Bayesian neural network unit priors and generalized Weibull-tail property

Mariia Vladimirova MARIIA.VLADIMIROVA@INRIA.FR
Julyan Arbel JULYAN.ARBEL@INRIA.FR
Stéphane Girard STEPHANE.GIRARD@INRIA.FR
Univ. Grenoble Alpes, Inria, CNRS, LJK, 38000 Grenoble, France

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

The connection between Bayesian neural networks and Gaussian processes gained a lot of attention in the last few years. Hidden units are proven to follow a Gaussian process limit when the layer width tends to infinity. Recent work has suggested that finite Bayesian neural networks may outperform their infinite counterparts because they adapt their internal representations flexibly. To establish solid ground for future research on finite-width neural networks, our goal is to study the prior induced on hidden units. Our main result is an accurate description of hidden units tails which shows that unit priors become heavier-tailed going deeper, thanks to the introduced notion of generalized Weibull-tail. This finding sheds light on the behavior of hidden units of finite Bayesian neural networks.

Keywords: generalized Weibull-tail, sub-Weibull, Bayesian neural networks

1. Introduction

Theoretical insights and the development of a comprehensive theory are often the driving force underlying the development of new and improved methods. Neural networks are powerful models, but they still lack comprehensive theoretical support. The Bayesian approach applied to neural networks is considered one of the best-suited frameworks to obtain theoretical explanations and improve the models.

Infinite-width Bayesian neural networks are well-studied. Induced priors in Bayesian neural networks with Gaussian weights are Gaussian processes when the number of hidden units per layer tends to infinity (Neal, 1996; Matthews et al., 2018; Lee et al., 2018; Garriga-Alonso et al., 2019). Stable distributions also lead to stable processes, which are generalizations of Gaussian ones (Favaro et al., 2020).

Since in some cases finite models perform better (Lee et al., 2018; Garriga-Alonso et al., 2019; Lee et al., 2020), there is a need for theoretical justifications. One of the ways is to study induced priors. By analyzing the priors over representations, Aitchison (2020) suggests that finite Bayesian neural networks may generalize better than their infinite counterparts because of their ability to learn representations. Another idea is to find the induced priors in the functional space (Wilson and Izmailov, 2020).

By bounding moments of distributions induced on units, Vladimirova et al. (2019) proved that hidden units have heavier-tailed upper bounds and follow sub-Weibull distributions

© 2021 M. Vladimirova, J. Arbel & S. Girard.

with an increasing tail parameter depending on the layer depth. Those bounds are optimal as they are achieved for shallow Bayesian neural networks; however, they are not accurate.

Recently, Zavatone-Veth and Pehlevan (2021); Noci et al. (2021) showed that there exists a precise description of induced unit priors through Meijer G-functions. These are full descriptions of priors at the unit level. The results are in accordance with the heavy-tailed nature and asymptotic expansions in a wide regime, but it has restrictions. First, the setting is simplified: linear or ReLU activation functions and Gaussian priors on weights. While Wilson and Izmailov (2020) argue that vague Gaussian priors in the parameter space induce useful function-space priors, in some cases, heavier-tailed priors can perform better (Fortuin et al., 2021). Second, it is hard to work with Meijer G-functions due to their complexity. Our goal is to obtain more general characterizations for hidden units.

We introduce a new concept for describing distributional properties of tails by extending the existing Weibull-tail characterization. A random variable X is called *Weibull-tail* (Gardes et al., 2011) with tail parameter $\beta > 0$, which is denoted by $X \sim \mathrm{WT}_{\mathbb{R}_+}(\beta)$, if its cumulative distribution function F_X satisfies

$$\overline{F}_X(x) = 1 - F_X(x) = \mathrm{e}^{-x^\beta l(x)}, \quad \text{for } x > 0, \tag{1}$$

where $l(x)$ is a slowly-varying function, i.e. it is a positive function such that for all $t > 0$ $\lim_{x\to\infty} \frac{l(tx)}{l(x)} = 1$. We note that the Weibull-tail property only considers the right tail of distributions. Here we adapt the Weibull-tail characterization to the whole space $\mathbb{R}$ by taking into consideration the left tail as well. Additionally, we introduce *generalized Weibull-tail* random variables with tail parameter $\beta > 0$ which have Weibull-tail upper and lower bounds for both tails (Definition 1). Such a characterization is easily interpretable and stable under basic operations, even for dependent random variables. The family of generalized Weibull-tail distributions covers a large variety of fundamental distributions such as Gaussian ($\beta = 2$), gamma ($\beta = 1$), Weibull ($\beta > 0$), to name a few, and turns out to be a key tool to describe distributional tails.

Contributions. We make the following contributions:

- We introduce a new notion of tail characteristics called *generalized Weibull-tail*, a version of Weibull-tail characteristics on $\mathbb{R}_+$ extended to variables on $\mathbb{R}$ (Section 2). The additional advantage of this notion is stability under basic operations such as multiplication by a constant and summation.
- With the results on generalized Weibull-tail characterization and dependence, we obtain an accurate characterization of the heavy-tailed nature of hidden units in Bayesian neural networks (Section 3). We establish these results under possibly heavy-tailed priors and relatively mild assumptions on the non-linearity. The conclusions of Vladimirova et al. (2019); Zavatone-Veth and Pehlevan (2021); Noci et al. (2021) which consider only Gaussian priors, mostly follow as corollaries of the obtained characterization. The comparison of different characterizations and related works are deferred to Sections 4 and 5.

2. Generalized Weibull-tail random variables

The study of the distributional tail behavior arises in many applied probability models of different areas, such as hydrology (Strupczewski et al., 2011), finance (Rachev, 2003) and

insurance risk theory (McNeil et al., 2015). Since exact distributions are not available in most cases, deriving asymptotic relationships for their tail probabilities becomes essential. In this context, an important role is played by so-called Weibull-tail distributions satisfying Equation (1) (Gardes et al., 2011; Gardes and Girard, 2016).

A majority of works focuses on right tails of distributions and develops a theory only applicable to right tails, while it is essential to study both right and left tails of distributions. We extend the notion of Weibull-tail on $\mathbb{R}_+$ to $\mathbb{R}$ and introduce generalized Weibull-tail random variables on $\mathbb{R}$ in the following definition. They are stable (under basic operations) extensions of Weibull-tail random variables on $\mathbb{R}$ (Appendix B).

