On Bayesian Neural Networks with Dependent and Possibly Heavy-Tailed Weights

Nicola Apollonio[*] Giovanni Franzina[†] Giovanni Luca Torrisi[‡]

Abstract

We consider fully connected and feedforward deep neural networks with dependent and possibly heavy-tailed weights, as introduced in [26], to address limitations of the standard Gaussian prior. It has been proved in [26] that, as the number of nodes in the hidden layers grows large, according to a sequential and ordered limit, the law of the output converges weakly to a Gaussian mixture. Among our results, we present sufficient conditions on the model parameters (the activation function and the associated Lévy measures) which ensure that the sequential limit is independent of the order. Next, we study the neural network through the lens of the posterior distribution with a Gaussian likelihood. If the random covariance matrix of the infinite-width limit is positive definite under the prior, we identify the posterior distribution of the output in the wide-width limit according to a sequential regime. Remarkably, we provide mild sufficient conditions to ensure the aforementioned invertibility of the random covariance matrix under the prior, thereby extending the results in [8]. We illustrate our findings using numerical simulations.

MSC 2010 Subject Classification: 68T07; 68W20; 60B10.
Keywords: Deep Neural Networks, Infinite-Width Limit, Gaussian Mixture, Random Matrices.

1 Introduction

A fully connected and feedforward neural network (which we simply refer to as a "neural network" throughout the paper) is the simplest neural network architecture and consists of a sequence of hidden layers stacked between an input layer and an output layer. Each node (or neuron) in a layer is connected to all the nodes in the subsequent layer. These networks are used to estimate unknown functions relating observed inputs to outputs. Once the parameters of the network, i.e. , biases and weights, have been estimated on the basis of a training dataset, network's output is a good approximation of the unknown target function. A neural network is called "deep" if it has more than one hidden layer and "shallow" if it has only one (we refer the reader to [17] for an introduction to neural networks).

The Bayesian approach to the analysis of neural networks allows to include in the model both a prior knowledge on the parameters and the observed data through a prior distribution on neural network's parameters and a likelihood function, respectively. Neal [32, 33] initiated the theoretical study of Bayesian neural networks by proving that, if a Bayesian shallow neural network is initialized with independent Gaussian parameters (i.e. , the prior is Gaussian), then the output of the neural network converges in distribution to a Gaussian process as the number of neurons in the hidden layers increases, i.e. , in the infinite-width limit. This result was extended to Bayesian deep neural networks two decades later (see [19, 25, 30]), and has only recently been made quantitative using optimal transport theory (see [6, 37]), the Stein method for Gaussian approximation (see [3, 4, 14]), and alternative techniques ([7, 11]). Another promising approach to analyze Bayesian neural networks is through the lens of large deviations. First results in this direction are given in [20, 27]. A different perspective is provided by the so-called mean field analysis of neural networks (see [18, 31]).

The emergence of Gaussian processes improved our understanding of how large neural networks work and how to make them more efficient. It also motivated the use of Bayesian regression inference methods, see [25].

[*]Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Roma, Italy. e-mail: nicola.apollonio@cnr.it
[†]Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Roma, Italy. e-mail: giovanni.franzina@cnr.it
[‡]Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Roma, Italy. e-mail: giovanniluca.torrisi@cnr.it

However, as noticed in [32] and [26], the connection with Gaussian processes also highlighted the limitations of Bayesian neural networks with a Gaussian prior. Indeed, there are at least three drawbacks with the choice of independent Gaussian weights: (*i*) Hidden layers do not represent hidden features that capture important aspects of the data; (*ii*) In the infinite-width limit the coordinates of the output become independent and identically distributed Gaussian processes, which is usually undesirable; (*iii*) The assumption of independent Gaussian weights is often unrealistic, as estimated weights of deep neural networks show dependencies and heavy tails (see [16, 29, 38]).

To circumvent these limitations, some authors ([5, 13, 23]) proposed priors which account for independent and identically distributed non-Gaussian weights. However, due to the independence assumption, in the infinite-width limit, the output of the neural network still converges to a stochastic process with independent coordinates. A more structured prior on the weights has been proposed by [1, 26]. Letting $W_{hj}^{(\ell+1)}$ denote the random weight between the node h at the layer $\ell+1$ and the node j at the layer ℓ, in [26] it is assumed that $W_{hj}^{(\ell+1)} := \sqrt{V_{n_\ell,j}^{(\ell)}} N_{hj}^{(\ell)}$, where $V_{n_\ell,j}^{(\ell)}$ are independent random variances, which are identically distributed over j, and $N_{hj}^{(\ell)}$ are independent and Gaussian distributed random variables, with mean 0 and variance $C_W > 0$, independent of the random variances. We refer the reader to Section 2.2 for a rigorous description of the model. The prior proposed in [26] is more general than the Gaussian one, which is retrieved setting $V_{n_\ell,j}^{(\ell)} := 1/n_\ell$, and it accounts for dependent and heavy-tailed weights. Indeed, for fixed ℓ and j, the weights $\{W_{hj}^{(\ell+1)}\}_h$ are stochastically dependent and, if the random variances are distributed according to the square of a heavy-tailed distribution with support on $(0,\infty)$, a simple computation shows that the corresponding weights are also heavy-tailed. It is proved in [26] that if the aggregate random variances at the level of the layer $\ell+1$ converge in distribution to an infinitely divisible law, as the number of nodes in the layer ℓ grows large, then the output of the neural network converges in law to a mixture of Gaussian processes whose coordinates are dependent but still identically distributed. Such a convergence takes place as the number of nodes in the hidden layers grows large, according to a sequential limit with a prescribed order (see Theorem 16 in [26] or Theorem 4.1 for the precise statement). In [26] it is left as challenging open problem to find more natural "limiting schemes" for the validity of such a convergence to a Gaussian mixture. Theorem 4.2 (whose proof is in Appendix) establishes sufficient conditions for the sequential limit of Theorem 16 in [26] to be independent of the order. This generalization is achieved via an alternate representation of the neural network (see Lemma 7.1), which has been used in [27] to study the large deviations of the output in the case of a Gaussian prior.

To the best of our knowledge, progress in the study of posterior Bayesian neural networks refers to models with a fixed variance for the Gaussian prior, see [10, 21, 22, 34, 37]. An exception is the recent work [9], where it is proved that, if the parameters of the Bayesian neural network follow a Gaussian prior and the variance of both the last hidden layer and the Gaussian likelihood function is distributed according to an Inverse-Gamma prior law, then the posterior Bayesian neural network converges to a Student-t process in the infinite-width limit.

In this paper we are concerned with posterior Bayesian neural networks with dependent and possibly heavy-tailed weights $\{W_{hj}^{(\ell+1)}\}$ as described above. If the likelihood is Gaussian, we identify the infinite-width posterior distribution of the output and show that, under a suitable probability measure, it is a Gaussian mixture whose coordinates are dependent and not identically distributed (see Theorem 5.1). This result entirely circumvents one of the limitations arising from the choice of a Gaussian prior. Such a result holds, however, under the assumption that the random covariance matrix of the infinite-width limit of the output is invertible under the prior distribution. Remarkably, we provide sufficient conditions that guarantee such an invertibility assumption, extending the results in [8] to a more general framework. (see Theorems 6.1 and 6.2). The posterior law of the output of the Bayesian neural network at a finite width is very complicated, especially for large values of the depth. One must resort to simulations to gain insight into this distribution. In the final section of the paper, we provide numerical illustrations of our main result.

The paper is organized as follows. Section 2 presents the Bayesian neural network model which is investigated in the article. Section 3 is devoted to preliminaries on matrices, infinitely divisible distributions and mixtures of Gaussian distributions. Section 4 first presents the main result in [26], i.e. , the Gaussian mixture approximation of the output of the prior Bayesian neural network in the infinite-width limit, according to a sequential limit with a prescribed order. It then presents a generalization of this result, providing sufficient conditions to guarantee that the order does not matter in the sequential limit. As the proof of this result is quite technical, we have deferred it to the Appendix. Section 5 studies the infinite-width limit of the posterior distribution of the output of the Bayesian neural network, providing identification of the limit in the case of a Gaussian likelihood. Section 6 addresses the challenging problem of invertibility of the prior random covariance matrix mentioned previously. A couple of numerical illustrations of our main result are provided in Section 7.

2 Neural networks

In this paper, for positive integers p and q, $\mathbb{R}^p$ is the real numerical vector space whose elements are the column vectors with p entries, while $\mathbb{R}^{p\times q}$ is the real vector space of real $p \times q$ matrices. The transposition operator $\top$ acts on $\mathbb{R}^p$ and $\mathbb{R}^{p\times q}$ in the usual way. Hereafter, we also use the notation $X \overset{\mathcal{L}}{=} Y$ to denote that two random elements X and Y have the same law and the notation $\overset{\mathcal{L}}{\to}$ to denote convergence in distribution. Usually in the sequel $\mathfrak{c}$ denotes a positive normalizing constant, which may vary from line to line.

2.1 Neural networks and statistical learning

From a mathematical point of view, fully connected and feedforward neural networks are a parametrized family of functions, recursively defined as follows. Let L, $n_0,\dots,\ n_{L+1} \in \mathbb{N}^*$ be integers. For $\ell = 1,\dots,L$, we set

$$Z_h^{(\ell+1)}(x) := B_h^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{hj}^{(\ell+1)} \sigma(Z_j^{(\ell)}(x)), \quad h = 1,\dots,n_{\ell+1}$$

$$Z_h^{(1)}(x) := B_h^{(1)} + \sum_{j=1}^{n_0} W_{hj}^{(1)} x_j, \quad h = 1,\dots,n_1$$

where $x \in \mathbb{R}^{n_0}$ is the input, $\sigma : \mathbb{R} \to \mathbb{R}$ (a measurable function) is the activation function, $\{B_h^{(\ell)}\}_{\ell=1,\dots,L+1;h=1,\dots,n_\ell}$ and $\{W_{hj}^{(\ell)}\}_{\ell=1,\dots,L+1;h=1,\dots,n_\ell;j=1,\dots,n_{\ell-1}}$ are real parameters called biases and weights, respectively. The network consists of L hidden layers stacked between the input layer (the layer 0) and the output layer (the layer $L+1$). On each hidden layer $\ell \in \{1,\dots,L\}$, there are n_ℓ nodes. The network is called deep if $L \geq 2$ and shallow if $L = 1$. Throughout the paper, we often apply σ to vectors, meaning that $\sigma(x) = (\sigma(x_1),\dots,\sigma(x_p))^\top$ for $x = (x_1,\dots,x_p)^\top$.

In statistical learning, neural networks are used to estimate unknown target functions. More precisely, for a fixed network architecture $(L,\ n_0,\dots,n_{L+1},\ \sigma)$, for a given unknown target function $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$ and a training dataset $\mathcal{D} := \{(x(i), y(i))\}_{i=1,\dots,d} \subset \mathbb{R}^{n_0} \times \mathbb{R}^{n_{L+1}}$, $d \in \mathbb{N}^*$, (i.e., couples of data $x(i)$ and outcomes $y(i) := f(x(i))$) the objective is to produce an estimate of the parameter $\Theta = (\{B_h^{(\ell)}\}, \{W_{hj}^{(\ell)}\})$, say $\widehat{\Theta}$, in such a way that the output of the neural network is a good estimate of f, i.e.,

$$(Z_1^{(L+1)}(x),\dots,Z_{n_{L+1}}^{(L+1)}(x))^\top\Big|_{\Theta=\widehat{\Theta}} \approx y, \quad \forall\ (x,y) \in \mathcal{D} \cup \mathcal{T}$$

where

$$\mathcal{T} := \{(x'(j), y'(j))\}_{j=1,\dots,t} \subset \mathbb{R}^{n_0} \times \mathbb{R}^{n_{L+1}},\ t \in \mathbb{N}^*$$

is a test set. Throughout the paper we encode the inputs (data) in the $n_0 \times d$ matrix $\mathbf{x} := (x(1)\dots x(d))$ and the outcomes (responses) in the $n_{L+1} \times d$ matrix $\mathbf{y} := (y(1)\dots y(d))$.

2.2 Neural networks with dependent and possibly heavy-tailed weights, and Bayesian statistical learning

From now on, all the random quantities are defined on a measurable space $(\Omega, \mathcal{F})$, on which different probability laws will be defined. Hereon $\mathcal{N}_1(\mu, v)$ denotes the one-dimensional Gaussian law with mean μ and variance v.

The prior knowledge on the parameter Θ is modeled via a prior probability measure $\mathbb{P}_{\text{prior}}$ on $(\Omega, \mathcal{F})$. In particular, throughout the paper we assume that, under $\mathbb{P}_{\text{prior}}$,

- For $\ell = 1, \dots, L+1$ and $h = 1, \dots, n_\ell$, $B_h^{(\ell)}$ are random variables with $B_h^{(\ell)} \sim \mathcal{N}_1(0, C_B)$, for a constant $C_B \geq 0$;

- For $\ell = 1, \dots, L+1$, $h = 1, \dots, n_\ell$ and $j = 1, \dots, n_{\ell-1}$, $W_{hj}^{(\ell)}$ are random variables defined by

$$W_{hj}^{(\ell)} := \sqrt{V_{n_{\ell-1},j}^{(\ell-1)}} N_{hj}^{(\ell)},$$

  where: for $j = 1, \dots, n_0$, $V_{n_0,j}^{(0)} := n_0^{-1}$, for $\ell = 2, \dots, L+1$, $\{V_{n_{\ell-1},j}^{(\ell-1)}\}_{j=1,\dots,n_{\ell-1}}$ are non-negative and identically distributed (over j) random variables, and for $\ell = 1, \dots, L+1$, $h = 1, \dots, n_\ell$ and $j = 1, \dots, n_{\ell-1}$, $N_{hj}^{(\ell)}$ are random variables with $N_{hj}^{(\ell)} \sim \mathcal{N}_1(0, C_W)$, for a constant $C_W > 0$;

- All the random variables $\{B_h^{(\ell)}, V_{n_{\ell-1},j}^{(\ell-1)}, N_{hj}^{(\ell)}\}$ are stochastically independent.

Under these distributional assumptions, one speaks of neural network with dependent and possibly heavy-tailed weights, see the seminal paper [26], where this prior has been introduced. Indeed, note that, for a fixed $\ell \in \{2, \dots, L+1\}$ and $j \in \{1, \dots, n_{\ell-1}\}$, the random weights $W_{1j}^{(\ell)}, \dots, W_{n_\ell j}^{(\ell)}$ are dependent, and that, if $V_{n_{\ell-1},j}^{(\ell-1)}$ is distributed as the square of a random variable with a heavy tail law with support on $(0, \infty)$, then the distribution of $W_{hj}^{(\ell)}$ is heavy tail. Note also that if $V_{n_{\ell-1},j}^{(\ell-1)} := \frac{1}{n_{\ell-1}}$, then we recover the usual Gaussian prior.

