Adv. Appl. Probab. **55,** 1415–1441 (2023)
*doi:*10.1017/apr.2023.3

α-STABLE CONVERGENCE OF HEAVY-/LIGHT-TAILED INFINITELY WIDE NEURAL NETWORKS

PAUL JUNG,* *Sam Houston State University*
HOIL LEE ,** *KAIST*
JIHO LEE,*** *Korea Science Academy of KAIST*
HONGSEOK YANG,**** *KAIST and Institute for Basic Science*

Abstract

We consider infinitely wide multi-layer perceptrons (MLPs) which are limits of standard deep feed-forward neural networks. We assume that, for each layer, the weights of an MLP are initialized with independent and identically distributed (i.i.d.) samples from either a light-tailed (finite-variance) or a heavy-tailed distribution in the domain of attraction of a symmetric α-stable distribution, where $\alpha \in (0, 2]$ may depend on the layer. For the bias terms of the layer, we assume i.i.d. initializations with a symmetric α-stable distribution having the same α parameter as that layer. Non-stable heavy-tailed weight distributions are important since they have been empirically seen to emerge in trained deep neural nets such as the ResNet and VGG series, and proven to naturally arise via stochastic gradient descent. The introduction of heavy-tailed weights broadens the class of priors in Bayesian neural networks. In this work we extend a recent result of Favaro, Fortini, and Peluchetti (2020) to show that the vector of pre-activation values at all nodes of a given hidden layer converges in the limit, under a suitable scaling, to a vector of i.i.d. random variables with symmetric α-stable distributions, $\alpha \in (0, 2]$.

Keywords: Heavy-tailed distribution; stable process; multi-layer perceptrons; infinite-width limit; weak convergence

2020 Mathematics Subject Classification: Primary 60F05
Secondary 60G52; 62M45

1. Introduction

Deep neural networks have brought remarkable progress in a wide range of applications, such as language translation and speech recognition, but a satisfactory mathematical answer on why they are so effective has yet to be found. One promising direction, which has been

Received 10 February 2022; revision received 18 January 2023.
* Postal address: Department of Mathematics and Statistics, 1905 University Ave, Huntsville, TX 77340, USA. Email address: phj001@shsu.edu
** Postal address: Department of Mathematical Sciences, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. Email address: hoil.lee@kaist.ac.kr
*** Postal address: Department of Mathematics and Computer Sciences, Korea Science Academy of KAIST, 105-47, Baegyanggwanmun-ro, Busanjin-gu, Busan 47162, Republic of Korea. Email address: efidiaf@gmail.com
**** Postal address: School of Computing and Kim Jaechul Graduate School of AI, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea; Discrete Mathematics Group, Institute for Basic Science, 55 Expo-ro, Yuseong-gu, Daejeon 34126, Republic of Korea. Email address: hongseok.yang@kaist.ac.kr

the subject of a large amount of recent research, is to analyze neural networks in an idealized setting where the networks have infinite widths and the so-called step size becomes infinitesimal. In this idealized setting, seemingly intractable questions can be answered. For instance, it has been shown that as the widths of deep neural networks tend to infinity, the networks converge to Gaussian processes, both before and after training, if their weights are initialized with independent and identically distributed (i.i.d.) samples from the Gaussian distribution [32, 24, 30, 33, 43]. (The methods used in these works can easily be adapted to show convergence to Gaussian processes when the initial weights are i.i.d. with finite variance.) Furthermore, in this setting, the training of a deep neural network (under the standard mean-squared loss) is shown to achieve zero training error, and an analytic form of a fully-trained network with zero error has been identified [17, 26]. These results, in turn, enable the use of tools from stochastic processes and differential equations to analyze deep neural networks in a novel way. They have also led to new high-performing data-analysis algorithms based on Gaussian processes [25].

One direction extending this line of research is to consider neural networks with possibly heavy-tailed initializations. Although these are not common, their potential for modeling heavy-tailed data was recognized early on by [41], and even the convergence of an infinitely wide yet shallow neural network under non-Gaussian α-stable initialization was shown in the 1990s [32]. Recently, Favaro, Fortini, and Peluchetti extended such convergence results from shallow to deep networks [4].

Favaro *et al.* [4] considered multi-layer perceptrons (MLPs) having large width n, and having i.i.d. weights with a symmetric α-stable (SαS) distribution of scale parameter σ_w. A random variable X is said to have an SαS distribution if its characteristic function takes the form, for $0 < \alpha \leq 2$,

$$\psi_X(t) := \mathbb{E}e^{itX} = e^{-|\sigma t|^\alpha},$$

for some constant $\sigma > 0$ called the scale parameter. In the special case $\alpha = 2$, X has a Gaussian distribution with variance $2\sigma^2$ (which differs from standard notation in this case, by a factor of 2).

The results of Favaro *et al.* [4] show that as n tends to ∞, the arguments of the nonlinear activation function ϕ, in any given hidden layer, converge jointly in distribution to a product of SαS(σ_ℓ) distributions with the same α parameter. The scale parameter σ_ℓ differs for each layer ℓ; however, an explicit form is provided as a function of σ_w, the input $\mathbf{x} = (x_1, \ldots, x_I)$, and the distribution of bias terms which have an SαS(σ_B) distribution for some $\sigma_B > 0$. Favaro *et al.* also show that as a function of $\mathbf{x}$, the joint distribution described above is an α-stable process, and they describe the spectral measure (see [38, Section 2.3]) of this process at the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Our work is a further extension of the work of [4]. We consider deep networks whose weights in a given layer are allowed to be initialized with i.i.d. samples from *either* a light-tailed (finite-variance) or heavy-tailed distribution, not necessarily stable, but in the domain of attraction of an SαS distribution. We show that as the widths of the networks increase, the networks at initialization converge to SαS processes.

One of our aims is to show universality, in the sense that the results also hold when the weights are i.i.d. and heavy-tailed, and in the domain of attraction of an SαS distribution. Such heavy-tailed (and non-stable) weight distributions are important in the context of deep neural networks, since they have been empirically seen to emerge from trained deep neural networks such as the ResNet and VGG series [28, 29] and have been shown to arise naturally via stochastic gradient descent [14, 16]. Also, such heavy-tailed distributions cover a wide

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

range of distributions, including for example some Pareto, inverse gamma, Fréchet, Student t, horseshoe, and beta-prime distributions. In particular, both Student t and horseshoe priors have been used for weights in Bayesian neural networks [8], since heavy tails can potentially improve the performance of priors [9]. Another of our goals is to fill a (minor) gap regarding one nontrivial step, and to clarify other details, of the proof in [4] and its companion paper [5] (see Lemma 3.1 below). Finally, we also generalize by considering a slightly more general case where the α parameter for the weights may depend on the layer it is in, including the case where it may be that $\alpha = 2$ for some layers. This provides, for instance, a proof of universality in the Gaussian case. Such a result for the non-Gaussian finite-variance weights is known in the 'folklore', but we are unaware of a published proof of it.

Notation. Let $\Pr(\mathbb{R})$ be the set of probability distributions on $\mathbb{R}$. In the sequel, for $\alpha \in (0, 2]$, let $\mu_{\alpha,\sigma} \in \Pr(\mathbb{R})$ denote an $S\alpha S(\sigma)$ distribution. We will typically use capital letters to denote random variables in $\mathbb{R}$. For example, a random weight in our neural network from layer $\ell - 1$ to layer ℓ is denoted by $W_{ij}^{(\ell)}$ and is henceforth assumed to be in the domain of attraction of $\mu_{\alpha,\sigma}$, which may depend on ℓ. One notable exception to this convention is our use of the capital letter L to denote a slowly varying function. We use the notation $|\cdot|^{\alpha\pm\epsilon}$ to denote the maximum of $|\cdot|^{\alpha+\epsilon}$ and $|\cdot|^{\alpha-\epsilon}$.

2. The model: heavy-tailed multi-layer perceptrons

At a high level, a neural network is just a parameterized function Y from inputs in $\mathbb{R}^I$ to outputs in $\mathbb{R}^O$ for some I and O. In this article, we consider the case that $O = 1$. The parameters Θ of the function consist of real-valued vectors $\mathbf{W}$ and $\mathbf{B}$, called weights and biases. These parameters are initialized randomly, and get updated repeatedly during the training of the network. We adopt the common notation $Y_\Theta(\mathbf{x})$, which expresses that the output of Y depends on both the input $\mathbf{x}$ and the parameters $\Theta = (\mathbf{W}, \mathbf{B})$.

Note that since Θ is set randomly, Y_Θ is a random function. This random-function viewpoint is the basis of a large body of work on Bayesian neural networks [32], which studies the distribution of this random function or its posterior conditioned on input–output pairs in training data. Our article falls into this body of work. We analyze the distribution of the random function Y_Θ at the moment of initialization. Our analysis is in the situation where Y_Θ is defined by an MLP, the width of the MLP is large (so the number of parameters in Θ is large), and the parameters Θ are initialized by possibly using heavy-tailed distributions. The precise description of the setup is given below.

2.1 (Layers.) We suppose that there are ℓ_{lay} layers, not including those for the input and output. Here, the subscript lay means 'layers'. The 0th layer is for the input and consists of I nodes assigned with deterministic values from the input $\mathbf{x} = (x_1, \ldots, x_I)$. We assume for simplicity that $x_i \in \mathbb{R}$. (None of our methods would change if we instead let $x_i \in \mathbb{R}^d$ for arbitrary finite d.) The layer $\ell_{\text{lay}} + 1$ is for the output. For layer ℓ with $1 \le \ell \le \ell_{\text{lay}}$, there are n_ℓ nodes for some $n_\ell \ge 2$.

2.2 (Weights and biases.) The MLP is fully connected, and the weights on the edges from layer $\ell - 1$ to ℓ are given by $\mathbf{W}^{(\ell)} = (W_{ij}^{(\ell)})_{1\le i\le n_\ell, 1\le j\le n_{\ell-1}}$. Assume that $\mathbf{W}^{(\ell)}$ is a collection of i.i.d. symmetric random variables in each layer, such that for each layer ℓ, (2.2.a) they are heavy-tailed, i.e. for all $t > 0$,

$$\mathbb{P}(|W_{ij}^{(\ell)}| > t) = t^{-\alpha_\ell} L^{(\ell)}(t), \qquad \text{for some } \alpha_\ell \in (0, 2], \tag{2.1}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

where $L^{(\ell)}$ is some slowly varying function, or
(2.2.b) $\mathbb{E}|W_{ij}^{(\ell)}|^2 < \infty$. (In this case, we set $\alpha_\ell = 2$ by default.)

Note that both (2.2.a) and (2.2.b) can hold at the same time. Even when this happens, there is no ambiguity about α_ℓ, which is set to be 2 in both cases. Our proof deals with the cases when $\alpha_\ell < 2$ and $\alpha_\ell = 2$ separately. (See below, in the definition of L_0.) We permit both the conditions (2.2.a) and (2.2.b) to emphasize that our result covers a mixture of both heavy-tailed and finite-variance (light-tailed) initializations.

Let $B_i^{(\ell)}$ be i.i.d. random variables with distribution $\mu_{\alpha_\ell, \sigma_{B^{(\ell)}}}$. Note that the distribution of $B_i^{(\ell)}$ is more constrained than that of $W_{ij}^{(\ell)}$. This is because the biases are not part of the normalized sum, and normalization is, of course, a crucial part of the stable limit theorem.

For later use in the $\alpha = 2$ case, we define a function $\widetilde{L}^{(\ell)}$ by

$$\widetilde{L}^{(\ell)}(x) := \int_0^x y\mathbb{P}(|W_{ij}^{(\ell)}| > y)\, dy.$$

Note that $\widetilde{L}^{(\ell)}$ is increasing. For the case (2.2.b), $\int_0^x y\mathbb{P}(|W_{ij}^{(\ell)}| > y)\, dy$ converges to a constant, namely to $1/2$ of the variance, and thus it is slowly varying. For the case (2.2.a), it is seen in Lemma A.1 that $\widetilde{L}^{(\ell)}$ is slowly varying as well.

For convenience, let

$$L_0 := \begin{cases} L^{(\ell)} & \text{if } \alpha_\ell < 2, \\ \widetilde{L}^{(\ell)} & \text{if } \alpha_\ell = 2. \end{cases}$$

We have dropped the superscript ℓ from L_0 as the dependence on ℓ will be assumed.