Definition 1 (Generalized Weibull-tail on $\mathbb{R}$) *A random variable X is generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta > 0$ if both its right and left tails are upper and lower bounded by some Weibull-tail functions with tail parameter β:*

$$\begin{aligned} \mathrm{e}^{-x^\beta l_1^r(x)} \leq & \overline{F}_X(x) \leq \mathrm{e}^{-x^\beta l_2^r(x)}, \qquad \text{for } x > 0 \text{ and } x \text{ large enough,} \\ \mathrm{e}^{-|x|^\beta l_1^l(|x|)} \leq & F_X(x) \leq \mathrm{e}^{-|x|^\beta l_2^l(|x|)}, \quad \text{for } x < 0 \text{ and } -x \text{ large enough,} \end{aligned}$$

where l_1^r, l_2^r, l_1^l and l_2^l are slowly-varying functions. We note $X \sim \mathrm{GWT}_{\mathbb{R}}(\beta)$.

This family includes widely used distributions such as Gaussian ($\beta = 2$), Laplace ($\beta = 1$) and generalized Gaussian distributions. These distributions are also symmetric (around 0), so their left and right tails are equal. Thus, it naturally leads to obtain tail characteristics for symmetric distributions by considering random variables whose absolute value is Weibull-tail on $\mathbb{R}_+$. See Appendix B for details. While Weibull-tail random variables are also generalized Weibull-tail, the opposite is not always true. Consider slowly-varying functions $l_1 \equiv 1$ and $l_2 \equiv 2$. Then function $l(\cdot) = 1 + \cos^2(\ln(\cdot))$ satisfies $l_1 \leq l \leq l_2$ but is not slowly varying. However, for any $\beta \geq 2$, function $\overline{F}_X(x) = \mathrm{e}^{-x^\beta l(x)}$ is the survival function (it is decreasing) of some random variable X and it satisfies $\mathrm{e}^{-x^\beta l_2(x)} \leq \overline{F}_X(x) \leq \mathrm{e}^{-x^\beta l_1(x)}$. Therefore, X is $\mathrm{GWT}(\beta)$ but not $\mathrm{WT}(\beta)$.

Next, we aim to obtain a tail characterization for the sum of generalized Weibull-tail variables on $\mathbb{R}$, where we allow random variables to be dependent. Further, we show that under the following assumption of positive dependence, the sum has a tail parameter equal to the minimum among the considered ones (Theorem 2.1).

Definition 2 (Positive dependence condition) *Random variables $X_1, \ldots, X_N$ satisfy the **positive dependence (PD) condition** if the following inequalities hold for all $z \in \mathbb{R}$ and some constant $C > 0$:*

$$\mathbb{P}\left(X_1 \geq 0, \ldots, X_{N-1} \geq 0 | X_N \geq z\right) \geq C, \quad \mathbb{P}\left(X_1 \leq 0, \ldots, X_{N-1} \leq 0 | X_N \leq z\right) \geq C.$$

Remark 2.1 *The choice of zeros and z in the PD condition is arbitrary: one can choose any $z_1, \ldots, z_N$ instead such that $z_1 + \cdots + z_N = z$. The choice of the N-th variable within $X_1, \ldots, X_N$ is also arbitrary. Besides, if random variables $X_1, \ldots, X_N$ are independent with non-zero right and left tails, then they satisfy the PD condition and the constant C is equal to the minimum between $\mathbb{P}(X_1 \leq 0) \ldots \mathbb{P}(X_{N-1} \leq 0)$ and $\mathbb{P}(X_1 \geq 0) \ldots \mathbb{P}(X_{N-1} \geq 0)$.*

There is a great variety of dependent distributions that obey this dependence property including pre-activations in Bayesian neural networks (see Lemma 3.1 for details). Note that the positive orthant dependence condition (POD, see Nelsen, 2007) implies the PD condition.

Theorem 2.1 (Sum of $\mathrm{GWT}_{\mathbb{R}}$ variables) *Let $X_1, \dots, X_N \in \mathrm{GWT}_{\mathbb{R}}$ with tail parameters $\beta_1, \dots, \beta_N$. If $X_1, \dots, X_N$ satisfy the PD condition of Definition 2, then, $X_1 + \cdots + X_N \in \mathrm{GWT}_{\mathbb{R}}(\beta)$ with $\beta = \min\{\beta_1, \dots, \beta_N\}$.*

All proofs are deferred in Appendix. The intuition of the PD condition is to prevent the tail of a sum from becoming lighter due to a negative connection. The simplest example of such negative dependence comes with counter-monotonicity where (X, Y) is such that $Y = -X$. Another less trivial example is $Y = X\mathbb{I}(|X| \leq m) - X\mathbb{I}(|X| > m)$ for some $m > 0$: the sum $X + Y = 2X\mathbb{I}(|X| \leq m)$ is a version of X truncated to the compact set $[-m, m]$. In both cases, it is easy to see that the PD condition does not hold. We conclude this section with a result on the product of independent generalized Weibull-tail random variables.

Theorem 2.2 (Product of $\mathrm{GWT}_{\mathbb{R}}$ variables) *Let $X_1, \dots, X_N \in \mathrm{GWT}_{\mathbb{R}}$ be independent symmetric with tail parameters $\beta_1, \dots, \beta_N$. Then, the product $X_1 \dots X_N \in \mathrm{GWT}_{\mathbb{R}}(\beta)$ with β such that $\frac{1}{\beta} = \frac{1}{\beta_1} + \cdots + \frac{1}{\beta_N}$.*

Now the obtained results can be applied to Bayesian neural networks, showing that hidden units are generalized Weibull-tail on $\mathbb{R}$.

3. Bayesian neural networks induced priors

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 ϕ. Given an input $\mathbf{x} \in \mathbb{R}^{H_0}$, the ℓ-th hidden layer consists of a vector whose size is called the width of the layer, denoted by H_ℓ. The coefficients of linear combinations are called weights, denoted by $w_{ij}^{(\ell)}$ for $i = 1, \dots, H_{\ell-1}, j = 1, \dots, H_\ell$. In Bayesian neural networks, weights are assigned some prior distribution (Neal, 1996). The pre-activations and post-activations of layer ℓ are respectively defined as

$$g_j^{(\ell)} = \sum_{i=1}^{H_{\ell-1}} w_{ij}^{(\ell)} h_i^{(\ell-1)}, \quad h_j^{(\ell)} = \phi(g_j^{(\ell)}) \tag{2}$$

where $h_i^{(0)}$ are elements of input vector $\mathbf{x}$, so $h_i^{(0)}$ are deterministic numerical object features.