The Bayesian appraoch to the statistical learning problem allows to incorporate in the model the training dataset $\mathcal{D}$ through a likelihood function $\mathcal{L}(\mathcal{D}, \Theta)$. Then the posterior knowledge on the parameter is summarized by the posterior probability measure

$$\mathrm{d}\mathbb{P}_{\text{posterior}} := \frac{\mathcal{L}(\mathcal{D}, \Theta)\mathrm{d}\mathbb{P}_{\text{prior}}}{\mathbb{E}_{\text{prior}}\mathcal{L}(\mathcal{D}, \Theta)},$$

where $\mathbb{E}_{\text{prior}}$ denotes the expectation under the prior probability measure and it is assumed $\mathbb{E}_{\text{prior}}\mathcal{L}(\mathcal{D}, \Theta) > 0$. If the law of Θ under the prior has density $\mathrm{p}_{prior}(\cdot)$, then the law of Θ under the posterior has density

$$\mathrm{p}_{\text{posterior}}(\theta) \propto \mathcal{L}(\mathcal{D}, \theta)\mathrm{p}_{\text{prior}}(\theta),$$

and one estimates the parameter Θ by maximizing $\mathrm{p}_{\text{posterior}}(\cdot)$ via an adequate variant of the Stochastic Gradient Descent (maximum a posteriori estimate), see e.g. [16].

3 Preliminaries

In this section we provide some preliminaries on matrices and infinitely divisible distributions, and we give the formal definition of Gaussian mixture distribution. As general references for the first two topics, we cite the monographs [24, 28] and [36], respectively.

3.1 Matrices

For positive integers p and q, both $\mathbb{R}^{p\times q}$ and $\mathbb{R}^{pq}$ are Euclidean spaces when equipped with the Frobenius inner product $\langle\cdot,\cdot\rangle_F$ and the standard dot product $\langle\cdot,\cdot\rangle$ defined, respectively by $\langle \mathbf{A}, \mathbf{B}\rangle_F = \mathrm{Tr}(\mathbf{A}^\top\mathbf{B})$, with $\mathrm{Tr}(\cdot)$ the trace operator, and $\langle a, b\rangle = a^\top b$. We denote by $\|\cdot\|_F$ and $\|\cdot\|$ the Frobenius norm and the standard Euclidean norm, induced by $\langle\cdot,\cdot\rangle_F$ and $\langle\cdot,\cdot\rangle$, respectively. For $\mathbf{A} \in \mathbb{R}^{p\times q}$, let $\mathrm{vec}(\mathbf{A})$ be the column vector in $\mathbb{R}^{pq}$ obtained by stacking the columns of $\mathbf{A}$ on top of each other, starting from the leftmost column. Since for $\mathbf{A} \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{q\times p}$, it holds that

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{vec}(\mathbf{A}^\top)^\top \mathrm{vec}(\mathbf{B}), \tag{1}$$

it follows that $\mathrm{vec} : \mathbb{R}^{p\times q} \to \mathbb{R}^{pq}$ is an isometry between the Euclidean spaces $(\mathbb{R}^{p\times q}, \langle\cdot,\cdot\rangle_F)$ and $(\mathbb{R}^{pq}, \langle\cdot,\cdot\rangle)$. In view of this bijection, we think of a random matrix $\mathbf{U}$ as a matrix-valued random variable whose law is induced by the random vector vec($\mathbf{U}$) (via vec^{-1}). Accordingly, the characteristic function of the random matrix $\mathbf{U} \in \mathbb{R}^{p\times q}$ is given by

$$\phi(\boldsymbol{\theta};\mathbf{U}) = \mathbb{E}\left(e^{\mathbf{i}\langle\boldsymbol{\theta},\mathbf{U}\rangle_F}\right) = \mathbb{E}\left(e^{\mathbf{i}\text{Tr}(\boldsymbol{\theta}^\top\mathbf{U})}\right)$$

where $\mathbf{i} := \sqrt{-1}$ and $\boldsymbol{\theta}$ ranges in $\mathbb{R}^{p\times q}$. The *Kronecker* (also known as *tensor*) product of matrices $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$ and $\mathbf{B} \in \mathbb{R}^{r\times s}$, is the $pr \times qs$ real matrix $\mathbf{A}\otimes\mathbf{B} = (a_{ij}\mathbf{B})_{1\le i\le p,\, 1\le j\le q}$. If $\mathbf{A}$ and $\mathbf{B}$ are invertible matrices, then $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$. Moreover, if $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are matrices such that $\mathbf{ABC}$ exists, then $\text{vec}(\mathbf{ABC}) = (\mathbf{C}^\top\otimes\mathbf{A})\text{vec}(\mathbf{B})$. Therefore, with $\mathbf{Id}$ the identity matrix of appropriate dimension, choosing $\mathbf{C} = \mathbf{Id}$ and $\mathbf{A}$ invertible, identity (1) and the preceding properties yield the useful relations

$$\text{vec}(\mathbf{B})^\top(\mathbf{Id}\otimes\mathbf{A})\text{vec}(\mathbf{B}) = \text{Tr}(\mathbf{B}^\top\mathbf{AB}) \text{ and } \text{vec}(\mathbf{B})^\top(\mathbf{Id}\otimes\mathbf{A})^{-1}\text{vec}(\mathbf{B}) = \text{Tr}(\mathbf{B}^\top\mathbf{A^{-1}B}). \tag{2}$$

For later purposes, we recall that, for every symmetric positive semi-definite matrix $\mathbf{A} = (a_{rs})_{r,s=1,\dots,d}$, there exists a unique symmetric positive semi-definite matrix $\mathbf{A}^\sharp = (a^\sharp_{rs})_{r,s=1,\dots,d}$ (the square-root of $\mathbf{A}$) such that $\mathbf{A}^\sharp\mathbf{A}^\sharp = \mathbf{A}$, i.e., $a_{rs} = \sum_{j=1}^d a^\sharp_{rj}a^\sharp_{js}$, $r,s = 1,\dots,d$.

For $\mathbf{A} \in \mathbb{R}^{p\times q}$, we denote by $rk(\mathbf{A})$ the rank of $\mathbf{A}$, i.e., the number of linearly independent columns or rows within $\mathbf{A}$. We denote by 0 the null vector of $\mathbb{R}^p$, by $\text{diag}_r(a_1,\dots,a_r)$, $r \in \mathbb{N}^*$, a diagonal $r\times r$ real matrix with diagonal elements $a_1,\dots,a_r$, by $\mathbf{Id}_p$ the $p\times p$ identity matrix, and we set $1_p := (1,\dots,1)^\top \in \mathbb{R}^p$, $p \in \mathbb{N}^*$.

The following elementary lemma can be easily proved.

Lemma 3.1. *Let $a_1,\dots,a_n \in \mathbb{R}^p$, with $n \ge p$, be p-dimensional (column) vectors and let $c_1,\dots,c_n$ be arbitrarily chosen positive numbers. Then the $p\times p$ matrix $\sum_{i=1}^n c_i a_i a_i^\top$ is positive definite if and only if the $p\times n$ matrix $(a_1,\dots,a_n)$ has rank equal to p.*

3.2 Infinitely divisible distributions

A real-valued random variable X is said to have an infinitely divisible law if, for each $n \in \mathbb{N}^*$, there exist independent and identically distributed $\mathbb{R}$-valued random variables $X_{n1},\dots,X_{nn}$ such that $X \overset{\mathcal{L}}{=} X_{n1} + \ldots + X_{nn}$.

If X is a non-negative random variable it turns out that X has an infinitely divisible distribution if and only if there exists a couple (a,ρ), where $a \ge 0$ is a non-negative constant and ρ is a Lévy measure on $(0,\infty)$, i.e., a Borel measure on $(0,\infty)$ with

$$\int_{(0,\infty)} \min\{1,x\}\rho(\mathrm{d}x) < \infty,$$

such that X has characteristic function of the form $\mathrm{e}^{\Psi(u)}$ with

$$\Psi(u) = \mathbf{i}ua + \int_{(0,\infty)} (\mathrm{e}^{\mathbf{i}ux} - 1)\rho(\mathrm{d}x).$$

We write $X = \text{ID}(a,\rho)$.

3.3 Gaussian mixtures

We denote by $\mathcal{N}_m(c,\mathbf{C})$ the m-dimensional Gaussian distribution with mean $c \in \mathbb{R}^m$ and covariance matrix $\mathbf{C}$. Let κ be a random vector with values on $\mathbb{R}^m$ and let $\mathbf{K}$ be a positive semi-definite and symmetric random matrix with values on $\mathbb{R}^{m\times m}$. We say that a random vector X, defined on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ and with values on $\mathbb{R}^m$, has the Gaussian mixture distribution (under $\mathbb{P}$) with random mean κ and random covariance matrix $\mathbf{K}$, denoted by $\mathcal{GM}^{\mathbb{P}}_m(\kappa,\mathbf{K})$, if $X \mid (\kappa,\mathbf{K}) \sim \mathcal{N}_m(\kappa,\mathbf{K})$. Note that if $\mathbb{P}(\det\mathbf{K} > 0) = 1$, then X has density (with respect to the Lebesgue measure) $\mathbb{E}[\varphi_{(\kappa,\mathbf{K})}(\xi)]$, $\xi \in \mathbb{R}^m$, being $\varphi_{(c,\mathbf{C})}$ the density of $\mathcal{N}_m(c,\mathbf{C})$. In the definition of Gaussian mixture we explicited the dependence on the probability measure $\mathbb{P}$ (with an upper index) since in this paper we will work with Gaussian mixtures under different probability measures.

3.4 Sequential limits

For a positive integer ℓ let $[\ell] := \{1, \ldots, \ell\}$. We think of $[\ell]$ either as a set of indices or as an interval of $\mathbb{N}$, linearly ordered by the natural order. Let $f : \mathbb{N}^\ell \to \mathbb{R}$ be a function, $S := \{i_1, \ldots, i_r\}$, $i_1 < i_2 \cdots < i_r$ be a linearly ordered subset of $[\ell]$ and $\beta : S \to [\ell]$ be an injection. By $\beta\overrightarrow{\lim} f$ we mean

$$\lim_{n_{\beta(i_r)}\to\infty} \lim_{n_{\beta(i_{r-1})}\to\infty} \cdots \lim_{n_{\beta(i_1)}\to\infty} f,$$

and we say that the limit is sequential along β. Note that if $S = [\ell]$, then β is a permutation π of $[\ell]$. In particular, if π is the identity, then we write $\overrightarrow{\lim}$ for $\pi\overrightarrow{\lim}$. If $\pi\overrightarrow{\lim} f = \overline{f}$, then we also write $f \overset{\pi}{\hookrightarrow} \overline{f}$. If β is the injection obtained by restricting π over S, then we write $\overline{f}_\beta$ for the sequential limit of f along β. By $(\pi \preceq j)$ we denote the injection obtained by restricting π on $[j]$. Analogously, $(\pi \succeq j)$ denotes the injection obtained by restricting π on $\{j, j+1, \ldots, \ell\}$. If $j = 0$, then $(\pi \preceq 0)$ is the unique permutation of the empty set and we set $(\pi \preceq 0)\overrightarrow{\lim} f = f$. Clearly, if $j = \ell$, then $(\pi \preceq j) = \pi$ while $(\pi \succeq j)$ is the unique permutation of $\{\pi(\ell)\}$. Moreover, if $j \neq \ell$, then

$$\pi\overrightarrow{\lim} f = (\pi \succeq j+1)\overrightarrow{\lim}\left[(\pi \preceq j)\overrightarrow{\lim} f\right]. \tag{3}$$

The preceding definitions can be adopted other notions of limits, for instance, if f a random object, we write $f \overset{\beta\mathcal{L}}{\hookrightarrow} \overline{f}_\beta$ to mean that f converges in law to $\overline{f}_\beta$ sequentially along β, and we write $f \overset{\beta\mathbb{P}}{\hookrightarrow} \overline{f}_\beta$ to mean that f converges $\mathbb{P}$-a.s. to $\overline{f}_\beta$ sequentially along β. When β is the identity permutation on $[\ell]$, we simply omit it in the corresponding symbol.

4 The infinite-width limit under the prior

We start introducing some notation. For $\ell = 1, \ldots, L+1$ and $i = 1, \ldots, d$, we set

$$Z^{(\ell)}(x(i)) := (Z_1^{(\ell)}(x(i)), \ldots, Z_{n_\ell}^{(\ell)}(x(i)))^\top,$$

and we consider the $n_\ell \times d$ random matrix

$$\mathbf{Z}^{(\ell)}(\mathbf{x}) := (Z^{(\ell)}(x(1)) \ldots Z^{(\ell)}(x(d))).$$

For later purposes, we vectorialize the $n_\ell \times d$ random matrix $\mathbf{Z}^{(\ell)}(\mathbf{x})$ defining the $n_\ell d$-dimensional random vector

$$Z^{(\ell)}(\mathbf{x}) := \mathrm{vec}((\mathbf{Z}^{(\ell)}(\mathbf{x}))^\top), \quad \ell = 1, \ldots, L+1.$$

The following Theorem 4.1 is the main result in [26] (see Theorem 16 therein). It extends to the case of dependent and possibly heavy-tailed weights the Gaussian behavior in the infinite-width limit of a neural network with a Gaussian prior (see the seminal paper [33] and the more recent contributions by [19, 25, 30]).

Theorem 4.1. *Assume that:*
(i) The activation function σ is continuous and such that

$$\forall\ z \in \mathbb{R},\ |\sigma(z)| \leq a_1 + a_2|z|^{a_3} \text{ for some positive constants } a_1, a_2, a_3 > 0.$$

(ii) $\forall\ \ell = 1, \ldots, L$, $\sum_{j=1}^{n_\ell} V_{n_\ell,j}^{(\ell)} \to \mathrm{ID}(a^{(\ell)}, \rho^{(\ell)})$ in distribution as $n_\ell \to \infty$, for some non-negative constant $a^{(\ell)}$ and Lévy measure $\rho^{(\ell)}(\cdot)$ on $(0, \infty)$.

Then, under $\mathbb{P}_{\text{prior}}$, we have

$$\overrightarrow{\lim} Z^{(L+1)}(\mathbf{x}) = G^{(L+1)}(\mathbf{x}) \sim \mathcal{GM}_{n_{L+1}d}^{\mathbb{P}_{\text{prior}}}(0, \mathbf{Id}_{n_{L+1}} \otimes \mathbf{K}^{(L+1)}(\mathbf{x})) \text{ in distribution.} \tag{4}$$

Here, $\mathbf{K}^{(L+1)}(\mathbf{x})$ is defined by the following stochastic recursion:

$$\mathbf{K}^{(1)}(\mathbf{x}) := (K^{(1)}(x(i), x(i')))_{1 \leq i, i' \leq d}, \tag{5}$$

where

$$K^{(1)}(x(i), x(i')) := C_B + C_W \frac{x(i)^\top x(i')}{n_0},$$

and

$$\mathbf{K}^{(\ell)}(\mathbf{x}) := C_B 1_d 1_d^\top + C_W \Bigg(a^{(\ell-1)} \mathbb{E}_{\text{prior}}[\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))^\top \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x})] + \sum_{j=1}^{N_{\ell-1}((0,\infty))} T_j^{(\ell-1)} \sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))^\top \Bigg), \tag{6}$$

for $\ell = 2, \ldots, L+1$.

