Stable behaviour of infinitely wide deep neural networks

Stefano Favaro
stefano.favaro@unito.it
Department ESOMAS
University of Torino
and Collegio Carlo Alberto

Sandra Fortini
sandra.fortini@unibocconi.it
Department of Decision Sciences
Bocconi University

Stefano Peluchetti
speluchetti@cogent.co.jp
Cogent Labs

Abstract

We consider fully connected feed-forward deep neural networks (NNs) where weights and biases are independent and identically distributed as symmetric centered stable distributions. Then, we show that the infinite wide limit of the NN, under suitable scaling on the weights, is a stochastic process whose finite-dimensional distributions are multivariate stable distributions. The limiting process is referred to as the stable process, and it generalizes the class of Gaussian processes recently obtained as infinite wide limits of NNs (Matthews et al., 2018b). Parameters of the stable process can be computed via an explicit recursion over the layers of the network. Our result contributes to the theory of fully connected feed-forward deep NNs, and it paves the way to expand recent lines of research that rely on Gaussian infinite wide limits.

1 Introduction

The connection between infinitely wide deep feed-forward neural networks (NNs), whose parameters at initialization are independent and identically distributed (iid) as scaled and centered Gaussian distributions, and Gaussian processes (GPs) is well known (Neal, 1995; Der and Lee, 2006; Lee et al., 2018; Matthews et al., 2018a,b; Yang, 2019). Recently, this intriguing connection has been exploited in many exciting research directions, including: i) Bayesian inference for GPs arising from infinitely wide networks (Lee et al., 2018; Garriga-Alonso et al., 2019); ii) kernel regression for infinitely wide networks which are trained with continuous-time gradient descent via the neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019); iii) analysis of the properties of infinitely wide networks as function of depth via the information propagation framework (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019). It has been shown a substantial gap between finite NNs and their corresponding infinite (wide) GPs counterparts in terms of empirical performance, at least on some of the standard benchmarks applications. Moreover, it has been shown to be a difficult task to avoid undesirable empirical properties arising in the context of very deep networks. Given that, there exists an increasing interest in expanding GPs arising in the limit of infinitely wide NNs as a way forward to close, or even reverse, this empirical performance gap and to avoid, or slow down, pathological behaviors in very deep NN.

Let $\mathcal{N}(\mu, \sigma^2)$ denote the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$. Following the celebrated work of Neal (1995), we consider the shallow NN

$$f_i^{(1)}(x) = \sum_{j=1}^{I} w_{i,j}^{(1)} x_j + b_i^{(1)}$$

$$f_i^{(2)}(x, n) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} w_{i,j}^{(2)} \phi(f_j^{(1)}(x)) + b_i^{(2)},$$

where ϕ is a nonlinearity, $i = 1, \ldots, n$, $w_{i,j}^{(1)}, w_{i,j}^{(2)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_w^2)$, $b_i^{(1)}, b_i^{(2)} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2)$ and $x \in \mathbb{R}^I$ is the input. It follows that

$$f_i^{(1)}(x) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2_{f^{(1)}}(x)\right)$$

$$f_i^{(2)}(x, n) | f^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2_{f^{(2)}}(x, n)\right)$$

$$\sigma^2_{f^{(1)}}(x) = \sigma_b^2 + \sigma_w^2 \frac{1}{I} \sum_{j=1}^{I} x_j^2$$

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

$$\sigma^2_{f^{(2)}}(x,n) = \sigma_b^2 + \sigma_w^2 \frac{1}{n}\sum_{j=1}^{n} \phi(f_j^{(1)}(x))^2.$$

If x' is another input we obtain bivariate Gaussian distributions

$$(f_i^{(1)}(x), f_i^{(1)}(x')) \overset{\text{iid}}{\sim} \mathcal{N}_2\left(0, \Sigma_{f^{(1)}}(x,x')\right)$$
$$(f_i^{(2)}(x,n), f_i^{(2)}(x',n)) | f^{(1)} \overset{\text{iid}}{\sim} \mathcal{N}_2\left(0, \Sigma_{f^{(2)}}(x,x',n)\right),$$

where

$$\Sigma_{f^{(1)}}(x,x') = \begin{bmatrix} \sigma^2_{f^{(1)}}(x) & c_{f^{(1)}}(x,x') \\ c_{f^{(1)}}(x,x') & \sigma^2_{f^{(1)}}(x') \end{bmatrix}$$
$$\Sigma_{f^{(2)}}(x,x',n) = \begin{bmatrix} \sigma^2_{f^{(2)}}(x,n) & c_{f^{(2)}}(x,x',n) \\ c_{f^{(2)}}(x,x',n) & \sigma^2_{f^{(2)}}(x',n) \end{bmatrix}$$
$$c_{f^{(1)}}(x,x') = \sigma_b^2 + \sigma_w^2 \frac{1}{I}\sum_{j=1}^{I} x_j x_j'$$
$$c_{f^{(2)}}(x,x',n) = \sigma_b^2 + \sigma_w^2 \frac{1}{n}\sum_{j=1}^{n} \phi(f_j^{(1)}(x))\phi(f_j^{(1)}(x')).$$

Let $\xrightarrow{\text{a.s.}}$ denote the almost sure convergence. By the strong law of large numbers we know that, as $n \to +\infty$, one has

$$\frac{1}{n}\sum_{j=1}^{n} \phi(f_j^{(1)}(x))^2 \xrightarrow{\text{a.s.}} \mathbb{E}[\phi(f_1^{(1)}(x))^2]$$
$$\frac{1}{n}\sum_{j=1}^{n} \phi(f_j^{(1)}(x))\phi(f_j^{(1)}(x')) \xrightarrow{\text{a.s.}} \mathbb{E}[\phi(f_1^{(1)}(x))\phi(f_1^{(1)}(x'))],$$

from which one can conjecture that in the limit of infinite width the stochastic processes $f_i^{(2)}(x)$ are distributed as iid (over i) centered GP with kernel $K(x,x') = \sigma_b^2 + \sigma_w^2\mathbb{E}[\phi(f_1^{(1)}(x))\phi(f_1^{(1)}(x'))]$. Provided that the nonlinear function ϕ is chosen so that $\phi(f_1^{(1)}(x))$ has finite second moment, Matthews et al. (2018b) made rigorous this argument and extended it to deep NNs.