2.3 (Scaling.) Fix a layer ℓ with $2 \le \ell \le \ell_{\text{lay}} + 1$, and let $n = n_{\ell-1}$ be the number of nodes at the layer $\ell - 1$. We will scale the random values at the nodes (pre-activation) by

$$a_n(\ell) := \inf\{t > 0 \colon t^{-\alpha_\ell} L_0(t) \le n^{-1}\}.$$

Then $a_n(\ell)$ tends to ∞ as n increases. One can check that $a_n(\ell) = n^{1/\alpha_\ell} G(n)$ for some slowly varying function G. If we consider, for example, power-law weights where $\mathbb{P}(|W_{ij}^{(\ell)}| > t) = t^{-\alpha_\ell}$ for $t \ge 1$, then $a_n(\ell) = n^{1/\alpha_\ell}$. For future purposes we record the well-known fact that, for $a_n = a_n(\ell)$,

$$\lim_{n\to\infty} n a_n^{-\alpha_\ell} L_0(a_n) = 1. \tag{2.2}$$

Let us quickly show (2.2). For the case (2.2.b), $t^2 L_0(t)$ becomes continuous and so $n a_n^{-\alpha_\ell} L_0(a_n)$ is simply 1. To see the convergence in the case (2.2.a), first note that as $\mathbb{P}(|W_{ij}^{(\ell)}| > t) = t^{-\alpha_\ell} L^{(\ell)}(t)$ is right-continuous, $n a_n^{-\alpha_\ell} L^{(\ell)}(a_n) \le 1$. For the reverse inequality, note that by (2.1) and the definition of a_n, for n large enough we have $\mathbb{P}\left(|W_{ij}^{(\ell)}| > \frac{1}{1+\epsilon} a_n\right) \ge 1/n$, and by the definition of a slowly varying function we have that

$$(1+2\epsilon)^{-\alpha_\ell} = \lim_{n\to\infty} \frac{\mathbb{P}\left(|W_{ij}^{(\ell)}| > \frac{1+2\epsilon}{1+\epsilon} a_n\right)}{\mathbb{P}\left(|W_{ij}^{(\ell)}| > \frac{1}{1+\epsilon} a_n\right)} \le \liminf_{n\to\infty} \frac{\mathbb{P}\left(|W_{ij}^{(\ell)}| > a_n\right)}{1/n}.$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

2.4 (Activation.) The MLP uses a nonlinear activation function $\phi(y)$. We assume that ϕ is continuous and bounded. The boundedness assumption simplifies our presentation; in Section 4 we relax this assumption so that for particular initializations (such as Gaussian or stable), more general activation functions such as ReLU are allowed.

2.5 (Limits.) We consider one MLP for each $(n_1, \dots, n_{\ell_{\text{lay}}}) \in \mathbb{N}^{\ell_{\text{lay}}}$. We take the limit of the collection of these MLPs in such a way that

$$\min(n_1, \dots, n_{\ell_{\text{lay}}}) \to \infty. \tag{2.3}$$

(Our methods can also handle the case where limits are taken from left to right, i.e., $\lim_{n_{\ell_{\text{lay}}} \to \infty} \cdots \lim_{n_1 \to \infty}$, but since this order of limits is easier to prove, we will focus on the former.)

2.6 (Hidden layers.) We write $\mathbf{n} = (n_1, \dots, n_{\ell_{\text{lay}}}) \in \mathbb{N}^{\ell_{\text{lay}}}$. For ℓ with $1 \le \ell \le \ell_{\text{lay}} + 1$, the pre-activation values at these nodes are given, for an input $\mathbf{x} \in \mathbb{R}^I$, recursively by

$$Y_i^{(1)}(\mathbf{x};\, \mathbf{n}) := Y_i^{(1)}(\mathbf{x}) := \sum_{j=1}^{I} W_{ij}^{(1)} x_j + B_i^{(1)},$$

$$Y_i^{(\ell)}(\mathbf{x};\, \mathbf{n}) := \frac{1}{a_{n_{\ell-1}}(\ell)} \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \phi(Y_j^{(\ell-1)}(\mathbf{x};\, \mathbf{n})) + B_i^{(\ell)}, \quad \ell \ge 2,$$

for each $n_{\ell-1} \in \mathbb{N}$ and $i \in \mathbb{N}$. Note that $Y_i^{(\ell)}(\mathbf{x};\, \mathbf{n})$ depends on only the coordinates $n_1, \dots, n_{\ell-1}$, but we may simply let it be constant in the coordinates $n_\ell, \dots, n_{\ell_{\text{lay}}}$. This will often be the case when we have functions of $\mathbf{n}$ in the sequel.

We often omit $\mathbf{n}$ and write $Y_i^{(\ell)}(\mathbf{x})$. When computing the output of the MLP with widths $\mathbf{n}$, one only needs to consider $i \le n_\ell$ for each layer ℓ. However, it is always possible to assign values to an extended MLP beyond $\mathbf{n}$, which is why we have assumed more generally that $i \in \mathbb{N}$. This will be important for the proofs, as we explain in the next paragraph.

Extending finite neural networks to infinite neural networks

Let us describe a useful construct for the proofs which allows us to leverage the natural exchangeability present in the model. For each $\mathbf{n} = (n_1, \dots, n_{\ell_{\text{lay}}})$, the MLP is finite and each layer has finite width. A key part of the proof is the application of de Finetti's theorem at each layer, which applies only in the case where one has an infinite sequence of random variables. As in [4], a crucial observation is that for each $\mathbf{n} = (n_1, \dots, n_{\ell_{\text{lay}}})$, we can extend the MLP to an infinite-width MLP by adding an infinite number of nodes at each layer that compute values in the same manner as nodes of the original MLP, but are ignored by nodes at the next layer. Thus, the finite-width MLP is embedded in an infinite-width MLP. This allows us to use de Finetti's theorem. With this in mind we will henceforth consider an infinite collection of weights $(W_{ij}^{(\ell)})_{ij \in \mathbb{N}^2}$, for any finite neural network.

3. Convergence to α-stable distributions

Our main results are summarized in the next theorem and its extension to the situation of multiple inputs in Theorem 5.1 in Section 5. They show that as the width of an MLP tends to infinity, the MLP becomes a relatively simple random object: the outputs of its ℓth layer become merely i.i.d. random variables drawn from a stable distribution, and the parameters of the distribution have explicit inductive characterizations.

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Let

$$c_\alpha := \lim_{M\to\infty} \int_0^M \frac{\sin u}{u^\alpha}\, du \quad \text{for } \alpha < 2 \quad \text{and} \quad c_2 = 1,$$

and let

$$\sigma_2^{\alpha_2} := (\sigma_2(\mathbf{x}))^{\alpha_2} := \sigma_{B^{(2)}}^{\alpha_2} + c_{\alpha_2} \int |\phi(y)|^{\alpha_2}\, \nu^{(1)}(dy), \quad \ell = 2, \tag{3.1}$$

$$\sigma_\ell^{\alpha_\ell} := (\sigma_\ell(\mathbf{x}))^{\alpha_\ell} := \sigma_{B^{(\ell)}}^{\alpha_\ell} + c_{\alpha_\ell} \int |\phi(y)|^{\alpha_\ell}\, \mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy), \quad \ell = 3, \dots, \ell_{\text{lay}} + 1,$$

where $\nu^{(1)} := \nu^{(1)}(\mathbf{x})$ is the distribution of $Y_1^{(1)}(\mathbf{x})$.

Theorem 3.1. *For each $\ell = 2, \dots, \ell_{\text{lay}} + 1$, the joint distribution of $(Y_i^{(\ell)}(\mathbf{x};\, \mathbf{n}))_{i\geq 1}$ converges weakly to $\bigotimes_{i\geq 1} \mu_{\alpha_\ell,\sigma_\ell}$ as $\min(n_1, \dots, n_{\ell_{\text{lay}}}) \to \infty$, with σ_ℓ inductively defined by (3.1). That is, the characteristic function of the limiting distribution is, for any finite subset $\mathcal{L} \subset \mathbb{N}$,*

$$\prod_{i\in\mathcal{L}} \psi_{B^{(2)}}(t_i) \exp\left(-c_{\alpha_2}|t_i|^{\alpha_2} \int |\phi(y)|^{\alpha_2}\, \nu^{(1)}(dy)\right), \quad \ell = 2,$$

$$\prod_{i\in\mathcal{L}} \psi_{B^{(\ell)}}(t_i) \exp\left(-c_{\alpha_\ell}|t_i|^{\alpha_\ell} \int |\phi(y)|^{\alpha_\ell}\, \mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right), \quad \ell = 3, \dots, \ell_{\text{lay}} + 1.$$

Remark 3.1. The integrals in Theorem 3.1 are well-defined since ϕ is bounded. For (possibly) unbounded ϕ, these integrals are well-defined as well under suitable assumptions on ϕ. See Section 4.

This theorem shows that, for a given data point $\mathbf{x}$, the individual layers of our MLP converge in distribution to a collection of i.i.d. stable random variables. The result is a universality counterpart to a similar result in [4] where, instead of general heavy-tailed weights on edges, one initializes precisely with stable weights. As already mentioned in the introduction, heavy-tailed initializations other than α-stable have been considered and discussed in previous literature. Later, in Theorem 5.1, we generalize this result to consider multiple data points $\mathbf{x}_1$, $\mathbf{x}_2, \dots, \mathbf{x}_k$.

Heuristic of the proof. The random variables $(Y_i^{(\ell)}(\mathbf{x};\, \mathbf{n}))_{i\in\mathbb{N}}$ are dependent only through the randomness of the former layer's outputs $(Y_j^{(\ell-1)}(\mathbf{x};\, \mathbf{n}))_{j\in\mathbb{N}}$. Just as in proofs in the literature for similar models, as the width grows to infinity, this dependence vanishes via an averaging effect.

Here, we briefly summarize the overarching technical points, from a bird's-eye view, in establishing this vanishing dependence; we also highlight what we believe are new technical contributions in relation to models with general heavy-tailed initializations.

By de Finetti's theorem, for each $\mathbf{n}$ there exists a random distribution $\xi^{(\ell-1)}(dy;\, \mathbf{n})$ such that the sequence $(Y_j^{(\ell-1)}(\mathbf{x}))_j$ is conditionally i.i.d. with common random distribution $\xi^{(\ell-1)}$. By first conditioning on $(Y_j^{(\ell-1)}(\mathbf{x}))_j$, we obtain independence among the summands of

$$Y_i^{(\ell)}(\mathbf{x}) = \frac{1}{a_{n_{\ell-1}}(\ell)} \sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)} \phi(Y_j^{(\ell-1)}(\mathbf{x})) + B_i^{(\ell)}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

as well as independence among the family $(Y_i^{(\ell)}(\mathbf{x}))_i$. Let $\alpha := \alpha_\ell$, $n := n_{\ell-1}$, and $a_n := a_{n_{\ell-1}}(\ell)$. Then, with the help of Lemma A.2, the conditional characteristic function of $Y_1^{(\ell)}(\mathbf{x})$ given $\xi^{(\ell-1)}$ is asymptotically equal to

$$e^{-\sigma_B^\alpha |t|^\alpha} \left(1 - \frac{b_n}{n} c_\alpha |t|^\alpha \int |\phi(y)|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)} \xi^{(\ell-1)}(dy;\, \mathbf{n})\right)^n, \tag{3.2}$$

where b_n is a deterministic constant that tends to 1. Assuming the inductive hypothesis, the random distribution $\xi^{(\ell-1)}$ converges weakly to $\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}$ as $\mathbf{n} \to \infty$ in the sense of (2.3), by Lemma 3.1 below. This lemma is intuitively obvious, but we have not seen it proved in any previous literature.

Next, since L_0 is slowly varying, one can surmise that the conditional characteristic function tends to

$$\exp\left(-\sigma_B^\alpha |t|^\alpha - c_\alpha |t|^\alpha \int |\phi(y)|^\alpha \mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right),$$

which is the characteristic function of the stable law we desire. Making the above intuition rigorous involves additional technicalities in the setting of general heavy-tailed weights: namely, we verify the convergence of (3.2) by proving uniform integrability of the integrand

$$|\phi(y)|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}$$

with respect to the family of distributions $\xi^{(\ell-1)}$ over the indices $\mathbf{n}$. In particular, by Lemma A.4, the integrand can be bounded by $O(|\phi(y)|^{\alpha\pm\epsilon})$ for small $\epsilon > 0$, and uniform integrability follows from the boundedness of ϕ. The joint limiting distribution converges to the desired stable law by similar arguments, which completes our top-level heuristic proof.

Before delving into the actual technical proof, we next present a key lemma mentioned in the above heuristic. Recall that de Finetti's theorem tells us that if a sequence $\mathbf{X} = (X_i)_{i\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ is exchangeable, then

$$\mathbb{P}(\mathbf{X} \in A) = \int \nu^{\otimes\mathbb{N}}(A)\, \pi(d\nu) \tag{3.3}$$

for some π which is a probability measure on the space of probability measures $\Pr(\mathbb{R})$. The measure π is sometimes called the *mixing measure*. The following lemma characterizes the convergence of exchangeable sequences by the convergence of their respective mixing measures. While intuitively clear, the proof of the lemma is not completely trivial.

Lemma 3.1. *For each $j \in \mathbb{N} \cup \{\infty\}$, let $\mathbf{X}^{(j)} = (X_i^{(j)})_{i\in\mathbb{N}}$ be an infinite exchangeable sequence of random variables with values in $\mathbb{R}$ (or more generally, a Borel space). Let π_j be the mixing measure on $\Pr(\mathbb{R})$ corresponding to $\mathbf{X}^{(j)}$, from (3.3). Then the family $(\mathbf{X}^{(j)})_{j\in\mathbb{N}}$ converges in distribution to $\mathbf{X}^{(\infty)}$ if and only if the family $(\pi_j)_{j\in\mathbb{N}}$ converges in the weak topology on $\Pr(\Pr(\mathbb{R}))$ to π_∞.*

The proof of the lemma is in the appendix. In the lemma, the topology on $\Pr(\Pr(\mathbb{R}))$ is formed by applying the weak-topology construction twice. We first construct the weak topology on $\Pr(\mathbb{R})$. Then we apply the weak-topology construction again, this time using $\Pr(\mathbb{R})$ instead of $\mathbb{R}$.

