Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, Julyan Arbel

Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions*

Mariia Vladimirova[1], Stéphane Girard[1], Hien Nguyen[2], Julyan Arbel[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France
[2]Department of Mathematics and Statistics, La Trobe University, Victoria 3086, Australia

Abstract

We propose the notion of *sub-Weibull* distributions, which are characterised by tails lighter than (or equally light as) the right tail of a Weibull distribution. This novel class generalises the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Sub-Weibull distributions are parameterized by a positive tail index θ and reduce to sub-Gaussian distributions for $\theta = 1/2$ and to sub-Exponential distributions for $\theta = 1$. A characterisation of the sub-Weibull property based on moments and on the moment generating function is provided and properties of the class are studied. An estimation procedure for the tail parameter is proposed and is applied to an example stemming from Bayesian deep learning.

1 Introduction and definition

Sub-Gaussian distributions, respectively sub-Exponential, are characterized by their tails being upper bounded by Gaussian, respectively Exponential, tails. More precisely, we say that a random variable X is sub-Gaussian, resp. sub-Exponential, if there exist positive constants a and b such that

$$\mathbb{P}(|X| \geq x) \leq a \exp(-bx^2), \quad \text{resp. } \mathbb{P}(|X| \geq x) \leq a \exp(-bx), \quad \text{for all } x > 0. \tag{1}$$

These properties have been intensely studied in the recent years due to their relationship with various fields of probability and statistics, including concentration, transportation and PAC-Bayes inequalities (Boucheron et al., 2013; Raginsky and Sason, 2013; van Handel, 2014; Catoni, 2007), the missing mass problem (Ben-Hamou et al., 2017), bandit problems (Bubeck and Cesa-Bianchi, 2012) and singular values of random matrices (Rudelson and Vershynin, 2010).

It is tempting to generalise (1) by considering the class of distributions satisfying

$$\mathbb{P}(|X| \geq x) \leq a \exp\left(-bx^{1/\theta}\right), \quad \text{for all } x > 0, \quad \text{for some } \theta, a, b > 0. \tag{2}$$

This is the goal of the present note. Since a *Weibull* random variable X on $\mathbb{R}_+$ is defined by a survival function, for $x > 0$,

$$\bar{F}(x) = \mathbb{P}(X \geq x) = \exp\left(-bx^{1/\theta}\right), \quad \text{for some } \ b > 0, \theta > 0, \tag{3}$$

we term a distribution satisfying (2) a *sub-Weibull* distribution (see Rinne, 2008, for a detailed account on the Weibull distribution).

*M. Vladimirova and J. Arbel are funded by Grenoble Alpes Data Institute, supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02). H. Nguyen is funded by the Australian Research Council grants: DE170101134 and DP180101192.

Corresponding author: mariia.vladimirova@inria.fr

Definition 1.1 (Sub-Weibull random variable)**.** *A random variable X, satisfying (2) for some positive a, b and θ, is called a sub-Weibull random variable with tail parameter θ, which is denoted by $X \sim subW(\theta)$.*

Interest in such heavier-tailed distributions than Gaussian or Exponential arises in our experience from their emergence in the field of Bayesian deep learning (Vladimirova et al., 2019). While writing this note, we were made aware of the preprint: Kuchibhotla and Chakrabortty (2018), which, independent of our work, also introduces sub-Weibull distributions but from a different perspective. The definition proposed by Kuchibhotla and Chakrabortty (2018) is based on Orlicz norm (building upon Wellner, 2017) and is equivalent to Definition 1.1. While Kuchibhotla and Chakrabortty (2018) focus on establishing tail bounds and rates of convergence for problems in high dimensional statistics, including covariance estimation and linear regression, under the sole sub-Weibull assumption, we are focused on proving sub-Weibull characterization properties. In addition, we illustrate their link with deep neural networks, not in the form of a model assumption as in Kuchibhotla and Chakrabortty (2018), but as a characterisation of the prior distribution of deep neural networks units. We further note that Hao et al. (2019) has also applied the notion of sub-Weibull distributions to the problem of constructing confidence bounds for bootstrapped estimators and cite Kuchibhotla and Chakrabortty (2018) and Vladimirova et al. (2019) as sources.

The outline of the paper is as follows. We define sub-Weibull distributions, and prove characteristic properties in Section 2, while some concentration properties are presented in Section 3. Finally, Section 4 provides an example of sub-Weibull distributions arising from Bayesian deep neural networks, as well as an estimation procedure for the tail parameter.