A key assumption underlying the interplay between infinite wide NNs and GPs is the finiteness of the variance of the parameters' distribution at initialization. In this paper we remove the assumption of finite variance by considering iid initializations based on stable distributions, which includes Gaussian initializations as a special case. We study the infinite wide limit of fully connected feed-forward NN in the following general setting: i) the NN is deep, namely the NN is composed of multiple layers; ii) biases and scaled weights are iid according to centered symmetric stable distributions; iii) the width of network's layers goes to infinity jointly on the layers, and not sequentially on each layer; iv) the convergence in distribution is established jointly for multiple inputs, namely the convergence concerns the class of finite dimensional distributions of the NN viewed as a stochastic process in function space. See Neal (1995) and Der and Lee (2006) for early works on NNs under stable initialization.

Within our setting, we show that the infinite wide limit of the NN, under suitable scaling on the weights, is a stochastic process whose finite-dimensional distributions are multivariate stable distributions (Samorodnitsky, 2017). This process is referred to as the stable process. Our result may be viewed as a generalization of the main result of Matthews et al. (2018b) to the context of stable distributions, as well as a generalization of results of Neal (1995) and Der and Lee (2006) to the context of deep NN. Our result contributes to the theory of fully connected feed-forward deep NNs, and it paves the way to extend the research directions i) ii) and iii) that rely on Gaussian infinite wide limits. The class of stable distributions is known to be especially relevant. Indeed while the contribution of each Gaussian weight vanishes as the width grows unbounded, some of the stable weights retains significant size, thus allowing them to represent "hidden features" (Neal, 1995).

The paper is structured as follows. Section 2 contains some preliminaries on stable distributions, whereas in Section 3 we define the class of feedforward NNs considered in this work. Section 4 contains our main result: as the width tends to infinity jointly on network's layers, the finite dimensional distributions of the NN converges to a multivariate stable distribution whose parameters are compute via a recursion over the layers. The convergence of the NN to the stable process then follows by finite-dimensional projections. In Section 5 we detail how our result extends previously established large width convergence results and comment on related work, whereas in Section 6 we discuss how our result applies to the research lines highlighted in i) ii) and iii) which relies on GP limits. In Section 7 we comment on future research directions. The Supplementary Material (SM) contains all the proofs (SM A,B,C), a preliminary numerical experiment on the recursion evaluation (SM D), an empirical investigation of the distribution of trained NN models' parameters (SM E). Code is available at https://github.com/stepelu/deep-stable.

2 Stable distributions

Let $\mathrm{St}(\alpha,\sigma)$ denote the symmetric centered stable distribution with stability parameter $\alpha \in (0,2]$ and scale parameter $\sigma > 0$, and let $S_{\alpha,\sigma}$ be a random variable distributed as $\mathrm{St}(\alpha,\sigma)$. That is, the characteristic function of $S_{\alpha,\sigma}\sim\mathrm{St}(\alpha,\sigma)$ is $\varphi_{S_{\alpha,\sigma}}(t)=\mathbb{E}[\mathrm{e}^{\mathrm{i}tS_{\alpha,\sigma}}]=\mathrm{e}^{-\sigma^{\alpha}|t|^{\alpha}}$. For any $\sigma>0$, a $S_{\alpha,\sigma}$ random variable with $0<\alpha<2$ has finite absolute moments $\mathbb{E}[|S_{\alpha,\sigma}|^{\alpha-\varepsilon}]$ for any $\varepsilon>0$, while $\mathbb{E}[|S_{\alpha,\sigma}|^{\alpha}]=+\infty$. Note that when $\alpha=2$, we have that $S_{2,\sigma}\sim\mathcal{N}(0,2\sigma^2)$. The random variable $S_{2,\sigma}$ has finite absolute moments of any order. For any $a\in\mathbb{R}$ we have the scaling identity $aS_{\alpha,1}\sim\mathrm{St}(\alpha,|a|)$.

We recall the definition of symmetric and centered multivariate stable distribution and its marginal distributions. First, let $\mathbb{S}^{k-1}$ be the unit sphere in $\mathbb{R}^k$. Let $\mathrm{St}_k(\alpha,\Gamma)$ denote the symmetric and centered k-dimensional stable distribution with stability $\alpha\in(0,2]$ and scale (finite) spectral measure Γ on $\mathbb{S}^{k-1}$, and let $\mathbf{S}_{\alpha,\Gamma}$ be a random vector of dimension $k\times 1$ distributed as $\mathrm{St}_k(\alpha,\Gamma)$. The characteristic function of $\boldsymbol{S}_{\alpha,\Gamma}\sim\mathrm{St}_k(\alpha,\Gamma)$ is

$$\varphi_{\boldsymbol{S}_{\alpha,\Gamma}}(\boldsymbol{t})=\mathbb{E}[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^T\boldsymbol{S}_{\alpha,\Gamma}}]=\exp\left\{-\int_{\mathbb{S}^{k-1}}|\boldsymbol{t}^T\boldsymbol{s}|^{\alpha}\Gamma(\mathrm{d}\boldsymbol{s})\right\}. \tag{1}$$

If $\boldsymbol{S}_{\alpha,\Gamma}\sim\mathrm{St}(\alpha,\Gamma)$ then the marginal distributions of $\boldsymbol{S}_{\alpha,\Gamma}$ are described as follows. Let $\mathbf{1}_r$ denote a vector of dimension $k\times 1$ with 1 in the r-the entry and 0 elsewhere. Then, the random variable corresponding to the r-th element of $\boldsymbol{S}_{\alpha,\Gamma}\sim\mathrm{St}(\alpha,\Gamma)$ can be defined as follows

$$\mathbf{1}_r^T\boldsymbol{S}_{\alpha,\Gamma}\sim\mathrm{St}(\alpha,\sigma(r)), \tag{2}$$

where

$$\sigma(r)=\left(\int_{\mathbb{S}^{k-1}}|\mathbf{1}_r^T\boldsymbol{s}|^{\alpha}\Gamma(\mathrm{d}\boldsymbol{s})\right)^{1/\alpha}. \tag{3}$$