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

In the proof of Theorem 3.1, we use the special case when the limiting sequence $\mathbf{X}^{(\infty)}$ is a sequence of i.i.d. random variables. In that case, by (3.3), it must be that π_∞ concentrates on a single element $\nu \in \Pr(\mathbb{R})$, i.e. it is a point mass, $\pi_\infty = \delta_\nu$, for some $\nu \in \Pr(\mathbb{R})$.

More specifically, we use the following corollary to Lemma.

Corollary 3.1. *In the setting of Theorem 3.1 , the joint distribution of the exchangeable sequence* $(Y_i^{(\ell-1)}(\mathbf{x}))_{i\geq 1}$ *converges weakly to the product measure* $\bigotimes_{i\geq 1}\mu_{\alpha,\sigma_{\ell-1}}$ *as the minimum of* $n_1, \ldots, n_{\ell_{\text{lay}}}$ *tends to* ∞ *if and only if the random probability measures* $(\xi^{(\ell-1)}(dy, \omega; \mathbf{n}))_{\mathbf{n}\in\mathbb{N}^{\ell_{\text{lay}}}}$ *defined in (3.8) converge weakly, in probability, to the deterministic probability measure* $\mu_{\alpha,\sigma_{\ell-1}}$.

Proof of Theorem 3.1. We start with a useful expression for the characteristic function conditioned on the random variables $\{Y_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\ldots,n_{\ell-1}}$:

$$\begin{aligned}\psi_{Y_i^{(\ell)}(\mathbf{x})|\{Y_j^{(\ell-1)}(\mathbf{x})\}_j}(t) &:= \mathbb{E}\left[\exp\left(itY_i^{(\ell)}(\mathbf{x})\right)\Big|\{Y_j^{(\ell-1)}(\mathbf{x})\}_j\right] \\ &= \mathbb{E}\left[\exp\left(it\left\{\frac{1}{a_{n_{\ell-1}}(\ell)}\sum_{j=1}^{n_{\ell-1}} W_{ij}^{(\ell)}\phi(Y_j^{(\ell-1)}(\mathbf{x})) + B_i^{(\ell)}\right\}\right)\Big|\{Y_j^{(\ell-1)}(\mathbf{x})\}_j\right] \\ &= e^{-\sigma|t|^{\alpha_\ell}}\prod_{j=1}^{n_{\ell-1}}\psi_{W_{ij}^{(\ell)}}\left(\frac{\phi(Y_j^{(\ell-1)}(\mathbf{x}))}{a_{n_{\ell-1}}(\ell)}t\right),\end{aligned} \tag{3.4}$$

where $\sigma := \sigma_{B^{(\ell)}}^{\alpha_\ell}$ and the argument on the right-hand side is random.
Case $\ell = 2$:

Let us first consider the case $\ell = 2$. Let $n = n_1$, $\alpha = \alpha_2$, $a_n = a_{n_1}(2)$, and $t \neq 0$. We first show the weak convergence of the one-point marginal distributions; i.e., we show that the distribution of $Y_i^{(2)}(\mathbf{x})$ converges weakly to $\mu_{\alpha,\sigma}$ for each i. Since $Y_j^{(1)}(\mathbf{x})$, $j = 1, \ldots, n$, are i.i.d., this is a straightforward application of standard arguments, which we include for completeness. Denote the common distribution of $Y_j^{(1)}(\mathbf{x})$, $j = 1, \ldots, n$, by $\nu^{(1)}$. Taking the expectation of (3.4) with respect to the randomness of $\{Y_j^{(1)}(\mathbf{x})\}_{j=1,\ldots,n}$, we have

$$\psi_{Y_i^{(2)}(\mathbf{x})}(t) = e^{-\sigma_{B^{(2)}}^{\alpha}|t|^\alpha}\left(\int \psi_W\left(\frac{\phi(y)}{a_n}t\right)\nu^{(1)}(dy)\right)^n,$$

where $\psi_W := \psi_{W_{ij}^{(2)}}$ for some/any i,j. From Lemma A.2, we have that

$$\psi_W(t) = 1 - c_\alpha|t|^\alpha L_0\left(\frac{1}{|t|}\right) + o\left(|t|^\alpha L_0\left(\frac{1}{|t|}\right)\right), \quad |t| \to 0,$$

for $c_\alpha = \lim_{M\to\infty}\int_0^M \sin u/u^\alpha\, du$ when $\alpha < 2$ and $c_2 = 1$. If $\phi(y) = 0$ then $\psi_W\left(\frac{\phi(y)}{a_n}t\right) = 1$. Otherwise, setting $b_n := na_n^{-\alpha}L_0(a_n)$, for fixed y with $\phi(y) \neq 0$ we have that, as $n \to \infty$,

$$\psi_W\left(\frac{\phi(y)}{a_n}t\right) = 1 - c_\alpha\frac{b_n}{n}|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)} + o\left(\frac{b_n}{n}|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\right). \tag{3.5}$$

By Lemma A.4 applied to $G(x) := x^{-\alpha}L_0(x)$ and $c = 1$, for any $\epsilon > 0$, there exist constants $b > 0$ and n_0 such that for all $n > n_0$ and all y with $\phi(y) \neq 0$,

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

$$|\phi(y)t|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)} = \frac{G\left(\frac{a_n}{|\phi(y)t|}\right)}{G(a_n)} \le b|\phi(y)t|^{\alpha\pm\epsilon}. \tag{3.6}$$

Since ϕ is bounded, the right-hand side of (3.5) is term-by-term integrable with respect to $\nu^{(1)}(dy)$. In particular, the integral of the error term can be bounded, for some small ϵ and large enough n, by

$$\int o\left(\frac{b_n}{n}|\phi(y)t|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\right) \nu^{(1)}(dy) \le o\left(b\frac{b_n}{n}\int |\phi(y)t|^{\alpha\pm\epsilon}\, \nu^{(1)}(dy)\right) = o\left(\frac{b_n}{n}\right).$$

(Set $|\phi(y)|^\alpha L_0(\frac{a_n}{|\phi(y)|}) = 0$ when $\phi(y) = 0$.) Thus, integrating both sides of (3.5) with respect to $\nu^{(1)}(dy)$ and taking the nth power, it follows that

$$\left(\int \psi_W\left(\frac{\phi(y)t}{a_n}\right) \nu^{(1)}(dy)\right)^n = \left(1 - c_\alpha \frac{b_n}{n}\int |\phi(y)t|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\, \nu^{(1)}(dy) + o\left(\frac{b_n}{n}\right)\right)^n.$$

From the bound in (3.6), we have, by dominated convergence, that as $n \to \infty$

$$\int |\phi(y)t|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\, \nu^{(1)}(dy) \to |t|^\alpha \int |\phi(y)|^\alpha\, \nu^{(1)}(dy).$$

Since $b_n = na_n^{-\alpha}L_0(a_n)$ converges to 1 by (2.2), we have that

$$\left(\int \psi_W\left(\frac{\phi(y)t}{a_n}\right) \nu^{(1)}(dy)\right)^n \to \exp\left(-c_\alpha|t|^\alpha \int |\phi(y)|^\alpha\, \nu^{(1)}(dy)\right).$$

Thus, the distribution of $Y_i^{(2)}(\mathbf{x})$ weakly converges to μ_{α,σ_2} where

$$\sigma_2^\alpha = \sigma_{B^{(2)}}^\alpha + c_\alpha \int |\phi(y)|^\alpha\, \nu^{(1)}(dy),$$

as desired.

Next we prove that the joint distribution of $(Y_i^{(2)}(\mathbf{x}))_{i\ge1}$ converges to the product distribution $\bigotimes_{i\ge1}\mu_{\alpha,\sigma_2}$. Let $\mathcal{L} \subset \mathbb{N}$ be a finite set. Let ψ_B denote the multivariate characteristic function for the $|\mathcal{L}|$-fold product distribution of $\mu_{\alpha,\sigma_{B^{(2)}}}$. For $\mathbf{t} = (t_i)_{i\in\mathcal{L}}$, conditionally on $\{Y_j^{(1)}(\mathbf{x})\}_{j=1,\dots,n}$,

$$\begin{aligned}
&\psi_{(Y_i^{(2)}(\mathbf{x}))_{i\in\mathcal{L}}|\{Y_j^{(1)}(\mathbf{x})\}_j}(\mathbf{t}) \qquad (3.7)\\
&:= \mathbb{E}\left[\exp\left(i\sum_{i\in\mathcal{L}} t_i Y_i^{(2)}(\mathbf{x})\right)\middle|\, \{Y_j^{(1)}(\mathbf{x})\}_j\right]\\
&= \mathbb{E}\left[\exp\left(i\sum_{i\in\mathcal{L}} B_i^{(2)}t_i\right)\right]\mathbb{E}\left[\exp\left(i\frac{1}{a_n}\sum_{j=1}^n\sum_{i\in\mathcal{L}} W_{ij}^{(2)}\phi(Y_j^{(1)}(\mathbf{x}))t_i\right)\middle|\, \{Y_j^{(1)}(\mathbf{x})\}_j\right]\\
&= \psi_B(\mathbf{t})\prod_{j=1}^n\prod_{i\in\mathcal{L}}\mathbb{E}\left[\exp\left(i\frac{1}{a_n}W_{ij}^{(2)}\phi(Y_j^{(1)}(\mathbf{x}))t_i\right)\middle|\, \{Y_j^{(1)}(\mathbf{x})\}_j\right]\\
&= \psi_B(\mathbf{t})\prod_{j=1}^n\prod_{i\in\mathcal{L}}\psi_W\left(\frac{\phi(Y_j^{(1)}(\mathbf{x}))t_i}{a_n}\right).
\end{aligned}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Taking the expectation over the randomness of $\{Y_j^{(1)}(\mathbf{x})\}_{j=1,\dots,n}$, we have

$$\begin{aligned}\frac{\psi_{(Y_i^{(2)}(\mathbf{x}))_{i\in\mathcal{L}}}(\mathbf{t})}{\psi_B(\mathbf{t})} &= \int \prod_{j=1}^{n}\prod_{i\in\mathcal{L}} \psi_W\left(\frac{\phi(y_j)t_i}{a_n}\right) \bigotimes_{j=1}^{n} \nu^{(1)}(dy_j)\\ &= \left(\int \prod_{i\in\mathcal{L}} \psi_W\left(\frac{\phi(y)t_i}{a_n}\right)\, \nu^{(1)}(dy)\right)^n.\end{aligned}$$

Now, since

$$\begin{aligned}&\prod_{i\in\mathcal{L}} \psi_W\left(\frac{\phi(y)t_i}{a_n}\right)\\ &\quad= 1 - c_\alpha \frac{b_n}{n}\sum_{i\in\mathcal{L}} |\phi(y)t_i|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t_i|}\right)}{L_0(a_n)} + o\left(\frac{b_n}{n}\sum_{i\in\mathcal{L}} |\phi(y)t_i|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t_i|}\right)}{L_0(a_n)}\right),\end{aligned}$$

it follows that

$$\begin{aligned}\frac{\psi_{(Y_i^{(2)}(\mathbf{x}))_{i\in\mathcal{L}}}(\mathbf{t})}{\psi_B(\mathbf{t})} &= \left(1 - c_\alpha \frac{b_n}{n}\sum_{i\in\mathcal{L}} \int |\phi(y)t_i|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t_i|}\right)}{L_0(a_n)}\, \nu^{(1)}(dy) + o\left(\frac{b_n}{n}\right)\right)^n\\ &\to \exp\left(-c_\alpha \sum_{i\in\mathcal{L}} |t_i|^\alpha \int |\phi(y)|^\alpha\, \nu^{(1)}(dy)\right)\\ &= \prod_{i\in\mathcal{L}} \exp\left(-c_\alpha |t_i|^\alpha \int |\phi(y)|^\alpha\, \nu^{(1)}(dy)\right).\end{aligned}$$

This proves the case $\ell = 2$.

Case $\ell > 2$:

The remainder of the proof uses induction on the layer ℓ, the base case being $\ell = 2$ proved above. Let $\ell > 2$. Also, let $n = n_{\ell-1}$, $\alpha = \alpha_\ell$, $a_n = a_{n_{\ell-1}}(\ell)$, $\sigma_B = \sigma_{B^{(\ell)}}$, and $t \neq 0$. Then $\{Y_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\dots,n}$ is no longer i.i.d.; however, it is still exchangeable. By de Finetti's theorem (see the end of Section 2), there exists a random probability measure

$$\xi^{(\ell-1)}(dy) := \xi^{(\ell-1)}(dy, \omega) := \xi^{(\ell-1)}(dy, \omega; \mathbf{n}) \tag{3.8}$$

such that given $\xi^{(\ell-1)}$, the random variables $Y_j^{(\ell-1)}(\mathbf{x})$, $j = 1, 2, \dots$, are i.i.d. with distribution $\xi^{(\ell-1)}(dy, \omega)$, where $\omega \in \Omega$ is an element of the probability space.

