Journal of Machine Learning Research 24 (2023) 1-78 Submitted 5/22; Revised 4/23; Published 9/23

Deep Neural Networks with Dependent Weights: Gaussian Process Mixture Limit, Heavy Tails, Sparsity and Compressibility

Hoil Lee HOIL.LEE@KAIST.AC.KR
Department of Mathematical Sciences, KAIST
Daejeon, South Korea

Fadhel Ayed FADHEL.AYED@GMAIL.COM
Huawei Technologies
Paris, France

Paul Jung PJUNG3@FORDHAM.EDU
Department of Mathematics, Fordham University
New York City, USA

Juho Lee JUHOLEE@KAIST.AC.KR
Kim Jaechul Graduate School of AI, KAIST
Daejeon, South Korea

Hongseok Yang HONGSEOK.YANG@KAIST.AC.KR
School of Computing and Kim Jaechul Graduate School of AI, KAIST
Discrete Mathematics Group, Institute for Basic Science (IBS)
Daejeon, South Korea

François Caron CARON@STATS.OX.AC.UK
Department of Statistics, University of Oxford
Oxford, United Kingdom

Editor: Daniel Roy

Abstract

This article studies the infinite-width limit of deep feedforward neural networks whose weights are dependent, and modelled via a mixture of Gaussian distributions. Each hidden node of the network is assigned a nonnegative random variable that controls the variance of the outgoing weights of that node. We make minimal assumptions on these per-node random variables: they are iid and their sum, in each layer, converges to some finite random variable in the infinite-width limit. Under this model, we show that each layer of the infinite-width neural network can be characterised by two simple quantities: a non-negative scalar parameter and a Lévy measure on the positive reals. If the scalar parameters are strictly positive and the Lévy measures are trivial at all hidden layers, then one recovers the classical Gaussian process (GP) limit, obtained with iid Gaussian weights. More interestingly, if the Lévy measure of at least one layer is non-trivial, we obtain a mixture of Gaussian processes (MoGP) in the large-width limit. The behaviour of the neural network in this regime is very different from the GP regime. One obtains correlated outputs, with non-Gaussian distributions, possibly with heavy tails. Additionally, we show that, in this regime, the weights are compressible, and some nodes have asymptotically non-negligible contributions, therefore representing important hidden features. Many sparsity-promoting neural network models can be recast as special cases of our approach, and we discuss their infinite-width limits; we also present an asymptotic analysis of the pruning error. We illustrate some of the benefits of the MoGP regime over the GP regime in terms of representation learning and compressibility on simulated, MNIST and Fashion MNIST datasets.

Keywords: Deep neural network, infinite-width, infinite divisibility, sparsity, compressibility, Gaussian process, triangular arrays, regular variation, pruning

1. Introduction

Two decades after the seminal work of Radford Neal (1996), the last few years have seen a renewed and growing interest in the analysis of (deep) neural networks, with random weights, in the infinite-width limit. When the weights are independently, identically distributed (iid) and suitably scaled Gaussian random variables, the random function associated to this random neural network converges to a Gaussian process (Neal, 1996; Lee et al., 2018; Matthews et al., 2018; Yang, 2019; Bracale et al., 2021). The connection to Gaussian processes has deepened our understanding of large neural networks, and motivated the use of Bayesian or kernel regression inference methods (Lee et al., 2018) or the development of kernel methods for training via gradient descent (Jacot et al., 2018) in the infinite-width limit.

While insightful, the Gaussian process connection also highlighted some of the limitations of large-width neural networks with iid Gaussian weights. As already noted by Neal (1995), *"with Gaussian priors the contributions of individual hidden units are all negligible, and consequently, these units do not represent 'hidden features' that capture important aspects of the data."* Additionally, the different dimensions of the output of the neural network become independent Gaussian processes in the infinite-width limit, which is generally undesirable. Finally, from a Bayesian perspective, the Gaussian independence assumption on weights is often seen as unrealistic: estimated weights of deep neural networks generally exhibit dependencies and heavy tails (Martin and Mahoney, 2019; Wenzel et al., 2020; Fortuin et al., 2021), and thus a family of prior distributions which allow for heavy tails is desirable. To alleviate some of these limitations, iid non-Gaussian random weights have been considered, either assuming stable (Neal, 1996; Der and Lee, 2006; Favaro et al., 2020), or more generally light-tailed/heavy-tailed distributions (Jung et al., 2023). However, due to the same iid assumption, some of the above limitations pertain, such as independence of the dimensions of the output.

We consider a more structured distribution on the weights of the neural network. We assume that weights emanating from a given node are dependent, where the dependency is captured via a scale mixture of Gaussians. More precisely, for a weight $W_{jk}^{(l+1)}$ between node $j = 1, \dots, p_l$ at hidden layer l and node $k = 1, \dots, p_{l+1}$ at hidden layer $l+1$, we assume that

$$W_{jk}^{(l+1)} = \sqrt{\lambda_{p_l,j}^{(l)}} V_{jk}^{(l+1)} \tag{1}$$

where $\lambda_{p_l,j}^{(l)}$, for $j = 1, \dots, p_l$, are nonnegative iid random variance parameters, one for each node $j = 1, \dots, p_l$ at layer l, and $V_{jk}^{(l+1)}$ are iid centred Gaussian random variables with variance $\sigma_v^2 > 0$. The per-node variance term $\lambda_{p_l,j}^{(l)}$ induces some dependency over the weights $W_{j1}^{(l+1)}, \dots, W_{jp_{l+1}}^{(l+1)}$ connected to node j. As we describe in the next paragraph, this assumption has been considered by a number of authors for training (finite) neural networks either (i) as a prior for Bayesian learning and pruning of neural networks, or (ii) as an implicit prior where a regularised empirical risk minimiser with group-sparse penalty is interpreted as a maximum a posteriori estimator, or (iii) as a random weight initialisation scheme for stochastic gradient descent.

A number of articles considered prior distributions of the form in Equation (1) for Bayesian learning of deep neural networks. Examples of distributions considered for the random variance $\lambda_{p_l,j}^{(l)}$ include the Bernoulli (Jantre et al., 2021), the horseshoe (Louizos et al., 2017; Ghosh et al., 2018, 2019; Popkes et al., 2019), the gamma (Scardapane et al., 2017; Wang et al., 2017), the inverse gamma (Ober and Aitchison, 2021), or the improper Jeffrey distributions (Louizos et al., 2017). See (Fortuin, 2021, Section 4.1) for a recent review. Distributions concentrated around 0, like the horseshoe, or with mass at 0, like the Bernoulli, favour more sparse-like representations, and they

have often been used for compression of deep neural networks, by pruning nodes based on the posterior distributions of the per-node variance parameter. Using a similar idea but with a slightly different formulation, Adamczewski and Park (2021) considered a joint Dirichlet distribution for the square root of the variances. In Section 6 and Appendix E.2, we discuss these examples in the context of our general framework.

These structured priors are also related to non-Bayesian estimators based on regularised empirical risk minimisation, where the estimator can be interpreted as a maximum a posteriori estimator under these priors. A typical example is the group lasso penalty on the weights of a neural network, used in a number of articles (Murray and Chiang, 2015; Scardapane et al., 2017; Wang et al., 2017; Ochiai et al., 2017), which can be interpreted as a negative log-prior on the weights when $\lambda_{p_l,j}^{(l)}$ follows a gamma distribution.

Finally, random weights of the form in Equation (1) have been used to initialise the weights in stochastic gradient descent algorithms, departing from the standard iid Gaussian initialisation commonly used for training deep neural networks (Glorot and Bengio, 2010). Blier et al. (2019) use per-node random learning rates in stochastic gradient descent. This is equivalent to using the prior in Equation (1) at initialisation, and then learning $V_{jk}^{(l+1)}$ while keeping the variances fixed after initialisation. A similar approach was considered by Wolinski et al. (2020b), but with deterministic variances.

As outlined above, neural networks with random weights of the form in Equation (1) have been extensively used in practice. A flurry of different distributions have been proposed for the random variance $\lambda_{p_l,j}^{(l)}$, and it is unclear which one we should choose in practice, and how this choice influences the properties of the resulting random neural network function.

The objective of this work is to analyse the infinite-width properties of feedforward neural networks with dependent weights of the form in Equation (1). Our work shows that the choice of the distribution of the per-node variance is crucial and can lead to fundamentally different infinite-width limits. Our main assumption is that, at each hidden layer l,

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \xrightarrow{\mathrm{d}} \Lambda^{(l)} \quad \text{as the width } p_l \to \infty, \tag{2}$$

where $\xrightarrow{\mathrm{d}}$ refers to convergence in distribution and $\Lambda^{(l)}$ is some nonnegative random variable, which may be constant. This assumption is natural as it implies that the activations and outputs of the neural network are almost surely finite in the infinite-width limit. Note that $\sum_{j=1}^{p_l} \mathrm{Var}\left(W_{jk}^{(l+1)} \middle| (\lambda_{p_l,j}^{(l)})_{j\geq 1}\right) = \sigma_v^2 \sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}$. Hence, the assumption in Equation (2) is similar to the commonly made assumption, in the iid case, that the sum of the variances of the incoming weights to a node converges to a constant in the infinite-width limit (Glorot and Bengio, 2010; He et al., 2015). The iid Gaussian case indeed arises as a special case by setting $\lambda_{p_l,j}^{(l)} = \frac{c}{p_l}$ for all $j = 1, \ldots, p_l$ for some $c > 0$. Note that $\Lambda^{(l)} = c$ is deterministic in this particular case.

The random variable $\Lambda^{(l)}$ is necessarily infinitely divisible (see Section 2), and parameterised by

(i) a location parameter $a^{(l)} \geq 0$ and
(ii) a Lévy measure $\rho^{(l)}$ on $(0, \infty)$.

We prove that, if $a^{(l)} > 0$ and the Lévy measures are trivially zero (that is $\int_0^\infty \rho^{(l)}(dx) = 0$) at all hidden layers l, then the limit is a Gaussian process (GP), as in the iid Gaussian case. As a consequence, all weights are uniformly small, with $\max_{j=1,\ldots,p_l} |W_{jk}^{(l+1)}| \to 0$ in probability. We show that this GP limit arises with a few models proposed in the literature, such as the group lasso (Scardapane et al., 2017; Wang et al., 2017) and inverse gamma (Ober and Aitchison, 2021) priors. These neural network models therefore are asymptotically equivalent to a model with iid Gaussian weights in the infinite-width limit.

More interestingly, if at least one of the Lévy measures is non-trivial, we obtain a very different behaviour, and the limit is now a *mixture of Gaussian processes* (MoGP), with a given random

covariance kernel. Under the MoGP regime, we show that the following results hold in the infinite-width limit, none of which hold for the iid Gaussian case.

- $\max_{j=1,\dots,p_l} |W_{jk}^{(l+1)}|$ converges in probability to a random variable which is not degenerately 0 (see Proposition 3). That is, some weights remain non-negligible asymptotically. It is natural to interpret this as being connected to nodes representing important hidden features.
- The different dimensions of the output remain dependent (see Theorems 8 and 16).
- The outputs are non-Gaussian, and may exhibit heavy tails depending on the behaviour of the Lévy measures at infinity (see Propositions 9 and 10).
- Pruning the network according to the variance parameter $\lambda_{p_l,j}^{(l)}$ at some level $\epsilon > 0$ sufficiently small, provides a finite, non-empty neural network with positive probability.[1] The resulting error associated to the pruned network can be related to the behaviour of the Lévy measure near 0 (see Corollary 13).
- The network is compressible: when pruning the network by removing a fixed proportion $(1 - \kappa) \in (0, 1)$ of nodes at each layer according to the variance parameter $\lambda_{p_l,j}^{(l)}$, the difference between the outputs of the pruned and unpruned networks converges to 0 in probability in the infinite-width limit (see Corollary 15).

Some illustrative examples. To give a sense of the range of results covered in this article, we now briefly present some illustrative examples in the case of a simple feedforward neural network with one hidden layer, d_{in}-dimensional input $\mathbf{x} = (x_1, \dots, x_{d_{\text{in}}})^T$, 2-dimensional output $(Z_1^{(2)}(\mathbf{x}; \mathbf{p}), Z_2^{(2)}(\mathbf{x}; \mathbf{p}))^T$, no bias, $\sigma_v = 1$ and rectified linear unit (ReLU) activation function. For $k = 1, 2$, the output is such that

$$Z_k^{(2)}(\mathbf{x}; p_1) = \sum_{j=1}^{p_1} \sqrt{\lambda_{p_1,j}^{(1)}} V_{jk}^{(2)} \max\left(0, \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} V_{ij}^{(1)} x_i\right).$$

More general deep neural networks and other examples are considered later in this article. As mentioned above, it is well known (see for instance (Lee et al., 2018)) that, if $\lambda_{p_1,j} = \frac{2}{p_1}$ (iid Gaussian weights, or He initialisation (He et al., 2015)), the outputs are asymptotically independent Gaussian processes with, for $k = 1, 2$,

$$\begin{pmatrix} Z_k^{(2)}(\mathbf{x}; p_1) \\ Z_k^{(2)}(\mathbf{x}'; p_1) \end{pmatrix} \xrightarrow{\text{d}} \mathcal{N}\left(0, \begin{pmatrix} \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}) & \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') \\ \mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') & \mathcal{K}^{(2)}(\mathbf{x}', \mathbf{x}') \end{pmatrix}\right) \text{ as } p_1 \to \infty \tag{3}$$

where the (deterministic) covariance kernel $\mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}')$ is defined by

$$\mathcal{K}^{(2)}(\mathbf{x}, \mathbf{x}') = \frac{\|\mathbf{x}\|\|\mathbf{x}'\|}{d_{\text{in}}} \times \frac{1}{\pi}\left(\sqrt{1 - \rho_{\mathbf{x},\mathbf{x}'}^2} + \left(\frac{\pi}{2} + \arcsin \rho_{\mathbf{x},\mathbf{x}'}\right)\rho_{\mathbf{x},\mathbf{x}'}\right), \tag{4}$$

with correlation $\rho_{\mathbf{x},\mathbf{x}'} = \frac{\mathbf{x}^T\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}$, see Appendix A.2 for background on ReLU kernels.

Consider now the following models for $p_1 \geq 2$:

(a) $\lambda_{p_1,j}^{(1)} \sim \text{IG}\left(2, \frac{2}{p_1}\right)$ (b) $\lambda_{p_1,j}^{(1)} \sim \text{Bernoulli}\left(\frac{2}{p_1}\right)$
(c) $\lambda_{p_1,j}^{(1)} \sim \text{Beta}\left(\frac{1}{p_1}, \frac{1}{2}\right)$ (d) $\lambda_{p_1,j}^{(1)} = \pi^2 \frac{U_j^2}{p_1^2}$ where $U_j \sim \text{Cauchy}_+(0, 1)$

where $\text{IG}(\beta_1, \beta_2)$ denotes the inverse gamma distribution with shape $\beta_1 > 0$ and scale $\beta_2 > 0$, and $\text{Cauchy}_+(0, 1)$ denotes the half-Cauchy distribution with pdf

$$f(u) = \frac{2}{\pi(1 + u^2)} \times \mathbf{1}_{\{u>0\}}. \tag{5}$$

1. Note that there is always some small probability of pruning everything and leaving an empty network.

Model	Limit process	Depend. outputs	Distribution of $Z_k^{(2)}(\mathbf{x}, p_1)$	Tail of $Z_k^{(2)}(\mathbf{x}, p_1)$	Number of active nodes	$\max \lvert W_{jk}^{(2)} \rvert \xrightarrow{\text{pr}} 0$	Tail of of $W_{jk}^{(2)}$	$(W_{(j)k}^{(2)})^2$ decrease in	Compressible
iid	GP	No	Gaussian	Expon.	∞	Yes	Expon.	–	No
(a)	GP	No	Gaussian	Expon.	∞	Yes	Expon.	–	No
(b)	MoGP	Yes	Compound Poisson	Expon.	Poisson(2)	No	Expon.	–	Yes
(c)	MoGP	Yes	Normal-gamma	Expon.	∞	No	Expon.	$O(e^{-cj})$	Yes
(d)	MoGP	Yes	Cauchy	Power-law	∞	No	Power-law	$O(j^{-2})$	Yes

Table 1: Summary of the properties of the neural network models for four different distributions on the per-node variances.

For all the above models (a-d), we have $\lambda_{p_1,j}^{(1)} \to 0$ in probability as $p_1 \to \infty$. For (a-c), $\mathbf{E}[\sum_j \lambda_{p_1,j}^{(1)}] \to 2$ as $p_1 \to \infty$ (the expectation is infinite for the horseshoe example (d)), as in the iid Gaussian case. However, the infinite-width limits are all very different.

Under the inverse gamma model (a), the infinite-width limit is the same as the iid Gaussian case. Under models (b-d), the infinite-width limit is a mixture of Gaussian processes, i.e. a Gaussian process with a random covariance kernel (see Theorem 16). These models illustrate some of the benefits of the MoGP regime. The outputs are now dependent in the infinite-width limit. The models (b-d) are compressible in the sense that the difference between the output of the pruned network and the output of the unpruned network vanishes in the infinite-width limit (see Theorem 5). This is not the case for the iid Gaussian model, nor for model (a). The weights as well as the outputs can have an exponential tail (b-c) or a power-law tail (d). The properties of the different models are summarised in Table 1. More details on these illustrative examples can be found in Appendix E.1.

Organisation of the article. In Section 2, we provide some background material on infinitely divisible random variables. The feedforward neural network model with dependent weights is described in Section 3, together with the asymptotic assumptions. We also show how the behaviour of the Lévy measure around zero and infinity tunes the properties of large and small weights. In Section 4, we give the asymptotic distribution of the outputs for a single input $\mathbf{x}$, in the case of ReLU-like activation functions. We discuss some of the implications of our result in terms of pruning and heavy tails, depending on the asymptotic properties of the model. In Section 5, the result is extended to multiple inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and general activation functions. In Section 6, we show how many models proposed in the literature can be formulated in our general framework, and present their limiting properties. In Section 7, we provide some illustrative experiments on Bayesian inference under this class of models, and in Section 8, we discuss related approaches. The Appendix contains the details of the illustrative example from above, further examples, most of the proofs, some additional background material and secondary lemmas. The code to reproduce the experiments is available at https://github.com/FadhelA/mogp.

Notations. For a random variable X, $X \sim F$ indicates that X is distributed according to F. For functions (or sequences) $a(x)$ and $b(x)$, we use the notation $a(x) \overset{x\to\infty}{\sim} b(x)$ for $\lim_{x\to\infty} a(x)/b(x) = 1$. The notation $\xrightarrow{\text{pr}}$ and $\xrightarrow{\text{d}}$ respectively mean 'convergence in probability' and 'convergence in distribution'. We also use the notation $X \overset{\text{d}}{=} Y$ to indicate that the two random variables X and Y have the same distribution. For two sequences of random variables X_n, Y_n, we write '$X_n \overset{n\to\infty}{\sim} Y_n$ in probability' for $X_n/Y_n \xrightarrow{\text{pr}} 1$ as $n \to \infty$.

2. Background Material on Infinitely Divisible Random Variables

A nonnegative random variable X is said to have an infinitely divisible distribution if, for every $n \in \mathbb{N}$, there exist iid nonnegative random variables $Y_{n1}, \ldots, Y_{nn}$ such that $X \overset{\text{d}}{=} \sum_{i=1}^n Y_{ni}$ (Sato, 1999). Examples of infinitely divisible nonnegative distributions are the lognormal, log-Cauchy, Pareto, gamma, betaprime, constant and positive stable distributions. (Appendix A.4 discusses the

last positive-stable case in detail.) If X is nonnegative and infinitely divisible, its distribution is uniquely characterised by a scalar $a \geq 0$ and a Lévy measure ρ on $(0, \infty)$ (that is, it is a Borel measure that satisfies $\int_0^\infty \min(1, x)\rho(dx) < \infty$). We write $X \sim \mathrm{ID}(a, \rho)$. The scalar a is a location parameter, and $X - a \sim \mathrm{ID}(0, \rho)$. The Lévy measure ρ may be

- Trivial, that is $\int_0^\infty \rho(dw) = 0$; in this case, $X = a$ is constant;
- Finite, that is $\int_0^\infty \rho(dw) < \infty$; in this case, $X \geq a$, with $\Pr(X = a) > 0$;
- Infinite, that is $\int_0^\infty \rho(dw) = \infty$; in this case, $X = a + Y$, where Y is an absolutely continuous random variable on $(0, \infty)$.

The Laplace transform is given, for any $t \geq 0$, by $\mathbf{E}[e^{-tX}] = e^{-ta-\psi(t)}$, where $\psi(t) := \int_0^\infty (1 - e^{-wt})\rho(dw)$. Infinitely divisible random variables are closely related to Poisson point processes. The random variable $X \sim \mathrm{ID}(a, \rho)$ admits the representation $X \overset{\mathrm{d}}{=} a + \sum_{i \geq 1} \xi_i$, where $(\xi_i)_{i \geq 1}$ are the points of a Poisson process on $(0, \infty)$ with mean measure ρ.

3. Statistical Model

3.1 Feedforward Neural Network

We consider a feedforward neural network (FFNN) with L hidden layers and $p_l \geq 1$ nodes at each layer l. We let $p_0 = d_{\text{in}}$ be the input dimension and $p_{L+1} = d_{\text{out}}$ be the output dimension. We write $\mathbf{p} = (p_1, \ldots, p_L)^T \in \mathbb{N}^L$. For l with $1 \leq l \leq L+1$, the pre-activation values at these nodes are given, for an input $\mathbf{x} = (x_1, \ldots, x_{d_{\text{in}}})^T \in \mathbb{R}^{d_{\text{in}}}$, recursively by

$$\begin{aligned} Z_k^{(1)}(\mathbf{x}; \mathbf{p}) &= \sum_{j=1}^{d_{\text{in}}} W_{jk}^{(1)} x_j + B_k^{(1)}, \\ Z_k^{(l)}(\mathbf{x}; \mathbf{p}) &= \sum_{j=1}^{p_{l-1}} W_{jk}^{(l)} \phi(Z_j^{(l-1)}(\mathbf{x}; \mathbf{p})) + B_k^{(l)}, \quad l \geq 2, \end{aligned} \tag{6}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the activation function, $W_{jk}^{(l)}$ is the weight between node j at layer $l-1$ and node k at layer l, and $B_k^{(l)}$ is the bias term of node k at layer l. The vector $(Z_1^{(L+1)}(\mathbf{x}; \mathbf{p}), \ldots, Z_{d_{\text{out}}}^{(L+1)}(\mathbf{x}; \mathbf{p}))^T$ is the output of the neural network for the input $\mathbf{x}$.

Let $\sigma_b \geq 0$. We assume that, for all $k \geq 1$ and $l = 1, \ldots, L+1$,

$$B_k^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_b^2\right) \tag{7}$$

if $\sigma_b > 0$, and $B_k^{(l)} = 0$ otherwise.

3.2 Distribution of the Weights

For $0 \leq l \leq L$, we assume that $W_{jk}^{(l+1)}$ follows a scale mixture of Gaussian distributions, with

$$W_{jk}^{(l+1)} = \sqrt{\lambda_{p_l, j}^{(l)}} V_{jk}^{(l+1)} \tag{8}$$

where

(a) for each layer $l = 0, \ldots, L$, $j = 1, \ldots, p_l$, and $k = 1, \ldots, p_{l+1}$,

$$V_{jk}^{(l+1)} \sim \mathcal{N}\left(0, \sigma_v^2\right) \text{ for some } \sigma_v > 0; \tag{9}$$

(b) $\lambda^{(0)}_{p_0,j} = \frac{1}{d_{\text{in}}}$, and for each layer $l = 1, \ldots, L$ and each node $j = 1, \ldots, p_l$ at layer l, $\lambda^{(l)}_{p_l,j} \geq 0$ is a (hidden) node variance parameter, with

$$\lambda^{(l)}_{p_l,j} \sim \mu^{(l)}_{p_l} \text{ for some probability distribution } \mu^{(l)}_{p_l} \text{ on } [0,\infty); \tag{10}$$

(c) all the random variables $\{\lambda^{(l)}_{p_l,j}, V^{(l)}_{jk}\}_{l,j,k}$ are assumed to be independent among themselves, and also with $\{B^{(l)}_k\}_{l,k}$.

3.3 Asymptotic Assumptions and Infinite Divisibility

As mentioned in the introduction, for any node k,

$$\sum_{j=1}^{p_l} \text{Var}\left(W^{(l+1)}_{jk} \,\middle|\, (\lambda^{(l)}_{p_l,j})_{j\geq 1}\right) = \sigma_v^2 \sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j}.$$

In order to have a.s. finite activations in the infinite-width limit, we need $\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j}$ to remain a.s. finite as p_l tends to infinity. To that end, recall from Equation (2) that

$$\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} \xrightarrow{\mathtt{d}} \Lambda^{(l)}$$

as $p_l \to \infty$, for some nonnegative random variable $\Lambda^{(l)}$. This natural and general assumption, together with the iid assumption, has two consequences.

(i) By (Kallenberg, 2002, Theorem 15.12), $\Lambda^{(l)}$ is necessarily an infinitely divisible random variable, characterised by a location parameter $a^{(l)} \geq 0$ and a Lévy measure $\rho^{(l)}$ on $(0,\infty)$. We express this by writing

$$\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} \xrightarrow{\mathtt{d}} \text{ID}(a^{(l)}, \rho^{(l)}). \tag{11}$$

(ii) By (Kallenberg, 2002, Lemma 15.13), we have $\lambda^{(l)}_{p_l,j} \xrightarrow{\mathtt{pr}} 0$ for any $j \geq 1$.

As we will show in the next subsections, the asymptotic properties of the neural network in the infinite-width limit are fully characterised by the activation function ϕ, the bias variance σ_b^2, the scaling factor σ_v^2 and the parameters $(a^{(l)}, \rho^{(l)})$ at each hidden layer $l = 1, \ldots, L$.

The following result shows that the infinite divisibility of the sum of per-node variances implies that the squared ℓ^2-norm of the vector of incoming weights of a node converges in distribution to an infinitely divisible random variable in the infinite-width limit. The proposition follows from Corollary 37 in the Appendix.

Proposition 1 *Let $l \in \{1, \ldots, L\}$. Assume Equations (8), (9) and (11) hold for some $\sigma_v > 0$, $a^{(l)} \geq 0$ and some Lévy measure $\rho^{(l)}$. Then, for any $k \geq 1$,*

$$\frac{1}{\sigma_v^2} \sum_{j=1}^{p_l} (W^{(l+1)}_{jk})^2 = \sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} \left(\frac{V^{(l+1)}_{jk}}{\sigma_v}\right)^2 \xrightarrow{\mathtt{d}} \text{ID}(a^{(l)}, \nu^{(l)})$$

where $\nu^{(l)}$ is a Lévy measure on $(0,\infty)$ defined by

$$\nu^{(l)}(dz) = \int_0^\infty \rho^{(l)}(dz/x)\, \text{Gamma}(x; 1/2, 1/2) dx, \tag{12}$$

where $\rho^{(l)}(dz/x)$ denotes the measure that assigns $\rho^{(l)}((a/x, b/x))$ to each interval $(a,b) \subseteq \mathbb{R}$.

Remark 2 *In the iid Gaussian case where* $\lambda^{(l)}_{p_l,j} = \frac{c}{p_l}$ *for some* c*, the sum of variances* $\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} = c$ *is constant and* $\frac{1}{\sigma_v^2}\sum_{j=1}^{p_l}(W^{(l+1)}_{jk})^2 = \frac{c}{p_l}\sum_{j=1}^{p_l}(V^{(l+1)}_{jk}/\sigma_v)^2 \xrightarrow{\text{pr}} c$*, where the convergence is by the law of large numbers.*

3.4 Properties of the Largest Weights in the Infinite-Width Limit

We discuss here some general structural properties of the FFNN in the infinite-width limit, depending on the parameters $a^{(l)}$ and $\rho^{(l)}$. In particular, we answer the following question: In which cases are the largest variances/weights of the FFNN asymptotically non-negligible?

We interpret a layer l to capture important features in the infinite-width limit if some of the per-node variances remain asymptotically non-negligible as $p_l \to \infty$. The following proposition, which follows from (Kallenberg, 2002, Theorem 15.29), shows that this arises whenever $\rho^{(l)}$ is a non-trivial Lévy measure.

Proposition 3 (Necessary and sufficient conditions for uniform convergence to 0) *Let* $l \in \{1, \dots, L\}$*. The following are equivalent:*

i) $\rho^{(l)}$ *is trivial;*

ii) $\max_j \lambda^{(l)}_{p_l,j} \xrightarrow{\text{pr}} 0$*;*

iii) for every $k \geq 1$*,* $\max_j |W^{(l+1)}_{jk}| \xrightarrow{\text{pr}} 0$*.*

The next proposition goes a bit further and describes the asymptotic distribution of the extreme weights. For a Lévy measure ρ on $(0, \infty)$, define the tail Lévy measure

$$\overline{\rho}(x) := \int_{(x,\infty)} \rho(dw) \text{ for all } x > 0.$$

For all $u > 0$, let $\overline{\rho}^{-1}(u) := \inf\{x > 0 : \overline{\rho}(x) < u\}$ denote the generalised inverse of $\overline{\rho}$, called the inverse tail Lévy intensity of ρ. Note that both $\overline{\rho}$ and $\overline{\rho}^{-1}$ are non-increasing functions, and are both equal to zero if ρ is trivial. The following proposition is a direct corollary of Proposition 30 in the Appendix and of Proposition 1 in the main text.

Proposition 4 (Extremes of the variances and weights) *Consider* $l \in \{1, \dots, L\}$*, and let* $\lambda^{(l)}_{p_l,(1)} \geq \lambda^{(l)}_{p_l,(2)} \geq \dots$ *be the order statistics of the per-node variances. Then, for any* $k \geq 1$*, as* $p_l \to \infty$*,*

$$\lambda^{(l)}_{p_l,(k)} \xrightarrow{\text{pr}} 0 \quad \text{if } \rho^{(l)} \text{ is trivial}; \qquad \lambda^{(l)}_{p_l,(k)} \xrightarrow{\text{d}} (\overline{\rho}^{(l)})^{-1}(G_k) \quad \text{otherwise},$$

where $G_k \sim \text{Gamma}(k, 1)$*. Here* $(\overline{\rho}^{(l)})^{-1}(G_k)$ *is a nonnegative random variable, non-degenerate at 0 if the Lévy measure is non-trivial. Additionally, let* $W^{(l+1)}_{(1),m} \geq W^{(l+1)}_{(2),m} \geq \dots$ *be the order statistics of the incoming weights of node* m *at layer* $l+1$*. Similarly, we have*

$$(W^{(l+1)}_{(k),m})^2 \xrightarrow{\text{pr}} 0 \quad \text{if } \rho^{(l)} \text{ (hence } \nu^{(l)}\text{) is trivial}; \qquad (W^{(l+1)}_{(k),m})^2 \xrightarrow{\text{d}} \sigma_v^2 \times (\overline{\nu}^{(l)})^{-1}(G_k) \quad \text{otherwise},$$

where $(\overline{\nu}^{(l)})^{-1}$ *is the inverse tail Lévy intensity of the measure* $\nu^{(l)}$ *defined in Equation* (12)*.*

What about the properties of small weights? One answer is given in Appendix D.1.

3.5 Compressibility of the Neural Network

About a decade ago, Gribonval et al. (2012) established a connection between heavy tails and compressibility in the compressed sensing literature. Recently, a series of works (Arora et al., 2018; Suzuki et al., 2019; Kuhn et al., 2021; Suzuki et al., 2020) have shown that the compressibility of a neural network is related to how well the network generalises, both from a theoretical and an empirical point of view. These two lines of works were brought together by Shin (2021); Barsbey et al. (2021), who proposed theoretical frameworks to establish a direct connection among the heavy tail index of the distribution of the weights of a neural network, the compressibility of the network and its generalisation properties. In the setting of our model, these studies on compressibility can be extended from the heavy-tailed case to the much larger class of models for which there is a non-trivial Lévy measure of the limiting infinitely divisible random variable of Equation (2).

Let $v_{p,(1)} \geq v_{p,(2)} \geq \cdots \geq v_{p,(p)}$ be the coordinates, reordered by size, of $\mathbf{v}(p) = (v_{p,1}, \ldots, v_{p,p})$. Motivated by similar notions in (Gribonval et al., 2012), we say that a sequence $(\mathbf{v}(p))_p$ is ℓ^2-compressible as $p \to \infty$ if for any $\kappa \in (0,1)$,

$$\lim_{p\to\infty} \frac{\sum_{j=1}^{p} \mathbf{1}_{\{v_{p,j} \leq v_{p,(\lfloor \kappa p \rfloor)}\}} v_{p,j}^2}{\sum_{j=1}^{p} v_{p,j}^2} = 0. \tag{13}$$

If $v_{p,i} \neq v_{p,j}$ when $i \neq j$, the indicator $\mathbf{1}_{\{v_{p,j} \leq v_{p,(\lfloor \kappa p \rfloor)}\}}$ retains the top κ-proportion of $v_{p,j}^2$ values.

To place this in the context of neural networks, we will say that layer l is compressible if Equation (13) holds in probability for the ℓ^2-norms of vectors of outgoing weights, for all nodes j in layer l. More precisely, for any $j = 1, \ldots, p_l$, denote the squared norm of the outgoing weights of the hidden node j at layer l by

$$T_j^{(l+1)} := \|W_{j,:}^{(l+1)}\|^2 = \sum_{k=1}^{p_{l+1}} \lambda_{p_l,j}^{(l)} (V_{j,k}^{(l+1)})^2$$

and let $T_{(1)}^{(l+1)} \geq \ldots \geq T_{(p_l)}^{(l+1)}$ denote the ordered values. Then, layer l is ℓ^2-norm-compressible if for every $\kappa \in (0,1)$,

$$\frac{\sum_{j=1}^{p_l} \mathbf{1}_{\{T_j^{(l+1)} \leq T_{(\lfloor \kappa p_l \rfloor)}^{(l+1)}\}} T_j^{(l+1)}}{\sum_{j=1}^{p_l} T_j^{(l+1)}} \xrightarrow{\text{pr}} 0 \text{ as } p_l \to \infty. \tag{14}$$

In our model, compressible layers are easily characterised simply by the value of $a^{(l)}$ as our next result shows.

Theorem 5 (Characterisation of compressibility) *For each layer l with $1 \leq l \leq L$, if $a^{(l)} = 0$, then for all $\kappa \in (0,1)$,*

$$\frac{\sum_{j=1}^{p_l} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}\}} \lambda_{p_l,j}^{(l)}}{\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}} \xrightarrow{\text{pr}} 0 \text{ as } p_l \to \infty, \tag{15}$$

where $\lambda_{p_l,(1)}^{(l)} \geq \lambda_{p_l,(2)}^{(l)} \geq \ldots \geq \lambda_{p_l,(p_l)}^{(l)}$ are the ordered per-node variance terms. In such a case, Equation (14) holds so that layer l is ℓ^2-norm-compressible.

3.6 Heavy Tail and Power-Law Properties of the Variances and Weights

A random variable X has a regularly varying tail if $\Pr(X > x) \overset{x\to\infty}{\sim} L(x)x^{-\tau}$ for some power-law exponent $\tau > 0$ and some slowly varying function L, that is, a function satisfying $L(\gamma x)/L(x) \to 1$

as $x \to \infty$ for all $\gamma > 0$. The simplest slowly varying function is the constant function $L(x) = c > 0$, and in this case we say that X has a power-law tail; to simplify the presentation, we restrict the presentation to this case here. The next proposition shows that, if the tail Lévy intensity decays polynomially at infinity, then the extremes of the per-node variance parameters and of the weights have power-law tails asymptotically.

Proposition 6 (Power law properties of the variances and weights) *Assume that for some $\tau > 0$ and some constant $c > 0$, $\overline{\rho}^{(l)}(x) \overset{x\to\infty}{\sim} cx^{-\tau}$. Then, for any $k, m \geq 1$,*

$$\lim_{p_l \to \infty} \Pr(\lambda^{(l)}_{p_l,(k)} > x) \overset{x\to\infty}{\sim} \frac{\overline{\rho}^{(l)}(x)^k}{k!} \overset{x\to\infty}{\sim} \frac{c^k}{k!} x^{-k\tau} \tag{16}$$

$$\lim_{p_l \to \infty} \Pr(|W^{(l)}_{(k),m}| > x) \overset{x\to\infty}{\sim} \frac{\overline{\nu}^{(l)}(x^2/\sigma_v^2)^k}{k!} \overset{x\to\infty}{\sim} \frac{\left(\frac{2^\tau \Gamma(\tau+1/2)}{\sqrt{\pi}} (\sigma_v^2)^\tau c\right)^k}{k!} x^{-2k\tau} \tag{17}$$

where $\overline{\nu}^{(l)}$ is the tail Lévy intensity of the measure $\nu^{(l)}$ defined in Equation (12).