The distribution $\mathrm{St}_k(\alpha,\Gamma)$ with characteristic function (1) allows for marginals which are not centered nor symmetric. However in the present work all the marginals will be centered and symmetric, and the spectral measure will often be a discrete measure, i.e., $\Gamma(\cdot)=\sum_{j=1}^{n}\gamma_j\delta_{\boldsymbol{s}_j}(\cdot)$for $n\in\mathbb{N}$, $\boldsymbol{s}_j\in\mathbb{S}^{k-1}$ and $\gamma_j\in\mathbb{R}$. In particular, under these specific assumptions, we have

$$\varphi_{\boldsymbol{S}_{\alpha,\Gamma}}(\boldsymbol{t})=\exp\left\{-\sum_{j=1}^{n}\gamma_j|\boldsymbol{t}^T\boldsymbol{s}_j|^{\alpha}\right\}.$$

See Samoradnitsky (2017) for a detailed account on $\boldsymbol{S}_{\alpha,\Gamma}\sim\mathrm{St}(\alpha,\Gamma)$.

3 Deep stable networks

We consider fully connected feed-forward NNs composed of $D\geq 1$ layers where each layer is of width $n\geq 1$. Let $w_{i,j}^{(l)}$ be the weights of the l-th layer, and assume that they are independent and identically distributed as $\mathrm{St}(\alpha,\sigma_w)$, a stable distribution with stability parameter $\alpha\in(0,2]$ and scale parameter $\sigma_w>0$. That is, the characteristic function of $w_{i,j}^{(l)}\sim\mathrm{St}(\alpha,\sigma_w)$ is

$$\varphi_{w_{i,j}^{(l)}}(t)=\mathbb{E}[\mathrm{e}^{\mathrm{i}tw_{i,j}^{(l)}}]=e^{-\sigma_w^{\alpha}|t|^{\alpha}}, \tag{4}$$

for any $i\geq 1$, $j\geq 1$ and $l\geq 1$. Let $b_i^{(l)}$ denote the biases of the l-th hidden layer, and assume that they are independent and identically distributed as $\mathrm{St}(\alpha,\sigma_b)$. That is, the characteristic function of the random variable $b_i^{(l)}\sim\mathrm{St}(\alpha,\sigma_b)$ is

$$\varphi_{b_i^{(l)}}(t)=\mathbb{E}[\mathrm{e}^{\mathrm{i}tb_i^{(l)}}]=e^{-\sigma_b^{\alpha}|t|^{\alpha}}, \tag{5}$$

for any $i\geq 1$ and $l\geq 1$. The random weights $w_{i,j}^{(l)}$ are independent of the biases $b_i^{(l)}$, for any $i\geq 1$, $j\geq 1$ and $l\geq 1$. That is,

$$(w_{i,j}^{(l)}+b_i^{(l)})\sim\mathrm{St}(\alpha,(\sigma_w^{\alpha}+\sigma_b^{\alpha})^{1/\alpha}).$$

Let $\phi:\mathbb{R}\to\mathbb{R}$ be a nonlinearity with a finite number of discontinuities and such that it satisfies the envelope condition

$$|\phi(s)|\leq(a+b|s|^{\beta})^{\gamma} \tag{6}$$

for every $s\in\mathbb{R}$, and for any parameter $a,b>0$, $\gamma<\alpha^{-1}$ and $\beta<\gamma^{-1}$. If $x\in\mathbb{R}^I$ is the input argument of the NN, then the NN is explicitly defined by means of

$$f_i^{(1)}(x,n)=f_i^{(1)}(x)=\sum_{j=1}^{I}w_{i,j}^{(1)}x_j+b_i^{(1)}, \tag{7}$$

and

$$f_i^{(l)}(x,n)=\frac{1}{n^{1/\alpha}}\sum_{j=1}^{n}w_{i,j}^{(l)}\phi(f_j^{(l-1)}(x,n))+b_i^{(l)} \tag{8}$$

for $l=2,\ldots,D$ and $i=1,\ldots,n$ in (7) and (8). The scaling of the weights in (8) will be shown to be the correct one to obtain non-degenerate limits as $n\to+\infty$.

4 Infinitely wide limits

We show that, as the width of the NN tends to infinity jointly on network's layers, the finite dimensional distributions of the network converge to a multivariate stable distribution whose parameters are compute via a suitable recursion over the network layers. Then, by combining this limiting result with standard arguments on finite-dimensional projections we obtain the large n limit of the stochastic process $(f_i^{(l)}(x^{(1)},n),\ldots,f_i^{(l)}(x^{(k)},n))_{i\geq 1}$ where $x^{(1)},\ldots,x^{(k)}$ are the inputs to the NN. In particular, let $\xrightarrow{w}$ denote the weak convergence. Then, we show that as $n\to+\infty$,

$$(f_i^{(l)}(x^{(1)},n),\ldots,f_i^{(l)}(x^{(k)},n))_{i\geq 1}\xrightarrow{w}\bigotimes_{i\geq 1}\mathrm{St}_k(\alpha,\Gamma(l)) \tag{9}$$

where $\bigotimes$ is the product measure. From now on k is the number of inputs, which is equal to the dimensionality of the finite dimensional distributions of interest for the stochastic processes $f_i^{(l)}$.Thorough the rest of the paper we assume that the assumptions introduced in Section 3 hold true. Hereafter we present a sketch of the proofs of our main result for a fixed index i and input x, and we defer to the SM for the complete proofs of our main results.

We start with a technical remark: in (7)-(8) the stochastic processes $f_i^{(l)}(x,n)$ are only defined for $i = 1,\dots,n$, while the limiting measure in (9) is the product measure on $i \geq 1$. This fact does not determine problems, as for each $\mathcal{L} \subset \mathbb{N}$ there is a n large enough such that for each $i \in \mathcal{L}$ the processes $f_i^{(l)}(x,n)$ are defined. In any case, the simplest solution consists in extending $f_i^{(l)}(x,n)$ from $i = 1,\dots,n$ to $i \geq 1$ in (7)-(8), and we will make this assumption in all the proofs.

