_{\mathcal{M}_n}$ of $\mathcal{M}_n$ converges weakly with probability 1 in Ω_{Xb} of (3.6) to a non-random measure $\nu_{\mathcal{M}}$ whose Stieltjes transform $f_{\mathcal{M}}$ (see (1.19)) can be obtained from the formulas

$$f_{\mathcal{M}}(z) = \int_0^\infty \frac{\nu_R(d\lambda)}{k(z)\lambda - z} = -z^{-1} + z^{-1} h(z) k(z), \tag{3.10}$$

where the pair (h,k) is a unique solution of the system of functional equations

$$h(z) = \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{k(z)\lambda - z} \tag{3.11}$$

$$k(z) = \int_0^\infty \frac{\lambda \nu_K(d\lambda)}{h(z)\lambda + 1}, \tag{3.12}$$

in which ν_R is defined in (3.3), ν_K is the probability distribution of $(\varphi'(\sqrt{q}\gamma))^2$ with q of (3.9) and the standard Gaussian random variable γ, i.e.,

$$\nu_K(\Delta) = \mathbf{P}\{(\varphi'(\sqrt{q}\gamma))^2 \in \Delta\},\ \Delta \in \mathbb{R}, \tag{3.13}$$

and we are looking for a solution of (3.11) – (3.12) in the class of pairs (h,k) of functions such that h is analytic outside the positive semi-axis, continuous and positive on the negative semi-axis and

$$\Im h(z)\Im z > 0,\ \Im z \neq 0;\ \sup_{\xi \geq 1} \xi h(-\xi) \in (0,\infty). \tag{3.14}$$

Remark 3.2. (i) To apply Theorem 3.1 to the proof of Theorem 2.1 we need a version of Theorem 3.1 in which its "parameters", i.e., R_n, hence S_n, in (3.1) – (3.3) and (possibly) $\{x_{\alpha n}\}_{\alpha=1}^n$ in (3.7) and (3.9) are random, defined for all n on the same probability space Ω_{Rx}, independent of Ω_{Xb} of (3.6) and satisfy conditions (3.2) – (3.3) and (3.9) with probability 1 on Ω_{Rx}, i.e., on a certain subspace $\overline{\Omega}_{Rx} \subset \Omega_{Rx}$, $\mathbf{P}(\overline{\Omega_{Rx}}) = 1$. In this case Theorem 3.1 is valid with probability 1 in $\Omega_{Xb} \times \Omega_{Rx}$. The corresponding argument is standard in random matrix theory (see, e.g. Section 2.3 of [17]) and similar to that presented in Remark 3.6 (i). In deed, let $\overline{\Omega}_{Xb}(\omega_{Rx}) \subset \Omega_{Xb}$, $\mathbf{P}(\overline{\Omega}_{Xb}(\omega_{Rx})) = 1$ be the subspace of Ω_{Xb} of (3.6) on which the theorem holds for a given realization $\omega_{Rx} \in \overline{\Omega}_{Rx}$ of the parameters. Then it follows from the Fubini theorem that Theorem 3.1 holds on a certain $\overline{\Omega} \subset \Omega_{Rx} \times \Omega_{Xb}$, $\mathbf{P}(\overline{\Omega}) = 1$. We will use this remark in the proof of Theorem 2.1. The obtained limiting NCM $\nu_{\mathcal{M}}$ is random in general due to the randomness of ν_R and q in (3.3) and (3.9) which are defined on the probability space Ω_{Rx} but do not depend on $\omega \in \overline{\Omega}_{Xb}$. We will use this remark in the proof of Theorem 2.1. Note, however, that in this case application the corresponding analogs of ν_R and q are not random, thus the limiting measure ν_{M^L} is a "genuine" non-random measure.

(ii) Repeating almost literally the proof of the theorem, one can treat a more general case where S_m is $m \times m$ positive definite matrix satisfying (3.2) – (3.3), K_n is the $n \times n$ diagonal matrix given by (3.7) – (3.9), X_n is a $n \times m$ Gaussian random matrix satisfying (1.6) and (cf. (1.13)) $\lim_{m\to\infty, n\to\infty} m/n = c \in (0,\infty)$. The corresponding modifications of the theorem are given in Remark 3.6 (ii).

(iii) The theorem is also valid for not necessarily Gaussian X_n and b_n (see [18] and Remark 2.2) (iii).

We will prove now Theorem 3.1

Proof. Lemma 3.9 (i) implies that the fluctuations of $\nu_{\mathcal{M}_n}$ vanish sufficiently fast as $n \to \infty$. This and the Borel-Cantelli lemma reduce the proof of the theorem to the proof of the weak convergence of the expectation

$$\overline{\nu}_{\mathcal{M}_n} := \mathbf{E}\{\nu_{\mathcal{M}_n}\} \tag{3.15}$$

of $\nu_{\mathcal{M}_n}$ to the limit $\nu_{\mathcal{M}}$ whose Stieltjes transform solves (3.10) – (3.14). It suffices to prove the tightness of the sequence $\{\overline{\nu}_{\mathcal{M}_n}\}_n$ of measures and the pointwise convergence on an open set of $\mathbb{C}\setminus\mathbb{R}_+$ of their Stieltjes transforms (cf. (1.19))

$$f_{\mathcal{M}_n}(z) := \int_0^\infty \frac{\overline{\nu}_{\mathcal{M}_n}(d\lambda)}{\lambda - z} \tag{3.16}$$

to the limit satisfying (3.10) – (3.14).

The tightness is guaranteed by the uniform in n boundedness of

$$\mu_n^{(1)} = \int_0^\infty \lambda \overline{\nu}_{\mathcal{M}_n}(d\lambda) \tag{3.17}$$

providing the uniform in n bounds for the tails of $\overline{\nu}_{\mathcal{M}_n}$.

According to the definition of the NCM (see, e.g. (1.15)), spectral theorem and (3.1) we have $\mu_n^{(1)} = \mathbf{E}\{n^{-1}\mathrm{Tr}\mathcal{M}_n\} = \mathbf{E}\{n^{-2}\mathrm{Tr}X_nR_nX_n^TK_n\}$ and then (2.20), (3.2) – (3.4) and (3.7) – (3.8) yield

$$\mu_n^{(1)} \le n^{-2}\Phi_1^2\mathbf{E}\{\mathrm{Tr}X_nR_nX_n^T\} = \Phi_1^2 n^{-1}\mathrm{Tr}R_n \le r_2^{1/2}\Phi_1^2. \tag{3.18}$$

This implies the tightness of $\{\overline{\nu}_{\mathcal{M}_n}\}_n$ and reduces the proof of the theorem to the proof of pointwise in $\mathbb{C}\setminus\mathbb{R}_+$ convergence of (3.16) to the limit determined by (3.10) – (3.12).

The above argument, reducing the analysis of the large size behavior of the eigenvalue distribution of random matrices to that of the expectation of the Stieltjes transform of the distribution, is widely used in random matrix theory (see [17], Chapters 3, 7, 18 and 19), in particular, while dealing with the sample covariance matrices. However, the matrix $\mathcal{M}_n$ of (3.1) differs essentially from the sample covariance matrices, since the "central" matrix K_n of (3.7) is random and dependent on X_n (data matrix according to statistics), while in the sample covariance matrix the analog of K_n is either deterministic or random but independent of X_n.

This is why the next, in fact, the main step of the proof of Theorem 3.1 is to show that in the limit $n \to \infty$ the Stieltjes transform (3.16) of (3.1) coincides with the Stieltjes transform f_{M_n} of the mean NCM $\overline{\nu}_{\mathsf{M}_n}$of the matrix

$$\mathsf{M}_n = S_nX_n^T\mathsf{K}_nX_nS_n, \tag{3.19}$$

where

$$\mathsf{K}_n = \{\delta_{jk}\mathsf{K}_{jn}\}_{j,k=1}^n,\ \mathsf{K}_{jn} = (\varphi'(q_n^{1/2}\gamma_j))^2, \tag{3.20}$$

φ is again defined in (3.7) – (3.8), $\{\gamma_j\}_{j=1}^n$ are independent standard Gaussian random variables and q_n is defined in (3.9).

This, crucial for the paper fact, is proved in Lemma 3.3 below provided that φ in (3.7) and (3.20) and S_n, hence R_n in (3.1) and (3.19) satisfy the conditions

$$\max_{x\in\mathbb{R}} |\varphi^{(p)}(x)| = \widetilde{\Phi}_p < \infty,\ p = 0, 1, 2, \tag{3.21}$$

and

$$\sup_n ||R_n|| = \rho < \infty. \tag{3.22}$$

Thus, since K_n, being random, is X_n-independent, the $n \to \infty$ limit of Stieltjes transform f_{M_n} of the mean NCM $\overline{\nu}_{\mathsf{M}_n}$ of (3.19) can be obtained by using one of the techniques of random matrix theory including those of free probability theory [5,13] or based on the Stieltjes transform, see [6,17] and references therein. We will present below the corresponding assertion as Lemma 3.5 and outline its proof based on the Stieltjes transform techniques.

Hence, Lemmas 3.3 and 3.5 imply that the limiting Stieltjes transform $f_{\mathcal{M}}$ of (3.10) can be expressed via a unique solution of the system (3.10) – (3.14), provided that φ and R_n in (3.1) satisfy the conditions (3.21) – (3.22), i.e., the assertion of Theorem 3.1 is proved under these conditions. Let us show that these technical conditions can be replaced by initial conditions (3.2) and (3.8) of the theorem.