Here, $\{\zeta_j^{(1)}(\mathbf{x})\}_{j\geq 1}$ *is a sequence of independent and identically distributed random vectors with* $\zeta_1^{(1)}(\mathbf{x}) \sim \mathcal{N}_d(0, \mathbf{K}^{(1)}(\mathbf{x}))$*, for* $\ell = 3, \ldots, L+1$*, given* $\mathbf{K}^{(\ell-1)}(\mathbf{x})$*,* $\{\zeta_j^{(\ell-1)}(\mathbf{x})\}_{j\geq 1}$ *is a sequence of independent and identically distributed random vectors with* $\zeta_1^{(\ell-1)}(\mathbf{x}) \sim \mathcal{N}_d(0, \mathbf{K}^{(\ell-1)}(\mathbf{x}))$*, and, for* $\ell = 2, \ldots, L+1$*,* $\{T_j^{(\ell-1)}\}_{j=1,\ldots,N_{\ell-1}((0,\infty))}$ *are the points of a Poisson process on* $(0, \infty)$ *with mean measure* $\rho^{(\ell-1)}$*, independent of the sequence* $\{\zeta_j^{(\ell-1)}(\mathbf{x})\}_{j\geq 1}$.

Let us represent $G^{(L+1)}(\mathbf{x})$ (the limiting random vector in (4)) as

$$G^{(L+1)}(\mathbf{x}) := \text{vec}((\mathbf{G}^{(L+1)}(\mathbf{x}))^\top), \tag{7}$$

where $\mathbf{G}^{(L+1)}(\mathbf{x})$ is the $n_{L+1} \times d$ random matrix

$$\mathbf{G}^{(L+1)}(\mathbf{x}) := (G^{(L+1)}(x(1)) \ldots G^{(L+1)}(x(d))),$$

and

$$G^{(L+1)}(x(i)) := (G_1^{(L+1)}(x(i)), \ldots, G_{n_{L+1}}^{(L+1)}(x(i)))^\top, \quad i = 1, \ldots, d.$$

Note that, under $\mathbb{P}_{\text{prior}}$, the rows of $\mathbf{G}^{(L+1)}(\mathbf{x})$ are pairwise dependent and uncorrelated, and identically distributed according to the Gaussian mixture law with mean 0 and random covariance matrix $\mathbf{K}^{(L+1)}(\mathbf{x})$. In Section 5 we will see that if we consider the posterior distribution with a Gaussian likelihood, then, under a suitable probability measure, the infinite-width limit of the posterior output has dependent and, in general, not identically distributed coordinates (see Theorem 5.1).

The limit in (4) not only is sequential, but it has also a prescribed order. Theorem 4.2 provides sufficient conditions which guarantee that in the sequential limit the order doesn't matter. Its proof is quite technical and postponed in Appendix.

Theorem 4.2. *Assume* $L \geq 2$ *and the condition (ii) of Theorem 4.1. If either*

$$\sigma \text{ is continuous and bounded} \tag{8}$$

or

$$\sigma \text{ satisfies the condition } (i) \text{ of Theorem 4.1 and } \rho^{(\ell)}((0,\infty)) < \infty \;\forall\; \ell = 2, \ldots, L, \tag{9}$$

then, for any fixed permutation π *of* $1, \ldots, L$*, under* $\mathbb{P}_{\text{prior}}$*,*

$$\pi \overrightarrow{\lim} Z^{(L+1)}(\mathbf{x}) = G^{(L+1)}(\mathbf{x}) \sim \mathcal{GM}_{n_{L+1}d}^{\mathbb{P}_{\text{prior}}}(0, \mathbf{Id}_{n_{L+1}} \otimes \mathbf{K}^{(L+1)}(\mathbf{x})) \text{ in distribution.}$$

Remark 4.3. *Note that, although Theorem 4.2 generalizes (under suitable assumptions) the claim (4), it does not imply that the weak convergence to a Gaussian mixture holds as* $\min\{n_1, \ldots, n_L\} \to \infty$.

5 The infinite-width limit under the posterior

Theorem 4.1 provides the distribution of the output of the neural network in the infinite-width limit, under the prior probability measure $\mathbb{P}_{\text{prior}}$ defined in Section 2.2. Here we give the distribution of the output of the neural network in the infinite-width limit, under the posterior probability measure $\mathbb{P}_{\text{posterior}}$ defined by

$$\mathrm{d}\mathbb{P}_{\text{posterior}} \propto \exp\left(-\sum_{i=1}^{d} \|Z^{(L+1)}(x(i)) - y(i)\|^2\right) \mathrm{d}\mathbb{P}_{\text{prior}}. \tag{10}$$

Hereon, (when the inverse matrices exist) we will consider the $n_{L+1}d \times n_{L+1}d$ random matrix

$$\mathbf{\Lambda}^{(L+1)}(\mathbf{x}) := \mathbf{Id}_{n_{L+1}} \otimes \mathbf{L}^{(L+1)}(\mathbf{x})^{-1},$$

where $\mathbf{L}^{(L+1)}(\mathbf{x})$ is the $d \times d$ random matrix

$$\mathbf{L}^{(L+1)}(\mathbf{x}) := 2\mathbf{Id}_d + \mathbf{K}^{(L+1)}(\mathbf{x})^{-1}.$$

We will also consider the $n_{L+1}d$-dimensional random vector

$$\lambda^{(L+1)}(\mathbf{x},\mathbf{y}) := \mathrm{vec}(\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y})^\top),$$

where $\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y})$ is the $n_{L+1} \times d$ random matrix

$$\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y}) := 2\mathbf{y}\mathbf{L}^{(L+1)}(\mathbf{x})^{-1}.$$

Hereafter, for a p-dimensional random element X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and with values on some metric space, we denote by $\mathbb{P}_X$ (or by $(\mathbb{P})_X$) the probability law induced by X on the metric space.

The following theorem holds.

Theorem 5.1. *Assume conditions* (*i*) *and* (*ii*) *of Theorem 4.1 and*

$$\mathbb{P}_{\text{prior}}(\det(\mathbf{K}^{(L+1)}(\mathbf{x})) > 0) = 1. \tag{11}$$

Then

$$\overrightarrow{\lim}(\mathbb{P}_{\text{posterior}})_{Z^{(L+1)}(\mathbf{x})} = \mathcal{GM}^{\mathbb{S}}_{n_{L+1}d}(\lambda^{(L+1)}(\mathbf{x},\mathbf{y}), \mathbf{\Lambda}^{(L+1)}(\mathbf{x})) \quad \textit{weakly.} \tag{12}$$

Here $\mathbb{S}$ denotes the probability measure

$$\mathrm{d}\mathbb{S} \propto \frac{\exp\left(\mathrm{Tr}(\mathbf{y}(\mathbf{Id}_d + (2\mathbf{K}^{(L+1)}(\mathbf{x}))^{-1})^{-1})\mathbf{y}^\top\right)}{(\det(\mathbf{Id}_d + 2\mathbf{K}^{(L+1)}(\mathbf{x})))^{n_{L+1}/2}} \mathrm{d}\mathbb{P}_{\text{prior}},$$

and $\mathcal{GM}^{\mathbb{S}}_{n_{L+1}d}(\lambda^{(L+1)}(\mathbf{x},\mathbf{y}), \mathbf{\Lambda}^{(L+1)}(\mathbf{x}))$ is a Gaussian mixture under $\mathbb{S}$, i.e., it has density

$$\mathbb{E}_{\mathbb{S}}[\varphi_{(\lambda^{(L+1)}(\mathbf{x},\mathbf{y}), \mathbf{\Lambda}^{(L+1)}(\mathbf{x}))}(\xi)], \quad \xi \in \mathbb{R}^{n_{L+1}d}$$

where $\mathbb{E}_{\mathbb{S}}$ denotes the mean with respect to $\mathbb{S}$.

As clarified by Lemma 5.5, the assumption (11), which will be explored in the next section, is crucial only to identify the limiting law.

Remark 5.2. *Let $G^{(L+1)}(\mathbf{x})$ be the vector defined in (7) and suppose that it is distributed according to the Gaussian mixture law in (12). Then, under the probability measure $\mathbb{S}$, the rows of $\mathbf{G}^{(L+1)}(\mathbf{x})$ are dependent and, in general, not identical distributed. Indeed, under $\mathbb{S}$, the i-th row of $\mathbf{G}^{(L+1)}(\mathbf{x})$ is distributed according to a Gaussian mixture with random covariance $\mathbf{L}^{(L+1)}(\mathbf{x})^{-1}$ and random mean given by the i-th row of $2\mathbf{y}\mathbf{L}^{(L+1)}(\mathbf{x})^{-1}$; note that the rows of this latter matrix are all equal if and only if the rows of $\mathbf{y}$ are such.*

Remark 5.3. *Note that*

$$\mathbb{S}_{\mathbf{K}^{(L+1)}(\mathbf{x})}(\mathrm{d}\mathbf{k}) \propto \frac{\exp\left(\mathrm{Tr}(\mathbf{y}(\mathbf{Id}_d+(2\mathbf{k})^{-1})^{-1})\mathbf{y}^\top\right)}{(\det(\mathbf{Id}_d+2\mathbf{k}))^{n_{L+1}/2}}(\mathbb{P}_{\mathrm{prior}})_{\mathbf{K}^{(L+1)}(\mathbf{x})}(\mathrm{d}\mathbf{k}),$$

is a probability measure on the space of positive definite $d\times d$ symmetric matrices which can be interpreted as a posterior with likelihood

$$\frac{\exp\left(\mathrm{Tr}(\mathbf{y}(\mathbf{Id}_d+(2\mathbf{k})^{-1})^{-1})\mathbf{y}^\top\right)}{(\det(\mathbf{Id}_d+2\mathbf{k}))^{n_{L+1}/2}}.$$

Finally, we remark that if $\mathbf{K}^{(L+1)}(\mathbf{x})$ is deterministic (as it happens when the Lévy measures are all equal to zero), then $\mathbb{S}=\mathbb{P}_{\mathrm{prior}}$.

Remark 5.4. *A simple modification of the proof of Theorem 5.1 shows that if $L\geq 2$, and we replace the condition (i) of Theorem 4.1 with either the assumption (8) or the assumption (9), then, for any permutation π of $[L]$, by Theorem 4.2 it follows that in (12) we can replace $\overrightarrow{\lim}$ with $\pi\overrightarrow{\lim}$.*

The proof of Theorem 5.1 exploits the following Lemma 5.5 which, for the sake of completeness, will be proved at the end of this section.

Let $\{X_n\}_{n\geq 1}$ and X be random variables defined of the probability space $(\Omega,\mathcal{F},\mathbb{P})$ and with values on some metric space. Consider on $(\Omega,\mathcal{F})$ the tilted probability measures:

$$\mathrm{d}\mathbb{Q}_n\propto g(X_n)\mathrm{d}\mathbb{P}\quad\text{and}\quad \mathrm{d}\mathbb{Q}\propto g(X)\mathrm{d}\mathbb{P}$$

where g is a bounded continuous function such that the normalizing constants are different from zero.

The following lemma holds.

Lemma 5.5. *Suppose that $\mathbb{P}_{X_n}\to\mathbb{P}_X$ weakly, as $n\to\infty$. Then $(\mathbb{Q}_n)_{X_n}\to\mathbb{Q}_X$ weakly, as $n\to\infty$.*

Proof. (Theorem 5.1) Thanks to Theorem 4.1 and Lemma 5.5, we only need to compute the law $\mathbb{Q}_{G^{(L+1)}(\mathbf{x})}$, where

$$\mathrm{d}\mathbb{Q}\propto \exp\left(-\sum_{i=1}^d \|G^{(L+1)}(x(i))-y(i)\|^2\right)\mathrm{d}(\mathbb{P}_{\mathrm{prior}})_{G^{(L+1)}(\mathbf{x})}$$

and

$$(\mathbb{P}_{\mathrm{prior}})_{G^{(L+1)}(\mathbf{x})}=\mathcal{GM}^{\mathbb{P}_{\mathrm{prior}}}_{n_{L+1}d}(0,\mathbf{Id}_{n_{L+1}}\otimes\mathbf{K}^{(L+1)}(\mathbf{x})).$$

Note that $\sum_{i=1}^d\|G^{(L+1)}(x(i))-y(i)\|^2=\|\mathbf{G}^{(L+1)}(\mathbf{x})-\mathbf{y}\|_F^2$. Let $\xi:=\mathrm{vec}(\boldsymbol{\xi}^\top)$, where $\boldsymbol{\xi}:=[\xi(1)\dots\xi(d)]$ and $\xi(i)\in\mathbb{R}^{n_{L+1}}$. By identity (2) ad the definition of Frobenius norm, we have

$$\begin{aligned}
\mathrm{d}\mathbb{Q}_{G^{(L+1)}(\mathbf{x})}(\xi) &= \mathfrak{c}\exp\left(-\sum_{i=1}^d\|\xi(i)-y(i)\|^2\right)\\
&\quad\times\mathbb{E}_{\mathrm{prior}}\left[(\det(\mathbf{Id}_{n_{L+1}}\otimes\mathbf{K}^{(L+1)}(\mathbf{x})))^{-1/2}\exp\left(-\frac{1}{2}\xi^\top(\mathbf{Id}_{n_{L+1}}\otimes\mathbf{K}^{(L+1)}(\mathbf{x}))^{-1}\xi\right)\right]\mathrm{d}\xi\\
&=\mathfrak{c}\,\mathbb{E}_{\mathrm{prior}}\left[\left(\frac{1}{\det\mathbf{K}^{(L+1)}(\mathbf{x})}\right)^{\frac{n_{L+1}}{2}}\exp\left(-\mathrm{Tr}[(\boldsymbol{\xi}-\mathbf{y})(\boldsymbol{\xi}-\mathbf{y})^\top]-\frac{1}{2}\mathrm{Tr}[\boldsymbol{\xi}(\mathbf{K}^{(L+1)}(\mathbf{x}))^{-1}\boldsymbol{\xi}^\top]\right)\right]\mathrm{d}\xi\\
&=\mathfrak{c}\,\mathbb{E}_{\mathrm{prior}}\left[\left(\frac{1}{\det\mathbf{K}^{(L+1)}(\mathbf{x})}\right)^{\frac{n_{L+1}}{2}}\exp\left(-\frac{1}{2}\mathrm{Tr}[2(\boldsymbol{\xi}-\mathbf{y})(\boldsymbol{\xi}-\mathbf{y})^\top-\boldsymbol{\xi}(\mathbf{K}^{(L+1)}(\mathbf{x}))^{-1}\boldsymbol{\xi}^\top]\right)\right]\mathrm{d}\xi. \qquad (13)
\end{aligned}$$

Next, for ease of notation set $\lambda=\lambda^{(L+1)}(\mathbf{x},\mathbf{y})$, $\boldsymbol{\lambda}=\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y})$, $\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x})$, $\mathbf{L}=\mathbf{L}^{(L+1)}(\mathbf{x})$ and $\mathbf{K}=\mathbf{K}^{(L+1)}(\mathbf{x})$ and observe that by identity (2) we have

$$\begin{aligned}
(\xi-\lambda)^\top\boldsymbol{\Lambda}^{-1}(\xi-\lambda)&=(\mathrm{vec}(\boldsymbol{\xi}^\top)-\mathrm{vec}(\boldsymbol{\lambda}^\top))^\top(\mathbf{Id}_{n_{L+1}}\otimes\mathbf{L})(\mathrm{vec}(\boldsymbol{\xi}^\top)-\mathrm{vec}(\boldsymbol{\lambda}^\top))\\
&=\mathrm{Tr}((\boldsymbol{\xi}-\boldsymbol{\lambda})\mathbf{L}(\boldsymbol{\xi}-\boldsymbol{\lambda})^\top). \qquad (14)
\end{aligned}$$