4.1 Large width asymptotics: $k = 1$

We characterize the limiting distribution of $f_i^{(l)}(x,n)$ as $n \to \infty$ for a fixed i and input x. We show that, as $n \to +\infty$,

$$f_i^{(l)}(x,n) \xrightarrow{w} \mathrm{St}(\alpha, \sigma(l)), \tag{10}$$

where the parameter $\sigma(l)$ is computed through the recursion:

$$\sigma(1) = \big(\sigma_b^\alpha + \sigma_w^\alpha \sum_{j=1}^{I} |x_j|^\alpha\big)^{1/\alpha}$$

$$\sigma(l) = \big(\sigma_b^\alpha + \sigma_w^\alpha \mathbb{E}_{f\sim q^{(l-1)}}[|\phi(f)|^\alpha]\big)^{1/\alpha}$$

and $q(l) = \mathrm{St}(\alpha, \sigma(l))$ for each $l \geq 1$. The generalization of this result to $k \geq 1$ inputs is given in Section 4.2.

Proof of (10). The proof exploits exchangeability of the sequence $(f_i^{(l)}(n,x))_{i\geq 1}$, an induction argument on the layer's index l for the directing (random measure) of $(f_i^{(l)}(n,x))_{i\geq 1}$, and some technical lemmas that are proved in SM. Recall that the input is a real-valued vector of dimension I. By means of (4) and (5), for $i \geq 1$:

$$\begin{aligned}
&\varphi_{f_i^{(1)}(x)}(t) \\
&= \mathbb{E}[e^{\mathrm{i}tf_i^{(1)}(x)}] \\
&= \mathbb{E}\left[\exp\left\{\mathrm{i}t\left[\sum_{j=1}^{I} w_{i,j}^{(1)} x_j + b_i^{(1)}\right]\right\}\right] \\
&= \exp\left\{-(\sigma_w^\alpha \sum_{j=1}^{I} |x_j|^\alpha + \sigma_b^\alpha)|t|^\alpha\right\},
\end{aligned}$$

i.e.

$$f_i^{(1)}(x) \overset{\mathrm{d}}{=} S_{\alpha,\left(\sigma_w^\alpha \sum_{j=1}^{I} |x_j|^\alpha + \sigma_b^\alpha\right)^{1/\alpha}};$$

and for $l = 2,\dots,D$

$$\begin{aligned}
&\varphi_{f_i^{(l)}(x,n)\,|\,\{f_j^{(l-1)}(x,n)\}_{j=1,\dots,n}}(t) \\
&= \mathbb{E}[e^{\mathrm{i}tf_i^{(l)}(x,n)}\,|\,\{f_j^{(l-1)}(x,n)\}_{j=1,\dots,n}] \\
&= \mathbb{E}\left[\exp\left\{\mathrm{i}t\left[\frac{1}{n^{1/\alpha}}\sum_{j=1}^{n} w_{i,j}^{(l)}\phi(f_j^{(l-1)}(x,n)) + b_i^{(l)}\right]\right\}\right. \\
&\qquad \left.\Bigg|\ \{f_j^{(l-1)}(x,n)\}_{j=1,\dots,n}\right] \\
&= \exp\left\{-(\frac{\sigma_w^\alpha}{n}\sum_{j=1}^{n} |\phi(f_j^{(l-1)}(x,n))|^\alpha + \sigma_b^\alpha)|t|^\alpha\right\},
\end{aligned}$$

i.e.,

$$\begin{aligned}
&f_i^{(l)}(x,n)\,|\,\{f_j^{(l-1)}(x,n)\}_{j=1,\dots,n} \\
&\overset{\mathrm{d}}{=} S_{\alpha,\left(\frac{\sigma_w^\alpha}{n}\sum_{j=1}^{n} |\phi(f_j^{(l-1)}(x,n))|^\alpha + \sigma_b^\alpha\right)^{1/\alpha}}.
\end{aligned}$$

It comes from (8) that, for every fixed l and for every fixed n the sequence $(f_i^{(l)}(n,x))_{i\geq 1}$ is exchangeable. In particular, let $p_n^{(l)}$ denote the directing (random) probability measure of the exchangeable sequence $(f_i^{(l)}(n,x))_{i\geq 1}$. That is, by de Finetti representation theorem, conditionally to $p_n^{(l)}$ the $f_i^{(l)}(n,x)$'s are iid as $p_n^{(l)}$. Now, consider the induction hypothesis that $p_n^{(l-1)} \xrightarrow{w} q^{(l-1)}$ as $n \to +\infty$, with $q^{(l-1)}$ be $\mathrm{St}(\alpha, \sigma(l-1))$, and the parameter $\sigma(l-1)$ will be specified. Therefore,

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{\mathrm{i}tf_i^{(l)}(x,n)}] \\
&= \mathbb{E}\left[\exp\left\{-|t|^\alpha\left(\frac{\sigma_w^\alpha}{n}\sum_{j=1}^{n} |\phi(f_j^{(l-1)}(x,n))|^\alpha + \sigma_b^\alpha\right)\right\}\right] \\
&= e^{-|t|^\alpha\sigma_b^\alpha}\,\mathbb{E}\left[\exp\left\{-|t|^\alpha\frac{\sigma_w^\alpha}{n}\sum_{j=1}^{n} |\phi(f_j^{(l-1)}(x,n))|^\alpha\right\}\right] \\
&= e^{-|t|^\alpha\sigma_b^\alpha}\,\mathbb{E}\left[\left(\int \exp\left\{-|t|^\alpha\frac{\sigma_w^\alpha}{n}|\phi(f)|^\alpha\right\} p_n^{(l-1)}(\mathrm{d}f)\right)^n\right].
\end{aligned} \tag{11}$$

where the first equality comes from plugging in the definition of $f_i^{(l)}(x,n)$, rewriting $\mathbb{E}[\exp(\sum_{j=1}^{n}\cdots)]$ as $\mathbb{E}[\prod_{j=1}^{n}\exp(\cdots)] = \prod_{j=1}^{n}\mathbb{E}[\exp(\cdots)]$ due to independence, computing the characteristic function for each term, and re-arranging. therein, since $(f_I^{(l-1)}(n,x))_{i\geq 1}$ is exchangeable there exists (de Finetti theorem) a random probability measure $p_n^{(l-1)}$ such that conditionally to $p_n^{(l-1)}$ the $f_I^{(l-1)}(n,x)$ are iid as $p_n^{(l-1)}$ which explains (11).