4. Infinite-Width Limit for a Single Input for Homogeneous Activation Functions

Definition 7 *A function $\phi : \mathbb{R} \to \mathbb{R}$ is positive homogeneous if and only if $\phi(\gamma x) = \gamma\phi(x)$ for all $\gamma > 0$ and $x \in \mathbb{R}$.*

The following standard activation functions are positive homogeneous:

$$\phi(x) = x \qquad \text{[Linear]} \tag{18}$$

$$\phi(x) = \max(x, 0) \qquad \text{[ReLU]} \tag{19}$$

$$\phi(x) = \begin{cases} x & x > 0 \\ \beta x & x \leq 0 \end{cases} \qquad \text{[Leaky ReLU]} \tag{20}$$

for some $\beta > 0$. Note that the tanh and sigmoid functions are not positive homogeneous. We present later in Theorem 16 more general assumptions that include these two cases.

4.1 Statement of the Main Theorem

We consider one FFNN for each $\mathbf{p} \in \mathbb{N}^L$. The following result is stated for positive homogeneous activation functions, which include many important particular cases, in particular the ReLU. A similar result holds under more general assumptions on ϕ. See Theorem 16.

Theorem 8 (Single input case, ReLU-type activation) *Consider the feedforward neural network model defined by Equations* (6) *to* (10). *Assume that the activation function ϕ is positive homogeneous and that, for all hidden layers $l = 1, \ldots, L$, we have*

$$\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} \overset{\mathrm{d}}{\to} \mathrm{ID}(a^{(l)}, \rho^{(l)}) \text{ as } p_l \to \infty$$

for some $a^{(l)} \geq 0$ and some Lévy measure $\rho^{(l)}$. Then, as $\min(p_1, \ldots, p_L) \to \infty$, for any $m \geq 1$, any layer $l = 1, \ldots, L+1$ and any input $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$,

$$\left(Z^{(l)}_1(\mathbf{x}; \mathbf{p}), \ldots, Z^{(l)}_m(\mathbf{x}; \mathbf{p})\right) \overset{\mathrm{d}}{\to} \mathbf{E}\left[\bigotimes_{k=1,\ldots,m} \mathcal{N}\left(0, \Sigma^{(l)}(\mathbf{x})\right)\right]. \tag{21}$$

Here, for each $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, $\left(\Sigma^{(1)}(\mathbf{x}), \ldots, \Sigma^{(L+1)}(\mathbf{x})\right)$ *is a Markov sequence of nonnegative random variables, defined recursively via the following stochastic recurrence equations:*

$$\Sigma^{(1)}(\mathbf{x}) := \sigma_b^2 + (\sigma_v^2 \|\mathbf{x}\|^2 / d_{\text{in}}) \tag{22}$$

$$\Sigma^{(l)}(\mathbf{x}) := \sigma_b^2 + \sigma_v^2 S^{(l-1)} \Sigma^{(l-1)}(\mathbf{x}) \quad \text{for } l = 2, \ldots, L+1, \tag{23}$$

where $S^{(1)}, \ldots, S^{(L)}$ *are independent random variables which additionally do not depend on the input* $\mathbf{x}$. *Moreover,* $S^{(l)} \sim \mathrm{ID}(c^{(l)}, \eta^{(l)})$ *where* $\eta^{(l)}$ *is a Lévy measure on* $(0, \infty)$ *with tail Lévy intensity*

$$\overline{\eta}^{(l)}(x) := \int_{\{z:\phi(z)\neq 0\}} \overline{\rho}^{(l)}(x/\phi(z)^2)\varphi(z)dz$$

when φ *denotes the pdf of the standard normal distribution, and* $c^{(l)}$ *is a nonnegative scalar defined by*

$$c^{(l)} := a^{(l)} \int_{-\infty}^{\infty} \phi(z)^2 \varphi(z) dz.$$

Example 1 *Recall that* $\overline{\nu}(x) = \int_0^\infty \overline{\rho}(x/z)\,\mathrm{Gamma}(z; 1/2, 1/2)dz$ *is the tail Lévy intensity of the Lévy measure in Equation* (12) *associated to the sum of the squares of the weights. For the linear activation function in Equation* (18)*, we have*

$$c^{(l)} = a^{(l)}, \qquad \overline{\eta}^{(l)}(x) = \overline{\nu}^{(l)}(x). \tag{24}$$

For the ReLU activation function in Equation (19)*, we have*

$$c^{(l)} = a^{(l)}/2, \qquad \overline{\eta}^{(l)}(x) = \overline{\nu}^{(l)}(x)/2. \tag{25}$$

For the leaky ReLU activation function in Equation (20)*, we have*

$$c^{(l)} = a^{(l)}(\beta^2 + 1)/2, \qquad \overline{\eta}^{(l)}(x) = (\overline{\nu}^{(l)}(x) + \overline{\nu}^{(l)}(x/\beta^2))/2. \tag{26}$$

4.2 Proof of Theorem 8

Denote $\Sigma^{(1)}(\mathbf{x}; \mathbf{p}) := \Sigma^{(1)}(\mathbf{x}) := \sigma_v^2 \|\mathbf{x}\|^2 / d_{\text{in}} + \sigma_b^2$ and, for each $l = 2, \ldots, L+1$,

$$\Sigma^{(l)}(\mathbf{x}; \mathbf{p}) := \sigma_b^2 + \sigma_v^2 \sum_{j=1}^{p_{l-1}} \lambda_{p_{l-1},j}^{(l-1)} \phi\left(Z_j^{(l-1)}(\mathbf{x}; \mathbf{p})\right)^2. \tag{27}$$

We have, for all $l = 2, \ldots, L+1$,

$$Z_k^{(1)}(\mathbf{x}; \mathbf{p}) = \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} V_{jk}^{(1)} x_j + B_k^{(1)}, \qquad Z_k^{(l)}(\mathbf{x}; \mathbf{p}) = \sum_{j=1}^{p_{l-1}} \sqrt{\lambda_{p_{l-1},j}^{(l-1)}} V_{jk}^{(l)} \phi(Z_j^{(l-1)}(\mathbf{x}; \mathbf{p})) + B_k^{(l)}.$$

Since the $V_{jk}^{(l)} \sim \mathcal{N}(0, \sigma_b^2)$ are independent among themselves, and also independent from the families $\{\lambda_j^{(l-1)}, Z_j^{(l-1)}(\mathbf{x}; \mathbf{p})\}_j$ and $\{B_k^{(l)}\}_k$, we may condition on $\Sigma^{(l)}(\mathbf{x}; \mathbf{p})$ to obtain, for all $l = 1, \ldots, L+1$,

$$(Z_1^{(l)}(\mathbf{x}; \mathbf{p}), \ldots, Z_{p_l}^{(l)}(\mathbf{x}; \mathbf{p})) \mid \Sigma^{(l)}(\mathbf{x}; \mathbf{p}) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \Sigma^{(l)}(\mathbf{x}; \mathbf{p})\right).$$

Hence,

$$(Z_1^{(l)}(\mathbf{x}; \mathbf{p}), \ldots, Z_{p_l}^{(l)}(\mathbf{x}; \mathbf{p})) \overset{\mathrm{d}}{=} \sqrt{\Sigma^{(l)}(\mathbf{x}; \mathbf{p})}(\varepsilon_1^{(l)}, \ldots, \varepsilon_{p_l}^{(l)}) \tag{28}$$

when $\varepsilon_k^{(l)} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. By the positive homogeneity of ϕ, Equation (27) can be rewritten as

$$\Sigma^{(l)}(\mathbf{x};\mathbf{p}) \overset{\mathrm{d}}{=} \sigma_b^2 + \sigma_v^2 \Sigma^{(l-1)}(\mathbf{x};\mathbf{p}) \sum_{j=1}^{p_{l-1}} \lambda_{p_{l-1},j}^{(l-1)} \phi(\varepsilon_j^{(l-1)})^2.$$

Let us set $S^{(l)}(p_l) := \sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \phi(\varepsilon_j^{(l)})^2$ for $l = 1, \ldots, L$. Then, the recurrence relation in Equation (27) defines a continuous map Ψ from $[0,\infty)^L$ to $[0,\infty)^{L+1}$ satisfying

$$(\Sigma^{(1)}(\mathbf{x};\mathbf{p}), \ldots, \Sigma^{(L+1)}(\mathbf{x};\mathbf{p})) \overset{\mathrm{d}}{=} \Psi(S^{(1)}(p_1), \ldots, S^{(L)}(p_L)).$$

Note that by the recurrence relation in Equation (23) and its relationship with Equation (27),

$$(\Sigma^{(1)}(\mathbf{x}), \ldots, \Sigma^{(L+1)}(\mathbf{x})) = \Psi(S^{(1)}, \ldots, S^{(L)}).$$

Also, note that the $S^{(l)}(p_l)$ are independent and do not depend on the input $\mathbf{x}$, and that the random variables $\{\phi(\varepsilon_j^{(l)})^2\}_{j=1,\ldots p_l}$ are independent. As $|\phi(x)| \leq C_{\text{Lip}}|x|$ for $C_{\text{Lip}} := \max(|\phi(1)|, |\phi(-1)|)$, we have $\mathbf{E}[\phi(\varepsilon_j^{(l)})^2] < \infty$. Additionally, $\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \overset{\mathrm{d}}{\to} \mathrm{ID}(a^{(l)}, \rho^{(l)})$ for each $l = 1, \ldots, L$. It follows from Corollary 37 in Appendix B that

$$(S^{(1)}(p_1), \ldots, S^{(L)}(p_L)) \overset{\mathrm{d}}{\to} \bigotimes_{l=1,\ldots,L} \mathrm{ID}(c^{(l)}, \eta^{(l)}) \tag{29}$$

as $\min(p_1, \ldots, p_L) \to \infty$. By the continuous mapping theorem, this implies

$$\Psi(S^{(1)}(p_1), \ldots, S^{(L)}(p_L)) \overset{\mathrm{d}}{\to} \Psi(S^{(1)}, \ldots, S^{(L)}).$$

Thus, $(\Sigma^{(1)}(\mathbf{x};\mathbf{p}), \ldots, \Sigma^{(L+1)}(\mathbf{x};\mathbf{p})) \overset{\mathrm{d}}{\to} (\Sigma^{(1)}(\mathbf{x}), \ldots, \Sigma^{(L+1)}(\mathbf{x}))$. The final result now follows.

4.3 Recursion for the Variance of the Limiting Outputs

Let $(\zeta_1^{(l)}(\mathbf{x}), \ldots, \zeta_m^{(l)}(\mathbf{x}))$ be the random variables that are distributed as $\bigotimes_{k=1,\ldots,m} \mathcal{N}(0, \Sigma^{(l)}(\mathbf{x}))$ when conditioned on $\Sigma^{(l)}$. Note that if we do not condition on $\Sigma^{(l)}$, these random variables $(\zeta_1^{(l)}(\mathbf{x}), \ldots, \zeta_m^{(l)}(\mathbf{x}))$ have the distribution $\mathbf{E}[\bigotimes_{k=1,\ldots,m} \mathcal{N}(0, \Sigma^{(l)}(\mathbf{x}))]$. Thus, they are the infinite-width pre-activations/outputs in Theorem 8. Assume that, for any l, $M_1^{(l)} := \int_0^\infty x \rho^{(l)}(dx) < \infty$, and $C_\phi := \mathbf{E}[\phi(X)^2] < \infty$, where $X \sim \mathcal{N}(0,1)$. Then,

$$\mathrm{Var}(\zeta_k^{(l)}(\mathbf{x})) = \mathbf{E}[\Sigma^{(l)}(\mathbf{x})] \tag{30}$$

where $\mathbf{E}[\Sigma^{(l)}(\mathbf{x})]$ follows the recursion

$$\mathbf{E}[\Sigma^{(1)}(\mathbf{x})] := \sigma_b^2 + (\sigma_v^2 \|\mathbf{x}\|^2 / d_{\text{in}}) \tag{31}$$

$$\mathbf{E}[\Sigma^{(l)}(\mathbf{x})] := \sigma_b^2 + \sigma_v^2 C_\phi (a^{(l-1)} + M_1^{(l-1)}) \mathbf{E}[\Sigma^{(l-1)}(\mathbf{x})] \quad \text{for } \ l = 2, \ldots, L+1. \tag{32}$$

In the particular cases where $\sigma_b = 0$, we obtain the simple expression

$$\mathrm{Var}(\zeta_k^{(l)}(\mathbf{x})) = \sigma_v^2 \frac{\|\mathbf{x}\|^2}{d_{\text{in}}} \prod_{l'=1}^{l-1} \sigma_v^2 C_\phi (a^{(l')} + M_1^{(l')}). \tag{33}$$

In order to avoid the variance of the pre-activations to explode/vanish as the depth increases, the pair $(a^{(l)}, \rho^{(l)})$ should be chosen such that $\sigma_v^2 C_\phi (a^{(l)} + M_1^{(l)}) = 1$. In the ReLU case, $C_\phi = 1/2$, and this reduces, if $\sigma_v = 1$, to $a^{(l)} + M_1^{(l)} = 2$. This is the configuration of the four examples presented in Section 1.

4.4 Regularly Varying Properties of the Activations and Outputs

We derive here results for the ReLU activation function. Similar results can be derived for other homogeneous activation functions. We have already shown in Proposition 6 that if the tail Lévy intensity decays polynomially at infinity, then the weights have power-law tails. The next proposition shows that if this is the case for all hidden layers, then the activations at each level and the outputs also have regularly varying tails, with an exponent which is twice the minimum of the exponents of the tail Lévy intensities in the layers below that level.

Proposition 9 *Let $L \geq 1$. Consider the same assumptions as in Theorem 8. Also, for $l = 1, \ldots, L$, assume that $\overline{\rho}^{(l)}$ has a power-law behaviour at infinity with exponent $\tau^{(l)}$, that is*

$$\overline{\rho}^{(l)}(x) \overset{x\to\infty}{\sim} c^{(l)} x^{-\tau^{(l)}} \tag{34}$$

for some positive constants $c^{(l)} > 0$. Then, for any $l = 1, \ldots, L$,

$$\Pr(S^{(l)} > u) \overset{u\to\infty}{\sim} \overline{\nu}^{(l)}(u)/2 \overset{u\to\infty}{\sim} \widetilde{c}^{(l)} u^{-\tau^{(l)}} \tag{35}$$

where

$$\overline{\nu}^{(l)}(u) = \int_0^\infty \overline{\rho}^{(l)}(u/z)\,\mathrm{Gamma}(z; 1/2, 1/2)dz, \qquad \widetilde{c}^{(l)} = c^{(l)} \times 2^{(\tau^{(l)}-1)}\Gamma(\tau^{(l)} + 1/2)/\sqrt{\pi}.$$

Also, for all $k \geq 1$ and $2 \leq l \leq L+1$, if we let $\zeta_k^{(l)}(\mathbf{x}) := \varepsilon_k^{(l)}\sqrt{\Sigma^{(l)}(\mathbf{x})}$ for $\varepsilon_k^{(l)} \sim \mathcal{N}(0,1)$, then

$$\Pr(\Sigma^{(l)}(\mathbf{x}) > u) \overset{u\to\infty}{\sim} u^{-\beta^{(l-1)}} L^{(l-1)}(u) \tag{36}$$

$$\Pr((\zeta_k^{(l)}(\mathbf{x}))^2 > u) \overset{u\to\infty}{\sim} u^{-\beta^{(l-1)}} L^{(l-1)}(u) \times 2^{\beta^{(l-1)}}(\Gamma(\beta^{(l-1)} + 1/2)/\Gamma(1/2)) \tag{37}$$

where $\beta^{(l-1)} = \min(\tau^{(1)}, \ldots, \tau^{(l-1)})$ and $L^{(l-1)}$ are some slowly varying functions.

In general, the slowly varying functions $L^{(l-1)}$ in Proposition 9 cannot be obtained analytically. An exception is when the hidden layers have the same asymptotic distribution and there is no bias, as we show in the next proposition.

Proposition 10 *Let $L \geq 1$. Consider the same assumptions as in Theorem 8. Additionally, assume that $\sigma_b = 0$, $a^{(l)} = a \geq 0$ and $\rho^{(l)} = \rho$ for all $l = 1, \ldots, L$ with $\overline{\rho}(x) \overset{x\to\infty}{\sim} cx^{-\tau}$ for some positive constant $c > 0$ and exponent $\tau > 0$. For $k \geq 1$ and $l = 2, \ldots, L+1$, let $\zeta_k^{(l)}(\mathbf{x}) := \varepsilon_k^{(l)}\sqrt{\Sigma^{(l)}(\mathbf{x})}$ for $\varepsilon_k^{(l)} \sim \mathcal{N}(0,1)$. Then, for $l = 2, \ldots, L+1$,*

$$\Pr(\Sigma^{(l)}(\mathbf{x}) > u) \overset{u\to\infty}{\sim} \left(\frac{\|\mathbf{x}\|^2}{d_{\text{in}}}\sigma_v^{2l}\right)^\tau \frac{\tau^{l-2}(\widetilde{c})^{l-1}}{(l-2)!} u^{-\tau}\log^{l-2} u$$

$$\Pr((\zeta_k^{(l)})^2 > u) \overset{u\to\infty}{\sim} \Pr(\Sigma^{(l)}(\mathbf{x}) > u) \times (2^\tau\Gamma(\tau + 1/2)/\Gamma(1/2))$$

where $\widetilde{c} = c \times (2^{\tau-1}\Gamma(\tau + 1/2)/\sqrt{\pi})$.

Note that in this particular case, the tails of the activations have the same exponent 2τ, but an additional log factor is added for each additional hidden layer after the first one, and so, the tails become slightly heavier as the network gets deeper.

4.5 Pruning of the Nodes of the Network

Suppose that we want to prune the nodes of the neural network in order to reduce the computational cost. We consider two different strategies for node pruning, both based on the values of the per-node variances $\lambda_{p_l,j}^{(l)}$. The first strategy, called *ϵ-pruning*, prunes nodes such that $\lambda_{p_l,j}^{(l)} \leq \epsilon$, for some fixed threshold $\epsilon > 0$. The second strategy, called *κ-pruning*, prunes nodes such that $\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}$ where the subscript $(\lfloor \kappa p_l \rfloor)$ denotes an order statistic: at layer l, $\lambda_{p_l,(1)}^{(l)} \geq \lambda_{p_l,(2)}^{(l)} \geq \ldots \geq \lambda_{p_l,(p_l)}^{(l)}$ denote the ordered values of $(\lambda_{p_l,j}^{(l)})_{j=1,\ldots,p_l}$. When there are no repeated values, κ-pruning is equivalent to pruning a proportion $(1-\kappa) \in (0,1)$ of the p_l nodes with lowest $\lambda_{p_l,j}^{(l)}$ values in each layer. The pruning strategies we employ here are related to the compressibility of a network discussed in Section 3.5. This connection was noted in Barsbey et al. (2021) where similar pruning schemes were discussed.

We start with an error bound of the pruned network that holds for both strategies with ϵ and κ. To this end, let $\lambda_{p_l}^{*(l)}, l = 1, 2, \ldots$, be nonnegative random variables. Consider the following pruned network:

$$\begin{aligned} Z_k^{*(1)}(\mathbf{x};\mathbf{p}) &:= Z_k^{*(1)}(\mathbf{x}) := \sum_{j=1}^{d_{\text{in}}} \frac{1}{\sqrt{d_{\text{in}}}} V_{jk}^{(1)} x_j + B_k^{(1)}, \\ Z_k^{*(l)}(\mathbf{x};\mathbf{p}) &:= \sum_{j=1}^{p_{l-1}} \left(\sqrt{\lambda_{p_{l-1},j}^{(l-1)}} \mathbf{1}_{\{\lambda_{p_{l-1},j}^{(l-1)} > \lambda_{p_{l-1}}^{*(l-1)}\}} \right) V_{jk}^{(l)} \phi(Z_j^{*(l-1)}(\mathbf{x};\mathbf{p})) + B_k^{(l)}, \quad l \geq 2. \end{aligned} \tag{38}$$

Namely, we prune a node if its node variance is less than or equal to the threshold $\lambda_{p_l}^{*(l)}$. For ϵ-pruning, $\lambda_{p_l}^{*(l)} = \epsilon$. In this case, we write $Z_k^{*(l)}(\mathbf{x};\mathbf{p}) = Z_k^{*(l)}(\mathbf{x};\mathbf{p},\epsilon)$ to emphasise the dependence of the network on ϵ. On the other hand, for κ-pruning, $\lambda_{p_l}^{*(l)} = \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}$. Similarly, we write $Z_k^{*(l)}(\mathbf{x};\mathbf{p}) = Z_k^{*(l)}(\mathbf{x};\mathbf{p},\kappa)$ to emphasise the dependence on κ.

Set $N_{p_l}^{(l)} := \mathbf{E}[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}]$. A key assumption used throughout this subsection on pruning is:

(UI) For all layers $l = 1, \ldots, L$,

$$\int_0^\infty u \rho^{(l)}(du) = M_1^{(l)} < \infty, \quad N_{p_l}^{(l)} < \infty \text{ for all } p_l, \quad \text{and} \quad N_{p_l}^{(l)} \to a^{(l)} + M_1^{(l)} \text{ as } p_l \to \infty.$$

In our setting, the assumption (UI) is equivalent to the uniform integrability of the family $\{\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}\}_{p_l}$ (see Appendix C.4).

We will also utilise the following assumptions in this subsection:

(A1) The activation function ϕ is positive homogeneous.

(A2) Equation (11) holds with $a^{(l)} = 0$ for all hidden layers $l = 1, \ldots, L$.

(A3) The Lévy measures of all layers are equal, $\rho^{(l)} = \rho$, and ρ satisfies $\overline{\rho}(u) \overset{u \to 0}{\sim} u^{-\alpha} L(1/u)$ for some $\alpha \in [0,1)$ and some slowly varying function L. In this case, $M_1 := M_1^{(l)}$ does not depend on l.

The following proposition gives a bound on the error of the above pruned network. The argument is a variant of the variance recursion given in Section 4.3. To state the proposition, recall that $C_{\text{Lip}} = \max(|\phi(1)|, |\phi(-1)|)$ and define $U^{(l)} := \sup_{\mathbf{p}} \mathbf{E}[(Z_1^{(l)}(\mathbf{x};\mathbf{p}))^2]$. We point out that $U^{(l)} < \infty$ under (UI) and (A1); see Lemma 38 in the Appendix.

Proposition 11 (Pruning error bound) *If (A1) holds, then the L^2-error between the pruned and unpruned networks satisfies*

$$\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \tag{39}$$
$$\leq \sigma_v^2 C_{\text{Lip}}^2 U^{(l)} A_{p_l}^{(l)} + (\sigma_v^2 C_{\text{Lip}}^2)^2 N_{p_l}^{(l)} U^{(l-1)} A_{p_{l-1}}^{(l-1)} + \cdots + (\sigma_v^2 C_{\text{Lip}}^2)^l N_{p_l}^{(l)} \cdots N_{p_2}^{(2)} U^{(1)} A_{p_1}^{(1)},$$

where $A_{p_l}^{(l)} := \mathbf{E}[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l}^{*(l)}\}}]$.

Remark 12 *To get a bound on $U^{(l)}$, note that the variance $\mathbf{E}[(Z_1^{(l)}(\mathbf{x};\mathbf{p}))^2]$ satisfies a similar recurrence relation to that described in Section 4.3. Namely,*

$$\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \leq \sigma_v^2 C_\phi N_{p_l}^{(l)} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p})\right)^2\right] + \sigma_b^2.$$

Also, the bound in Equation (39) *holds when the supremum in $U^{(l)}$ for each l is taken for $\mathbf{p}'$ with $\min \mathbf{p}' \geq \min \mathbf{p}$. See the proofs of Lemma 38 and Proposition 11 for details. In the particular case where $\sigma_b = 0$, $\text{Var}(\zeta_1^{(l)}(\mathbf{x})) > 0$, $\min \mathbf{p}$ is sufficiently large and (A2) holds, if the supremum of $U^{(l)}$ is taken over $\mathbf{p}'$ with $\min \mathbf{p}' \geq \min \mathbf{p}$ for every l, then $U^{(l)}$ satisfies*

$$U^{(l)} \leq 2\,\text{Var}(\zeta_1^{(l)}(\mathbf{x})) = 2\sigma_v^2 \frac{\|\mathbf{x}\|^2}{d_{\text{in}}} \prod_{l'=1}^{l-1} \sigma_v^2 C_\phi M_1^{(l')}.$$

4.5.1 ϵ-Pruning

Let $\lambda_{p_l}^{*(l)} = \epsilon$ for some $\epsilon > 0$. At layer l, this means that we keep the hidden nodes j such that $\lambda_{p_l,j}^{(l)} > \epsilon$. (We do not let ϵ, the pruning level, depend on the layer here just to simplify presentation; lifting this restriction would not invalidate our results to be presented next.) It should be noted that, when the limiting unpruned network is infinite, this pruning strategy produces a finite network.

To analyse the error between the unpruned network and the ϵ-pruned network in Equation (38), we investigate the limit of the L^2 pruning error $\mathbf{E}[(Z_k^{*(l)}(\mathbf{x};\mathbf{p},0) - Z_k^{*(l)}(\mathbf{x};\mathbf{p},\epsilon))^2]$ and show that, under assumptions (UI) and (A1-A3), this error remains small in the limit as $\min(p_1,\ldots,p_L) \to \infty$. This comes as a corollary to Proposition 11.

Corollary 13 (Single input case, ϵ-pruning) *Consider pruned FFNNs defined by Equations* (7), (9), (10) *and* (38) *with $\lambda_{p_l}^{*(l)} = \epsilon$. Suppose (UI) and (A1-A3) hold. Then, for all $\delta \in (0, 1-\alpha)$, there exists $\epsilon_0(\delta) > 0$ such that if $\epsilon < \epsilon_0(\delta)$, we have, for each $l = 1,\ldots,L$ and any $k \geq 1$,*

$$\lim_{\min \mathbf{p} \to \infty} \mathbf{E}\left[\left|Z_k^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_k^{*(l+1)}(\mathbf{x};\mathbf{p})\right|^2\right] \leq D(l) \cdot \epsilon^{1-(\alpha+\delta)},$$

where

$$D(l) = \frac{\sigma_v^2 C_{\text{Lip}}^2}{1-(\alpha+\delta)} (U^{(l)} + (\sigma_v^2 C_{\text{Lip}}^2 M_1) U^{(l-1)} + \cdots + (\sigma_v^2 C_{\text{Lip}}^2 M_1)^{l-1} U^{(1)})$$

is a constant not depending on ϵ.

Although the pruning error is controlled mostly by the pruning level ϵ, the error can vary according to the constant $D(l)$ which depends on the number of previous layers l. The deeper our network gets, the larger the pruning error becomes. In other words, the pruning error is small at shallow layers, but it accumulates and gets larger at deeper layers.

In the particular case $\sigma_b = 0$ (no bias), combining Corollary 13 with Remark 12, we obtain

$$\lim_{\min \mathbf{p} \to \infty} \mathbf{E}\left[\left|Z_k^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_k^{*(l+1)}(\mathbf{x};\mathbf{p})\right|^2\right] \leq \left(\frac{\sigma_v^2 C_{\text{Lip}}^2}{1-(\alpha+\delta)}\right) \cdot \left(\sum_{l'=0}^{l-1}\left(\frac{C_{\text{Lip}}^2}{C_\phi}\right)^{l'}\right) \cdot 2\,\text{Var}(\zeta_1^{(l)}(\mathbf{x})) \cdot \epsilon^{1-(\alpha+\delta)}$$

where C_ϕ is as in Section 4.3.

Remark 14 *In Appendix C.4, we prove Corollary 13 in a slightly more general setting where we allow for different $\rho^{(l)}$'s in different layers. The trade-off is that, if we confine $\rho^{(l)} = \rho$ for some ρ as in (A3), then $\epsilon_0(\delta)$ depends only on δ and not on L, thus one can possibly add more layers after $L+1$. On the contrary, if we allow for different $\rho^{(l)}$'s as in the proof, then $\epsilon_0(\delta, L)$ depends not only on δ but also on L, so adding more layers requires changing ϵ_0.*

4.5.2 κ-Pruning

For fixed $\kappa \in (0,1)$, let $\lambda_{p_l}^{*(l)} = \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}$. That is, κ-pruning discards nodes j at layer l with $\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}$.

The next result shows that, under assumptions (UI) and (A1-A2) including the compressibility of layers (A2; see Section 3.5), the error between the unpruned output and the κ-pruned output in Equation (38) converges to 0, no matter what the value $\kappa \in (0,1)$ is. Again, this comes as a corollary to Proposition 11.

Corollary 15 (Single input case, κ-pruning) *Consider pruned FFNNs defined by Equations* (7), (9), (10) *and* (38) *with $\lambda_{p_l}^{*(l)} = \lambda_{p_l,(\lfloor \kappa p_l \rfloor)}^{(l)}$. Suppose (UI) and (A1-A2) hold. Then, for each $l = 1, \ldots, L$ and for any $\kappa \in (0,1)$ and any $k \geq 1$,*

$$\mathbf{E}\left[\left|Z_k^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_k^{*(l+1)}(\mathbf{x};\mathbf{p})\right|^2\right] \to 0 \text{ as } \min \mathbf{p} \to \infty.$$

This result states that, if $a^{(l)} = 0$ for all $l = 1, \ldots, L$, the neural network is compressible: the difference between the output of the κ-pruned network and that of the unpruned network vanishes in probability as the width of the network goes to infinity. This is not generally the case if $a^{(l)} > 0$. If, in addition, almost surely no node variances are repeated (so κ-pruning prunes a $(1-\kappa)$-proportion of nodes), we do not obtain the vanishing error, which occurs when $a^{(l)} = 0$. For instance, consider a network with one hidden layer. Then, the L^2-error is

$$\mathbf{E}\left[\left|Z_1^{(2)}(\mathbf{x};\mathbf{p}) - Z_1^{*(2)}(\mathbf{x};\mathbf{p})\right|^2\right] = \sigma_v^2 \mathbf{E}\left[\sum_{j=1}^{p_1} \lambda_{p_1,j}^{(1)} \mathbf{1}_{\{\lambda_{p_1,j}^{(1)} \leq \lambda_{p_1,(\lfloor \kappa p_1 \rfloor)}^{(1)}\}}\right] \mathbf{E}\left[\phi^2(Z^{(1)}(\mathbf{x}))\right]$$

which is not guaranteed to converge to 0 for all $\kappa \in (0,1)$ when $a > 0$. See Proposition 32.

In the iid Gaussian case, our κ-pruning strategy prunes every node due to the repeated node variance $\frac{c_1}{p_l}$, so that the pruning error trivially does not vanish. In practice, one prunes the iid Gaussian case by removing nodes using instead

$$T_{p,j}^{(l)} := \|W_{j,:}\|^2 = \lambda_{p,j}^{(l)} \sum_{k=1}^{p_{l+1}} (V_{j,k}^{(l+1)})^2$$

Denote by Z^{**}, the network defined in a similar way to Equation (38) but where $T_{p,j}$ is used for pruning instead of $\lambda_{p,j}$. Then, it can be shown that in the iid Gaussian case, the error is non-vanishing, i.e.

$$\limsup \mathbf{E}\left[\left|Z_k^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_k^{*(l+1)}(\mathbf{x};\mathbf{p})\right|^2\right] > 0.$$

5. Infinite-Width Limit for Multiple Inputs in the General Case

We prove the convergence theorem for multiple inputs under a more general assumption for the activation function ϕ than the positive homogeneity assumption in Theorem 8. Any positive homogeneous function such as ReLU satisfies this generalisation, as well as the classical tanh or sigmoid functions. This assumption is called a "polynomial envelope" condition by (Matthews et al., 2018), and commonly used in the context of analysing infinitely-wide neural networks either implicitly or explicitly. In (Neal, 1996; Lee et al., 2018), the authors are implicitly exploiting this assumption by considering tanh and ReLU mainly, and in (Favaro et al., 2020; Jung et al., 2023), they explicitly considered a weaker version of this assumption.

Theorem 16 (Multi-input case) *Consider the feedforward neural network model defined by Equations* (6) *to* (10)*. Assume that the activation function ϕ is continuous and satisfies the so-called polynomial envelope condition: for all $z \in \mathbb{R}$, $|\phi(z)| \leq A + B|z|^C$ for some $A, B, C > 0$. Assume that, for all hidden layers $l = 1, \ldots, L$, we have*

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \xrightarrow{\mathrm{d}} \mathrm{ID}(a^{(l)}, \rho^{(l)}) \text{ as } p_l \to \infty$$

for some $a^{(l)} \geq 0$ and some Lévy measure $\rho^{(l)}$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n inputs, where $\mathbf{x}_i \in \mathbb{R}^{d_{\text{in}}}$. Define $\vec{Z}_k^{(l)}(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{p}) := (Z_k^{(l)}(\mathbf{x}_1; \mathbf{p}), \ldots, Z_k^{(l)}(\mathbf{x}_n; \mathbf{p}))^T \in \mathbb{R}^n$, the associated k-th outputs. Then, for all $l = 1, \ldots, L+1$ and all $m \geq 1$, as $\mathbf{p} \to \infty$ in the order $\lim_{p_L \to \infty} \ldots \lim_{p_1 \to \infty}$,

$$\left(\vec{Z}_k^{(l)}(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mathbf{p})\right)_{k=1,\ldots,m} \xrightarrow{\mathrm{d}} \mathbf{E}\left[\bigotimes_{k=1,\ldots,m} \mathcal{N}(0, \Sigma^{(l)})\right].$$

Here, $\Sigma^{(l)}$ is a random n-by-n positive semi-definite matrix defined by $\Sigma_{ij}^{(l)} = K^{(l)}(\mathbf{x}_i, \mathbf{x}_j)$, for $1 \leq i, j \leq n$, where $K^{(l)} : \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ is a random covariance kernel. The sequence of random kernels $(K^{(1)}, \ldots, K^{(L+1)})$ is a Markov sequence whose distribution can be defined recursively, for $l = 1, \ldots, L$, by:

$$K^{(1)}(\mathbf{x}, \mathbf{x}') := \sigma_b^2 + \sigma_v^2 \frac{\mathbf{x}^T \mathbf{x}'}{d_{\text{in}}} \tag{40}$$

$$K^{(l+1)}(\mathbf{x}, \mathbf{x}') := \sigma_b^2 + \sigma_v^2 a^{(l)} \mathbf{E}\left[\phi(\zeta_1^{(l)}(\mathbf{x}))\phi(\zeta_1^{(l)}(\mathbf{x}'))\middle| K^{(l)}\right] + \sigma_v^2 \sum_{j \geq 1} \widetilde{\lambda}_j^{(l)} \phi\left(\zeta_j^{(l)}(\mathbf{x})\right) \phi\left(\zeta_j^{(l)}(\mathbf{x}')\right)$$

where $\{\widetilde{\lambda}_j^{(l)}\}_{j \geq 1}$ are the points of a Poisson point process on $(0, \infty)$ with mean measure $\rho^{(l)}$ and, for $j \geq 1$,

$$\zeta_j^{(l)} \mid K^{(l)} \overset{iid}{\sim} \mathrm{GP}(0, K^{(l)}).$$

Here $\mathrm{GP}(\mu, K)$ denotes a Gaussian process on $\mathbb{R}^{d_{\text{in}}}$, i.e., a random element of $\mathcal{M} = \{f : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}\}$, with mean $\mu \in \mathcal{M}$ and covariance function $K : \mathbb{R}^{d_{\text{in}}} \times \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$.

Remark 17 *The limit in the above theorem is taken in sequential order from the first layer to the last layer. Extending the theorem to a different and more natural limiting scheme, such as $\min(p_1, \ldots, p_L) \to \infty$, is non-trivial. For instance, although the proof of Theorem 8 handles the case $\min(p_1, \ldots, p_L) \to \infty$, it heavily relies on positive homogeneity of the activation function so as to rephrase the outputs of hidden nodes in some layer l as a vector of independent Gaussian random variables that is scaled by a random scalar (Equation* (28)*). Since the positive homogeneity of ϕ does not let us move a matrix M from $\phi(Mv)$ to the outside in any form, It is difficult to obtain an analogous result in the case of multiple inputs. We expect that a different approach, such as the use of exchangeability (Favaro et al., 2020; Matthews et al., 2018), is needed for such extension of our result, and we leave this as one of the remaining future challenges.*

When $\rho^{(l)}$ is trivial for all $l = 1, \dots, L$, the kernels are deterministic, and one recovers a Gaussian process. Otherwise, we obtain a mixture of Gaussian processes, where the mixture comes from the randomness of the kernel $K^{(l)}$. We now discuss some of the properties of the random kernel.