2 Sub-Weibull distributions: characteristic properties

Let X be a random variable. When the k-th moment of X exists, $k \geq 1$, we denote $\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k}$. The following theorem states different equivalent distribution properties, such as tail decay, growth of moments, and inequalities involving the moment generating function (MGF) of $|X|^{1/\theta}$. The proof of this result shows how to transform one type of information about the random variable into another. Our characterizations are inspired by the characterizations of sub-Gaussian and sub-Exponential distributions in Vershynin (2018).

Theorem 2.1 (Sub-Weibull equivalent properties)**.** *Let X be a random variable. Then the following properties are equivalent:*

1. *The tails of X satisfy*

$$\exists K_1 > 0 \quad \textit{such that} \quad \mathbb{P}(|X| \geq x) \leq 2\exp\left(-(x/K_1)^{1/\theta}\right) \quad \textit{for all } x \geq 0.$$

2. *The moments of X satisfy*

$$\exists K_2 > 0 \quad \textit{such that} \quad \|X\|_k \leq K_2 k^{\theta} \quad \textit{for all } k \geq 1.$$

3. *The MGF of $|X|^{1/\theta}$ satisfies*

$$\exists K_3 > 0 \quad \textit{such that} \quad \mathbb{E}\left[\exp\left((\lambda|X|)^{1/\theta}\right)\right] \leq \exp\left((\lambda K_3)^{1/\theta}\right)$$

   for all λ such that $0 < \lambda \leq 1/K_3$.

4. *The MGF of $|X|^{1/\theta}$ is bounded at some point, namely*

$$\exists K_4 > 0 \quad \textit{such that} \quad \mathbb{E}\left[\exp\left((|X|/K_4)^{1/\theta}\right)\right] \leq 2.$$

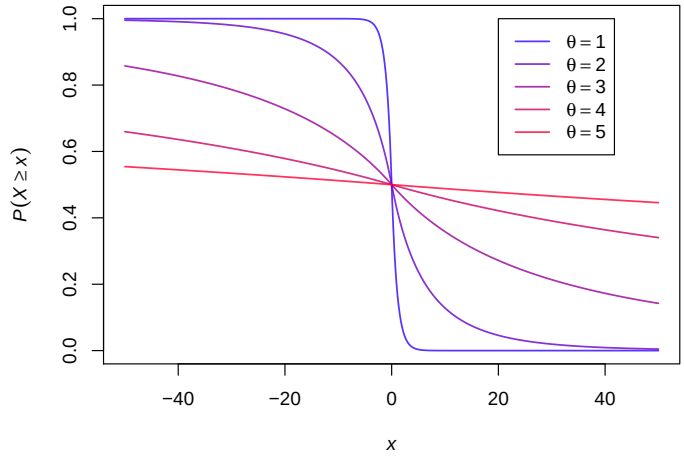

Figure 1: Illustration of sub-Weibull survival curves on $\mathbb{R}$ with varying tail parameters θ.

An illustration of symmetric sub-Weibull distributions is represented in Figure 1 for different values of the tail parameter θ. Since only the tail is relevant for the illustration, the sub-Weibull survival functions S_θ depicted here are obtained from the survival functions S_θ^W of Weibull random variables with shape parameter $1/\theta$, scale parameter 1. More specifically, only the part of S_θ^W to the right of the 95-th quantile is considered, and it is symmetrized around zero:

$$S_\theta(x) = 10\mathrm{e}^{-(x+\log(20)^\theta)^{1/\theta}} \quad \text{if } x \geq 0, \quad \text{and} \quad S_\theta(x) = 1 - 10\mathrm{e}^{-(-x+\log(20)^\theta)^{1/\theta}} \quad \text{if } x < 0.$$

Proof. In the proof we use the notation $\lesssim$ between two positive sequences $(a_k)_k$ and $(b_k)_k$, writing $a_k \lesssim b_k$, if there exists a constant $C > 0$ such that for all integer k, $a_k \leq Cb_k$.