As before, we start by proving convergence of the marginal distribution. Taking the conditional expectation of (3.4), given $\xi^{(\ell-1)}$, we have

$$\begin{aligned}\psi_{Y_i^{(\ell)}(\mathbf{x})|\xi^{(\ell-1)}}(t) &:= \mathbb{E}\left[\psi_{Y_i^{(\ell)}(\mathbf{x})|\{Y_j^{(\ell-1)}(\mathbf{x})\}_j}(t)\,\middle|\, \xi^{(\ell-1)}\right]\\ &= e^{-\sigma_B^\alpha |t|^\alpha}\mathbb{E}\left[\prod_{j=1}^{n} \psi_{W_{ij}^{(\ell)}}\left(\frac{\phi(Y_j^{(\ell-1)}(\mathbf{x}))}{a_n}t\right)\,\middle|\, \xi^{(\ell-1)}\right]\\ &= e^{-\sigma_B^\alpha |t|^\alpha}\left(\int \psi_W\left(\frac{\phi(y)}{a_n}t\right)\, \xi^{(\ell-1)}(dy)\right)^n,\end{aligned}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

where $\psi_W := \psi_{W_{ij}^{(\ell)}}$ for some/any i,j. Using Lemma A.2 and Lemma A.4 again, we get

$$\left(\int \psi_W\left(\frac{\phi(y)t}{a_n}\right)\xi^{(\ell-1)}(dy)\right)^n = \tag{3.9}$$
$$\left(1-c_\alpha\frac{b_n}{n}\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\xi^{(\ell-1)}(dy)+o\left(\frac{b_n}{n}\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\xi^{(\ell-1)}(dy)\right)\right)^n.$$

Note that these are random integrals since $\xi^{(\ell-1)}(dy)$ is random, whereas the corresponding integral in the case $\ell=2$ was deterministic. Also, each integral on the right-hand side is finite almost surely since ϕ is bounded. By the induction hypothesis, the joint distribution of $(Y_i^{(\ell-1)}(\mathbf{x}))_{i\geq 1}$ converges weakly to the product measure $\bigotimes_{i\geq 1}\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}$. We claim that

$$\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\xi^{(\ell-1)}(dy)\xrightarrow{p}|t|^\alpha\int|\phi(y)|^\alpha\,\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy). \tag{3.10}$$

To see this, note that

$$\left|\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\xi^{(\ell-1)}(dy)-\int|\phi(y)t|^\alpha\,\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right| \tag{3.11}$$
$$\leq\left|\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\xi^{(\ell-1)}(dy)-\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right|$$
$$+\left|\int|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)-\int|\phi(y)t|^\alpha\,\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right|.$$

First, consider the first term on the right-hand side of the above. By Corollary 3.1, the random measures $\xi^{(\ell-1)}$ converge weakly, in probability, to $\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}$ as $\mathbf{n}\to\infty$ in the sense of (2.3), where $\mathbf{n}\in\mathbb{N}^{\ell_{\text{lay}}}$. Also, by Lemma A.4, we have

$$|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\leq b|\phi(y)t|^{\alpha\pm\epsilon} \tag{3.12}$$

for large n. For any subsequence $(\mathbf{n}_j)_j$, there is a further subsequence $(\mathbf{n}_{j_k})_k$ along which, ω-almost surely, $\xi^{(\ell-1)}$ converges weakly to $\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}$. To prove that the first term on the right-hand side of (3.11) converges in probability to 0, it is enough to show that it converges almost surely to 0 along each subsequence $(\mathbf{n}_{j_k})_k$. Fix an ω-realization of the random distributions $(\xi^{(\ell-1)}(dy,\omega;\mathbf{n}))_{\mathbf{n}\in\mathbb{N}^{\ell_{\text{lay}}}}$ such that convergence along the subsequence $(\mathbf{n}_{j_k})_k$ holds. Keeping ω fixed, view $g(y_\mathbf{n})=|\phi(y_\mathbf{n})t|^{\alpha\pm\epsilon}$ as a random variable where the parameter $y_\mathbf{n}$ is sampled from the distribution $\xi^{(\ell-1)}(dy,\omega;\mathbf{n})$. Since ϕ is bounded, the family of these random variables is uniformly integrable. Since $\xi^{(\ell-1)}(dy,\omega;\mathbf{n})$ converges weakly to $\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}$ along the subsequence, the Skorokhod representation and Vitali convergence theorem [37, p. 94] guarantee the convergence of the first term on the right-hand side of (3.11) to 0 as $\mathbf{n}$ tends to ∞.

Now, for the second term, since

$$\lim_{n\to\infty}|\phi(y)t|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}=|\phi(y)t|^\alpha$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

for each y and ϕ is bounded, we can use dominated convergence via (3.12) to show that the second term on the right-hand side of (3.11) also converges to 0, proving the claim.

Having proved (3.10), we have

$$\left(1+\frac{1}{n}\left(-c_\alpha b_n \int |\phi(y)t|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t|}\right)}{L_0(a_n)}\,\xi^{(\ell-1)}(dy)+o(b_n)\right)\right)^n \xrightarrow{p} \exp\left(-c_\alpha|t|^\alpha \int |\phi(y)|^\alpha\,\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right)$$

and hence

$$\psi_{Y_i^{(\ell)}(\mathbf{x})|\xi^{(\ell-1)}}(t) \xrightarrow{p} e^{-\sigma_B^\alpha|t|^\alpha}\exp\left(-c_\alpha|t|^\alpha \int |\phi(y)|^\alpha\,\mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right).$$

Thus, the limiting distribution of $Y_i^{(\ell)}(\mathbf{x})$, given $\xi^{(\ell-1)}$, is μ_{α,σ_ℓ} with

$$\sigma_\ell^\alpha = \sigma_B^\alpha + c_\alpha \int |\phi(y)|^\alpha\,\mu_{\alpha,\sigma_{\ell-1}}(dy).$$

Recall that characteristic functions are bounded by 1. Thus, by taking the expectation of both sides and using dominated convergence, we can conclude that the (unconditional) characteristic function converges to the same expression and thus the (unconditional) distribution of $Y_i^{(\ell)}(\mathbf{x})$ converges weakly to μ_{α,σ_ℓ}.

Finally, we prove that the joint distribution converges weakly to the product $\bigotimes_{i\geq 1}\mu_{\alpha,\sigma_\ell}$. Let $\mathcal{L}\subset\mathbb{N}$ be a finite set and $\mathbf{t}=(t_i)_{i\in\mathcal{L}}$. Conditionally on $\{Y_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\dots,n}$,

$$\psi_{(Y_i^{(\ell)}(\mathbf{x}))_{i\in\mathcal{L}}|\{Y_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\dots,n}}(\mathbf{t}) = \psi_B(\mathbf{t})\prod_{j=1}^n\prod_{i\in\mathcal{L}}\psi_W\left(\frac{\phi(Y_j^{(\ell-1)}(\mathbf{x}))t_i}{a_n}\right). \tag{3.13}$$

Taking the expectation with respect to $\{Y_j^{(\ell-1)}(\mathbf{x})\}_{j=1,\dots,n}$, we have

$$\begin{aligned}\frac{\psi_{(Y_i^{(\ell)}(\mathbf{x}))_{i\in\mathcal{L}}}(\mathbf{t})}{\psi_B(\mathbf{t})} &= \mathbb{E}\int\prod_{j=1}^n\prod_{i\in\mathcal{L}}\psi_W\left(\frac{\phi(y_j)t_i}{a_n}\right)\bigotimes_{j\geq 1}\xi^{(\ell-1)}(dy_j)\\ &= \mathbb{E}\left(\int\prod_{i\in\mathcal{L}}\psi_W\left(\frac{\phi(y)t_i}{a_n}\right)\xi^{(\ell-1)}(dy)\right)^n.\end{aligned}$$

Now since

$$\prod_{i\in\mathcal{L}}\psi_W\left(\frac{\phi(y)t_i}{a_n}\right)\sim 1-c_\alpha\frac{b_n}{n}\sum_{i\in\mathcal{L}}|\phi(y)t_i|^\alpha\frac{L_0\left(\frac{a_n}{|\phi(y)t_i|}\right)}{L_0(a_n)},$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

a similar argument to that of convergence of the marginal distribution shows that

$$\begin{aligned}\frac{\psi_{(Y_i^{(\ell)}(\mathbf{x}))_{i\in\mathcal{L}}}(\mathbf{t})}{\psi_B(\mathbf{t})} &\sim \mathbb{E}\left(1 - c_\alpha \frac{b_n}{n}\sum_{i\in\mathcal{L}}\int |\phi(y)t_i|^\alpha \frac{L_0\left(\frac{a_n}{|\phi(y)t_i|}\right)}{L_0(a_n)}\,\xi^{(\ell-1)}(dy)\right)^n \\ &\to \exp\left(-c_\alpha \sum_{i\in\mathcal{L}} |t_i|^\alpha \int |\phi(y)|^\alpha\, \mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right) \\ &= \prod_{i\in\mathcal{L}} \exp\left(-c_\alpha |t_i|^\alpha \int |\phi(y)|^\alpha\, \mu_{\alpha_{\ell-1},\sigma_{\ell-1}}(dy)\right),\end{aligned}$$

completing the proof. □

4. Relaxing the boundedness assumption

As we mentioned earlier in Remark 3.1, the boundedness assumption on ϕ can be relaxed, as long as it is done with care. It is known that the growth rate of the activation function ϕ affects the behavior of the network at deeper layers. If ϕ grows too fast, then the variance will quickly become too large at deeper layers, causing chaotic behavior of the network at those deeper layers. If, on the other hand, ϕ grows too slowly, then the variance will become too small, causing the network to behave as if it were not random [13, 15, 36]. Thus, it is important to find an appropriate growth rate for the activation function. Before presenting our result, we first present a counterexample where, for heavy-tailed initializations, we cannot use a function which grows linearly. This shows the subtlety of our relaxation.

Remark 4.1. Consider the case where $\phi = \mathrm{ReLU}$, $\mathbb{P}(|W_{ij}^{(\ell)}| > t) = t^{-\alpha}$ for $t \geq 1$, $0 < \alpha < 2$, and $\sigma_B = 0$. For an input $\mathbf{x} = (1, 0, \ldots, 0) \in \mathbb{R}^I$, we have

$$\begin{aligned}Y_i^{(1)}(\mathbf{x}) &= W_{i1}^{(1)}, \\ Y_i^{(2)}(\mathbf{x}) &= \frac{1}{a_n}\sum_{j=1}^n W_{ij}^{(2)} W_{j1}^{(1)} \mathbf{1}_{\{W_{j1}^{(1)}>0\}}, \\ a_n &= n^{1/\alpha}.\end{aligned}$$

Let us calculate the distribution function of $W_{ij}^{(2)} W_{j1}^{(1)} \mathbf{1}_{\{W_{j1}^{(1)}>0\}}$. For $z \geq 1$,

$$\begin{aligned}&\mathbb{P}(W_{ij}^{(2)} W_{j1}^{(1)} \mathbf{1}_{\{W_{j1}^{(1)}>0\}} \leq z) \\ &= \mathbb{P}(W_{j1}^{(1)} \leq 0) + \int_1^z \frac{\alpha w_1^{-\alpha-1}}{2}\mathbb{P}\left(W_{ij}^{(2)} \leq \frac{z}{w_1}\right) dw_1 + \int_z^\infty \frac{\alpha w_1^{-\alpha-1}}{2}\mathbb{P}\left(W_{ij}^{(2)} \leq -1\right) dw_1 \\ &= 1 - \frac{1}{4}z^{-\alpha} - \frac{1}{4}\alpha z^{-\alpha}\log z.\end{aligned}$$

Similarly, for $z \leq -1$,

$$\mathbb{P}(W_{ij}^{(2)} W_{j1}^{(1)} \mathbf{1}_{\{W_{j1}^{(1)}>0\}} < z) = \frac{1}{4}\alpha(-z)^{-\alpha}\log(-z) + \frac{1}{4}(-z)^{-\alpha}.$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Thus,

$$\mathbb{P}(|W_{ij}^{(2)} W_{j1}^{(1)} \mathbf{1}_{\{W_{j1}^{(1)}>0\}}| > z) = \frac{1}{2} z^{-\alpha} \, (1 + \alpha \log z) \, .$$

Let $\hat{a}_n := \inf\{x\colon \ x^{-\alpha}(1+\alpha \log x)/2 \le n^{-1}\}$. Then $n\hat{a}_n^{-\alpha}(1+\alpha \log \hat{a}_n)/2 \to 1$ as $n \to \infty$, which leads to

$$\frac{\hat{a}_n}{n^{1/\alpha}} \sim ((1+\alpha \log \hat{a}_n)/2)^{1/\alpha} \to \infty$$

when n is large. Thus, $\hat{a}_n$ is of strictly larger order than $n^{1/\alpha}$, which shows that $Y_i^{(2)}(\mathbf{x})$ does not converge using the suggested normalization.

However, despite the remark, one can modify the scaling to $a_n = n^{1/\alpha} L(n)$ where $L(n)$ is a nonconstant slowly varying factor, in order to make the network converge at initialization. For details, we refer to [6], where the authors handle the convergence of shallow ReLU networks with stable weights.