The following proposition is an immediate consequence of the Campbell theorem for Poisson random measures, together with results regarding the ReLU activation function (Cho and Saul, 2009); see Appendix A.2.

Proposition 18 (Conditional mean and variance of the kernel) *For any $l \geq 1$ and $n \geq 1$, let $M_n^{(l)} = \int_0^\infty x^n \rho^{(l)}(dx)$. We have*

$$\begin{aligned}
\mathbf{E}\left[K^{(l+1)}(\mathbf{x},\mathbf{x}')\middle| K^{(l)}\right] &= \sigma_b^2 + \sigma_v^2(M_1^{(l)} + a^{(l)})\mathbf{E}\left[\phi(\zeta_1^{(l)}(\mathbf{x}))\phi(\zeta_1^{(l)}(\mathbf{x}'))\middle| K^{(l)}\right] \\
\mathrm{Var}\left[K^{(l+1)}(\mathbf{x},\mathbf{x}')\middle| K^{(l)}\right] &= \sigma_v^4 M_2^{(l)}\mathbf{E}\left[\phi(\zeta_1^{(l)}(\mathbf{x}))^2\phi(\zeta_1^{(l)}(\mathbf{x}'))^2\middle| K^{(l)}\right]
\end{aligned}$$

where

$$\begin{pmatrix} \zeta_1^{(l)}(\mathbf{x}) \\ \zeta_1^{(l)}(\mathbf{x}') \end{pmatrix}\middle| K^{(l)} \overset{iid}{\sim} \mathcal{N}\left(0, \begin{pmatrix} K^{(l)}(\mathbf{x},\mathbf{x}) & K^{(l)}(\mathbf{x},\mathbf{x}') \\ K^{(l)}(\mathbf{x},\mathbf{x}') & K^{(l)}(\mathbf{x}',\mathbf{x}') \end{pmatrix}\right). \tag{41}$$

In the ReLU case, we have the analytic expressions

$$\begin{aligned}
\mathbf{E}\left[K^{(l+1)}(\mathbf{x},\mathbf{x}')\middle| K^{(l)}\right] &= \sigma_b^2 + \sigma_v^2(M_1^{(l)} + a^{(l)})\frac{\sqrt{K^{(l)}(\mathbf{x},\mathbf{x})K^{(l)}(\mathbf{x}',\mathbf{x}')}}{2\pi}\kappa_1(\rho^{(l)}_{\mathbf{x},\mathbf{x}'}) \\
\mathrm{Var}\left[K^{(l+1)}(\mathbf{x},\mathbf{x}')\middle| K^{(l)}\right] &= \sigma_v^4 M_2^{(l)}\frac{K^{(l)}(\mathbf{x},\mathbf{x})K^{(l)}(\mathbf{x}',\mathbf{x}')}{2\pi}\kappa_2(\rho^{(l)}_{\mathbf{x},\mathbf{x}'})
\end{aligned}$$

where $\rho^{(l)}_{\mathbf{x},\mathbf{x}'} = K^{(l)}(\mathbf{x},\mathbf{x}')/\sqrt{K^{(l)}(\mathbf{x},\mathbf{x})K^{(l)}(\mathbf{x}',\mathbf{x}')}$ and

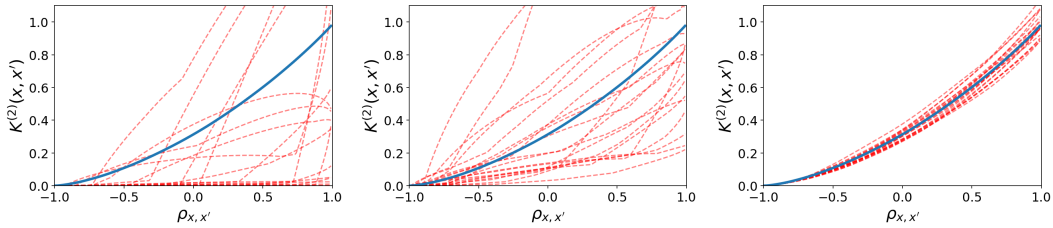
$$\kappa_n(\rho) = \begin{cases} \sqrt{1-\rho^2} + \left(\frac{\pi}{2} + \arcsin\rho\right)\rho & \text{if } n = 1 \\ 3\sqrt{1-\rho^2}\rho + \left(\frac{\pi}{2} + \arcsin\rho\right)(1+2\rho^2) & \text{if } n = 2. \end{cases} \tag{42}$$

Example 2 *Assume that $\sigma_v = 1$ and $\sigma_b = 0$. Consider the model $\lambda_{p,j}^{(l)} \sim \mathrm{Beta}(\beta/p, \beta/2)$ for some $\beta > 0$. This generalises the example (c) introduced in Section 1, with an additional parameter $\beta > 0$. As will be shown later in Section 6.5, $\sum_j \lambda_{p,j}^{(l)}$ converges in distribution to a random variable $\Lambda^{(l)} \sim \mathrm{ID}(0,\rho)$ where $\rho(dx) = \beta x^{-1}(1-x)^{(\beta/2)-1}\mathbf{1}_{\{x\in(0,1)\}}dx$. This is a beta Lévy measure, with moments $M_k = \beta\frac{\Gamma(k)\Gamma(\beta/2)}{\Gamma(k+\beta/2)}$, so that $M_1 = \mathbf{E}[\Lambda^{(l)}] = 2$ and $M_2 = \mathrm{Var}(\Lambda^{(l)}) = 4/(2+\beta)$. It follows that*

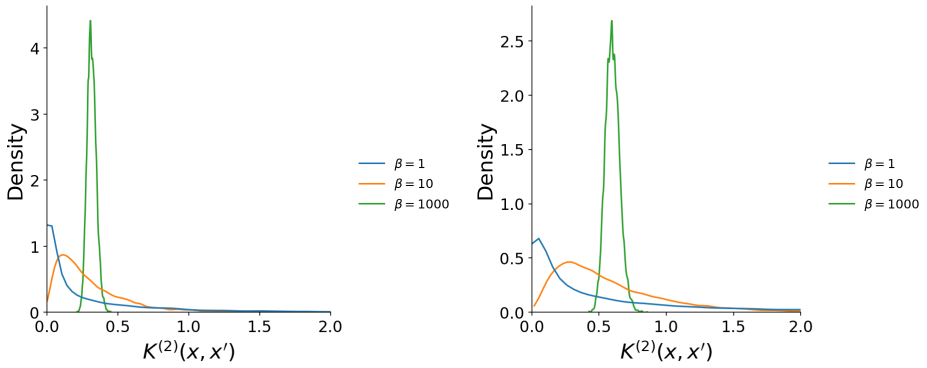
$$\mathbf{E}\left[K^{(2)}(\mathbf{x},\mathbf{x}')\right] = \mathcal{K}^{(2)}(\mathbf{x},\mathbf{x}'), \qquad \mathrm{Var}\left[K^{(2)}(\mathbf{x},\mathbf{x}')\right] = \frac{2}{\pi(2+\beta)}\frac{\|\mathbf{x}\|^2\|\mathbf{x}'\|^2}{{d_{\mathrm{in}}}^2}\kappa_2(\rho^{(1)}_{\mathbf{x},\mathbf{x}'}),$$

where $\mathcal{K}^{(2)}$ is the GP ReLU kernel given in Equation (4). Thus, the random kernel $K^{(2)}$ is centred on $\mathcal{K}^{(2)}$, and the parameter β controls the variance of the kernel. Realisations of the kernel $K^{(2)}$ for different values of β and with $\|\mathbf{x}\|\|\mathbf{x}'\|/d_{\mathrm{in}} = 1$ are given in Figure 1.

In Appendix D.2, we further discuss a special case of Theorem 16 when the limiting infinitely divisible distribution of $\sum_{j=1}^{p_l}\lambda_{p_l,j}^{(l)}$ is an α-stable distribution.



(a) $\beta = 1$ (b) $\beta = 10$ (c) $\beta = 1000$



(d) $\rho_{\mathbf{x},\mathbf{x}'} = 0$ (e) $\rho_{\mathbf{x},\mathbf{x}'} = 0.5$

Figure 1: (a-c) Dashed red lines represent 20 realisations of the kernel $K^{(2)}(\mathbf{x}, \mathbf{x}')$, as a function of the correlation $\rho_{\mathbf{x},\mathbf{x}'} = \frac{\mathbf{x}^T\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}$, when $\|\mathbf{x}\|\|\mathbf{x}'\|/d_{\text{in}} = 1$, $\sigma_v = 1$, $\sigma_b = 0$, for the beta model in Example 2 with (a) $\beta = 1$, (b) $\beta = 10$ and (c) $\beta = 1000$. The solid blue line represents the GP ReLU kernel in Equation (4). The random kernels $K^{(2)}$ are centred on the GP ReLU kernel, and the variance decreases with the tuning parameter β. (d-e) Distribution of $K^{(2)}(\mathbf{x}, \mathbf{x}')$ for different values of β, for (d) $\rho_{\mathbf{x},\mathbf{x}'} = 0$ and (e) $\rho_{\mathbf{x},\mathbf{x}'} = 0.5$.

Name	Mixture's name	μ_p	a	Lévy measure	Support	Finite?	Exp. α	Exp. τ
Determ.	Gaussian	$\delta_{c_1/p}$	c_1	0	–	–	–	–
Bernoulli	Spike and Slab	$\left(1-\frac{c}{p}\right)\cdot\delta_0+\frac{c}{p}\delta_1$	0	$c\delta_1$	$\{1\}$	Yes	0	–
Gamma	Group lasso	Gamma $\left(\frac{p_{l+1}+1}{2},\frac{p_l(p_{l+1}+1)}{2c_1}\right)$	c_1	0	–	–	–	–
Beta	Normal-beta	Beta $\left(\frac{1}{p},\frac{1}{c}\right)$	0	$x^{-1}(1-x)^{1/c-1}$	(0,1)	No	–	–
Inv.-Gamma	Multivariate t	IG $(2,2/p)$	2	0	–	–	–	–
Beta prime	Horseshoe	$\frac{2p}{\pi^2}x^{-1/2}(1+\frac{4xp^2}{\pi^2})^{-1}$	0	$\frac{1}{2}x^{-3/2}$	$(0,\infty)$	No	1/2	1/2
Gen. BFRY	Normal -gen. BFRY	See Equation (48)	0	$\frac{\eta x^{-1-\tau}}{\Gamma(1-\alpha)}\gamma(\tau-\alpha,x)$	$(0,\infty)$	No	$\alpha\in(0,1)$	$\tau>\alpha$

Table 2: List of models and their limiting location parameter and Lévy measure.

6. Examples

In this section, we provide examples of models used in the literature, and the associated parameters of the limiting infinitely divisible random variable of Equation (2). In some cases, we use a different scaling so that the limit exists, and is not degenerate at 0. Table 2 summarises the properties of these models. Further discussions on these and additional example models can be found in Appendix E.2. To simplify notation, we often drop the layer index l fully or partially in the rest of this section, writing e.g. $\lambda_{p,j}\sim\mu_p$.

6.1 Constant Variance (iid Gaussian/Weight Decay/L2 Regularisation)

The standard iid Gaussian model is obtained as a special case when $\lambda_{p,j}\sim\delta_{c_1/p}$ for some constant $c_1>0$ and so the weights W_{jk} are iid $\mathcal{N}(0,(c_1\sigma_v^2)/p)$. In this case, $\sum_j\lambda_{p,j}=c_1$, so that $\sum_j\lambda_{p,j}\xrightarrow{\mathtt{d}}\mathrm{ID}(c_1,0)$. The weights (and variances) converge uniformly to 0, i.e. for any $k\geq 1$, $\max_{j=1,\dots,p}(|W_{jk}|)\xrightarrow{\mathtt{pr}}0$.

6.2 Bernoulli Prior

For some $c>0$, consider $\lambda_{p_l,j}^{(l)}\sim\mathrm{Bernoulli}(c/p_l)$ for every $p_l\geq c$. This corresponds to a marginal spike and slab distribution for $W_j^{(l)}=(W_{j1}^{(l)},\dots,W_{jp_{l+1}}^{(l)})$, with

$$W_j^{(l)}\sim\left(1-\frac{c}{p_l}\right)\cdot\delta_0+\frac{c}{p_l}\cdot\mathcal{N}(0,\sigma_v^2I_{p_{l+1}}).$$

Such a prior has been used by Jantre et al. (2021) for pruning Bayesian neural networks. In that case, $\sum_j\lambda_{p_l,j}^{(l)}\xrightarrow{\mathtt{d}}\mathrm{ID}(0,c\delta_1)$. That is, the location parameter a is zero, and the Lévy measure $\rho=c\delta_1$ is finite and discrete.

6.3 Group Lasso Prior

We consider that[2] $\lambda_{p_l,j}^{(l)}\sim\mathrm{Gamma}((p_{l+1}+1)/2,\,b_{p_l}/2)$, where b_{p_l} is an inverse-scale parameter that depends on the layer's width. Such a distribution leads to the so-called group lasso distribution (Raman et al., 2009; Casella et al., 2010) over the weights $(W_{jk}^{(l+1)})_{jk}$, which have joint marginal density

$$f(w)\propto\exp\left(-\frac{\sqrt{b_{p_l}}}{\sigma_v}\sum_{j=1}^{p_l}\sqrt{\sum_{k=1}^{p_{l+1}}w_{jk}^2}\right). \tag{43}$$

2. Note that in this case, $\lambda_{p_l,j}^{(l)}$ depends on the size p_{l+1} of the upper layer as well. However, we show here that, for a specific choice of b_{p_l}, at the infinite-width limit with respect to p_l, this dependency on p_{l+1} disappears. For clarity, we keep the superscript/subscript l in this subsection.

The regularisation term

$$-\log f(w) = \left(\frac{\sqrt{b_{p_l}}}{\sigma_v} \sum_{j=1}^{p_l} \sqrt{\sum_{k=1}^{p_{l+1}} w_{jk}^2} \right) + C \tag{44}$$

is known as the group lasso penalty, introduced by Yuan and Lin (2006) for regression models. This penalty has been used as a regulariser for neural networks by Scardapane et al. (2017) and Wang et al. (2017). The group lasso distribution in Equation (43) has been used as a sparsity-promoting prior in Bayesian learning of sparse neural networks by de Jong (2018).

Scardapane et al. (2017) suggested to set $b_{p_l} = p_l$. However, this assumption implies that $\lim_{p_l \to \infty} \sum_j \lambda_{p_l,j}^{(l)} = p_{l+1} + 1$ almost surely, which diverges if $p_{l+1} \to \infty$. This fact has been noted by Wolinski et al. (2020a) who suggested the different scaling $b_{p_l} = p_l(p_{l+1}+1)/c_1$ (with $c_1\sigma_v = 1$). Setting $b_{p_l} = p_l(p_{l+1}+1)/c_1$, we obtain $\sum_j \lambda_{p_l,j}^{(l)} \xrightarrow{\text{pr}} c_1$ as $p_l \to \infty$. Thus,

$$\sum_j \lambda_{p_l,j}^{(l)} \xrightarrow{\text{d}} \text{ID}(c_1, 0).$$

6.4 Inverse Gamma Prior and Similar Models

We consider here, as in (Ober and Aitchison, 2021), that the variances follow an inverse gamma distribution

$$\lambda_{p,j} \sim \text{IG}(2, 2/p). \tag{45}$$

Note that this is equivalent to

$$\lambda_{p,j} = Y_j/p \tag{46}$$

where $Y_1, Y_2, \ldots,$ are iid IG$(2,2)$. By the law of large numbers, $\sum_j \lambda_{p,j} \xrightarrow{\text{pr}} 2$ or equivalently, $\sum_j \lambda_{p,j} \xrightarrow{\text{d}} \text{ID}(2,0)$. More generally, any model of the form in Equation (46) where $Y_1, Y_2, \ldots$ are iid random variables with finite mean, satisfies $\sum_j \lambda_{p,j} \xrightarrow{\text{d}} \text{ID}(\mathbf{E}[Y_1], 0)$.

6.5 Beta Model and Beta Lévy Measure

Consider $\lambda_{p,j} \sim \text{Beta}(\eta/p, b)$ where $\eta, b > 0$. An application of Theorem 29 in Appendix B yields $\sum_j \lambda_{p,j} \xrightarrow{\text{d}} \text{ID}(0, \rho)$, where $\rho(dx) = \eta x^{-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}}dx$ is a Beta Lévy measure (Hjort, 1990). The measure is infinite with bounded support.

6.6 Horseshoe Model

In the horseshoe model (Carvalho et al., 2010), we assume the independent random variables $Y_1, Y_2, \ldots$ that have the same distribution as $Y = T^2$, where $T \sim \text{Cauchy}_+(0,1)$ is a half-Cauchy random variable, with pdf given by Equation (5). The random variable $Y \sim \text{Betaprime}(1/2, 1/2)$ is a beta prime random variable (with both shape parameters equal to $1/2$), with pdf

$$f_Y(y) = \frac{1}{\pi\sqrt{y}(1+y)}.$$

Its survival function satisfies

$$\Pr(Y > y) \overset{y\to\infty}{\sim} (2y^{-1/2})/\pi,$$

and therefore Y has a power-law tail at infinity with exponent $\alpha = 1/2$. Let $c > 0$ be some scaling parameter. Setting

$$\lambda_{p,j} = (c\pi^2 Y_j)/(4p^2),$$

we obtain

$$\sum_j \lambda_{p,j} \xrightarrow{\mathrm{d}} \mathrm{ID}(0,\rho) = \mathrm{IG}(1/2, c\pi/4). \tag{47}$$

where $\rho(dx) = (\sqrt{c}/2)x^{-3/2}\mathbf{1}_{\{x>0\}}dx$. The tail Lévy intensity $\overline{\rho}(x)$ in this case has power-law tails at 0 and ∞, with exponent $1/2$.[3]

6.7 Generalised Gamma Pareto Model

The model described in Section 6.6 allows us to obtain a Lévy measure which has power-law tails with the same exponent α at 0 and ∞. We describe here a model that permits power-law tails with different exponents. Let $\lambda_{p,j} = \beta_j \zeta_{p,j}$ where

$$\beta_j \sim \mathrm{Pareto}(\tau, 1), \qquad \zeta_{p,j} \sim \mathrm{etBFRY}\left(\alpha, \left(\frac{p\alpha\tau}{\eta(\tau-\alpha)}\right)^{1/\alpha}, 1\right) \tag{48}$$

for $\alpha \in (0,1)$, $\tau > \alpha$ and $\eta > 0$. Here $\mathrm{Pareto}(\tau, c)$ denotes the Pareto distribution with pdf $f(x) = \tau c^\tau x^{-\tau-1}\mathbf{1}_{\{x>c\}}$. Also, $\mathrm{etBFRY}(\alpha, t, \xi)$ denotes an exponentially tilted BFRY distribution (Lee et al., 2016; Bertoin et al., 2006), with pdf

$$g(s) = \frac{\alpha s^{-1-\alpha} e^{-\xi s}(1-e^{-ts})}{\Gamma(1-\alpha)((t+\xi)^\alpha - \xi^\alpha)}.$$

We can sample easily from this distribution by inversion. The variances $\lambda_{p,j}$ follow a generalised BFRY distribution with density:

$$f_p(x) = \frac{\tau\alpha x^{-\tau-1}}{\Gamma(1-\alpha)((t+1)^\alpha - 1)}\left(\gamma(\tau-\alpha, x) - \frac{\gamma(\tau-\alpha, (t+1)x)}{(t+1)^{\tau-\alpha}}\right),$$

where $t = \left(\frac{p\alpha\tau}{\eta(\tau-\alpha)}\right)^{1/\alpha}$ and $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ denotes the lower incomplete gamma function

Under this model, $\sum_j \lambda_{p,j} \xrightarrow{\mathrm{d}} \mathrm{ID}(0,\rho)$ where the limiting Lévy measure ρ is a generalised gamma Pareto measure, introduced by Ayed et al. (2019, 2020):

$$\rho(dx) = \frac{\eta}{\Gamma(1-\alpha)} x^{-1-\tau}\gamma(\tau-\alpha, x)dx.$$

As shown by Ayed et al. (2020), the tail Lévy intensity of this measure shows power-law behaviours at both 0 and ∞:

$$\overline{\rho}(x) \overset{x\to 0}{\sim} c_1 x^{-\alpha}, \qquad \overline{\rho}(x) \overset{x\to\infty}{\sim} c_2 x^{-\tau}$$

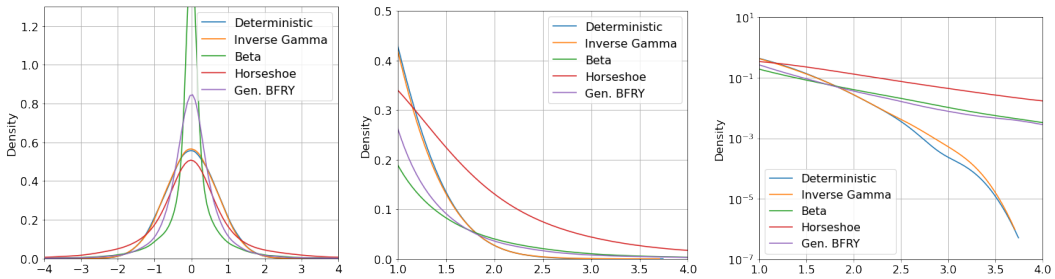
for some constants $c_1, c_2 > 0$. The exponents $\alpha \in (0,1)$ and $\tau > \alpha$ here can take different values, allowing for different asymptotic behaviours for small and large weights.

7. Illustrative Experiments

7.1 MoGP at Initialisation

In this subsection, we illustrate the key benefits of the MoGP regime as well as our main results through simulations; we consider a FFNN model defined by Equations (6) to (10), with no bias,

3. We say that a Lévy measure ρ on $(0,\infty)$ has a power-law tail at ∞ with exponent $\tau \in \mathbb{R}$ if its tail Lévy intensity $\overline{\rho}$ satisfies $\overline{\rho}(x) \overset{x\to\infty}{\sim} cx^{-\tau}$. Similarly, we say that ρ has a power-law tail at 0 with exponent $\alpha > 0$ if $\overline{\rho}(x) \overset{x\to 0}{\sim} cx^{-\alpha}$.



(a) Full distributions (b) Tails (c) Tails in log scale

Figure 2: MoGP output distribution

Width	Deterministic	Inverse Gamma	Beta	Horseshoe	Gen. BFRY
100	0.019994	0.113897	0.320444	0.691159	0.33325
500	0.00539	0.028584	0.281498	0.434425	0.219763
1000	0.005495	0.015217	0.279571	0.995462	0.316032
2000	0.001844	0.004522	0.297515	0.253737	0.235673

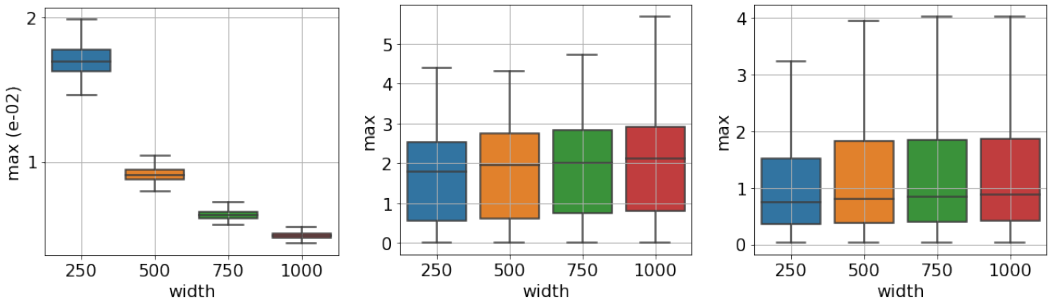
Table 3: MoGP output correlation.

$\sigma_v = 1$, ReLU activation and univariate inputs. For the variance distributions, we consider five of the examples described in Section 6, namely the deterministic, inverse-gamma, beta, horseshoe and generalised BFRY models. For all models except the horseshoe, we set the parameters such that $\mathbb{E}[\sum_j \lambda_{p_1,j}] \to 1$ as $p_1 \to \infty$. For the horseshoe model, we take $\lambda_{p_1,j} = (\frac{\pi}{2})^2 \frac{U_j^2}{p_1^2}$ where $U_j \sim \text{Cauchy}_+(0,1)$. Unless otherwise stated, the neural networks have a single hidden layer, recovering the illustrative example described in the introduction.

Output distribution. Figure 2 shows the distribution of the output with a large width $p_1 = 2000$. We use 50000 samples from the model to draw the plots, each corresponding to a random realisation of the weights. The figure confirms the limiting behaviour described in Theorem 8: the deterministic and inverse-gamma converge to the same Gaussian Process (the orange and blue lines overlap), whereas MoGP regimes offer a wider class of output distributions. In particular, when we examine the densities in log-log scale, we can notice that the beta, horseshoe and generalised BFRY exhibit a density with a power-law tail (straight line in log-log scale), whereas the deterministic and inverse-gamma exhibit a light-tailed density.

Dependence of the dimensions on the output. Another key consequence of Theorem 8 is that in the GP regime, the different dimensions of the output are asymptotically independent, while this is not the case in the MoGP regime. For a two-dimensional output FFNN, we report in Table 3 the empirical correlation between $(Z_1^{(2)}(x;p_1))^2$ and $(Z_2^{(2)}(x;p_1))^2$ when $p_1 \to \infty$ for the different models using 5000 random samples. The empirical results confirm the theoretical ones: we can see that for the deterministic and inverse-gamma models, the correlation converges to zero, while this is not the case for the other models.

Distribution of the largest weight. Proposition 4 describes another benefit of the MoGP regime: when the Levy measure is trivial, i.e. in the GP regime, the largest weight in each layer converges in probability to zero, while this is not the case in the MoGP regime. Figure 3 empirically validates this result; we show the evolution of the distribution of $\max_{1\leq j,k\leq p} |W_{jk}^{(2)}|$ as the width p_1 grows. This property can have a significant impact on the performance of the models since some weights remain non-negligible asymptotically and can be connected to nodes representing important



(a) Deterministic (b) Beta (c) Gen. BFRY

Figure 3: Distribution of the largest weight when the width increases.

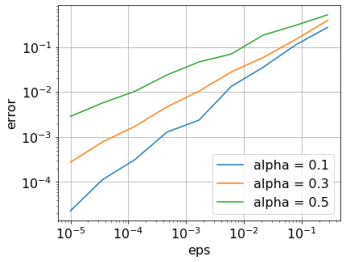


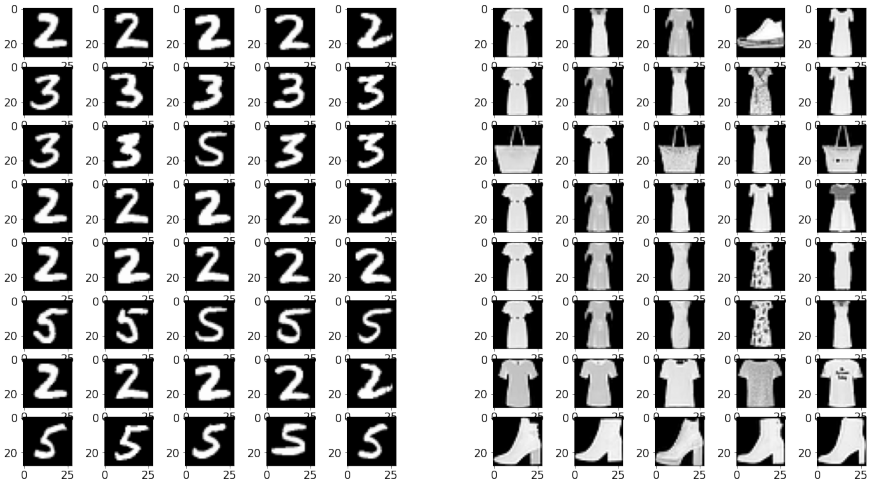
Figure 4: Expected truncation error as a function of ϵ (in log-log scale).

hidden features. This, coupled with a heavy-tailed distribution of the nodes, can favour specialisation of the neurons, with benefits for pruning and feature learning. We refer the reader to Sections 7.2 and 7.3 for experiments with real data where our proposed framework is used either in a frequentist or Bayesian fashion.

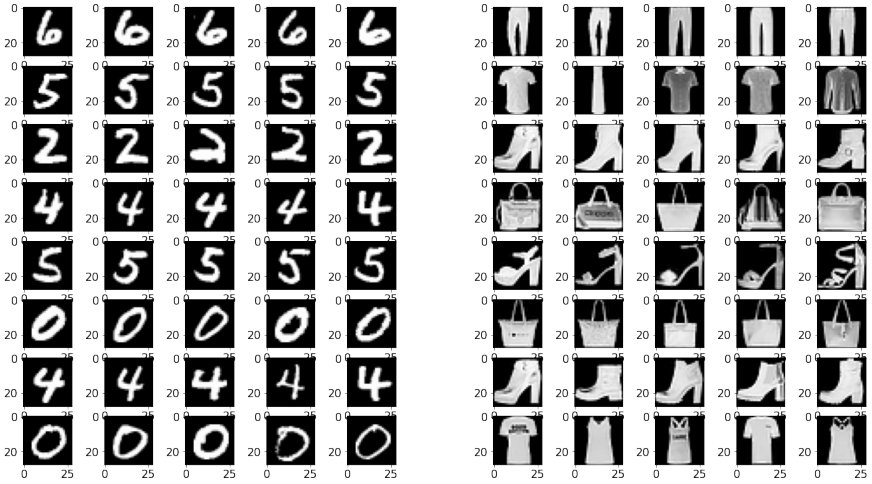
Truncation error. In Figure 4, we illustrate Corollary 13 with a generalised BFRY model with different values of α (and $\tau = 5$ is fixed). The expectation is estimated using 1000 simulations with width $p_1 = 2000$ and depth $L = 3$. The empirical results match well the theoretical bound. In particular, in log-log scale, we get an empirical slope of 0.492 for $\alpha = 0.5$, an empirical slope of 0.691 for $\alpha = 0.3$ and an empirical slope of 0.920 for $\alpha = 0.1$. Therefore, the slope is approximately equal to $1 - \alpha$, which confirms the theoretical rate of decay of the expected pruning error as a function of the truncation level ϵ. We get similar results with different depths L.

In Appendix F, we report further experiments analysing the vanishing/exploding gradient phenomenon in the MoGP context for deep networks (up to 20 layers). We also show how it can be alleviated with the right choice of model parameters.

The following two subsections describe how one can use our proposed framework with real data, either as a regularisation term or as a prior for a Bayesian Neural Network. We illustrate the discussed benefits of the proposed framework on compressibility and feature learning. The datasets considered in our experiments are MNIST and Fashion MNIST. Both datasets correspond to an image classification problem with 10 classes (digits for MNIST and clothes type for Fashion MNIST). The images are grey-scale and split between training and test sets, composed of 60000 and 10000 examples.

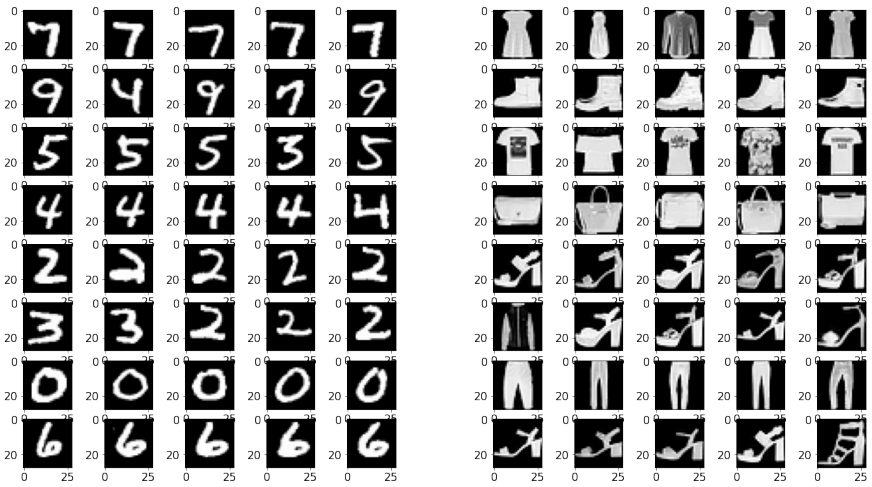


(a) Deterministic



(b) Horseshoe

Figure 5: Visualisation of the top-8 neurons of the first hidden layer of models trained on MNIST (left) and Fashion MNIST (right) — deterministic and horseshoe cases. Each row corresponds to a neuron. The elements of the row correspond to the ordered 5 images that maximise the neuron output.



(a) Gen. BFRY

Figure 6: Visualisation of the top-8 neurons of the first hidden layer of models trained on MNIST (left) and Fashion MNIST (right) — generalised BFRY case. Each row corresponds to a neuron. The elements of the row correspond to the ordered 5 images that maximise the neuron output.

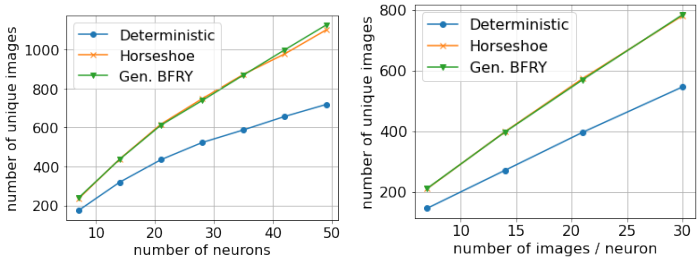
7.2 MoGP as a Regularisation

The most straightforward application of the MoGP framework is to use the prior as a regularisation term to add to the loss. We consider FFNN models f_θ with ReLU activation and three hidden layers, all having the same width $p = 2000$. The parameters $\theta = (\lambda_{p,j}^{(l)}, V_{jk}^{(l)}, B_k^{(l)})$ are trained using Adam optimisation for 50 epochs to minimise the objective function:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \gamma \left(\sum_{jkl} \log \pi_V(V_{jk}^{(l)}) + \sum_{lk} \log \pi_B(B_k^{(l)}) + \sum_{jl} \log \pi_\lambda(\lambda_{p,j}^{(l)}) \right) \tag{49}$$

where ℓ is a loss function, π_V and π_B are, respectively, the densities of zero-mean Gaussian distributions with variance σ_v^2 and σ_b^2, and π_λ is the density of the finite-dimensional approximation of the limiting infinitely divisible distribution. The parameter γ controls the weight of the prior. Notice that when $\gamma = 1$, we recover the maximum a posteriori estimator when ℓ is a log likelihood. In our experiments, we take $\sigma_v = \sigma_b = 1$ and $\gamma = 0.2$, and ℓ is the cross-entropy loss. We consider three examples detailed in Section 6, namely, the deterministic, the horseshoe and the generalised BFRY models. The deterministic and the horseshoe parameters are as in the simulated experiments. For the generalised BFRY, we set $\alpha = 0.8$ and $\tau = 5$. We bring to the reader's attention that for the deterministic model, the variance distribution is a Dirac at $1/p$; therefore, the network is trained with a similar parameterisation as the Neural Tangent Kernel framework (Jacot et al., 2018).

Feature learning. For each variance distribution, we train a network on MNIST and another on Fashion MNIST to minimise Equation (49). For all the models, we reach a test accuracy of approximately 98% on MNIST and 88% on Fashion MNIST, which are standard performances for feedforward models. Exact numbers can be found in the compressibility paragraph hereafter. We visualise the top-8 neurons of the first hidden layer of each model by plotting the 5 input images that maximise the neuron output. For the horseshoe and generalised BFRY, the top neurons are the ones with the highest variance $\lambda_k^{(1)}$. For the deterministic, since all the variances are equal, the top neurons are selected according to $\sum_k (W_{jk}^{(1)})^2$ (using this metric for the horseshoe and generalised



(a) (b)

Figure 7: Number of unique images when a) representing each neuron using the top-30 images that maximise the output of the top neurons, while varying the number of top neurons, and b) representing each neuron by a varying number of images, while fixing the number of neurons to thirty. Average of five runs for models trained on the MNIST dataset.