On the other hand

$$\mathrm{Tr}((\boldsymbol{\xi}-\boldsymbol{\lambda})\mathbf{L}(\boldsymbol{\xi}-\boldsymbol{\lambda})^\top) = \mathrm{Tr}(\boldsymbol{\xi}\mathbf{L}\boldsymbol{\xi}^\top) - 2\mathrm{Tr}(\boldsymbol{\lambda}\mathbf{L}\boldsymbol{\xi}^\top) + \mathrm{Tr}(\boldsymbol{\lambda}\mathbf{L}\boldsymbol{\lambda}^\top).$$

By definition $\boldsymbol{\lambda}^{(L+1)} = 2\mathbf{y}\mathbf{L}^{-1}$. Hence $\boldsymbol{\lambda}\mathbf{L}\boldsymbol{\xi}^\top = 2\mathbf{y}\boldsymbol{\xi}^\top$. Plugging this expression in the previous identity and re-arranging in (14) yields

$$\mathrm{Tr}[\boldsymbol{\xi}\mathbf{L}\boldsymbol{\xi}^\top - 4\mathbf{y}\boldsymbol{\xi}^\top] = (\xi-\lambda)^\top\boldsymbol{\Lambda}^{-1}(\xi-\lambda) - \mathrm{Tr}(\boldsymbol{\lambda}\mathbf{L}\boldsymbol{\lambda}^\top). \tag{15}$$

A straightforward computation shows

$$\mathrm{Tr}[2(\boldsymbol{\xi}-\mathbf{y})(\boldsymbol{\xi}-\mathbf{y})^\top + \boldsymbol{\xi}\mathbf{K}^{-1}\boldsymbol{\xi}^\top] = \mathrm{Tr}[\boldsymbol{\xi}\mathbf{L}\boldsymbol{\xi}^\top - 4\boldsymbol{\xi}\mathbf{y}^\top] + 2\mathrm{Tr}(\mathbf{y}\mathbf{y}^\top).$$

On combining this latter relation with (15) we have

$$\mathrm{Tr}[2(\boldsymbol{\xi}-\mathbf{y})(\boldsymbol{\xi}-\mathbf{y})^\top + \boldsymbol{\xi}\mathbf{K}^{-1}\boldsymbol{\xi}^\top] = (\xi-\lambda)^\top\boldsymbol{\Lambda}^{-1}(\xi-\lambda) - \mathrm{Tr}(\boldsymbol{\lambda}\mathbf{L}\boldsymbol{\lambda}^\top) + 2\mathrm{Tr}(\mathbf{y}\mathbf{y}^\top). \tag{16}$$

Plugging the right-hand side of (16) in (14) yields

$$\begin{aligned}
&\mathrm{d}\mathbb{Q}_{G^{(L+1)}(\mathbf{x})}(\xi)\\
&= \mathfrak{c}\mathbb{E}_{\mathrm{prior}}\Bigg[(\det\mathbf{K}^{(L+1)}(\mathbf{x}))^{-n_{L+1}/2}\exp\left(\frac{1}{2}\mathrm{Tr}(\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y})\mathbf{L}^{(L+1)}(\mathbf{x})(\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y}))^\top)\right)\\
&\qquad\times\exp\left(-\frac{1}{2}(\xi-\lambda^{(L+1)}(\mathbf{x},\mathbf{y}))^\top(\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x}))^{-1}(\xi-\lambda^{(L+1)}(\mathbf{x},\mathbf{y}))\right)\Bigg]\mathrm{d}\xi\\
&= \mathfrak{c}\mathbb{E}_{\mathrm{prior}}\Bigg[\frac{(\det\mathbf{K}^{(L+1)}(\mathbf{x}))^{-n_{L+1}/2}\exp\left(\frac{1}{2}\mathrm{Tr}(\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y})\mathbf{L}^{(L+1)}(\mathbf{x})(\boldsymbol{\lambda}^{(L+1)}(\mathbf{x},\mathbf{y}))^\top)\right)}{(2\pi)^{-(n_{L+1}d)/2}\det(\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x}))^{-1/2}}\\
&\quad\times(2\pi)^{-(n_{L+1}d)/2}\det(\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x}))^{-1/2}\exp\left(-\frac{1}{2}(\xi-\lambda^{(L+1)}(\mathbf{x},\mathbf{y}))^\top(\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x}))^{-1}(\xi-\lambda^{(L+1)}(\mathbf{x},\mathbf{y}))\right)\Bigg]\mathrm{d}\xi\\
&= \mathbb{E}_{\mathbb{S}}[\varphi_{(\lambda^{(L+1)}(\mathbf{x},\mathbf{y}),\boldsymbol{\Lambda}^{(L+1)}(\mathbf{x}))}(\xi)]\mathrm{d}\xi,
\end{aligned}$$

and the proof is completed. □

Proof. (Lemma 5.5) Set

$$\mathfrak{c}_n := (\mathbb{E}[g(X_n)])^{-1} \quad\text{and}\quad \mathfrak{c} := (\mathbb{E}[g(X)])^{-1}.$$

Let S be the metric space where the random variables X_n, $n \geq 1$, and X take values and let $f : S \to \mathbb{R}$ be a bounded and continuous function. We have

$$\begin{aligned}
&\int_S f(\xi)\mathrm{d}\mathbb{Q}_{n,X_n}(\xi) - \int_S f(\xi)\mathrm{d}\mathbb{Q}_X(\xi) = \mathfrak{c}_n\int_S f(\xi)g(\xi)\mathrm{d}\mathbb{P}_{X_n}(\xi) - \mathfrak{c}\int_S f(\xi)g(\xi)\mathrm{d}\mathbb{P}_X(\xi)\\
&= \mathfrak{c}_n\int_S f(\xi)g(\xi)(\mathrm{d}\mathbb{P}_{X_n}(\xi)-\mathrm{d}\mathbb{P}_X(\xi)) + (\mathfrak{c}_n-\mathfrak{c})\int_S f(\xi)g(\xi)\mathrm{d}\mathbb{P}_X(\xi) =: \mathfrak{c}_n\, I_n + (\mathfrak{c}_n-\mathfrak{c})\, I.
\end{aligned}$$

Therefore

$$\Big|\int_S f(\xi)\mathrm{d}\mathbb{Q}_{n,X_n}(\xi) - \int_S f(\xi)\mathrm{d}\mathbb{Q}_X(\xi)\Big| \leq \mathfrak{c}_n\,|I_n| + |\mathfrak{c}_n-\mathfrak{c}|\,|I| \leq \mathfrak{c}_n\,|I_n| + \|f\|_\infty\|g\|_\infty\,|\mathfrak{c}_n-\mathfrak{c}|.$$

Since $g(\cdot)$ and $f(\cdot)g(\cdot)$ are bounded and continuous functions by the weak convergence of $\mathbb{P}_{X_n}$ to $\mathbb{P}_X$ it follows

$$\lim_{n\to\infty}\mathfrak{c}_n = \mathfrak{c} \quad\text{and}\quad \lim_{n\to\infty} I_n = 0,$$

and the proof is completed. □

6 Sufficient conditions for (11)

Condition (11) is crucial to identify the distribution of the infinite-width limit of the posterior output of the neural network (see Theorem 5.1). The following results provide sufficient conditions for (11).

Theorem 6.1. *Assume σ Lipschitz continuous and non constant and*

$$a^{(\ell-1)} = 0 \Rightarrow \mathbb{P}_{\text{prior}}(N_{\ell-1}((0,\infty)) = \infty) = 1,\ \forall\ \ell \in \{2,\dots,L+1\}. \tag{17}$$

If moreover

$$\text{The data } x(1),\dots,x(d) \text{ are linearly independent vectors, with } n_0 \geq d, \tag{18}$$

then (11) holds.

Theorem 6.2. *Assume σ Lipschitz continuous and nonlinear and*

$$a^{(1)} > 0 \text{ and } a^{(\ell-1)} = 0 \Rightarrow \mathbb{P}_{\text{prior}}(N_{\ell-1}((0,\infty)) = \infty) = 1,\ \forall\ \ell \in \{3,\dots,L+1\}. \tag{19}$$

If moreover

$$\text{If } C_B > 0 \text{ then the data } x(1),\dots,x(d) \text{ are all distinct} \tag{20}$$

and

$$\text{If } C_B = 0 \text{ then the data } x(1),\dots,x(d) \text{ are pairwise non-proportional,} \tag{21}$$

then (11) holds.

The proof of these theorems relies on the following proposition which is proved later on in this section.

Proposition 6.3. *Assume σ Lipschitz continuous and non constant, and let $\ell \in \{2,\dots,L+1\}$ be fixed. If*

$$\mathbb{P}_{\text{prior}}(\det(\mathbf{K}^{(\ell-1)}(\mathbf{x})) > 0) = 1 \tag{22}$$

and

$$a^{(\ell-1)} = 0 \Rightarrow \mathbb{P}_{\text{prior}}(N_{\ell-1}((0,\infty)) = \infty) = 1, \tag{23}$$

then

$$\mathbb{P}_{\text{prior}}(\det(\mathbf{K}^{(\ell)}(\mathbf{x})) > 0) = 1. \tag{24}$$

The proof of Proposition 6.3 exploits the following results, which are proved at the end of the section.

Proposition 6.4. *Let σ be Lipschitz continuous and non constant, and let X be a d-dimensional random vector with a density with respect to the Lebesgue measure which is strictly positive on $\mathbb{R}^d$. Then there exist $A = A_\sigma \in \mathcal{B}(\mathbb{R}^d)$ (A does not depend on X) and $\varphi_{\sigma(X)} : A \to [0,\infty)$ measurable such that*

$$\mathbb{P}(\sigma(X) \in C) = \int_C \varphi_{\sigma(X)}(y)\mathrm{d}y, \text{ for any } C \in \mathcal{B}(A), \text{ and } \mathbb{P}(\sigma(X) \in A) > 0. \tag{25}$$

Proposition 6.5. *Let X be a real-valued m-dimensional (column) random vector and suppose that its law and the Lebesgue measure on $\mathbb{R}^m$ are not singular and that $\mathbb{E}[XX^\top] < \infty$ (i.e., all the entries of the matrix $\mathbb{E}[XX^\top]$ are finite). Then the matrix $\mathbb{E}[XX^\top]$ is positive definite.*

Proposition 6.6. *Let $X_1,\dots,X_r$ be r (column) random vectors with values in $\mathbb{R}^d$, $1 \leq r \leq d$. If the law of $\mathrm{vec}((X_1,\dots,X_r))$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{dr}$, then the random vectors $\{X_1,\dots,X_r\}$ are linearly independent almost surely.*

6.1 Proof of Theorems 6.1 and 6.2

Proof of Theorem 6.1. The assumption (18) implies $\det(\mathbf{K}^{(1)}(\mathbf{x})) > 0$ (indeed, it is easily verified that if the columns of a matrix $\mathbf{A}$ are linearly independent then the Gram matrix $\mathbf{A}^\top\mathbf{A}$ is positive definite). The claim follows by Proposition 6.3. □

Proof of Theorem 6.2. By Theorems 6 and 7 in [8] we have that the matrix $\mathbb{E}_{\text{prior}}[\sigma(\zeta_1^{(1)}(\mathbf{x}))\sigma(\zeta_1^{(1)}(\mathbf{x}))^\top]$ is positive definite. Recall that

$$\mathbf{K}^{(2)}(\mathbf{x}) = C_B 1_d 1_d^\top + C_W\left(a^{(1)}\mathbb{E}_{\text{prior}}[\sigma(\zeta_1^{(1)}(\mathbf{x}))\sigma(\zeta_1^{(1)}(\mathbf{x}))^\top] + \sum_{j=1}^{N_1((0,\infty))} T_j^{(1)}\sigma(\zeta_j^{(1)}(\mathbf{x}))\sigma(\zeta_j^{(1)}(\mathbf{x}))^\top\right).$$

Since $a^{(1)} > 0$, the matrix $C_B 1_d 1_d^\top$ is positive semi-definite and the random matrix

$$\sum_{j=1}^{\infty} T_j^{(1)}\sigma(\zeta_j^{(1)}(\mathbf{x}))\sigma(\zeta_j^{(1)}(\mathbf{x}))^\top \text{ is positive semi-definite, } \mathbb{P}_{\text{prior}}\text{-a.s.}$$

we have $\mathbb{P}_{\text{prior}}(\det(\mathbf{K}^{(2)}(\mathbf{x})) > 0) = 1$. The claim follows by Proposition 6.3. □

6.2 Proof of Propositions 6.3, 6.4, 6.5 and 6.6

Proof of Proposition 6.3. We assume $C_B > 0$. The proof in the case when $C_B = 0$ is similar, and therefore omitted. Recall the definition of $\mathbf{K}^{(\ell)}(\mathbf{x})$, $\ell = 2, \dots, L+1$, in the statement of Theorem 4.1, and note that it consists of the sum of three positive semi-definite matrices. Since by assumption (23) we have that $\mathbb{P}_{\text{prior}}(N_{\ell-1}((0,\infty)) = \infty) = 1$ when $a^{(\ell-1)} = 0$, the claim follows if we prove that

$$\mathbb{E}_{\text{prior}}[\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))^\top \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x})] \text{ is positive definite, } \mathbb{P}_{\text{prior}}\text{-a.s.,} \tag{26}$$

and

$$\det\left(\sum_{j=1}^{\infty} T_j^{(\ell-1)}\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))^\top\right) > 0, \quad \mathbb{P}_{\text{prior}}\text{-a.s.} \tag{27}$$

Proof of (26). By (22), for $(\mathbb{P}_{\text{prior}})_{\mathbf{K}^{(\ell-1)}(\mathbf{x})}$-almost all $\mathbf{k}$, under the prior, $\zeta_1^{(\ell-1)}(\mathbf{x}) \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}$ has a Gaussian density (which is obviously strictly positive on $\mathbb{R}^d$). So, by Proposition 6.4 we have that the law of $\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})) \,|\mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}$ under $\mathbb{P}_{\text{prior}}$ and the Lebesgue measure are not singular. By the Lipschitzianity of σ and the Gaussianity of $\zeta_1^{(\ell-1)}(\mathbf{x}) \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}$, all the entries of

$$\mathbb{E}_{\text{prior}}[\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))^\top \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}]$$

are finite. The claim then follows by Proposition 6.5.
Proof of (27). Note that, for $n \geq d$ arbitrarily fixed, by Lemma 3.1 we have

$$\begin{aligned}\Big\{rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) = d\Big\} &\equiv \Big\{\sum_{j=1}^{n} T_j^{(\ell-1)}\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))^\top \quad \text{is positive definite}\Big\} \\ &\subset \Big\{\sum_{j\geq 1} T_j^{(\ell-1)}\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))^\top \quad \text{is positive definite}\Big\}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{P}_{\text{prior}}\left(\det\left(\sum_{j\geq 1} T_j^{(\ell-1)}\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))\sigma(\zeta_j^{(\ell-1)}(\mathbf{x}))^\top\right) > 0\right) \\ &\qquad \geq \mathbb{P}_{\text{prior}}\left(rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) = d\right), \quad \text{for each } n \geq d.\end{aligned} \tag{28}$$