The main ingredient for Theorem 2.1 is the positive dependence condition of Definition 2. The product of hidden units and weights satisfies the positive dependence condition:

Lemma 3.1 *Let $X_1, \dots, X_N$ be some possibly dependent random variables and $W_1, \dots, W_N$ be symmetric, mutually independent and independent from $X_1, \dots, X_N$, then random variables $X_1 W_1, \dots, X_N W_N$ satisfy the PD condition.*

Along with Theorem 2.1, the previous lemma implies that neural network hidden units are generalized Weibull-tail with tail parameter depending on those of the weights.

Theorem 3.1 (Hidden units are GWT) *Consider a Bayesian neural network as described in Equation (2) with ReLU activation function. Let ℓ-th layer weights be independent symmetric generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta_w^{(\ell)}$. Then, ℓ-th layer pre-activations are generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta^{(\ell)}$ such that $\frac{1}{\beta^{(\ell)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell)}}$.*

Note that the most popular case of weight prior, iid Gaussian (Neal, 1996), corresponds to $\mathrm{GWT}_{\mathbb{R}}(2)$ weights. This leads to units of layer ℓ which are $\mathrm{GWT}_{\mathbb{R}}(\frac{2}{\ell})$.

To illustrate this theorem, we have built neural networks of 4 hidden layers, with 4 hidden units on each layer. We used a fixed input $\mathbf{x}$ 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 from which units were obtained by forward evaluation with the ReLU non-linearity. This process was iterated $n = 10^6$ times. Note that for a $\mathrm{GWT}_{\mathbb{R}}$ random variable, $\mathbb{P}(X \geq x) = \mathrm{e}^{-x^\beta l(x)}$ so the tail parameter can be expressed as:

$$\beta = \frac{\log(-\log \mathbb{P}(X \geq x))}{\log x} - \frac{\log l(x)}{\log x}. \tag{3}$$

In Figure 1, we plot $\log(-\log \mathbb{P}(X \geq x))$ as a function of $\log x$. We see that the obtained tail parameters approximations are increasing for the increasing layer number and visually correspond to the theoretical tail parameter

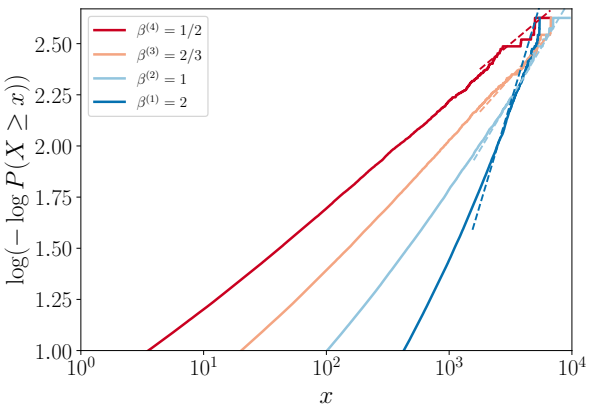

Figure 1: *Solid lines*: approximations of tail parameters $\beta^{(\ell)}$ based on Equation (3) where X are hidden units of layers $\ell = 1, 2, 3, 4$ corresponding theoretically to generalized Weibull-tail with tail parameters $\beta^{(\ell)} = 2, 1, 2/3, 1/2$, under the independent Gaussian weights assumption. *Dashed lines*: linear regressions with coefficients equal to the theoretical tail parameters $\beta^{(\ell)} = 2/\ell$ and manually selected biases to approach the solid lines for visual comparison.

4. Comparison of different characterizations

4.1. Generalized Weibull-tail vs sub-Weibull

Some of the commonly used techniques to study the tail behavior is to consider probability tail bounds such as sub-Gaussian, sub-exponential, or their generalization to sub-Weibull

distributions (Vladimirova et al., 2020; Kuchibhotla and Chakrabortty, 2018). A non-negative random variable X is called sub-Weibull with tail parameter $\theta > 0$ if its survival function is upper-bounded by that of a Weibull distribution:

$$\overline{F}_X(x) \leq a\mathrm{e}^{-bx^{1/\theta}}, \quad \text{for } x > 0 \text{ and some } a, b > 0. \tag{4}$$

This property ensures the existence of the moment generating function as well as bounds on moments. In contrast, the Weibull-tail property characterizes the survival or density functions without a hand on moments. While tail parameters in Equation (1) and (4) of generalized Weibull-tail and sub-Weibull properties respectively are different, there exist connections. Notice that for any constants $a, b, \beta > 0$, function $l(x) = b - \frac{\log a}{x^\beta} \geq 0$ is slowly-varying for x large enough and $a\mathrm{e}^{-bx^\beta} = \mathrm{e}^{-x^\beta l(x)}$. It means that if a random variable X is sub-Weibull with parameter $\theta = 1/\beta > 0$, satisfying Equation (4), then the survival function of X is upper-bounded by a Weibull-tail function with tail parameter β and slowly-varying function $l(x) = 1$, satisfying Equation (1). If random variable X is generalized Weibull-tail with tail parameter β, then from the last item of Proposition A.1, for $a_1, a_2 > 0$ we have

$$a_1\mathrm{e}^{-x^{\beta_1}} \leq \overline{F}_X(x) = \mathrm{e}^{-x^\beta l(x)} \leq a_2\mathrm{e}^{-x^{\beta_2}},$$

or $\mathrm{GWT}_{\mathbb{R}_+}(\beta) \subset \mathrm{SubW}(1/\beta_2)$ and $\mathrm{GWT}_{\mathbb{R}_+}(\beta) \not\subset \mathrm{SubW}(1/\beta_1)$ for x large enough and $\forall(\beta_1, \beta_2)$ such that $0 < \beta_2 < \beta < \beta_1$, as illustrated on Figure 2.

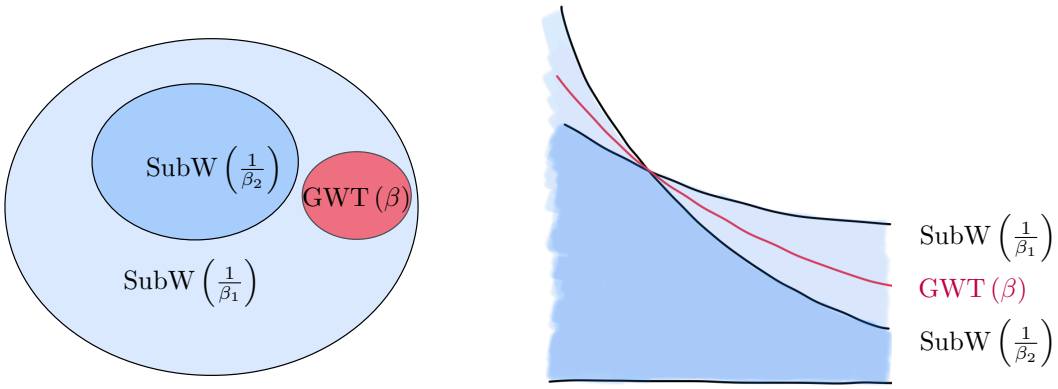

Figure 2: Relation between sub-Weibull and generalized Weibull-tail characteristics.