Truncation (i.e., $1-\kappa$)	Deterministic	Horseshoe	Gen. BFRY
0.0%	97.44 ($\pm$0.05)	97.94 ($\pm$0.09)	98.00 ($\pm$0.07)
80.0%	95.58 ($\pm$0.70)	97.94 ($\pm$0.09)	98.00 ($\pm$0.07)
90.0%	71.70 ($\pm$11.2)	97.94 ($\pm$0.09)	98.00 ($\pm$0.07)
95.0%	23.90 ($\pm$12.0)	97.94 ($\pm$0.09)	98.00 ($\pm$0.07)
98.0%	12.12 ($\pm$3.95)	64.22 ($\pm$13.1)	65.74 ($\pm$6.98)
98.5%	10.36 ($\pm$0.63)	44.14 ($\pm$10.9)	50.76 ($\pm$2.90)

Table 4: MNIST truncation accuracy: Average accuracy and standard deviation (between parenthesis) using five independent runs.

BFRY leads to similar results). Figures 5 and 6 reveal a key distinction between the MoGP regime (horseshoe and generalised BFRY) and the standard GP regime (deterministic). In the former regime, the top neurons tend to be more specialised: each neuron learns a different feature. In the latter regime, several top neurons learn the same features, which is materialised by almost equal lines in Figures 5 and 6, such as the neurons 1, 4, 5, and 7 of the network trained on MNIST with the deterministic model. To validate this phenomenon, we repeat the training of each model five times. In Figure 7, we plot the evolution of the average total number of unique images among the representative ones as a function of the number of top neurons, and also as a function of the number of images considered per neuron. The total number of unique images is interpreted as a simple metric to quantify the diversity of the learned features. The curves validate our hypothesis; in the MoGP regime, the top neurons learn more heterogeneous features.

Compressibility. We expect the higher diversity of the features learnt by the top neurons to affect the compressibility of the networks. We compare the degradation of the accuracy of the models with node pruning. For the horseshoe and generalised BFRY, we prune the nodes as described in Section 3.5 using the node variances $\lambda_j^{(l)}$. For the deterministic model, we use $\sum_k (W_{jk}^{(1)})^2$. For each layer, a given fraction κ of the nodes is kept. The mean and standard deviation of the accuracies are reported in Table 4 for MNIST and Table 5 for Fashion MNIST. As expected, the horseshoe and generalised BFRY outperform the deterministic model, with a slight advantage for the latter. What is even more interesting is that the accuracies of the pruned generalised BFRY models have a smaller variance. Though both the horseshoe and generalised BFRY have a power-law tail, the

Truncation (i.e., $1-\kappa$)	Deterministic	Horseshoe	Gen. BFRY
0.0%	87.98 ($\pm$0.19)	88.70 ($\pm$0.20)	88.54 ($\pm$0.19)
80.0%	86.24 ($\pm$0.80)	88.70 ($\pm$0.20)	88.54 ($\pm$0.19)
90.0%	60.24 ($\pm$5.14)	88.68 ($\pm$0.19)	88.56 ($\pm$0.18)
95.0%	19.64 ($\pm$7.01)	88.50 ($\pm$0.12)	88.40 ($\pm$0.25)
98.0%	10.84 ($\pm$1.17)	76.56 ($\pm$3.35)	77.24 ($\pm$2.32)
98.5%	10.26 ($\pm$0.58)	58.26 ($\pm$14.2)	60.44 ($\pm$3.42)

Table 5: Fashion MNIST truncation accuracy: Average accuracy and standard deviation (between parenthesis) using five independent runs.

tail of the horseshoe is heavier; in particular, the distribution has an infinite expectation, which is not the case for the generalised BFRY. This can explain the difference between the models in terms of variances. We believe this simple experiment serves as motivation to further explore the MoGP regime beyond the horseshoe model, as different limiting distributions can offer valuable practical advantages.

In Appendix F, we empirically verify on the Cifar10 dataset that using the MoGP framework as a regularisation also improves the compressibility of convolutional neural networks.

7.3 MoGP in a Fully Bayesian Setting

We further demonstrate the MoGP in a fully Bayesian setting, where we simulate the posterior distribution of a FFNN with MoGP priors on the weights. Let f_θ be a FFNN with ReLU activation and three hidden layers of width $p = 2000$. The log joint-density for classification with this FFNN is then given as:

$$\log g(\theta) = \sum_i \log h(y_i, f_\theta(x_i)) + \sum_{j,k,l} \log \pi_V(v_{jk}^{(l)}) + \sum_{l,k} \log \pi_B(b_k^{(l)}) + \sum_{j,l} \log \pi_\lambda(\lambda_j^{(l)}). \tag{50}$$

We consider the C-way classification problem where $y_i \in \{1, \dots, C\}$ and $h(y, f_\theta(x))$ is the categorical likelihood, i.e., $h(y, f_\theta(x)) = \text{softmax}(f_\theta(x))_y$, with $\text{softmax}(f_\theta(x))_c = \frac{\exp(f_\theta(x)_c)}{\sum_{c'=1}^{C} \exp(f_\theta(x)_{c'})}$ to get proper probability vectors.

We compare the deterministic, the horseshoe and the generalised BFRY models on MNIST and Fashion MNIST datasets. We infer the posteriors of the network weights via Stochastic Gradient Hamiltonian Monte-Carlo (SGHMC) (Chen et al., 2014) with batch size set to 100. We run the samplers for 100 epochs through datasets and collect samples every 2 epochs after 50 burn-in epochs. Following Zhang et al. (2020), we adopt a simple cosine-annealed step size with a single cycle and set the first half of the epochs as an exploration stage (updating without noise for quick convergence to a local minimum). For the generalised BFRY, considering the importance of the hyperparameter α, we introduce a uniform prior on it and infer its value along with the model parameters. We run every experiment five times and averaged results.

Compressibility. As in Section 7.2, we first compare the predictive classification accuracies of the FFNN models under a varying truncation ratio. For all models, we collect 25 samples after 50 burn-in epochs (collecting a sample at the end of every 2 epochs after the burn-in), and the test accuracy is measured with Monte-Carlo estimates of predictive distributions,

$$p(y_*|x_*, \mathcal{D}) \approx \frac{1}{S} \sum_{s=1}^{S} \text{softmax}(f_{\theta^{(s)}}(x_*))_{y_*}, \tag{51}$$

Truncation (i.e., $1-\kappa$)	Deterministic	Horseshoe	Gen. BFRY
0.0%	90.23 ($\pm$0.12)	97.83 ($\pm$0.07)	97.78 ($\pm$0.10)
80.0%	13.14 ($\pm$3.18)	97.93 ($\pm$0.06)	97.71 ($\pm$0.10)
90.0%	9.98 ($\pm$0.82)	97.73 ($\pm$0.04)	97.72 ($\pm$0.05)
95.0%	9.70 ($\pm$0.89)	97.59 ($\pm$0.06)	97.68 ($\pm$0.03)
98.0%	9.46 ($\pm$0.52)	87.97 ($\pm$3.74)	54.90 ($\pm$4.13)
98.5%	9.46 ($\pm$0.52)	89.29 ($\pm$3.58)	57.69 ($\pm$11.15)

Table 6: Predictive classification accuracy on MNIST dataset under various truncation ratio.

Truncation (i.e., $1-\kappa$)	Deterministic	Horseshoe	Gen. BFRY
0.0%	80.65 ($\pm$0.12)	87.72 ($\pm$0.12)	87.75 ($\pm$0.05)
80.0%	10.00 ($\pm$0.00)	87.57 ($\pm$0.28)	87.28 ($\pm$0.29)
90.0%	10.00 ($\pm$0.00)	87.43 ($\pm$0.28)	87.01 ($\pm$0.29)
95.0%	10.00 ($\pm$0.00)	87.27 ($\pm$0.32)	86.64 ($\pm$0.54)
98.0%	10.00 ($\pm$0.00)	80.65 ($\pm$3.73)	81.85 ($\pm$3.78)
98.5%	10.00 ($\pm$0.00)	59.34 ($\pm$6.02)	68.34 ($\pm$6.71)

Table 7: Predictive classification accuracy on Fashion MNIST under various truncation ratio.

where $\mathcal{D}$ is the training set and $\theta_1, \ldots, \theta_S$ are samples collected from SGHMC. As in Section 7.2, we prune the nodes with respect to the magnitude of the node variances $\lambda_j^{(l)}$ and measure the test accuracies of the pruned networks. Tables 6 and 7 summarise the results. FFNNs with horseshoe or generalised BFRY priors are more robust to truncation; both maintain decent classification accuracies even when 95% of the neurons are truncated. For the generalised BFRY, we present the posterior samples of the hyperparameter α in Figure 8. The posteriors are well concentrated around the range $[0.7, 0.8]$.

Feature learning. Finally, to demonstrate the feature learning aspect of MoGP priors, we conduct a transfer learning experiment. We start by taking an external dataset, not used during training, and split it into two halves D_1 and D_2. Then, we sort the neurons of the first hidden layers of the

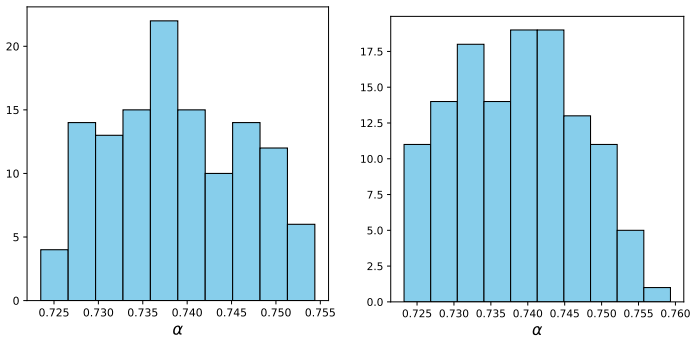


Figure 8: Inferred α values for gen. BFRY on MNIST (left) and Fashion MNIST (right).

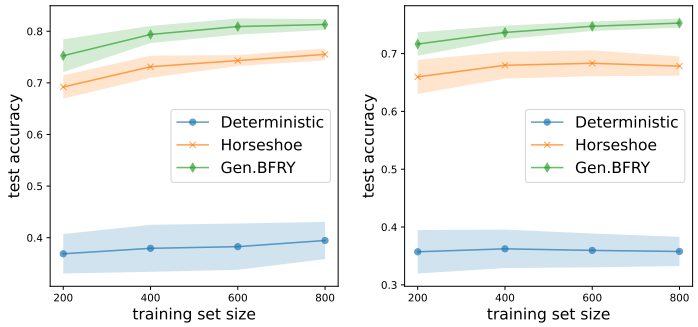


Figure 9: Comparing feature learning aspect of FFNN models via transfer learning. Results for MNIST (left) and Fashion MNIST (right).

trained FFNNs with respect to their average magnitudes of activations over the images in D_1, and select the top-k activated neurons. Next, for each of these top-k activated neurons, we select m images from D_1 which most strongly activate this neuron, and use the combined collection of these images (over all the top-k activated neurons) to form a subset that will then be used for transfer learning. Each image in this subset is then represented by a vector consisting of the activations of the top-activated neurons. For instance, choosing the top-30 activated neurons and the top-10 activating images per neuron, results in 300 images in total in this subset (assuming no overlaps in top-activating images). Each image in the subset is represented as a 30-dimensional vector which is just the concatenation of activations from the top-30 activated neurons. Our hypothesis here is that if a trained neural network exhibits feature learning, the subset of images selected in this way is representative enough to express the important features in the set D_1, and so, if we train a new classifier based on this subset, the resulting model should generalise well to D_2. To validate this hypothesis, we train a light-weight FFNN with one hidden layer using the selected subset with the selected vectors of activations. Then we evaluate the test accuracy of the trained light-weight FFNN using the vector activation form of D_2 computed from the neurons selected with D_1. Since we are working with a Bayesian model, for each configuration, we have multiple samples of parameters. We first compute the Monte-Carlo estimates of the average activations using those samples, and then use the estimates to sort neurons, select images and form vectorized activations. We perform this transfer learning experiment with varying subset sizes, repeating all experiments five times per configuration and take average results. Figure 9 summarises the result. As we expect, the horseshoe and generalised BFRY models transfer well, while the deterministic model fails to transfer. In particular, the test accuracy of the deterministic model does not increase as the size of the training set increases, demonstrating that the model does not exhibit feature learning.

We comment that unlike our results in Section 7.2, the top-5 activating images for each of the top-8 activated neurons do not show a noticeable difference between the deterministic case and the rest in this fully Bayesian setting. See Figure 15 in Appendix F for the visualisation of those images in the deterministic, the horseshoe and the generalised BFRY cases.

8. Discussion and Further Directions

Models with iid non-Gaussian weights. Neal (1996)'s seminal work on infinite-width limits of shallow neural networks includes the case where weights are initialised by an iid symmetric stable distribution. A subsequent thorough analysis (Der and Lee, 2006) showed that such networks converge to stable processes as the widths increase, and this analysis was generalised from shallow networks to deep networks (Favaro et al., 2020; Bracale et al., 2021; Jung et al., 2023). This line of work (the iid model) and our dependent-weight model can both lead to infinite-width limits

that are heavy-tailed non-Gaussian processes; however, the sources of heavy-tails in the two cases are different. In the former (the iid model), the source is the use of a non-Gaussian, heavy-tailed distribution for initialising weights, which corresponds to assuming in our setup that the $V_{jk}^{(l)}$ are sampled independently from a heavy-tailed distribution, instead of a Gaussian distribution. In the latter (dependent-weight model), on the other hand, the source is the use of a per-node random variance that is shared by the weights of all outgoing edges from a given node. As a result, the limiting networks of these two classes of models have different properties. In the former, nodes in a layer of a limiting network are always independent, while in the latter, they are usually dependent. Also, in the former, the limiting networks exist largely due to a normalisation adjusted for the sum of heavy-tailed non-Gaussian random variables, while in the latter, the limiting networks exist because the random variances of nodes in a layer remain summable at the infinite-width limit. One interesting future direction is to study the generalisation of our model class where the $V_{jk}^{(l)}$ are sampled independently from a stable or heavy-tailed distribution using techniques developed in (Neal, 1996; Der and Lee, 2006; Favaro et al., 2020; Bracale et al., 2021; Jung et al., 2023). Such a generalisation would be a good starting point for analysing the pruning of not only network nodes but also edges. Also, our tools for handling the limiting random variables with infinitely divisible distributions, and for analysing the tail behaviour and pruning of the limiting neural networks, may help extend results of that line of work. Finally, we point out that Lee et al. (2022) studied the infinite-width limits of neural networks where weights are initialised by iid Gaussians but the shared variance of weights in the last readout layer is made random. This simple change caused the limiting networks to be mixtures of Gaussian processes, increasing the expressivity of those networks without sacrificing the tractability of Gaussian processes for inference much. Our work differs from Lee et al. (2022) in that our models use per-node random variances, while theirs use per-layer random variances.

Row-column exchangeable models. Tsuchida et al. (2019) introduced a general class of deep FFNNs with dependent weights, and analysed their infinite-width limits. They consider[4] that, at a given layer, the weights take the form $W_{jk} = F_{jk}/\sqrt{p}$, where the (infinite) array of random variables (F_{jk}) is assumed to be row-column exchangeable (RCE), that is $(F_{\pi_1(j)\pi_2(k)}) \overset{\mathrm{d}}{=} (F_{jk})$ for any permutations π_1 and π_2 of $\mathbb{N}$. By the Aldous-Hoover theorem, any RCE array admits the representation $F_{jk} = f(A, B_j, C_k, D_{jk})$ for some measurable function f and some iid random variables A, B_j, C_k, D_{jk}. A crucial thing to note is that here none of the random variables depends on the width p. In contrast, our model assumes that the rows are independent, and their distributions may depend on p in a non-trivial way. The models presented in Section 6.4, which converge to a GP, fall into this RCE class, as they can be expressed as $W_{jk} = \frac{\sqrt{Y_j}}{\sqrt{p}} V_{jk}$. More generally, Tsuchida et al. (2019) show that RCE models converge to a MoGP. The properties of the infinite-width neural network are, however, very different from those obtained in this article in the MoGP regime. A key difference between the infinitely divisible models considered here and RCE models is that in the latter, the weights still converge uniformly to 0 in the infinite-width limit: $\max_{j=1,\dots,p} |W_{jk}| = \frac{1}{\sqrt{p}} \max_{j=1,\dots,p} |F_{jk}| \overset{\mathrm{pr}}{\to} 0$ as $p \to \infty$. This follows from the fact that (F_{jk}) are exchangeable in j, and thus conditionally iid. RCE therefore cannot exhibit the behaviour described in Proposition 4.

Lévy adaptive regression kernels. In the case of a single hidden layer with location parameter $a = 0$, the infinite-width model falls in the class of Lévy adaptive regression kernels (Wolpert et al., 2011), which is explored by Jang et al. (2017) for kernel learning.

Lévy measures and sparsity-promoting priors. The link between Lévy measures and sparsity promoting models has been explored in the high-dimensional (sparse) linear regression setting, see e.g. the works of Caron and Doucet (2008); Polson and Scott (2012); McCullagh and Polson (2018).

4. Under the assumtion that $\mathbf{E}[F_{jk}] = 0$

MoGP as a stationary distribution for trained networks with SGD. Understanding the statistical properties of trained neural networks is a complex endeavor due to the non-convexity of deep learning problems. The past decade has seen considerable efforts to provide theoretical tools to understand neural networks trained with SGD. A fruitful line of research proposes to view the SGD dynamics with a fixed step size as a stationary stochastic process (Mandt et al., 2016; Hu et al., 2017; Chaudhari and Soatto, 2018; Zhu et al., 2019). The intuition is that with a constant step size, SGD first moves towards a valley of the objective function and then bounces around because of sampling noise in the gradient estimate. This stationary distribution governs the statistical properties of the trained networks. The core question is then how to model the stationary distribution. In Shin (2021); Barsbey et al. (2021), the authors base their models on the fact that when the networks are trained with large learning rates or small batch sizes, heavy-tailed stationary distributions emerge (Hodgkinson and Mahoney, 2021; Gurbuzbalaban et al., 2021). Under such a hypothesis, the authors derive a generalisation bound as a function of the tail index of the weights. Both papers assume that the weights of the trained network are asymptotically independent when the width is large enough. In Shin (2021), the weights follow a Pareto distribution, whereas in Barsbey et al. (2021), the weights follow a generic heavy-tailed distribution. One might explore whether our MoGP framework might be used as a possible stationary distribution of a trained network, so that similar generalisation bounds could then be derived (see Section 3.5 for more details). Our proposed framework has the additional benefit of readily taking into account the summability of the weights as the width goes to infinity, whereas other models would need to impose an additional scaling factor.

Infinite networks with bottlenecks. In order to allow for feature/representation learning in infinitely-wide neural networks, Aitchison (2020) recently proposed the use of infinite neural networks with bottlenecks. The idea is to consider the same iid Gaussian assumptions on the weights as for NNGP, but to set one layer (the bottleneck layer) to have a fixed number of hidden nodes, while taking the number of nodes in other layers to infinity. As shown by Aitchison (2020), the resulting model no longer converges to a Gaussian process, but rather to a Gaussian process with a random kernel, and this then allows for representation learning. The training dynamics of such a model is also partially analysed (Littwin et al., 2021).

Our general approach allows to construct neural networks with bottlenecks similar to (Aitchison, 2020); the difference is that in our case, the limiting model is obtained by taking the widths of all layers to infinity, including the bottleneck layer, but the number of active nodes in the bottleneck layer remains finite in this limit, and converges to a Poisson distribution. For example, one way to achieve this is to use, in the bottleneck layer, the Bernoulli prior described in Section 6.2 where, for $p_l \geq c$, $\lambda^{(l)}_{p_l,j} \sim \text{Bernoulli}(c/p_l)$, for some $c > 0$. The limit has parameters $a^{(l)} = 0$ and $\rho^{(l)} = c\delta_1$. As p_l tends to infinity, the number of active hidden nodes of the bottleneck layer, that is the number of nodes with $\lambda_{p_l,j} > 0$, converges to a Poisson random variable: $\sum_{j=1}^{p_l} \mathbf{1}_{\{\lambda_{p_l,j}>0\}} \xrightarrow{\text{d}} \text{Poisson}(c)$. Using $\lambda^{(l)}_{p_l,j} = c_2/p_l$ for all other layers, we obtain a model where the bottleneck layer has a random number of hidden nodes.

Deep Gaussian processes. Deep Gaussian processes (Damianou and Lawrence, 2013; Bui et al., 2016; Dunlop et al., 2018) are hierarchical models where each layer is described by a latent variable Gaussian process. That is, they denote a random function $g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form $g(\mathbf{x}) = (\mathbf{f}^{(L+1)} \circ \ldots \circ \mathbf{f}^{(1)})(\mathbf{x})$, where for each $l = 1, \ldots, L+1$, the l-th GP layer $\mathbf{f}^{(l)}$ has width p_l and is defined as follows:

$$\mathbf{f}^{(l)} : \mathbb{R}^{p_{l-1}} \to \mathbb{R}^{p_l}, \qquad \mathbf{f}^{(l)}(\mathbf{z}) = (f^{(l)}_1(\mathbf{z}), \ldots, f^{(l)}_{p_l}(\mathbf{z}))$$

with $f^{(l)}_j \overset{\text{iid}}{\sim} \text{GP}(0, k^{(l)})$ for some kernel $k^{(l)} : \mathbb{R}^{p_{l-1}} \times \mathbb{R}^{p_{l-1}} \to \mathbb{R}$ of the l-th layer. As noted by a number of authors, a finite neural network with iid Gaussian weights is a deep Gaussian process. Similarly, for our finite model, *conditionally on the set of variances* $(\lambda^{(l)}_{p_l,j})_{l=1,\ldots,L;j=1,\ldots,p_l}$,

the model can be seen as a deep Gaussian process, whose kernels $k^{(l)}$ depends on $(\lambda^{(l)}_{p_l,j})_{j=1,\dots,p_l}$. Assume $a^{(l)} = 0$. In the infinite limit, as is apparent from Theorem 16, we still have a deep Gaussian process, *conditionally on the point processes* $\{\widetilde{\lambda}^{(l)}_j\}_{j\geq 1}$ with mean measure $\rho^{(l)}$. The Gaussian process kernels are given by

$$k^{(1)}(\mathbf{x},\mathbf{x}') := \sigma_b^2 + \sigma_v^2 \frac{\mathbf{x}^T\mathbf{x}'}{d_{\text{in}}}, \qquad k^{(l+1)}(\mathbf{z},\mathbf{z}') := \sigma_b^2 + \sigma_v^2 \sum_{j\geq 1} \widetilde{\lambda}^{(l)}_j \phi(z_j)\phi(z'_j) \text{ for } l = 1,\dots,L.$$

Neural tangent kernels. An interesting direction would be to investigate the behaviour of the neural network during training by gradient descent when initialised with the MoGP model. In the iid Gaussian case, it has been shown that the evolution of the neural network can be described by a fixed kernel (neural tangent kernel) in the infinite-width limit (Jacot et al., 2018), and that this kernel remains constant over training. We conjecture that in the MoGP case, when the Lévy measure is non-trivial, we would obtain a random kernel in the infinite-width limit, where this kernel evolves over training.

Acknowledgments

We would like to extend our gratitude to the Action Editor, Dan Roy, and the two anonymous reviewers for their detailed and insightful feedback. In particular, their contributions significantly enhanced the clarity and rigour of Theorem 16's statement and proof. HL and PJ were funded in part by the National Research Foundation of Korea (NRF) grants NRF-2017R1A2B200195215 and NRF-2019R1A5A102832421. HL was funded in part by NRF grant NRF-2023R1A2C100584311. HY was supported by the Engineering Research Center Program through the National Research Foundation of Korea funded by the Korean government MSIT (NRF-2018R1A5A1059921) and also by the Institute for Basic Science (IBS-R029-C1). JL was supported by an Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government MSIT (No.2019-0-00075), Artificial Intelligence Graduate School Program (KAIST).

Appendices

Organisation of the appendices. Appendix A contains background material on positive homogeneous functions, ReLU kernels, regularly varying functions, stable distributions, conditions for the convergence to infinitely divisible random variables and Lévy measures. Appendix B gives and proves some limit theorems on independent triangular arrays. These results form the building blocks of the proofs of the main theorems and propositions, which are given in Appendix C. Appendix D provides additional secondary theoretical results on the properties of small weights in our model, and on the infinite-width limit for multiple inputs in the symmetric α-stable case. Appendix E provides details of the derivations for the examples given in Sections 1 and 6 and additional properties concerning these examples. It also provides a general recipe to construct novel examples, and describes new concrete examples. Finally, Appendix F provides additional experimental results, including stability with respect to depth and our experimental findings on convolutional neural networks.

Appendix A. Background Material

A.1 Background on Positive Homogeneous Functions

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a positive homogeneous function. That is, for all $\alpha > 0$,

$$\phi(\alpha x) = \alpha \phi(x).$$

Define $C_{\text{Lip}} := \max(|\phi(1)|, |\phi(-1)|)$. Then, ϕ is C_{Lip}-Lipschitz continuous and satisfies $|\phi(x)| \leq C_{\text{Lip}}|x|$ for all x.

Proof We first show that $|\phi(x)| \leq C_{\text{Lip}}|x|$ for all x. When x is 0, $\phi(0) = \alpha\phi(0)$ for all $\alpha > 0$ and so $\phi(0) = 0$. When x is not 0,

$$|\phi(x)| = |x| \left|\phi\left(\frac{x}{|x|}\right)\right| \leq |x| \max(|\phi(1)|, |\phi(-1)|) = |x| C_{\text{Lip}}.$$

Next we show the Lipschitz continuity. Let $x, y \in \mathbb{R}$. If both x and y are nonnegative, then

$$|\phi(x) - \phi(y)| = |x\phi(1) - y\phi(1)| = |\phi(1)||x - y| \leq C_{\text{Lip}}|x - y|.$$

If both x and y are nonpositive, then

$$|\phi(x) - \phi(y)| = |(-x)\phi(-1) - (-y)\phi(-1)| = |\phi(-1)||x - y| \leq C_{\text{Lip}}|x - y|.$$

The remaining case is that one of x and y is positive and the other is negative. Without loss of generality, we may assume that x is positive. Then,

$$|\phi(x) - \phi(y)| = \big||x|\phi(1) - |y|\phi(-1)\big| \leq |x||\phi(1)| + |y||\phi(-1)| \leq C_{\text{Lip}}(|x| + |y|) = C_{\text{Lip}}|x - y|.$$

■

A.2 Background on ReLU Kernels

Following (Cho and Saul, 2009), we have, for $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}(0, \Sigma)$ and $\alpha \geq 0$,

$$\mathbf{E}\left[\max(0, X)^\alpha \max(0, Y)^\alpha\right] = \frac{1}{2\pi} \left(\Sigma_{11}\Sigma_{22}\right)^{\alpha/2} J_\alpha(\theta)$$

where

$$J_\alpha(\theta) = \Gamma(\alpha+1)\sin^{2\alpha+1}(\theta)\int_0^{\pi/2}\frac{\cos^\alpha(x)}{(1-\cos(\theta)\cos(x))^{\alpha+1}}dx$$

with

$$\theta = \arccos(\rho) \quad \text{and} \quad \rho = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}.$$

Furthermore, if $\alpha \in \mathbb{N}$,

$$J_\alpha(\theta) = (-1)^\alpha \sin^{2\alpha+1}(\theta)\left(\frac{1}{\sin(\theta)}\frac{\partial}{\partial\theta}\right)^\alpha\left(\frac{\pi-\theta}{\sin(\theta)}\right)$$

In particular, we have

$$\begin{aligned}
J_0(\theta) &= \pi - \theta \\
J_1(\theta) &= \sin(\theta) + (\pi-\theta)\cos(\theta) \\
J_2(\theta) &= 3\sin(\theta)\cos(\theta) + (\pi-\theta)(1+2\cos^2(\theta)).
\end{aligned}$$

For $\alpha = 1/2$, we show that

$$J_{1/2}(\theta) = \sqrt{\frac{\pi}{2}}\left(2\,\text{EllipE}\left(\frac{\cos(\theta)+1}{2}\right) - (1-\rho)\,\text{EllipK}\left(\frac{\cos(\theta)+1}{2}\right)\right)$$

where EllipK and EllipE are respectively the complete elliptical integrals of the first and second kind, defined by

$$\text{EllipK}(m) = \int_0^{\pi/2}(1-m\sin^2(t))^{-1/2}dt = \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-mt^2)}} \tag{52}$$

$$\text{EllipE}(m) = \int_0^{\pi/2}(1-m\sin^2(t))^{1/2}dt = \int_0^1 \frac{\sqrt{1-mt^2}}{\sqrt{1-t^2}}dt \tag{53}$$

which can be computed efficiently using the arithmetic-geometric mean. Writing the above expressions in terms of the correlation ρ, we obtain

$$\mathbf{E}\left[\max(0,X)^\alpha \max(0,Y)^\alpha\right] = \frac{1}{2\pi}\left(\Sigma_{11}\Sigma_{22}\right)^{\alpha/2}\kappa_\alpha(\rho)$$

where

$$\kappa_\alpha(\rho) = J_\alpha(\arccos(\rho)) = \begin{cases} \frac{\pi}{2} + \arcsin(\rho) & \text{if } \alpha = 0 \\ \sqrt{\frac{\pi}{2}}\left(2\,\text{EllipE}(\frac{\rho+1}{2}) - (1-\rho)\,\text{EllipK}(\frac{\rho+1}{2})\right) & \text{if } \alpha = 1/2 \\ \sqrt{1-\rho^2} + \left(\frac{\pi}{2} + \arcsin(\rho)\right)\rho & \text{if } \alpha = 1 \\ 3\sqrt{1-\rho^2}\rho + \left(\frac{\pi}{2} + \arcsin(\rho)\right)(1+2\rho^2) & \text{if } \alpha = 2 \end{cases} \tag{54}$$

using the identities $\pi - \arccos(x) = \frac{\pi}{2} + \arcsin(x)$ and $\sin(\arccos(x)) = \cos(\arcsin(x)) = \sqrt{1-x^2}$.

We also have

$$\mathbf{E}\left[\max(0,X)^{2\alpha}\right] = \frac{1}{2}\mathbf{E}\left[X^{2\alpha}\right] = (\Sigma_{11})^\alpha \frac{2^{\alpha-1}\Gamma(\alpha+1/2)}{\Gamma(1/2)} = \begin{cases} \frac{1}{2} & \text{if } \alpha = 0 \\ \sqrt{\frac{\Sigma_{11}}{2\pi}} & \text{if } \alpha = 1/2 \\ \frac{\Sigma_{11}}{2} & \text{if } \alpha = 1 \\ \frac{3}{2}\Sigma_{11}^2 & \text{if } \alpha = 2 \end{cases}.$$

Proof All of the above results are from (Cho and Saul, 2009), except for $J_{1/2}$ (or equivalently $\kappa_{1/2}$). Write $\kappa_{1/2}(\rho) = \Gamma(3/2)(1-\rho^2)f(\rho)$, where

$$\begin{aligned} f(\rho) &:= \int_0^{\pi/2} \frac{\sqrt{\cos(x)}}{(1-\rho\cos(x))^{3/2}} dx \\ &= \int_0^1 \frac{\sqrt{v}}{(1-\rho v)^{3/2}\sqrt{1-v^2}} dv. \end{aligned}$$

Integrating with respect to ρ and using Fubini's theorem,

$$\int_a^\rho f(u)du = 2\int_0^1 \frac{1}{\sqrt{v}(1-\rho v)^{1/2}\sqrt{1-v^2}} dv + const. \tag{55}$$

Using the change of variables $u = \sqrt{\frac{1+\frac{1}{v}}{2}}$, we have that

$$\begin{aligned} \int_0^1 \frac{1}{\sqrt{v}(1-\rho v)^{1/2}\sqrt{1-v^2}} dv &= \int_1^\infty \frac{4u}{(2u^2-1)^2\sqrt{\frac{1}{2u^2-1}}\sqrt{1-\frac{\rho}{2u^2-1}}\sqrt{1-\frac{1}{(2u^2-1)^2}}} du \\ &= \int_1^\infty \frac{4u}{\sqrt{2}\sqrt{u^2-\frac{\rho+1}{2}}\sqrt{2u^2-2}\sqrt{2u^2}} du \\ &= \sqrt{2}\int_1^\infty \frac{du}{\sqrt{u^2-\frac{\rho+1}{2}}\sqrt{u^2-1}} \\ &= \sqrt{2}\,\mathrm{EllipK}\left(\frac{\rho+1}{2}\right). \end{aligned}$$

Differentiating Equation (55) with respect to ρ gives

$$f(\rho) = \frac{\sqrt{2}}{1-\rho^2}\left(2\,\mathrm{EllipE}\left(\frac{\rho+1}{2}\right) - (1-\rho)\,\mathrm{EllipK}\left(\frac{\rho+1}{2}\right)\right).$$

■

A.3 Useful Lemmas on Regularly Varying Random Variables

A nonnegative random variable X is said to be regularly varying with index $\alpha \geq 0$ if and only if

$$\Pr(X > x) \overset{x\to\infty}{\sim} x^{-\alpha}L(x)$$

where L is a slowly varying function. That is, its survival function (a.k.a. complementary cumulative distribution function) $\overline{F}(x) := \Pr(X > x)$ is regularly varying with index $-\alpha$.

Lemma 19 *(Jessen and Mikosch, 2006, Lemma 4.1.(i)) Assume that X_1 and X_2 are independent nonnegative random variables and that X_1 is regularly varying with index $\alpha > 0$. If either X_2 is regularly varying with index $\alpha > 0$ or $\Pr(X_2 > x) = o(\Pr(X_1 > x))$, then X_1X_2 is regularly varying with index $\alpha > 0$.*

Lemma 20 *(Jessen and Mikosch, 2006, Lemma 4.1.(iv)) Let $X_1, \ldots, X_p$ be iid nonnegative random variables with survival function satisfying $\Pr(X_1 > x) \overset{x\to\infty}{\sim} cx^{-\alpha}$ for some $c > 0$. Then,*

$$\Pr(X_1 \ldots X_p > x) \overset{x\to\infty}{\sim} \frac{\alpha^{p-1}c^p}{(p-1)!} x^{-\alpha}\log^{p-1} x.$$

Lemma 21 *(Jessen and Mikosch, 2006, Lemma 4.2) Assume that X_1 and X_2 are nonnegative independent random variables and that X_1 is regularly varying with index $\alpha > 0$.*

A.1 If there exists $\epsilon > 0$ such that $\mathbf{E}[X_2^{\alpha+\epsilon}] < \infty$, then

$$\Pr(X_1 X_2 > x) \overset{x\to\infty}{\sim} \mathbf{E}[X_2^{\alpha}] \Pr(X_1 > x). \tag{56}$$

A.2 If $\Pr(X_1 > x) \overset{x\to\infty}{\sim} cx^{-\alpha}$ and $\mathbf{E}[X_2^{\alpha}] < \infty$, then Equation (56) holds.

Proposition 22 *(Bingham et al., 1989, Theorem 8.2.1 p.341) Let $X \sim \mathrm{ID}(a, \rho)$ be a nonnegative infinitely divisible random variable. Let $\overline{F}(x) = 1 - F(x)$ be its survival function, where F is the cdf of X. Then, for all $\alpha \geq 0$, the tail Lévy intensity $\overline{\rho}$ is regularly varying with index $-\alpha$ if and only if $\overline{F}$ is also. Furthermore, in that case,*

$$\overline{\rho}(x) \overset{x\to\infty}{\sim} \overline{F}(x).$$

Proposition 23 *(Resnick, 2005, Proposition 5(iii)) Let G be a regularly varying function with index $\alpha \in \mathbb{R}$, and $(a_p)_p$ and $(b_p)_p$ be two sequences that satisfy $0 < a_p \to \infty$, $0 < b_p \to \infty$ and $a_p \overset{p\to\infty}{\sim} cb_p$ for some $0 < c < \infty$. Then,*

$$G(a_p) \overset{p\to\infty}{\sim} c^{\alpha} G(b_p).$$

Lemma 24 *(Feller, 1971, Lemma 2, VIII.8) If L is slowly varying at infinity, then for any $\delta > 0$, there exists x_0 such that $x^{-\delta} < L(x) < x^{\delta}$ for all $x > x_0$.*

A.4 Background on Positive Stable Random Variables

A (possibly degenerate) positive strictly stable random variable $X \sim \mathrm{Stable}(\alpha, \gamma)$ with stability exponent $\alpha \in (0, 1]$ and scale parameter $\gamma > 0$ has Laplace transform

$$\mathbf{E}[e^{-tX}] = e^{-(\gamma t)^{\alpha}}, \text{ for } t \geq 0.$$

It satisfies, for any $n \geq 1$, $\sum_{i=1}^{n} X_i \overset{\mathrm{d}}{=} n^{1/\alpha} X$ where $X_1, \ldots, X_n$ are iid copies of X, and is an important example of an infinitely divisible random variable. If $\alpha = 1$, $X = \gamma$ is degenerate at γ; for $\alpha \in (0, 1)$, it is non-degenerate with support $(0, \infty)$. In general, X is infinitely divisible with

$$X \sim \begin{cases} \mathrm{ID}(0, \rho_{\mathrm{stable}}(\,\cdot\,; \alpha, \gamma/\Gamma(1-\alpha)^{1/\alpha})) & \text{if } \alpha \in (0, 1) \\ \mathrm{ID}(\gamma, 0) & \text{if } \alpha = 1 \end{cases}$$

where $\rho_{\mathrm{stable}}(dx; \alpha, c)$ denotes the following alpha–stable Lévy measure on $(0, \infty)$ with stability exponent $\alpha \in (0, 1)$ and parameter $c > 0$:

$$\rho_{\mathrm{stable}}(dx; \alpha, c) := \alpha c^{\alpha} x^{-\alpha-1} \mathbf{1}_{\{x>0\}} dx. \tag{57}$$

In the special case $\alpha = \frac{1}{2}$, we have $\mathrm{Stable}(1/2, \gamma) = \mathrm{IG}(\frac{1}{2}, \frac{\gamma}{4})$, that is, the stable distribution with scale parameter γ corresponds to the inverse gamma distribution with shape $1/2$ and scale $\gamma/4$.