We will begin with (3.8). For any φ having a piecewise continuous derivative and satisfying (3.8) introduce

$$\begin{aligned}\varphi_\varepsilon(x) &= (2\pi)^{-1/2}\int e^{-y^2/2}\varphi(x+\varepsilon y)dy \\ &= (2\pi\varepsilon^2)^{-1/2}\int e^{-(x-y)^2/2\varepsilon^2}\varphi(y)dy,\ \varepsilon>0.\end{aligned} \tag{3.23}$$

Then φ_ε and φ'_ε converge to φ and φ' as $\varepsilon \to 0$ uniformly on a compact set of $\mathbb{R}$ (except the discontinuity points of φ') and

$$\sup_{x\in\mathbb{R}}|\varphi_\varepsilon^{(p)}(x)| \le \Phi_p,\ p=0,1,\ \sup_{x\in\mathbb{R}}|\varphi_\varepsilon^{''}(x)| \le \Phi_1/\varepsilon. \tag{3.24}$$

Hence, φ_ε satisfies (3.21) with $\widetilde{\Phi}_p = \Phi_p,\ p = 0,1$ and $\widetilde{\Phi}_2 = \Phi_1/\varepsilon < \infty$ and the assertion of theorem is valid for φ_ε according to the above argument.

Let $\nu_{\mathcal{M}}$ be the measure whose Stieljes transform satisfies (3.10) – (3.12) with ν_R such that supp $\nu_R \subset [0,\rho],\ \rho<\infty$ (cf. (3.22)), φ of (3.13) be satisfying (3.8), $\nu_{\mathcal{M}^\varepsilon}$ be the analogous measure with φ_ε instead of φ in (3.13), $\overline{\nu}_{\mathcal{M}_n}$ be the mean NCM of (3.1) and $\overline{\nu}_{\mathcal{M}_n^\varepsilon}$ be the mean NCM of the matrix (3.1) with φ_ε instead of φ in (3.7), i.e., with

$$K_n^\varepsilon = \{\delta_{jk}K_{jn}^\varepsilon\}_{j,k=1}^n,\ K_{jn}^\varepsilon = \left(\varphi'_\varepsilon\left(n^{-1/2}\sum_{\alpha=1}^n X_{j\alpha}x_{\alpha n}+b_j\right)\right)^2, \tag{3.25}$$

instead of K_{jn} of (3.7). We write then for any n-independent $z \in \mathbb{C}\setminus\mathbb{R}_+$

$$\begin{aligned}|f_{\mathcal{M}}(z)-f_{\mathcal{M}_n}(z)| &\le |f_{\mathcal{M}}(z)-f_{\mathcal{M}^\varepsilon}(z)| \\ &+|f_{\mathcal{M}^\varepsilon}(z)-f_{\mathcal{M}_n^\varepsilon}(z)|+|f_{\mathcal{M}_n^\varepsilon}(z)-f_{\mathcal{M}_n}(z)|.\end{aligned} \tag{3.26}$$

According to Lemma 3.12 (ii), the measure whose Stieltjes transform solves (3.10) – (3.12) is weakly continuous in ν_K. Besides, it follows from (3.13) that ν_K is weakly continuous in φ' with respect to the bounded point-wise convergence of φ'. Hence, the first term on the right of (3.34) vanishes as $\varepsilon \to 0$. Next, the theorem proved above under conditions (3.21) – (3.22) implies that the second term on the right vanishes as $n \to \infty$ for any n-independent $\varepsilon > 0$. We conclude that the l.h.s. of (3.26) vanishes as $n \to \infty$ if the third term on the right of (3.26) vanishes as $\varepsilon \to 0$ uniformly in n:

$$f_{\mathcal{M}_n^\varepsilon}(z)-f_{\mathcal{M}_n}(z)\to 0,\ \varepsilon\to 0,\ \zeta=\mathrm{dist}(z,\mathbb{R}_+)\ge\zeta_0>0. \tag{3.27}$$

Denoting $\mathcal{G} = (\mathcal{M}_n - z)^{-1}$, $\mathcal{G}_\varepsilon = (\mathcal{M}_n^\varepsilon - z)^{-1}$ and using the resolvent identity $\mathcal{G}_\varepsilon - \mathcal{G} = \mathcal{G}(\mathcal{M}_n - \mathcal{M}_n^\varepsilon)\mathcal{G}_\varepsilon$ and the relations $f_{\mathcal{M}_n}(z) = \mathbf{E}\{n^{-1}\mathrm{Tr}\mathcal{G}\}$ and $f_{\mathcal{M}_n^\varepsilon}(z) = \mathbf{E}\{n^{-1}\mathrm{Tr}\mathcal{G}_\varepsilon\}$, we get

$$f_{\mathcal{M}_n^\varepsilon}(z) - f_{\mathcal{M}_n}(z) = n^{-1}\mathbf{E}\{\mathrm{Tr}\mathcal{G}_\varepsilon\mathcal{G}(\mathcal{M}_n - \mathcal{M}_n^\varepsilon)\}$$

$$= n^{-2}\sum_{j=1}^n \mathbf{E}\{(XS\mathcal{G}_\varepsilon\mathcal{G}SX^T)_{jj}(K_{jn} - K_{jn}^\varepsilon)\}. \tag{3.28}$$

Now, (3.22), Schwarz inequality for expectations and the bounds

$$|K_j| \le \Phi_1^2,\ ||\mathcal{G}|| \le \zeta^{-1},\ ||\mathcal{G}_\varepsilon|| \le \zeta^{-1},\ \zeta = \mathrm{dist}\{z, \mathbb{R}_+\} \ge \zeta_0 > 0, \tag{3.29}$$

where we used the bound

$$||(A - z)^{-1}|| \le \zeta^{-1} \tag{3.30}$$

valid for any positive definite A, yield for the r.h.s. of (3.28)

$$\rho(\zeta n)^{-2}\sum_{j=1}^n \mathbf{E}\{||X^{(j)}||^2|K_{jn} - K_{jn}^\varepsilon)|\}$$

$$\le \rho(\zeta n)^{-2}\sum_{j=1}^n \mathbf{E}^{1/2}\{||X^{(j)}||^4\}\mathbf{E}^{1/2}\{|K_{jn} - K_{jn}^\varepsilon|^2\},$$

where $X^{(j)} = \{X_{j\alpha}\}_{\alpha=1}^n$, $j = 1, ..., n$ are the columns of the $n \times n$ matrix X. Taking into account that

$$||X^{(j)}||^2 = \sum_{\alpha=1}^n X_{j\alpha}^2 \tag{3.31}$$

and that $\{X_{j\alpha}\}_{\alpha=1}^n$ are independent standard Gaussian (see (1.6)), we obtain

$$\mathbf{E}\{||X^{(j)}||^2\} = n,\ \mathbf{E}\{||X^{(j)}||^4\} = n(n+2) \le Cn^2,\ C \ge 3. \tag{3.32}$$

Since, in addition, $\{(K_{jn} - K_{jn}^\varepsilon)\}_{j=1}^n$ are i.i.d. random variables, we have in view of (3.7), (3.23) and (3.32):

$$|f_{\mathcal{M}_n^\varepsilon}(z) - f_{\mathcal{M}_n}(z)| \le C^{1/2}\rho\zeta^{-2}\mathbf{E}^{1/2}\{|K_{1n} - K_{1n}^\varepsilon|^2\}$$
$$\le C^{1/2}\rho\zeta^{-2}((2\pi)^{-1/2}\int e^{-y^2/2}|\varphi'(x) - \varphi'(x + \varepsilon y)|^2\Gamma_n(dx)dy)^{1/2},$$

where Γ_n is the probability law of the argument of φ' in (3.7) and (3.25). Since $\{X_{j\alpha}\}_{j,\alpha=1}^n$ and $\{b_j\}_{j=1}^n$ are independent standard Gaussian, $\Gamma_n(dx) = g_n(x)dx$, where g_n is the density of the Gaussian distribution of zero mean and variance q_n of (3.9), the r.h.s. of the above expression tends to zero as $\varepsilon \to 0$ uniformly in $n \to \infty$. This proves (3.27), hence, justifies the replacement of (3.21) by the condition (3.8) of the theorem.

Next, we will replace (3.22) by condition of (3.2) of the theorem. This is, in fact, a known procedure of random matrix theory. In our case it is a version of the procedure given in the first part of proof of Theorem 7.2.2 (or Theorem 19.1) in [17]. Here is an outline of the procedure. Let R_n be a general (i.e., not satisfying in general (3.22)) positive definite matrix such that (3.2) – (3.3) hold with certain r_2 and the limiting measure ν_R. For any positive integer p introduce the truncated matrix $R_n^{(p)}$ having the same eigenvectors as R_n and eigenvalues $R_\alpha^{(p)} = \max\{R_\alpha, p\}$, $\alpha = 1, 2, ..., n$, where $\{R_\alpha\}_{\alpha=1}^n$ are the eigenvalues of R_n. Then $R_n^{(p)}$ satisfies (3.22) with $\rho = p$, its NCM $\nu_{R_n^{(p)}}$ satisfies (3.2) – (3.3) with the weak limit $\nu_{R^{(p)}} := \lim_{n\to\infty} \nu_{R_n^{(p)}}$ coinciding with ν_R inside $[0, p)$, equals zero outside $[0, p]$ and such that

$$\lim_{p\to\infty} \nu_{R^{(p)}} = \nu_R. \tag{3.33}$$

Denote by $\mathcal{M}_n^{(p)}$ the matrix (3.1) with $R_n^{(p)}$ instead of R_n, by $\overline{\nu}_{\mathcal{M}_n^{(p)}}$ its mean NCM and by $\nu_{\mathcal{M}^{(p)}}$ its limit as $n \to \infty$ with a fixed $p > 0$. We will use now an argument analogous to that used above to prove the replacement of (3.21) by (3.8). We write (cf. (3.26))

$$|\nu_{\mathcal{M}} - \overline{\nu}_{\mathcal{M}_n}| \le |\nu_{\mathcal{M}} - \nu_{\mathcal{M}^{(p)}}| + |\nu_{\mathcal{M}^{(p)}} - \overline{\nu}_{\mathcal{M}_n^{(p)}}| + |\overline{\nu}_{\mathcal{M}_n^{(p)}} - \overline{\nu}_{\mathcal{M}_n}|. \tag{3.34}$$