By Proposition 6.4, for $(\mathbb{P}_{\text{prior}})_{\mathbf{K}^{(\ell-1)}(\mathbf{x})}$-almost all $\mathbf{k}$, there exist $A = A_\sigma \in \mathcal{B}(\mathbb{R}^d)$ and $\varphi_{\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\,|\,\mathbf{K}^{(\ell-1)}(\mathbf{x})=\mathbf{k}} : A \to [0,\infty)$ measurable such that

$$\mathbb{P}_{\text{prior}}(\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})) \in C \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}) = \int_C \varphi_{\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\,|\,\mathbf{K}^{(\ell-1)}(\mathbf{x})=\mathbf{k}}(y)\mathrm{d}y, \ \forall\, C \in \mathcal{B}(A) \tag{29}$$

(indeed, since $\mathbf{K}^{(\ell-1)}(\mathbf{x})$ is positive definite $\mathbb{P}_{\text{prior}}$-a.s. , for $(\mathbb{P}_{\text{prior}})_{\mathbf{K}^{(\ell-1)}(\mathbf{x})}$-almost all $\mathbf{k}$, the random vector $\zeta_1^{(\ell-1)}(\mathbf{x})\,|\,\mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}$, which has the Gaussian distribution $\mathcal{N}_d(0,\mathbf{k})$, has a strictly positive density). Set

$$p_{\mathbf{k}} := \mathbb{P}_{\text{prior}}(\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})) \in A \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}) = \int_A \varphi_{\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}))\,|\,\mathbf{K}^{(\ell-1)}(\mathbf{x})=\mathbf{k}}(y)\mathrm{d}y > 0. \tag{30}$$

For all $n \geq d$, define the event

$$D_{n,d} := \{\text{the matrix } (\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x}))) \text{ has at least } d \text{ columns in } A\}.$$

We will prove later on that

$$\mathbb{P}_{\text{prior}}(D_{n,d} \cap \{rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) \leq d-1\}) = 0. \tag{31}$$

Thus, by assuming (31),

$$\begin{aligned}\mathbb{P}_{\text{prior}}(D_{n,d}) &= \mathbb{P}_{\text{prior}}(D_{n,d} \cap \{rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) \geq d\})\\ &\leq \mathbb{P}_{\text{prior}}(rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x}))))) = d).\end{aligned}$$

So (27) follows by (28) and this latter relation if we prove that $\lim_{n\to\infty} \mathbb{P}_{\text{prior}}(D_{n,d}) = 1$. Due to the dominated convergence theorem, in turn, this latter limit holds if

$$\lim_{n\to\infty} \mathbb{P}_{\text{prior}}(D_{n,d} \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}) = 1.$$

In order to prove this latter relation, we note that since, under $\mathbb{P}_{\text{prior}}$, given $\{\mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}\}$, the random vectors $\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x}))$ are independent and identically distributed, recalling the definition of $p_{\mathbf{k}}$ in (30), we have that

$$\begin{aligned}\mathbb{P}_{\text{prior}}(D_{n,d} \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}) &= \mathbb{P}_{\text{prior}}\Big((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x}), \ldots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))\\ &\qquad\qquad \text{has at least } d \text{ columns in } A \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}\Big)\\ &= \sum_{k=d}^{n} \binom{n}{k} p_{\mathbf{k}} a^k (1-p_{\mathbf{k}})^{n-k}.\end{aligned}$$

Since

$$\lim_{n\to\infty} \sum_{k=0}^{d} \binom{n}{k} p_{\mathbf{k}}^k (1-p_{\mathbf{k}})^{n-k} = 0,$$

we have

$$1 = \lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p_{\mathbf{k}}^k (1-p_{\mathbf{k}})^{n-k} = \lim_{n\to\infty} \sum_{k=d}^{n} \binom{n}{k} p_{\mathbf{k}}^k (1-p_{\mathbf{k}})^{n-k} = \lim_{n\to\infty} \mathbb{P}_{\text{prior}}(D_{n,d} \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}),$$

and the proof is completed.

It remains to prove (31). We start noticing that

$$
\begin{aligned}
&D_{n,d} \cap \{rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) \leq d-1\} \\
&\equiv \Bigg\{ \exists\, j \in \{d, \dots, n\} \text{ and } i_1, \dots, i_j \in \{1, \dots, n\} \text{ such that } \sigma(\zeta_{i_h}^{(\ell-1)}(\mathbf{x})) \in A \;\forall\, h = 1, \dots, j \\
&\qquad\qquad \text{and } rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x}))))) \leq d-1 \Bigg\} \\
&\subseteq \bigcup_{j=d}^{n} \bigcup_{\{i_1,\dots,i_j\} \subseteq \{1,\dots,n\}} \Bigg\{ \sigma(\zeta_{i_h}^{(\ell-1)}(\mathbf{x})) \in A \;\forall\, h = 1, \dots, j \\
&\qquad\qquad \text{and } \sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x})) \text{ are linearly dependent} \Bigg\}.
\end{aligned}
$$

Set

$$L_j := \{(x_1, \dots, x_j) \in \mathbb{R}^{d \times j} : \; x_1, \dots, x_j \quad \text{are linearly dependent}\}.$$

Using the union bound, it follows

$$
\begin{aligned}
&\mathbb{P}_{\text{prior}}(D_{n,d} \cap \{rk((\sigma(\zeta_1^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_n^{(\ell-1)}(\mathbf{x})))) \leq d-1\}) \\
&\leq \sum_{j=d}^{n} \sum_{\{i_1,\dots,i_j\} \subseteq \{1,\dots,n\}} \mathbb{P}_{\text{prior}}\Bigg(\sigma(\zeta_{i_h}^{(\ell-1)}(\mathbf{x})) \in A \;\forall\, h = 1, \dots, j \\
&\qquad\qquad \text{and } \sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x})) \text{ lin. dep.} \Bigg) \\
&= \sum_{j=d}^{n} \sum_{\{i_1,\dots,i_j\} \subseteq \{1,\dots,n\}} \mathbb{P}_{\text{prior}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))) \in A^j \cap L_j),
\end{aligned}
$$

and so it suffices to prove that

$$\text{For each } j = d, \dots, n,\; \mathbb{P}_{\text{prior}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))) \in A^j \cap L_j \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}) = 0. \tag{32}$$

Define the probability measure

$$
\begin{aligned}
&\eta_{\mathbf{k}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x})))^{-1}(C)) \\
&:= \frac{\mathbb{P}_{\text{prior}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))) \in C \cap A^j \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k})}{\mathbb{P}_{\text{prior}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))) \in A^j \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k})} \\
&\quad = p_{\mathbf{k}}^{-j} \mathbb{P}_{\text{prior}}((\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))) \in C \cap A^j \,|\, \mathbf{K}^{(\ell-1)}(\mathbf{x}) = \mathbf{k}), \quad C \in \mathcal{B}(\mathbb{R}^{d \times j}).
\end{aligned} \tag{33}
$$

Here the equality follows by (30) and the fact that, under the prior, given $\mathbf{K}^{(\ell-1)}(\mathbf{x})$, the random vectors $\zeta_1^{(\ell-1)}(\mathbf{x}), \dots, \zeta_n^{(\ell-1)}(\mathbf{x})$ are independent and identically distributed. Due to (29), combining this conditional independence with (33) we have that, under $\eta_{\mathbf{k}}$, the joint law of $\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))$ is absolutely continuous with respect to the Lebesgue measure, and so by Proposition 6.6 the random vectors $\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))$ are linearly independent $\eta_{\mathbf{k}}$-a.s.. Therefore

$$\eta_{\mathbf{k}}\left(\left(\sigma(\zeta_{i_1}^{(\ell-1)}(\mathbf{x})), \dots, \sigma(\zeta_{i_j}^{(\ell-1)}(\mathbf{x}))\right)^{-1} (L_j) \right) = 0,$$

which gives (32). $\square$

Proof of Proposition 6.4. We will prove that (25) holds with

$$A := \mathbb{R}^d \setminus S \quad \text{and} \quad \varphi_{\sigma(X)}(y) := \begin{cases} \sum_{x\in\sigma^{-1}(y)} \frac{\varphi_X(x)}{J_\sigma(x)} & \text{if } \sigma^{-1}(y) \text{ is a finite set} \\ +\infty & \text{otherwise} \end{cases}$$

where $J_\sigma(x)$ is the absolute value of the determinant of the Jacobian matrix of σ,

$$S := \{y \in \mathbb{R}^d : \exists\, x \in \sigma^{-1}(y) \text{ with } J_\sigma(x) = 0\}$$

and φ_X is the density of X. Let $B \in \mathcal{B}(\mathbb{R}^d \setminus S)$ be arbitrarily fixed. By the area formula applied to non-negative and measurable functions u (see Theorem 3.2.2 in [15] and Theorem 2.93 together with the following remark in [2]), we have

$$\int_{\mathbb{R}^d} u(x)J_\sigma(x)\mathrm{d}x = \int_{\mathbb{R}^d} \sum_{x\in\sigma^{-1}(y)} u(x)\mathrm{d}y. \tag{34}$$

By taking

$$u(x) := \frac{\varphi_X(x)}{J_\sigma(x)}\mathbf{1}\{\sigma(x) \in B\},$$

we get

$$\mathbb{P}(\sigma(X) \in B) = \int_{\sigma^{-1}(B)} \varphi_X(x)\mathrm{d}x = \int_B \varphi_{\sigma(X)}(y)\mathrm{d}y.$$

It remains to prove that $\mathbb{P}(\sigma(X) \in \mathbb{R}^d \setminus S) > 0$. This claim is equivalent to $\mathbb{P}(\sigma(X) \in S) < 1$ i.e. $\mathbb{P}(X \in \mathbb{R}^d \setminus \sigma^{-1}(S)) > 0$. Note that

$$\sigma^{-1}(S) = T_0 \cup T_1, \text{ where } T_0 := \{x \in \sigma^{-1}(S) : J_\sigma(x) \neq 0\} \text{ and } T_1 := \{x \in \sigma^{-1}(S) : J_\sigma(x) = 0\}.$$

We will show later on that

$$\mathrm{Leb}^{(d)}(T_0) = 0, \tag{35}$$

where $\mathrm{Leb}^{(d)}$ denotes the Lebesgue measure on $\mathbb{R}^d$. Therefore, by the absolute continuity of X, we have

$$\mathbb{P}(X \in \mathbb{R}^d \setminus \sigma^{-1}(S)) = \mathbb{P}(X \in \mathbb{R}^d \setminus T_1),$$

and we reduce ourselves to prove that $\mathbb{P}(X \in \mathbb{R}^d \setminus T_1) > 0$. Since by assumption X has density φ_X such that $\varphi_X > 0$ on $\mathbb{R}^d$ this latter inequality is equivalent to $\mathrm{Leb}^{(d)}(\mathbb{R}^d \setminus T_1) > 0$. In the final part of this proof, we show that this latter claim is guaranteed by the fact that σ is non constant and Lipschitz. Reasoning by contradiction suppose that $\mathrm{Leb}^{(d)}(\mathbb{R}^d \setminus T_1) = 0$, then for $\mathrm{Leb}^{(d)}$-almost all $x \in \mathbb{R}^d$ we have $x \in T_1$ and so $J_\sigma(x) = 0$. Since σ acts componentwise on vectors the Jacobian matrix of σ at $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ is diagonal with diagonal elements $\sigma'(x_i)$, $i = 1, \dots, d$. Therefore, for $\mathrm{Leb}^{(1)}$-almost all $x \in \mathbb{R}$ it holds $\sigma'(x) = 0$. Since σ is Lipschitz continuous this implies that σ is constant, which is a contradiction. We finally prove (35). By the area formula (34) with

$$u(x) := \frac{1}{J_\sigma(x)}\mathbf{1}\{x \in T_0\}$$

we have

$$\mathrm{Leb}^{(d)}(T_0) = \int_{\mathbb{R}^d} \sum_{x\in\sigma^{-1}(y)} \frac{1}{J_\sigma(x)}\mathbf{1}\{x \in T_0\}\mathrm{d}y = \int_{\sigma(T_0)} \sum_{x\in\sigma^{-1}(y)} \frac{1}{J_\sigma(x)}\mathrm{d}y \leq \int_S \sum_{x\in\sigma^{-1}(y)} \frac{1}{J_\sigma(x)}\mathrm{d}y = 0,$$

where the vanishing of the latter integral follows noticing that the Lipschitzianity of σ implies $\mathrm{Leb}^{(d)}(S) = 0$, which, in turn, follows by the (co)area formula (34) with the choice $u(x) = \mathbf{1}\{x \in T_1\}$ (this is the weak Morse-Sard property, see Remark *(ii)* after Theorem 1 in [12, Section 3.4.2]). □

Proof of Proposition 6.5. We start noticing that the matrix $\mathbb{E}[XX^\top]$ is positive semi-definite. Reasoning by contradiction, suppose that there exists $v \in \mathbb{R}^m \setminus \{0\}$ such that $\mathbb{E}|v^\top X|^2 = 0$. Hence, $v^\top X = 0$ $\mathbb{P}$-a.s., i.e. $\mathbb{P}_X(H_v) = 1$, where by H_v we are denoting the hyperplane $\{y \in \mathbb{R}^m : \ v^\top y = 0\}$. Let A^c denote the complement of a set A. Since $\mathbb{P}_X$ and the Lebesgue measure are not singular, letting $\mathbb{P}_{X,\ll}$ denote the continuous part in the Lebesgue decomposition of $\mathbb{P}_X$ with respect to the Lebesgue measure, we have that there exists a Borel set $C \subseteq \mathbb{R}^m$ such that $\mathbb{P}_{X,\ll}(C) > 0$. Therefore, since $\mathbb{P}_{X,\ll}(H_v) = 0$, we have

$$\begin{aligned} 0 = 1 - \mathbb{P}_X(H_v) = \mathbb{P}_X(H_v^c) \geq \mathbb{P}_{X,\ll}(H_v^c) \geq \mathbb{P}_{X,\ll}(C \cap H_v^c) \\ = \mathbb{P}_{X,\ll}(C \cap H_v^c) + \mathbb{P}_{X,\ll}(B \cap H_v) = \mathbb{P}_{X,\ll}(C) > 0. \end{aligned}$$