Remark 25 *Standard definitions of positive stable random variables are given for non-degenerate random variables, with $\alpha \in (0, 1)$. See e.g. (Samorodnitsky and Taqqu, 1994; Janson, 2011). We include here the degenerate case $\alpha = 1$, as this case relates to the Gaussian process limit, as we show in Section 5. For $\alpha \in (0, 1)$, our parameterisation $\mathrm{Stable}(\alpha, \gamma)$ corresponds to the standard four-parameters parameterisation (Janson, 2011, Theorem 3.3) $S_{\alpha}(\gamma_0, \beta_0, \delta_0)$, with $\beta_0 = 1$, $\delta_0 = 0$ and $\gamma_0 = \gamma(\cos(\pi\alpha/2))^{1/\alpha}$.*

A.5 Background on Lévy Measures on $(0, \infty)$

The generalised gamma Lévy measure (Hougaard, 1986; Brix, 1999) has three parameters $\eta > 0$, $\alpha \in (-\infty, 1)$ and $\tau > 0$ if $\alpha \leq 0$ and $\tau \geq 0$ if $\alpha \in (0, 1)$. It is defined by

$$\rho_{\text{gg}}(dx; \eta, \alpha, \tau) = \eta \frac{1}{\Gamma(1-\alpha)} x^{-\alpha-1} e^{-\tau x} dx. \tag{58}$$

It is finite when $\alpha < 0$, and infinite otherwise. It admits as special cases the gamma measure when $\alpha = 0$ and the positive stable measure if $\alpha \in (0, 1)$ and $\tau = 0$. We denote the gamma measure and (scaled) positive stable measures as

$$\rho_{\text{gamma}}(dx; \eta, \tau) := \rho_{\text{gg}}(dx; \eta, 0, \tau) = \eta x^{-1} e^{-\tau x} dx, \tag{59}$$

$$\rho_{\text{stable}}(dx; \alpha, c) := \rho_{\text{gg}}(dx; \alpha c^{\alpha} \Gamma(1-\alpha), \alpha, 0) = \alpha c^{\alpha} x^{-\alpha-1} dx. \tag{60}$$

Here $\eta, \tau > 0$ for the gamma measure, and $\alpha \in (0, 1)$ and $c > 0$ for the stable Lévy measure. The c is called a scaling parameter. Note that $\text{ID}(0, \rho_{\text{gamma}}(\,\cdot\,; \eta, \tau)) = \text{Gamma}(\eta, \tau)$. Additionally, if $X \sim \text{ID}(0, \rho_{\text{stable}}(\,\cdot\,; \alpha, c))$, then X is a positive stable random variable with parameter $\alpha \in (0, 1)$, with Laplace transform $\mathbf{E}[e^{-tX}] = e^{-(\gamma t)^{\alpha}}$ for $\gamma := c \cdot \Gamma(1-\alpha)^{1/\alpha}$.

The stable beta Lévy measure with parameter $\eta > 0$, $\alpha \in (-\infty, 1)$, $\phi > -\alpha$ is defined as (Hjort, 1990; Thibaux and Jordan, 2007; Teh and Gorur, 2009)

$$\rho_{\text{sb}}(dx; \eta, \alpha, \phi) = \eta \frac{\Gamma(1+\phi)}{\Gamma(1-\alpha)\Gamma(\phi+\alpha)} x^{-\alpha-1} (1-x)^{\phi+\alpha-1} \mathbf{1}_{\{x \in (0,1)\}} dx. \tag{61}$$

It is infinite if $\alpha \geq 0$ and finite otherwise. If $\alpha = 0$, this is known as the beta Lévy measure.

The scaled stable beta measure has the additional scale parameter $c > 0$, and is defined as

$$\rho_{\text{ssb}}(dx; \eta, \alpha, \phi, c) = \eta \frac{\Gamma(1+\phi) c^{\alpha}}{\Gamma(1-\alpha)\Gamma(\phi+\alpha)} x^{-\alpha-1} (1-x/c)^{\phi+\alpha-1} \mathbf{1}_{\{x \in (0,c)\}} dx. \tag{62}$$

The following proposition derives some connections between the scaled stable beta measure and the generalised gamma measure, and some invariance property of the scaled stable measure. Similar expressions were obtained by Griffin and Leisen (2017) for constructing dependent completely random measures.

Proposition 26 (Gamma-function integral formulas) *Let* $\text{Gamma}(x; a, b)$ *denote the pdf of a Gamma random variable with shape parameter* a *and inverse scale parameter* b*. Let* $\kappa > 0$*,* $\eta > 0$*,* $\alpha \in (-\infty, 1)$*,* $c > 0$*,* $\phi > \max(0, -\alpha)$ *and* $b > 0$*. Then,*

$$\rho_{\text{gg}}\left(dx; \frac{\eta\kappa\phi}{(bc)^{\alpha}}, \alpha, bc\right) = dx \times \kappa \int_0^{\infty} \frac{1}{z} \text{Gamma}\left(\frac{x}{z}; \phi, b\right) \rho_{\text{ssb}}(dz; \eta, \alpha, \phi, 1/c), \tag{63}$$

$$\rho_{\text{stable}}\left(dx; \alpha, \frac{c}{b}\left(\frac{\kappa\Gamma(\phi+\alpha)}{\Gamma(\phi)}\right)^{1/\alpha}\right) = dx \times \kappa \int_0^{\infty} \frac{1}{z} \text{Gamma}\left(\frac{x}{z}; \phi, b\right) \rho_{\text{stable}}(dz; \alpha, c). \tag{64}$$

Proof Let

$$\begin{aligned}
\nu(x) &:= \kappa \int_0^{\infty} \frac{1}{z} \text{Gamma}\left(\frac{x}{z}; \phi, b\right) \rho_{\text{ssb}}(dz; \eta, \alpha, \phi, 1/c) \\
&= \eta\kappa \frac{c^{-\alpha} b^{\phi} \Gamma(1+\phi)}{\Gamma(\phi)\Gamma(\phi+\alpha)\Gamma(1-\alpha)} \int_0^{1/c} z^{-1} (x/z)^{\phi-1} e^{-xb/z} z^{-\alpha-1} (1-cz)^{\alpha+\phi-1} dz \\
&= \eta\kappa \frac{c^{-\alpha} b^{\phi} \phi}{\Gamma(\phi+\alpha)\Gamma(1-\alpha)} x^{\phi-1} \int_0^{1/c} z^{-\phi-\alpha-1} e^{-xb/z} (1-cz)^{\alpha+\phi-1} dz.
\end{aligned}$$

Using the change of variable $u = \frac{1}{z} - c$ so that $z = \frac{1}{u+c}$ and $dz = \frac{-du}{(u+c)^2}$, we obtain

$$\begin{aligned}
\nu(x) &= \eta\kappa \frac{c^{-\alpha} b^{\phi} \phi}{\Gamma(\phi+\alpha)\Gamma(1-\alpha)} x^{\phi-1} \int_0^{\infty} (u+c)^{\phi+\alpha-1} e^{-xb(u+c)} \left(\frac{u}{u+c}\right)^{\alpha+\phi-1} du \\
&= \eta\kappa \frac{c^{-\alpha} b^{\phi} \phi}{\Gamma(\phi+\alpha)\Gamma(1-\alpha)} x^{\phi-1} \int_0^{\infty} e^{-xb(u+c)} u^{\alpha+\phi-1} du \\
&= \eta\kappa \frac{c^{-\alpha} b^{\phi} \phi}{\Gamma(\phi+\alpha)\Gamma(1-\alpha)} x^{\phi-1} e^{-bcx} \frac{\Gamma(\alpha+\phi)}{(bx)^{\alpha+\phi}} \\
&= \eta\kappa \frac{c^{-\alpha} b^{-\alpha} \phi}{\Gamma(1-\alpha)} x^{-\alpha-1} e^{-bcx},
\end{aligned}$$

which is the density of the generalised gamma Lévy measure with parameters $(\eta\kappa(bc)^{-\alpha}\phi, \alpha, bc)$ at x.

Similarly,

$$\begin{aligned}
\nu_2(x) &:= \kappa \int_0^{\infty} \frac{1}{z} \operatorname{Gamma}\left(\frac{x}{z}; \phi, b\right) \rho_{\text{stable}}(dz; \alpha, c) \\
&= \kappa \frac{c^{\alpha} \alpha b^{\phi}}{\Gamma(\phi)} \int_0^{\infty} z^{-1} (x/z)^{\phi-1} e^{-xb/z} z^{-\alpha-1} dz \\
&= \kappa \frac{c^{\alpha} \alpha b^{\phi}}{\Gamma(\phi)} x^{\phi-1} \int_0^{\infty} e^{-xbu} u^{\phi+\alpha-1} du \\
&= \kappa \frac{c^{\alpha} \alpha \Gamma(\phi+\alpha)}{\Gamma(\phi) b^{\alpha}} x^{-\alpha-1} = \left(\frac{\kappa^{1/\alpha} c \Gamma(\phi+\alpha)^{1/\alpha}}{\Gamma(\phi)^{1/\alpha} b}\right)^{\alpha} \alpha x^{-\alpha-1}.
\end{aligned}$$

■

The following are corollaries of the above proposition, with $\kappa = \phi = b = \frac{1}{2}$, in combination with Corollary 37.

Corollary 27 *Let $X_1, X_2, \ldots,$ be iid standard normal random variables, and $(\xi_i)_i$ be the points of a Poisson point process with mean measure $\rho_{\text{ssb}}(dz; \eta, \alpha, 1/2, 1/c)$ for $\eta > 0$, $c > 0$ and $\alpha > -\frac{1}{2}$. Then,*

$$\sum_{i \geq 1} \xi_i \sim \text{ID}(0, \rho_{\text{ssb}}(\,\cdot\,; \eta, \alpha, 1/2, 1/c)),$$

$$\sum_{i \geq 1} \xi_i \max(0, X_i)^2 \sim \text{ID}\left(0, \rho_{\text{gg}}\left(\,\cdot\,; \eta 2^{\alpha-2} c^{-\alpha}, \alpha, c/2\right)\right).$$

In particular, if additionally $\alpha = 0$,

$$\sum_{i \geq 1} \xi_i \max(0, X_i)^2 \sim \operatorname{Gamma}\left(\frac{\eta}{4}, \frac{c}{2}\right).$$

Corollary 28 *Let $X_1, X_2, \ldots,$ be iid standard normal random variables, and $(\xi_i)_i$ be the points of a Poisson point process with mean measure $\rho_{\text{stable}}(dz; \alpha, c)$ for some $\alpha \in (0,1)$ and $c > 0$. Then,*

$$\sum_{i \geq 1} \xi_i \sim \text{ID}(0, \rho_{\text{stable}}(\,\cdot\,; \alpha, c)),$$

$$\sum_{i \geq 1} \xi_i \max(0, X_i)^2 \sim \text{ID}\left(0, \rho_{\text{stable}}\left(\,\cdot\,; \alpha, 2c\left(\frac{\Gamma(1/2+\alpha)}{2\sqrt{\pi}}\right)^{1/\alpha}\right)\right),$$

so that

$$\sum_{i\geq 1}\xi_i \max(0, X_i)^2 \overset{\mathrm{d}}{=} 2\left(\frac{\Gamma(1/2+\alpha)}{2\sqrt{\pi}}\right)^{1/\alpha}\sum_{i\geq 1}\xi_i.$$

Appendix B. Some Limit Theorems on Independent Triangular Arrays

Throughout the paper, we use a necessary and sufficient condition for sums of random variables to converge to an infinitely divisible random variable, and also a sufficient condition for such convergence when all the random variables involved have densities. These two conditions are summarised in the following theorem. All the proofs of the examples in Section 6 and Appendix E.2 rely on these conditions; details are given in Appendix E.3.

Theorem 29 (Necessary and sufficient conditions for convergence to $\mathrm{ID}(a,\rho)$) *Let $(X_{p,j})_{p\geq 1, j=1,\ldots,p}$ be a triangular array of nonnegative real random variables, where for each $p \geq 1$, the random variables $X_{p,1},\ldots,X_{p,p}$ are iid. Let $a \geq 0$ and ρ be a Lévy measure on $(0,\infty)$. Then, $\sum_{j=1}^{p} X_{p,j} \xrightarrow{\mathsf{d}} \mathrm{ID}(a,\rho)$ if and only if the following two conditions hold:*

(i) $p\Pr(X_{p,1} > x) \to \overline{\rho}(x)$ for all $x > 0$ such that $\rho(\{x\}) = 0$, and

(ii) $p\mathbf{E}[X_{p,1}\mathbf{1}_{\{X_{p,1}\leq h\}}] \to a + \int_0^h x\rho(dx)$ for any $h > 0$ with $\rho(\{h\}) = 0$.

If every $X_{p,1}$ is an absolutely continuous random variable with density f_p, and ρ is absolutely continuous with density ϱ and support S, then condition (i) is implied by the following three conditions:

(a) $pf_p(x) \to \varrho(x)$ for all $x > 0$,

(b) for any $x_0 > 0$, there exists C_{x_0} such that $\frac{pf_p(x)}{\varrho(x)} \leq C_{x_0}$ for all $x \in [x_0,\infty) \cap S$, and

(c) for any $x_0 > 0$, $\int_{[x_0,\infty)\setminus S} f_p(x)dx = o(1/p)$.

In Theorem 29, we have included the second part since in practice, for continuous random variables, conditions (a-c) will be easier to check than the condition (i).

Proof The first part of the theorem is a corollary of Theorem 15.28 in (Kallenberg, 2002). We focus on the second part for absolutely continuous random variables.

Let $(0,\infty] = (0,\infty)\cup\{\infty\}$ denote the set of positive reals with the addition of ∞, which is called the set of extended positive reals. Note that for any $a > 0$, $[a,\infty]$ is a compact set of $(0,\infty]$. Let $C_K^+((0,\infty])$ denote the set of continuous functions $f : (0,\infty] \to \mathbb{R}_+$ with compact support. Note that the functions are necessarily bounded as $f(\infty) \in \mathbb{R}_+$.

Let $(X_{p,j})$ be a triangular array of random variables such that for every p, $(X_{p,j})_{j=1,\ldots,p}$ is an iid sequence of random variables from μ_p. By (Kallenberg, 2002, Theorem 15.29), $p\Pr(X_{p,1} > x) \to \overline{\rho}(x)$ for all $x > 0$ such that $\rho(\{x\}) = 0$ is equivalent to

$$\eta_p := \sum_{j=1}^{p} \delta_{X_{p,j}} \xrightarrow{\mathsf{d}} \eta \text{ on } (0,\infty]$$

where η is a Poisson random measure with mean measure ρ. This is equivalent to showing that

$$\mathbf{E}[e^{-\eta_p(g)}] \to \mathbf{E}[e^{-\eta(g)}]$$

for all $g \in C_K^+((0,\infty])$. Pick $g \in C_K^+((0,\infty])$. Let $S \subseteq (0,\infty)$ be the support of ρ. We have

$$\begin{aligned}
\mathbf{E}[e^{-\eta_p(g)}] &= \mathbf{E}\left[e^{-\sum_{j=1}^{p} g(X_{p,j})}\right] \\
&= \mathbf{E}\left[e^{-g(X_{p,1})}\right]^p \\
&= \left(\int_0^\infty e^{-g(x)} f_p(x)dx\right)^p
\end{aligned}$$

$$= \left(1 - \int_0^\infty (1 - e^{-g(x)}) f_p(x) dx\right)^p$$
$$= \left(1 - \frac{1}{p}\int_0^\infty (1 - e^{-g(x)}) p f_p(x) dx\right)^p$$
$$= \left(1 - \frac{1}{p}\left[\int_S (1 - e^{-g(x)}) \frac{p f_p(x)}{\varrho(x)} \varrho(x) dx + \int_{\mathbb{R}_+ \setminus S} (1 - e^{-g(x)}) p f_p(x) dx\right]\right)^p.$$

Since g has compact support on $(0, \infty]$, there exists $x_0 > 0$ such that $g(x) = 0$ for $x < x_0$. Then, by assumption, $\frac{p f_p(x)}{\varrho(x)} \leq C_{x_0}$ for all $x \in S \cap [x_0, \infty)$ and $p \geq 1$. Also, again by assumption, $\frac{p f_p(x)}{\varrho(x)} \to 1$ as $p \to \infty$. What we have proved so far lets us use the dominated convergence theorem and derive the following convergence:

$$\int_{S \cap [x_0, \infty)} (1 - e^{-g(x)}) \frac{p f_p(x)}{\varrho(x)} \varrho(x) dx \to \int_{S \cap [x_0, \infty)} (1 - e^{-g(x)}) \varrho(x) dx.$$

Additionally, $\int_{\mathbb{R}_+ \setminus S} (1 - e^{-g(x)}) p f_p(x) dx \leq \int_{\mathbb{R}_+ \setminus S} p f_p(x) dx = o(1)$. Hence,

$$\int_S (1 - e^{-g(x)}) \frac{p f_p(x)}{\varrho(x)} \varrho(x) dx + \int_{\mathbb{R}_+ \setminus S} (1 - e^{-g(x)}) p f_p(x) dx$$
$$= \int_{S \cap [x_0, \infty)} (1 - e^{-g(x)}) \frac{p f_p(x)}{\varrho(x)} \varrho(x) dx + \int_{\mathbb{R}_+ \setminus S} (1 - e^{-g(x)}) p f_p(x) dx$$
$$\to \int_S (1 - e^{-g(x)}) \varrho(x) dx + 0 = \int_0^\infty (1 - e^{-g(x)}) \varrho(x) dx.$$

Recall that for any real sequence $(a_p)_{p \geq 1}$ converging to a, we have $(1 - \frac{a_p}{p})^p \to e^{-a}$. Thus,

$$\left(1 - \frac{1}{p}\int_S (1 - e^{-g(x)}) \frac{p f_p(x)}{\varrho(x)} \varrho(x) dx\right)^p \to e^{-\int_0^\infty (1 - e^{-g(x)}) \varrho(x) dx} = \mathbf{E}[e^{-\eta(g)}].$$

■

Proposition 30 (Extremes of triangular arrays and infinite divisibility) *Let*

$$(X_{p,j})_{p \geq 1, j = 1, \ldots, p}$$

be a triangular array of independent nonnegative real random variables such that for each p, $(X_{p,j})_{j=1,\ldots,p}$ are iid. Assume $\sum_{j=1}^p X_{p,j} \xrightarrow{\mathsf{d}} \mathrm{ID}(a, \rho)$ for some $a \geq 0$ and some Lévy measure ρ on $(0, \infty)$. Let

$$\overline{\rho}^{-1}(u) := \inf\{x > 0 : \overline{\rho}(x) < u\},$$

the inverse tail Lévy intensity of ρ. For each $p \geq 1$, let $X_{p,(1)} \geq X_{p,(2)} \geq \ldots \geq X_{p,(p)}$ denote the order statistics of $(X_{p,j})_j$. Then, the asymptotic behaviour of $X_{p,(k)}$, as $p \to \infty$, is solely characterised by ρ (not a) with $X_{p,(1)} \xrightarrow{\mathsf{pr}} 0$ if ρ is trivial, and if ρ is non-trivial,

$$\left(X_{p,(k)}\right)_{k \geq 1} \xrightarrow{\mathsf{d}} \left(\overline{\rho}^{-1}(G_k)\right)_{k \geq 1}$$

with $(G_k)_{k \geq 1}$ being ordered points of a standard rate one Poisson process on $(0, \infty)$ with $G_k \sim \mathrm{Gamma}(k, 1)$. In particular, for non-trivial ρ, $\overline{\rho}^{-1}(G_k)$ is a nonnegative random variable, non-degenerate at 0, with cumulative density function F_k defined by

$$F_k(x) = e^{-\overline{\rho}(x)} \sum_{i=0}^{k-1} \frac{\overline{\rho}(x)^i}{i!} \quad \text{for any } x > 0 \text{ with } \rho(\{x\}) = 0$$

and

$$F_k(0) = \begin{cases} e^{-\overline{\rho}(0)} \sum_{i=0}^{k-1} \frac{\overline{\rho}(0)^i}{i!} & \text{if } \rho \text{ is finite,} \\ 0 & \text{if } \rho \text{ is infinite.} \end{cases}$$

Proof From (Kallenberg, 2002, Theorem 15.29), $\sum_{j=1}^{p} X_{p,j} \xrightarrow{\mathrm{d}} \mathrm{ID}(a, \rho)$ implies that for any pre-compact Borel subsets $B_1, \ldots, B_k$ of the extended real half-line $(0, \infty] = (0, \infty) \cap \{\infty\}$,

$$\Big(\#\{j \mid X_{p,j} \in B_i\}\Big)_{i=1}^{k} \xrightarrow{\mathrm{d}} \Big(\eta(B_i)\Big)_{i=1}^{k}$$

where η is a Poisson random measure with mean measure ρ. In particular, $\big(X_{p,(j)}\big)_{j=1}^{k}$ converges in distribution to the joint distribution of the first k arrival times, going backwards in time from ∞, of a Poisson process with intensity ρ. This is because for

$$0 < x_k < y_k < x_{k-1} < \cdots < x_1 < y_1 < x_0 = \infty$$

such that $\rho(\{x_k, y_k, \ldots, x_1, y_1\}) = 0$,

$$\begin{aligned}
&\Pr\left(\bigcap_{j=1}^{k} \{X_{p,(j)} \in (x_j, y_j)\}\right) \\
&\quad \to \Pr\left(\{\eta(x_k, y_k) \geq 1\} \cap \bigcap_{j=1}^{k-1} \{\eta(x_j, y_j) = 1\} \cap \bigcap_{j=1}^{k} \{\eta(y_j, x_{j-1}) = 0\}\right) \\
&\quad = \big(1 - e^{-[\overline{\rho}(x_k) - \overline{\rho}(y_k)]}\big) \prod_{j=1}^{k-1} e^{-[\overline{\rho}(x_j) - \overline{\rho}(y_j)]} [\overline{\rho}(x_j) - \overline{\rho}(y_j)] \prod_{j=1}^{k} e^{-[\overline{\rho}(y_j) - \overline{\rho}(x_{j-1})]} \\
&\quad = \big(1 - e^{-[\overline{\rho}(x_k) - \overline{\rho}(y_k)]}\big) e^{-\overline{\rho}(y_k)} \prod_{j=1}^{k-1} [\overline{\rho}(x_j) - \overline{\rho}(y_j)] \\
&\quad = \big(e^{-\overline{\rho}(y_k)} - e^{-\overline{\rho}(x_k)}\big) \prod_{j=1}^{k-1} [\overline{\rho}(x_j) - \overline{\rho}(y_j)]
\end{aligned}$$

and the final expression on the right-hand side above can be seen to be the joint distribution of those k arrival times. It remains to calculate the limiting distribution of the marginal $X_{p,(k)}$.

For any $x > 0$ such that $\rho(\{x\}) = 0$,

$$\begin{aligned}
\Pr(X_{p,(k)} \leq x) &= \sum_{i=0}^{k-1} \Pr(\#\{j \mid X_{p,j} > x\} = i) \\
&\to \sum_{i=0}^{k-1} \Pr(\eta(x, \infty) = i) = \sum_{i=0}^{k-1} \frac{\overline{\rho}(x)^i e^{-\overline{\rho}(x)}}{i!}.
\end{aligned}$$

The value at 0 follows due to the right continuity of the cdf. Finally, using the identity, for any $\lambda > 0$,

$$\sum_{i=0}^{k-1} \frac{\lambda^i e^{-\lambda}}{i!} = \frac{\lambda^k}{\Gamma(k)} \int_1^{\infty} u^{k-1} e^{-u\lambda} du$$

we obtain

$$\sum_{i=0}^{k-1} \frac{\overline{\rho}(x)^i e^{-\overline{\rho}(x)}}{i!} = \frac{\overline{\rho}(x)^k}{\Gamma(k)} \int_1^{\infty} u^{k-1} e^{-u\overline{\rho}(x)} du = \Pr(G_k \geq \overline{\rho}(x)) = \Pr(x \geq \overline{\rho}^{-1}(G_k))$$

where the last equality follows from the definition of the inverse tail intensity $\overline{\rho}^{-1}$ and the absolute continuity of G_k. ■

Lemma 31 (Lévy continuity theorem for triangular arrays) *Let $(X_{p,i})_{p\geq 1, i=1,\ldots,p}$ be a triangular array of nonnegative scalar random variables such that for every p, $(X_{p,i})_{i=1,\ldots,p}$ is an iid sequence of random variables from a probability distribution μ_p on $[0,\infty)$. Let $a \geq 0$ and ρ be a Lévy measure on $(0,\infty)$. Then,*

$$\sum_{i=1}^{p} X_{p,i} \xrightarrow{\mathtt{d}} \mathrm{ID}(a,\rho) \text{ as } p \to \infty$$

if and only if, for any $t \geq 0$,

$$\int_0^\infty (1-e^{-tx}) p\mu_p(dx) \to at + \int_0^\infty (1-e^{-wt})\rho(dw)$$

pointwise as $p \to \infty$.

Proof Recall that if $S \sim \mathrm{ID}(a,\rho)$, then $\mathbf{E}[e^{-tS}] = e^{-at-\psi(t)}$ where $\psi(t) = \int_0^\infty (1-e^{-wt})\rho(dw)$. By the Lévy continuity theorem for Laplace transforms of nonnegative random variables, $\sum_{i=1}^p X_{p,i} \xrightarrow{\mathtt{d}} \mathrm{ID}(a,\rho)$ if and only if, for any $t \geq 0$,

$$\begin{aligned}
\mathbf{E}\left[e^{-t\sum_{i=1}^p X_{p,i}}\right] &= \mathbf{E}\left[e^{-tX_{p,1}}\right]^p \\
&= \left(\int_0^\infty e^{-tx}\mu_p(dx)\right)^p \\
&= \left(1 - \frac{1}{p}\int_0^\infty (1-e^{-tx}) p\mu_p(dx)\right)^p \to e^{-at-\psi(t)},
\end{aligned}$$

which holds if and only if $\int_0^\infty (1-e^{-tx})p\mu_p(dx) \to at + \psi(t)$. ■

Proposition 32 (Compressibility of triangular arrays) *Let*

$$(X_{p,j})_{p\geq 1, j=1,\ldots,p}$$

be a triangular array of nonnegative real random variables such that for each p, $(X_{p,j})_{j=1,\ldots,p}$ are iid. Assume $\sum_{j=1}^p X_{p,j} \xrightarrow{\mathtt{d}} \mathrm{ID}(a,\rho)$ for some $a \geq 0$ and some Lévy measure ρ on $(0,\infty)$. For each $p \geq 1$, let $X_{p,(1)} \geq X_{p,(2)} \geq \ldots \geq X_{p,(p)}$ be the ordered values. Then, for every $\kappa \in (0,1)$,

$$X_{p,(\lfloor \kappa p \rfloor)} \xrightarrow{\mathtt{pr}} 0 \quad \text{as } p \to \infty. \tag{65}$$

Moreover, if $a = 0$, then for each $\kappa \in (0,1)$,

$$\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} \xrightarrow{\mathtt{pr}} 0 \quad \text{as } p \to \infty. \tag{66}$$

Proof We first prove Equation (65). Suppose to the contrary that $X_{p,(\lfloor \kappa p \rfloor)}$ does not converge to 0. Then, there exist $\epsilon > 0$ and $\eta \in (0,1)$ such that ϵ is a continuity point of ρ (i.e. $\rho(\{\epsilon\}) = 0$) and for any p_0, there exists $p > p_0$ such that

$$\Pr(X_{p,(\lfloor \kappa p \rfloor)} > \epsilon) > \eta.$$

Hence,

$$p \Pr(X_{p,1} > \epsilon) = \mathbf{E}\left[\sum_j \mathbf{1}_{\{X_{p,j} > \epsilon\}}\right] > \eta \lfloor \kappa p \rfloor. \tag{67}$$

But $p \Pr(X_{p,1} > \epsilon) \to \overline{\rho}(\epsilon) < \infty$, which gives a contradiction.

Now, suppose $a = 0$ and choose $\kappa \in (0, 1)$. We prove Equation (66) by showing that, for any $\epsilon > 0$, as $p \to \infty$,

$$\Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon\right) \to 0.$$

Since $a = 0$, it follows that for any continuity point $h > 0$ of ρ,

$$p\mathbf{E}[X_{p,1}\mathbf{1}_{\{X_{p,1} \leq h\}}] \to \int_0^h x\rho(dx)$$

as $p \to \infty$. (See the first part of Theorem 29, which is a corollary of Theorem 15.28 in (Kallenberg, 2002).) Thus, for any $\eta > 0$, there exist a continuity point $h_0 > 0$ of ρ and $p_0 \in \mathbb{N}$ such that for all $p \geq p_0$,

$$p\mathbf{E}[X_{p,1}\mathbf{1}_{\{X_{p,1} \leq h_0\}}] < \eta. \tag{68}$$

Consider $\epsilon > 0$ and $\gamma > 0$. Define $\eta := (\gamma\epsilon)/2$. Let h_0 and p_0 be such that Equation (68) holds for η. Note that

$$\begin{aligned}\Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon\right) &= \Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon \text{ and } X_{p,(\lfloor \kappa p \rfloor)} \leq h_0\right) \\ &\quad + \Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon \text{ and } X_{p,(\lfloor \kappa p \rfloor)} > h_0\right).\end{aligned} \tag{69}$$

We will prove that for all sufficiently large p, each summand on the right-hand side above is bounded by $\gamma/2$. The next derivation uses Markov's inequality and bounds the first summand in Equation (69) for all $p \geq p_0$:

$$\begin{aligned}\Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon \text{ and } X_{p,(\lfloor \kappa p \rfloor)} \leq h_0\right) &\leq \Pr\left(\sum_j X_{p,j}\mathbf{1}_{\{X_{p,j} \leq h_0\}} > \epsilon\right) \\ &\leq \frac{p\mathbf{E}[X_{p,1}\mathbf{1}_{\{X_{p,1} \leq h_0\}}]}{\epsilon} < \frac{\eta}{\epsilon} = \frac{\gamma\epsilon}{2\epsilon} = \frac{\gamma}{2}.\end{aligned}$$

The bound for the second summand in Equation (69) follows from Equation (65). There is $p_1 \in \mathbb{N}$ such that for all $p \geq p_1$,

$$\Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon \text{ and } X_{p,(\lfloor \kappa p \rfloor)} > h_0\right) \leq \Pr\left(X_{p,(\lfloor \kappa p \rfloor)} > h_0\right) < \frac{\gamma}{2}.$$

Bringing these two bounds together, we can conclude that for all $p \geq \max(p_0, p_1)$,

$$\Pr\left(\sum_{j=1}^{p} \mathbf{1}_{\{X_{p,j} \leq X_{p,(\lfloor \kappa p \rfloor)}\}} X_{p,j} > \epsilon\right) < \gamma,$$

as desired.

■

For the rest of this section, we consider the space $\mathbb{K}_d$ of positive semi-definite $n \times n$ matrices. To state the results, we need to recall a few definitions regarding cones.

Let $\mathbb{H}$ be a finite-dimensional Hilbert space, and $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ be its norm and inner product. A nonempty convex subset $\mathbb{S}$ of $\mathbb{H}$ is a *cone* if $\lambda \geq 0$ and $x \in \mathbb{S}$ implies $\lambda x \in \mathbb{S}$. A cone is *proper* if $x = 0$ whenever x and $-x$ are in $\mathbb{S}$. The *dual cone* $\mathbb{S}'$ of $\mathbb{S}$ is defined as $\mathbb{S}' = \{y \in \mathbb{H} : \langle y, x \rangle \geq 0 \text{ for all } x \in \mathbb{S}\}$. Examples of proper cones include $[0, \infty)$, $[0, \infty)^d$, and the set $\mathbb{K}_d$ of positive semi-definite d-by-d matrices with real entries.

Let $\mathbb{S}$ be a proper convex cone of $\mathbb{H}$. Denote by $\mathbb{S} \cup \{\Delta\}$ the one-point compactification of the cone. Let $C(\mathbb{S} \cup \{\Delta\})$ be the set of continuous functions $f : \mathbb{S} \cup \{\Delta\} \to \mathbb{R}$. For a function $f \in C(\mathbb{S} \cup \{\Delta\})$, define $\|f\| = \sup_{x \in \mathbb{S} \cup \{\Delta\}} |f(x)|$. Similarly, for a function $g : \mathbb{S} \to \mathbb{R}$, define $\|g\| = \sup_{x \in \mathbb{S}} |g(x)|$.

The next proposition expresses Lévy's continuity theorem for $\mathbb{K}_d$, the cone of positive semi-definite d-by-d matrices with real entries. In this case, the ambient space $\mathbb{H}$ is that of symmetric d-by-d matrices with real entries, and its inner product and norm are inherited from the Euclidean space $\mathbb{R}^{d \times d}$. We point out that with respect to this ambient space $\mathbb{H}$, the cone $\mathbb{K}_d$ is self-dual: $(\mathbb{K}_d)' = \mathbb{K}_d$.

Lemma 33 (Lévy continuity theorem on $\mathbb{K}_d$) *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures on $\mathbb{K}_d$. If $\widetilde{\mu}_n(\theta) = \int e^{-\langle x, \theta \rangle} \mu_n(dx)$ converges pointwise to $\widetilde{\mu}(\theta) = \int e^{-\langle x, \theta \rangle} \mu(dx)$ for every $\theta \in \mathbb{K}_d$, then $\mu_n \xrightarrow{\mathrm{d}} \mu$.*

Proof The proof follows that of Theorem 5.3 in (Kallenberg, 2002).

Assume that $\widetilde{\mu}_n(\theta)$ converges to $\widetilde{\mu}(\theta)$ for every $\theta \in \mathbb{K}_d$. We have, using $e^{-t} \leq \frac{1}{2}$ for all $t \geq 1$,

$$1 - \widetilde{\mu}_n(\theta) = \int_{\mathbb{K}_d} (1 - e^{-\langle x, \theta \rangle}) \, \mu_n(dx) \geq \frac{1}{2} \mu_n(\{x : \langle x, \theta \rangle \geq 1\}).$$

Hence, for all $\theta \in \mathbb{K}_d$ and $r > 0$, we have $\mu_n(\{x : \langle x, \theta \rangle \geq r\}) \leq 2(1 - \widetilde{\mu}_n(\theta / r))$, which then implies

$$\limsup_n \mu_n(\{x : \langle x, \theta \rangle \geq r\}) \leq \lim_{n \to \infty} 2(1 - \widetilde{\mu}_n(\theta / r)) = 2(1 - \widetilde{\mu}(\theta / r)).$$

Taking $r \to \infty$ and using the continuity of $\widetilde{\mu}$ at 0, we obtain that

$$\lim_{r \to \infty} \limsup_n \mu_n(\{x : \langle x, \theta \rangle \geq r\}) = 0.$$

From this, straightforward calculations show that

$$\lim_{r \to \infty} \limsup_n \mu_n(\{x : \|x\| \geq r\}) = 0,$$

that is, the sequence $(\mu_n)_n$ is tight. As a result, for any $\varepsilon > 0$, we may choose large r that $\mu_n(\{x : \|x\| \geq r\}) \leq \varepsilon$ for all n and $\mu(\{x : \|x\| \geq r\}) \leq \varepsilon$.