Despite the above remark, there is still room to relax the boundedness assumption on ϕ. Note that, in the proof of Theorem 3.1, we used boundedness (in a critical way) to prove the claim (3.10). In particular, boundedness gave us that the family of random variables $|\phi(y)|^{\alpha+\epsilon}$ with respect to the random distribution $\xi^{(\ell-1)}(dy, \omega; \mathbf{n})$ is y -uniformly integrable ω-almost surely. We make this into a direct assumption on ϕ as follows. Let $n := n_{\ell-2}$ and $a_n := a_{n_{\ell-2}}(\ell-1)$. Suppose

(UI1) for $\ell = 2$, there exists $\epsilon_0 > 0$ such that $|\phi(Y_j^{(1)})|^{\alpha_2+\epsilon_0}$ is integrable;

(UI2) for $\ell = 3, \ldots, \ell_{\text{lay}} + 1$, there exists $\epsilon_0 > 0$ such that for any array $(c_{\mathbf{n},j})_{\mathbf{n},j}$ satisfying

$$\sup_{\mathbf{n}} \frac{1}{n} \sum_{j=1}^{n} |c_{\mathbf{n},j}|^{\alpha_{\ell-1}+\epsilon_0} < \infty, \tag{4.1}$$

we have uniform integrability of the family

$$\left\{ \left| \phi\left(\frac{1}{a_n} \sum_{j=1}^{n} c_{\mathbf{n},j} W_j^{(\ell-1)} \right) \right|^{\alpha_\ell+\epsilon_0} \right\}_{\mathbf{n}} \tag{4.2}$$

over $\mathbf{n}$.

If ϕ is bounded, then the above is obviously satisfied. It is not clear whether there is a simpler description of the family of functions that satisfies this assumption (see [1]); however, we now argue that this is general enough to recover the previous results of Gaussian weights or stable weights.

In [30] (as well as many other references), the authors consider Gaussian initializations with an activation function ϕ satisfying the so-called polynomial envelope condition. That is, $|\phi(y)| \le a + b|y|^m$ for some $a, b > 0$ and $m \ge 1$ and $W \sim \mathcal{N}(0, \sigma^2)$. In this setting, we have $a_n \sim \sigma\sqrt{n/2}$ and $\alpha = 2$ for all ℓ, and $c_{\mathbf{n},j} = c_{\mathbf{n},j}^{(\ell-2)} = \phi(Y_j^{(\ell-2)}(\mathbf{x}; \mathbf{n}))$. Conditioning on $(Y_j^{(\ell-2)})_j$ and assuming that (4.1) holds almost surely, let us show that ϕ satisfying the polynomial envelope condition also satisfies our uniform integrability assumptions (UI1) and (UI2) almost surely. For $\ell = 2$, the distribution of

$$Y_i^{(1)} = \sum_{j=1}^{I} W_{ij}^{(1)} x_j + B_i^{(1)}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

is Gaussian, and thus $|\phi(Y_j^{(1)})|^{2+\epsilon_0} \le C_0 + C_1 |Y_j^{(1)}|^{m(2+\epsilon_0)}$ is integrable. For $\ell \ge 3$, note that

$$S_{\mathbf{n}}^{(\ell-1)} := \frac{1}{a_n} \sum_{j=1}^{n} c_{\mathbf{n},j} W_j^{(\ell-1)} \sim \mathcal{N}\left(0, \frac{2}{n} \sum_{j=1}^{n} c_{\mathbf{n},j}^2\right),$$

where the variance is uniformly bounded over $\mathbf{n}$ if we assume (4.1). For $\theta > 1$, let $\nu := m(2+\epsilon_0)\theta$; the νth moment of S_n can be directly calculated and is known to be

$$2^{\nu/2} \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1+\nu}{2}\right) \left(\frac{2}{n} \sum_{j=1}^{n} c_{\mathbf{n},j}^2\right)^{\nu/2}.$$

This is uniformly bounded over $\mathbf{n}$, and hence $|\phi(S_n)|^{2+\epsilon_0}$ is uniformly integrable over $\mathbf{n}$. This shows that ϕ satisfying the polynomial envelope condition meets (UI1) and (UI2) assuming (4.1).

In [4], the authors consider the case where $W^{(\ell)}$ is an SαS random variable with scale parameter σ_ℓ, i.e., with characteristic function $e^{-\sigma_\ell^\alpha |t|^\alpha}$. They use the envelope condition $|\phi(y)| \le a + b|y|^\beta$ where $\beta < 1$. For the more general case where we have different α_ℓ-stable weights for different layers ℓ, this envelope condition can be generalized to $\beta < \min_{\ell \ge 2} \alpha_{\ell-1}/\alpha_\ell$. In this case, $a_n^{\alpha_\ell} \sim (\sigma_\ell^{\alpha_\ell} n)/c_{\alpha_\ell}$ and $c_{\mathbf{n},j} = c_{\mathbf{n},j}^{(\ell-2)} = \phi(Y_j^{(\ell-2)}(\mathbf{x};\, \mathbf{n}))$. Again, conditioning on $(Y_j^{(\ell-2)})_j$ and assuming (4.1), let us show that ϕ under this generalized envelope condition satisfies the uniform integrability assumptions (UI1) and (UI2) above. For $\ell = 2$, the distribution of

$$Y_j^{(1)} = \sum_{j=1}^{I} W_{ij}^{(1)} x_j + B_i^{(1)}$$

is α_1-stable. By the condition on β, there are δ and ϵ_0 satisfying $\beta(\alpha_2 + \epsilon_0) \le \alpha_1 - \delta$ so that

$$|\phi(Y_j^{(1)})|^{\alpha_2+\epsilon} \le C_0 + C_1 |Y_j^{(1)}|^{\alpha_1-\delta},$$

which is integrable. For $\ell \ge 3$, the distribution of $S_{\mathbf{n}}^{(\ell-1)} := a_n^{-1} \sum_j c_{\mathbf{n},j} W_j^{(\ell-1)}$ becomes a symmetric $\alpha_{\ell-1}$-stable distribution with scale parameter

$$\left(\frac{c_{\alpha_{\ell-1}}}{n} \sum_{j=1}^{n} |c_{\mathbf{n},j}|^{\alpha_{\ell-1}}\right)^{1/\alpha_{\ell-1}},$$

which is uniformly bounded over $\mathbf{n}$ assuming (4.1). Since $\beta < \min_{\ell \ge 2} \alpha_{\ell-1}/\alpha_\ell$, it follows that, for some $\theta > 1$, there exist small $\epsilon_0 > 0$ and $\delta > 0$ such that

$$\left|\phi\left(S_{\mathbf{n}}^{(\ell-1)}\right)\right|^{(\alpha_\ell+\epsilon_0)\theta} \le C_0 + C_1 \left|S_{\mathbf{n}}^{(\ell-1)}\right|^{\beta(\alpha_\ell+\epsilon_0)\theta} \le C_0 + C_1 \left|S_{\mathbf{n}}^{(\ell-1)}\right|^{\alpha_{\ell-1}-\delta}.$$

It is known (see for instance [39]) that the expectation of $|S_{\mathbf{n}}^{(\ell-1)}|^\nu$ with $\nu < \alpha_{\ell-1}$ is

$$K_\nu \left(\frac{c_{\alpha_{\ell-1}}}{n} \sum_{j=1}^{n} |c_{\mathbf{n},j}|^{\alpha_{\ell-1}}\right)^{\nu/\alpha_{\ell-1}},$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

where K_ν is a constant that depends only on ν (and $\alpha_{\ell-1}$). As this is bounded uniformly over $\mathbf{n}$, the family

$$\left\{\left|\phi\left(S_{\mathbf{n}}^{(\ell-1)}\right)\right|^{\alpha_\ell+\epsilon_0}\right\}_{\mathbf{n}}$$

is uniformly integrable. Thus our ϕ, under the generalized envelope condition, satisfies (UI1) and (UI2).

Let us now see that $c_{\mathbf{n},j}$ satisfies the condition (4.1) in both the Gaussian and the symmetric stable case. For $\ell = 3$, $c_{\mathbf{n},j} = \phi(Y_j^{(1)})$ satisfies (4.1) by the strong law of large numbers since $|\phi(Y_j^{(1)})|^{\alpha_2+\epsilon_0}$ is integrable. For $\ell > 3$, an inductive argument shows that the family $\{|\phi(Y_j^{(\ell-2)})|^{\alpha_{\ell-1}+\epsilon_0}\}_{\mathbf{n}}$ is uniformly integrable, which leads to (4.1). The details of this inductive argument are contained in the following proof.

Proof of Theorem 3.1 under (UI1) and (UI2). We return to the claim in (3.10) to see how the conditions (UI1) and (UI2) are sufficient, even when ϕ is unbounded. We continue to let $n := n_{\ell-2}$. Choose a sequence $\{(n, \mathbf{n})\}_n$, where $\mathbf{n} = \mathbf{n}(n)$ depends on n and $\mathbf{n} \to \infty$ as $n \to \infty$ in the sense of (2.3). Note that (i) to evaluate the limit as $\mathbf{n} \to \infty$, it suffices to show that the limit exists consistently for any choice of sequence $\{\mathbf{n}(n)\}_n$ that goes to infinity, and (ii) we can always pass to a subsequence (not depending on ω), since we are concerned with convergence in probability. Therefore, below we will show almost sure uniform integrability over some infinite subset of an arbitrary index set of the form $\{(n, \mathbf{n}(n)) : n \in \mathbb{N}\}$.

Let $a_n := a_{n_{\ell-2}}(\ell - 1)$. Proceeding as in (3.11) and (3.12), we need to show that the family $|\phi(y_{\mathbf{n}})|^{\alpha+\epsilon}$ where $y_{\mathbf{n}} \sim \xi^{(\ell-1)}(dy, \omega; \mathbf{n})$ is uniformly integrable. Since $\{a_n^{-1} \sum_j \phi(Y_j^{(\ell-2)}) W_{ij}^{(\ell-1)}\}_i$ is conditionally i.i.d. given $\{Y_j^{(\ell-2)}\}_j$, the random distribution $\xi^{(\ell-1)}(dy, \omega; \mathbf{n})$ is the law of $a_n^{-1} \sum_j \phi(Y_j^{(\ell-2)}) W_{ij}^{(\ell-1)}$ given $\{Y_j^{(\ell-2)}\}_j$, by the uniqueness of the directing random measure (see [20, Proposition 1.4]). Thus, by (UI2), it suffices to check that $n^{-1} \sum_j |\phi(Y_j^{(\ell-2)})|^{\alpha_{\ell-1}+\epsilon_0}$ is uniformly bounded for $\ell = 3, \ldots, \ell_{\text{lay}} + 1$. For $\ell = 3$, since $|\phi(Y_j^{(1)})|^{\alpha_2+\epsilon_0}$ is integrable by (UI1),

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} |\phi(Y_j^{(1)})|^{\alpha_2+\epsilon_0} < \infty$$

by the strong law of large numbers, and hence the normalized sums are almost surely bounded. For $\ell > 3$, we proceed inductively. By the inductive hypothesis, we have

$$\sup_{\mathbf{n}} \frac{1}{n_{\ell-3}} \sum_{j=1}^{n_{\ell-3}} |\phi(Y_j^{(\ell-3)})|^{(\alpha_{\ell-2}+\epsilon_0)(1+\epsilon')} < \infty$$

by adjusting $\epsilon_0, \epsilon' > 0$ appropriately. By (UI2), we have that the family

$$\{|\phi(y_{\mathbf{n}})|^{(\alpha_{\ell-1}+\epsilon_0)(1+\epsilon'')} : y_{\mathbf{n}} \sim \xi^{(\ell-2)}(dy; \mathbf{n})\}$$

is almost surely uniformly integrable for some $\epsilon'' > 0$. Since the $Y_j^{(\ell-2)}$ are conditionally i.i.d. with common distribution $\xi^{(\ell-2)}(dy; \mathbf{n})$ given $\xi^{(\ell-2)}(dy, \omega; \mathbf{n})$, by Lemma A.6 we have that

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{j=1}^{n} |\phi(Y_j^{(\ell-2)})|^{\alpha_{\ell-1}+\epsilon_0} - \int |\phi(y)|^{\alpha_{\ell-1}+\epsilon_0} \xi^{(\ell-2)}(dy; \mathbf{n})\right| \geq \delta \;\middle|\; \xi^{(\ell-2)}\right) \to 0$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

almost surely. By the dominated convergence theorem we can take expectations on both sides to conclude that

$$\left|\frac{1}{n}\sum_{j=1}^{n}|\phi(Y_j^{(\ell-2)})|^{\alpha_{\ell-1}+\epsilon_0} - \int |\phi(y)|^{\alpha_{\ell-1}+\epsilon_0}\xi^{(\ell-2)}(dy;\ \mathbf{n})\right| \to 0$$

in probability, so by passing to a subsequence we have that the convergence holds for almost every ω. Since

$$\sup_{\mathbf{n}} \int |\phi(y)|^{\alpha_{\ell-1}+\epsilon_0}\xi^{(\ell-2)}(dy;\ \mathbf{n}) < \infty$$

almost surely, we have also that

$$\sup_{\mathbf{n}} \frac{1}{n}\sum_{j=1}^{n}|\phi(Y_j^{(\ell-2)})|^{\alpha_{\ell-1}+\epsilon_0} < \infty$$

almost surely, proving our claim. □

5. Joint convergence with different inputs

In this section, we extend Theorem 3.1 to the joint distribution of k different inputs. We show that the k -dimensional vector $(Y_i^{(\ell)}(\mathbf{x}_1;\mathbf{n}), \ldots, Y_i^{(\ell)}(\mathbf{x}_k;\ \mathbf{n}))$ converges, and we represent the limiting characteristic function via a finite measure Γ_ℓ on the unit sphere $S_{k-1} = \{x \in \mathbb{R}^k : |x| = 1\}$, called the spectral measure. This extension to k inputs is needed for our convergence result to be applied in practice, since practical applications involve multiple inputs: a network is trained on a set of input–output pairs, and the trained network is then used to predict the output of a new unseen input. For instance, as suggested in the work on infinitely wide networks with Gaussian initialization [24, 25], such an extension is needed to perform Bayesian posterior inference and prediction with heavy-/light-tailed infinitely wide MLPs, where the limiting process in the multi-input extension is conditioned on k_0 input–output pairs, with $k_0 < k$, and then the resulting conditional or posterior distribution of the process is used to predict the outputs of the process for $k - k_0$ inputs.