It follows then from Lemma 3.12 (ii) and (3.33) that solution of (3.11) – (3.12), hence (3.10), with $\nu_{R^{(p)}}$ instead ν_R converges pointwise in $\mathbb{C} \setminus \mathbb{R}_+$ as $p \to \infty$ to that of (3.10) – (3.12) with the "genuine" ν_R satisfying (3.2) (see also (3.102)). Thus, the first term on the right vanishes as $p \to \infty$. Next, since the theorem is valid under condition (3.22), hence (3.2) – (3.3), and $R_n^{(p)}$ satisfies (3.22) with $\rho = p$, the second term on the right vanishes as $n \to \infty$ for any n-independent $p > 0$. Thus, it suffices to prove that

$$\overline{\nu}_{\mathcal{M}_n} - \overline{\nu}_{\mathcal{M}_n^{(p)}}$$

tends weakly to zero as $p \to \infty$ uniformly in $n \to \infty$ (cf. (3.27)). The expectations $\overline{\nu}_{\mathcal{M}_n}$ and $\overline{\nu}_{\mathcal{M}_n^{(p)}}$ coincide with those $\overline{\nu}_{M_n}$ and $\overline{\nu}_{M_n^{(p)}}$ of matrices $M_n = D_n X_n R_n X_n^T D_n$ and $M_n^{(p)} = D_n X_n R_n^{(p)} X_n^T D_n$ (cf. (1.12). Writing M_n as the sum of the rank-one matrices (cf. (1.12) and (3.91))

$$M_n = \sum_{\alpha=1}^n Y_\alpha \otimes Y_\alpha,\ Y_\alpha = \{Y_{j\alpha}\}_{j=1}^n,\ Y_{j\alpha} = (D_n X_n S_n)_{j\alpha} \tag{3.35}$$

and using the analogous representation for $M_n^{(p)}$, we conclude that

$$\operatorname{rank}(M_n - M_n^{(p)}) \le \sharp\{R_\alpha : R_\alpha > p,\ \alpha = 1, 2, ..., n\}$$

and then the min-max principle of linear algebra and the definition of a NCM (see, e.g. (1.15)) yield for any interval Δ of spectral axis

$$|\overline{\nu}_{\mathcal{M}_n}(\Delta) - \overline{\nu}_{\mathcal{M}_n^{(p)}}(\Delta)| \le \nu_{R_n}([p, \infty)). \tag{3.36}$$

This estimate and (3.3) imply the weak convergence of the r.h.s. to zero as $p \to \infty$ uniformly in n, hence, the weak convergence of $\overline{\nu}_{\mathcal{M}_n}$ to $\nu_{\mathcal{M}}$ as $n \to \infty$ and the coincidence of the Stieltjes transform of $\nu_{\mathcal{M}}$ with that given by (3.10) – (3.11) under condition (3.2). $\square$

We will prove now an assertion which is used in the proof of the theorem and which is central in this work since it shows the mathematical mechanism of the coincidence of the limiting eigenvalue distribution of "non-linear" random matrix

$\mathcal{M}_n$ of (3.1), where K_n of (3.7) depends nonlinearly on X_n, and a conventional for random matrix theory matrix M_n of (3.19), where the analog K_n of K_n is random but independent of X_n matrix given by (3.20).

Lemma 3.3. *Consider the matrices $\mathcal{M}_n$ and M_n given by (3.1) and (3.19) and such that:*

- *the matrix S_n in $\mathcal{M}_n$ and M_n is diagonal, positive definite and satisfies (3.22);*
- *the random matrix X_n in $\mathcal{M}_n$ and M_n is Gaussian and given by (3.4);*
- *the matrix K_n in $\mathcal{M}_n$ is defined in (3.7) with φ satisfying (3.21);*
- *the matrix K_n in M_n is defined in (3.20) with the same φ satisfying (3.21).*

Denote by $\overline{\nu}_{\mathcal{M}_n}$ and $\overline{\nu}_{\mathsf{M}_n}$ the mean NCM of $\mathcal{M}_n$ and M_n, by $f_{\mathcal{M}_n}$ and f_{M_n} the Stieltjes transforms of $\overline{\nu}_{\mathcal{M}_n}$ and $\overline{\nu}_{\mathsf{M}_n}$ and

$$\Delta_n(z) := f_{\mathcal{M}_n}(z) - f_{\mathsf{M}_n}(z),\ z \in \mathbb{C} \setminus \mathbb{R}_+. \tag{3.37}$$

Then we have for any n-independent z, $\zeta := \operatorname{dist}\{z, \mathbb{R}_+\} > 0$:

$$\lim_{n\to\infty} \Delta_n(z) = 0. \tag{3.38}$$

Proof. Writing

$$f_{\mathcal{M}_n} = \mathbf{E}\{n^{-1}\mathrm{Tr}\mathcal{G}_n(z)\},\ f_{\mathsf{M}_n} = \mathbf{E}\{n^{-1}\mathrm{Tr}\mathsf{G}_n(z)\} \tag{3.39}$$

where

$$\mathcal{G}_n(z) = (\mathcal{M}_n - z)^{-1},\ \mathsf{G}_n(z) = (\mathsf{M}_n - z)^{-1},\ z \in \mathbb{C} \setminus \mathbb{R}_+ \tag{3.40}$$

are the corresponding resolvents, we obtain from (3.37)

$$\Delta_n(z) = \mathbf{E}\{n^{-1}\mathrm{Tr}(\mathcal{G}_n(z) - \mathsf{G}_n(z))\}. \tag{3.41}$$

Note that the symbol $\mathbf{E}\{...\}$ in (3.41) and below denotes the expectation with respect to the "old" collections $\{X_{j\alpha}\}_{j,\alpha=1}^n$ and $\{b_j\}_{j=1}^n$ of (3.4) and (3.5) as well as with respect to the "new" collection $\{\gamma_j\}_{j=1}^n$ of (3.20) of independent standard Gaussian variables.

Set for $j = 1, ..., n$

$$\begin{aligned} \eta_j(t) &= t^{1/2}\eta_j + (1-t)^{1/2} q_n^{1/2}\gamma_j,\ t \in [0,1], \\ \eta_j &= n^{-1/2}\sum_{\alpha=1}^n X_{j\alpha}x_{\alpha n} + b_j, \end{aligned} \tag{3.42}$$

$$K_n(t) = \{\delta_{jk}K_{jn}(t)\}_{j,k=1}^n,\ K_{jn}(t) = (\varphi'(\eta_j(t)))^2 \tag{3.43}$$

and

$$\mathcal{M}_n(t) = S_n X_n^T K_n(t) X_n S_n,\ \mathcal{G}_n(z,t) = (\mathcal{M}_n(t) - z)^{-1}. \tag{3.44}$$

Then $\mathcal{M}_n(1) = \mathcal{M}_n$, $\mathcal{M}_n(0) = \mathsf{M}_n$ and by using the formula

$$\frac{d}{dt}A^{-1}(t) = -A^{-1}(t)\frac{d}{dt}A(t)A^{-1}(t), \tag{3.45}$$

valid for any matrix function A invertible uniformly in t, we obtain in view of (3.40) and (3.43):

$$\Delta_n(z) = \frac{1}{n}\int_0^1 \frac{d}{dt}\mathbf{E}\{\mathrm{Tr}\mathcal{G}_n(z,t)\}dt = -\frac{1}{n}\int_0^1 \mathbf{E}\{\mathrm{Tr}\mathcal{G}_n^2(z,t)\dot{\mathcal{M}}(t)\}dt,$$

where

$$\dot{\mathcal{M}}_n(t) = \frac{d}{dt}\mathcal{M}_n(t) =: \{\dot{\mathcal{M}}_{\alpha\beta}(t)\}_{\alpha,\beta=1}^n,$$

$$\dot{\mathcal{M}}_{\alpha\beta}(t) = \frac{1}{2n}\sum_{j=1}^n (S_nX_n^T\dot{K}_n(t))_{\alpha j}(t^{-1/2}\eta_j - (1-t)^{-1/2}q_n^{1/2}\gamma_j)(X_nS_n)_{j\beta},$$

and according to (3.20) – (3.21), and (3.43)

$$\dot{K}_{jn}(t) = 2(\varphi'\varphi'')(x)|_{x=\eta_j(t)}. \tag{3.46}$$

By using (3.45) again, we get

$$\Delta_n(z) = \delta_n'(z),$$
$$\delta_n(z) = \frac{1}{2n^2}\sum_{j=1}^n \int_0^1 \mathbf{E}\{F_j(z,t)(t^{-1/2}\eta_j - (1-t)^{-1/2}q_n^{1/2}\gamma_j)\}t^{-1/2}dt, \tag{3.47}$$

where

$$F_j(z,t) = (X_nS_n\mathcal{G}_n(z,t)S_nX_n^T\dot{K}_n(t))_{jj}. \tag{3.48}$$

It suffices to prove that

$$\max_{z\in O}|\delta_n(z)| = o(1),\ n\to\infty, \tag{3.49}$$

where O is an open set lying strictly inside $\mathbb{C}\setminus\mathbb{R}_+$. Indeed, since F_j is analytic in $\mathbb{C}\setminus\mathbb{R}_+$, δ_n is analytic in O and any such bound implies (3.38) by the Cauchy theorem.

To deal with the expectation in the r.h.s. of the second equality in (3.47), we take into account that $\{X_{j\alpha}\}_{\alpha=1}^n$ and γ_j are independent Gaussian random variables (see (3.4)) and (3.20)) and use the simple differentiation formula

$$\mathbf{E}\{\xi f(\xi)\} = \mathbf{E}\{f'(\xi)\} \tag{3.50}$$

valid for the standard Gaussian random variable and any differentiable $f:\mathbb{R}\to\mathbb{C}$ with a polynomially bounded derivative.

The formula, applied to η_j's and γ_j's in the integrand of (3.47), yields

$$\mathbf{E}\{F_j(z,t)(t^{-1/2}\eta_j - (1-t)^{-1/2}q_n^{1/2}\gamma_j)\}$$
$$= (tn)^{-1/2}\sum_{\alpha=1}^n \mathbf{E}\left\{\frac{\partial F_j}{\partial X_{j\alpha}}\right\}x_{\alpha n}, \tag{3.51}$$

where the partial derivative in the r.h.s. denotes the "explicit" derivative (not applicable to $X_{j\alpha}$ in the argument of K_{jn} and $\dot{K}_{jn}$ of (3.43)).