1 ⇒ 2. Assume Property 1 holds. We have, for any $k \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[|X|^k\right] &\overset{(a)}{=} \int_0^\infty \mathbb{P}\left(|X|^k > x\right) \mathrm{d}x = \int_0^\infty \mathbb{P}\left(|X| > x^{1/k}\right) \mathrm{d}x \\
&\overset{(b)}{\leq} \int_0^\infty 2\exp\left(-(x/K_1)^{1/(k\theta)}\right) \mathrm{d}x = 2K_1 k\theta \int_0^\infty \mathrm{e}^{-u} u^{k\theta-1} \mathrm{d}u = 2K_1 k\theta\, \Gamma\left(k\theta\right) = 2K_1\, \Gamma\left(k\theta+1\right) \\
&\overset{(c)}{\lesssim} 2K_1 \left(k\theta+1\right)^{k\theta+1} = 2K_1 \left(k\theta+1\right)\left(k\theta+1\right)^{k\theta},
\end{aligned}$$

where (a) is by the so-called integral identity for $|X|^k$, (b) is by Property 1, and (c) comes from Stirling's formula, yielding $\Gamma(u) \lesssim u^u$. Taking the k-th root of the expression above yields

$$\|X\|_k \lesssim (2K_1)^{1/k} (k\theta+1)^{1/k} \left(k\theta+1\right)^\theta \lesssim k^\theta,$$

which is to Property 2.

2 ⇒ 3. Assume Property 2 holds. Recalling the Taylor series expansion of the exponential function, we obtain

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}|X|^{1/\theta}\right)\right] = \mathbb{E}\left[1 + \sum_{k=1}^\infty \frac{(\lambda^{1/\theta}|X|^{1/\theta})^k}{k!}\right] = 1 + \sum_{k=1}^\infty \frac{\lambda^{k/\theta}\mathbb{E}[|X|^{k/\theta}]}{k!}.$$

Property 2 guarantees that $\mathbb{E}[|X|^{k/\theta}] \leq K_2^{k/\theta}(k/\theta)^k$ for some K_2 and for any $k \geq \theta$. There exist only a finite number of integers k that are less than θ, thus there exists a constant $\tilde{K}_2$ such that $\mathbb{E}[|X|^{k/\theta}] \leq \tilde{K}_2^{k/\theta}(k/\theta)^k$ for integers $k \geq \theta$. Defining $K_2' = \max\{K_2, \tilde{K}_2\}$ implies $\mathbb{E}[|X|^{k/\theta}] \leq K_2'^{k/\theta}(k/\theta)^k$ for all integers $k \geq 1$. By Stirling's approximation we have $k! \gtrsim (k/\mathrm{e})^k$. Substituting these two bounds, we get

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}|X|^{1/\theta}\right)\right] \lesssim 1 + \sum_{k=1}^\infty \frac{\lambda^{k/\theta} K_2'^{k/\theta}(k/\theta)^k}{(k/\mathrm{e})^k} = \sum_{k=0}^\infty K_2'^{k/\theta}(\mathrm{e}\lambda^{1/\theta}/\theta)^k = \frac{1}{1 - \mathrm{e}(K_2'\lambda)^{1/\theta}/\theta}$$

provided that $\mathrm{e}(K_2'\lambda)^{1/\theta}/\theta < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the inequality $\frac{1}{1-x} \leq \mathrm{e}^{2x}$, which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta}X^{1/\theta}\right)\right] \leq \exp\left(2\mathrm{e}(K_2'\lambda)^{1/\theta}/\theta\right)$$

for all λ satisfying $0 < \lambda \leq \frac{1}{K_2'}\left(\frac{\theta}{2\mathrm{e}}\right)^{\theta}$ and some positive K_2' defined above. This yields Property 3 with $K_3 = K_2'(2\mathrm{e}/\theta)^{\theta}$.

3 $\Rightarrow$ **4.** Assume Property 3 holds. Take $\lambda = 1/K_4$, where $K_4 \geq K_3/(\ln 2)^{\theta}$. This yields Property 4.

4 $\Rightarrow$ **1.** Assume Property 4 holds. Then, by Markov's inequality and Property 3, we obtain

$$\mathbb{P}\big(|X| > x\big) = \mathbb{P}\left(\mathrm{e}^{(|X|/K_4)^{1/\theta}} > \mathrm{e}^{(x/K_4)^{1/\theta}}\right) \leq \frac{\mathbb{E}[\mathrm{e}^{(|X|/K_4)^{1/\theta}}]}{\mathrm{e}^{(x/K_4)^{1/\theta}}} \leq 2\mathrm{e}^{-(x/K_4)^{1/\theta}}.$$

This proves Property 1 with $K_1 = K_4$. □

Informally, the tails of a $subW(\theta)$ distribution are dominated by (i.e. decay at least as fast as) the tails of a Weibull variable with shape parameter[1] equal to $1/\theta$. Sub-Gaussian and sub-Exponential variables, which are commonly used, are special cases of sub-Weibull random variables with tail parameter $\theta = 1/2$ and $\theta = 1$, respectively, see Table 1. Sub-Gaussian distributions are

Table 1: Sub-Gaussian, sub-Exponential and sub-Weibull distributions comparison in terms of tail $P(|X| \geq x)$ and moment condition, with K_1 and K_2 some positive constants. The first two are a special case of the last with $\theta = 1/2$ and $\theta = 1$, respectively.