This is a contradiction, and the proof is completed. □

Proof of Proposition 6.6. Recall that, for $1 \leq r \leq d$, the vectors $a_1, \ldots, a_r$ of $\mathbb{R}^d$ are linearly dependent if and only if every $r \times r$ submatrix of $\mathbf{A} := (a_1, \ldots, a_r)$ has determinant equal to zero. Let $\mathfrak{J}_r$ be the collection of subsets of $\{1, \ldots, d\}$ with cardinality equal to r. For $J \in \mathfrak{J}_r$, $J = \{j_1, \ldots j_r\}$, let $p_J : \mathbb{R}^d \to \mathbb{R}^r$ be the projection $(x_1, \ldots, x_d)^\top \mapsto (x_{j_1}, \ldots, x_{j_r})^\top$. Clearly, for every $J \in \mathfrak{J}_r$, $\widetilde{p}_J(\mathbf{A}) := (p_J(a_1), \ldots, p_J(a_r))$ is an $r \times r$ submatrix of $\mathbf{A}$, and every $r \times r$ submatrix of $\mathbf{A}$ arises in this way for some $J \in \mathfrak{J}_r$. Note that $a_1, \ldots, a_r$ are linearly dependent if and only if $f(\mathbf{A}) = 0$ where $f : \mathbb{R}^{d \times r} \to [0, \infty)$ is defined by

$$f(\mathbf{A}) := \sum_{J \in \mathfrak{J}_r} |\det(\widetilde{p}_J(\mathbf{A}))|.$$

It is easily realized that f is a continuous function (as composition of continuous functions), and so f is measurable. Set $\mathbf{X} = (X_1, \ldots, X_r)$ and

$$E := \{\text{the random vectors } X_1, \ldots, X_r \text{ are linearly independent}\}.$$

We have

$$\mathbb{P}(E) = 1 - \mathbb{P}(f(\mathbf{X}) = 0) \geq 1 - \mathbb{P}(\det(\widetilde{p}_{\{1,\ldots,r\}}(\mathbf{X})) = 0).$$

Since $\det(\widetilde{p}_{\{1,\ldots,r\}}(\cdot))$ is a polynomial function, by the absolute continuity of the law of $\mathrm{vec}((X_1, \ldots, X_r))$ with respect to the Lebesgue measure we have

$$\mathbb{P}(\det(\widetilde{p}_{\{1,\ldots,r\}}(\mathbf{X})) = 0) = 0.$$

Here, we also used the classical fact that every non-identically zero polynomial function $p : \mathbb{R}^m \to \mathbb{R}$ is non-zero almost everywhere, which can be seen by induction on m in view of Fubini's theorem (see, e.g., [15, Section 2.6.5]). Therefore $\mathbb{P}(E) = 1$, and the proof is completed. □

7 Numerical illustrations

In this section we provide numerical illustrations of Theorem 5.1. More precisely, we consider two specific models of neural networks with dependent and heavy-tailed weights, we sample from the posterior distribution of the output at a finite and at an infinite width and we verify the convergence, as the number of nodes in the hidden layers grows large. We start introducing the models.

7.1 Model 1

The first model that we analyze is a neural network with dependent weights, as defined in Sections 2.1 and 2.2, with $C_B > 0$ and $V_{n_\ell,j}^{(\ell)} := Y_j^{(\ell)}/n_\ell$, $j = 1, \ldots, n_\ell$, $\ell = 1, \ldots, L$, where, under $\mathbb{P}_{\text{prior}}$, $\{Y_j^{(\ell)}\}_{j \geq 1, \ell = 1, \ldots, L}$ are independent random variables and $\{Y_j^{(\ell)}\}_{j \geq 1}$ are identically distributed with $\mathbb{P}_{\text{prior}}(Y_1^{(\ell)} > 0) = 1$ and $\mathbb{E}_{\text{prior}} Y_1^{(\ell)} \in (0, \infty)$. By the law of the large numbers, for each $\ell = 1, \ldots, L$, we have

$$\sum_{j=1}^{n_\ell} V_{n_\ell,j}^{(\ell)} \to \mathbb{E}_{\text{prior}} Y_1^{(\ell)}, \quad \mathbb{P}_{\text{prior}}\text{-a.s., as } n_\ell \to \infty.$$

Therefore, the assumption (ii) of Theorem 4.1 holds with $a^{(\ell)} = \mathbb{E}_{\text{prior}} Y_1^{(\ell)}$ and $\rho^{(\ell)} \equiv 0$. If the inputs are e.g. distinct and the activation function σ is e.g. Lipschitz continuous and nonlinear, then all the hypotheses of Theorem 6.2 are fullfilled and the corresponding assumption (11) holds.

To make the model more realistic, under the prior, the weights should be heavy-tailed (see the introduction and the references [29, 38]). For instance, this happens if, under $\mathbb{P}_{\text{prior}}$, $\{Y_j^{(\ell)}\}_{j\geq 1, \ell=1,\ldots,L}$ are independent and identically distributed random variables with $Y_1^{(1)} \stackrel{\mathcal{L}}{=} (WE)^2$, where WE has the Weibull distribution with parameters $(1, 1/2)$, i.e., it has density (with respect to the Lebesgue measure)

$$f_{WE}(x) := \frac{1}{2} x^{-1/2} \mathrm{e}^{-\sqrt{x}}, \quad x > 0$$

(we could work with other heavy-tailed distributions, however, since we are interested in simulating the model, we prefer to make a specific choice of the parameters soon at this stage). Indeed, in such a case, the Laplace transform of $W_{11}^{(2)}$ on $\mathbb{R} \setminus \{0\}$ is equal to infinity, as a simple calculation shows. Note also that a standard computation gives $\mathbb{E}_{\text{prior}}(WE)^2 = 24$, and so in this specific case $a^{(\ell)} = 24$, $\ell = 1, \ldots, L$.

7.2 Model 2

The second model that we analyze is a neural network with dependent weights, as defined in Sections 2.1 and 2.2, with $C_B > 0$ and $V_{n_\ell,j}^{(\ell)} := \frac{\pi^2}{n_\ell^2} Y_j^{(\ell)}$, $j = 1, \ldots, n_\ell$, $\ell = 1, \ldots, L$, where, under $\mathbb{P}_{\text{prior}}$, $\{Y_j^{(\ell)}\}_{j\geq 1, \ell=1,\ldots,L}$ is a family of independent and identically distributed random variables with $Y_1^{(1)} \stackrel{d}{=} (HC)^2$, where HC denotes a random variable distributed according to the half-Cauchy law, i.e., with probability density (with respect to the Lebesgue measure)

$$f_{HC}(x) := \frac{2}{\pi(1+x^2)} \mathbf{1}\{x > 0\}.$$

Note that $\mathbb{P}_{\text{prior}}(Y_1^{(1)} > 0) = 1$, $\ell = 1, \ldots, L$. Under $\mathbb{P}_{\text{prior}}$, it turns out (see Appendix E in [26]) that, for $\ell = 1, \ldots, L$,

$$\sum_{j=1}^{n_\ell} V_{n_\ell,j}^{(\ell)} \to \mathrm{ID}(0, \rho), \quad \text{in distribution, as } n_\ell \to \infty,$$

where

$$\rho(\mathrm{d}x) := x^{-3/2} \mathbf{1}\{x > 0\} \mathrm{d}x. \tag{36}$$

Therefore, the assumption (ii) of Theorem 4.1 holds with $a^{(\ell)} = 0$ and $\rho^{(\ell)} = \rho$ for each ℓ. Note that, for any $\varepsilon > 0$,

$$\rho((\varepsilon, \infty)) = 2\varepsilon^{-1/2} \quad \text{and} \quad \rho((0, \varepsilon]) = +\infty,$$

and therefore a Poisson process with mean measure ρ have infinitely many points on $(0, \infty)$. Consequently, the assumption (17) holds. If the inputs are such that (18) is satisfied and the activation function σ is Lipschitz and non constant, then all the hypotheses of Theorem 6.1 are fullfilled and the corresponding assumption (11) holds.

Here again, to make the model more realistic, under the prior, the weights should be heavy-tailed (see the introduction and the references [29, 38]). This is the case for the Model 2. Indeed, the Laplace transform of $W_{11}^{(2)}$ on $\mathbb{R} \setminus \{0\}$ is equal to infinity, as can be easily verified by a simple computation.

7.3 Simulation of Model 1

We consider the Model 1 with $C_B := 1$, $C_W := 1$, $L := 2$, $n_0 := 4$, $n_3 := 1$, $n_1 = n_2 = n \in \{2, 4, 8, 16, 32\}$, $x(1) := (1, 0, 0, 0)^\top$, $x(2) := (0, 1, 0, 0)^\top$, $x(3) := (0, 0, 1, 0)^\top$, $d := 3$, $y(1) := f(x(1))$, $y(2) := f(x(2))$, $y(3) := f(x(3))$, where $f(v_1, v_2, v_3, v_4) := v_1^2 + 2v_1v_2 + v_3^2$, and activation function $\sigma(x) := \max\{0, x\}$. Then

$$\mathbf{K}^{(1)}(\mathbf{x}) := (K^{(1)}(x(i), x(i')))_{1\leq i,i'\leq 3}, \qquad \text{where} \qquad K^{(1)}(x(i), x(i')) := 1 + \frac{1}{4} x(i)^\top x(i'),$$

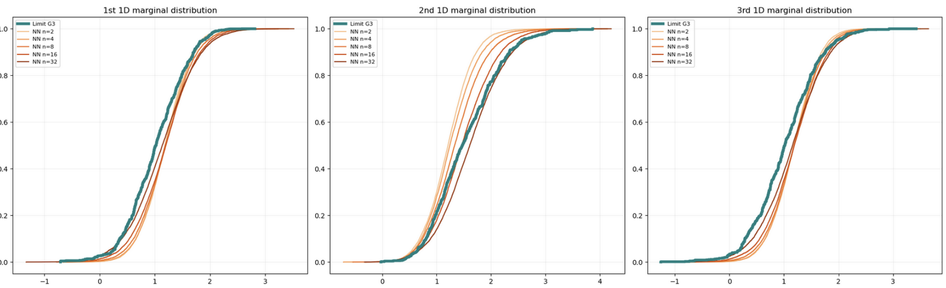

Figure 1: **Simulation of Model 1**. The non-green curves are estimates of the one dimensional marginal distribution functions (with $n_1 = n$ nodes in first hidden layer and $n_2 = 2n$ in second hidden layer, for different values of $n = 4, 8, 16, 32$) of the distribution function of the posterior law of the 3-variate output $Z^{(3)}(\mathbf{x})$. The green curves are the one dimensional marginal distribution functions of the distribution function of the 3-variate infinite-width limit $G^{(3)}(\mathbf{x})$. All the parameters are specified in Section 7.3.

and for $\ell = 1, 2$, given $\zeta_1^{(\ell)}(\mathbf{x}) \sim \mathcal{N}_3(0, \mathbf{K}^{(\ell)}(\mathbf{x}))$,

$$\mathbf{K}^{(\ell+1)}(\mathbf{x}) = 1_3 1_3^\top + 24\mathbb{E}_{\text{prior}}[\max\{\zeta_1^{(\ell)}(\mathbf{x}), 0\} \max\{\zeta_1^{(\ell)}(\mathbf{x}), 0\}^\top] \,.$$

By Theorem 5.1 if we denote by $G^{(3)}(\mathbf{x})$ the infinite-width limit of the posterior law of the output, then

$$G^{(3)}(\mathbf{x}) \sim \mathcal{N}_3(\lambda^{(3)}(\mathbf{x}, \mathbf{y}), \mathbf{\Lambda}^{(3)}(\mathbf{x})).$$

Since, for $A \in \mathcal{B}(\mathbb{R}^3)$,

$$\mathbb{P}_{\text{posterior}}(Z^{(3)}(\mathbf{x}) \in A) = \mathfrak{c}\mathbb{E}_{\text{prior}}\left[\mathbf{1}_A(Z^{(3)}(\mathbf{x})) \exp\left(-\sum_{i=1}^{3} \|Z^{(3)}(x(i)) - y(i)\|^2\right)\right],$$

if the number of nodes is kept fixed, we sample from the posterior law of $Z^{(3)}(\mathbf{x})$ simulating this random variable under the prior, and then using a Monte Carlo estimator. See Figure 1 for a numerical validation of Theorem 5.1.

7.4 Simulation of Model 2

We consider the Model 2 with $C_B := 1$, $C_W := 1$, $L = 1$, $n_0 := 4$, $n_1 \in \{2, 4, 8, 16, 32\}$, $n_2 := 1$, $d := 3$, $x(1) := (1, 0, 0, 0)^\top$, $x(2) := (0, 1, 0, 0)^\top$, $x(3) := (0, 0, 1, 0)^\top$, $y(1) := f(x(1))$, $y(2) := f(x(2))$, $y(3) := f(x(3))$, where $f(v_1, v_2, v_3, v_4) = v_1^2 + 2v_1v_2 + v_3^2$, and activation function $\sigma(x) := \max\{0, x\}$. Then

$$\mathbf{K}^{(1)}(\mathbf{x}) := (K^{(1)}(x(i), x(i')))_{1 \le i, i' \le 3}, \qquad \text{where} \qquad K^{(1)}(x(i), x(i')) := 1 + \frac{1}{4} x(i)^\top x(i'),$$

and

$$\mathbf{K}^{(2)}(\mathbf{x}) := 1_3 1_3^\top + \sum_{j=1}^{\infty} T_j^{(1)} \sigma(\zeta_j^{(1)}(\mathbf{x})) \sigma(\zeta_j^{(1)}(\mathbf{x}))^\top.$$

Here, $\{\zeta_j^{(1)}(\mathbf{x})\}_{j \ge 1}$ is a sequence of i.i.d. random vectors with $\zeta_1^{(1)}(\mathbf{x}) \sim \mathcal{N}_3(0, \mathbf{K}^{(1)}(\mathbf{x}))$, independent of $\{T_j^{(1)}\}_{j \ge 1}$, which are the points of a Poisson process on $(0, \infty)$ with mean measure ρ defined by (36).