By using the formula

$$\frac{\partial\mathcal{G}_{\beta\gamma}}{\partial X_{j\alpha}} = -\frac{1}{n}(\mathcal{G}SX^TK)_{\beta j}(S\mathcal{G})_{\alpha\gamma} - \frac{1}{n}(\mathcal{G}S)_{\beta\alpha}(KXS\mathcal{G})_{j\gamma}, \tag{3.52}$$

which follows from (3.45) and where we omitted the subindex n in all the matrices and denoted $\mathcal{G} = \mathcal{G}_n(z,t),\ K = K_n(t)$ (see (3.44) and (3.43)), we obtain

$$(tn)^{-1/2}\sum_{\alpha=1}^n \mathbf{E}\left\{\frac{\partial F_j}{\partial X_{j\alpha}}\right\}x_{\alpha n}$$

$$= \frac{2}{(tn)^{1/2}} \mathbf{E}\{(\dot{K}X\mathcal{G}_S x)_j (1 - n^{-1}(KX\mathcal{G}_S X^T)_{jj})\}, \tag{3.53}$$

where

$$\mathcal{G}_S = S\mathcal{G}S. \tag{3.54}$$

We have then via (3.21), (3.22), (3.29), (3.43) and (3.46)

$$|K_j| \le \widetilde{\Phi}_1^2, \ |\dot{K}_j| \le 2\widetilde{\Phi}_1\widetilde{\Phi}_2, \ ||\mathcal{G}_S|| \le \rho\zeta^{-1}. \tag{3.55}$$

This and (3.32) imply that the r.h.s. of (3.53) admits the bounds

$$\frac{4\widetilde{\Phi}_1\widetilde{\Phi}_2\rho}{\zeta(tn)^{1/2}} ||x|| \mathbf{E}\{||X^{(j)}|| + (\widetilde{\Phi}_1)^2 \rho (\zeta n)^{-1} ||X^{(j)}||^3\}$$

$$\le \frac{4\widetilde{\Phi}_1\widetilde{\Phi}_2\rho}{\zeta t^{1/2}} ||x|| (1 + C^{3/4}\rho\zeta^{-1}(\widetilde{\Phi}_1)^2), \tag{3.56}$$

and since, according to (3.9), $||x|| = O(n^{1/2})$, we combine the above bound with (3.51) and (3.53) to conclude that the expectation in the r.h.s. of (3.47) is $\varepsilon_n(z,t)$, where $\varepsilon_n(z,t) = O(n^{1/2})$, $n \to \infty$ uniformly in $t \in [0,1]$ and z belonging to an open set O lying strictly inside $\mathbb{C} \setminus \mathbb{R}_+$.

We have proved (3.49), hence (3.38), both with the r.h.s. of the order $O(n^{-1/2})$ uniformly in $z \in O \subset \mathbb{C} \setminus \mathbb{R}_+$. □

Remark 3.4. The "interpolating" random variable (3.42) implements a simple version of the "interpolation" procedure used in [16], Theorem 5.7 and in [17], Sections 18.3 - 18.4 and 19.1 - 19.2 to pass from the Gaussian random matrices to matrices with i.i.d. entries. The procedure can be viewed as a manifestation of the so-called Lindeberg principle, see [8] for related results and references.

We will find now the limiting eigenvalue distribution of a class of random matrices containing (3.19) and used in the proof of Theorem 3.1. In particular, we obtain functional equations (3.10) – (3.12) determining uniquely the Stieltjes transform of the distribution, hence, the distribution. We will use for these, more general, matrices the same notation M_n. Note that we give here a rather simple version of the assertion sufficient to prove Theorem 3.1. For more general versions see, e.g. [6] and references therein.

Lemma 3.5. *Consider the $n \times n$ random matrix*

$$\mathsf{M}_n = n^{-1} S_n X_n^T \mathsf{K}_n X_n S_n, \tag{3.57}$$

(see (3.19) – (3.20)), where S_n satisfies (3.2) and (3.3), X_n has standard Gaussian entries (see (3.4)) and K_n is a $n \times n$ positive definite matrix such that (cf. (3.2) – (3.3))

$$\sup_n n^{-1} \mathrm{Tr} \mathsf{K}_n^2 \le k_2 < \infty, \tag{3.58}$$

$$\lim_{n\to\infty} \nu_{\mathsf{K}_n} = \nu_{\mathsf{K}}, \ \nu_{\mathsf{K}}(\mathbb{R}) = 1, \tag{3.59}$$

where ν_{K_n} is the Normalized Counting Measure of K_n, ν_{K} is a non-negative and not concentrated ar zero measure (cf. (3.2) – (3.3)) and lim *denotes the weak convergence of probability measures.*

Then the Normalized Counting Measure $\nu_{\mathbf{M}_n}$ of M_n converges weakly with probability 1 to a non-random measure ν_{M}, $\nu_{\mathsf{M}}(\mathbb{R}_+) = 1$ and its Stieltjes transform f_{M} (see (1.19)) can be obtained from the system (3.11) – (3.12) in which ν_K is replaced by ν_{K} of (3.58) – (3.59) and which is uniquely solvable in the class of pairs (h, k) of functions such that h is analytic outside the positive semi-axis, continuous and positive on the negative semi-axis and satisfies (3.14).

Remark 3.6. (i) It is easy to check that the assertions of the lemma remain valid with probability 1 in the case where the "parameters" of the theorem, i.e., S_n, hence R_n, in (3.2) – (3.3) and K_n (3.58) – (3.59) are random, defined for all n on the same probability space $\Omega_{R\mathsf{K}}$, independent of $X_n = \{X_{j\alpha}\}_{j,\alpha=1}^n$ for every n and satisfies conditions (3.2) – (3.3) and (3.58) – (3.59) with probability 1 on $\Omega_{R\mathsf{K}}$, i.e., on a certain subspace $\overline{\Omega}_{R\mathsf{K}} \subset \Omega_{R\mathsf{K}}$, $\mathbf{P}(\overline{\Omega}_{R\mathsf{K}}) = 1$. This follows from an argument analogous to that presented in Remark 3.2 (i). In this case $\mathbf{E}\{...\}$ denotes the expectation with respect to X_n.

(ii) Repeating almost literally the proof of the lemma, one can treat a more general case where S_m is a $m \times m$ positive definite matrix satisfying (3.2) – (3.3), K_n is a $n \times n$ positive definite matrix satisfying (3.58) – (3.59), X_n is a $n \times m$ Gaussian random matrix satisfying (1.6) and (cf. (1.13))

$$\lim_{m\to\infty, n\to\infty} m/n = c \in (0, \infty). \tag{3.60}$$

In this case the Stieltjes transform f_{M} of the limiting NCM is again uniquely determined by three functional equations, where the first and the third coincide with (3.10) and (3.12) while the second is (3.11) in which $k(z)$ is replaced by $k(z)c^{-1}$ (see, e.g. [6]) and references therein.

(iii) The lemma is also valid for not necessarily Gaussian X_n (see [6, 18] and references therein for more general cases of the theorem and their properties. If, however, we confine ourselves to the Gaussian case, then we can reformulate our result in terms of correlated Gaussian entries. Indeed, let $Z_n = \{Z_{j\alpha}\}_{j,\alpha=1}^n$ be a Gaussian matrix with

$$\mathbf{E}\{Z_{j\alpha}\} = 0, \ \mathbf{E}\{Z_{j_1\alpha_1} Z_{j_2\alpha_2}\} = C_{j_1\alpha_1, j_2\alpha_2},$$

and a separable covariance matrix $C_{j_1\alpha_1, j_2\alpha_2} = \mathsf{K}_{j_1 j_2} R_{\alpha_1\alpha_2}$, i.e., $C = \mathsf{K} \otimes R$ and $\mathsf{K}_n = \{\mathsf{K}_{j_1 j_2}\}_{j_1, j_2=1}^n$ and $R_n = \{R_{\alpha_1\alpha_2}\}$ as in the lemma. Writing $\mathsf{K}_n = \mathsf{D}_n^2$, $R_n = S_n^2$ and denoting $Z_n = S_n X_n D_n$, we can view as a data matrix and then the corresponding sample covariance matrix $Z_n^T Z_n$ is (3.57) of spatial-temporal correlated time series.

Here is the proof of Lemma 3.5.

Proof. As it was in the proof of Theorem 3.1, Lemma 3.9 (i) together with the Borel-Cantelli lemma reduce the proof of the theorem to that of the weak convergence of the expectation

$$\overline{\nu}_{\mathsf{M}_n} = \mathbf{E}\{\nu_{\mathsf{M}_n}\} \tag{3.61}$$

of ν_{M_n}.

Next, it follows from the condition of the lemma that the argument analogous to that proving (3.18) yields

$$\int_0^\infty \lambda \overline{\nu}_{\mathsf{M}_n}(d\lambda) \leq k_2^{1/2} r_2^{1/2} < \infty,$$

hence, the tightness of measures $\{\overline{\nu}_{\mathsf{M}_n}\}_n$ and, in turn, reduces the proof of the lemma to that of the pointwise convergence in $\mathbb{C} \setminus \mathbb{R}_+$ of their Stieltjes transforms f_{M_n} to the limit f satisfying (3.10) – (3.12). Moreover, the analyticity of f_{M_n}, f_{M}, h and k in $\mathbb{C}\setminus\mathbb{R}_+$ (see Lemma 3.12) allows us to confine ourselves to the open negative semi-axis

$$I_- = \{z \in \mathbb{C} : z = -\xi,\ 0 < \xi < \infty\}. \tag{3.62}$$

Thus, we will mean and often write explicitly below that $z \in I_-$.