Distribution	Tails	Moments
Sub-Gaussian	$\mathbb{P}(\lvert X\rvert \geq x) \leq 2\mathrm{e}^{-(x/K_1)^2}$	$\lVert X\rVert_k \leq K_2\sqrt{k}$
Sub-Exponential	$\mathbb{P}(\lvert X\rvert \geq x) \leq 2\mathrm{e}^{-x/K_1}$	$\lVert X\rVert_k \leq K_2 k$
Sub-Weibull	$\mathbb{P}(\lvert X\rvert \geq x) \leq 2\mathrm{e}^{-(x/K_1)^{1/\theta}}$	$\lVert X\rVert_k \leq K_2 k^{\theta}$

sub-Exponential as well. Such inclusion properties are generalized to sub-Weibull distributions in the following proposition.

Proposition 2.1 (Inclusion)**.** *Let θ_1 and θ_2 such that $0 < \theta_1 \leq \theta_2$ be two sub-Weibull tail parameters. The following inclusion holds*

$$subW(\theta_1) \subset subW(\theta_2).$$

Proof. For $X \sim subW(\theta_1)$, there exists some constant $K_2 > 0$ such that for all $k > 0$, $\|X\|_k \leq K_2 k^{\theta_1}$. Since $k^{\theta_1} \leq k^{\theta_2}$ for all $k \geq 1$, this yields $\|X\|_k \leq K_2 k^{\theta_2}$, which by definition implies $X \sim subW(\theta_2)$. □

Let a random variable X follow a sub-Weibull distribution with tail parameter θ. Due to the property of inclusion from Proposition 2.1, the sub-Weibull definition provides an upper bound for the tail. In order to address the question of a lower bound on the tail parameter of some sub-Weibull distribution, we rely on an optimal tail parameter for that distribution through the use of the moment Property 2 of Theorem 2.1. To this aim, we introduce the notation of asymptotic equivalence between two positive sequences $(a_k)_k$ and $(b_k)_k$ as by $a_k \asymp b_k$, defined by the existence of constants $D \geq d > 0$ such that

$$d \leq \frac{a_k}{b_k} \leq D, \quad \text{for all } k \in \mathbb{N}. \tag{4}$$

We can now introduce the notion of optimal sub-Weibull tail coefficient, and provide a moment-based condition for optimality to hold.

[1] Weibull distributions are commonly parameterized by a shape parameter κ. Here we use instead $\theta = 1/\kappa$ for the convenience that the larger the tail parameter θ, the heavier the tails of the sub-Weibull distribution.

Proposition 2.2 (Optimal sub-Weibull tail coefficient and moment condition)**.** *Let $\theta > 0$ and let X be a random variable satisfying the following asymptotic equivalence on moments*

$$\|X\|_k \asymp k^\theta.$$

Then X is sub-Weibull distributed with optimal tail parameter θ, in the sense that for any $\theta' < \theta$, X is not subW(θ').

Proof. By the upper bound of the asymptotic equivalence assumption on moments, X satisfies Property 2 of Theorem 2.1, so $X \sim subW(\theta)$. Let $\theta' < \theta$. By the lower bound of the asymptotic equivalence assumption on moments, there does not exist any constant K_2 such that $\|X\|_k \leq K_2 k^{\theta'}$ for any $k \in \mathbb{R}$, so X is not sub-Weibull with tail parameter θ'. This concludes the proof. □

When sub-Gaussian and sub-Exponential properties are studied, it is often assumed that the distribution is centered. However, note that these properties, as well as the sub-Weibull property, are tail properties, and as such, they are not affected by translation. Non-centered random variables can always be centered by subtracting the mean without affecting their tail property. In particular, variable centering does not change the optimal tail parameter of a sub-Weibull distribution. Scaling procedure does not influence the tail neither: multiplication by some fixed number will change the coefficient K_1 in the tail Property 1 of Theorem 2.1, but not the tail parameter.

It is known that the product of sub-Gaussian random variables is sub-exponential (cf. Vershynin 2018, Lemma 2.7.7). By the same virtue, we have the fact that the class of sub-Weibull random variables is closed under multiplication and addition.