It was recently shown in Vladimirova et al. (2019) that hidden units of Bayesian neural networks with iid Gaussian priors are sub-Weibull with tail parameter proportional to the hidden layer number, that is $\theta = \frac{\ell}{2}$. It means that the unit distributions of hidden layer ℓ can be upper-bounded by some Weibull distributions $a\mathrm{e}^{-x^{2/\ell}}$ for all ℓ. For larger tail parameter θ, Weibull distribution $a\mathrm{e}^{-x^{1/\theta}}$ is heavier-tailed but being sub-Weibull does not guarantee the heaviness of the tails. However, this upper bound is optimal in the sense that it is achieved for neural networks with one hidden unit per layer.

From Theorem 3.1, for neural networks with independent Gaussian weights, hidden units of ℓ-th layer are generalized Weibull-tail with tail parameter $\beta = 1/\theta = 2/\ell$ so they have upper and lower bounds of the form $\mathrm{e}^{-x^{2/\ell}l(x)}$ up to a constant where l is some slowly-varying function. Therefore, it proves that hidden units are heavier-tailed as going deeper for any finite numbers of hidden units per layer.

4.2. Meijer G-functions description

In Springer and Thompson (1970) it was shown that the probability density function of the product of independent normal variables could be expressed through a Meijer G-function. It resulted in an accurate description of induced unit priors given Gaussian priors on weights and linear or ReLU activation function (Zavatone-Veth and Pehlevan, 2021; Noci et al., 2021). It is a full description of function-space priors but under strong assumptions, requiring Gaussian priors on weights and linear or ReLU activation functions, and with convoluted expressions. In contrast, we provide results for many distributions, including heavy-tailed ones, and our results can be extended to smooth activation functions, such as PReLU, ELU, Softplus.

5. Future applications

Cold posterior effect and priors. It was recently empirically found that Gaussian priors led to the *cold posterior effect* in which a tempered "cold" posterior, obtained by exponentiating the posterior to some power largely greater than one, performs better than an untempered one (Wenzel et al., 2020). The performed Bayesian inference is considered sub-optimal due to the need for cold posteriors, and the model is deemed misspecified. From that angle, Wenzel et al. (2020) suggested that Gaussian priors might not be a good choice for Bayesian neural networks. In some works, data augmentation is argued to be the main reason for this effect (Izmailov et al., 2021; Nabarro et al., 2021) as the increased amount of observed data naturally leads to higher posterior contraction (Izmailov et al., 2021). At the same time, even considering the data augmentation for some models, the cold posterior effect is still present. In addition, Aitchison (2021) demonstrates that the problem might originate in the wrong likelihood of the models and that modifying only the likelihood based on data curation mitigates the cold posterior effect. Nabarro et al. (2021) hypothesize that using an appropriate prior incorporating knowledge of data augmentation might provide a solution. Moreover, heavy-tailed priors have been shown to mitigate the cold posterior effect (Fortuin et al., 2021). According to Theorem 3.1, heavier-tailed priors lead to even heavier-tailed induced priors in function-space. Thus, the heavy-tail property of distributions in function-space might be a highly beneficial feature. Fortuin et al. (2021) also proposed correlated priors for convolutional neural networks since trained weights are empirically strongly correlated. Correlated priors improve overall performance but do not alleviate the cold posterior effect. Our theory can be extended to correlated weight priors. This direction is promising for further uncovering the effect of weight prior on function-space prior.

Edge of Chaos. An active line of research studies the propagation of deterministic inputs in neural networks (Poole et al., 2016; Schoenholz et al., 2017; Hayou et al., 2019). The main idea is to explore the covariance between pre-activations for two given different data points. Poole et al. (2016) and Schoenholz et al. (2017) obtained recurrence relations under the assumption of Gaussian initialization and Gaussian pre-activations. They conclude that there is a critical line, so-called *Edge of Chaos*, separating signal propagation into two regions. The first one is an ordered phase in which all inputs end up asymptotically correlated. The second is a chaotic phase in which all inputs end up asymptotically independent. To propa-

gate the information deeper in a neural network, one should choose Gaussian prior variances corresponding to the separating line. Hayou et al. (2019) show that the smoothness of the activation function also plays an important role. Since this line of works considers Gaussian priors not only on the weights but also on the pre-activations, it is closely related to a wide regime where the number of hidden units per layer tends to infinity. Given that hidden units are heavier-tailed with depth, we speculate that future research will focus on finding better approximations of the pre-activation functions in recurrence relations obtained for finite-width neural networks.

6. Conclusion

We extend the theory on induced distributions in Bayesian neural networks and establish an accurate and easily interpretable characterization of hidden units tails. The obtained results confirm the heavy-tailed nature of hidden units for different weight priors.

References

Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. In *International Conference on Machine Learning*, 2020.

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.

Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*. Number 27. Cambridge university press, 1989.

Stefano Favaro, Sandra Fortini, and Peluchetti Stefano. Stable behaviour of infinitely wide deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

Laurent Gardes and Stéphane Girard. On the estimation of the functional Weibull tail-coefficient. *Journal of Multivariate Analysis*, 146:29–45, 2016.

Laurent Gardes, Stéphane Girard, and Armelle Guillou. Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1): 429–444, 2011.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. In *International Conference on Learning Representations*, 2019.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*, 2019.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, 2021.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Jaehoon Lee, Jascha Sohl-Dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. In *International Conference on Neural Information Processing Systems*, 2020.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools*. Princeton university press, 2015.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in Bayesian neural networks and the cold posterior effect. *arXiv preprint arXiv:2106.05586*, 2021.

Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 1996.

Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

Lorenzo Noci, Gregor Bachmann, Kevin Roth, Sebastian Nowozin, and Thomas Hofmann. Precise characterization of the prior predictive distribution of deep ReLU networks. *arXiv preprint arXiv:2106.06615*, 2021.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *International Conference on Neural Information Processing Systems*, 2016.