Now, let $\xrightarrow{p}$ denote the convergence in probability. The following technical lemmas (Appendix A):

L1) for each $l \geq 2$ $\Pr[p_n^{(l-1)} \in I] = 1$, with $I = \{p : \int |\phi(f)|^\alpha p(\mathrm{d}f) < +\infty\}$;

L2) $\int |\phi(f)|^\alpha p_n^{(l-1)}(\mathrm{d}f) \xrightarrow{p} \int |\phi(f)|^\alpha q^{(l-1)}(\mathrm{d}f)$, as $n \to +\infty$;

L3) $\int |\phi(f)|^\alpha [1 - e^{-|t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha}] p_n^{(l-1)}(\mathrm{d}f) \xrightarrow{p} 0$, as $n \to +\infty$.

together with Lagrange theorem, are the main ingredients for proving (10) by studying the large n asymptotic behavior of (11). By combining (11) with lemma L1,

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}tf_i^{(l)}(x,n)}] = e^{-|t|^\alpha \sigma_b^\alpha} \mathbb{E}\Bigg[\mathbb{1}_{\{(p_n^{(l-1)} \in I)\}} \times \left(\int \exp\left\{ -|t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha \right\} p_n^{(l-1)}(\mathrm{d}f) \right)^n \Bigg].$$

By means of Lagrange theorem, there exists $\theta_n \in [0,1]$ such that

$$\begin{aligned} &\exp\left\{ -|t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha \right\} \\ &\quad = 1 - |t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha \\ &\qquad + |t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha \left(1 - \exp\left\{ -\theta_n |t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha \right\} \right). \end{aligned}$$

Now, since

$$\begin{aligned} 0 &\leq \int |\phi(f)|^\alpha [1 - e^{-\theta_n |t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha}] p_n^{(l-1)}(\mathrm{d}f) \\ &\leq \int |\phi(f)|^\alpha [1 - e^{-|t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha}] p_n^{(l-1)}(\mathrm{d}f), \end{aligned}$$

$$\begin{aligned} &\mathbb{E}[\mathrm{e}^{\mathrm{i}tf_i^{(l)}(x,n)}] \leq e^{-|t|^\alpha \sigma_b^\alpha} \mathbb{E}\Bigg[\mathbb{1}_{\{(p_n^{(l-1)} \in I)\}} \\ &\times \Bigg(1 - |t|^\alpha \frac{\sigma_w^\alpha}{n} \int |\phi(f)|^\alpha p_n^{(l-1)}(\mathrm{d}f) \\ &+ |t|^\alpha \frac{\sigma_w^\alpha}{n} \int |\phi(f)|^\alpha [1 - e^{-|t|^\alpha \frac{\sigma_w^\alpha}{n} |\phi(f)|^\alpha}] p_n^{(l-1)}(\mathrm{d}f) \Bigg)^n \Bigg]. \end{aligned}$$

Finally, recall the fundamental limit $\mathrm{e}^x = \lim_{n \to +\infty}(1 + x/n)^n$. This, combined with L2 and L3 leads to

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}tf_i^{(l)}(x,n)}] \to e^{-|t|^\alpha [\sigma_b^\alpha + \sigma_w^\alpha \int |\phi(f)|^\alpha q^{(l-1)}(\mathrm{d}f)]},$$

as $n \to +\infty$. That is, we proved that the large n limiting distribution of $f_i^{(l)}(x,n)$ is $\mathrm{St}(\alpha, \sigma(l))$, where we set

$$\sigma(l) = \left(\sigma_b^\alpha + \sigma_w^\alpha \int |\phi(f)|^\alpha q^{(l-1)}(\mathrm{d}f) \right)^{1/\alpha}$$

□

4.2 Large width asymptotics: $k \geq 1$

We establish the convergence in distribution of $(f_i^{(l)}(x^{(1)}, n), \ldots, f_i^{(l)}(x^{(k)}, n))$ as $n \to +\infty$ for a fixed i and k inputs $x^{(1)}, \ldots, x^{(k)}$. This result, combined with standard arguments on finite-dimensional projections, then establishes the convergence of the NN to the stable process. Precisely, we show that, as $n \to +\infty$, one has

$$(f_i^{(l)}(x^{(1)}, n), \ldots, f_i^{(l)}(x^{(k)}, n)) \xrightarrow{w} \mathrm{St}_k(\alpha, \Gamma(l)), \quad (12)$$

where the spectral measure $\Gamma(l)$ is computed through the recursion:

$$\Gamma(1) = \sigma_b^\alpha ||1||^\alpha \delta_{\frac{\mathbf{1}}{||\mathbf{1}||}} + \sigma_w^\alpha \sum_{j=1}^{I} ||\boldsymbol{x}_j||^\alpha \delta_{\frac{\boldsymbol{x}_j}{||\boldsymbol{x}_j||}} \quad (13)$$

$$\Gamma(l) = \sigma_b^\alpha ||1||^\alpha \delta_{\frac{\mathbf{1}}{||\mathbf{1}||}} + \sigma_w^\alpha \mathbb{E}_{f \sim q^{(l-1)}} [||\phi(f)||^\alpha \delta_{\frac{\phi(f)}{||\phi(f)||}}] \quad (14)$$

and $q(l) = \mathrm{St}_k(\alpha, \Gamma(l))$ for each $l \geq 1$, where $\boldsymbol{x}_j = [x_j^{(1)}, \ldots, x_j^{(k)}] \in \mathbb{R}^k$. Here (and in all the expressions involving the function δ) we make use of the notational assumption that if $\lambda = 0$ in $\lambda\delta_\bullet$, then $\lambda\delta_\bullet = 0$. This assumption allows us to avoid making the notation more cumbersome than necessary to explicitly exclude the case of $\phi(f) = 0$, for which $\phi(f)/||\phi(f)||$ is undefined. We omit the sketch of the proof of (12), as it is a step-by-step parallel of the proof of (10) with the added complexities due to the multivariate stable distributions. The reader can refer to the SM for the full proof.