For simplicity, we use the following notation:

- $\vec{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ where $\mathbf{x}_j \in \mathbb{R}^I$.
- $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^k$.
- $\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\ \mathbf{n}) = (Y_i^{(\ell)}(\mathbf{x}_1;\ \mathbf{n}), \ldots, Y_i^{(\ell)}(\mathbf{x}_k;\ \mathbf{n})) \in \mathbb{R}^k$, for $i \in \mathbb{N}$.
- $\phi(\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\ \mathbf{n})) = (\phi(Y_i^{(\ell)}(\mathbf{x}_1;\ \mathbf{n})), \ldots, \phi(Y_i^{(\ell)}(\mathbf{x}_k;\ \mathbf{n}))) \in \mathbb{R}^k$.
- $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^k$.
- For any given j, let the law of the k -dimensional vector $\mathbf{Y}_j^{(\ell)}(\vec{\mathbf{x}})$ be denoted by $\nu_k^{(\ell)}$ (which does not depend on j). Its projection onto the s th component $Y_i^{(\ell)}(\mathbf{x}_s;\ \mathbf{n})$ is denoted by $\nu_{k,s}^{(\ell)}$ for $1 \le s \le k$, and the projection onto two coordinates, the ith and jth, is denoted by $\nu_{k,ij}^{(\ell)}$. The limiting distribution of $\mathbf{Y}_j^{(\ell)}(\vec{\mathbf{x}})$ is denoted by $\mu_k^{(\ell)}$, and the projections are similarly denoted by $\mu_{k,s}^{(\ell)}$ and $\mu_{k,ij}^{(\ell)}$.

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

- A centered k-dimensional multivariate Gaussian with covariance matrix M is denoted by $\mathcal{N}_k(M)$.
- For $\alpha < 2$, we denote the k -dimensional SαS distribution with spectral measure Γ by $\mathrm{S}_\alpha\mathrm{S}_k(\Gamma)$. For those not familiar with the spectral measure of a multivariate stable law, Appendix C provides background.

Recall that

$$c_\alpha = \lim_{M\to\infty} \int_0^M \frac{\sin u}{u^\alpha} du \quad \text{and} \quad c_2 = 1.$$

Theorem 5.1. *Let $(Y_i^{(\ell)}(\,\cdot\,;\,\mathbf{n}))_{i\geq 1}$ be defined as in Section 2, and $(\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\,\mathbf{n}))_{i\geq 1}$ as above. Then, for each $\ell = 2, \ldots, \ell_{\text{lay}} + 1$, the joint distribution of the random variables $(\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\,\mathbf{n}))_{i\geq 1}$ converges weakly to $\mu_k^{(\ell)}$ as given below:*

- *For $\alpha_\ell < 2$, $\mu_k^{(\ell)} = \bigotimes_{i\geq 1} \mathrm{S}\alpha_\ell\mathrm{S}_k(\Gamma_\ell)$, where Γ_ℓ is defined by*

$$\Gamma_2 = \left\|\sigma_{B^{(2)}}\mathbf{1}\right\|^{\alpha_2} \delta_{\frac{\mathbf{1}}{\|\mathbf{1}\|}} + c_{\alpha_2} \int \|\phi(\mathbf{y})\|^{\alpha_2}\ \delta_{\frac{\phi(\mathbf{y})}{\|\phi(\mathbf{y})\|}}\ \nu_k^{(1)}(d\mathbf{y}) \tag{5.1}$$

and

$$\Gamma_\ell = \left\|\sigma_{B^{(\ell)}}\mathbf{1}\right\|^{\alpha_\ell} \delta_{\frac{\mathbf{1}}{\|\mathbf{1}\|}} + c_{\alpha_\ell} \int \|\phi(\mathbf{y})\|^{\alpha_\ell}\ \delta_{\frac{\phi(\mathbf{y})}{\|\phi(\mathbf{y})\|}}\ \mu_k^{(\ell-1)}(d\mathbf{y}) \tag{5.2}$$

for $\ell > 2$.

- *For $\alpha_\ell = 2$, $\mu_k^{(\ell)} = \bigotimes_{i\geq 1} \mathcal{N}_k(M_\ell)$, where*

$$\begin{aligned}(M_2)_{ii} &= \mathbb{E}|B_i^{(2)}|^2 + \frac{1}{2}\int |\phi(y)|^2\ \nu_{k,i}^{(1)}(dy),\\ (M_2)_{ij} &= \frac{1}{2}\int \phi(y_1)\phi(y_2)\ \nu_{k,ij}^{(1)}(dy_1 dy_2),\end{aligned} \tag{5.3}$$

and

$$\begin{aligned}(M_\ell)_{ii} &= \mathbb{E}|B_i^{(\ell)}|^2 + \frac{1}{2}\int |\phi(y)|^2\ \mu_{k,i}^{(\ell-1)}(dy),\\ (M_\ell)_{ij} &= \frac{1}{2}\int \phi(y_1)\phi(y_2)\ \mu_{k,ij}^{(\ell-1)}(dy_1 dy_2)\end{aligned} \tag{5.4}$$

for $\ell > 2$.

As mentioned below the statement of Theorem 3.1, this theorem finally shows that the individual layers of an MLP initialized with arbitrary heavy-/light-tailed weights have a limit, as the width tends to infinity, which is a stable process in the parameter $\mathbf{x}$.

Proof. Let $\mathbf{t} = (t_1, \ldots, t_k)$. We again start with the expression

$$\begin{aligned}&\psi_{\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}})|\{\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})\}_{j\geq 1}}(\mathbf{t})\\ &\quad = \mathbb{E}\left[e^{i\langle \mathbf{t}, \mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}})\rangle}\,\middle|\,\{\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})\}_{j\geq 1}\right]\end{aligned} \tag{5.5}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

$$
\begin{aligned}
&= \mathbb{E}\left[\exp\left(i\left\langle \mathbf{t}, \frac{1}{a_n}\sum_{j=1}^{n} W_{ij}^{(\ell)}\phi(\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})) + B_i^{(\ell)}\mathbf{1}\right\rangle\right)\middle| \{\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})\}_{j\geq 1}\right]\\
&= \mathbb{E}e^{iB_i^{(\ell)}\langle \mathbf{t},\mathbf{1}\rangle}\prod_{j=1}^{n}\mathbb{E}\left[\exp\left(i\frac{1}{a_n}W_{ij}^{(\ell)}\left\langle \mathbf{t}, \phi(\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}}))\right\rangle\right)\middle| \{\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})\}_{j\geq 1}\right]\\
&= \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle)\left(\psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}}))\rangle\right)\right)^n.
\end{aligned}
$$

Here ψ_B and ψ_W are characteristic functions of the random variables $B_i^{(\ell)}$ and $W_{ij}^{(\ell)}$ for some/any i,j.

Case $\ell = 2$:

As before, let $n = n_1$, $\alpha = \alpha_2$, and $a_n = a_{n_1}(2)$. As in Theorem 3.1, $(\mathbf{Y}_j^{(1)}(\vec{\mathbf{x}}))_{j\geq 1}$ is i.i.d, and thus

$$
\begin{aligned}
\psi_{\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}})}(\mathbf{t}) &= \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle)\mathbb{E}\left(\psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}}))\rangle\right)\right)^n\\
&= \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle)\int\left(\psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right)\right)^n \nu_k^{(1)}(d\mathbf{y}).
\end{aligned}
$$

As before,

$$
\begin{aligned}
&\left(\psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right)\right)^n\\
&\quad= \left(1 - c_\alpha\frac{b_n}{n}|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\frac{L_0\left(\frac{a_n}{|\langle \mathbf{t},\phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)} + o\left(\frac{b_n}{n}|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\frac{L_0\left(\frac{a_n}{|\langle \mathbf{t},\phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)}\right)\right)^n.
\end{aligned}
$$

The main calculation needed to extend the proof of Theorem 3.1 to the situation involving $\vec{\mathbf{x}}$ is as follows. Assuming the uniform integrability in Section 4, we have, for some $b > 0$ and $0 < \epsilon < \epsilon_0$,

$$
\begin{aligned}
\int|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\frac{L_0\left(\frac{a_n}{|\langle \mathbf{t},\phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)}\nu_k^{(1)}(d\mathbf{y}) &\leq b\int|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha\pm\epsilon}\ \nu_k^{(1)}(d\mathbf{y})\\
&= \int b\left|\sum_{s=1}^{k} t_s\phi(y_s)\right|^{\alpha\pm\epsilon}\nu_k^{(1)}(d\mathbf{y})\\
&\leq \int bc_k\sum_{s=1}^{k}|t_s\phi(y_s)|^{\alpha\pm\epsilon}\ \nu_k^{(1)}(d\mathbf{y})\\
&= bc_k\sum_{s=1}^{k}\int|t_s\phi(y_s)|^{\alpha\pm\epsilon}\ \nu_{k,s}^{(1)}(dy) < \infty.
\end{aligned}
\tag{5.6}
$$

It thus follows that

$$
\begin{aligned}
&\int|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\frac{L_0\left(\frac{a_n}{|\langle \mathbf{t},\phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)}\nu_k^{(1)}(d\mathbf{y}) \to \int|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\ \nu_k^{(1)}(d\mathbf{y}), \quad \text{and}\\
&\int o\left(\frac{b_n}{n}|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha\frac{L_0\left(\frac{a_n}{|\langle \mathbf{t},\phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)}\right)\nu_k^{(1)}(d\mathbf{y}) = o\left(\frac{b_n}{n}\right).
\end{aligned}
$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Therefore,

$$\psi_B(\langle \mathbf{t}, \mathbf{1}\rangle)\mathbb{E}\left(\psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right)\right)^n \to \exp\left(-\sigma_B^\alpha |\langle \mathbf{t}, \mathbf{1}\rangle|^\alpha\right)\exp\left(-c_\alpha \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \ \nu_k^{(1)}(d\mathbf{y})\right). \tag{5.7}$$

Let $\|\cdot\|$ denote the standard Euclidean norm. Observe that for $\alpha < 2$,

$$c_\alpha \, |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha = c_\alpha \int_{S^{k-1}} |\langle \mathbf{t}, \mathbf{s}\rangle|^\alpha \ \|\phi(\mathbf{y})\|^\alpha \ \delta_{\frac{\phi(\mathbf{y})}{\|\phi(\mathbf{y})\|}}(d\mathbf{s}).$$

Thus, by Theorem C.1, we have the convergence $\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\ \mathbf{n}) \xrightarrow{w} \mathrm{S}_\alpha \mathrm{S}_k(\Gamma_2)$ where Γ_2 is defined by (5.1).

For $\alpha = 2$, we have

$$\exp\left(-c_\alpha \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^2 \nu_k^{(1)}(d\mathbf{y})\right) = \exp\left(-\frac{1}{2}\langle \mathbf{t}, M_2\mathbf{t}\rangle\right)$$

where M_2 is given by (5.3), which is equal to the characteristic function of $\mathcal{N}(M_2)$.

Extending the calculations in (3.7), the convergence $(\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\ \mathbf{n}))_{i\geq 1} \xrightarrow{w} \bigotimes_{i\geq 1} \mathrm{S}_\alpha \mathrm{S}_k(\Gamma_2)$ follows similarly.

Case $\ell > 2$:

Similarly to (3.8), let $\xi^{(\ell-1)}(d\mathbf{y}, \omega)$ be a random distribution such that, given $\xi^{(\ell-1)}$, the random vectors $\mathbf{Y}_j^{(\ell-1)}(\vec{\mathbf{x}})$, $j = 1, 2, \ldots$, are i.i.d. with distribution $\xi^{(\ell-1)}(d\mathbf{y})$.