Svetlozar Todorov Rachev. *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance, Book 1*. Elsevier, 2003.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Melvin Dale Springer and William E. Thompson. The distribution of products of beta, gamma and Gaussian random variables. *SIAM Journal on Applied Mathematics*, 18(4): 721–737, 1970.

Witold G Strupczewski, Krzysztof Kochanek, Iwona Markiewicz, Ewa Bogdanowicz, Stanislaw Weglarczyk, and Vijay P Singh. On the tails of distributions of annual peak flow. *Hydrology Research*, 42(2-3):171–192, 2011.

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in Bayesian neural networks at the unit level. In *International Conference on Machine Learning*, 2019.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *International Conference on Neural Information Processing Systems*, 2020.

Jacob A Zavatone-Veth and Cengiz Pehlevan. Exact priors of finite neural networks. *arXiv preprint arXiv:2104.11734*, 2021.

Appendix A. Slowly and regularly-varying functions theory

The set of regular-varying functions with index $\rho \in \mathbb{R}$ is denoted by $\mathcal{RV}_\rho$. Note that for $\rho = 0$, the set $\mathcal{RV}_0$ boils down to the set of slowly-varying functions. In particular, any function $r \in \mathcal{RV}_\rho$ can be written $r(x) = x^\rho l(x)$, where l is slowly-varying.

Definition 3 (Regularly-varying function) *Let r be a positive function. Then $r \in \mathcal{RV}_\rho$ if for all $t > 0$ $\lim_{x\to\infty} \frac{r(tx)}{r(x)} = t^\rho$.*

Proposition A.1 *(Bingham et al., 1989, Proposition 1.3.6) Let $l, l_1, \ldots, l_k$ be slowly-varying functions. Then:*

1. *$(\log l(x))/\log x \to 0$ as $x \to \infty$.*
2. *l^α varies slowly for every $\alpha \in \mathbb{R}$.*
3. *$l_1 l_2$, $l_1 + l_2$, and (if $l_2(x) \to \infty$ as $x \to \infty$) $l_1 \circ l_2$ vary slowly.*
4. *If $f(x_1, \ldots, x_k)$ is a rational function with positive coefficients, $f(l_1, \ldots, l_k)$ varies slowly.*
5. *For any $\alpha > 0$, $x^\alpha l(x) \to \infty$, $x^{-\alpha} l(x) \to 0$ $(x \to \infty)$.*

Lemma A.1 *If $l_1(x)$ is slowly-varying, then $l_2(x) = l_1(x^a)$ is slowly-varying for $a > 0$.*

Lemma A.2 *If l_1, l_2 vary slowly, so does $\max\{l_1, l_2\}$ and $\min\{l_1, l_2\}$.*

Lemma A.3 *Let r_1 and r_2 be regularly-varying functions with parameters $\beta_1 > 0$ and $\beta_2 > 0$. Then, the function r such that $\mathrm{e}^{-r} = \mathrm{e}^{-r_1} + \mathrm{e}^{-r_2}$, is regularly-varying with parameter $\beta = \min\{\beta_1, \beta_2\}$.*

Proof Let us express function r from the statement: $r(x) = -\log\varphi(x) = -\log\left(\mathrm{e}^{-r_1(x)} + \mathrm{e}^{-r_2(x)}\right)$. If $\beta_1 \neq \beta_2$, without loss of generality, let us assume that $\beta_1 < \beta_2$, then

$$\varphi(x) = \mathrm{e}^{-r_1(x)}\left(1 + \mathrm{e}^{-r_2(x)+r_1(x)}\right) = \mathrm{e}^{-r_1(x)}\left(1 + \mathrm{e}^{r_2(x)\left(-1+\frac{r_1(x)}{r_2(x)}\right)}\right).$$

Notice that $\frac{r_1}{r_2} \in \mathcal{RV}_{\beta_1-\beta_2}$. For $\beta_1 < \beta_2$, $\frac{r_1(x)}{r_2(x)} \to 0$ when $x \to \infty$. The expression in the exponent $r_2(x)\left(-1 + \frac{r_1(x)}{r_2(x)}\right) \simeq$ [1] $- r_2(x) \to -\infty$, so the exponent $\mathrm{e}^{r_2(x)\left(-1+\frac{r_1(x)}{r_2(x)}\right)} \to 0$. Then $\varphi(x) \simeq \mathrm{e}^{-r_1(x)}$ and $r(x) = -\log\varphi(x) \simeq r_1(x)$. It means that for the case $\beta_1 < \beta_2$, $r \in \mathcal{RV}_{\beta_1}$.

Let us consider the case with equal parameters $\beta_1 = \beta_2 = \beta$, then $r_1(x) = x^\beta l_1(x)$, $r_2(x) = x^\beta l_2(x)$ with slowly-varying l_1 and l_2. With $l = \min\{l_1, l_2\}$, we can write

$$\varphi(x) = \mathrm{e}^{-x^\beta l(x)}\left(1 + \mathrm{e}^{-x^\beta|l_2(x)-l_1(x)|}\right).$$

Consider the logarithm of the latter expression

$$-\log\varphi(x) = x^\beta\left(l(x) + x^{-\beta}\log\left(1 + \mathrm{e}^{-x^\beta|l_2(x)-l_1(x)|}\right)\right).$$

Since the function l is slowly-varying by Lemma A.2 and $0 \leq \mathrm{e}^{-x^\beta|l_2(x)-l_1(x)|} \leq 1$, then $r(x) = -\log\varphi(x) \simeq x^\beta l(x) \in \mathcal{RV}_\beta$. ■

Appendix B. Weibull-tail properties on $\mathbb{R}$

Let us firstly introduce a notion of generalized Weibull-tail random variable which has an additional property of stability:

Definition 4 (Generalized Weibull-tail on $\mathbb{R}_+$) *A random variable X is called generalized Weibull-tail with tail parameter $\beta > 0$ if its survival function $\overline{F}_X$ is bounded by Weibull-tail functions of tail parameter β with possibly different slowly-varying functions l_1 and l_2:*

$$\mathrm{e}^{-x^\beta l_1(x)} \leq \overline{F}_X(x) \leq \mathrm{e}^{-x^\beta l_2(x)}, \quad \text{for } x > 0. \tag{5}$$

We note $X \sim GWT_{\mathbb{R}_+}(\beta)$.

Now we define a random variable whose both right and left tails are Weibull-tail on $\mathbb{R}_+$.