4.3 Finite-dimensional projections

In Section 4.2 we obtained the convergence in law of $f_i^{(l)}(x^{(1)}, n), \ldots, f_i^{(l)}(x^{(k)}, n)$ for k inputs and a generic i to a multivariate Stable distribution. Let refer to this random vector as $f_i(x)$. Now, we derive the limiting behavior in law of $f_i(x)$ jointly over all $i = 1, \ldots$ (again for a given k-dimensional input). It is enough to study the convergence of $f_1(x), \ldots, f_n(x)$ for a generic $n \geq 1$. That is, it is enough to establish the convergence of the finite dimensional distributions (over i: we consider here $f_i(x)$ as random sequence over i). See Billingsley (1999) for details.

To establish the convergence of the finite dimensional distributions (over i) it then suffices to establish the convergence of linear combinations. More precisely, let $\boldsymbol{X} = [x^{(1)}, \dots, x^{(k)}] \in \mathbb{R}^{I \times k}$. We show that, as $n \to +\infty$,

$$(f_i^{(l)}(\boldsymbol{X}, n))_{i \geq 1} \xrightarrow{w} \bigotimes_{i \geq 1} \mathrm{St}_k(\alpha, \Gamma(l)),$$

by proving the large n asymptotic behavior of any finite linear combination of the $f_i^{(l)}(\boldsymbol{X}, n)$'s, for $i \in \mathcal{L} \subset \mathbb{N}$. Following the notation of Matthews et al. (2018b), let

$$T^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n) = \sum_{i \in \mathcal{L}} p_i [f_i^{(l)}(\boldsymbol{X}, n) - b_i^{(l)} \mathbf{1}].$$

Then, we write

$$\begin{aligned} &T^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n) \\ &= \sum_{i \in \mathcal{L}} p_i \left[\frac{1}{n^{1/\alpha}} \sum_{j=1}^{n} w_{i,j}^{(l)} (\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n)) \right] \\ &= \frac{1}{n^{1/\alpha}} \sum_{j=1}^{n} \gamma_j^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n), \end{aligned}$$

where

$$\gamma_j^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n) = \sum_{i \in \mathcal{L}} p_i w_{i,j}^{(l)} (\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n)).$$

Then,

$$\begin{aligned} &\varphi_{T^{(l)}(\mathcal{L}, p, \mathbf{X}, n) \,|\, \{f_j^{(l-1)}(\boldsymbol{X}, n)\}_{j=1,\dots,n}}(\boldsymbol{t}) \\ &= \mathbb{E}[e^{\mathrm{i}\boldsymbol{t}^T T^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n)} \,|\, \{f_j^{(l-1)}(\boldsymbol{X}, n)\}_{j=1,\dots,n}] \\ &= \prod_{j=1}^{n} \prod_{i \in \mathcal{L}} \mathrm{e}^{-\frac{p_i^\alpha \sigma_w^\alpha}{n} |\boldsymbol{t}^T (\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n))|^\alpha} \end{aligned}$$

That is,

$$T^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n) \,|\, \{f_j^{(l-1)}(\boldsymbol{X}, n)\}_{j=1,\dots,n} \overset{\mathrm{d}}{=} \boldsymbol{S}_{\alpha, \Gamma^{(l)}}$$

where

$$\Gamma^{(l)} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i \in \mathcal{L}} ||p_i \sigma_w (\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n))||^\alpha \delta_{\frac{\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n)}{||\phi \circ f_j^{(l-1)}(\boldsymbol{X}, n)||}}$$

Then, along lines similar to the proof of the large n asymptotics for the i-th coordinate, we have the following

$$\begin{aligned} &\mathbb{E}[\mathrm{e}^{\mathrm{i}\boldsymbol{t}^T T^{(l)}(\mathcal{L}, p, \boldsymbol{X}, n)}] \to \exp\Bigg\{ - \int \int_{\mathbb{S}^{k-1}} |\boldsymbol{t}^T \boldsymbol{s}|^\alpha \\ &\times \left(\sum_{i \in \mathcal{L}} ||p_i \sigma_w (\phi \circ f)||^\alpha \delta_{\frac{\phi \circ f}{||\phi \circ f||}} \right) (\mathrm{d}\boldsymbol{s}) q^{(l-1)}(\mathrm{d}f) \Bigg\} \end{aligned}$$

as $n \to +\infty$. This complete the proof of the limiting behaviour (9).

5 Related work

For the classical case of Gaussian weights and biases, and more in general for finite-variance iid distributions, the seminal work is that of Neal (1995). Here the author establishes, among other notable contributions, the connection between infinitely wide shallow (1 hidden layer) NNs and centered GPs. We reviewed the essence of it in Section 1.

This result is extended in Lee et al. (2018) to deep NNs where the width $n(l)$ of each layer l goes to infinity sequentially, starting from lowest layer, i.e. $n(1)$ to $n(D)$. The sequential nature of the limits reduces the task to a sequential application of the approach of Neal (1995). The computation of the GP kernel for each layer l involves a recursion, and the authors propose a numerical method to approximate the integral involved in each step of the recursion. The case where each $n(l)$ goes to infinity jointly, i.e. $n(l) = n$, is considered in Matthews et al. (2018a) under more restrictive hypothesis, which are relaxed in Matthews et al. (2018b). While this setting is most representative of a sequence of increasingly wide networks, the theoretical analysis is considerably more complicated as it does not reduce to a sequential application of the classical multivariate central limit theorem.

Going beyond finite-variance weight and bias distributions, Neal (1995) also introduced preliminary results for infinitely wide shallow NNs when weights and biases follow centered symmetric stable distributions. These results are refined in Der and Lee (2006) which establishes the convergence to a stable process, again in the setting of shallow NNs.