Proposition 2.3 (Closure under multiplication and addition)**.** *Let X and Y be sub-Weibull random variables with tail parameters θ_1 and θ_2, respectively. Then, XY and $X + Y$ are sub-Weibull with respective tail parameters $\theta_1 + \theta_2$ and $\max(\theta_1, \theta_2)$.*

Proof. Let us focus on the closure under addition property. Property 2 in Theorem 2.1 and the triangular inequality entail that there exist $K_{2,1} > 0$ and $K_{2,2} > 0$ such that, for all $k \geq 1$,

$$\|X + Y\|_k \leq \|X\|_k + \|Y\|_k \leq K_{2,1} k^{\theta_1} + K_{2,2} k^{\theta_2} \leq (K_{2,1} + K_{2,2}) k^{\max(\theta_1, \theta_2)}.$$

The conclusion follows from Theorem 2.1. □

Furthermore, we may establish that the class of sub-Weibull random variables is larger than the class of bounded random variables but is smaller than the class of random variables with finite p-th moment, for every $p \in [0, \infty)$. This result is comparable to Remark 2.7.14 of Vershynin (2018), which establishes the same relationship regarding the class of sub-Gaussian random variables.

3 Concentration properties

In this section, we focus on theoretical results about the sum of sub-Weibull random variables, including concentration properties.

Proposition 3.1 (Sum of sub-Weibull random variables)**.** *Let $X_1, \ldots, X_n$ be sub-Weibull random variables with tail parameter θ. Then the sum $\sum_{i=1}^n X_i$ is sub-Weibull with tail parameter θ.*

Proof. Using Property 2 in Theorem 2.1 combined with Minkowski's inequality yields (for any $k \geq 1$):

$$\left\| \sum_{i=1}^n X_i \right\|_k \leq \sum_{i=1}^n \|X_i\|_k \leq \sum_{i=1}^n K_{2,i} k^\theta = K_{2,\Sigma} k^\theta, \tag{5}$$

where $K_{2,\Sigma} = \sum_{i=1}^n K_{2,i}$. It implies that the sum $X_1 + \cdots + X_n$ satisfies the sub-Weibull property with tail parameter at most θ. In addition, if $X_1, \ldots, X_n$ are from the same distribution, i.e. $\|X_i\|_k \leq K_2 k^\theta$ for all $i \in \{1, \ldots, n\}$ with some constant $K_2 > 0$, then $K_{2,\Sigma} = nK_2$. □

Using Proposition 2.3, and the proof of Property 2 in Theorem 2.1, we have the following Hoeffding-type concentration inequality regarding the sum $\sum_{i=1}^n X_i$.

Corollary 3.1 (Concentration of the sum). *Let that $X_1, \ldots, X_n$ be identically distributed sub-Weibull random variables with tail parameter θ. Then, for all $x \geq nK_\theta$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq x\right) \leq \exp\left(-\left(\frac{x}{nK_\theta}\right)^{1/\theta}\right),$$

for some constant K_θ dependent on θ.

Proof. For all real $k \geq 1$, the bound (5) in the proof of Proposition 3.1 yields

$$\mathbb{E}\left(\left|\sum_{i=1}^{n} X_i\right|^k\right) \leq \left[K_{2,\Sigma} k^\theta\right]^k = \left[nK_2 k^\theta\right]^k,$$

and thus, by Markov's inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right|^k \geq t^k\right) \leq \frac{\mathbb{E}\left|\sum_{i=1}^{n} X_i\right|^k}{t^k} \leq \frac{\left(nK_2\right)^k \left(k^\theta\right)^k}{t^k}.$$

Choose t so that $\exp(-k) = \left(nK_2\right)^k \left(k^\theta\right)^k / t^k$. Then, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq enK_2 k^\theta\right) \leq \exp(-k). \tag{6}$$

Let $K_\theta = eK_2$ and notice that $nK_\theta k^\theta = x$ implies that $k = \left(\frac{x}{nK_\theta}\right)^{1/\theta}$ and thus, for $x \geq nK_\theta$ (since $k \geq 1$):

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq x\right) \leq \exp\left(-\left(\frac{x}{nK_\theta}\right)^{1/\theta}\right),$$

which is the expected result. □

Let us remark that, alternatively, letting $\exp(-k) = \alpha$ in (6) implies, for $0 < \alpha < 1/e$, the confidence statement