1. We say functions $f \simeq g$ if and only if $f/g \to 1$.

Definition 5 (Weibull-tail on $\mathbb{R}$) *A random variable X on $\mathbb{R}$ is Weibull-tail on $\mathbb{R}$ with tail parameter $\beta > 0$ if both its right and left tails are Weibull-tail with tail parameter β:*

$$\overline{F}_X(x) = 1 - F_X(x) = \mathrm{e}^{-x^\beta l_1(x)}, \quad x > 0$$

$$F_X(x) = \mathrm{e}^{-|x|^\beta l_2(|x|)}, \quad x < 0,$$

where l_1 and l_2 are slowly-varying functions. We note $X \sim WT_{\mathbb{R}}(\beta)$.

Lemma B.1 *(i) If random variable X on $\mathbb{R}$ is $WT_{\mathbb{R}}(\beta)$, then $|X|$ is $WT_{\mathbb{R}_+}(\beta)$.*

(ii) If X is asymmetric but both tails are Weibull-tail and $|X|$ is $WT_{\mathbb{R}_+}(\beta)$, then one of the tails (right or left) is $WT_{\mathbb{R}_+}(\beta)$ and the other one is $WT_{\mathbb{R}_+}(\beta')$ where $\beta' \geq \beta$.

(iii) For symmetric distributions, X is $WT_{\mathbb{R}}(\beta)$ if and only if $|X|$ is $WT_{\mathbb{R}_+}(\beta)$.

Proof

(i) For $x > 0$, the cumulative distribution function of $|X|$ is the following

$$F_{|X|}(x) = F_X(x) - F_X(-x).$$

Then, $\overline{F}_{|X|}(x)$ can be expressed as a sum of the right and left tails:

$$\overline{F}_{|X|}(x) = \overline{F}_X(x) + F_X(-x).$$

If X is $\mathrm{WT}_{\mathbb{R}}(\beta)$, then $|X|$ is $\mathrm{WT}_{\mathbb{R}_+}(\beta)$ as a consequence of Definition 5 and Lemma A.3.

(ii) Let $|X|$ be $\mathrm{WT}_{\mathbb{R}_+}(\beta)$ and X has Weibull left and right tails with different tail parameters. Without loss of generality, assume that $F_X(-x)$ is Weibull-tail with tail parameter $\beta' < \beta$. According to Lemma A.3, the sum survival function $\overline{F}_{|X|}(x)$ will be Weibull-tail with tail parameter $\beta' = \min\{\beta', \beta\}$. We obtained a contradiction and $F_X(-x)$ must have tail parameter greater or equal β. If both tail parameters are greater than β, then the tails sum have the tail parameter equal to the minimum tail parameter among them which is greater than β. It means that at least one tail must have tail parameter β.

(iii) For symmetric distributions $\overline{F}_X(x) = F_X(-x)$ for any x, then $\frac{1}{2}\overline{F}_{|X|}(x) = \overline{F}_X(x) = F_X(-x)$ and we have the equality.

■

Lemma B.2 *(i) If a random variable X is $GWT_{\mathbb{R}}(\beta)$, then $|X|$ is $GWT_{\mathbb{R}_+}(\beta)$.*

(ii) For symmetric distributions, X is $GWT_{\mathbb{R}}(\beta)$ if and only if $|X|$ is $GWT_{\mathbb{R}_+}(\beta)$.

Proof Similarly as in Lemma B.1, we obtain $\overline{F}_{|X|}(x) = \overline{F}_X(x) + F_X(-x)$ for $x \geq 0$.

(i) If X is $\mathrm{GWT}_{\mathbb{R}}(\beta)$, then right and left tails are upper and lower-bounded by some Weibull-tail functions. Then, the sum of the right and left tails is upper and lower-bounded by these Weibull-tail functions and $|X| \sim \mathrm{GWT}_{\mathbb{R}_+}(\beta)$.

(ii) For symmetric distributions we have $\overline{F}_X(x) = F_X(-x)$ for all x, then $\frac{1}{2}\overline{F}_{|X|}(x) = \overline{F}_X(x) = F_X(-x)$ and we have the equality.

■

Lemma B.3 (Power and multiplication by a constant) *If $X \sim GWT_{\mathbb{R}}(\beta)$ and the distribution of X is symmetric, then $a|X|^b \sim GWT_{\mathbb{R}_+}\left(\frac{\beta}{b}\right)$ for $a, b > 0$.*

Proof According to Lemma B.1, $|X| \sim \mathrm{GWT}_{\mathbb{R}_+}(\beta)$. For $a, b > 0$, the tail of $Y = a|X|^b$ is

$$\mathbb{P}\Big(a|X|^b \geq y\Big) = \mathbb{P}\left(|X|^b \geq \frac{y}{a}\right) = \mathbb{P}\Big(|X| \geq \left(\frac{y}{a}\right)^{1/b}\Big).$$

Since $|X|$ is generalized Weibull-tail on $\mathbb{R}_+$ with tail parameter β, $\mathrm{e}^{-x^\beta l_1(x)} \leq \mathbb{P}\left(|X| \geq x\right) \leq \mathrm{e}^{-x^\beta l_2(x)}$, where l_1 and l_2 are slowly-varying functions, it implies

$$\mathrm{e}^{-y^{\beta/b}\tilde{l}_1(y)} \leq \mathbb{P}\Big(a|X|^b \geq y\Big) \leq \mathrm{e}^{-y^{\beta/b}\tilde{l}_2(y)},$$

where $\tilde{l}_i(y) = \frac{l_i((y/a)^{1/b})}{a^{\beta/b}}$, $i = 1, 2$ are slowly-varying functions by Lemma A.1. It leads to the statement of the lemma. ■

Theorem 2.1 (Sum of $\mathrm{GWT}_{\mathbb{R}}$ variables) *Let $X_1, \ldots, X_N \in \mathrm{GWT}_{\mathbb{R}}$ with tail parameters $\beta_1, \ldots, \beta_N$. If $X_1, \ldots, X_N$ satisfy the PD condition of Definition 2, then, $X_1 + \cdots + X_N \in \mathrm{GWT}_{\mathbb{R}}(\beta)$ with $\beta = \min\{\beta_1, \ldots, \beta_N\}$.*

Proof Let us start with $N = 2$. For any random variables X and Y, the following upper bound holds:

$$\mathbb{P}(X + Y \geq z) \leq \mathbb{P}(X \geq z/2) + \mathbb{P}(Y \geq z/2) \leq 2\max\{\mathbb{P}(X \geq z/2), \mathbb{P}(Y \geq z/2)\}.$$

The PD condition leads to a lower bound for the sum:

$$\mathbb{P}(X + Y \geq z) \geq \mathbb{P}(X \geq 0, Y \geq z) = \mathbb{P}(X \geq 0 | Y \geq z)\mathbb{P}(Y \geq z) \geq C\,\mathbb{P}(Y \geq z),$$

where constant $C > 0$. Thus, the sum survival function $\overline{F}_Z(z) = \mathbb{P}(X + Y \geq z)$ has the following bounds for the right tail:

$$C\,\overline{F}_Y(z) \leq \overline{F}_Z(z) \leq 2\max\{\overline{F}_X(z/2), \overline{F}_Y(z/2)\},$$

where $\overline{F}_X$ and $\overline{F}_Y$ are the survival functions of X and Y.