Note first that since $\{X_{j\alpha}\}_{j,\alpha=1}^n$ are standard Gaussian, we can assume without loss of generality that S_n and K_n are diagonal, i.e.,

$$S_n = \{\delta_{\alpha\beta} S_{\alpha n}\}_{\alpha,\beta=1}^n,\ \mathsf{K}_n = \{\delta_{jk} \mathsf{K}_{jn}\}_{j,k=1}^n. \tag{3.63}$$

Given $j = 1, ..., n$, consider the $n \times n$ matrix

$$H^{(j)} = \{H^{(j)}_{\alpha\beta}\}_{\alpha,\beta=1}^n,\ H^{(j)}_{\alpha\beta} := n^{-1}(\mathsf{G}SX^T)_{\alpha j}(\mathsf{K}XS)_{j\beta} \tag{3.64}$$

and we omit here and below the subindex n in the notation of matrices and their entries.

It follows from (3.39) – (3.40) and the resolvent identity

$$\mathsf{G} = -z^{-1} + z^{-1}\mathsf{G}\mathsf{M}, \tag{3.65}$$

implying

$$z^{-1}\sum_{j=1}^n \mathbf{E}\{H^{(j)}\} = z^{-1}\mathbf{E}\{\mathsf{G}\mathsf{M}\}, \tag{3.66}$$

that it suffices to find the $n \to \infty$ limit of $\mathbf{E}\{n^{-1}\mathrm{Tr}H^{(j)}\}$.

To this end we will apply to the expectation in the r.h.s. of (3.64) the Gaussian differentiation formula (3.50). We compute the derivative of $\mathsf{G}_{\alpha\gamma}$ with respect to $X_{j\gamma}$ by using an analog of (3.52) and we obtain

$$\begin{aligned} \mathbf{E}\{H^{(j)}_{\alpha\beta}\} &= n^{-1}\mathbf{E}\{\mathsf{G}_{\alpha\beta}\}S_\beta^2\mathsf{K}_j - \mathbf{E}\{h_n(z)H^{(j)}_{\alpha\beta}\}\mathsf{K}_j \\ &\quad -n^{-2}\mathbf{E}\{(\mathsf{G}S^2\mathsf{G}SX^T)_{\alpha j}(XS)_{j\beta}\}\mathsf{K}_j^2, \end{aligned} \tag{3.67}$$

where

$$h_n(z) = n^{-1}\mathrm{Tr}S\mathsf{G}S = n^{-1}\mathrm{Tr}\mathsf{G}R,\ R = S^2. \tag{3.68}$$

We write

$$h_n = \overline{h}_n + (h_n - \overline{h}_n),\ \overline{h}_n = \mathbf{E}\{h_n\} \tag{3.69}$$

in the r.h.s. of (3.67) and get

$$\begin{aligned} \mathbf{E}\{H^{(j)}_{\alpha\beta}\} &= n^{-1}\mathbf{E}\{\mathsf{G}_{\alpha\beta}\}S_\beta^2 Q_j - \mathbf{E}\{(h_n - \overline{h}_n)H^{(j)}_{\alpha\beta}\}Q_j \\ &\quad -n^{-2}\mathbf{E}\{(\mathsf{G}S^2\mathsf{G}SX^T)_{\alpha j}(XS)_{j\beta}\}\mathsf{K}_j Q_j, \end{aligned} \tag{3.70}$$

where

$$Q_j(z) = \mathsf{K}_j(\overline{h}_n(z)\mathsf{K}_j + 1)^{-1} \tag{3.71}$$

is well defined for $z = -\xi < 0$. Indeed, since R is positive definite, it follows from (3.68), the spectral theorem and (3.2) that $\overline{h}_n$ admits the representation

$$h_n(z) = \int_0^\infty \frac{\mu(d\lambda)}{\lambda - z},\ \mu \geq 0,\ \mu(\mathbb{R}_+) = n^{-1}\mathrm{Tr}R_n \leq r_2^{1/2} < \infty. \tag{3.72}$$

Thus, we have in view of (3.62)

$$0 < h_n(-\xi) \leq r_2^{1/2}/\xi < \infty, \tag{3.73}$$

and then the positivity of K_j of and (3.63) imply

$$0 < Q_j(-\xi) \leq \mathsf{K}_j,\ \xi > 0. \tag{3.74}$$

We then sum (3.70) over $j = 1, ..., n$ and denote

$$H_{\alpha\beta} = \sum_{j=1}^n H_{\alpha\beta}^{(j)},\ H = \{H_{\alpha\beta}\}_{\alpha,\beta=1}^n \tag{3.75}$$

yielding

$$\mathbf{E}\{H\} = \mathbf{E}\{\mathsf{G}\}k_n(-\xi)R - T,\ T = T^{(1)} + T^{(2)}, \tag{3.76}$$

where

$$k_n(-\xi) := n^{-1}\sum_{j=1}^n Q_j = \int \frac{\lambda\nu_{\mathsf{K}_n}(d\lambda)}{\overline{h}_n(-\xi)\lambda + 1}, \tag{3.77}$$

ν_{K_n} is the NCM of K_n (see (3.59)) and

$$\begin{aligned} T^{(1)} &= n^{-1}\mathbf{E}\{(h_n - \overline{h}_n)\mathsf{G}SX^T\mathsf{K}QXS\}, \\ T^{(2)} &= -n^{-2}\mathbf{E}\{\mathsf{G}S^2\mathsf{G}SX^T\mathsf{K}QXS\}. \end{aligned} \tag{3.78}$$

Plugging now (3.76) into (3.66) and the obtained expression in the r.h.s. of expectation of (3.65), we get

$$\mathbf{E}\{\mathsf{G}\}(k_n(z)R - z) = 1 - T. \tag{3.79}$$

The matrix $(k_n(-\xi)R + \xi)$ is invertible uniformly in n. Indeed, since R is positive definite and $k_n(-\xi),\ \xi > 0$ is positive in view of (3.73) and (3.77), we have uniformly in $n \to \infty$:

$$||(k_n(-\xi)R + \xi)||^2 \geq \xi^2. \tag{3.80}$$

Thus, we can write instead of (3.79)

$$\mathbf{E}\{\mathsf{G}\} = \overline{\mathsf{G}} - \overline{\mathsf{G}}T,\ \overline{\mathsf{G}} = (k_n(-\xi)R + \xi)^{-1} \tag{3.81}$$

yielding in view of the spectral theorem for R_n

$$f_{\mathsf{M}_n}(-\xi) = \int_0^\infty \frac{\nu_{R_n}(d\lambda)}{k_n(-\xi)\lambda + \xi} + t_n(-\xi),\ t_n(-\xi) = -n^{-1}\mathrm{Tr}\overline{\mathsf{G}}T, \tag{3.82}$$

where ν_{R_n} is the NCM of R_n (see (3.3)).

Next, multiplying (3.81) by R and applying to the result the operation $n^{-1}\mathrm{Tr}$, we obtain in view of (3.68) and (3.69)

$$\overline{h}_n(-\xi) = \int_0^\infty \frac{\lambda\nu_{R_n}(d\lambda)}{k_n(-\xi)\lambda+\xi} + \widetilde{t}_n(-\xi),\ \widetilde{t}_n(-\xi) = -n^{-1}\mathrm{Tr}\overline{\mathsf{G}}RT. \tag{3.83}$$

The integral terms in the r.h.s. of (3.77), (3.82) and (3.83) are obviously the prelimit versions of the r.h.s. of (3.10) – (3.12). Thus we have to show that the remainder terms t_n and $\widetilde{t}_n$ in (3.82) and (3.83) vanish as $n\to\infty$ under the condition (3.62) and to carry out the limiting transition in the integral terms of (3.77), (3.82) and (3.83). The second procedure is quite standard in random matrix theory and based on (3.3) and (3.59), the compactness of sequences of bounded analytic functions with respect to the uniform convergence on a compact set of complex plane, the compactness of sequences on probability measures with respect to the weak convergence and the unique solvability of the system (3.11) – (3.12) proved in Lemma 3.12 (see, e.g. [17] for a number of examples of the procedure).

Thus, we will deal with the remainders in (3.82) – (3.83). We will assume for time being that the matrix $R_n = S_n^2$ of (3.2) is uniformly bounded in n (see (3.22)). This assumption can be removed at the end of the proof by using an argument analogous to that used at the end of proof of Theorem 3.1. Recall that we are assuming that $z=-\xi\in I_-$ of (3.62).

We will start with the contribution

$$t_n^{(1)} = -n^{-2}\mathbf{E}\{(h_n-\overline{h}_n)\mathrm{Tr}S\mathsf{G}\overline{\mathsf{G}}S\mathsf{B}\},\ \mathsf{B} = X^T\mathsf{K}QX\ge 0, \tag{3.84}$$

of $T^{(1)}$ in (3.78) to $t_n(-\xi)$ of (3.82). We have from (2.20), (3.22), (3.29) and (3.84):

$$\begin{aligned} n^{-2}|\mathrm{Tr}S\mathsf{G}\overline{\mathsf{G}}S\mathsf{B}| &\le \rho(\xi n)^{-2}\mathrm{Tr}\mathsf{B} \\ &\le \rho(\xi n)^{-2}\sum_{j=1}^n ||X^{(j)}||^2\mathsf{K}_jQ_j, \end{aligned}$$

where $X^{(j)}$ is the jth column of X. This, Schwarz inequality for expectations, (3.32), (3.58) and (3.74) yield

$$\begin{aligned} |t_n^{(1)}| &\le \rho\xi^{-2}n^{-2}\sum_{j=1}^n \mathsf{K}_jQ_j\mathbf{E}^{1/2}\{||X^{(j)}||^4\}\mathbf{E}^{1/2}\{|h(-\xi)-\overline{h}_n(-\xi)|^2\} \\ &\le \rho k_2\xi^{-2}C^{1/2}\mathbf{E}^{1/2}\{|h_n(-\xi)-\overline{h}_n(-\xi)|^2\} \end{aligned}$$

and then an analog of Lemma 3.9 (iii) for M_n and (3.73) implies for every $\xi>0$

$$|t_n^{(1)}| = O(n^{-1/2}),\ n\to\infty. \tag{3.85}$$

Similarly, we have for the contribution

$$t_n^{(2)} = n^{-3}\mathbf{E}\{\mathrm{Tr}S\overline{\mathsf{G}}\mathsf{G}S^2\mathsf{G}S\mathsf{B}\}$$

of $T^{(2)}$ of (3.78) to t_n in (3.83) by (2.20), (3.22) and (3.29): $|t_n^{(2)}|\le\rho^2\xi^{-3}n^{-3}\mathbf{E}\{\mathrm{Tr}\mathsf{B}\}$ and then for every $\xi>0$

$$t_n^{(2)} = O(n^{-1}),\ n\to\infty. \tag{3.86}$$

Combining now (3.85) – (3.86), we obtain $t_n(-\xi) = O(n^{-1/2}),\ n\to\infty,\ \xi>0$.