By Theorem 5.1 if we denote by $G^{(2)}(\mathbf{x})$ the infinite-width limit of the posterior law of the output, it holds

$$G^{(2)}(\mathbf{x}) \sim \mathcal{GM}_3^{\mathbb{S}}(\lambda^{(2)}(\mathbf{x}, \mathbf{y}), \mathbf{\Lambda}^{(2)}(\mathbf{x})).$$

Therefore, the probability of the event $\{G^{(2)}(\mathbf{x}) \in A\}$, $A \in \mathcal{B}(\mathbb{R}^3)$ is equal to

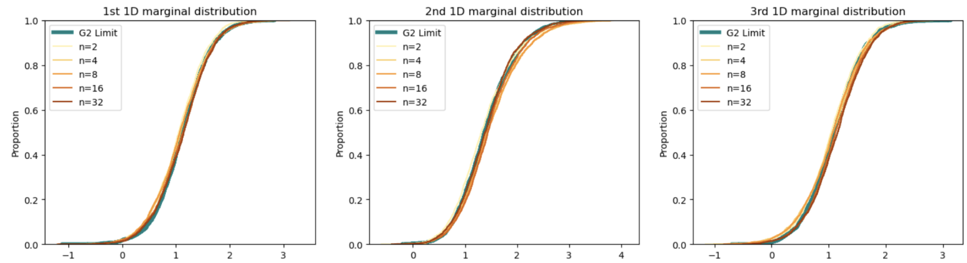

Figure 2: **Simulation of Model 2**. The non-green curves are estimates of the one dimensional marginal distribution functions (for different values $n_1 = n = 2, 4, 8, 16, 32$ of the number of nodes in the hidden layer) of the posterior law of the 3-variate output $Z^{(2)}(\mathbf{x})$. The green curves are the one dimensional marginal distribution functions of the distribution function of the 3-variate infinite-width limit $G^{(2)}(\mathbf{x})$. All the parameters are specified in Section 7.4.

$$\mathbb{E}_{\mathbb{S}}\left[\int_A \varphi_{(\lambda^{(2)}(\mathbf{x},\mathbf{y}),\mathbf{\Lambda}^{(2)}(\mathbf{x}))}(\xi)\mathrm{d}\xi\right]$$
$$= \mathfrak{c}\mathbb{E}_{\text{prior}}\left[\frac{\exp\left(\mathrm{Tr}(\mathbf{y}(\mathbf{Id}_3 + (2\mathbf{K}^{(2)}(\mathbf{x}))^{-1})^{-1})\mathbf{y}^\top\right)}{(\det(\mathbf{Id}_3 + 2\mathbf{K}^{(2)}(\mathbf{x})))^{n_2/2}}\int_A \varphi_{(\lambda^{(2)}(\mathbf{x},\mathbf{y}),\mathbf{\Lambda}^{(2)}(\mathbf{x}))}(\xi)\mathrm{d}\xi\right].$$

So, we sample from the law of $G^{(2)}(\mathbf{x})$ simulating the random matrix $\mathbf{K}^{(2)}(\mathbf{x})$ under the prior, and then using a Monte Carlo estimator. To sample the random matrix $\mathbf{K}^{(2)}(\mathbf{x})$ (under the prior), we note that letting $\{T^{(1)}_{(j)}\}_{j\geq 1}$, $T^{(1)}_{(1)} > T^{(1)}_{(2)} > \ldots$, denote the sequence of points $\{T^{(1)}_j\}_{j\geq 1}$ ordered in decreasing way, one has

$$\sum_{j\geq 1} T^{(1)}_j \sigma(\zeta^{(1)}_j(\mathbf{x}))\sigma(\zeta^{(1)}_j(\mathbf{x}))^\top \overset{\mathcal{L}}{=} \sum_{j\geq 1} T^{(1)}_{(j)} \sigma(\zeta^{(1)}_j(\mathbf{x}))\sigma(\zeta^{(1)}_j(\mathbf{x}))^\top,$$

and, as noticed in Appendix E.3.2 of [26],

$$T^{(1)}_{(j)} = \frac{4}{(\sum_{k=1}^j E_k)^2},$$

where $\{E_k\}_{k\geq 1}$ is a sequence of independent random variables with the exponential law with mean 1.

If the number of nodes is kept fixed, we sample from the posterior law of $Z^{(\ell)}(\mathbf{x})$, $\ell = 1, 2$, simulating these random variables under the prior, and then using a Monte Carlo estimator as described in Section 7.3.

See Figure 2 for another numerical validation of Theorem 5.1.

Appendix: Proof of Theorem 4.2

The proof of Theorem 4.2 is based on two lemmas. The first one, Lemma 7.1, provides an alternative representation of the neural network; its proof is omitted since it is similar to the proof of Lemma 3 in [27]. The second one, Lemma 7.2, is concerned with the weak limit of triangular arrays; its proof can be found in [26], see Corollary 41 and the last lines of the proof of Theorem 16 therein.

Let $\{\widetilde{N}^{(\ell)}_{jr}\}_{j,\ell\geq 1; r=1,\ldots,d}$ be a family of independent standard normal random variables, independent of the family $\{V^{(\ell-1)}_{n_{\ell-1},j}\}_{\ell=1,\ldots,L+1; j=1,\ldots n_{\ell-1}}$. For $\ell = 1, \ldots, L$, we define the symmetric $d \times d$ positive semi-definite random matrices

$$\mathbf{H}^{(\ell)}(\mathbf{x}) = C_B 1_d 1_d^\top + C_W \sum_{j=1}^{n_\ell} V^{(\ell)}_{n_\ell,j} \sigma(\gamma^{(\ell-1)}_j(\mathbf{x}))\sigma(\gamma^{(\ell-1)}_j(\mathbf{x}))^\top \tag{37}$$

and

$$\mathbf{H}^{(0)}(\mathbf{x}) = \mathbf{K}^{(1)}(\mathbf{x}) := C_B + \frac{C_W}{n_0}\mathbf{x}^\top\mathbf{x}.$$

Here, for $\ell = 1, \ldots, L$ and $j = 1, \ldots, n_\ell$,

$$\gamma_j^{(\ell-1)}(\mathbf{x}) := \left(\sum_{t=1}^{d} \widetilde{N}_{jt}^{(\ell)} H_{t1}^{(\ell-1),\sharp}(\mathbf{x}), \ldots, \sum_{t=1}^{d} \widetilde{N}_{jt}^{(\ell)} H_{td}^{(\ell-1),\sharp}(\mathbf{x}) \right)^\top.$$

Moreover, for $\ell = 0, \ldots, L$ and $r, s = 1, \ldots, d$, we denote by $H_{rs}^{(\ell)}(\mathbf{x})$ the rs-entry of $\mathbf{H}^{(\ell)}(\mathbf{x})$ and by $H_{rs}^{(\ell),\sharp}(\mathbf{x})$ the rs-entry of $[\mathbf{H}^{(\ell)}(\mathbf{x})]^\sharp$ (i.e., the square-root of $\mathbf{H}^{(\ell)}(\mathbf{x})$). Note that, given $\mathbf{H}^{(\ell-1)}(\mathbf{x})$, $\{\gamma_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\ldots,n_\ell}$ are independent and identically distributed with law $\mathcal{N}_d(0, \mathbf{H}^{(\ell-1)}(\mathbf{x}))$. Note also that

$$H_{rs}^{(\ell)}(\mathbf{x}) := C_B + C_W \sum_{j=1}^{n_\ell} V_{n_\ell j}^{(\ell)} \sigma\left(\sum_{t=1}^{d} \widetilde{N}_{jt}^{(\ell)} H_{tr}^{(\ell-1),\sharp}(\mathbf{x}) \right) \sigma\left(\sum_{t=1}^{d} \widetilde{N}_{jt}^{(\ell)} H_{ts}^{(\ell-1),\sharp}(\mathbf{x}) \right).$$

If we set $\boldsymbol{\gamma}^{(\ell-1)}(\mathbf{x}) = [\gamma_1^{(\ell-1)}(\mathbf{x}), \ldots, \gamma_{n_\ell}^{(\ell-1)}(\mathbf{x})] \in \mathbb{R}^{d \times n_\ell}$, $\mathbf{V}^{(\ell)} = \operatorname{diag}\left(V_{n_\ell,1}^{(\ell)}, \ldots, V_{n_\ell,n_\ell}^{(\ell)}\right) \in \mathbb{R}^{n_\ell \times n_\ell}$ and $\widetilde{\mathbf{N}}^{(\ell)} = (\widetilde{N}_{ij}^{\ell}) \in \mathbb{R}^{n_\ell \times d}$, then

$$\mathbf{H}^{(\ell)}(\mathbf{x}) = C_B 1_d 1_d^\top + C_W \sigma\left(\widetilde{\mathbf{N}}^{(\ell)}[\mathbf{H}^{(\ell-1)}(\mathbf{x})]^\sharp\right)^\top \mathbf{V}^{(\ell)} \sigma\left(\widetilde{\mathbf{N}}^{(\ell)}[\mathbf{H}^{(\ell-1)}(\mathbf{x})]^\sharp\right) \tag{38}$$

$$\boldsymbol{\gamma}^{(\ell-1)}(\mathbf{x}) = \left(\widetilde{\mathbf{N}}^{(\ell)}[\mathbf{H}^{(\ell-1)}(\mathbf{x})]^\sharp\right)^\top. \tag{39}$$

Recall that the characteristic function of a radom matrix $\mathbf{U}$ is denoted by $\phi(\boldsymbol{\theta}; \mathbf{U})$.

Lemma 7.1. *Under the foregoing assumptions and notation, it holds*

$$\mathbf{Z}^{(L+1)}(\mathbf{x}) \stackrel{\mathcal{L}}{=} \widetilde{\mathbf{N}}^{(L+1)}[\mathbf{H}^{(L)}(\mathbf{x})]^\sharp.$$

Hence

$$\phi(\boldsymbol{\theta}; \mathbf{Z}^{(L+1)}(\mathbf{x})) = \mathbb{E}\left[\exp\left(-\frac{1}{2}\mathrm{Tr}\big[\boldsymbol{\theta}^\top \mathbf{H}^{(L)}(\mathbf{x})\boldsymbol{\theta}\big]\right)\right], \quad \boldsymbol{\theta} \in \mathbb{R}^{d \times n_{L+1}}. \tag{40}$$

Lemma 7.2. *Let $\{V_{n,j}\}_{n\geq 1, j=1,\ldots,n}$ be a sequence of non-negative random variables such that*

$$\sum_{j=1}^{n} V_{n,j} \stackrel{\mathcal{L}}{\to} \mathrm{ID}(a, \rho), \quad \textit{as } n \to \infty$$

where $a \geq 0$ is a non-negative constant and ρ is a Lévy measure on $(0, \infty)$, and let $\{\xi_j\}_{j\geq 1}$ be a sequence of independent random vectors with law $\mathcal{N}_d(0, \mathbf{C})$, independent of $\{V_{n,j}\}_{n\geq 1, j=1,\ldots,n}$. If σ satisfies the assumption (i) of Theorem 4.1, then

$$\sum_{j=1}^{n} V_{n,j}\sigma(\xi_j)\sigma(\xi_j)^\top \stackrel{\mathcal{L}}{\to} a\mathbb{E}[\sigma(\xi_1)\sigma(\xi_1)^\top] + \sum_{j=1}^{N(0,\infty)} T_j \sigma(\xi_j)\sigma(\xi_j)^\top, \quad \textit{as } n \to \infty$$

where $N = \{T_j\}_{j\geq 1}$ is a Poisson process on $(0, \infty)$ with mean measure ρ, independent of $\{\xi_j\}_{j\geq 1}$.

Proof of Theorem 4.2. For ease of notation, we omit the explicit reference to the input matrix $\mathbf{x}$ in every symbol it occurs. Since vec and $\top$ are isometries of Euclidean spaces, we equivalently prove the theorem for the random matrices $\mathbf{Z}^{(L+1)}$ and $\mathbf{G}^{(L+1)}$ rather that for the random vectors $Z^{(L+1)}$ ad $G^{(L+1)}$. Let $\mathcal{S}_+^d$ be the space of the order d symmetric and positive semi-definite real matrices endowed with the Frobenius norm, and note that, for any fixed $\boldsymbol{\theta} \in \mathbb{R}^{d\times d}$, the mapping $\mathcal{S}_+^d \ni \mathbf{A} \mapsto \exp\left(-\frac{1}{2}\mathrm{Tr}\big[\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\big]\right)$ is real-valued, continuous and bounded. Hence, by (40) and the definition of weak convergence it holds that,

$$\mathbf{H}^{(L)} \stackrel{\pi\mathcal{L}}{\longrightarrow} \mathbf{K}^{(L+1)} \Longrightarrow \pi\overrightarrow{\lim}\phi(\boldsymbol{\theta}; \mathbf{Z}^{(L+1)}) = \mathbb{E}\left[\exp\left(-\frac{1}{2}\mathrm{Tr}\big[\boldsymbol{\theta}^\top \mathbf{K}^{(L+1)}\boldsymbol{\theta}\big]\right)\right] = \phi(\boldsymbol{\theta}; \mathbf{G}^{(L+1)}).$$

Therefore to prove the theorem it suffices to prove that $\mathbf{H}^{(L)} \xhookrightarrow{\pi\mathcal{L}} \mathbf{K}^{(L+1)}$, for every permutation π of $[L]$. We proceed by induction on the number of layers L, the basis of the induction being $\mathbf{H}^{(1)} \xhookrightarrow{\pi\mathcal{L}} \mathbf{K}^{(2)}$ which reduces to

$$\mathbf{H}^{(1)} \xrightarrow[n_1\to\infty]{\mathcal{L}} \mathbf{K}^{(2)} \tag{41}$$

because, for $L = 1$, the unique permutation of $[L]$ is the identity. We defer the proof of (41) after the proof of the inductive step. So, let $L \geq 3$ and assume

$$\mathbf{H}^{(L-1)} \xhookrightarrow{\tau\mathcal{L}} \mathbf{K}^{(L)} \quad \text{for every permutation } \tau \text{ of } [L-1]. \tag{42}$$

Clearly, by the Continuity Theorem the claim follows if we prove

$$\pi \lim \phi(\boldsymbol{\theta}; \mathbf{H}^{(L)}) = \phi(\boldsymbol{\theta}; \mathbf{K}^{(L+1)}) \quad \forall \boldsymbol{\theta} \in \mathbb{R}^{d\times d} \text{ and every permutation } \pi \text{ of } [L], \tag{43}$$

assuming (41) (to be proved) and (42). Let π be any permutation of $[L]$ and let $j^* = \pi^{-1}(L)$. Hence $(\pi \preceq j^* - 1) = (\tau \preceq j^* - 1)$ for some permutation τ of $[L-1]$. Since the mapping $\mathbf{A} \to \mathbf{A}^\sharp$ is continuous on $\mathcal{S}^d_+$, the inductive step (42) yields $[\mathbf{H}^{(L-1)}]^\sharp \xhookrightarrow{(\pi\preceq j^*-1)\mathcal{L}} \mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}$. Since, by construction, $\mathbf{H}^{(L-1)}$ is independent of $\mathbf{V}^{(L)}$ and $\widetilde{\mathbf{N}}^{(L)}$ and σ is continuous, by (38) it follows that

$$\sigma\left(\widetilde{\mathbf{N}}^{(L)}[\mathbf{H}^{(L-1)}]^\sharp\right)^\top \mathbf{V}^{(L)} \sigma\left(\widetilde{\mathbf{N}}^{(L)}[\mathbf{H}^{(\ell-1)}]^\sharp\right) \xhookrightarrow{(\pi\preceq j^*-1)\mathcal{L}} \sigma\left(\widetilde{\mathbf{N}}^{(L)}[\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}]^\sharp\right)^\top \mathbf{V}^{(L)} \sigma\left(\widetilde{\mathbf{N}}^{(L)}[\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}]^\sharp\right).$$

Note that also $[\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}]^\sharp$ is independent of $\mathbf{V}^{(L)}$ and $\widetilde{\mathbf{N}}^{(L)}$. Therefore, if $\boldsymbol{\Xi} := [\xi_1, \ldots, \xi_{n_L}] \in \mathbb{R}^{d\times n}$ is the matrix defined by $\boldsymbol{\Xi} = \widetilde{\mathbf{N}}^{(L)}[\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}]^\sharp$, then $\boldsymbol{\Xi}$ is independent of $\mathbf{V}^{(L)}$ and, given $[\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}]^\sharp$, $\{\xi_1, \ldots, \xi_{n_L}\}$ are independent and identically distributed random vectors in $\mathbb{R}^d$ with law $\mathcal{N}_d(0, \mathbf{K}^{(L)}_{(\pi\preceq j^*-1)})$ each. Consequently, in view of (37), one has

$$(\pi \preceq j^* - 1)\overrightarrow{\lim}\phi(\boldsymbol{\theta}; \mathbf{H}^{(L)}) = \mathbb{E}\left[\exp\left(\mathbf{i}\langle\boldsymbol{\theta}, C_B \mathbf{1}_d \mathbf{1}_d^\top + C_W \sum_{j=1}^{n_L} V^{(L)}_{n_L j}\sigma(\xi_j)\sigma(\xi_j)^\top\rangle_F\right)\right].$$