Now fix a bounded continuous function $f : \mathbb{K}_d \to \mathbb{R}$, with $\|f\| \leq m < \infty$. Pick an arbitrary $\varepsilon > 0$. Let $r > 0$ be the real such that $\mu_n(\{x : \|x\| \geq r\}) \leq \varepsilon$ for all n and $\mu(\{x : \|x\| \geq r\}) \leq \varepsilon$. Define f_r to be the restriction of f to the ball $\{\|x\| \leq r\}$, and extend f_r to a continuous function $\widetilde{f}$ on $\mathbb{K}_d \cup \{\Delta\}$ with $\left\|\widetilde{f}\right\| \leq m$. For instance, one can set

$$\widetilde{f}(x) = f\left(\frac{rx}{\|x\|}\right) \times \max(r + 1 - \|x\|, 0) \text{ for } \|x\| \geq r.$$

By Lemma 34, there exists some function $g(x) = \sum_{j=1}^{p} \lambda_j e^{-\langle x,\theta_j\rangle}$ with $\theta_j \in \mathbb{K}_d$ and $\lambda_j \in \mathbb{R}$ such that $\left\|\widetilde{f} - g\right\| \leq \varepsilon$. We obtain

$$\begin{aligned}|\mu_n f - \mu_n g| &\leq \left|\mu_n f - \mu_n \widetilde{f}\right| + \left|\mu_n \widetilde{f} - \mu_n g\right| \\ &\leq \mu_n(\{x : \|x\| \geq r\}) \left\|f - \widetilde{f}\right\| + \left\|\widetilde{f} - g\right\| \\ &\leq (2m+1)\varepsilon,\end{aligned}$$

and similarly for μ. Thus,

$$\begin{aligned}|\mu_n f - \mu f| &\leq |\mu_n f - \mu_n g| + |\mu_n g - \mu g| + |\mu f - \mu g| \\ &\leq |\mu_n g - \mu g| + 2(2m+1)\varepsilon.\end{aligned}$$

By the pointwise convergence of $\widetilde{\mu}_n$,

$$\begin{aligned}|\mu_n g - \mu g| &\leq \sum_{j=1}^{p} |\lambda_j| |\widetilde{\mu}_n(\theta_j) - \widetilde{\mu}(\theta_j)| \\ &\to 0\end{aligned}$$

as $n \to \infty$. Letting $n \to \infty$ and then $\varepsilon \to 0$, we obtain $\mu_n f \to \mu f$. As f was chosen arbitrarily, this proves $\mu_n \xrightarrow{\mathtt{d}} \mu$. ■

We state the following version of Lemma 5.4 in (Kallenberg, 2002) for the cone of d-by-d positive semi-definite matrices.

Lemma 34 (Stone-Weierstrass theorem on $\mathbb{K}_d$) *Every continuous function $g : \mathbb{K}_d \cup \{\Delta\} \to \mathbb{R}$ can be approximated uniformly by linear combinations of the functions $x \longmapsto e^{-\langle x,\theta\rangle}$ for $\theta \in \mathbb{K}_d$.*

We omit the proof of Lemma 34 as it directly follows from the general Stone-Weierstrass theorem.

Lemma 35 *Let (X_{pi}) be a triangular array of nonnegative scalar random variables such that for every p, $(X_{pi})_{i=1,\ldots,p}$ is an iid sequence of random variables from μ_p. Suppose also that (Y_i) is an iid sequence in the cone $\mathbb{K}_n$ of positive semi-definite n-by-n matrices with $Y_i \sim F$ and $\mathbf{E}[\|Y_1\|] < \infty$. If $\sum_{i=1}^{p} X_{pi}$ converges in distribution to $\mathrm{ID}(a,\rho)$ as $p \to \infty$, then as $p \to \infty$,*

$$\sum_{i=1}^{p} X_{pi} Y_i \xrightarrow{\mathtt{d}} \mathrm{ID}(A,\nu) \tag{70}$$

where $A \in \mathbb{K}_n$ and the Lévy measure ν on $\mathbb{K}_n \setminus \{0\}$ are defined by

$$A := a\mathbf{E}[Y_1], \qquad \nu(dz) := \int_0^\infty F\left(\frac{dz}{x}\right)\rho(dx).$$

Remark 36 *The measure ν defined above is indeed a Lévy measure. Obviously, $\nu(\{0\}) = 0$. To see that $\int \min(\|z\|,1)\,\nu(dz) < \infty$, note that*

$$\int \min(\|z\|,1)\,\nu(dz) = \int_{\mathbb{K}_n}\int \min(x\|y\|,1)\mathbf{1}_{\{xy\neq 0\}}\,\rho(dx)\,F(dy).$$

Here, note that, for nonnegative a and b, $\min(1,ab) \leq \max(1,a)\min(1,b)$. It follows that

$$\begin{aligned}\int_{\mathbb{K}_n}\int \min(x\|y\|,1)\mathbf{1}_{\{xy\neq 0\}}\,\rho(dx)\,F(dy) &\leq \int_{\mathbb{K}_n}\int \max(1,\|y\|)\min(1,x)\mathbf{1}_{\{xy\neq 0\}}\,\rho(dx)\,F(dy) \\ &\leq \int_{\mathbb{K}_n}(1+\|y\|)\,F(dy)\int \min(1,x)\,\rho(dx) < \infty.\end{aligned}$$

Proof Let $\theta \in \mathbb{K}_n' = \mathbb{K}_n$. We calculate the following Laplace transform at θ:

$$\begin{aligned}\mathbf{E}\left[\exp\left(-\left\langle \theta, \sum_{i=1}^{p} X_{pi} Y_i \right\rangle\right)\right] &= \left(\mathbf{E}\left[\exp\left(-\langle \theta, X_{p1} Y_1\rangle\right)\right]\right)^p \\ &= \left(\int_{\mathbb{K}_n} \int e^{-\langle \theta, y\rangle x}\, \mu_p(dx)\, F(dy)\right)^p \\ &= \left(1 - \frac{1}{p}\int_{\mathbb{K}_n} \int \left(1 - e^{-\langle \theta, y\rangle x}\right)\, p\mu_p(dx)\, F(dy)\right)^p.\end{aligned}$$

Now, for a fixed $\theta, y \in \mathbb{K}_n$, Lemma 31 implies that

$$\int \left(1 - e^{-\langle \theta, y\rangle x}\right)\, p\mu_p(dx) \to a\langle \theta, y\rangle + \int \left(1 - e^{-\langle \theta, y\rangle x}\right)\, \rho(dx).$$

Note that, as a function of y, the above integral on the left-hand side is dominated by an integrable function. Specifically, we get, for large enough p's,

$$\begin{aligned}\int \left(1 - e^{-\langle \theta, y\rangle x}\right)\, p\mu_p(dx) &\le \int \min(1, \langle \theta, y\rangle x)\, p\mu_p(dx) \\ &\le \max(1, \langle \theta, y\rangle) \int \min(1, x)\, p\mu_p(dx) \\ &\le (1 + \|\theta\|\|y\|)\left(a + \int \min(1, x)\, \rho(dx) + \epsilon\right),\end{aligned}$$

where we used the fact that $\min(1, ab) \le \max(1, a)\min(1, b)$ for nonnegative a and b. As $\mathbf{E}[\|Y_1\|] < \infty$, the last expression is integrable with respect to $F(dy)$. By dominated convergence, it follows that

$$\begin{aligned}&\mathbf{E}\left[\exp\left(-\left\langle \theta, \sum_{i=1}^{p} X_{pi} Y_i \right\rangle\right)\right] \\ &\to \exp\left(-\left(\langle \theta, a\mathbf{E}[Y_1]\rangle + \int_{K_n} \int \left(1 - e^{-\langle \theta, xy\rangle}\right)\, \rho(dx)\, F(dy)\right)\right) \\ &= \exp\left(-\left(\langle \theta, A\rangle + \int \left(1 - e^{-\langle \theta, z\rangle}\right)\, \nu(dz)\right)\right),\end{aligned}$$

which is the Laplace transform of $\mathrm{ID}(A, \nu)$. By Lemma 33, (see (Davydov et al., 2008, Theorem 5.4)), the Laplace transform uniquely determines the distribution, so the proof is completed. ■

In the special case when $n = 1$, we have the following.

Corollary 37 *Let (X_{pi}) be a triangular array of nonnegative scalar random variables such that for every p, $(X_{pi})_{i=1,\ldots,p}$ is an iid sequence of random variables from μ_p. Let $Y_1, Y_2, \ldots$ be a sequence of iid nonnegative random variables $Y_i \in [0, \infty)$ with $Y_i \sim F$ and $\mathbf{E}[Y_1] < \infty$. Let $a \ge 0$ and ρ be a Lévy measure on $(0, \infty)$. Then,*

$$\sum_{i=1}^{p} X_{pi} \xrightarrow{\mathtt{d}} \mathrm{ID}(a, \rho) \tag{71}$$

implies

$$\sum_{i=1}^{p} X_{pi} Y_i \xrightarrow{\mathtt{d}} \mathrm{ID}(c, \nu) \tag{72}$$

where $c \geq 0$ and the Lévy measure ν on $(0, \infty)$ are defined by

$$c := a\mathbf{E}[Y_1], \qquad \nu(dz) := \int_0^\infty F\left(\frac{dz}{x}\right) \rho(dx).$$

When F has a density f with respect to the Lebesgue measure, the Levy measure ν can be expressed as

$$\nu(dz) := \int_0^\infty \rho(dz/x) f(x) dx.$$

In this paper, we often used this equivalent form of ν.

Appendix C. Proofs of the Main Theorems and Propositions

C.1 Proof of Theorem 5

Assume $a^{(l)} = 0$. Then, Equation (15) follows directly from Proposition 32 in the Appendix and Slutsky's theorem.

Note that $T_j^{(l+1)} = \lambda_{p_l,j}^{(l)} Y_j$, where the random variables $Y_j := \sum_{k=1}^{p_{l+1}} (V_{j,k}^{(l+1)})^2$ are iid, and do not depend on p_l. If $a^{(l)} = 0$, then Corollary 37 implies that $\sum_{j=1}^{p_l} T_j^{(l+1)} \to \mathrm{ID}(0, \nu)$ for some Lévy measure ν. Equation (14) then follows similarly from Proposition 32 and Slutsky's theorem.

C.2 Proof of Proposition 6

We assume here more generally that $\overline{\rho}^{(l)}(x) \overset{x\to\infty}{\sim} x^{-\tau} L(x)$, where L is a slowly varying function. Using Proposition 4 and Proposition 30 in the Appendix with the fact that $1 - \sum_{i=0}^{k-1} \frac{x^i e^{-x}}{i!} \overset{x\to 0}{\sim} \frac{x^k}{k!}$, we obtain the first equivalence relations in Equations (16) and (17). We have

$$\begin{aligned}
\overline{\nu}^{(l)}(x) &= \int_0^\infty \overline{\rho}^{(l)}(x/z)\,\mathrm{Gamma}\left(z; \frac{1}{2}, \frac{1}{2}\right) dz \\
&= x \int_0^\infty \overline{\rho}^{(l)}(1/u)\,\mathrm{Gamma}\left(ux; \frac{1}{2}, \frac{1}{2}\right) du \\
&= \frac{1}{\sqrt{2\pi}} x^{1/2} \int_0^\infty \overline{\rho}^{(l)}(1/u) u^{-1/2} e^{-ux/2} du.
\end{aligned}$$

We have $\overline{\rho}^{(l)}(1/u)u^{-1/2} \overset{u\to 0}{\sim} u^{\tau-1/2} L(1/u)$. Using a Tauberian theorem (Feller, 1971, Chapter 13, Theorem 2), we obtain

$$\begin{aligned}
\overline{\nu}^{(l)}(x) &\overset{x\to\infty}{\sim} \frac{1}{\sqrt{2\pi}} x^{1/2} \times \Gamma(\tau + 1/2)(x/2)^{-\tau-1/2} L(x/2) \\
&\overset{x\to\infty}{\sim} \frac{2^\tau \Gamma(\tau + 1/2)}{\sqrt{\pi}} x^{-\tau} L(x).
\end{aligned}$$

The second equivalence relations in Equations (16) and (17) now follow directly.

C.3 Proofs of Propositions 9 and 10

Proof of Proposition 9. First consider Equation (35). We have $S^{(l)} \sim \mathrm{ID}(\frac{a^{(l)}}{2}, \frac{\nu^{(l)}}{2})$ as mentioned in Equation (25). Also, we have shown in the proof of Proposition 6 that Equation (34) implies

$$\overline{\nu}^{(l)}(x) \overset{x\to\infty}{\sim} \frac{2^{\tau^{(l)}} \Gamma(\tau^{(l)} + 1/2)}{\sqrt{\pi}} c^{(l)} x^{-\tau^{(l)}}.$$

Equation (35) then follows from Proposition 22. For $l = 2$, we have

$$\Sigma^{(2)}(\mathbf{x}) = \sigma_b^2 + \sigma_v^2 S^{(1)} \Sigma^{(1)}(\mathbf{x}) \tag{73}$$

where $\Sigma^{(1)}(\mathbf{x}) = \sigma_b^2 + \sigma_v^2 \frac{\|\mathbf{x}\|^2}{d_{\text{in}}}$ is fixed. Thus,

$$\Pr\left(\Sigma^{(2)}(\mathbf{x}) > u\right) \overset{u\to\infty}{\sim} \Pr(S^{(1)} > u) \cdot (\sigma_v^2 \Sigma^{(1)}(\mathbf{x}))^{\tau^{(1)}} \tag{74}$$

and the random variable $\Sigma^{(2)}(\mathbf{x})$ has a power-law tail with exponent $\beta^{(1)} = \tau^{(1)}$. We can now proceed by induction. If $\Sigma^{(l)}(\mathbf{x})$ has a regularly varying tail with exponent $\beta^{(l-1)}$ and $S^{(l)}$ has a

regularly varying tail with exponent $\tau^{(l)}$, then $\Sigma^{(l+1)}(\mathbf{x}) = \sigma_b^2 + \sigma_v^2 S^{(l)}\Sigma^{(l)}(\mathbf{x})$ also has regularly varying tail with exponent $\beta^{(l)} = \min(\beta^{(l-1)}, \tau^{(l)})$ by Lemma 19. Finally, we have, using Lemma 21,

$$\begin{aligned}\Pr\left((\zeta_k^{(l)})^2 > u\right) &\overset{u\to\infty}{\sim} \Pr\left(\Sigma^{(l)}(\mathbf{x}) > u\right)\mathbf{E}\left[(\varepsilon_k^{(l)})^{2\beta^{(l-1)}}\right] \\ &\overset{u\to\infty}{\sim} \Pr\left(\Sigma^{(l)}(\mathbf{x}) > u\right) \times 2^{\beta^{(l-1)}}\frac{\Gamma(\beta^{(l-1)}+1/2)}{\Gamma(1/2)}.\end{aligned}$$

Proof of Proposition 10. If $\sigma_b = 0$, then

$$\Sigma^{(l+1)}(\mathbf{x}) = \frac{\|\mathbf{x}\|^2}{d_{\text{in}}}\sigma_v^{2(l+1)}\prod_{j=1}^{l} S^{(j)}$$

where $S^{(1)}, \ldots, S^{(L)}$ are iid and

$$\Pr\left(S^{(j)} > u\right) \overset{u\to\infty}{\sim} \widetilde{c}u^{-\tau}.$$

Here

$$\widetilde{c} = \frac{2^{\tau-1}\Gamma(\tau+1/2)}{\sqrt{\pi}}c.$$

It follows from Lemma 20 that

$$\Pr\left(\prod_{j=1}^{l} S^{(j)} > u\right) \overset{u\to\infty}{\sim} \frac{\tau^{l-1}(\widetilde{c})^l}{(l-1)!}u^{-\tau}\log^{l-1} u.$$

Therefore, for $l \geq 1$,

$$\Pr\left(\Sigma^{(l+1)}(\mathbf{x}) > u\right) \overset{u\to\infty}{\sim} \left(\frac{\|\mathbf{x}\|^2}{d_{\text{in}}}\sigma_v^{2(l+1)}\right)^{\tau}\frac{\tau^{l-1}(\widetilde{c})^l}{(l-1)!}u^{-\tau}\log^{l-1} u.$$

Finally, we have, using Lemma 21,

$$\Pr\left((\zeta_k^{(l+1)})^2 > u\right) \overset{u\to\infty}{\sim} \Pr\left(\Sigma^{(l+1)}(\mathbf{x}) > u\right) \times 2^{\tau}\frac{\Gamma(\tau+1/2)}{\Gamma(1/2)}.$$

C.4 Proofs of Proposition 11 and Corollaries 13 and 15

Recall the assumption (UI): For all layers $l = 1, \ldots, L$,

$$\int_0^\infty u\rho^{(l)}(du) = M_1^{(l)} < \infty, \quad N_{p_l}^{(l)} < \infty \text{ for all } p_l, \quad \text{and} \quad N_{p_l}^{(l)} \to a^{(l)} + M_1^{(l)} \text{ as } p_l \to \infty.$$

As mentioned earlier, this is equivalent to the uniform integrability of the family

$$\left\{\sum_{j=1}^{p_l}\lambda_{p_l,j}^{(l)}\right\}_{p_l}. \tag{75}$$

This can be seen by Durrett (2019, Theorem 4.6.3) and Skorokhod representation.

We start by introducing a lemma that results from this assumption.

Lemma 38 *Suppose (UI) and (A1) hold. For all $l = 1, \ldots, L+1$, we have*

$$\sup_{\mathbf{p}} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p})\right)^2\right] < \infty.$$

Proof Recall $C_\phi = \mathbf{E}[\phi(\epsilon)^2]$ for $\epsilon \sim \mathcal{N}(0,1)$. Note that

$$\begin{aligned}\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] &= \sigma_v^2 \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \phi^2(Z_j^{(l)}(\mathbf{x};\mathbf{p}))\right] + \sigma_b^2 \\ &= \sigma_v^2 \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)}\right] \mathbf{E}\left[\phi^2(Z_1^{(l)}(\mathbf{x};\mathbf{p}))\right] + \sigma_b^2 \\ &= \sigma_v^2 C_\phi N_{p_l}^{(l)} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p})\right)^2\right] + \sigma_b^2.\end{aligned}$$

Apply this recurrence repeatedly and note that $\mathbf{E}[(Z_1^{(1)}(\mathbf{x}))^2] = \Sigma^{(1)}(\mathbf{x}) = \sigma_v^2\|\mathbf{x}\|^2/d_{\text{in}} + \sigma_b^2$ is a constant. Then, we get

$$\begin{aligned}&\mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p})\right)^2\right] \\ &= \sigma_v^2 \frac{\|\mathbf{x}\|^2}{d_{\text{in}}} \left(\prod_{l'=1}^{l-1} \sigma_v^2 C_\phi N_{p_{l'}}^{(l')}\right) + \sigma_b^2 \left(\prod_{l'=1}^{l-1} \sigma_v^2 C_\phi N_{p_{l'}}^{(l')} + \cdots + \sigma_v^2 C_\phi N_{p_{l-1}}^{(l-1)} + 1\right).\end{aligned}$$

Since $N_{p_l}^{(l)} \to a^{(l)} + M_1^{(l)}$ as $p_l \to \infty$ for each $l = 1, \ldots, L$, we have

$$\begin{aligned}&\sup_{\mathbf{p}} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p})\right)^2\right] \\ &\le 2\left[\sigma_v^2 \frac{\|\mathbf{x}\|^2}{d_{\text{in}}} \left(\prod_{l'=1}^{l-1} \sigma_v^2 C_\phi (a^{(l')} + M_1^{(l')})\right)\right. \\ &\quad \left. + \sigma_b^2 \left(\prod_{l'=1}^{l-1} \sigma_v^2 C_\phi (a^{(l')} + M_1^{(l')}) + \cdots + \sigma_v^2 C_\phi (a^{(l-1)} + M_1^{(l-1)}) + 1\right)\right] \\ &< \infty.\end{aligned}$$

where the supremum is taken for all $\mathbf{p}$'s with large enough $\min \mathbf{p}$. ■

Proof of Proposition 11. Note that, for all p_l,

$$\begin{aligned}&\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \\ &= \sigma_v^2 \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} > \lambda_{p_l}^{*(l)}\}} \left(\phi(Z_j^{(l)}(\mathbf{x};\mathbf{p}) - \phi(Z_j^{*(l)}(\mathbf{x};\mathbf{p}))\right)^2\right] \\ &\quad + \sigma_v^2 \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \le \lambda_{p_l}^{*(l)}\}} \phi^2(Z_j^{(l)}(\mathbf{x};\mathbf{p}))\right] \\ &\le \sigma_v^2 C_{\text{Lip}}^2 N_{p_l}^{(l)} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l)}(\mathbf{x};\mathbf{p})\right)^2\right] \\ &\quad + \sigma_v^2 C_{\text{Lip}}^2 \mathbf{E}\left[(Z_1^{(l)}(\mathbf{x};\mathbf{p}))^2\right] \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \le \lambda_{p_l}^{*(l)}\}}\right].\end{aligned}$$

If $U^{(l)} := \sup_{\mathbf{p}} \mathbf{E}\left[(Z_1^{(l)}(\mathbf{x};\mathbf{p}))^2\right] < \infty$ for all $l = 1, \ldots, L$, we get the following recurrence relation:

$$\begin{aligned}
&\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \\
&\leq \sigma_v^2 C_{\mathrm{Lip}}^2 N_{p_l}^{(l)} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l)}(\mathbf{x};\mathbf{p})\right)^2\right] + \sigma_v^2 C_{\mathrm{Lip}}^2 U^{(l)} \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l}^{*(l)}\}}\right] \\
&=: \sigma_v^2 C_{\mathrm{Lip}}^2 N_{p_l}^{(l)} \mathbf{E}\left[\left(Z_1^{(l)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l)}(\mathbf{x};\mathbf{p})\right)^2\right] + \sigma_v^2 C_{\mathrm{Lip}}^2 U^{(l)} A_{p_l}^{(l)}
\end{aligned}$$

Noting that $Z_1^{(1)}(\mathbf{x};\mathbf{p}) = Z_1^{*(1)}(\mathbf{x};\mathbf{p})$, it inductively follows that

$$\begin{aligned}
&\mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \\
&\leq \sigma_v^2 C_{\mathrm{Lip}}^2 U^{(l)} A_{p_l}^{(l)} + (\sigma_v^2 C_{\mathrm{Lip}}^2)^2 N_{p_l}^{(l)} U^{(l-1)} A_{p_{l-1}}^{(l-1)} + \cdots + (\sigma_v^2 C_{\mathrm{Lip}}^2)^l N_{p_l}^{(l)} \cdots N_{p_2}^{(2)} U^{(1)} A_{p_1}^{(1)}.
\end{aligned} \tag{76}$$

On the other hand, if $U^{(l)} = \infty$ for some l, the above inequality holds trivially.

Proof of Corollary 13. As mentioned in Remark 14, we prove the corollary in the setting where, for each $l = 1, \ldots, L$, the tail Lévy measure satisfies

$$\overline{\rho}^{(l)} \overset{u \to 0}{\sim} u^{-\alpha^{(l)}} L^{(l)}\left(\frac{1}{u}\right) \tag{77}$$

where $\alpha^{(l)} \in [0, 1)$ and $L^{(l)}$ is a slowly varying function. We also let $M_1^{(l)} = \int_0^\infty u\rho^{(l)}(du)$, $\alpha = \max_l \alpha^{(l)}$ and $0 < \delta < 1 - \alpha$.

From Theorem 29, it is straightforward to check that, for each $l = 1, \ldots, L$,

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \epsilon\}} \xrightarrow{\mathrm{d}} \mathrm{ID}(0, \rho^{(l)}|_{(0,\epsilon]}), \tag{78}$$

as $p_l \to \infty$, where we denote by $\rho^{(l)}|_{(0,\epsilon]}$ the measure $\rho^{(l)}$ restricted to $(0, \epsilon]$. (Technically, we need to assume that ϵ is a continuity point of $\rho^{(l)}$. See Remark 40.) As the family in Equation (75) is uniformly integrable, it follows that, for each $l = 1, \ldots, L$,

$$A_{p_l}^{(l)} = \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \epsilon\}}\right] \to \int x \mathbf{1}_{\{x \leq \epsilon\}} \rho^{(l)}(dx)$$

as $p_l \to \infty$. To finish the proof, we introduce a corollary to Lemma 24.

Corollary 39 *Recall that $\alpha = \max_l \alpha^{(l)}$ and $0 < \delta < 1 - \alpha$. Then, for each $l = 1, \ldots, L$, there exists $\epsilon^{(l)} > 0$ such that*

$$\overline{\rho}^{(l)}(x) \leq x^{-(\alpha^{(l)} + \delta)}$$

for all $x \leq \epsilon^{(l)}$. Consequently, for all $l = 1, \ldots, L$,

$$\overline{\rho}^{(l)}(x) \leq x^{-(\alpha + \delta)}$$

for all $x \leq \epsilon_0$, where $\epsilon_0 = \min_l \epsilon^{(l)} > 0$.

Proof Since $\overline{\rho}^{(l)}(x) \sim x^{-\alpha^{(l)}} L^{(l)}(1/x)$, for any $\eta > 0$, we can find $\tilde{\epsilon}^{(l)}$ such that for all $x \leq \tilde{\epsilon}^{(l)}$

$$\overline{\rho}^{(l)}(x) \leq (1+\eta) x^{-\alpha^{(l)}} L^{(l)}(1/x).$$

Note that $(1+\eta)L^{(l)}(1/x)$ is still a slowly varying function. By Lemma 24, there exists $\epsilon^{(l)} \leq \tilde{\epsilon}^{(l)}$ such that $(1+\eta)L^{(l)}(1/x) \leq x^{-\delta}$ for all $x \leq \epsilon^{(l)}$. Thus, for every $x \leq \epsilon^{(l)}$, we have

$$\overline{\rho}^{(l)}(x) \leq x^{-(\alpha^{(l)}+\delta)}.$$

The second statement automatically follows. ■

By Corollary 39, for any $0 < \delta < 1 - \alpha$, there exists $\epsilon_0(\delta)$ such that for $\epsilon < \epsilon_0(\delta)$,

$$\int x \mathbf{1}_{\{x \leq \epsilon\}} \, \rho^{(l)}(dx) \leq \int_0^{\epsilon} \overline{\rho}^{(l)}(t) \, dt \leq \frac{1}{1-(\alpha+\delta)} \epsilon^{1-(\alpha+\delta)}.$$

Thus, for each $l = 1, \ldots, L$,

$$\lim_{p_l \to \infty} A_{p_l}^{(l)} \leq \frac{1}{1-(\alpha+\delta)} \epsilon^{1-(\alpha+\delta)}.$$

By taking the limit of Equation (76) as $\min \mathbf{p} \to \infty$, we get

$$\begin{aligned}
&\lim_{\min \mathbf{p} \to \infty} \mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] \\
&\leq (\sigma_v^2 C_{\text{Lip}}^2 U^{(l)} + (\sigma_v^2 C_{\text{Lip}}^2)^2 M^{(l)} U^{(l-1)} + \cdots + (\sigma_v^2 C_{\text{Lip}}^2)^l M^{(l)} \cdots M^{(2)} U^{(1)}) \frac{\epsilon^{1-(\alpha+\delta)}}{1-(\alpha+\delta)} \\
&= D(l) \epsilon^{1-(\alpha+\delta)}
\end{aligned}$$

where

$$D(l) := \frac{\sigma_v^2 C_{\text{Lip}}^2}{1-(\alpha+\delta)} (U^{(l)} + (\sigma_v^2 C_{\text{Lip}}^2) M^{(l)} U^{(l-1)} + \cdots + (\sigma_v^2 C_{\text{Lip}}^2)^{l-1} M^{(l)} \cdots M^{(2)} U^{(1)})$$

is a constant not depending on ϵ.

Remark 40 *When checking Equation* (78)*, it is necessary that the pruning level ϵ is a continuity point of $\rho^{(l)}$ for all $l = 1, \ldots, L$. However, if ϵ is not a continuity point of some $\rho^{(l)}$, we can find such a continuity point that is arbitrarily close to ϵ (or arbitrarily small; see Corollary 39).*

Proof of Corollary 15. By Proposition 32, for each $l = 1, \ldots, L$ and any $\kappa \in (0,1)$,

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l,(\kappa p_l)}^{(l)}\}} \xrightarrow{\text{pr}} 0$$

as $p_l \to \infty$. As the family in Equation (75) is uniformly integrable, it follows that, for each $l = 1, \ldots, L$ and $\kappa \in (0,1)$,

$$A_{p_l}^{(l)} = \mathbf{E}\left[\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \mathbf{1}_{\{\lambda_{p_l,j}^{(l)} \leq \lambda_{p_l,(\kappa p_l)}^{(l)}\}}\right] \to 0$$

as $p_l \to \infty$. Thus, by taking the limit of Equation (76) as $\min \mathbf{p} \to \infty$, we get

$$\lim_{\min \mathbf{p} \to \infty} \mathbf{E}\left[\left(Z_1^{(l+1)}(\mathbf{x};\mathbf{p}) - Z_1^{*(l+1)}(\mathbf{x};\mathbf{p})\right)^2\right] = 0.$$

C.5 Proof of Theorem 16

Let us denote $\vec{\mathbf{x}} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the tuple of the n inputs in the theorem. Throughout this subsection, let $\Sigma^{(l)}(\vec{\mathbf{x}}) \in \mathbb{R}^{n\times n}$ be a (possibly random) covariance matrix depending on $\vec{\mathbf{x}}$, and let $(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}}))_{j\geq 1}$ be iid centred Gaussian random vectors in $\mathbb{R}^n$, given the covariance matrix $\Sigma^{(l)}(\vec{\mathbf{x}})$. Since the matrix $\phi(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}}))^T$ is clearly positive semi-definite, we obtain the following corollary from Lemma 35. Consider $\vec{z} := (z_1, \ldots, z_n)^T \in \mathbb{R}^n$. For a given activation function ϕ, define the map

$$L_{\vec{z}}(u) = u\phi(\vec{z})\phi(\vec{z})^T : [0,\infty) \to \mathbb{R}^{n\times n}.$$

Corollary 41 *Let $(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}}))_j$ be iid centred Gaussian random vectors in $\mathbb{R}^n$ with covariance $\Sigma^{(l)}(\vec{\mathbf{x}})$. Then, for $l \geq 1$,*

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \phi\left(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}})\right) \phi\left(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}})\right)^T$$

converges in distribution, as $p_l \to \infty$, to $S^{(l)}(\vec{\mathbf{x}}) \sim \mathrm{ID}\left(\tilde{a}^{(l)}, \tilde{\rho}^{(l)}\right)$ concentrating on the cone $\mathbb{K}_n$ of $n \times n$ positive semi-definite matrices, where, for $l \geq 1$, $\tilde{a}^{(l)}$ and $\tilde{\rho}^{(l)}$ are defined as

$$\tilde{a}^{(l)} := a^{(l)}\mathbf{E}\left[\phi\left(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}})\right)\phi\left(\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}})\right)^T\right]$$

$$\tilde{\rho}^{(l)}(B) := \int (L_{\vec{z}})_\star(\rho^{(l)})(B)\,\xi(d\vec{z})$$

with ξ being the measure of $\mathcal{N}(0, \Sigma^{(l)}(\vec{\mathbf{x}}))$ (i.e., the law of $\vec{\zeta}_j^{(l)}(\vec{\mathbf{x}})$) and $(L_{\vec{z}})_\star(\rho^{(l)})$ denoting the pushforward of $\rho^{(l)}$ under the map $L_{\vec{z}}$.

Note that the above corollary is later used in the setting of a random covariance matrix. However, when it is used, we condition on the covariance, and thus for all intents and purposes, the covariance is nonrandom.

Remark 42 *Throughout the paper, we use two 'dual' perspectives of viewing the measure ν. For a distribution F and a measure ρ, the measure $\nu(dz) = \int \rho(dx)F(dz/x)$ is a mixture of a pushforward distribution $F(dz/x)$ (which maps each Borel set $B \subseteq \mathbb{K}_n$ to $F(\{z/x : z \in B\})$) with mixing measure $\rho(dx)$. Note that, in some other places, we use $\nu(dx) = \int (L_z)_\star(\rho)(dx)F(dz)$, where $L_z(x) = xz : \mathbb{R} \to \mathbb{K}_n$ for $x \in [0,\infty)$ and $z \in \mathbb{K}_n$, which can be described as a mixture of a pushforward of ρ by L_z with mixing measure $F(dz)$. Indeed, both definitions refer to the same measure: for a Borel set $B \subset \mathbb{K}_n$, $\nu(B) = (\rho \bigotimes F)(\{(x,z)\colon xz \in B\})$, where $\bigotimes$ denotes the product of measures.*

Proof of Theorem 16. Recall that $\vec{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is the tuple of the n inputs in the theorem. Let $\vec{t}_k \in \mathbb{R}^n$ for $k \leq m$ and $\vec{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$.

We start by calculating, for $m \in \mathbb{N}$, $l \geq 2$ and $\mathbf{p} \in \mathbb{N}^L$, the conditional characteristic function of the size-m tuple of n-dimensional random vectors

$$(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m} := (Z_k^{(l)}(\mathbf{x}_1;\mathbf{p}), \ldots, Z_k^{(l)}(\mathbf{x}_n;\mathbf{p}))_{k\leq m},$$

given $\{\lambda_{p_{l-1},j}^{(l-1)}\}_j$ and $\{\vec{Z}_j^{(l-1)}\}_j = \{\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p})\}_j$. Since both $\{V_{jk}^{(l)}\}_{jk}$ and $\{B_k^{(l)}\}_k$ are iid and also independent of each other, we have

$$\mathbf{E}\left[\prod_{k\leq m} \exp\left(i\langle \vec{t}_k, B_k^{(l)}\vec{1}\rangle + i\sum_{j=1}^{p_{l-1}} \left\langle \vec{t}_k, \sqrt{\lambda_{p_{l-1},j}^{(l-1)}} V_{jk}^{(l)} \phi(\vec{Z}_j^{(l-1)})\right\rangle\right) \middle| \{\lambda_{p_{l-1},j}^{(l-1)}\}_j, \{\vec{Z}_j^{(l-1)}\}_j\right]$$

$$= \prod_{k\leq m} \mathbf{E}\Big[\exp\Big(i\langle \vec{t}_k, B_k^{(l)}\vec{1}\rangle\Big)\Big] \prod_{j=1}^{p_{l-1}} \mathbf{E}\Big[\exp\Big(i\Big\langle \vec{t}_k, \sqrt{\lambda_{p_{l-1},j}^{(l-1)}} V_{jk}^{(l)} \phi(\vec{Z}_j^{(l-1)})\Big\rangle\Big)\Big| \{\lambda_{p_{l-1},j}^{(l-1)}\}_j, \{\vec{Z}_j^{(l-1)}\}_j\Big]$$

$$= \prod_{k\leq m} \exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \left[\sigma_b^2 \vec{1}\vec{1}^T + \sigma_v^2 \sum_{j=1}^{p_{l-1}} \lambda_{p_{l-1},j}^{(l-1)} \phi(\vec{Z}_j^{(l-1)})\phi(\vec{Z}_j^{(l-1)})^T\right]\vec{t}_k\right\rangle\right).$$

Thus, the (unconditional) characteristic function $\psi_{(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m}}$ satisfies

$$\psi_{(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m}}(\vec{t}_1,\ldots,\vec{t}_m)$$
$$= \mathbf{E}\left[\prod_{k\leq m} \exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \left[\sigma_b^2 \vec{1}\vec{1}^T + \sigma_v^2 \sum_{j=1}^{p_{l-1}} \lambda_{p_{l-1},j}^{(l-1)} \phi(\vec{Z}_j^{(l-1)})\phi(\vec{Z}_j^{(l-1)})^T\right]\vec{t}_k\right\rangle\right)\right]. \quad (79)$$

We now prove the following convergence in distribution using induction on the layer $l = 1,\ldots,(L+1)$: for all $m \in \mathbb{N}$, where we henceforth assume without loss of generality that $m \leq p_l$,

$$\left(\lim_{p_{l-1}\to\infty} \cdots \lim_{p_1\to\infty} (\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m}\right) \overset{\mathrm{d}}{=} (\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\leq m}$$

where $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\leq m}$ is the size-m tuple of the random vectors $\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}) = (\zeta_k^{(l)}(\mathbf{x}_1),\ldots,\zeta_k^{(l)}(\mathbf{x}_n))^T$ in $\mathbb{R}^n$ defined as follows. When conditioned on $\Sigma^{(l)}$, the tuple $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\leq m}$ is distributed as $\otimes_{k\leq m}\mathcal{N}(0,\Sigma^{(l)})$, where $\Sigma^{(l)}$ is a random covariance matrix defined recursively by the kernels in Equation (40). In other words, conditioned on $\Sigma^{(l)}$, the m components of $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\leq m}$ are iid Gaussian vectors with covariance matrix $\Sigma^{(l)}$, while if we do not condition, then the joint distribution $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\leq m}$ is a convex mixture of iid Gaussian vectors with covariance matrix $\Sigma^{(l)}$ with the mixture being governed by the randomness of $\Sigma^{(l)}$. Since $(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m}$ does not depend on $p_l,\ldots,p_L$, (as long as p_l is bigger or equal to m) the above convergence implies the claim of the theorem. In our inductive proof, we explicitly denote the dependency of $\Sigma^{(l)}$ on the inputs $\vec{\mathbf{x}}$ by writing $\Sigma^{(l)}(\vec{\mathbf{x}})$. Also, we fix $m \in \mathbb{N}$.