By using a similar argument, we find that $\widetilde{t}_n(-\xi) = O(n^{-1/2}),\ n \to \infty,\ \xi > 0$. This and (3.82) – (3.83) with $z = -\xi < 0$ lead to (3.10) and (3.11). Multiplying (3.11) by k and using the first equality in (3.10), we obtain the second equality.

The unique solvability of system (3.11) – (3.12) is proved in Lemma 3.12. $\square$

It is convenient to write the equations (3.11) – (3.12) in a compact form similar to that of free probability theory [5, 13]. This, in particular, makes explicit the symmetry and the transitivity of the binary operation (2.1).

Corollary 3.7. *Let ν_{K}, ν_R and ν_{M} be the probability measures (i.e., non-negative measures of the total mass 1) entering (3.10) – (3.12) and m_{K}, m_R and m_{M} be their moment generating functions (see (1.20) – (1.21)). Then their functional inverses z_{M}, z_{K} and z_R of the corresponding moment generating functions are related as follows*

$$z_{\mathsf{M}}(m) = z_{\mathsf{K}}(m) z_R(m) m^{-1}, \tag{3.87}$$

or, writing $z_A(m) = m\sigma_A(m),\ A = \mathsf{M}, \mathsf{K}, R$,

$$\sigma_{\mathsf{M}}(m) = \sigma_{\mathsf{K}}(m)\sigma_R(m) \tag{3.88}$$

Proof. It follows from (3.11) – (3.12) and (1.21) that

$$\begin{aligned} m_{\mathsf{K}}(-h(z)) &= -h(z)k(z),\ m_R(k(z)z^{-1}) = -h(z)k(z), \\ m_{\mathsf{M}}(z^{-1}) &= -h(z)k(z). \end{aligned} \tag{3.89}$$

Now the first and the third relations (3.89) yield $m_{\mathsf{K}}(-h(z^{-1})) = m_{\mathsf{M}}(z)$, hence $z_{\mathsf{K}}(m) = -h(z_{\mathsf{M}}^{-1}(m))$, and then the second and the third relations yield $m_R(k(z^{-1})z) = m_{\mathsf{M}}(z)$, hence $z_R(\mu) = k(z_{\mathsf{M}}^{-1}(m))z_{\mathsf{M}}(m)$. Multiplying these two relations and using once more the third relation in (3.89), we obtain

$$z_{\mathsf{K}}(m)z_R(m) = -k(z_{\mathsf{M}}^{-1}(m))h(z_{\mathsf{M}}^{-1}(m))z_{\mathsf{M}}(m) = z_{\mathsf{M}}(m)m$$

and (3.87) and (3.88) follows. $\square$

Remark 3.8. In the case of rectangular matrices X_n in (3.1), described in Remark 3.6 (ii), the analogs of (3.87) and (3.88) are

$$z_{\mathsf{M}}(m) = z_{\mathsf{M}}(cm)z_R(cm)m^{-1},\ \sigma_{\mathsf{M}}(m) = c^2\sigma_{\mathsf{K}}(cm)\sigma_R(cm). \tag{3.90}$$

Lemma 3.9. *Let $\mathcal{M}_n$ be given by (3.1) in which $X_n = \{X_{j\alpha}\}_{j,\alpha=1}^n$ of (3.4) and $b_n = \{b_{j\alpha}\}_j^n$ of (3.5) are i.i.d. random variables. Denote $\nu_{\mathcal{M}_n}$ the Normalized Counting Measure of $\mathcal{M}_n$ (see, e.g. (1.15)), $g_n(z)$ its Stieltjes transform*

$$g_n(z) = n^{-1}\mathrm{Tr}(\mathcal{M}_n - z)^{-1},\ \zeta = \mathrm{dist}(z, \mathbb{R}_+) > 0,$$

and (see (3.68))

$$h_n(z) = n^{-1}\mathrm{Tr}S_n(\mathcal{M}_n - z)^{-1}S_n,\ \zeta = \mathrm{dist}(z, \mathbb{R}_+) > 0,$$

where S_n is a positive definite matrix satisfying (3.2) with $R_n = S_n^2$. Then we have:

(i) *for any nindependent interval Δ of spectral axis*

$$\mathbf{E}\{|\nu_{\mathcal{M}_n}(\Delta) - \mathbf{E}\{\nu_{\mathcal{M}_n}(\Delta)\}|^4\} \le C_1/n^2,$$

where C_1 is an absolute constant;

(ii) *for any n-independent z with $\zeta > 0$*

$$\mathbf{E}\{|g_n - \mathbf{E}\{g_n\}|^4\} \le C_2/n^2\zeta^4,$$

where C_2 is an absolute constant;

(iii) *for any n-independent z with $\zeta > 0$*

$$\mathbf{E}\{|h_n - \mathbf{E}\{h_n\}|^4\} \le C_3 r_2^2/n^2\zeta^4,$$

where C_3 is an absolute constant and r_2 is defined in (3.2).

Proof. It follows from (3.1) that (cf. (3.35))

$$\mathcal{M}_n = \sum_{j=1}^{n} \mathcal{Y}_j \otimes \mathcal{Y}_j,\ \mathcal{Y}_j = \{\mathcal{Y}_{j\alpha}\}_{\alpha=1}^{n},\ \mathcal{Y}_{j\alpha} = n^{-1/2}(D_n X S_n)_{j\alpha}. \tag{3.91}$$

It is easy to see that $\{\mathcal{Y}_j\}_{j=1}^n$ are independent. This allows us to use the martingale bounds given in Sections 18.2 and 19.1 of [17] and implying the assertions of the lemma in view of (3.30) and (3.72) . □

Remark 3.10. (i) The independence of random vectors $\mathcal{Y}_j$ in (3.91) is the main reason to pass from the matrices M_n^l given by (2.11) and (2.13) to the matrices $\mathcal{M}_n$ given by (2.12), (2.16) and (3.1).

(ii) The lemma is valid for an arbitrary (not necessarily Gaussian) collection (3.4) and (3.5) of i.i.d. random variables as well as for random but independent of (3.4) and (3.5) S_n and $\{x_{j,\alpha}\}_{j,\alpha=1}^n$, see Remarks 3.2 (i) and (iii) and [18]. It is also valid for matrices M_n of (3.19).

The next lemma deals with asymptotic properties of the vectors of activations x^l in the lth layer, see (1.2). It is an extended version (treating the convergence with probability 1) of assertions proved in [12, 22, 26].

Lemma 3.11. *Let $y^l = \{y_j^l\}_{j=1}^n$, $l = 1, 2, ...$ be post-affine random vectors defined in (1.2) – (1.6) with x^0 satisfying (2.2), $\chi : \mathbb{R} \to \mathbb{R}$ be a bounded piecewise continuous function and Ω_l be defined in (1.9). Set*

$$\chi_n^l = n^{-1} \sum_{j_l=1}^{n} \chi(y_{j_l}^l),\ l \ge 1. \tag{3.92}$$

Then there exists $\overline{\Omega}_l \subset \Omega_l$, $\mathbf{P}(\overline{\Omega}_l) = 1$ such that for every $\omega_l \in \overline{\Omega}_l$ (i.e., with probability 1) the limit

$$\chi^l := \lim_{n\to\infty} \chi_n^l,\ l = 1, 2, ..., \tag{3.93}$$

exists, is non-random and given by the formula

$$\chi^l = \int_{-\infty}^{\infty} \chi(\gamma\sqrt{q^l})\Gamma(d\gamma),\ l = 1, 2, ..., \tag{3.94}$$

valid on $\overline{\Omega}_l$ with $\Gamma(d\gamma) = (2\pi)^{-1/2}e^{-\gamma^2/2}d\gamma$ being the standard Gaussian probability distribution and q^l defined recursively by the formula

$$q^l = \int_{-\infty}^{\infty} \varphi^2(\gamma\sqrt{q^{l-1}})\Gamma(d\gamma) + \sigma_b^2,\ l = 2, 3, ... \tag{3.95}$$

and by q_1 of (2.2).

In particular, we have with probability 1:

(i) *for the activation vector $x^l = \{x_j^l\}_{j=1}^n$ of the lth layer (see (1.2)):*

$$\lim_{n\to\infty} n^{-1}\sum_{j_l=1}^{n}(x_{j_l}^l)^2 = q^{l+1} - \sigma_b^2,\ l = 1, 2, ... \tag{3.96}$$

(ii) *for the weak limit ν_{K^l} of the Normalized Counting Measure $\nu_{K_n^l}$ of diagonal random matrix K_n^l of (1.22): ν_{K^l} is the probability distribution of the random variable $(\varphi'(\gamma\sqrt{q_l}))^2$.*

Proof. Set $l = 1$ in (3.92) Since $\{b_{j_1}^1\}_{j_1=1}^n$ and $\{X_{j_1,j_0}^1\}_{j_1,j_0=1}^n$ are i.i.d. Gaussian random variables satisfying (1.5) – (1.6), it follows from (1.2) that the components of $y^1 = \{y_{j_1}^1\}_{j_1=1}^n$ are also i.i.d. Gaussian random variables of zero mean and variance q_n^1 of (2.2). Since χ is bounded, the collection $\{\chi(y_{j_1}^1)\}_{j_1=1}^n$ consists of bounded i.i.d random variables defined for all n on the same probability space Ω_1 generated by (1.7) and (1.8) with $l = 1$. This allows us to apply to $\{\chi(y_{j_1}^1)\}_{j_1}^n$ the strong Law of Large Numbers implying (3.93) with $l = 1$ together with the formula

$$\begin{aligned} \chi^1 &= \lim_{n\to\infty} \mathbf{E}\{\chi(y_1^1)\} \\ &= \lim_{n\to\infty}\int_{-\infty}^{\infty}\chi(\gamma\sqrt{q_n^1})\Gamma(d\gamma) = \int_{-\infty}^{\infty}\chi(\gamma\sqrt{q^1})\Gamma(d\gamma) \end{aligned} \tag{3.97}$$

for the limit, both valid with probability 1, i.e., on a certain $\overline{\Omega}_1 \subset \Omega_1 = \Omega^1$, $\mathbf{P}(\Omega_1) = 1$, see (1.9). This yields (3.94) for $l = 1$.

