Understanding Priors in Bayesian Neural Networks at the Unit Level

Mariia Vladimirova [1 2] **Jakob Verbeek** [1] **Pablo Mesejo** [3] **Julyan Arbel** [1]

Abstract

We investigate deep Bayesian neural networks with Gaussian weight priors and a class of ReLU-like nonlinearities. Bayesian neural networks with Gaussian priors are well known to induce an $\mathcal{L}^2$, "weight decay", regularization. Our results characterize a more intricate regularization effect at the level of the unit activations. Our main result establishes that the induced prior distribution on the units before and after activation becomes increasingly heavy-tailed with the depth of the layer. We show that first layer units are Gaussian, second layer units are sub-exponential, and units in deeper layers are characterized by sub-Weibull distributions. Our results provide new theoretical insight on deep Bayesian neural networks, which we corroborate with simulation experiments.

1. Introduction

Neural networks (NNs), and their deep counterparts (Goodfellow et al., 2016), have largely been used in many research areas such as image analysis (Krizhevsky et al., 2012), signal processing (Graves et al., 2013), or reinforcement learning (Silver et al., 2016), just to name a few. The impressive performance provided by such machine learning approaches has greatly motivated research that aims at a better understanding the driving mechanisms behind their effectiveness. In particular, the study of the NNs distributional properties through Bayesian analysis has recently gained much attention.

Bayesian approaches investigate models by assuming a prior distribution on their parameters. Bayesian machine learning refers to extending standard machine learning approaches with posterior inference, a line of research pioneered by works on Bayesian neural networks (Neal, 1992; MacKay, 1992). There is a large variety of applications, e.g. gene selection (Liang et al., 2018), and the range of models is now very broad, including e.g. Bayesian generative adversarial networks (Saatci & Wilson, 2017). See Polson & Sokolov (2017) for a review. The interest of the Bayesian approach to NNs is at least twofold. First, it offers a principled approach for modeling uncertainty of the training procedure, which is a limitation of standard NNs which only provide point estimates. A second main asset of Bayesian models is that they represent regularized versions of their classical counterparts. For instance, maximum a posteriori (MAP) estimation of a Bayesian regression model with double exponential (Laplace) prior is equivalent to Lasso regression (Tibshirani, 1996), while a Gaussian prior leads to ridge regression. When it comes to NNs, the regularization mechanism is also well appreciated in the literature, since they traditionally suffer from overparameterization, resulting in overfitting.

Central in the field of regularization techniques is the *weight decay* penalty (Krogh & Hertz, 1991), which is equivalent to MAP estimation of a Bayesian neural network with independent Gaussian priors on the weights. Dropout has recently been suggested as a regularization method in which neurons are randomly turned off (Srivastava et al., 2014), and Gal & Ghahramani (2016) proved that a neural network with arbitrary depth and non-linearities, with dropout applied before every weight layer, is mathematically equivalent to an approximation to the probabilistic deep Gaussian process (Damianou & Lawrence, 2013), leading to the consideration of such NNs as Bayesian models.

This paper is devoted to the investigation of hidden units prior distributions in Bayesian neural networks under the assumption of independent Gaussian weights. We first describe a fully connected neural network architecture as illustrated in Figure 1. Given an input $\boldsymbol{x} \in \mathbb{R}^N$, the ℓ-th hidden layer unit activations are defined as

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

where $\boldsymbol{W}^{(\ell)}$ is a weight matrix including the bias vector. A nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$ is applied element-wise, which is called nonlinearity, $\boldsymbol{g}^{(\ell)} = \boldsymbol{g}^{(\ell)}(\boldsymbol{x})$ is a vector of pre-nonlinearities, and $\boldsymbol{h}^{(\ell)} = \boldsymbol{h}^{(\ell)}(\boldsymbol{x})$ is a

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France [2]Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia [3]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain. Correspondence to: Mariia Vladimirova <mariia.vladimirova@inria.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

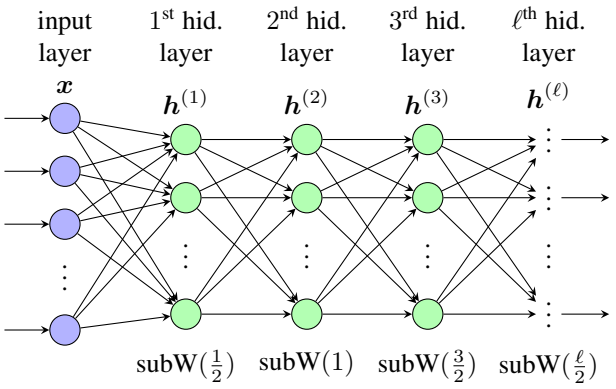

Figure 1. Neural network architecture and characterization of the ℓ-layer units prior distribution as sub-Weibull distribution with tail parameter $\ell/2$, see Definition 3.3.

vector of post-nonlinearities. When we refer to either pre- or post-nonlinearities, we will use the notation $\boldsymbol{U}^{(\ell)}$.