Case $l = 1$:
Note that in this case, for all $\mathbf{p} \in \mathbb{N}^L$, both $(\vec{Z}_k^{(1)}(\vec{\mathbf{x}};\mathbf{p}))_{k\geq 1}$ and $(\vec{\zeta}_k^{(1)}(\vec{\mathbf{x}}))_{k\geq 1}$ are iid even without conditioning since $\Sigma^{(1)}$ is nonrandom. Each component follows the law $\mathcal{N}(0,\Sigma^{(1)}(\vec{\mathbf{x}}))$ and

$$(\vec{Z}_k^{(1)}(\vec{\mathbf{x}};\mathbf{p}))_{k\leq m} \overset{\mathrm{d}}{=} (\vec{\zeta}_k^{(1)}(\vec{\mathbf{x}}))_{k\leq m},$$

holds for all $\mathbf{p} \in \mathbb{N}^L$.

Case $l \geq 2$:
By the induction hypothesis, we have the following convergence in distribution: for all $m' \in \mathbb{N}$, where we assume without loss of generality that $m' \leq p_{l-1}$,

$$\left(\lim_{p_{l-2}\to\infty} \cdots \lim_{p_1\to\infty} (\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\leq m'}\right) \overset{\mathrm{d}}{=} (\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\leq m'}, \quad (80)$$

where the left-hand side is interpreted as $(\vec{Z}_j^{(1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\leq m'}$ when $l = 2$.

For all non-negative reals $c_1,\ldots,c_{p_{l-1}}$, and for all $m \in \mathbb{N}$ and $\vec{t}_1,\ldots,\vec{t}_m \in \mathbb{R}^n$, the function $h : \mathbb{R}^{n\times p_{l-1}} \to (0,\infty)$ defined by

$$h\left((\vec{z}_j)_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(c_j)_{j\le p_{l-1}}\right)= \prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k,\left[\sigma_b^2\vec{1}\vec{1}^T+\sigma_v^2\sum_{j=1}^{p_{l-1}}c_j\phi(\vec{z}_j)\phi(\vec{z}_j)^T\right]\vec{t}_k\right\rangle\right), \quad (81)$$

is continuous in $(\vec{z}_j)_{j\le p_{l-1}}$, non-negative, and bounded by 1. Thus, we can use the induction hypothesis for the $(l-1)$-th layer in Equation (80) and, from the definition of convergence in distribution, deduce the following convergence for this bounded continuous function h:

$$\lim_{p_{l-2}\to\infty}\cdots\lim_{p_1\to\infty}\mathbf{E}\left[h\left((\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}'))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(c_j)_{j\le p_{l-1}}\right)\right]= \mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(c_j)_{j\le p_{l-1}}\right)\right].$$

Also, by Equation (79),

$$\begin{aligned}&\psi_{(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\le m}}(\vec{t}_1,\ldots,\vec{t}_m)\\ &\quad=\mathbf{E}\left[h\left((\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\right]\\ &\quad=\mathbf{E}\left[\mathbf{E}\left[h\left((\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right]\right].\end{aligned}$$

Since h is bounded, by the dominated convergence theorem, we have

$$\begin{aligned}&\lim_{p_{l-2}\to\infty}...\lim_{p_1\to\infty}\psi_{(\vec{Z}_k^{(l)}(\vec{\mathbf{x}};\mathbf{p}))_{k\le m}}(\vec{t}_1,\ldots,\vec{t}_m)\\ &=\lim_{p_{l-2}\to\infty}...\lim_{p_1\to\infty}\mathbf{E}\left[\mathbf{E}\left[h\left((\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right]\right]\\ &=\mathbf{E}\left[\lim_{p_{l-2}\to\infty}...\lim_{p_1\to\infty}\mathbf{E}\left[h\left((\vec{Z}_j^{(l-1)}(\vec{\mathbf{x}};\mathbf{p}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right]\right]\\ &=\mathbf{E}\left[\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right]\right]\\ &=\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\right].\end{aligned}$$

To complete the inductive step, we now show that

$$\lim_{p_{l-1}\to\infty}\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\right]=\psi_{(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\le m}}(\vec{t}_1,\ldots,\vec{t}_m).$$

Since h is bounded, by the dominated convergence theorem, we have

$$\begin{aligned}&\lim_{p_{l-1}\to\infty}\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\right]\\ &\quad=\lim_{p_{l-1}\to\infty}\mathbf{E}\left[\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]\right]\\ &\quad=\mathbf{E}\left[\lim_{p_{l-1}\to\infty}\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]\right]\end{aligned}$$

where the nested conditional expectation has the following form by the definition of h:

$$\begin{aligned}&\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}};(\vec{t}_k)_{k\le m},(\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]=\\ &\quad\mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle\vec{t}_k,\left[\sigma_b^2\vec{1}\vec{1}^T+\sigma_v^2\sum_{j=1}^{p_{l-1}}\lambda_{p_{l-1},j}^{(l-1)}\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))^T\right]\vec{t}_k\right\rangle\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right].\end{aligned}$$

Note that by Corollary 41, when conditioned on $\Sigma^{(l-1)}(\vec{\mathbf{x}})$,

$$\left(\sum_{j=1}^{p_{l-1}} \lambda_{p_{l-1},j}^{(l-1)} \phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))^T\right) \xrightarrow{\mathrm{d}} S^{(l-1)}(\vec{\mathbf{x}}) \in \mathbb{R}^{n\times n} \text{ as } p_{l-1}\to\infty,$$

where the random matrix $S^{(l-1)}(\vec{\mathbf{x}})$ has an infinitely divisible distribution with Lévy characteristic $(\tilde{a}^{(l-1)}, \tilde{\rho}^{(l-1)})$ as defined in Corollary 41. Since $\prod_{k\le m}\exp(-y_k/2)$ is bounded by 1 for all non-negative reals $y_1,\dots,y_k$, the above convergence implies that as p_{l-1} tends to ∞,

$$\begin{aligned}\mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \left[\sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2\sum_{j=1}^{p_{l-1}}\lambda_{p_{l-1},j}^{(l-1)}\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))^T\right]\vec{t}_k\right\rangle\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]\\ \to \mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \left[\sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2 S^{(l-1)}(\vec{\mathbf{x}})\right]\vec{t}_k\right\rangle\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right].\end{aligned}$$

Thus,

$$\begin{aligned}&\lim_{p_{l-1}\to\infty}\mathbf{E}\left[h\left((\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))_{j\le p_{l-1}}; (\vec{t}_k)_{k\le m}, (\lambda_{p_{l-1},j}^{(l-1)})_{j\le p_{l-1}}\right)\right]\\ &= \mathbf{E}\left[\mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \left[\sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2 S^{(l-1)}(\vec{\mathbf{x}})\right]\vec{t}_k\right\rangle\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]\right]\\ &= \mathbf{E}\left[\mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \Sigma^{(l)}(\vec{\mathbf{x}})\vec{t}_k\right\rangle\right)\,\middle|\,\Sigma^{(l-1)}(\vec{\mathbf{x}})\right]\right]\\ &= \mathbf{E}\left[\prod_{k\le m}\exp\left(-\frac{1}{2}\left\langle \vec{t}_k, \Sigma^{(l)}(\vec{\mathbf{x}})\vec{t}_k\right\rangle\right)\right]\\ &= \psi_{(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\le m}}(\vec{t}_1,\dots,\vec{t}_m).\end{aligned}$$

The last equality follows from the definition of $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\le m}$ and the fact that when conditioned on $\Sigma^{(l)}(\vec{\mathbf{x}})$, the random variables $(\vec{\zeta}_k^{(l)}(\vec{\mathbf{x}}))_{k\le m}$ are iid with each component having the distribution $\mathcal{N}(0,\Sigma^{(l)}(\vec{\mathbf{x}}))$. The justification of the second equality is slightly more involved. It follows from the boundedness of $\prod_{k\le m}\exp(-y_k/2)$ for all non-negative reals $y_1,\dots,y_k$, and the below conditional distributional equality: when conditioned on $\Sigma^{(l-1)}(\vec{\mathbf{x}})$,

$$\sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2 S^{(l-1)}(\vec{\mathbf{x}}) \overset{\mathrm{d}}{=} \Sigma^{(l)}(\vec{\mathbf{x}}),$$

which follows from the definition of $\tilde{\rho}^{(l-1)}$. To see this, condition on $\Sigma^{(l-1)}(\vec{\mathbf{x}})$. Then, as $S^{(l-1)}(\vec{\mathbf{x}})$ follows $\mathrm{ID}(\tilde{a}^{(l-1)},\tilde{\rho}^{(l-1)})$, it can be represented as $\tilde{a}^{(l-1)} + \sum_{j\ge1} X_j$ where $(X_j)_{j\ge1}$ are points in $\mathbb{R}^{n\times n}$ which result from the pushforward of a Poisson process on $(0,\infty)$ with mean measure $\rho^{(l-1)}$, as described in Corollary 41. Since $\tilde{\rho}^{(l-1)}$ is a pushforward of $\rho^{(l-1)}(du)$, X_j can be represented as $\widetilde{\lambda}_j^{(l-1)}\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))^T$. Thus, under our assumed conditioning on $\Sigma^{(l-1)}(\vec{\mathbf{x}})$, we have

$$\sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2 S^{(l-1)}(\vec{\mathbf{x}}) \overset{\mathrm{d}}{=} \sigma_b^2\vec{1}\vec{1}^T + \sigma_v^2\tilde{a}^{(l-1)} + \sigma_v^2\sum_{j\ge1}\widetilde{\lambda}_j^{(l-1)}\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))\phi(\vec{\zeta}_j^{(l-1)}(\vec{\mathbf{x}}))^T \quad \overset{\mathrm{d}}{=} \Sigma^{(l)}(\vec{\mathbf{x}}).$$

This completes the proof of the inductive case.

Appendix D. Additional Theoretical Results

D.1 Properties of Small Weights in our Model

The following proposition characterises the rate of decay of the variances/weights, under a polynomial decay of the tail Lévy measure at 0.

Proposition 43 (Asymptotic properties of small variances and weights) *Assume $\rho^{(l)}$ is an infinite Lévy measure with tail $\overline{\rho}^{(l)}(x) \overset{x\to 0}{\sim} cx^{-\alpha}$ for some $\alpha \in (0,1)$ and some constant $c > 0$. Let $\Phi^{(l)}_{(k)}$ and $\Psi^{(l+1)}_{(k),m}$ be random variables in Proposition 4 such that $\lambda^{(l)}_{p_l,(k)} \xrightarrow{\mathtt{d}} \Phi^{(l)}_{(k)}$ and $(W^{(l+1)}_{(k),m})^2 \xrightarrow{\mathtt{d}} \Psi^{(l+1)}_{(k),m}$ as p_l tends to ∞. Then, in probability,*

$$\Phi^{(l)}_{(k)} \overset{k\to\infty}{\sim} (\overline{\rho}^{(l)})^{-1}(k) \overset{k\to\infty}{\sim} c^{1/\alpha} k^{-1/\alpha} \tag{82}$$

and for any $m \geq 1$, in probability,

$$\Psi^{(l+1)}_{(k),m} \overset{k\to\infty}{\sim} \sigma_v^2 \times (\overline{\nu}^{(l)})^{-1}(k) \overset{k\to\infty}{\sim} \left(\frac{2^\alpha \Gamma(\alpha+1/2)}{\sqrt{\pi}} (\sigma_v^2)^\alpha c\right)^{1/\alpha} k^{-1/\alpha}. \tag{83}$$

Proof By Proposition 4, $\lambda^{(l)}_{p_l,(k)} \xrightarrow{\mathtt{d}} (\overline{\rho}^{(l)})^{-1}(G_k)$ as $p_l \to \infty$, where $G_k \sim \text{Gamma}(k,1)$. Additionally, by the law of large numbers $\frac{G_k}{k} \xrightarrow{\mathtt{pr}} 1$ as $k \to \infty$. Also, $\overline{\rho}^{(l)}(x) \overset{x\to 0}{\sim} cx^{-\alpha}$ implies $(\overline{\rho}^{(l)})^{-1}(x) \overset{x\to\infty}{\sim} (x/c)^{-1/\alpha}$. The rest follows from properties of regularly varying functions, see Proposition 23 in Appendix A.3. Additionally, similarly to the proof of Proposition 6, $\overline{\rho}^{(l)}(x) \overset{x\to 0}{\sim} cx^{-\alpha}$ implies

$$\overline{\nu}^{(l)}(x) \overset{x\to 0}{\sim} \frac{2^\alpha \Gamma(\alpha+1/2)}{\sqrt{\pi}} x^{-\alpha} c$$

concluding the proof. ■

D.2 Infinite-Width Limit for Multiple Inputs in the Symmetric α-Stable Case

If, for some $\alpha \in (0,1)$,

$$\sum_{j=1}^{p_l} \lambda^{(l)}_{p_l,j} \xrightarrow{\mathtt{d}} \text{Stable}(\alpha, 1) \text{ as } p_l \to \infty,$$

then the Poisson point process $\{\widetilde{\lambda}^{(l)}_j\}_{j\geq 1}$ in Theorem 16 has mean measure

$$\rho^{(l)}(du) = \frac{\alpha}{\Gamma(1-\alpha)} u^{-\alpha-1} du.$$

Let $\mathbb{K}_n$ denote the set of n-by-n positive semi-definite matrices and define the limit in distribution

$$\zeta^{(l)}_k(\mathbf{x}) = \lim_{p_{l-1}\to\infty} \dots \lim_{p_1\to\infty} Z^{(l)}_k(\mathbf{x}; \mathbf{p})$$

which we recall is conditionally Gaussian given the previous layers $1, \dots, l-1$. If $\phi(\zeta^{(l)}_k(\mathbf{x}))$ has sufficient moments, then for all n inputs $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^{d_{\text{in}}}$, the n-by-n random matrix

$$\begin{aligned} R^{(l)} &:= \left(K^{(l)}(\mathbf{x}_i, \mathbf{x}_j) - \sigma_b^2\right)_{i,j=1}^n \\ &= \left(\sigma_v^2 \sum_{k\geq 1} \widetilde{\lambda}^{(l)}_k \phi\left(\zeta^{(l)}_k(\mathbf{x}_i)\right) \phi\left(\zeta^{(l)}_k(\mathbf{x}_j)\right)\right)_{i,j=1}^n \end{aligned} \tag{84}$$

has a (strictly) α-stable distribution in $\mathbb{K}_n$ with characteristic exponent α, i.e.,

$$c_1^{1/\alpha} R_1^{(l)} + c_2^{1/\alpha} R_2^{(l)} \stackrel{\mathrm{d}}{=} (c_1 + c_2)^{1/\alpha} R^{(l)},$$

where $R_1^{(l)}$ and $R_2^{(l)}$ are independent copies of $R^{(l)}$. This allows us to strengthen Theorem 16 in this special case. For a single input $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, we obtain a more precise form of the limiting output distribution.

Theorem 44 (α-stable case, further results) *Suppose that in the l-th hidden layer,*

$$\sum_{j=1}^{p_l} \lambda_{p_l,j}^{(l)} \stackrel{\mathrm{d}}{\to} \text{Stable}(\alpha, 1) \text{ as } p_l \to \infty \tag{85}$$

for $\alpha \in (0, 1]$. Then, for any inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, Theorem 16 holds with the random matrix $(K^{(l)}(\mathbf{x}_i, \mathbf{x}_j) - \sigma_b^2)_{i,j=1}^n$ in Equation (84) having a conditional distribution, given $K^{(1)}, \ldots, K^{(l-1)}$, that is α-stable in $\mathbb{K}_n$. Moreover, in the case $n = 1$ with the single input $\mathbf{x}$, the conditional distribution of $K^{(l)}(\mathbf{x}, \mathbf{x}) - \sigma_b^2$ is $\text{Stable}(\alpha, r^{(l)}(\mathbf{x}))$ *where*

$$r^{(l)}(\mathbf{x}) := \sigma_v^2 \cdot \left(\mathbf{E}\left[\left| \phi(\zeta_1^{(l)}(\mathbf{x})) \right|^{2\alpha} \,\middle|\, K^{(1)}, \ldots, K^{(l-1)} \right] \right)^{1/\alpha}.$$

Thus, given $K^{(1)}, \ldots, K^{(l-1)}$, the random variance $K^{(l)}(\mathbf{x}, \mathbf{x})$ has the same conditional distribution as

$$\sigma_b^2 + \sigma_v^2 \cdot \tilde{\Lambda} \cdot \left(\mathbf{E}\left[\left| \phi(\zeta_1^{(l)}(\mathbf{x})) \right|^{2\alpha} \,\middle|\, K^{(1)}, \ldots, K^{(l-1)} \right] \right)^{1/\alpha}$$

for $\tilde{\Lambda} \sim \text{Stable}(\alpha, 1)$.

Proof We assume, without loss of generality, that $(\widetilde{\lambda}_k^{(l)})_{k \geq 1}$ is ordered:

$$\widetilde{\lambda}_1^{(l)} \geq \widetilde{\lambda}_2^{(l)} \geq \cdots.$$

By Proposition 30, the order statistics $(\lambda_{p_l,(1)}^{(l)}, \ldots, \lambda_{p_l,(p_l)}^{(l)}, 0, \ldots)$ converge in distribution to $(\widetilde{\lambda}_k^{(l)})_{k \geq 1}$, a Poisson point process with intensity measure $\rho^{(l)}(du) = \alpha\Gamma(1-\alpha)^{-1} u^{-\alpha-1} du$. Such a Poisson process takes the form $((G_k \Gamma(1-\alpha))^{-1/\alpha})_{k \geq 1}$ where $(G_k)_{k \geq 1}$ is a standard rate-one Poisson point process on $(0, \infty)$ (LePage et al., 1981; Davydov et al., 2008).

Given $K^{(1)}, \ldots, K^{(l-1)}$, we have that $\left((\zeta_k^{(l)}(\mathbf{x}_1), \ldots, \zeta_k^{(l)}(\mathbf{x}_n))^T \right)_{k \geq 1}$ are iid with common distribution $\mathcal{N}(0, \Sigma^{(l)})$. Since ϕ satisfies the polynomial envelope condition, we have, conditioned on $K^{(1)}, \ldots, K^{(l-1)}$, that $\phi(\zeta_1^{(l)}(\mathbf{x}_i))\phi(\zeta_1^{(l)}(\mathbf{x}_j))$ has conditional moments of all order for all pairs (i, j) (importantly, recall that we defined the conditional expectation via regular conditional probabilities). By the three-series theorem, $\sum_{k \geq 1} \widetilde{\lambda}_k^{(l)} \phi(\zeta_k^{(l)}(\mathbf{x}_i))\phi(\zeta_k^{(l)}(\mathbf{x}_j))$ converges for all pairs (i, j). Therefore, by Theorem 2 in (LePage et al., 1981), $R^{(l)}$ is α-stable.

For the special case $n = 1$ with the single input $\mathbf{x}$, setting $U_k := \phi^2(\zeta_k^{(l)}(\mathbf{x}))$, by Corollary 37, the Lévy-Khintchine formula and the integral formula (Samorodnitsky and Taqqu, 1994, p. 15)

$$y^\alpha = \int_0^\infty (1 - e^{-xy}) \frac{\alpha}{\Gamma(1-\alpha)} x^{-\alpha-1} dx \quad \text{for} \quad 0 < \alpha < 1$$

we have that

$$\mathbf{E}\left[\exp\left(-t \sum_k \widetilde{\lambda}_k^{(l)} U_k \right) \right] = \exp\left(-\mathbf{E}\left[\int_0^\infty \left(1 - e^{-txU_1}\right) \frac{\alpha}{\Gamma(1-\alpha)} x^{-\alpha-1} dx \right] \right)$$

$$= \exp\left(-t^{\alpha}\mathbf{E}[U_1^{\alpha}]\right).$$

■

Appendix E. Additional Details on the Examples

E.1 Detailed Illustration of the Main Results on the Simple Model in Section 1

Recall the four models briefly discussed in Section 1: for $p_1 \geq 2$,

$$\text{(a) } \lambda^{(1)}_{p_1,j} \sim \text{IG}\left(2, \tfrac{2}{p_1}\right) \qquad \text{(b) } \lambda^{(1)}_{p_1,j} \sim \text{Bernoulli}\left(\tfrac{2}{p_1}\right)$$
$$\text{(c) } \lambda^{(1)}_{p_1,j} \sim \text{Beta}\left(\tfrac{1}{p_1}, \tfrac{1}{2}\right) \qquad \text{(d) } \lambda^{(1)}_{p_1,j} = \pi^2 \tfrac{U_j^2}{p_1^2} \text{ where } U_j \sim \text{Cauchy}_+(0,1)$$

where $\text{IG}(\beta_1, \beta_2)$ denotes the inverse gamma distribution with shape $\beta_1 > 0$ and scale $\beta_2 > 0$, and $\text{Cauchy}_+(0,1)$ denotes the half-Cauchy distribution with pdf in Equation (5). The 50 largest values of a realisation of $(\lambda^{(1)}_{p_1,j})_{j=1,\ldots,p_1}$ for a neural network of width $p_1 = 5000$ are represented in Figure 10 under these models.

These four models have different infinite-width limits. Under the inverse gamma model (a), the infinite-width limit is the same as the iid Gaussian case, and Equation (3) holds. Under models (b-d), the infinite-width limit is a mixture of Gaussian processes (see Theorem 16). That is, for each case $s \in \{b, c, d\}$,

$$\begin{pmatrix} Z^{(2)}_k(\mathbf{x}; p_1) \\ Z^{(2)}_k(\mathbf{x}'; p_1) \end{pmatrix} \xrightarrow{\text{d}} \mathcal{N}\left(0, \begin{pmatrix} K^{(2)}_s(\mathbf{x}, \mathbf{x}) & K^{(2)}_s(\mathbf{x}, \mathbf{x}') \\ K^{(2)}_s(\mathbf{x}, \mathbf{x}') & K^{(2)}_s(\mathbf{x}', \mathbf{x}') \end{pmatrix}\right) \text{ as } p_1 \to \infty \tag{86}$$

where $K^{(2)}_b$, $K^{(2)}_c$ and $K^{(2)}_d$ are *random* covariance kernels defined by

$$K^{(2)}_s(\mathbf{x}, \mathbf{x}') := \sum_{j \geq 1} \widetilde{\lambda}^{(1)}_{(j),s} \max\left(0, \zeta^{(1)}_{(j),s}(\mathbf{x})\right) \max\left(0, \zeta^{(1)}_{(j),s}(\mathbf{x}')\right). \tag{87}$$

Here $\widetilde{\lambda}^{(1)}_{(1),s} \geq \widetilde{\lambda}^{(1)}_{(2),s} \geq \ldots \geq 0$ are random weights defined by (see details in Appendix E.3)

$$\widetilde{\lambda}^{(1)}_{(j),b} = \begin{cases} 1 & \text{if } j \leq N^{(1)} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } N^{(1)} \sim \text{Poisson}(2) \tag{88}$$

$$\widetilde{\lambda}^{(1)}_{(j),c} = \prod_{k=1}^{j} \beta_k \qquad \text{where } \beta_k \overset{\text{iid}}{\sim} \text{Beta}(2,1) \tag{89}$$

$$\widetilde{\lambda}_{(j),d} = \left(\sum_{k=1}^{j} E_j\right)^{-2} \qquad \text{where } E_k \overset{\text{iid}}{\sim} \text{Exponential}(2) \tag{90}$$

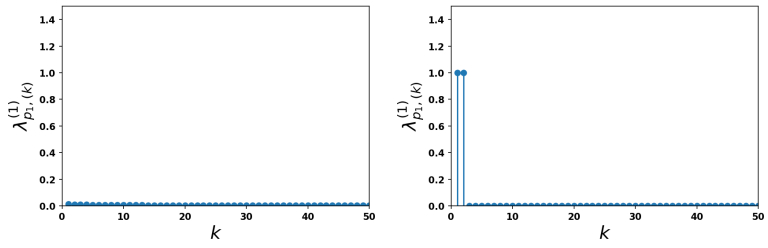
and, for $j \geq 1$, $\zeta^{(1)}_{(j),s} \overset{\text{iid}}{\sim} \text{GP}(0, K^{(1)})$ with $K^{(1)}(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^T \mathbf{x}'}{d_{\text{in}}}$.

We have $\mathbf{E}[K^{(2)}_b] = \mathbf{E}[K^{(2)}_c] = \mathcal{K}^{(2)}$. In the case (d), $\mathbf{E}[K^{(2)}_d]$ is undefined.

Due to the shared random covariance kernel, the outputs are therefore dependent in the infinite-width limit for examples (b-d). In the case (b), only a finite (random) number of nodes are active (that is, such that $\lambda^{(1)}_{p_1,j} > 0$) in the infinite-width limit; the infinite-width network is equivalent to a finite network with a Poisson(2) number of hidden nodes. In the cases (c-d), an infinite number of nodes are active in the infinite-width limit. The marginal random variances take the form $K^{(2)}_s(\mathbf{x}, \mathbf{x}) = (S^{(1)}_s \|\mathbf{x}\|^2)/d_{\text{in}}$ for $s \in \{b, c, d\}$, where (see Theorem 8, Appendices E.3.1 and E.3.2, and Corollary 27)

- $S^{(1)}_b \sim \text{Gamma}(\frac{N}{2}, \frac{1}{2})$ with $N \sim \text{Poisson}(1)$[5];

5. with the convention that $S^{(1)}_b = 0$ if $N = 0$.



(a) Inverse Gamma

(b) Bernoulli

(c) Beta

(d) Horseshoe

Figure 10: The 50 largest values $\lambda^{(1)}_{p_1,(1)} \geq \lambda^{(1)}_{p_1,(2)} \geq \ldots \geq \lambda^{(1)}_{p_1,(50)}$ of $(\lambda^{(1)}_{p_1,j})_{j=1,\ldots,p_1}$ in a neural network of width $p_1 = 5000$ for examples (a-d). The y-scale is the same for all the plots. For (a), all the values are non-zero and very small, of order $(2/p_1)$. For (b), only a small number of values are non-zero, all being equal to 1. For (c-d), all the values are non-zero. For model (c), $\lambda^{(1)}_{p_1,(k)}$ decreases exponentially fast with k, while it decreases in $O(k^{-2})$ for model (d).

- $S_c^{(1)} \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$;
- $S_d^{(1)} \sim \text{IG}(\frac{1}{2}, \frac{1}{2})$ is an inverse gamma/Lévy random variable.

In the case (c), $Z_k^{(2)}(\mathbf{x}; p_1)$ converges in distribution to a normal-gamma[6] random variable. For (d), it converges to a Cauchy random variable, and both dimensions of the output therefore have power-law tails, with exponent 1. For model (a), for any $k = 1, 2$, $\max_{j=1,\dots,p_1}(W_{jk}^{(2)})^2 \xrightarrow{\text{pr}} 0$, and the weights are all asymptotically small. For each case $s \in \{b, c, d\}$, $\max_j(W_{jk}^{(2)})^2 \xrightarrow{\text{d}} M_s$ where M_s is a random variable whose cdf is analytically available (see Section 3.4). In particular, in the horseshoe example (d), M_d follows a scaled Fréchet distribution. For models (b-d), let $(W_{(1),k}^{(2)})^2 \geq (W_{(2),k}^{(2)})^2 \geq \dots$ be the ordered weights connected to an output k. In the infinite-width limit, $W_{(j),k}^{(2)}$ decreases exponentially fast with j for model (c), while it decreases in $O(j^{-2})$ for model (d) (see Appendix D.1). These properties are of importance when using such models for pruning the nodes/edges of large networks. A related important property is that of the compressibility of the network. Let $\lambda_{p,(1)} \geq \lambda_{p,(2)} \geq \dots$ be the ordered per-node variance terms. For some compression level $\kappa \in (0, 1)$, let

$$Z_k^{*(2)}(\mathbf{x}; p_1, \kappa) := \sum_{j=1}^{p_1} \sqrt{\lambda_{p_1,j}^{(1)}} \mathbf{1}_{\{\lambda_{p_1,j}^{(1)} > \lambda_{p_1,(\lfloor \kappa p_1 \rfloor)}^{(1)}\}} V_{jk}^{(2)} \max\left(0, \frac{1}{\sqrt{d_{\text{in}}}} \sum_{i=1}^{d_{\text{in}}} V_{ij}^{(1)} x_i\right)$$

be the neural network obtained by pruning a $(1 - \kappa)$ proportion of nodes with the smallest $\lambda_{p_1,j}^{(1)}$ values. The models (b-d) are compressible in the sense that the difference between the pruned output $Z_k^{*(2)}(\mathbf{x}; p_1, \kappa)$ and the unpruned output $Z_k^{(2)}(\mathbf{x}; p_1)$ vanishes in probability in the infinite-width limit (see Theorem 5). This is not the case for the iid Gaussian model, nor for model (a), which are not compressible. The properties of the different models are summarised in Table 1.

E.2 Additional Examples

In Section 6, we discussed examples of models used in the literature, and the associated parameters of the limiting infinitely divisible random variable of Equation (2). In this part of the appendix, we provide additional examples and a general recipe behind some of these models and the Horseshoe model. As in Section 6, we use a different scaling in some cases so that the limit exists without being degenerate at 0.

All the proofs of the examples in this subsection rely on Theorem 29. Details are given in Appendix E.3. To simplify notation, we often drop the layer index l fully or partially in the rest of this subsection, writing e.g. $\lambda_{p,j} \sim \mu_p$.

E.2.1 Generalised Spike and Slab Prior

As a generalisation of the Bernoulli case in Section 6.2, we can consider the following spike and slab prior for the variances. Let $c > 0$ and $\tilde{c} \geq 0$. Consider

$$\lambda_{p,j} \sim \left(1 - \frac{c}{p}\right) \cdot \delta_{\tilde{c}/p} + \frac{c}{p} \cdot H$$

for some probability distribution H (slab) on $(0, \infty)$. We have

$$\sum_j \lambda_{p,j} \xrightarrow{\text{d}} \text{ID}(\tilde{c}, cH).$$

6. Various authors use the name normal-gamma to mean different things. Here we mean a Gaussian mixture distribution where the variance is governed by a gamma distribution.

Name	Mixture's name	μ_p	a	Lévy measure	Support	Finite?	Exp. α	Exp. τ
Determ.	Gaussian	$\delta_{c_1/p}$	c_1	0	–	–	–	–
Bernoulli	Spike and Slab	$\left(1-\frac{c}{p}\right)\cdot\delta_0+\frac{c}{p}\delta_1$	0	$c\delta_1$	$\{1\}$	Yes	0	–
Gamma	Group lasso	Gamma $\left(\frac{p_{l+1}+1}{2},\frac{p_l(p_{l+1}+1)}{2c_1}\right)$	c_1	0	–	–	–	–
Beta	Normal-beta	Beta $\left(\frac{1}{p},\frac{1}{c}\right)$	0	$x^{-1}(1-x)^{1/c-1}$	(0,1)	No	–	–
Inv.-Gamma	Multivariate t	IG $(2,2/p)$	2	0	–	–	–	–
Inv.-Gamma	Multivariate t	IG $\left(\alpha,\Gamma(1+\alpha)^{-1/\alpha}p^{-1/\alpha}\right)$	0	$\alpha x^{-\alpha-1}$	$(0,\infty)$	No	$\alpha\in(0,1)$	$\tau=\alpha$
Beta prime	Horseshoe	$\frac{2p}{\pi^2}x^{-1/2}(1+\frac{4xp^2}{\pi^2})^{-1}$	0	$\frac{1}{2}x^{-3/2}$	$(0,\infty)$	No	1/2	1/2
Resc. Beta Prime	Reg. Horseshoe	See Equation (95)	0	$\frac{x^{-3/2}}{\pi}(1-\frac{x}{c^2})^{-1/2}$	$(0,c^2)$	No	1/2	–
Gen. BFRY	Normal -gen. BFRY	See Equation (48)	0	$\frac{\eta x^{-1-\tau}}{\Gamma(1-\alpha)}\gamma(\tau-\alpha,x)$	$(0,\infty)$	No	$\alpha\in(0,1)$	$\tau>\alpha$

Table 8: List of models and their limiting location parameter and Lévy measure, with its properties.

E.2.2 Stable Limit and the Horseshoe Model

We describe here a family of models whose limit is a positive stable random variable, defined in Appendix A.4, which is a special kind of infinitely divisible random variable. The horseshoe model (Carvalho et al., 2010) in Section 6.6, which has been used by Louizos et al. (2017); Ghosh et al. (2018, 2019); Popkes et al. (2019) as a Bayesian prior for the weights of a deep neural network, arises as a special case.

Models converging to a stable distribution. We consider that

$$\lambda_{p,j}=\frac{Y_j}{(c_1p)^{1/\alpha}} \tag{91}$$

where $Y_1,Y_2,\ldots,$ are iid nonnegative random variables with cdf F and survival function $\overline{F}=1-F$ satisfying

$$\overline{F}(y)\overset{y\to\infty}{\sim}y^{-\alpha}c_1 \tag{92}$$

for some index $\alpha\in(0,1)$ and some positive constant c_1.[7] We have (Feller (1971, Theorem XVII.5.3), see also Janson (2011, Example 5.5))

$$\sum_j\lambda_{p,j}\xrightarrow{\mathrm{d}}\mathrm{ID}(0,\rho_{\mathrm{stable}}(\,\cdot\,;\alpha,1))=\mathrm{Stable}(\alpha,\Gamma(1-\alpha)^{1/\alpha}). \tag{93}$$

In this case, the limit Lévy measure is the positive stable Lévy measure with tail Lévy intensity $\overline{\rho}_{\mathrm{stable}}(u;\alpha,1)=u^{-\alpha}$. It has power-law tails at 0 and ∞, but with the same exponent α, thus lacking some flexibility. This limitation will be addressed by our later example which permits different exponents at 0 and ∞. There is a lot of flexibility in the choice of the distribution F. For example, all the following distributions have tails that satisfy Equation (92) for some constant $c_1>0$:

$$Y_j\sim\mathrm{Pareto}(\alpha,c),\quad Y_j\sim\mathrm{IG}(\alpha,1),\quad Y_j\sim\mathrm{Fréchet}(\alpha),\quad Y_j\sim\mathrm{Betaprime}(c,\alpha)$$

where $c>0$. Pareto(α,c) denotes the Pareto distribution with pdf $f(x)=\alpha c^\alpha x^{-\alpha-1}\mathbf{1}_{\{x>c\}}$, Fréchet$(\alpha)$ denotes the Fréchet distribution with cdf $F(x)=e^{-x^{-\alpha}}$ and Betaprime(c,α) denotes the beta prime distribution with pdf $f(x)=x^{c-1}(1+x)^{-c-\alpha}\frac{\Gamma(c+\alpha)}{\Gamma(c)\Gamma(\alpha)}$. Combining Equation (91) with $Y_j\sim\mathrm{IG}(\alpha,1)$ gives

$$\lambda_{p,j}\sim\mathrm{IG}\left(\alpha,\frac{1}{\Gamma(1+\alpha)^{1/\alpha}p^{1/\alpha}}\right).$$

While this model appears similar to the model in Equation (45), the asymptotic properties of the two models are very different.