Let X and Y be generalized Weibull-tail on $\mathbb{R}$ of parameters β_x and β_y, then for $z > 0$ large enough

$$C\,\mathrm{e}^{-z^{\beta_y} l_1^r(z)} \leq \overline{F}_Z(z) \leq 2\mathrm{e}^{-z^{\min\{\beta_x, \beta_y\}} l_2^r(z)},$$

where $l_1^r = l_1^y$ is the slowly-varying function appearing in the right tail lower bound of generalized Weibull-tail Y and $l_2^r(z) = \min\{\frac{1}{2^{\beta_x}} l_2^x(z/2), \frac{1}{2^{\beta_y}} l_2^y(z/2)\}$ is the minimum among

slowly-varying functions, where l_2^x and l_2^y are slowly-varying functions in the right tails upper bounds of X and Y. According to Lemma A.2, $l_2^r(z)$ is also slowly-varying. Similarly, we can get bounds for the left tail. Therefore, $X+Y$ is generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta = \min\{\beta_x, \beta_y\}$.

Similarly as above when $N=2$, bounds for the right tail of the sum $Z = X_1 + \cdots + X_N$ with survival function $\overline{F}_Z = \mathbb{P}(Z \geq z)$ are:

$$C\overline{F}_N(z) \leq \overline{F}_Z(z) \leq N \max\{\overline{F}_1(z/N), \ldots, \overline{F}_N(z/N)\},$$

where $\overline{F}_i$ is the survival functions of X_i and constant $C > 0$. The rest of the proof is identical to the one of the case $N = 2$. ■

The case when distributions have only right tails (or only left tails), can be considered as a particular case of the last theorem: a sum of non-negative generalized Weibull-tail random variables is non-negative generalized Weibull-tail with tail parameter equal to the minimum among those of the terms.

Theorem 2.2 (Product of independent $\mathrm{GWT}_{\mathbb{R}}$ variables) *Let $X_1, \ldots, X_N \in \mathrm{GWT}_{\mathbb{R}}$ be independent symmetric with tail parameters $\beta_1, \ldots, \beta_N$. Then, the product $X_1 \ldots X_N \in \mathrm{GWT}_{\mathbb{R}}(\beta)$ with β such that $\frac{1}{\beta} = \frac{1}{\beta_1} + \cdots + \frac{1}{\beta_N}$.*

Proof Consider two independent symmetric generalized Weibull-tail random variables with tail parameters $\beta_x = \frac{1}{\theta_x}$ and $\beta_y = \frac{1}{\theta_y}$, $X \sim \mathrm{GWT}_{\mathbb{R}}\left(\frac{1}{\theta_x}\right)$ and $Y \sim \mathrm{GWT}_{\mathbb{R}}\left(\frac{1}{\theta_y}\right)$. From Lemma B.1 and since random variables X and Y are symmetric, it is equivalent to $|X| \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\frac{1}{\theta_x}\right)$ and $|Y| \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\frac{1}{\theta_y}\right)$.

The product of independent symmetric distributions is symmetric since $Z = XY =^d (-X)Y =^d -Z$. From Lemma B.1, $Z \sim \mathrm{GWT}_{\mathbb{R}}(\beta)$ if and only if $|Z| \sim \mathrm{GWT}_{\mathbb{R}_+}(\beta)$.

Our goal is to show that for some slowly-varying functions l_1 and l_2, there exist upper and lower bounds for the survival function of $|Z|$ and z large enough as follows:

$$\mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_1(z)} \leq \overline{F}_{|Z|}(z) = \mathbb{P}(|XY| \geq z) \leq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_2(z)}. \tag{6}$$

(i) *Upper bound.* First, notice that from the concavity of the logarithm, we have $\ln(pu + (1-p)v) \geq p \ln u + (1-p)\ln v$ for any $u, v > 0$ and $p \in (0,1)$. Then $pu + (1-p)v \geq u^p v^{1-p}$. The change of variables $x = u^p$, $y = v^{1-p}$ implies $px^{1/p} + (1-p)y^{1/(1-p)} \geq xy$. From the latter equation, an upper bound of the product tail is

$$\mathbb{P}(|XY| \geq z) \leq \mathbb{P}\left(p|X|^{1/p} + (1-p)|Y|^{1/(1-p)} \geq z\right). \tag{7}$$

Lemma B.3 implies that $p|X|^{1/p} \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\frac{p}{\theta_x}\right)$ and $(1-p)|Y|^{1/1-p} \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\frac{1-p}{\theta_y}\right)$. Taking $p = \frac{\theta_x}{\theta_x+\theta_y}$ and $1-p = \frac{\theta_y}{\theta_x+\theta_y}$, yields a sum of two independent non-negative generalized Weibull-tail random variables with tail parameter $\frac{1}{\theta_x+\theta_y}$ on the right-hand side of Equation (7). By Theorem 2.1, this sum is generalized Weibull-tail with the same tail parameter $\frac{1}{\theta_x+\theta_y}$. It means that there exists a slowly-varying function l_2 such that the tail of product absolute value $|XY|$ is upper-bounded by

$$\mathbb{P}(|XY| \geq z) \leq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_2(z)}. \tag{8}$$

(ii) *Lower bound.* By independence of $|X|$ and $|Y|$ we have

$$\mathbb{P}(|XY| \geq z) \geq \mathbb{P}\left(|X| \geq z^{\frac{\theta_x}{\theta_x+\theta_y}}\right) \mathbb{P}\left(|Y| \geq z^{\frac{\theta_y}{\theta_x+\theta_y}}\right).$$

Since $|X|$ and $|Y|$ are generalized Weibull-tail on $\mathbb{R}_+$, we can define function $l_1(z) = l_1^x(z^{\theta_x/(\theta_x+\theta_y)}) + l_1^y(z^{\theta_y/(\theta_x+\theta_y)})$ with l_1^x and l_1^y being slowly-varying functions in the lower bounds of generalized Weibull-tail $|X|$ and $|Y|$. Then, l_1 is slowly-varying by Lemma A.1 and we have

$$\mathbb{P}(|XY| \geq z) \geq \mathrm{e}^{-z^{\frac{1}{\theta_x+\theta_y}} l_1(z)}. \tag{9}$$

Combining together Equations (8) and (9) and Definition 4 with Lemma B.1 implies the statement of the theorem. ■

Appendix C. Bayesian neural network properties

Proofs of Section 3.