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \leq nK_\theta \left(\log\frac{1}{\alpha}\right)^\theta\right) \geq 1 - \alpha.$$

Under the same conditions as Proposition 3.1, under the restriction that $\theta \leq 1$, Boucheron et al. (2013) proposed the inequality

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq x\right) \leq K_\theta \exp\left(-\frac{1}{K_\theta} \min\left\{\frac{x^2}{n}, \frac{x^{1/\theta}}{n^{(1-\theta)/\theta}}\right\}\right), \tag{7}$$

where K_θ is again a constant that only depends on θ. We observe that $(1-\theta)/\theta < 1/\theta$ and thus the bound of (7) converges faster to zero than that of Corollary 3.1 in all cases where they are comparable. A version of this result also appears in Lemma 3.5 of Adamczak et al. (2011). Further results regarding the concentration of the sum of sub-Weibull random variables are available in Kuchibhotla and Chakrabortty (2018) and Hao et al. (2019). See also Bakhshizadeh et al. (2020) for further concentration results regarding heavy-tailed distributions.

4 Application to Bayesian neural networks

This section gives an example of sub-Weibull variables that arise in deep learning, more specifically in the context of Bayesian deep neural networks, as originally described in Vladimirova et al. (2019).

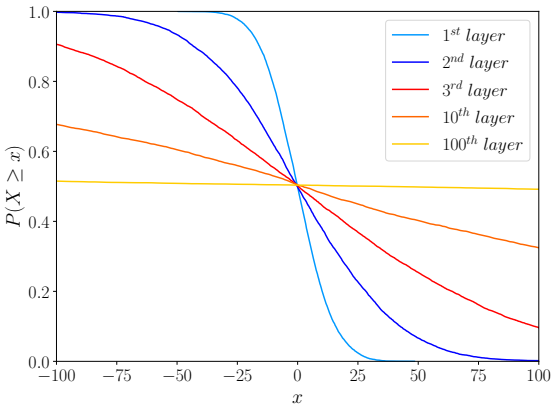

Figure 2: Illustration of layers $\ell = 1, 2, 3, 10$ and 100 hidden units marginal prior distributions tails. According to Theorem 2.1 of Vladimirova et al. (2019), they correspond respectively to $subW(1/2)$, $subW(1)$, $subW(3/2)$, $subW(5)$ and $subW(50)$.

4.1 Distribution at the level of units

We first describe so-called fully connected neural networks. Neural networks are hierarchical models made of layers: an input, several hidden layers and an output. Each layer following the input layer consists of units which are linear combinations of previous layer units transformed by a nonlinear function, often referred to as the nonlinearity or activation function denoted by $\phi : \mathbb{R} \to \mathbb{R}$. Given an input $\boldsymbol{x} \in \mathbb{R}^N$ (for instance an image made of N pixels) the ℓ-th hidden layer consists of two vectors whose size is called the width of layer, denoted by H_ℓ. The vector of units before application of the non-linearity is called pre-nonlinearity, and is denoted by $\boldsymbol{g}^{(\ell)} = \boldsymbol{g}^{(\ell)}(\boldsymbol{x})$, while the vector obtained after element-wise application of ϕ is called post-nonlinearity and is denoted by $\boldsymbol{h}^{(\ell)} = \boldsymbol{h}^{(\ell)}(\boldsymbol{x})$. More specifically, these vectors are defined as

$$\boldsymbol{g}^{(\ell)}(\boldsymbol{x}) = \boldsymbol{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}), \quad \boldsymbol{h}^{(\ell)}(\boldsymbol{x}) = \phi(\boldsymbol{g}^{(\ell)}(\boldsymbol{x})), \tag{8}$$

where $\boldsymbol{W}^{(\ell)}$ is a weight matrix of dimension $H_\ell \times H_{\ell-1}$ including a bias vector, with the convention that $H_0 = N$, the input dimension.

Neural networks perform state-of-the-art results in many areas. Researchers aim at better understanding the driving mechanisms behind their effectiveness. In particular, the study of the neural networks distributional properties through Bayesian analysis, where weights are assumed to follow a prior distribution, has attracted a lot of attention in the recent years. The main focus of research in the field shows that a *Bayesian deep neural network converges in distribution to a Gaussian process when the width of all the layers goes to infinity.* See for instance Matthews et al. (2018); Lee et al. (2018) for the main proofs. Further research such as Hayou et al. (2019) builds upon the limiting Gaussian process property of neural networks, as well as on the notion of *edge of chaos* developed by Schoenholz et al. (2017), in order to devise novel architecture rules for neural networks.