7. More generally, one could replace the constant c_1 by a slowly varying function L.

Horseshoe distribution. The horseshoe model (Carvalho et al., 2010) arises as another special case. One assumes that the random variables Y_j have the same distribution as $Y = T^2$, where $T \sim \text{Cauchy}_+(0,1)$ is a half-Cauchy random variable, with pdf given by Equation (5). The random variable $Y \sim \text{Betaprime}(1/2, 1/2)$ is a beta prime random variable (with both shape parameters equal to $1/2$), with pdf

$$f_Y(y) = \frac{1}{\pi\sqrt{y}(1+y)}.$$

Its survival function satisfies

$$\Pr(Y > y) \overset{y\to\infty}{\sim} \frac{2}{\pi} y^{-1/2},$$

and therefore Y has a power-law tail at infinity with exponent $\alpha = 1/2$. Let $c > 0$ be some scaling parameter. Setting

$$\lambda_{p,j} = c \times \frac{\pi^2}{4} \frac{Y_j}{p^2},$$

we obtain

$$\sum_j \lambda_{p,j} \xrightarrow{\mathtt{d}} \text{ID}(0, \rho_{\text{stable}}(\,\cdot\,; 1/2, c)) = \text{Stable}(1/2, c\pi) = \text{IG}(1/2, c\pi/4). \tag{94}$$

The limit is therefore a stable distribution, with exponent $\frac{1}{2}$, which is also inverse gamma with shape parameter $1/2$ in this case. The tail Lévy intensity $\overline{\rho}_{\text{stable}}(x; 1/2, c)$ has power-law tails at 0 and ∞, with exponent $1/2$.

E.2.3 Regularised Horseshoe and Stable Beta Process

Ghosh et al. (2018, 2019) considered Bayesian learning of neural networks using regularised horseshoe priors. In this case, we have

$$\lambda_{p,j} = \frac{c^2 \frac{T_j^2}{p^2}}{c^2 + \frac{T_j^2}{p^2}}$$

where $T_j \sim \text{Cauchy}_+(0,1)$ and $c > 0$. Note that $\lambda_{p,j} \in (0, c^2)$ is bounded, with pdf

$$f_p(x) = \frac{p}{\pi} \cdot x^{-\frac{1}{2}} (1 - x/c^2)^{-\frac{3}{2}} \left(1 + \frac{p^2 x}{(1 - x/c^2)}\right)^{-1} \mathbf{1}_{\{x<c^2\}}. \tag{95}$$

We have, for any $x > 0$,

$$\lim_{p\to\infty} p f_p(x) = \frac{1}{\pi} \cdot x^{-\frac{3}{2}} \left(1 - x/c^2\right)^{-\frac{1}{2}} \mathbf{1}_{\{x<c^2\}}.$$

An application of Theorem 29 gives

$$\sum_j \lambda_{p,j} \xrightarrow{\mathtt{d}} \text{ID}(0, \rho)$$

where

$$\rho(du) = \frac{1}{\pi} \cdot u^{-\frac{3}{2}} \left(1 - u/c^2\right)^{-\frac{1}{2}} \mathbf{1}_{\{u<c^2\}} du$$

is a (scaled) stable beta Lévy measure (Teh and Gorur, 2009). The Lévy measure has bounded support, and the associated tail Lévy intensity increases polynomially at 0, with exponent $\alpha = 1/2$,

$$\overline{\rho}(x) \overset{x\to 0}{\sim} \frac{2}{\pi} x^{-1/2}.$$

The limiting random variable $\text{ID}(0, \rho)$ has support $(0, \infty)$.

E.2.4 General Models with Arbitrary Limiting Lévy Measure

It may be of interest to set the limiting constant a and the Lévy measure ρ, so that they satisfy a number of properties, and then pick a distribution μ_p that makes $\sum_j \lambda_{p,j} \xrightarrow{\mathtt{d}} \mathrm{ID}(a,\rho)$.

First note that if $\sum_j \lambda_{p,j} \xrightarrow{\mathtt{d}} \mathrm{ID}(0,\rho)$, then $\sum_j(\lambda_{p,j}+\frac{a}{p}) \xrightarrow{\mathtt{d}} \mathrm{ID}(a,\rho)$, so without loss of generality, we restrict the discussion to models with $a=0$.

If ρ is finite, then $H(dx)=\rho(dx)/\overline{\rho}(0)$ is a probability distribution, and one can simply set

$$\mu_p = \frac{\overline{\rho}(0)}{p} H + \left(1 - \frac{\overline{\rho}(0)}{p}\right)\delta_0.$$

If ρ is infinite, on the other hand, one can resort to the construction of Perman et al. (1992) (see also (Lee et al., 2019)), with

$$\mu_p(du) = \frac{(1-e^{-u\psi^{-1}(p)})}{p}\rho(du)$$

where ψ^{-1} is the generalised inverse of $\psi(t)=\int_0^\infty(1-e^{-ut})\rho(du)$, the Laplace exponent of ρ.

E.3 Derivations of the Beta and Horseshoe Examples in Sections 1 and 6

E.3.1 Beta Model (c) in Sections 1 and 6.5

Consider

$$\lambda_{p,j} \sim \mathrm{Beta}\left(\frac{\eta}{p}, b\right).$$

The pdf of $\lambda_{p,j}$ is

$$\begin{aligned} f_p(x) &= \frac{\Gamma(\eta/p+b)}{\Gamma(\eta/p)\Gamma(b)} x^{\eta/p-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}} \\ &= \frac{\eta}{p}\frac{\Gamma(\eta/p+b)}{\Gamma(\eta/p+1)\Gamma(b)} x^{\eta/p-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}}. \end{aligned}$$

We have

$$\begin{aligned} pf_p(x) &= \eta\frac{\Gamma(\eta/p+b)}{\Gamma(\eta/p+1)\Gamma(b)} x^{\eta/p-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}} \\ &\to \eta x^{-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}} = \varrho(x). \end{aligned}$$

This is the density of a stable beta measure ρ_{sb} with parameters $(\frac{\eta}{b},0,b)$. For $x\in(0,1)$,

$$\frac{pf_p(x)}{\varrho(x)} = \frac{\Gamma(\eta/p+b)}{\Gamma(\eta/p+1)\Gamma(b)} x^{\eta/p}\mathbf{1}_{\{x\in(0,1)\}} \le \frac{\Gamma(\eta/p+b)}{\Gamma(\eta/p+1)\Gamma(b)} \le 2$$

for p large enough. Finally, μ_p and ρ have the same support. Thus, Theorem 29 implies that $\sum_j \lambda_{p,j} \xrightarrow{\mathtt{d}} \mathrm{ID}(0,\rho)$ with $\rho(dx)=\eta x^{-1}(1-x)^{b-1}\mathbf{1}_{\{x\in(0,1)\}}dx$. The tail Lévy intensity satisfies

$$\overline{\rho}(x) \overset{x\to 0}{\sim} \eta\log(1/x).$$

Let $(\widetilde{\lambda}_{(1)} \ge \widetilde{\lambda}_{(2)} \ge \ldots)$ denote the ordered weights of a Poisson point process on $(0,\infty)$ with mean measure ρ. In the case $b=1$, they admit the simple inverse-Lévy/stick-breaking construction (Teh and Gorur, 2009)

$$\widetilde{\lambda}_{(j)} = \prod_{k=1}^{j}\beta_k, \quad \text{where } \beta_1,\beta_2\ \ldots \text{ are iid } \mathrm{Beta}(\eta,1).$$

E.3.2 Horseshoe Example (d) in Section 1 and Appendices E.1 and E.2.2

Let $\lambda_{p,j} = c\frac{\pi^2}{4}\frac{U_j^2}{p^2}$ where $U_j \sim \text{Cauchy}_+(0,1)$, with $c = 4$. From Appendix E.2.2,

$$\sum_j \lambda_{p,j} \xrightarrow{\text{d}} \text{ID}(0, \rho_{\text{stable}}(\,\cdot\,; 1/2, c)) = \text{Stable}(1/2, c\pi) = \text{IG}(1/2, c\pi/4).$$

Let $(\widetilde{\lambda}_{(1)} \geq \widetilde{\lambda}_{(2)} \geq \ldots)$ denote the ordered weights of a Poisson point process on $(0,\infty)$ with mean measure $\rho_{\text{stable}}(\,\cdot\,; 1/2, c)$. They admit the inverse-Lévy representation

$$\widetilde{\lambda}_{(j)} = \frac{c}{\left(\sum_{k=1}^{j} E_k\right)^2}$$

where $E_1, E_2, \ldots$, are iid exponential random variables with unit rate.

From Corollary 28,

$$\sum_j \lambda_{p,j} \max(0, X_j)^2 \xrightarrow{\text{d}} \text{ID}(0, \rho_{\text{stable}}(\,\cdot\,; 1/2, c/(2\pi))) = \text{Stable}(1/2, c/2) = \text{IG}(1/2, c/8)$$

where $X_1, X_2 \ldots$, are iid $N(0,1)$ random variables. Using Proposition 4, for any $x > 0$,

$$\Pr(\max_j \lambda_{p,j} \leq x) \to e^{-(x/c)^{-1/2}}$$

which is the cdf of a Fréchet random variable with shape parameter $\alpha = 1/2$ and scale parameter c. Similarly, since $\sigma_v = 1$,

$$\Pr(\max_j W_{jk}^2 \leq x) \to e^{-(x/(2c/\pi))^{-1/2}}$$

which is the cdf of a Fréchet random variable with shape parameter $\alpha = 1/2$ and scale parameter $2c/\pi$. Equivalently,

$$\Pr(\max_j |W_{jk}| \leq y) \to e^{-\left(y/\sqrt{2c/\pi}\right)^{-1}}$$

which is the cdf of a Fréchet random variable with shape parameter $\alpha = 1$ and scale parameter $\sqrt{2c/\pi}$.

Appendix F. Additional Experiments

F.1 Stability of MoGP for Deep Networks

A common practical problem that may emerge with deep models is that of the vanishing/exploding norm of the output/gradient. In this subsection, we empirically investigate these phenomena for the ReLU activation in the MoGP context. We consider the FFNN model with the generalised BFRY variance distribution described in Section 7.1, with a fixed width of $p = 1000$ and univariate input and output. The parameters of the generalised BFRY are fixed as in Section 7.1; here, only σ_b and σ_v vary. For each depth, we simulate 500 random initialisations of the model. We investigate the stability of forward passes by looking at the empirical distribution of the norm of the output $Z^{(L+1)}(\mathbf{x})$ as the depth $L+1$ increases. For backward passes, we look at the empirical distribution of the norm of the gradient of the loss with respect to the input weights $W^{(1)}$.

Theorem 8 states that the variance of the output follows the stochastic recurrence equation

$$\Sigma^{(l)} = \sigma_b^2 + \sigma_v^2 S^{(l-1)} \Sigma^{(l-1)},$$

where $S^{(l-1)} := \sum_{j=1}^{p} \lambda_j^{(l-1)} \phi(\varepsilon_j^{(l-1)})^2$ and $\varepsilon_j^{(l-1)}$ are i.i.d standard normals. It is well known that if $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] < 0$, the random process $\Sigma^{(l)}$ is strictly stationary ergodic (Buraczewski et al., 2013). If we further assume that $\sigma_b > 0$, the forward pass is stable (non-vanishing and non-exploding), as illustrated in Figure 11. However, one can notice that, similarly to standard initialisations, a non-zero bias leads to an unstable backward pass. With $\sigma_b = 0$, taking parameters such that $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] < 0$ or $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] > 0$ is not practical for deep networks as such parameters lead respectively to vanishing or exploding outputs and gradients (see Figure 12). As illustrated in Figure 13, taking parameters such that $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] = 0$ leads to relatively stable forward and backward passes even at depth 20. However, we do observe an increased variance of the distributions. In rare runs, this may cause difficulties in training the models.

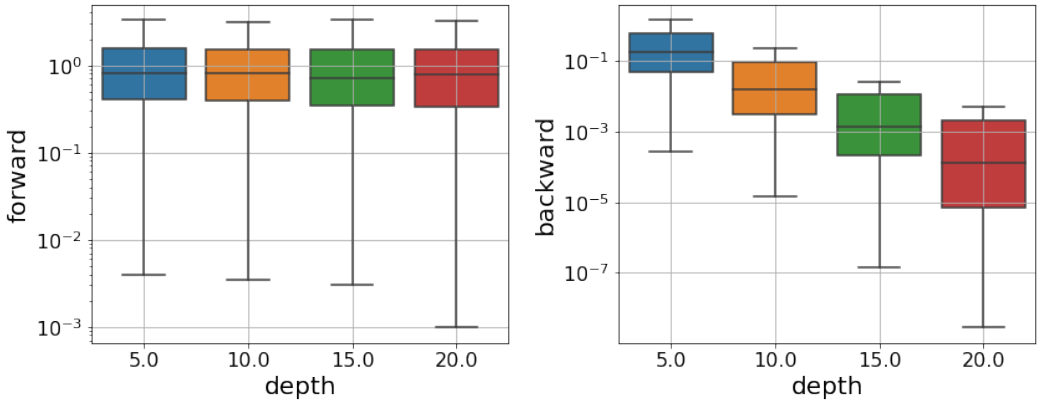
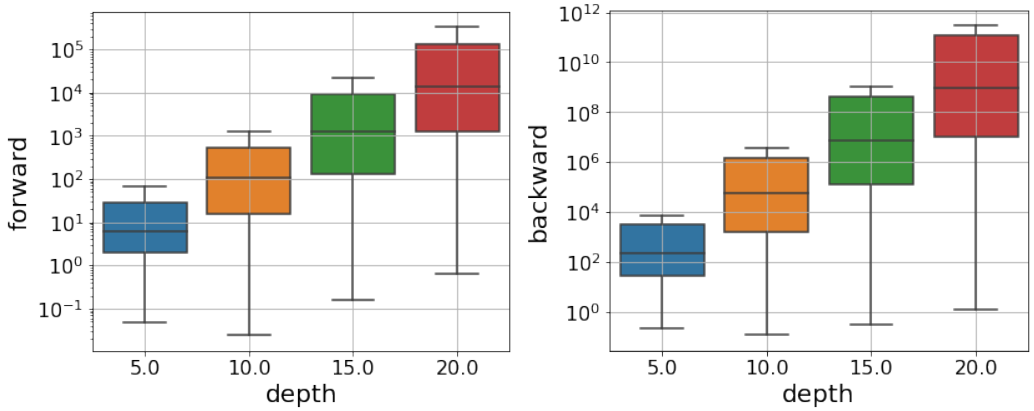


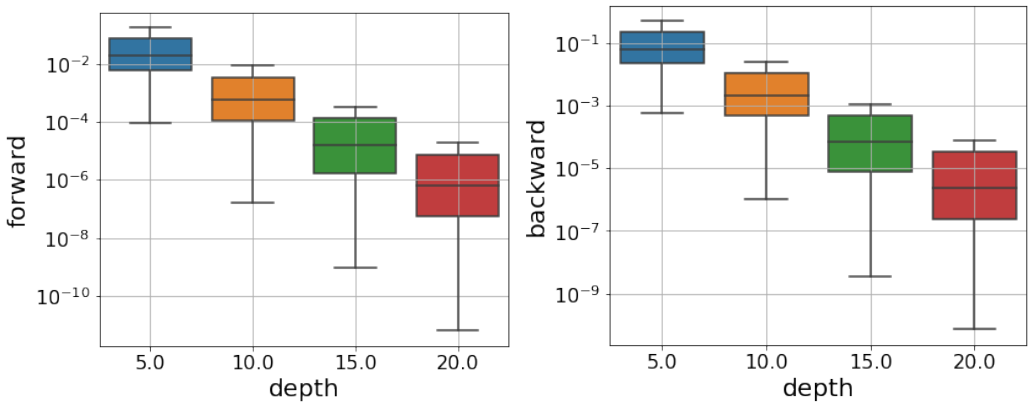
Figure 11: Stability of the forward and backward passes as the depth increases. Initialisation with non-zero bias $\sigma_b = 1$ and $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] = -1$.

F.2 Using MoGP as a Regularisation for Convolutional Layers

In this subsection, we empirically illustrate how the MoGP framework can be used with convolutional neural networks (CNN) to promote compressibility. Here we consider the more challenging dataset Cifar10. The CNN model consists of two convolutional layers (`Conv1` and `Conv2`) and a final fully-connected layer (`fc`). Each convolutional layer is followed by a max pooling layer, with kernel size



(a) $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] = 1$



(b) $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] = -1$

Figure 12: Stability of the forward and backward passes as the depth increases. Initialisation with zero bias $\sigma_b = 0$.

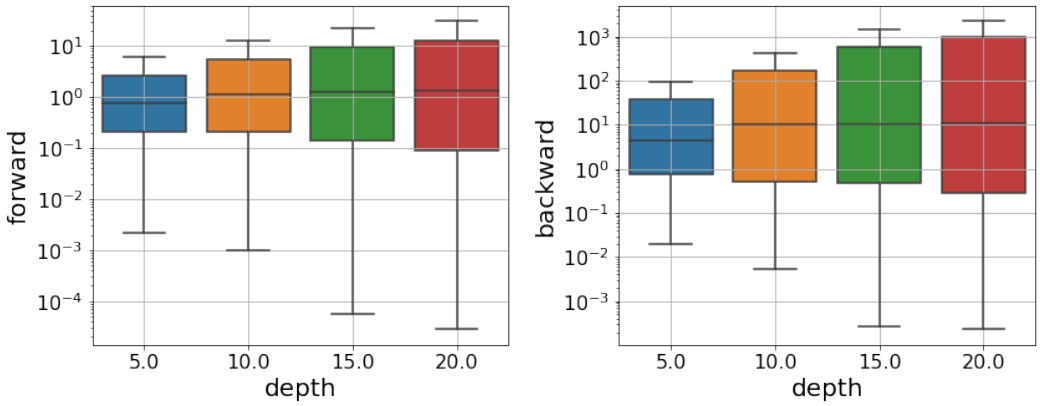


Figure 13: Stability of the forward and backward passes as the depth increases. Initialisation with $\sigma_b = 0$ and $\mathbf{E}\big[\log \sigma_v^2 S^{(l)}\big] = 0$.

2×2. The weights of each convolutional layer $l \in \{1, 2\}$ is a tensor of dimensions

$$n_f^{(l)} \times C_{in}^{(l)} \times K_x^{(l)} \times K_y^{(l)},$$

where $n_f^{(l)}$ is the number of filters of the layer, $C_{in}^{(l)}$ is the number of input channels (number of channels of the input "image") and $K_x^{(l)} \times K_y^{(l)}$ is the kernel size. For both `Conv1` and `Conv2`, we take $K_x = K_y = 3$. We note that $n_f^{(1)}$, the number of filters of `Conv1`, is equal to $C_{in}^{(2)}$ the number of input channels of `Conv2`. In this CNN setting, the input channels of a convolutional layer play the role of the input nodes in the FFNN setting. We reproduce the experiment of Section 7.2 and assign to each input channel j of `Conv2`, a scale $\lambda_j^{\texttt{Conv2}}$ and associate penalisation. Notice that if we prune a fraction $1 - \kappa$ of the input channels of `Conv2`, then only a fraction κ of the output filters of `Conv1` and the input channels of `Conv2` are active; the total number of parameters of `Conv1` and `Conv2` after compression is hence reduced to a fraction κ of the original number. We also assign a scale $\lambda_j^{\texttt{fc}}$ to each input node j of the final fully-connected layer. This enables to compress the full model and not only the convolutional layers. Both convolutional layers have 512 filters. The results are reported in Table 9 and Figure 14. We can see that the conclusion of the FFNN setting still hold in this CNN setting: using the MoGP framework as a regularisation during training leads to more compressible models. We believe that similar results would hold for more complex architectures such as ResNets, but leave this open question for future work.

Truncation (i.e., $1 - \kappa$)	Deterministic	Horseshoe	Gen. BFRY
0%	70.58 (±0.11)	69.66 (±0.12)	69.58 (±0.24)
10%	68.26 (±0.33)	69.52 (±0.20)	69.52 (±0.34)
20%	66.13 (±0.68)	68.81 (±0.28)	68.36 (±0.46)
50%	52.15 (±2.9)	60.95 (±1.8)	59.35 (±0.57)
80%	29.10 (±1.0)	44.34 (±5.3)	40.52 (±2.7)
90%	19.25 (±1.6)	32.60 (±3.46)	29.23 (±2.7)

Table 9: Classification accuracy on Cifar10 under various truncation ratio for a CNN model.

F.3 Further Results from MNIST and Fashion-MNIST

Figure 15 shows the top-5 activating images for each of the top-8 activating neurons in the deterministic, the horseshoe and the generalised BFRY cases in our full Bayesian experiments.

References

K. Adamczewski and M. Park. Dirichlet pruning for convolutional neural networks. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS'21), pages 3637–3645, 2021.

L. Aitchison. Why bigger is not always better: on finite and infinite neural networks. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), pages 156–164, 2020.

S. Arora, R. Ge, B. Neyshabur, and Y. Zhang. Stronger generalization bounds for deep nets via a compression approach. In Proceedings of the 35th International Conference on Machine Learning (ICML'18), pages 254–263, 2018.

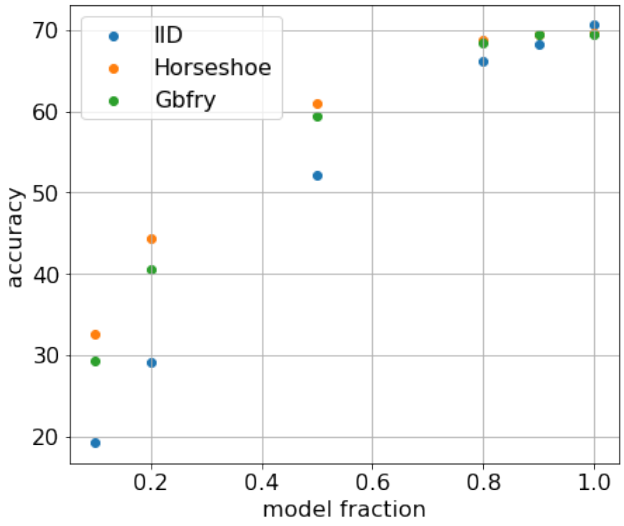


Figure 14: Classification accuracy on Cifar10 with a CNN model.

F. Ayed, J. Lee, and F. Caron. Beyond the Chinese restaurant and Pitman-Yor processes: Statistical models with double power-law behavior. In Proceedings of the 36th International Conference on Machine Learning (ICML'19), pages 395–404, 2019.

F. Ayed, J. Lee, and F. Caron. The normal-generalised gamma-Pareto process: A novel pure-jump Lévy process with flexible tail and jump-activity properties. arXiv preprint arXiv:2006.10968, 2020.

M. Barsbey, M. Sefidgaran, M. A. Erdogdu, G. Richard, and U. Simsekli. Heavy tails in SGD and compressibility of overparametrized neural networks. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS'21), pages 29364–29378, 2021.

J. Bertoin, T. Fujita, B. Roynette, and M. Yor. On a particular class of self-decomposable random variables: the durations of Bessel excursions straddling independent exponential times. 2006.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. Regular Variation. Cambridge university press, 1989.

L. Blier, P. Wolinski, and Y. Ollivier. Learning with random learning rates. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'19), pages 449–464, 2019.

D. Bracale, S. Favaro, S. Fortini, and S. Peluchetti. Large-width functional asymptotics for deep Gaussian neural networks. arXiv preprint arXiv:2102.10307, 2021.

A. Brix. Generalized gamma measures and shot-noise Cox processes. Advances in Applied Probability, 31(4):929–953, 1999.

T. Bui, D. Hernández-Lobato, J. Hernandez-Lobato, Y. Li, and R. Turner. Deep Gaussian processes for regression using approximate expectation propagation. In Proceedings of the 33rd International Conference on Machine Learning (ICML'16), pages 1472–1481, 2016.

D. Buraczewski, E. Damek, T. Mikosch, and J. Zienkiewicz. Large deviations for solutions to stochastic recurrence equations under kesten's condition. 2013.

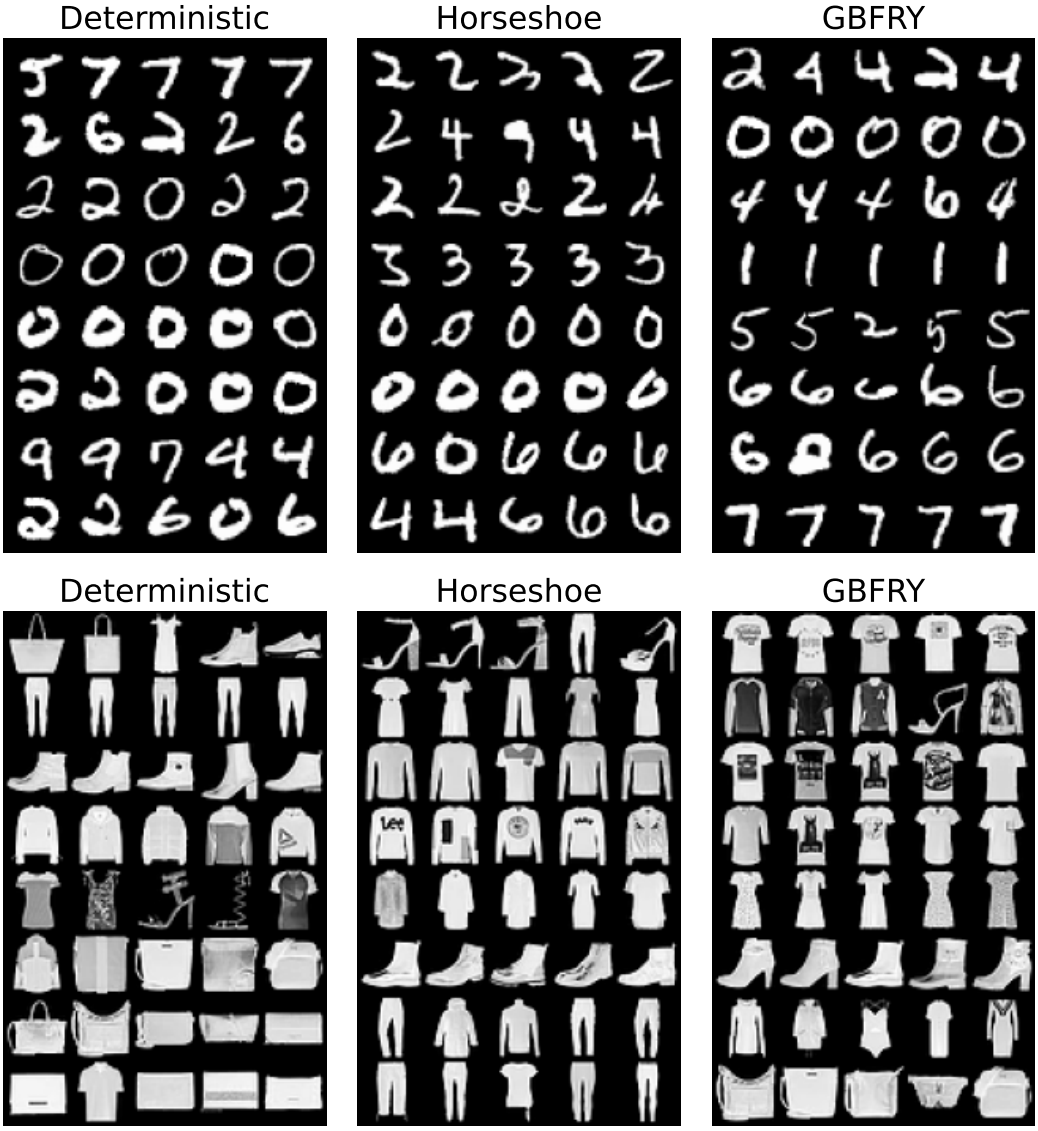


Figure 15: Top activating neurons and corresponding top activationg images. MNIST (left column) and Fashion MNIST (right column). Rows are neurons and columns are corresponding top activating images. GBFRY represents generalised BFRY.

F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In Proceedings of the 25th International Conference on Machine learning (ICML'08), pages 88–95, 2008.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Biometrika, 97(2):465–480, 2010.

G. Casella, M. Ghosh, J. Gill, and M. Kyung. Penalized regression, standard errors, and Bayesian lassos. Bayesian analysis, 5(2):369–411, 2010.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In Proceedings of the 2018 Information Theory and Applications Workshop (ITA'18), pages 1–10. IEEE, 2018.

T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte-Carlo. In Proceedings of the 31st International Conference on Machine Learning (ICML'14), pages 1683–1691, 2014.

Y. Cho and L. Saul. Kernel methods for deep learning. In Proceedings of the 23rd Conference on Neural Information Processing Systems (NeurIPS'09), pages 342–350, 2009.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS'13), pages 207–215, 2013.

Y. Davydov, I. Molchanov, and S. Zuyev. Strictly stable distributions on convex cones. Electronic Journal of Probability, 13:259–321, 2008.

S. de Jong. Compressing Neural Networks. PhD thesis, University of Cambridge, 2018.

R. Der and D. Lee. Beyond Gaussian processes: On the distributions of infinite networks. In Proceedings of the 20th Conference on Neural Information Processing Systems (NeurIPS'06), pages 275–282, 2006.

M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup. How deep are deep Gaussian processes? Journal of Machine Learning Research, 19(54):1–46, 2018.

R. Durrett. Probability: Theory and Examples, volume 49. Cambridge university press, 2019.

S. Favaro, S. Fortini, and S. Peluchetti. Stable behaviour of infinitely wide deep neural networks. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20), pages 1137–1146, 2020.

W. Feller. An Introduction to Probability Theory and Its Application Vol II. John Wiley and Sons, 1971.

V. Fortuin. Priors in Bayesian deep learning: A review. arXiv preprint arXiv:2105.06868, 2021.

V. Fortuin, A. Garriga-Alonso, F. Wenzel, G. Rätsch, R. Turner, M. van der Wilk, and L. Aitchison. Bayesian neural network priors revisited. arXiv preprint arXiv:2102.06571, 2021.

S. Ghosh, J. Yao, and F. Doshi-Velez. Structured variational learning of Bayesian neural networks with horseshoe priors. In Proceedings of the 35th International Conference on Machine Learning (ICML'18), pages 1744–1753, 2018.

S. Ghosh, J. Yao, and F. Doshi-Velez. Model selection in Bayesian neural networks via horseshoe priors. Journal of Machine Learning Research, 20(182):1–46, 2019.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS'10), pages 249–256, 2010.

R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. IEEE Transactions on Information Theory, 58(8):5016–5034, 2012.

J. E. Griffin and F. Leisen. Compound random measures and their use in Bayesian non-parametrics. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(2):525–545, 2017.

M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in SGD. In Proceedings of the 38th International Conference on Machine Learning (ICML'21), pages 3964–3975, 2021.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV'15), pages 1026–1034, 2015.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. The Annals of Statistics, pages 1259–1294, 1990.

L. Hodgkinson and M. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In Proceedings of the 38th International Conference on Machine Learning (ICML'21), pages 4262–4274, 2021.

P. Hougaard. Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73(2):387–396, 1986.

W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. arXiv preprint arXiv:1705.07562, 2017.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS'18), pages 8571–8580, 2018.

P. A. Jang, A. Loeb, M. Davidow, and A. G. Wilson. Scalable Lévy process priors for spectral kernel learning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS'17), pages 3940–3949, 2017.

S. Janson. Stable distributions. arXiv preprint arXiv:1112.0220, 2011.

S. Jantre, S. Bhattacharya, and T. Maiti. Layer adaptive node selection in Bayesian neural networks: Statistical guarantees and implementation details. arXiv preprint arXiv:2108.11000, 2021.

H. A. Jessen and T. Mikosch. Regularly varying functions. Publications de L'institut Mathematique, 80(94):171–192, 2006.

P. Jung, H. Lee, J. Lee, and H. Yang. α-stable convergence of heavy-tailed infinitely-wide neural networks. Advances in Applied Probability, 2023.

O. Kallenberg. Foundations of Modern Probability. Springer, second edition, 2002.

L. Kuhn, C. Lyle, A. N. Gomez, J. Rothfuss, and Y. Gal. Robustness to pruning predicts generalization in deep neural networks. arXiv preprint arXiv:2103.06002, 2021.

H. Lee, E. Yoon, H. Yang, and J. Lee. Scale mixtures of neural network Gaussian processes. In Proceedings of the 10th International Conference on Learning Representations (ICLR'22), 2022.

J. Lee, L. F. James, and S. Choi. Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS'16), pages 3162–3170, 2016.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

J. Lee, X. Miscouridou, and F. Caron. A unified construction for series representations and finite approximations of completely random measures. arXiv preprint arXiv:1905.10733, 2019.

R. LePage, M. Woodroofe, and J. Zinn. Convergence to a stable distribution via order statistics. The Annals of Probability, pages 624–632, 1981.

E. Littwin, O. Saremi, S. Zhai, V. Thilak, H. Goh, J. M. Susskind, and G. Yang. Implicit acceleration and feature learning in infinitely wide neural networks with bottlenecks. CoRR, abs/2107.00364, 2021.

C. Louizos, K. Ullrich, and M. Welling. Bayesian compression for deep learning. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS'17), pages 3288–3298, 2017.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In Proceedings of the 33rd International Conference on Machine Learning (ICML'16), pages 354–363, 2016.

C. H. Martin and M. W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. arXiv preprint arXiv:1901.08276, 2019.

A. G. d. G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

P. McCullagh and N. Polson. Statistical sparsity. Biometrika, 105(4):797–814, 2018.

K. Murray and D. Chiang. Auto-sizing neural networks: With applications to n-gram language models. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15), pages 908–916, 2015.

R. M. Neal. Bayesian Learning for Neural Networks. PhD thesis, Department of Computer Science, University of Toronto, 1995.

R. M. Neal. Priors for infinite networks. In Bayesian Learning for Neural Networks, pages 29–53. Springer New York, 1996.

S. W. Ober and L. Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In Proceedings of the 38th International Conference on Machine Learning (ICML'21), pages 8248–8259, 2021.

T. Ochiai, S. Matsuda, H. Watanabe, and S. Katagiri. Automatic node selection for deep neural networks using group lasso regularization. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), pages 5485–5489. IEEE, 2017.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. Probability Theory and Related Fields, 92(1):21–39, 1992.

N. Polson and J. Scott. Local shrinkage rules, Lévy processes and regularized regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(2):287–311, 2012.

A.-L. Popkes, H. Overweg, A. Ercole, Y. Li, J. M. Hernández-Lobato, Y. Zaykov, and C. Zhang. Interpretable outcome prediction with sparse Bayesian neural networks in intensive care. arXiv preprint arXiv:1905.02599, 2019.

S. Raman, T. Fuchs, P. Wild, E. Dahl, and V. Roth. The Bayesian group-lasso for analyzing contingency tables. In Proceedings of the 26th International Conference on Machine Learning (ICML'09), pages 881–888, 2009.

S. Resnick. Heavy tailed analysis eurandom summer 2005. 2005.

G. Samorodnitsky and M. Taqqu. Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance: Stochastic Modeling. Chapman & Hall, 1994.

K. Sato. Lévy Processes and Infinitely Divisible Distributions. Cambridge university press, 1999.

S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. Neurocomputing, 241:81–89, 2017.

J. Y. Shin. Compressing heavy-tailed weight matrices for non-vacuous generalization bounds. arXiv preprint arXiv:2105.11025, 2021.

T. Suzuki, H. Abe, and T. Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. arXiv preprint arXiv:1909.11274, 2019.

T. Suzuki, H. Abe, T. Murata, S. Horiuchi, K. Ito, T. Wachi, S. Hirai, M. Yukishima, and T. Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20), pages 2839–2846, 2020.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In Proceedings of the 23rd Conference on Neural Information Processing Systems (NeurIPS'09), pages 1838–1846, 2009.

R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS'07), pages 564–571, 2007.

R. Tsuchida, F. Roosta, and M. Gallagher. Richer priors for infinitely wide multi-layer perceptrons. arXiv preprint arXiv:1911.12927, 2019.

J. Wang, C. Xu, X. Yang, and J. M. Zurada. A novel pruning algorithm for smoothing feedforward neural networks based on group lasso method. IEEE Transactions on Neural Networks and Learning Systems, 29(5):2012–2024, 2017.

F. Wenzel, 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? In Proceedings of the 37th International Conference on Machine Learning (ICML'20), pages 10248–10259, 2020.

P. Wolinski, G. Charpiat, and Y. Ollivier. An equivalence between Bayesian priors and penalties in variational inference. arXiv preprint arXiv:2002.00178, 2020a.

P. Wolinski, G. Charpiat, and Y. Ollivier. Asymmetrical scaling layers for stable network pruning. 2020b.

R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels. Annals of Statistics, 39(4):1916–1962, 2011.

G. Yang. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS'19), pages 9947–9960, 2019.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67, 2006.

R. Zhang, C. Li, J. Zhang, C. Chen, and A. G. Wilson. Cyclical stochastic gradient MCMC for Bayesian deep learning. In Proceedings of the 8th International Conference on Learning Representations (ICLR'20), 2020.

Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Proceedings of the 36th International Conference on Machine Learning (ICML'19), pages 7654–7663, 2019.