Lemma 3.1 *Let $X_1, \ldots, X_N$ be some possibly dependent random variables and $W_1, \ldots, W_N$ be symmetric, mutually independent and independent from $X_1, \ldots, X_N$, then random variables $X_1W_1, \ldots, X_NW_N$ satisfy the PD condition.*

Proof The joint probability for the right tail $\mathbb{P}\left(\bigcap_{i=1}^N W_iX_i \geq z_i\right)$ can be expressed as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{P}\left(\bigcap_{i=1}^N W_ix_i \geq z_i \middle| \bigcap_{j=1}^N X_j = x_j\right) f(x_1, \ldots, x_N)\, \mathrm{d}x_1 \ldots \mathrm{d}x_N. \tag{10}$$

Independence between W_i and X_j yields

$$\mathbb{P}\left(\bigcap_{i=1}^N W_ix_i \geq z_i \middle| \bigcap_{j=1}^N X_j = x_j\right) = \mathbb{P}\left(\bigcap_{i=1}^N W_ix_i \geq z_i\right) = \prod_{i=1}^N \mathbb{P}\left(W_ix_i \geq z_i\right),$$

where the last equality is due to the mutual independence of weights $W_1, \ldots, W_N$. Let $z_1 = \cdots = z_{N-1} = 0$ and $z_N = z$. If $x_i = 0$, the probability $\mathbb{P}\left(W_ix_i \geq 0\right) = 1$. If $x_i \neq 0$, then, due to the symmetry of W_i, the probability $\mathbb{P}\left(W_ix_i \geq 0\right) = \frac{1}{2}$. Thus, the following lower bound holds:

$$\mathbb{P}\left(W_ix_i \geq 0\right) \geq \frac{1}{2} \quad \text{for all } i \in \{1, \ldots, N-1\}.$$

Notice that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{P}\left(W_Nx_N \geq z\right) f(x_1, \ldots, x_N)\, \mathrm{d}x_1 \ldots \mathrm{d}x_N = \mathbb{P}\left(W_NX_N \geq z\right).$$

Substituting the latter equations into Equation (10) leads to the lower bound:

$$\mathbb{P}\left(\bigcap_{i=1}^{N-1} W_i X_i \geq 0, W_N X_N \geq z\right) \geq \frac{1}{2^{N-1}} \mathbb{P}\left(W_N X_N \geq z\right).$$

By the conditional probability definition, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{N-1} W_i X_i \geq 0 \middle| W_N X_N \geq z\right) = \frac{\mathbb{P}\left(\bigcap_{i=1}^{N} W_i X_i \geq 0\right)}{\mathbb{P}\left(W_N X_N \geq z\right)} \geq \frac{1}{2^{N-1}}.$$

The proof for the left tail is identical. ■

Theorem 3.1 *Consider a Bayesian neural network as described in Equation (2) with ReLU activation function. Let ℓ-th layer weights be independent symmetric generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta_w^{(\ell)}$. Then, ℓ-th layer pre-activations are generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta^{(\ell)}$ such that $\frac{1}{\beta^{(\ell)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell)}}$.*

Proof The goal is to show that $h_j^{(\ell)} \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\beta^{(\ell)}\right)$ where $\frac{1}{\beta^{(\ell)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell)}}$. We proceed by induction on the layer depth ℓ.

(i) *First hidden layer*

For $\ell = 1$, by Lemma 3.1, products $w_{ij}^{(1)} h_i^{(0)}$ for $i = 1, \ldots, H_1$ satisfy the positive dependence condition of Definition 2, thus the sum $g_j^{(1)} = \sum_i w_{ij}^{(1)} h_i^{(0)}$ is generalized Weibull-tail on $\mathbb{R}$ with tail parameter $\beta^{(1)} = \beta_w^{(1)}$. Since ReLU function does not change the right tail and post-activations lie completely on $\mathbb{R}_+$, we have $h_j^{(1)} \sim \mathrm{GWT}_{\mathbb{R}_+}(\beta^{(1)})$.

(ii) *Step of induction*

Let $h_j^{(\ell-1)} \sim \mathrm{GWT}_{\mathbb{R}_+}\left(\beta^{(\ell-1)}\right)$ where $\frac{1}{\beta^{(\ell-1)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell-1)}}$. Since $h_j^{(\ell-1)}$ is non-negative and $w_{ij}^{(\ell)}$ are symmetric, their product is symmetric by symmetry of the weights: $h_j^{(\ell-1)} w_{ij}^{(\ell)}$ has the same distribution as $h_j^{(\ell-1)}\left(-w_{ij}^{(\ell)}\right) = -h_j^{(\ell-1)} w_{ij}^{(\ell)}$. By Theorem 2.2 and independence of random variables $h_j^{(\ell-1)}$ and $w_{ij}^{(\ell)}$, we obtain that $h_j^{(\ell-1)} w_{ij}^{(\ell)} \sim \mathrm{GWT}_{\mathbb{R}}(\beta^{(\ell)})$ where $\frac{1}{\beta^{(\ell)}} = \frac{1}{\beta^{(\ell-1)}} + \frac{1}{\beta_w^{(\ell)}} = \frac{1}{\beta_w^{(1)}} + \cdots + \frac{1}{\beta_w^{(\ell-1)}} + \frac{1}{\beta_w^{(\ell)}}$.

According to Lemma 3.1, $w_{ij}^{(\ell)} h_i^{(\ell-1)}$, $i = 1, \ldots, H_\ell$ satisfy the dependence condition of Definition 2. Thus, the sum $g_j^{(\ell)} = \sum_{i=1}^{H_\ell} w_{ij}^{(\ell)} h_i^{(\ell-1)}$ is symmetric generalized Weibull-tail with the same tail parameter. The ReLU activation does not impact the right tail, therefore, $h_j^{(\ell)} \sim \mathrm{GWT}_{\mathbb{R}_+}(\beta^{(\ell)})$.

■