On the other hand, with $n_{\pi(j^*)} = n_L$, by Lemma 7.2 we have that

$$C_B \mathbf{1}_d \mathbf{1}_d^\top + C_W \sum_{j=1}^{n_L} V^{(L)}_{n_L j}\sigma(\xi_j)\sigma(\xi_j)^\top \xrightarrow[n_L\to\infty]{\mathcal{L}} \mathbf{K}^{(L+1)}_{(\pi\preceq j^*)}$$

where, referring to (6), $\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)}$ is defined similarly to $\mathbf{K}^{(L+1)}$ but with the ξ_j's in place of the $\zeta^{(L)}_j$'s and $\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}$ in place of $\mathbf{K}^{(L)}$, i.e.

$$\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)} := C_B 1_d 1_d^\top + C_W\left(a^{(L)}\mathbb{E}[\sigma(\xi_1)\sigma(\xi_1)^\top \,|\, \mathbf{K}^{(L)}_{(\pi\preceq j^*-1)}] + \sum_{j=1}^{N_L((0,\infty))} T^{(L)}_j \sigma(\xi_j)\sigma(\xi_j)^\top\right). \tag{44}$$

Since $L = \pi(j^*)$ we have proved that

$$(\pi \preceq j^*)\overrightarrow{\lim}\phi(\boldsymbol{\theta}; \mathbf{H}^L) = \lim_{n_L\to\infty}\left[(\pi \preceq j^* - 1)\overrightarrow{\lim}\phi(\boldsymbol{\theta}; \mathbf{H}^{(L)})\right] = \mathbb{E}\left[\exp\left(\mathbf{i}\mathrm{Tr}\left[\boldsymbol{\theta}^\top \mathbf{K}^{(L+1)}_{(\pi\preceq j^*)}\right]\right)\right]. \tag{45}$$

If $j^* = L$, then we are done. Indeed, since $(\pi \preceq j^*) = \pi$ and $(\pi \preceq j^* - 1) = \tau$ for some permutation τ of $[L-1]$, it follows that $\mathbf{K}^{(L)}_{(\pi\preceq j^*-1)} = \mathbf{K}^{(L)}$ by (42). Hence $\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)} = \mathbf{K}^{(L+1)}$ by (44), after recalling (6). We thus assume $j^* \neq L$ and prove that $\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)} \xhookrightarrow{(\pi\succeq j^*+1)\mathcal{L}} \mathbf{K}^{(L+1)}$. Since $(\pi \succeq j^*+1) = (\tau \succeq j^*+1)$ for some permutation τ of $[L]$, it follows that $\mathbf{K}^{(L)}_{(\pi\preceq j^*)} \xhookrightarrow{(\pi\succeq j^*+1)\mathcal{L}} \mathbf{K}^{(L)}$ by (42) . A simple computation with characteristic functions shows

$$(\xi_1, \ldots, \xi_{n_L}) \xhookrightarrow{(\pi\succeq j^*+1)\mathcal{L}} (\zeta^{(L)}_1, \ldots, \zeta^{(L)}_{n_L}). \tag{46}$$

By (46) and Skorohod's representation theorem, there exists a probability space, say $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, and copies of the random variables ξ and ζ (still denoted, by a little abuse of notation, ξ and ζ, respectively), where the convergence is $\tilde{\mathbb{P}}$-a.s. Recall that if $\{X_n\}_{n\geq 1}$ is a sequence of d-dimensional Gaussian random vectors and X is a d-dimensional Gaussian random vector, one has that if $X_n \xrightarrow{\mathcal{L}} X$, then, for any $p_1, \dots, p_k \geq 0$, $0 \leq k \leq d$, $\{i_1, \dots, i_k\} \subseteq \{1, \dots, d\}$ it holds

$$\mathbb{E}|X_n(i_1)|^{p_1} \dots |X_n(i_k)|^{p_k} \to \mathbb{E}|X(i_1)|^{p_1} \dots |X(i_k)|^{p_k}.$$

Since σ is continuous and is either bounded or with subpolynomial growth, using the dominated convergence theorem in the bounded case and the generalized dominated convergence theorem in the case of a subpolynomial growth, one has that

$$\tilde{\mathbb{E}}\left[\sigma(\xi_1)\sigma(\xi_1)^\top \,|\, \mathbf{K}^{(L)}_{(\pi\preceq j^*)}\right] \xrightarrow{(\pi\succeq j^*+1)\tilde{\mathbb{P}}} \tilde{\mathbb{E}}\left[\sigma(\zeta_1^{(L)}\sigma(\zeta_1^{(L)})^\top \,|\, \mathbf{K}^{(L)}\right].$$

Using the continuity of σ, one also has that

$$\sum_{j=1}^{N_L((0,\infty))} T_j^{(L)}\sigma(\xi_j)\sigma(\xi_j)^\top \xrightarrow{(\pi\succeq j^*+1)\tilde{\mathbb{P}}} \sum_{j=1}^{N_L((0,\infty))} T_j^{(L)}\sigma(\zeta_j^{(L)})\sigma(\zeta_j^{(L)})^\top$$

where, with a little abuse of notation we still denote by $N_L = \{T_j^{(L)}\}$ a copy of $N_L = \{T_j^{(L)}\}$ on $\tilde{\Omega}$. This convergence is clear if $\rho^{(L)}((0,\infty)) < \infty$. In the case when σ is bounded this convergence holds by the dominated convergence theorem. Indeed, being $\rho^{(L)}$ a Lévy measure, one has that the versions on $\tilde{\Omega}$ of the random variables

$$\sum_{j=1}^{N_L((0,1))} T_j^{(L)} \quad \text{and} \quad \sum_{j:\, T_j^{(L)}\in[1,\infty)} T_j^{(L)}$$

are finite $\tilde{\mathbb{P}}$-a.s. (the first sum has a finite mean and the second sum has a finite number of addends). So, by (44), the version of $\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)}$ on $\tilde{\Omega}$ converges to the version of $\mathbf{K}^{(L+1)}$ on $\tilde{\Omega}$, $\tilde{\mathbb{P}}$-a.s.. Therefore, by the dominated convergence theorem, in view of (3) and (45), one has

$$\begin{aligned}
\pi\overrightarrow{\lim}\phi(\boldsymbol{\theta};\mathbf{H}^{(L)}) &= (\pi \succeq j^*+1)\overrightarrow{\lim}\left[(\pi \preceq j^*)\overrightarrow{\lim}\phi(\boldsymbol{\theta};\mathbf{H}^{L})\right] \\
&= (\pi \succeq j^*+1)\overrightarrow{\lim}\mathbb{E}\left[\exp\left(\mathbf{i}\mathrm{Tr}\left[\boldsymbol{\theta}^\top\mathbf{K}^{(L+1)}_{(\pi\preceq j^*)}\right]\right)\right] \\
&= \mathbb{E}\left[\exp\left(\mathbf{i}\mathrm{Tr}\left[\boldsymbol{\theta}^\top\mathbf{K}^{(L+1)}\right]\right)\right]
\end{aligned}$$

and (43) is proved. It remains to prove (41). By construction $\mathbf{H}^{(0)} = \mathbf{K}^{(1)}$. Hence, with $\ell = L = 1$, by (39), for $j = 1\dots, n_1$, it holds that $\gamma_j^{(0)} \stackrel{\mathcal{L}}{=} \zeta_j^{(1)}$, where $\zeta_j^{(1)}$ has law $\mathcal{N}_d(0, \mathbf{K}^{(1)})$. Consequently, by (37), one has

$$\mathbf{H}^{(1)} = C_B 1_d 1_d^\top + C_W \sum_{j=1}^{n_1} V^{(1)}_{n_1 j}\sigma(\zeta_j^{(1)})\sigma(\zeta_j^{(1)})^\top.$$

Hence, after recalling (6), by Lemma 7.2, $\mathbf{K}^{(2)}$ is the limit in law of $\mathbf{H}^{(1)}$ as $n_1 \to \infty$. The proof is thus completed. □

Acknowledgments. The authors acknowledge the support of the Istituto Nazionale di Alta Matematica - GNAMPA group.

References

[1] S Adams, S. Patanè, M. Lahijanian and L. Laurenti Finite Neural Networks as Mixtures of Gaussian Processes: From Provable Error Bounds to Prior Selection. *arXiv:2407.18707*, 2024.

[2] L. Ambrosio, N. Fusco and D. Pallara. *Functions of Bounded Variations and Free Discontinuity Problems* Oxford University Press, London, 2000.

[3] N. Apollonio, D. De Canditiis, G. Franzina, P. Stolfi and G.L. Torrisi. Normal Approximation of Random Gaussian Neural Networks. *Stochastic Systems,* to appear, 2025.

[4] K. Balasubramanian, L. Goldstein and A. Salim. Gaussian Random Field Approximation via Stein's Method with Applications to Wide Random Neural Networks. *Appl. Comput. Harm. Anal.,* 72, 2024.

[5] K. Balasubramanian and N. Ross. Finite-Dimensional Gaussian Approximation for Deep Neural Networks: Universality in Random Weights. *arXiv:2507.12686*, 2025.

[6] A. Basteri and D. Trevisan. Quantitative Gaussian Approximation of Randomly Initialized Deep Neural Networks. *Machine Learning,* 113, 6373-6393, 2024.

[7] A. Bordino, S. Favaro and S. Fortini. Non-Asymptotic Approximations of Gaussian Neural Networks Via Second-Order Poincaré Inequalities. In *Proceedings of Machine Learning Research* (AABI24), 2024. 2023

[8] L. Carvalho, J.L. Costa, J. Mourão and G. Oliveira. The Positivity of the Neural Tangent Kernel. *SIAM Journal on Mathematics of Data Science,* 7, 2025.

[9] F. Caporali, S. Favaro and D. Trevisan. Student-t Processes as Infinite-Width Limits of Posterior Bayesian Neural Networks. *arXiv:2502.0427*, 2025.

[10] L. Celli and G. Peccati. Entropic Bounds for Conditionally Gaussian Vectors and Application to Neural Networks. *arXiv:2504.08335*, 2025.

[11] R. Eldan, D. Mikulincer and T. Schramm,. Non-asymptotic Approximations of Neural Networks by Gaussian Processes. *In Conference on Learning Theory*, pp. 1754-1775, 2021.

[12] L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions.* CRC Press, Boca Raton, Florida, 1992.

[13] S. Favaro, S. Fortini and S. Peluchetti. Deep Stable Neural Networks: Large-Width Asymptotics and Convergence Rates. *Bernoulli*, 29, 2574-2597, 2023.

[14] S. Favaro, B. Hanin, D. Marinucci, I. Nourdin and G. Peccati. Quantitative CLTs in Deep Neural Networks. *Probability Theory and Related Fields,* to appear, 2025.

[15] H. Federer. *Geometric Measure Theory.* Springer, New York, 1969.

[16] V. Fortuin, A. Garriga-Alonso, S.W. Ober, F. Wenzel, G. Rätsch, R.E. Turner, M. van der Wilk and L. Aitchison Bayesian Neural Network Priors Revisited *International Conference on Learning Representations*, 2022.

[17] I. Goodfellow, Y. Bengio and A. Courville. *Deep Learning* MIT Press, 2016.

[18] K. Hajjar, L. Chizat and C. Giraud. Training Integrable Parameterizations of Deep Neural Networks in the Infinite-Width Limit. *Journal of Machine Learning Research,* 25, 2024.

[19] B. Hanin. Random Neural Networks in the Infinite Width Limit as Gaussian Processes. *The Annals of Applied Probability,* 33, 4798-4819, 2023.

[20] C. Hirsch and D. Willhalm. Large Deviations of One-Hidden-Layer Neural Networks. *Stochastics and Dynamics,* 24, 2024.

[21] J. Hron, Y. Bahri, R. Novak, J. Pennington, J. Sohl-Dickstein. Exact Posterior Distributions of Wide Bayesian Neural Networks. *In: Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[22] P. Izmailov, S. Vikram, M.D. Hoffman and A.G. Wilson. What Are Bayesian Neural Network Posteriors Really Like? *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[23] P. Jung, H. Lee, J. Lee and H. Yang. α-Stable Convergence of Heavy-Tailed Infinitely Wide Neural Networks. *Advances in Applied Probability*, 55, 1415-1441, 2023.

[24] P. Lancaster, M. Tismenetsky. *Theory of Matrices: With Applications.* San Diego University Press, San Diego, 1985.

[25] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, J. Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[26] H. Lee, F. Ayed, P. Jung, J. Lee, H. Yang and F. Caron. Deep Neural Networks with Dependent Weights: Gaussian Process Mixture Limit, Heavy Tails, Sparsity and Compressibility. *Journal of Machine Learning Research,* 24, 78 pp., 2023.

[27] C. Macci, B. Pacchiarotti and G. L. Torrisi. Large and Moderate Deviations for Gaussian Neural Networks. *Journal of Applied Probability*, Published online 2025:1-20. doi:10.1017/jpr.2025.10024

[28] K.V. Mardia, J.T. Kent and J.M. Bibby. *Multivariate Analysis Probability and Mathematical Statistics.* Academic Press, Waltham, 1995.

[29] C.H. Martin and M.W. Mahoney. Traditional and Heavy-Tailed Self Regularization in Neural Network Models. *International Conference on Machine Learning*, 2019.

[30] A.G.d. G. Matthews, J. Hron, M. Rowland, R.E. Turner, Z. Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[31] S. Mei, A. Montanari and P.M. Nguyen. A Mean Field View of the Landscape of Two-Layers Neural Network. *Proceedings of the National Academy of Sciences (PNAS)*, 2018.

[32] R.M. Neal. Bayesian Learning for Neural Networks. *PhD Thesis*, Department of Computer Science, University of Toronto, 1995.

[33] R.M. Neal. Priors for Infinite Networks. *In: Bayesian Learning for Neural Networks. Lecture Notes in Statistics* , 118, 29-53, Springer, New York, 1996.

[34] L. Pezzetti, S. Favaro and S. Peluchetti. Function-space MCMC for Bayesian wide neural networks. *arXiv:2408.14325*, 2024.

[35] T. Rolski, H. Schmidli, V. Schmidt, J. Teugels. *Stochastic Processes for Insurance and Finance* Wiley, Chichester, 1999.

[36] K. Sato. *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press, Cambridge, 1999.

[37] D. Trevisan. Wide Deep Neural Networks with Gaussian Weights are Very Close to Gaussian Processes. *arXiv:2312.11737*, 2025.

[38] F. Wentzel, K. Roth, B.S. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton and S. Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? *Proceedings of the 37th International Conference on Machine Learning*, pp. 10248-10259, 2020.