In contrast, Theorem 2.1 of Vladimirova et al. (2019) shows that the *non asymptotic* (i.e. for finite width neural networks) prior distribution of units from the ℓ-th layer (both before and after activation, $\boldsymbol{g}^{(\ell)}$ and $\boldsymbol{h}^{(\ell)}$) induced by a standard Gaussian prior on the weights $\boldsymbol{W}^{(\ell)}$, is *sub-Weibull with tail parameter* $\ell/2$. Therefore, the deeper the layer is, the heavier-tailed is the units distribution. This result puts into perspective the infinite width Gaussian process property which might be far from holding for real world neural networks.

In order to illustrate the sub-Weibull property of units prior distributions, we performed the following experiment with a deep neural network of 100 layers. We considered a Bayesian neural network with independent standard Gaussian priors on the weights, where layers are composed of $H_\ell = 1000 - 10(\ell - 1)$ hidden units, $\ell = 1, \ldots, 100$, with the ReLU nonlinearity ϕ defined by $\phi(x) = \max(0, x)$ (see (8)). The input vector $\boldsymbol{x}$ contains 10^4 numbers sampled from independent standard Gaussian. In order to evaluate the units prior distributions, we used a Monte Carlo

approximation, where the input vector $\boldsymbol{x}$ was kept fixed, and the weights were sampled from independent standard Gaussian $n = 10^5$ times. Figure 2 illustrates the survival function of pre-nonlinearity hidden units $\boldsymbol{g}^{(\ell)}$ for layers $\ell = 1, 2, 3, 10$ and 100. This indicates that the prior tails of the units get heavier when they originate from deeper layers, in accordance with the main result of Vladimirova et al. (2019).

4.2 Tail parameter estimation

We conclude this section by suggesting a statistical estimation procedure for the tail parameter θ. We adopt an approach that relies on Weibull random variables with tail parameter $\theta > 0$ and scale parameter $\lambda > 0$, which are $subW(\theta)$. The cumulative distribution function is $F(y) = 1 - \mathrm{e}^{-(y/\lambda)^{1/\theta}}$ for all $y \geq 0$, with corresponding quantile function $q(t) = \lambda \left(-\log(1-t)\right)^{\theta}$. Taking the logarithm, we obtain the following expression

$$\log q(t) = \theta \log\log \frac{1}{1-t} + \log \lambda,$$

showing an affine relationship involving the log quantiles and θ as slope parameter. This suggests a simple estimation procedure for θ by linear regression. More specifically, assume we have n iid observations $Y_1, \ldots, Y_n$, and denote the order statistics by $Y_{i,n}$. Consider a fraction $k \leq n$ of the largest observations, that are the order statistics $Y_{n-i+1,n}$ for $1 \leq i \leq k$. These are approximating the quantiles of order $\frac{n-i}{n} = 1 - \frac{i}{n}$, respectively. As a consequence, we consider the slope coefficient in the regression of $\log Y_{n-i+1,n}$ against $\log\log(n/i)$, for $1 \leq i \leq k$, as an estimate of θ. This estimator was first introduced by Gardes and Girard (2008) in the context of Weibull-tail distributions. We visualize the alignment in a quantile-quantile plot in log-log scale in Figure 3.

We considered the following experiment in order to illustrate the estimation procedure. We have built neural networks of 10 hidden layers, with $H_\ell = H$ hidden units on each layers, and made H vary in $\{1, 3, 10\}$. We used a fixed input $\boldsymbol{x}$ of size 10^4, which can be thought of as an image of dimension 100×100. This input was sampled once for all with standard Gaussian entries. In order to obtain samples from the prior distribution of the neural network units, we have sampled the weights from independent centered Gaussians with variance set to 2, from which units were obtained by forward evaluation with the ReLU non-linearity. This process was iterated $n = 10^5$ times and the number of largest observations selected is $k = 10^3$. We show in Figure 3 the θ estimates obtained by using pre-nonlinearities $\boldsymbol{g}^{(\ell)}$ of layers ℓ in $\{1, 3, 10\}$ as data, for H varying in $\{1, 3, 10\}$. We can see that the theoretical result from Vladimirova et al. (2019) that states that $\theta = \ell/2$ is well in line with the estimates obtained with networks of width $H = 1$. When the width increases, the estimates for θ tend to decrease, narrowing the gap to the lower bound of $1/2$. These results illustrate the large width results by Matthews et al. (2018); Lee et al. (2018), which state that in the large width regime, the unit distribution tends to a Gaussian, thus with a limiting θ parameter of $1/2$.