Contributions. In this paper, we extend the theoretical understanding of feedforward fully connected NNs by studying prior distributions at the units level, under the assumption of independent and normally distributed weights. Our contributions are the following:

(i) As our main contribution, we prove in Theorem 3.1 that under some conditions on the activation function ϕ, a Gaussian prior on the weights induces a sub-Weibull distribution on the units (both pre- and post-nonlinearities) with optimal tail parameter $\theta = \ell/2$, see Figure 1. The condition on ϕ essentially imposes that ϕ strikes at a linear rate to $+\infty$ or $-\infty$ for large absolute values of the argument, as ReLU does. In the case of bounded support ϕ, like sigmoid or tanh, the units are bounded, making them *de facto* sub-Gaussian[1]

(ii) We offer an interpretation of the main result from a more elaborate regularization scheme at the level of the units in Section 4.

In the remainder of the paper, we first discuss related work, and then present our main contributions starting with the necessary statistical background and theoretical results (i), then moving to intuitions and interpretation (ii), and ending up with the description of the experiments and the discussion of the results obtained. More specifically, Section 3 states our main contribution, Theorem 3.1, with a proof sketch while additional technical results are deferred to Supplementary material. Section 4 illustrates penalization techniques, providing an interpretation for the theorem. Section 5 describes the experiments. Conclusions and directions for future work are presented in Section 6.

[1]A trivial version of our main result holds, see Remark 3.1.

2. Related work

Studying the distributional behaviour of feedforward networks has been a fruitful avenue for understanding these models, as pioneered by the works of Radford Neal (Neal, 1992; 1996) and David MacKay (MacKay, 1992). The first results in the field addressed the limiting setting when the number of units per layer tends to infinity, also called the wide regime. Neal (1996) proved that a single hidden layer neural network with normally distributed weights tends in distribution in the wide limit either to a Gaussian process (Rasmussen & Williams, 2006) or to an α-stable process, depending on how the prior variance on the weights is rescaled. In recent works, Matthews et al. (2018b), or its updated version Matthews et al. (2018a), and Lee et al. (2018) extend the result of Neal to more-than-one-layer neural networks: when the number of hidden units grows to infinity, deep neural networks (DNNs) also tend in distribution to the Gaussian process, under the assumption of Gaussian weights for properly rescaled prior variances. For the rectified linear unit (ReLU) activation function, the Gaussian process covariance function is obtained analytically (Cho & Saul, 2009). For other nonlinear activation functions, Lee et al. (2018) use a numerical approximation algorithm. This Gaussian process approximation is used for instance by Hayou et al. (2019) for improving neural networks training strategies. Novak et al. (2019) extend the results by proving the Gaussian process limit for convolutional neural networks.

Various distributional properties are also studied in NNs regularization methods. The *dropout* technique (Srivastava et al., 2014) was reinterpreted as a form of approximate Bayesian variational inference (Kingma et al., 2015; Gal & Ghahramani, 2016). While Gal & Ghahramani (2016) built a connection between dropout and the Gaussian process, Kingma et al. (2015) proposed a way to interpret Gaussian dropout. They suggested *variational dropout* where each weight of a model has its individual dropout rate. *Sparse variational dropout* (Molchanov et al., 2017) extends variational dropout to all possible values of dropout rates, and leads to a sparse solution. The approximate posterior is chosen to factorize either over rows or individual entries of the weight matrices. The prior usually factorizes in the same way, and the choice of the prior and its interaction with the approximating posterior family are studied by Hron et al. (2018). Performing dropout can be used as a Bayesian approximation but, as noted by Duvenaud et al. (2014), it has no regularization effect on infinitely-wide hidden layers.

Recent work by Bibi et al. (2018) provides the expression of the first two moments of the output units of a one layer NN.

Obtaining the moments is a first step towards characterizing the full distribution. However, the methodology of Bibi et al. (2018) is limited to the first two moments and to single-layer NNs, while we address the problem in more generality for deep NNs.

3. Bayesian neural networks have heavy-tailed deep units

The deep learning approach uses stochastic gradient descent and error back-propagation in order to fit the network parameters $(\mathbf{W}^{(\ell)})_{1\leq\ell\leq L}$, where ℓ iterates over all network layers. In the Bayesian approach, the parameters are random variables described by probability distributions.