Taking the conditional expectation of (5.5) given $\xi^{(\ell-1)}$, we get

$$\psi_{\mathbf{Y}_i^{(\ell)}|\xi^{(\ell-1)}}(\mathbf{t}) = \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle)\mathbb{E}\left[\left(\int \psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right)\xi^{(\ell-1)}(d\mathbf{y})\right)^n \ \middle|\ \xi^{(\ell-1)}\right]$$

for any i. Here,

$$\int \psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right)\xi^{(\ell-1)}(d\mathbf{y}) \sim 1 - c_\alpha \frac{b_n}{n}\int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\xi^{(\ell-1)}(d\mathbf{y}).$$

From the induction hypothesis, $(\mathbf{Y}_i^{(\ell-1)}(\vec{\mathbf{x}}))_{i\geq 1}$ converges weakly either to $\bigotimes_{i\geq 1} \mathrm{S}_\alpha \mathrm{S}(\Gamma_{\ell-1})$ or to $\bigotimes_{i\geq 1} \mathcal{N}_k(M_\ell)$. We claim that

$$\int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\xi^{(\ell-1)}(d\mathbf{y}) \xrightarrow{p} \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \mu_k^{(\ell-1)}(d\mathbf{y}).$$

To see this, note that

$$\begin{aligned}
&\left|\int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\xi^{(\ell-1)}(d\mathbf{y}) - \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \mu_k^{(\ell-1)}(d\mathbf{y})\right| \qquad (5.8)\\
&\quad \leq \left|\int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\xi^{(\ell-1)}(d\mathbf{y}) - \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\mu_k^{(\ell-1)}(d\mathbf{y})\right|\\
&\quad + \left|\int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \frac{L_0(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|})}{L_0(a_n)}\mu_k^{(\ell-1)}(d\mathbf{y}) - \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^\alpha \mu_k^{(\ell-1)}(d\mathbf{y})\right|.
\end{aligned}$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Now, the uniform integrability assumption in Section 4 combined with (5.6) shows that

$$|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha} \frac{L_0\left(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)}$$

is uniformly integrable with respect to the family $(\xi^{(\ell-1)})_n$, and thus the first term on the right-hand side of (5.8) converges in probability to 0. Also, from (5.6) and the fact that

$$\lim_{n\to\infty} |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha} \frac{L_0\left(\frac{a_n}{|\langle \mathbf{t}, \phi(\mathbf{y})\rangle|}\right)}{L_0(a_n)} = |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha}$$

for each $\mathbf{y}$, dominated convergence gives us convergence to 0 of the second term. Therefore,

$$\left(\int \psi_W\left(\frac{1}{a_n}\langle \mathbf{t}, \phi(\mathbf{y})\rangle\right) \xi^{(\ell-1)}(d\mathbf{y})\right)^n \xrightarrow{p} \exp\left(-c_\alpha \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha} \mu_k^{(\ell-1)}(d\mathbf{y})\right),$$

and consequently,

$$\psi_{\mathbf{Y}_i^{(\ell)}|\xi^{(\ell-1)}}(\mathbf{t}) \xrightarrow{p} \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle) \exp\left(-c_\alpha \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha} \mu_k^{(\ell-1)}(d\mathbf{y})\right).$$

Finally, noting that the characteristic function is bounded by 1 and using dominated convergence, we get

$$\psi_{\mathbf{Y}_i^{(\ell)}}(\mathbf{t}) \xrightarrow{p} \psi_B(\langle \mathbf{t}, \mathbf{1}\rangle) \exp\left(-c_\alpha \int |\langle \mathbf{t}, \phi(\mathbf{y})\rangle|^{\alpha} \mu_k^{(\ell-1)}(d\mathbf{y})\right),$$

where the right-hand side is the characteristic function of $\mathrm{S}_\alpha\mathrm{S}_k(\Gamma_\ell)$ (or $\mathcal{N}_k(M_\ell)$ for $\alpha = 2$), with Γ_ℓ and M_ℓ given by (5.2) and (5.4), respectively.

The proof of $(\mathbf{Y}_i^{(\ell)}(\vec{\mathbf{x}};\ \mathbf{n}))_{i\geq 1} \xrightarrow{w} \bigotimes_{i\geq 1} \mathrm{S}_\alpha\mathrm{S}_k(\Gamma_\ell)$ (or $\bigotimes_{i\geq 1} \mathcal{N}_k(M_\ell)$ in the case $\alpha = 2$) follows similarly to the calculations following (3.13). □

6. Conclusion and future directions

We have considered a deep feed-forward neural network whose weights are i.i.d. heavy-tailed or light-tailed random variables (Section 2). If the activation function is bounded and continuous, then as the width goes to infinity, the joint pre-activation values in a given layer of the network, for a given input, converge in distribution to a product of i.i.d. SαS random variables (Theorem 3.1), whose scale parameter is inductively defined by (3.1). This is generalized to multiple inputs (Theorem 5.1), where the pre-activation values converge to a multivariate SαS distribution whose spectral measure (or, in the case $\alpha = 2$, the covariance matrix) is inductively defined by (5.1)–(5.4). These results show that an initialization using any i.i.d. heavy-/light-tailed weights can be treated similarly to an α-stable prior assumption in the context of Bayesian modeling. In Section 4, we sought a more general assumption on the activation function, beyond boundedness. This is of importance because if the activation function is not carefully chosen, then the initialized variances may exhibit erratic behavior as the number of layers grows: either collapsing to zero (so that pre-activation values at deeper layers saturate), or exploding to infinity [13, 15, 36]. Unlike the case of Gaussian initialization, our model in general does not allow the use of ReLU. The trade-off is that we allow the use of arbitrary

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

heavy-/light-tailed distributions for network weights, which is favorable for encoding heavy-tailed behaviors of neural networks that are known to arise in well-known trained networks [28, 42, 9].

Gradient descent on an infinitely wide deep network with the L^2-loss function is related to the kernel method via the neural network Gaussian process (NNGP) kernel [31, 26] and the neural tangent kernel (NTK) [17, 2]. One interesting future direction is to generalize this relationship with the kernel method to our model, in particular, by finding an appropriate counterpart of the NTK. For shallow networks with stable weights and ReLU activation, it has been shown that the NTK converges in distribution as the width tends to infinity [6], and the network dynamics have been explained in terms of the kernel method. Another possible future direction is to relax the independence assumptions on the weights. For instance, it should be possible to extend the infinite-width limit result to the case of exchangeable weights in each layer. Indeed, in [40], the authors consider row–column exchangeable random variables for network weights in each layer and analyze the infinite-width limit of such a network. Some authors have also proposed structured recipes for designing a network with dependent weights while ensuring that the weights are partially exchangeable. One particular way is to consider a scale mixture of Gaussians for the weight distribution [18, 34, 27, 11, 12]. Infinite-width limits of these networks with Gaussian scale mixture weights have also been studied, at least in part, by [23]. However, it would be more challenging to generalize the infinite-width limit result to a network with general dependent structures for weights.

Appendix A. Auxiliary lemmas

Lemma A.1. *If L is slowly varying, then*

$$\widetilde{L}(x) = \int_0^x t^{-1} L(t)\, dt$$

is also slowly varying.

Proof. If $\widetilde{L}$ is bounded, then since $\widetilde{L}$ is increasing, $\widetilde{L}(x)$ converges as $x \to \infty$. Thus $\widetilde{L}$ is slowly varying. If $\widetilde{L}$ is not bounded, then by L'Hôpital's rule,

$$\lim_{x\to\infty} \frac{\widetilde{L}(\lambda x)}{\widetilde{L}(x)} = \lim_{x\to\infty} \frac{\int_0^x y^{-1} L(\lambda y)\, dy}{\int_0^x y^{-1} L(y)\, dy} = \lim_{x\to\infty} \frac{L(\lambda x)}{L(x)} = 1.$$

□

The next four lemmas are standard results; we give references for their proofs. In particular, the next lemma is a standard result concerning the characteristic functions of heavy-tailed distributions ([35, Theorem 1 and Theorem 3]; see also [3, Equation 3.8.2]).

Lemma A.2. *If W is a symmetric random variable with tail probability $\mathbb{P}(|W| > t) = t^{-\alpha} L(t)$ where $0 < \alpha \le 2$ and L is slowly varying, then the characteristic function $\psi_W(t)$ of W satisfies*

$$\psi_W(t) = 1 - c_\alpha |t|^\alpha L\left(\frac{1}{|t|}\right) + o\left(|t|^\alpha L\left(\frac{1}{|t|}\right)\right), \quad t \to 0,$$

where

$$c_\alpha = \lim_{M\to\infty} \int_0^M \frac{\sin u}{u^\alpha}\, du = \frac{\pi/2}{\Gamma(\alpha)\sin(\pi\alpha/2)},$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

for $\alpha < 2$, *and*

$$\psi_W(t) = 1 - |t|^2 \widetilde{L}\left(\frac{1}{|t|}\right) + o\left(|t|^2 \widetilde{L}\left(\frac{1}{|t|}\right)\right), \quad t \to 0,$$

where

$$\widetilde{L}(x) = \int_0^x y\mathbb{P}(|W| > y)\, dy = \int_0^x y^{-1} L(y)\, dy,$$

for $\alpha = 2$.

We next state a standard result about slowly varying functions [7, Section VIII.8, Lemma 2].

Lemma A.3. *If L is slowly varying, then for any fixed $\epsilon > 0$ and all sufficiently large x,*

$$x^{-\epsilon} < L(x) < x^{\epsilon}.$$

Moreover, the convergence

$$\frac{L(tx)}{L(t)} \to 1$$

as $t \to \infty$ is uniform in finite intervals $0 < a < x < b$.

An easy corollary of the above lemma is the following result, which we single out for convenience [35, Lemma 2].

Lemma A.4. *If $G(t) = t^{-\alpha} L(t)$ where $\alpha \geq 0$ and L is slowly varying, then for any given positive ϵ and c, there exist a and b such that*

$$\frac{G(\lambda t)}{G(t)} < \frac{b}{\lambda^{\alpha+\epsilon}} \quad \text{for } t \geq a,\ 0 < \lambda \leq c,$$
$$\frac{G(\lambda t)}{G(t)} < \frac{b}{\lambda^{\alpha-\epsilon}} \quad \text{for } t \geq a,\ \lambda \geq c.$$

In particular, for sufficiently large $t > 0$, we have

$$\frac{G(\lambda t)}{G(t)} \leq b(1/\lambda)^{\alpha \pm \epsilon}$$

for all $\lambda > 0$, where we define $x^{\alpha \pm \epsilon} := \max\left(x^{\alpha+\epsilon}, x^{\alpha-\epsilon}\right)$.

The next lemma concerns the convolution of distributions with regularly varying tails [7, Section VIII.8, Proposition].

Lemma A.5. *For two distributions F_1 and F_2 such that as $x \to \infty$*

$$1 - F_i(x) = x^{-\alpha} L_i(x)$$

with L_i slowly varying, the convolution $G = F_1 * F_2$ has a regularly varying tail such that

$$1 - G(x) \sim x^{-\alpha}(L_1(x) + L_2(x)).$$

Lemma A.6. *Let $\{X_{kn}: k \in \mathbb{N}\}$ be i.i.d. with $\mathbb{E}X_{1n} = 0$ for each $n \in \mathbb{N}$. If the family $\{|X_{1n}|^p : n \in \mathbb{N}\}$ is uniformly integrable for some $p > 1$, then as $n \to \infty$, we have*

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

$$S_n := \frac{1}{n}\sum_{k=1}^{n} X_{kn} \to 0$$

in probability.

Proof. For $M > 0$, let

$$Y_{kn} := X_{kn}\mathbf{1}_{[|X_{kn}|\le M]} - \mathbb{E}\big(X_{kn}\mathbf{1}_{[|X_{kn}|\le M]}\big), \qquad Z_{kn} := X_{kn}\mathbf{1}_{[|X_{kn}|> M]} - \mathbb{E}\big(X_{kn}\mathbf{1}_{[|X_{kn}|> M]}\big),$$
$$T_n := \frac{1}{n}\sum_{k=1}^{n} Y_{kn}, \qquad U_n := \frac{1}{n}\sum_{k=1}^{n} Z_{kn}.$$

By Markov's inequality,

$$\mathbb{P}\big(|T_n| \ge \delta\big) \le \frac{\mathbf{Var}\, Y_{1n}}{n\delta^2} \le \frac{4M^2}{n\delta^2},$$

and

$$\mathbb{P}\big(|U_n| \ge \delta\big) \le \frac{\mathbb{E}|U_n|^p}{\delta^p} \le \frac{\mathbb{E}|Z_{1n}|^p}{\delta^p}.$$

Thus, we have

$$\limsup_{n\to\infty} \mathbb{P}\big(|S_n| \ge 2\delta\big) \le \frac{1}{\delta^p}\sup_n \mathbb{E}|Z_{1n}|^p.$$

By the uniform integrability assumption, the right-hand side can be made arbitrarily small by increasing M. □

Appendix B. Proof of Lemma 3.1

First suppose $(\pi_j)_{j\in\mathbb{N}}$ converges to π_∞ in the weak topology on $\mathrm{Pr}(\mathrm{Pr}(\mathbb{R}))$. We want to show that $(\mathbf{X}^{(j)})_{j\in\mathbb{N}}$ converges in distribution to $\mathbf{X}^{(\infty)}$. By [19, Theorem 4.29], convergence in distribution of a sequence of random variables is equivalent to showing that for every $m > 0$ and all bounded continuous functions $f_1, \dots, f_m$, we have

$$\mathbb{E}\Big[f_1(X_1^{(j)}) \cdots f_m(X_m^{(j)})\Big] \to \mathbb{E}\Big[f_1(X_1^{(\infty)}) \cdots f_m(X_m^{(\infty)})\Big]$$

as $j \to \infty$. Rewriting the above using (3.3), we must show that as $j \to \infty$,

$$\int_{\mathrm{Pr}(\mathbb{R})} \left(\int_{\mathbb{R}^m} \prod_{i=1}^{m} f_i(x_i)\, \nu^{\otimes m}(d\mathbf{x})\right) \pi_j(d\nu) \longrightarrow \int_{\mathrm{Pr}(\mathbb{R})} \left(\int_{\mathbb{R}^m} \prod_{i=1}^{m} f_i(x_i)\, \nu^{\otimes m}(d\mathbf{x})\right) \pi_\infty(d\nu).$$

But this follows since $\nu \mapsto \int_{\mathbb{R}^m} \prod_{i=1}^m f_i(x_i)\, \nu^{\otimes m}(d\mathbf{x})$ is a bounded continuous function on $\mathrm{Pr}(\mathbb{R})$ with respect to the weak topology.