The present paper can be considered a generalization of the work of Matthews et al. (2018b) to the context of weights and biases distributed according to centered and symmetric stable distributions. Our proof follows different arguments from the proof of Matthews et al. (2018b), and in particular it does not rely on the central limit theorem for exchangeable sequences (Blum et al., 1958). Hence, since the Gaussian distribution is a special case of the stable distribution, our proof provides an alternative and self-contained proof to the result of Matthews et al. (2018b). It should be noted that our envelope condition (6) is more restrictive than the linear envelope condition of Matthews et al. (2018b). For the classical Gaussian setting the conditions on the activation function have been weakened in the work of Yang (2019).

Finally, there has been recent interest in using heavy-tailed distributions for gradient noise (Simsekli et al., 2019) and for trained parameter distributions (Martin and Mahoney, 2019). In particular, Martin and

Mahoney (2019) includes an empirical analysis of the parameters of pre-trained convolutional architectures (which we also investigate in SM E) supportive of heavy-tailed distributions. Results of this kind are compatible with the conjecture that stochastic processes arising from NNs whose parameters are heavy-tailed might be closer representations of their finite, high-performing, counterparts.

6 Future applications

6.1 Bayesian inference

Infinitely wide NNs with centered iid Gaussian initializations, and more in general finite variance centered iid initializations, gives rise to iid centered GPs at every layer l. Let us assume that weights and biases are distributed as in Section 1, and let us assume L layers ($L-1$ hidden layers). Each centered GPs is characterized by its covariance kernel function. Let us denote by $f^{(l)}$ such GPs for the layer $2 \leq l \leq L$. Over two inputs x and x' the distribution of $(f^{(l)}(x), f^{(l)}(x'))$ is characterized by the variances $q_x^{(l)} = \mathbb{V}[f^{(l)}(x)]$, $q_{x'}^{(l)} = \mathbb{V}[f^{(l)}(x')]$ and by the covariance $c_{x,x'}^{(l)} = \mathbb{C}[f^{(l)}(x), f^{(l)}(x')]$. These quantities satisfy

$$q_x^{(l)} = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[\phi\left(\sqrt{q_x^{(l-1)}}z\right)^2\right] \tag{15}$$

$$c_{x,x'}^{(l)} = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[\phi\left(\sqrt{q_x^{(l-1)}}z\right) \times \phi\left(\sqrt{q_{x'}^{(l-1)}}\left(\rho_{x,x'}^{(l-1)}z + \sqrt{1-(\rho_{x,x'}^{(l-1)})^2}z'\right)\right)\right] \tag{16}$$

where z and z' are independent standard Gaussian distributions $\mathcal{N}(0,1)$,

$$\rho^{(l)} = \frac{c_{x,x'}^{(l)}}{\sqrt{q_x^{(l)} q_{x'}^{(l)}}} \tag{17}$$

with initial conditions $q_x^{(1)} = \sigma_b^2 + \sigma_w^2 \|x\|^2$ and $c_{x,x'}^{(1)} = \sigma_b^2 + \sigma_w^2 \langle x, x' \rangle$.

To perform prediction via $\mathbb{E}[f^{(L)}(x^*)|x^*, \mathcal{D}]$, it is necessary to compute these recursions for all ordered pairs of data points x, x' in the training dataset $\mathcal{D}$, and for all pairs x^*, x with $x \in \mathcal{D}$. Lee et al. (2018) proposes an efficient quadrature solution that keeps the computational requirements manageable for an arbitrary activation ϕ.

In our setting, the corresponding recursion is defined by (13)-(14), which is a more computationally challenging problem with respect to the Gaussian setting. A sketch of a potential approach is as follows. Over the training data points and test points, $f^{(1)} \sim \mathrm{St}_k(\alpha, \Gamma(1))$ where k is equal to the size of training and test datasets combined. As $\Gamma(1)$ is a discrete measure exact simulations algorithms are available with a computational cost of $\mathcal{O}(I)$ per sample (Nolan, 2008). We can thus generate M samples $\widetilde{f}_j^{(1)}$, $j = 1, \ldots, M$, in $\mathcal{O}(IM)$, and use these to approximate $f^{(2)} \sim \mathrm{St}_k(\alpha, \Gamma(2))$ with $\mathrm{St}_k(\alpha, \widetilde{\Gamma}(2))$ with $\widetilde{\Gamma}(2)$ being

$$\widetilde{\Gamma}(2) = \sigma_b^\alpha ||1||^\alpha \delta_{\frac{\mathbf{1}}{||\mathbf{1}||}} + \sigma_w^\alpha \sum_{j=1}^{M} ||\phi(\widetilde{f}_j^{(1)})||^\alpha \delta_{\frac{\phi(\widetilde{f}_j^{(1)})}{||\phi(\widetilde{f}_j^{(1)})||}}$$

We can repeat this procedure by generating (approximate) random samples $\widetilde{f}_j^{(2)}$, with a cost of $\mathcal{O}(M^2)$, that in turn are used to approximate $\Gamma(3)$ and so on. In this procedure the errors can accumulate across the layers, as in Lee et al. (2018). This may be ameliorated by using quasi random number generators of Joe and Kuo (2008), as the sampling algorithms for multivariate stable distributions (Weron, 1996; Weron et al., 2010; Nolan, 2008) are all implemented as transformations of uniform distributions. The use of QRNG effectively defines a quadrature scheme for the integration problem. We report in the SM preliminary results regarding the numerical approximation of the recursion defined by (13)-(14).

This leaves us with the problem of computing a statistic of $f^{(L)}(x^*)|(x^*, \mathcal{D})$ or sampling from it, to perform prediction. Again, it could be beneficial to leverage on the discreteness of $\widetilde{\Gamma}(L)$. For example, these multivariate stable random variables can be expressed as suitable linear transformations of independent stable random variables (Samoradnitsky, 2017), and results expressing stable variables as mixtures of Gaussian variables are available in Samoradnitsky (2017).