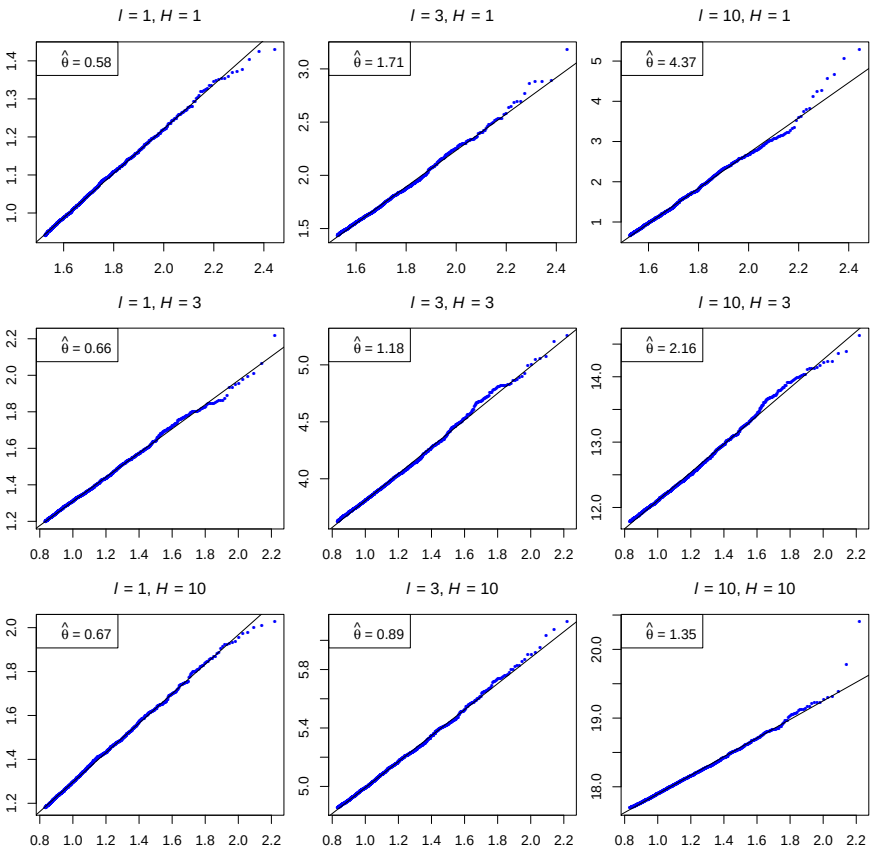

Figure 3: Illustration of the estimation procedure for the tail parameter θ for the prior distribution of Bayesian neural networks units. Layers ℓ of level 1, 3 and 10 on the left, middle and right panels, respectively. Width H (for all considered layers ℓ) equal to 1, 3 and 10 on the top, middle and bottom rows, respectively. Note that the theoretical tail parameter value stated by Vladimirova et al. (2019) is $\theta = \ell/2$, while the infinite width ($H \to \infty$) limiting value is $\theta = 1/2$.

References

Adamczak, R., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2011). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34:61–88.

Bakhshizadeh, M., Maleki, A., and H. de la Pena, V. (2020). Sharp concentration results for heavy-tailed distributions. *arXiv preprint arXiv:2003.13819*.

Ben-Hamou, A., Boucheron, S., and Ohannessian, M. I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23:249–287.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems*. Foundations and Trends® in Machine Learning. Now Publishers, Delft.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Monograph Series. Institute of Mathematical Statistics, Lithuania.

Gardes, L. and Girard, S. (2008). Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference*, 138:1416–1427.

Hao, B., Yadkori, Y. A., Wen, Z., and Cheng, G. (2019). Bootstrapping upper confidence bound. In *Advances in Neural Information Processing Systems*.

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

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.

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*.

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

Raginsky, M. and Sason, I. (2013). Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10:1–246.

Rinne, H. (2008). *The Weibull Distribution: a Handbook*. CRC Press, Boca Raton.

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *International Congress of Mathematicians*. World Scientific.

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

van Handel, R. (2014). Probability in high dimension. Technical report, Princeton University, Princeton.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding priors in Bayesian neural networks at the unit level. In *International Conference on Machine Learning*.

Wellner, J. A. (2017). The Bennett-Orlicz Norm. *Sankhya A*, 79:355–383.