Consider now the case $l = 2$. Since $\{X^1, b^1\}$ and $\{X^2, b^2\}$ are independent collections of random variables, we can fix $\omega_1 \in \overline{\Omega}_1$ (a realization of $\{X^1, b^1\}$) and apply to χ_n^2 of (3.92) the same argument as that for the case $l = 1$ above to prove that for every $\omega_1 \in \overline{\Omega}_1$ there exists $\overline{\Omega}^2(\omega^1) \subset \Omega^2$, $\mathbf{P}(\overline{\Omega}^2(\omega^1) = 1$ on which we have (3.93) for $l = 2$ with some (cf. (3.97))

$$\chi^2(\omega^1, \omega^2) = \lim_{n\to\infty}\mathbf{E}_{\{X^2,b^2\}}\{\chi(y_1^2)\} \tag{3.98}$$

where $\mathbf{E}_{\{X^2,b^2\}}\{...\}$ denotes the expectation with respect to $\{X^2, b^2\}$ only. Now the Fubini theorem implies that there exists $\overline{\Omega}_2 \subset \Omega_2 = \Omega^1 \otimes \Omega^2$, $\mathbf{P}(\overline{\Omega}_2) = 1$ on which we have (3.93) with $l = 2$.

Using once more the independence of $\{X^1, b^1\}$ and $\{X^2, b^2\}$, we can compute the r.h.s. of (3.98) by observing that if $\{X^2, b^2\}$ are Gaussian, then, according to (1.2), y_1^2 is also Gaussian of zero mean and variance (cf. (2.2))

$$q_n^2 = n^{-1}\sum_{j_1=1}^{n}(x_{j_1}^1)^2 + \sigma_b^2, \tag{3.99}$$

or, in view of (1.2),

$$q_n^2 = n^{-1}\sum_{j_1=1}^{n}(\varphi(y_{j_1}^1))^2 + \sigma_b^2. \tag{3.100}$$

The first term on the right is a particular case of (3.92) with $\chi = (\varphi)^2$ and $l = 1$, thus, according to (3.97), the limiting form of the above relation is (3.95) with $l = 2$ for every $\omega_1 \in \overline{\Omega}_1$ and we have

$$\chi^2 = \lim_{n\to\infty} \int_{-\infty}^{\infty} \chi(\gamma\sqrt{q_n^2})\Gamma(d\gamma) = \int_{-\infty}^{\infty} \chi(\gamma\sqrt{q^2})\Gamma(d\gamma)$$

i.e., formula (3.94) for $l = 2$ valid on $\overline{\Omega}_2 \subset \Omega_2 = \Omega^1 \otimes \Omega^2$, $\mathbf{P}(\overline{\Omega}_2) = 1$, i.e., with probability 1.

This proves the validity (3.93) – (3.95) for $l = 2$ with probability 1. Analogous argument applies for $l = 3, 4, ...$

The proof of item (i) is, in fact, that of (3.95), see (3.99) – (3.100) for $l = 2$, for $l \geq 3$ the proof is analogous.

Let us prove item (ii) of the lemma, i.e., the weak convergence with probability 1 of the Normalized Counting Measure $\nu_{K_n^l}$ of K_n^l in (1.22) to the probability distribution of $(\varphi'(\gamma\sqrt{q_l}))^2$. It suffices to prove the validity with probability 1 of the relation

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \psi(\lambda)\nu_{K_n^l}(d\lambda) = \int_{-\infty}^{\infty} \psi(\lambda)\nu_{K^l}(d\lambda)$$

for any bounded and piece-wise continuous $\psi : \mathbb{R} \to \mathbb{R}$.

In view of (1.2), (1.11) and (1.22) the relation can be written in the form

$$\lim_{n\to\infty} n^{-1} \sum_{j_l=1}^{n} \psi((\varphi'(y_{j_l}^l))^2) = \int_{-\infty}^{\infty} \psi((\varphi'(\gamma\sqrt{q_{l-1}}))^2)\Gamma(d\gamma),\ l \geq 1.$$

This is a particular case of (3.93) – (3.95) for $\chi = \psi \circ \varphi'^2$, hence, assertion (ii) follows. $\square$

The next lemma provides the unique solvability of the system (3.11) – (3.12). Note that in the course of proving Lemma 3.5 it was proved that the system has at least one solution.

Lemma 3.12. *The system (3.11) – (3.12) with ν_R and ν_K satisfying*

(3.101) $$\nu_K(\mathbb{R}_+) = 1,\ \nu_R(\mathbb{R}_+) = 1$$

and (cf. ((3.2))

(3.102) $$\int_0^{\infty} \lambda^2 \nu_K(d\lambda) = \kappa_2 < \infty,\ \int_0^{\infty} \lambda^2 \nu_R(d\lambda) = \rho_2 < \infty$$

has a unique solution in the class of pairs of functions (h, k) defined in $\mathbb{C} \setminus \mathbb{R}_+$ and such that h is analytic in $\mathbb{C} \setminus \mathbb{R}_+$, continuous and positive on the open negative semi-axis and satisfies (3.14) with r_2 replaced by ρ_2 of (3.102).

Besides:

(i) *the function k is analytic in $\mathbb{C} \setminus \mathbb{R}_+$, continuous and positive on the open negative semi-axis and (cf. (3.14))*

(3.103) $$\Im k(z)\Im z < 0 \text{ for } \Im z \neq 0,\ 0 < k(-\xi) \leq \kappa_2^{1/2} \text{ for } \xi > 0$$

with κ_2 of (3.102);

(ii) *if the sequence* $\{\nu_{K^{(p)}}, \nu_{R^{(p)}}\}_p$ *has uniformly in* p *bounded second moments (see (3.102)) and converges weakly to* (ν_K, ν_R) *also satisfying (3.102), then the sequences of the corresponding solutions* $\{h^{(p)}, k^{(p)}\}_p$ *of the system (3.11) – (3.12) converges pointwise in* $\mathbb{C} \setminus \mathbb{R}_+$ *to the solution* (h, k) *of the system corresponding to the limiting measures* (ν_K, ν_R).

Proof. We will start with the proof of assertion (i). It follows from (3.12), (3.102) and the analyticity of h in $\mathbb{C} \setminus \mathbb{R}_+$ that k is also analytic in $\mathbb{C} \setminus \mathbb{R}_+$. Next, for any solution of (3.11) – (3.12) we have from (3.12) with $\Im z \neq 0$

$$\Im k(z) = -\Im h(z) \int_0^\infty \frac{\lambda^2 \nu_K(d\lambda)}{|h(z)\lambda + 1|^2} \tag{3.104}$$

and then (3.14) yields (3.103) for $\Im z \neq 0$, while (3.12) with $z = -\xi < 0$

$$k(-\xi) = \int_0^\infty \frac{\lambda \nu_K(d\lambda)}{h(-\xi)\lambda + 1},$$

the positivity of $h(-\xi)$ (see (3.14)), (3.101) and Schwarz inequality yield (3.103) for $z = -\xi$.

Let us prove now that the system (3.11) – (3.12) is uniquely solvable in the class of pairs of functions (h, k) analytic in $\mathbb{C} \setminus \mathbb{R}_+$ and satisfying (3.14) and (3.103).

Denote $\mathbb{C}_+$ and $\mathbb{C}_-$ the upper and lower open half-planes. Consider first the case $z \in \mathbb{C}_+$ of the system (3.11) – (3.12). To this end introduce the map

$$F : \{h \in \mathbb{C}_+\} \times \{k \in \mathbb{C}_-\} \times \{z \in \mathbb{C}_+\} \to \mathbb{C} \times \mathbb{C} \tag{3.105}$$

defined by

$$\begin{aligned} F_1(h, k, z) &= h - \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{k\lambda - z}, h \\ F_2(h, k, z) &= k - \int_0^\infty \frac{\lambda \nu_K(d\lambda)}{h\lambda + 1}. \end{aligned} \tag{3.106}$$

The map is well defined in the indicated domain, since there $\Im|k\lambda - z| > \lambda|\Im k|$ and $\Im|1 + h\lambda| > \lambda \Im h$, hence the absolute values of the integrals in F_1 and that in F_2 are bounded from above by $|\Im k|^{-1} < \infty$ and $(\Im h)^{-1} < \infty$ respectively. The equation

$$F(h, k, z) = 0 \tag{3.107}$$

is in fact (3.11) – (3.12). We will apply now to the equation the implicit function theorem. To this end we have to prove that the Jacobian of F, i.e., 2×2 matrix of derivatives of F with respect to h and k, is invertible. It is easy to find that the determinant of the Jacobian is

$$1 - I(h)J(k, z) \tag{3.108}$$

with

$$I(h) = \int_0^\infty \frac{\lambda^2 \nu_K(d\lambda)}{(h\lambda + 1)^2} \leq A(h), \ J(k, z) = \int_0^\infty \frac{\lambda^2 \nu_R(d\lambda)}{(k\lambda - z)^2} \leq B(k, z)$$

and

$$0 \ < \ A(h) := \int_0^\infty \frac{\lambda^2 \nu_K(d\lambda)}{|h\lambda + 1|^2} \leq (\Im h)^{-2} < \infty,$$

$$0 \quad < \quad B(k,z) := \int_0^\infty \frac{\lambda^2 \nu_R(d\lambda)}{|k\lambda - z|^2} \le (\Im k)^{-2} < \infty,$$

where we used (3.101) to obtain the second inequality.