We now prove the reverse direction. We assume $(\mathbf{X}^{(j)})_{j\in\mathbb{N}}$ converges in distribution to $\mathbf{X}^{(\infty)}$ and must show that $(\pi_j)_{j\in\mathbb{N}}$ converges to π_∞.

In order to show this, we first claim that the family $(\pi_j)_{j\in\mathbb{N}}$ is tight. By [21, Theorem 4.10] (see also [10, Theorem A.6]), such tightness is equivalent to the tightness of the expected measures

$$\left(\int \nu^{\otimes\mathbb{N}}\, \pi_j(d\nu)\right)_{j\in\mathbb{N}}.$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

But these are just the distributions of the family $(\mathbf{X}^{(j)})_{j\in\mathbb{N}}$, which we have assumed converges in distribution. Hence its distributions are tight.

Let us now return to proving that $(\pi_j)_{j\in\mathbb{N}}$ converges to π_∞. Suppose to the contrary that this is not the case. Since the family $(\pi_j)_{j\in\mathbb{N}}$ is tight, by Prokhorov's theorem there must be another limit point of this family, $\tilde{\pi} \neq \pi_\infty$, and a subsequence $(j_n)_{n\in\mathbb{N}}$ such that

$$\pi_{j_n} \xrightarrow{\mathrm{w}} \tilde{\pi}$$

as $n \to \infty$. By the first part of our proof, this implies that $(\mathbf{X}^{(j_n)})_{n\in\mathbb{N}}$ converges in distribution to an exchangeable sequence with distribution $\int \nu^{\otimes\mathbb{N}}\, \tilde{\pi}(d\nu)$. However, by assumption we have that $(\mathbf{X}^{(j)})_{j\in\mathbb{N}}$ converges in distribution to $\mathbf{X}^{(\infty)}$, which has distribution $\int \nu^{\otimes\mathbb{N}}\, \pi_\infty(d\nu)$. Thus, it must be that

$$\int \nu^{\otimes\mathbb{N}}\, \tilde{\pi}(d\nu) = \int \nu^{\otimes\mathbb{N}}\, \pi_\infty(d\nu).$$

But [20, Proposition 1.4] tells us that the measure π in (3.3) is unique, contradicting $\tilde{\pi} \neq \pi_\infty$. Thus, it must be that $(\pi_j)_{j\in\mathbb{N}}$ converges to π_∞.

Appendix C. Multivariate stable laws

This section contains some basic definitions and properties related to multivariate stable distributions, to help familiarize readers with these concepts. The material in this section comes from the monograph [38] and also from [22].

Definition C.1. *A probability measure μ on $\mathbb{R}^k$ is said to be **(jointly) stable** if for all $a, b \in \mathbb{R}$ and two independent random variables X and Y with distribution μ, there exist $c \in \mathbb{R}$ and $v \in \mathbb{R}^k$ such that*

$$aX + bY \overset{d}{=} cX + v.$$

*If μ is symmetric, then it is said to be **symmetric stable**.*

Similarly to the one-dimensional case, there exists a constant $\alpha \in (0, 2]$ such that $c^\alpha = a^\alpha + b^\alpha$ for all a,b, which we call the **index of stability**. The distribution μ is multivariate Gaussian in the case $\alpha = 2$.

Theorem C.1. *Let $\alpha \in (0, 2)$. A random variable $\mathbf{X}$ taking values in $\mathbb{R}^k$ is symmetric stable if and only if there exists a finite measure Γ on the unit sphere $S_{k-1} = \{x \in \mathbb{R}^k : |x| = 1\}$ such that*

$$\mathbb{E} \exp\Big(i\langle \mathbf{t}, \mathbf{X}\rangle\Big) = \exp\Big(-\int_{S_{k-1}} |\langle \mathbf{t}, \mathbf{s}\rangle|^\alpha \Gamma(d\mathbf{s})\Big) \tag{C.1}$$

*for all $\mathbf{t} \in \mathbb{R}^k$. The measure Γ is called the **spectral measure** of $\mathbf{X}$, and the distribution is denoted by $\mathrm{S}_\alpha\mathrm{S}_k(\Gamma)$.*

In the case $k = 1$, the measure Γ is always of the form $c_1\delta_1 + c_{-1}\delta_{-1}$. Thus, the characteristic function reduces to the familiar form

$$\mathbb{E}e^{itX} = e^{-|\sigma t|^\alpha}.$$

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press

Acknowledgements

We thank François Caron and Juho Lee for sugggesting the paper [4] to us.

Funding information

P. Jung and H. Lee were funded in part by the National Research Foundation of Korea (NRF) grant NRF-2017R1A2B2001952. P. Jung, H. Lee, and J. Lee were funded in part by the NRF grant NRF-2019R1A5A1028324. H. Yang was supported by the Engineering Research Center Program, through the NRF, funded by the Korean government's Ministry of Science and ICT (NRF-2018R1A5A1059921), and also by the Institute for Basic Science (IBS-R029-C1).

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] ALDOUS, D. J. (1986). Classical convergence of triangular arrays, stable laws and Schauder's fixed-point theorem. *Adv. Appl. Prob.* **18**, 9–14.

[2] ARORA, S. *et al.* (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 8141–8150.

[3] DURRETT, R. (2019). *Probability: Theory and Examples*. Cambridge University Press.

[4] FAVARO, S., FORTINI, S. AND PELUCHETTI, S. (2020). *Stable behaviour of infinitely wide deep neural networks*. Preprint. Available at https://arxiv.org/abs/2003.00394.

[5] FAVARO, S., FORTINI, S. AND PELUCHETTI, S. (2021). *Deep stable neural networks: large-width asymptotics and convergence rates*. Preprint. Available at https://arxiv.org/abs/2108.02316.

[6] FAVARO, S., FORTINI, S. AND PELUCHETTI, S. (2022). Neural tangent kernel analysis of shallow α-stable ReLU neural networks. Preprint. Available at https://arxiv.org/abs/2206.08065.

[7] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. **2**. John Wiley, New York.

[8] FORTUIN, V. (2021). *Priors in Bayesian deep learning: a review*. Preprint. Available at https://arxiv.org/abs/2105.06868.

[9] FORTUIN, V. *et al.* (2021). *Bayesian neural network priors revisited*. Preprint. Available at https://arxiv.org/abs/2102.06571.

[10] GHOSAL, S. AND VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

[11] GHOSH, S., YAO, J. AND DOSHI-VELEZ, F. (2018). Structured variational learning of Bayesian neural networks with horseshoe priors. In *Proc. 35th International Conference on Machine Learning* (PMLR 80), eds J. Dy and A. Krause, Proceedings of Machine Learning Research, pp. 1744–1753.

[12] GHOSH, S., YAO, J. AND DOSHI-VELEZ, F. (2019). Model selection in Bayesian neural networks via horseshoe priors. *J. Mach. Learning Res.* **20**, 1–46.

[13] GLOROT, X. AND BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learning Res.* **9**, 249–256.

[14] GURBUZBALABAN, M., SIMSEKLI, U. AND ZHU, L. (2021). The heavy-tail phenomenon in SGD. In *Proc. 38th International Conference on Machine Learning* (PMLR 139), eds M. Meila and T. Zhang, Proceedings of Machine Learning Research, pp. 3964–3975.

[15] HE, K., ZHANG, X., REN, S. AND SUN, J. (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proc. 2015 IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers, pp. 1026–1034.

[16] HODGKINSON, L. AND MAHONEY, M. (2021). Multiplicative noise and heavy tails in stochastic optimization. In *Proc. 38th International Conference on Machine Learning* (PMLR 139), eds M. Meila and T. Zhang, Proceedings of Machine Learning Research, pp. 4262–4274.

[17] JACOT, A., HONGLER, C. AND GABRIEL, F. (2018). Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 8580–8589.

[18] JANTRE, S., BHATTACHARYA, S. AND MAITI, T. (2021). *Layer adaptive node selection in Bayesian neural networks: statistical guarantees and implementation details*. Preprint. Available at https://arxiv.org/abs/2108.11000.
[19] KALLENBERG, O. (2002). *Foundations of Modern Probability*. Springer, Cham.
[20] KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York.
[21] KALLENBERG, O. (2017). *Random Measures, Theory and Applications* Vol. **1**. Springer, Cham.
[22] KUELBS, J. (1973). A representation theorem for symmetric stable processes and stable measures on *H*. *Z. Wahrscheinlichkeitsth.* **26**, 259–271.
[23] LEE, H., YUN, E., YANG, H. AND LEE, J. (2022). Scale mixtures of neural network Gaussian processes. In *Proc. 10th International Conference on Learning Representations (ICLR 2022)*. Available at https://openreview.net/forum?id=YVPBh4k78iZ.
[24] LEE, J. *et al.* (2018). Deep neural networks as Gaussian processes. In *Proc. 6th International Conference on Learning Representations (ICLR 2018)*. Available at https://openreview.net/forum?id=B1EA-M-0Z.
[25] LEE, J. *et al.* (2020). Finite versus infinite neural networks: an empirical study. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 15156–15172.
[26] LEE, J., XIAO, L., SCHOENHOLZ, S. S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J. AND PENNINGTON, J. (2019). In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 8570–8581.
[27] LOUIZOS, C., ULLRICH, K. AND WELLING, M. (2017). Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 3288–3298.
[28] MARTIN, C. AND MAHONEY, M. (2019). Traditional and heavy tailed self regularization in neural network models. In *Proc. 36th International Conference on Machine Learning* (PMLR 97), eds K. Chaudhuri and R. Salakhutdinov, Proceedings of Machine Learning Research, pp. 4284–4293.
[29] MARTIN, C. H. AND MAHONEY, M. W. (2020). Heavy-tailed universality predicts trends in test accuracies for very large pre-trained deep neural networks. In *Proc. 2020 SIAM International Conference on Data Mining (SDM)*, Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 505–513.
[30] MATTHEWS, A. G. de G. *et al.* (2018). Gaussian process behaviour in wide deep neural networks. In *Proc. 6th International Conference on Learning Representations (ICLR 2018)*. Available at https://openreview.net/forum?id=H1-nGgWC-.
[31] MATTHEWS, A. G. de G., Hron, J., Turner, R. E. and Ghahramani, Z. (2017). Sample-then-optimize posterior sampling for Bayesian linear models. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference*. Available at http://approximateinference.org/2017/accepted/MatthewsEtAl2017.pdf.
[32] NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, Berlin, Heidelberg.
[33] NOVAK, R. *et al.* (2019). Bayesian deep convolutional networks with many channels are Gaussian processes. In *Proc. 7th International Conference on Learning Representations (ICLR 2019)*. Available at https://openreview.net/forum?id=B1g30j0qF7.
[34] OBER, S. W. AND AITCHISON, L. (2021). Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *Proc. 38th International Conference on Machine Learning* (PMLR 139), eds M. Meila and T. Zhang, Proceedings of Machine Learning Research, pp. 8248–8259.
[35] PITMAN, E. (1968). On the behaviour of the characteristic function of a probability distribution in the neighbourhood of the origin. *J. Austral. Math. Soc.* **8**, 423–443.
[36] ROBERTS, D. A., YAIDA, S. AND HANIN, B. (2022). *The Principles of Deep Learning Theory*. Cambridge University Press.
[37] ROYDEN, H. L. AND FITZPATRICK, P. (2010). *Real Analysis*, 4th edn. Macmillan, New York.
[38] SAMORODNITSKY, G. AND TAQQU, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, Boca Raton, FL.
[39] SHANBHAG, D. N. AND SREEHARI, M. (1977). On certain self-decomposable distributions. *Z. Wahrscheinlichkeitsth.* **38**, 217–222.
[40] TSUCHIDA, R., ROOSTA, F. AND GALLAGHER, M. (2019). *Richer priors for infinitely wide multi-layer perceptrons*. Preprint. Available at https://arxiv.org/abs/1911.12927.
[41] WAINWRIGHT, M. J. AND SIMONCELLI, E. P. (1999). Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, MIT Press, pp. 855–861.
[42] WENZEL, F. *et al.* (2020). How good is the Bayes posterior in deep neural networks really? Preprint. Available at https://arxiv.org/abs/2002.02405.
[43] YANG, G. (2019). Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Neural Information Processing Systems Foundation, San Diego, CA, pp. 9947–9960.

https://doi.org/10.1017/apr.2023.3 Published online by Cambridge University Press