6.2 Neural tangent kernel

In Section 6.1 we reviewed how the connection with GPs makes it possible to perform Bayesian inference directly on the limiting process. This corresponds to a "weakly-trained" regime of NNs, in the sense that the point (mean) predictions are equivalent to assuming an l_2 loss function, and fitting only a terminal linear layer to the training data, i.e. performing a kernel regression (Arora et al., 2019). The works of Jacot et al. (2018), Lee et al. (2019) and Arora et al. (2019) consider "fully-trained" NNs with l_2 loss and continuous-time gradient descent. Under Gaussian initialization assumptions it is shown that as the width of the NN goes to infinity, the point predictions corresponding by such fully trained networks are given again by a kernel regression but with respect to a different kernel, the neural tangent kernel.

In the derivation of the neural tangent kernel, one important point is that the gradients are not computed with respect to the standard model parameters, i.e. the the weights and biases entering the affine transforms. Instead they are "reparametrized gradients" which are computed with respect to parameters initialized as $\mathcal{N}(0,1)$, with any scaling (standard deviation) defined by parameter multiplication. It would thus be interesting to study whether a corresponding neural tangent kernel can be defined for the case of stable distributions with $0<\alpha<2$, and whether the parametrization of (7)-(8) is the appropriate one to do so.

6.3 Information propagation

The recursions (15)-(16) define the evolution over depth of the distribution of $f^{(l)}$ for two points x, x' when weights and biases are distributed as in Section 1. The information propagation framework studies the behavior of $q_x^{(l)}$ and $\rho_{x,x'}^{(l)}$ as $l \to +\infty$. It is shown in Poole et al. (2016) and Schoenholz et al. (2017) that the (σ_w, σ_b) positive quadrant is divided in two regions: a stable phase where $\rho_{x,x'}^{(l)} \to 1$ and a chaotic phase where $\rho_{x,x'}^{(l)}$ converges to a random variable (in the $\phi = \tanh$ case, in other cases the limiting processes may fail to exist). Thus in the stable phase $f^{(l)}$ is eventually perfectly correlated over inputs (and in most cases perfectly constant), while in the chaotic phase it is almost everywhere discontinuous. The work of Hayou et al. (2019) formalizes these results and investigates the case where (σ_w, σ_b) is on the curve separating the stable from the chaotic phase, i.e. the edge of chaos. Here it is shown that the behavior is qualitatively similar to that of the stable case, but with a lower rate of convergence with respect to depth. Thus in all cases the distribution of $f^{(l)}$ eventually collapse to degenerate and inexpressive distributions as depth increases.

In this context it would be interesting to study what is the impact of the use of stable distributions. All results mentioned above holds for the Gaussian case, which corresponds to $\alpha = 2$. Thus this further analysis would study the case $0 < \alpha < 2$, resulting in a triplet $(\sigma_w, \sigma_b, \alpha)$. Even though it seems hard to escape the course of depth under iid initializations, it might be that the use of stable distributions, with their not-uniformly-vanishing relevance at unit level (Neal, 1995), might slow down the rate of convergence to the limiting regime.

7 Conclusions

Within the setting of fully connected feed-forward deep NNs with weights and biases iid as centered and symmetric stable distributions, we proved that the infinite wide limit of the NN, under suitable scaling on the weights, is a stable process. This result contributes to the theory of fully connected feed-forward deep NNs, generalizing the work of Matthews et al. (2018b). We presented an extensive discussion on how our result can be used to extend recent lines of research which relies on GP limits.

On the theoretical side further developments of our work are possible. Firstly, Matthews et al. (2018b) performs an empirical analysis of the rates of convergence to the limiting process as function of depth with the respect to the MMD discrepancy (Gretton et al., 2012). Having proved the convergence of the finite dimensional distributions to multivariate stable distributions, the next step would be to establish the rate of convergence with respect to a metric of choice as function of the stability index α and depth l. Secondly, all the established convergence results (this paper included) concern the convergence of the finite dimensional distributions of the NN layers. For the countable case, which is the case of the components $i \geq 1$ in each layer, this is equivalent to the convergence in distribution of the whole process (over all the i) with respect to the product topology. However, the input space being R^I it is not countable. Hence, for a given i, the convergence of the finite dimensional distributions (i.e. over a finite collection of inputs) is not enough to establish the convergence in distribution of the stochastic process seen as a random function on the input (with respect to an appropriate metric). This is also the case for results concerning the convergence to GPs. It would thus be worthwhile to complete this theoretical line of research by establishing such result for any $0 < \alpha \leq 2$. As a side result, doing so is likely to provide estimates on the smoothness proprieties of the limiting stochastic processes.

8 Acknowledgements

We wish to thank the three anonymous reviewers and the meta reviewer for their valuable feedback. Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

References

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32*.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley-Interscience, 2nd edition.

Der, R. and Lee, D. D. (2006). Beyond gaussian processes: On the distributions of infinite networks. In *Advances in Neural Information Processing Systems*, pages 275–282.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. (2019). Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Hayou, S., Doucet, A., and Rousseau, J. (2019). On the impact of the activation function on deep neural networks training. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2672–2680.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8571–8580.

Joe, S. and Kuo, F. Y. (2008). Notes on generating sobol sequences. *ACM Transactions on Mathematical Software (TOMS)*, 29(1):49–57.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as gaussian processes. In *International Conference on Learning Representations*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems 32*.

Martin, C. H. and Mahoney, M. W. (2019). Traditional and heavy-tailed self regularization in neural network models. *arXiv preprint arXiv:1901.08276*.

Matthews, A. G. d. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018a). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018b). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto.

Nolan, J. P. (2008). An overview of multivariate stable distributions. *Online: http://academic2. american.edu/ jpnolan/stable/overview.pdf*.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016). Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pages 3360–3368.

Samoradnitsky, G. (2017). *Stable non-Gaussian random processes: stochastic models with infinite variance*. Routledge.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. (2017). Deep information propagation. In *International Conference on Learning Representations*.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. *arXiv preprint arXiv:1901.06053*.

Weron, R. (1996). On the chambers-mallows-stuck method for simulating skewed stable random variables. *Statistics & probability letters*, 28(2):165–171.

Weron, R. et al. (2010). Correction to: ”on the chambers–mallows–stuck method for simulating skewed stable random variables”. Technical report, University Library of Munich, Germany.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*.