On the other hand, the imaginary part of (3.106) – (3.107) yield (cf. (3.104))

$$A(h) = -\frac{\Im k}{\Im h},\ B(k,z) = -\frac{\Im h}{\Im k} + \frac{\Im z}{\Im k} C(k,z),$$

where

$$0 < C(k,z) = \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{|k\lambda - z|^2} < \infty,\ \Im z > 0. \tag{3.109}$$

This implies

$$0 < A(h)B(k,z) = 1 - \Im z (\Im h)^{-1} C(k,z) \tag{3.110}$$

and since $C(k,z)\,\Im z(\Im h)^{-1} > 0$ in view of (3.105) and (3.109), we have for the determinant (3.108)

$$|1 - I(h)J(k,z)| \ge 1 - A(h)B(k,z) = C(k,z)\,\Im z (\Im h)^{-1} > 0. \tag{3.111}$$

Thus, the Jacobian of the map (3.105) – (3.106) is invertible and the system (3.11) – (3.12) is uniquely solvable in $\mathbb{C}_+$. The proof for $\mathbb{C}_-$ is analogous.

Assume now that $z = -\xi,\ \xi > 0$. Here we consider the map

$$\widetilde{F} : \{h \in \mathbb{R}_+ \setminus \{0\}\} \times \{k \in \mathbb{R}_+\} \times \{z = -\xi \in \mathbb{R}_- \setminus \{0\}\} \to \mathbb{R} \times \mathbb{R}$$

defined by (3.106) with $h > 0,\ k > 0,\ z = -\xi < 0$. It is easy to find that the map is well defined since $k\lambda + \xi > k\lambda \ge 0,\ 1 + h\lambda > h\lambda \ge 0$, hence the integrals in (3.105) with $h > 0,\ k > 0,\ z = -\xi < 0$ are positive and bounded from above by $k^{-1} < \infty$ and $h^{-1} < \infty$ respectively. Moreover, since in this case we have

$$\begin{aligned} I(h) \quad &= \quad A(h) = \int_0^\infty \frac{\lambda^2 \nu_K(d\lambda)}{(h\lambda + 1)^2}, \\ J(k,-\xi) \quad &= \quad B(k,-\xi) = \int_0^\infty \frac{\lambda^2 \nu_R(d\lambda)}{(k\lambda + \xi)^2}, \end{aligned}$$

the determinant of the Jacobian of $\widetilde{F}$ is now (cf. (3.108))

$$1 - A(h)B(k,-\xi),\quad h > 0,\ k > 0,\ \xi > 0.$$

Set in (3.110) $h = h' + i\varepsilon,\ k = k' - i\varepsilon,\ z = -\xi + i\varepsilon$ where $h' > 0,\ k' > 0,\ \xi > 0,\ \varepsilon > 0$ and carry out the limit $\varepsilon \to 0$. We obtain (cf. (3.111))

$$1 - A(h')B(k',-\xi) = C(k',-\xi) = \int_0^\infty \frac{\lambda \nu_R(d\lambda)}{(k\lambda + \xi)^2} > 0.$$

Besides, it follows from (3.11) – (3.12) that (3.107) is valid for $h = 0,\ k = k_1$ and $z = \infty$, where k_1 is the first moment of ν_K. This proves the unique solvability of (3.11) – (3.12) in $\mathbb{C} \setminus \mathbb{R}_+$.

Let us prove assertion (ii) of the lemma. Since $h^{(p)}$ and $k^{(p)}$ are analytic and uniformly in p bounded outside the closed positive semiaxis, there exist subsequences $\{h^{(p_j)}, k^{(p_j)}\}_j$ converging pointwise in $\mathbb{C} \setminus \mathbb{R}_+$ to a certain analytic pair $(\widetilde{h}, \widetilde{k})$. Let us show that $(\widetilde{h}, \widetilde{k}) = (h, k)$. It suffices to consider real negative $z = -\xi > 0$ (see (3.62)). Write for the analog of (3.12) for $\nu_{K^{(p)}}$:

$$\begin{aligned} k^{(p)} &= \int_0^\infty \frac{\lambda \nu_{K^{(p)}}(d\lambda)}{h^{(p)}\lambda + 1} \\ &= \int_0^\infty \frac{\lambda \nu_{K^{(p)}}(d\lambda)}{\widetilde{h}\lambda + 1} + (\widetilde{h} - h^{(p)}) \int_0^\infty \frac{\lambda^2 \nu_{K^{(p)}}(d\lambda)}{(h^{(p)}\lambda + 1)(\widetilde{h}\lambda + 1)}. \end{aligned}$$

Putting here $p = p_j \to \infty$, we see that the l.h.s. converges to $\widetilde{k}$, the first integral on the right converges to the r.h.s of (3.12) with $\widetilde{h}$ instead of h since $\nu_{K^{(p)}}$ converges weakly to ν_K, the integrand is bounded and continuous and the second integral is bounded in p since $h^{(p)}(-\xi) > 0$, $\widetilde{h}(-\xi) > 0$ and the second moment of $\nu_{K^{(p)}}$ is bounded in p according to (3.102), hence, the second term vanishes as $p = p_j \to \infty$. An analogous argument applied to (3.11) show $(\widetilde{h}, \widetilde{k})$ is a solution of (3.11) – (3.12) and then the unique solvability of the system implies that $(\widetilde{h}, \widetilde{k}) = (h, k)$. $\square$

Acknowledgment. We are grateful to Dr. M. Simbirsky for introducing us to fascinating field of machine learning and interesting discussions. We are also grateful to the referee for the careful reading of the manuscript and for suggestions which helped us to improve considerably our presentation.

References

[1] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices.* Springer, New York, 2010.

[2] Y. Bengio, A. Courville and P. Vincent, *Representation learning: A review and new perspectives*, IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013), 1798–1828.

[3] N. Buduma, *Fundamentals of Deep Learning.* O'Reilly, Boston, 2017.

[4] A. L. Caterini and D. E. Chang, *Deep Neural Networks in a Mathematical Framework.* Springer, Heidelberg, 2018.

[5] A. Chakrabarty, S. Chakraborty and R. S. Hazra, *A note on the folklore of free independence*, (2018), http://arxiv.org/abs/1802.00952.

[6] R. Couillet and W. Hachem, *Analysis of the limiting spectral measure of large random matrices of the separable covariance type,* Random Matrices Theory Appl. **3** (2014), 1450016.

[7] R. Giryes, G. Sapiro and A. M. Bronstein, *Deep neural networks with random Gaussian weights: A universal classification strategy?* IEEE Trans. Signal Process **64** (2016), 3444–3457.

[8] F. Gotze, H. Kosters and A Tikhomirov, *Asymptotic spectra of matrix-valued functions of independent random matrices and free probability*, Random Matrices Theory Appl. **4** (2015) 1550005.

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

[10] A. Ling and R.C. Qiu, *Spectrum concentration in deep residual learning: a free probability approach*, IEEE Acess **7** (2019), 105212–105223.

[11] C. H. Martin and M. W. Mahoney, *Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior*, 2017, http://arxiv.org/abs/1710.09533

[12] A. G. de G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. *Gaussian process behaviour in wide deep neural networks*, (2018), http://arxiv.org/abs/1804.1127100952

[13] J. A. Mingo and R. Speicher, *Free Probability and Random Matrices.* Springer, Berlin, 2017.

[14] R. B. Muirhead, *Aspects of Multivariate Statistical Theory.* Wiley, N.Y., 2005.

[15] R. Müller, *On the asymptotic eigenvalue distribution of concatenated vector-valued fading channels.* IEEE Trans. Inf. Theory **48** (2002), 2086–2091.

[16] L. Pastur, *Eigenvalue distribution of random matrices.* In: *Random Media 2000* Proceedings of the Mandralin Summer School, June 2000, Poland, Interdisciplinary Centre of Mathematical and Computational Modeling, Warsaw, 2007, pp.93 – 206

[17] L. Pastur and M. Shcherbina, *Eigenvalue Distribution of Large Random Matrices*, AMS, Providence, 2011.

[18] L. Pastur and V. Slavin, *On random matrices arising in deep neural networks: General i.i.d. case* (in preparation).

[19] D. Petz and F. Hiai, *Semicircle Law, Free Random Variables and Entropy.* AMS, Providence, 2000.

[20] J. Pennington and Y. Bahri, *Geometry of neural network loss surfaces via random matrix theory*, Proc. Mach. Learn. Res. (PMLR) **70** (2017), 2798–2806.

[21] J. Pennington, S. Schoenholz, and S. Ganguli, *The emergence of spectral universality in deep networks.* Proc. Mach. Learn. Res. (PMLR) **84** (2018), 1924–1932.

[22] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos.* In: Advances In Neural Information Processing Systems, 2016, pp. 3360–3368.

[23] A. M. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Y. Ng, *On random weights and unsupervised feature learning.* In: *ICML 2011* Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, June 28 - July 2, 2011, pp. 1089–1096.

[24] S. Scardapane and D. Wang, *Randomness in neural networks: an overview*, WIREs Data Mining Knowl. Discov, 2017, 7:e1200. doi: 10.1002/widm.1200.

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

[26] S. S. Schoenholz, J. Gilmer, S. Ganguli and J. Sohl-Dickstein, *Deep information propagation*, (2016), `http://arxiv.org/abs/1611.01232`.

[27] A. Shrestha and A. Mahmood. *Review of deep learning algorithms and architectures.* IEEE Acess **7** (2019), 53040–53065.

[28] R. Vershynin, *High-Dimensional Probability. An Introduction with Applications in Data Science.* Cambridge University Press, Cambridge, 2018.

Manuscript received December 16 2019
revised April 3 2020

L. Pastur
B. Verkin Institute for Low Temperature Physics and Engineering, 47 Science Avenue, 61103, Kharkiv, Ukraine
E-mail address: `pastur@ilt.kharkov.ua`