3.1. Assumptions on neural network

We assume a prior distribution on the model parameters, that are the weights $\boldsymbol{W}$. In particular, let all weights (including biases) be independent and have zero-mean normal distribution

$$W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2), \tag{2}$$

for all $1 \leq \ell \leq L$, $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$, with fixed variance σ_w^2. Given some input $\boldsymbol{x}$, such prior distribution induces by forward propagation (1) a prior distribution on the pre-nonlinearities and post-nonlinearities, whose *tail properties* are the focus of this section. To this aim, the nonlinearity ϕ is required to span at least half of the real line as follows. We introduce an extended version of the nonlinearity assumption from Matthews et al. (2018a):

Definition 3.1 (Extended envelope property for nonlinearities)**.** *A nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property if there exist $c_1, c_2 \geq 0$, $d_1, d_2 > 0$ such that the following inequalities hold*

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \textit{for all } u \in \mathbb{R}_+ \textit{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \textit{for all } u \in \mathbb{R}. \end{aligned} \tag{3}$$

The interpretation of this property is that ϕ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$, at least linearly (first line of (3)), and also at most linearly (second line of (3)). Of course, compactly supported nonlinearities such as sigmoid and tanh do not satisfy the extended envelope property but the majority of other nonlinearities do, including ReLU, ELU, PReLU, and SeLU.

We need to recall the definition of asymptotic equivalence between numeric sequences which we use to describe characterization properties of distributions:

Definition 3.2 (Asymptotic equivalence for sequences)**.** *Two sequences a_k and b_k are called asymptotic equivalent and denoted as $a_k \asymp b_k$ if there exist constants $d > 0$ and $D > 0$ such that*

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

The extended envelope property of a function yields the following asymptotic equivalence:

Lemma 3.1. *Let a nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ obey the extended envelope property. Then for any symmetric random variable X the following asymptotic equivalence holds*

$$\|\phi(X)\|_k \asymp \|X\|_k, \quad \textit{for all } k \geq 1, \tag{5}$$

where $\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k}$ is a k-th norm of X.

The proof can be found in the supplementary material.

3.2. Main theorem

This section postulates the rigorous result with a proof sketch. In the supplementary material one can find proofs of intermediate lemmas.

Firstly, we define the notion of *sub-Weibull* random variables (Kuchibhotla & Chakrabortty, 2018; Vladimirova & Arbel, 2019).

Definition 3.3 (Sub-Weibull random variable)**.** *A random variable X satisfying for all $x > 0$ and for some $\theta > 0$*

$$\mathbb{P}(|X| \geq x) \leq a \exp\left(-x^{1/\theta}\right), \tag{6}$$

is called a sub-Weibull random variable with so-called tail parameter θ, which is denoted by $X \sim \textbf{\textit{subW}}(\theta)$.

Sub-Weibull distributions are characterized by tails lighter than (or equally light as) Weibull distributions; in the same way as sub-Gaussian or sub-exponential distributions correspond to distributions with tails lighter than Gaussian and exponential distributions, respectively. Sub-Weibull distributions are parameterized by a positive tail index θ and are equivalent to sub-Gaussian for $\theta = 1/2$ and sub-exponential for $\theta = 1$. To describe a tail lower bound through some sub-Weibull distribution family, i.e. a distribution of X to have the tail heavier than some sub-Weibull, we define the optimal tail parameter for that distribution as the positive parameter θ characterized by:

$$\|X\|_k \asymp k^\theta. \tag{7}$$

Then X is sub-Weibull distributed with optimal tail parameter θ, in the sense that for any $\theta' < \theta$, X is not sub-Weibull with tail parameter θ' (see Vladimirova & Arbel, 2019, for a proof).

The following theorem postulates the main results.

Theorem 3.1 (Sub-Weibull units)**.** *Consider a feed-forward Bayesian neural network with Gaussian priors (2) and with nonlinearity ϕ satisfying the extended envelope condition of Definition 3.1. Then conditional on the input $\boldsymbol{x}$, the marginal prior distribution[2] induced by forward propagation (1) on any unit (pre- or post-nonlinearity) of the ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,*

$$U_m^{(\ell)} \sim \textbf{\textit{subW}}(\ell/2),$$

*where a **subW** distribution is defined in Definition 3.3, and $U_m^{(\ell)}$ is either a pre-nonlinearity $g_m^{(\ell)}$ or a post-nonlinearity $h_m^{(\ell)}$.*

Proof. The idea is to prove by induction with respect to hidden layer depth ℓ that pre- and post-nonlinearities satisfy the asymptotic moment equivalence

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2} \text{ and } \|h^{(\ell)}\|_k \asymp k^{\ell/2}.$$

The statement of the theorem then follows by the moment characterization of optimal sub-Weibull tail coefficient in Equation (7).

According to Lemma 1.1 from from the supplementary material, centering does not harm tail properties, then, for simplicity, we consider zero-mean distributions $W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2)$.

Base step: Consider the distribution of the first hidden layer pre-nonlinearity $g = g^{(1)}$. Since weights $\boldsymbol{W}_m$ follow normal distribution and $\boldsymbol{x}$ is a feature vector, then each hidden unit $\boldsymbol{W}_m^\intercal \boldsymbol{x}$ follow also normal distribution

$$g = \boldsymbol{W}_m^\intercal \boldsymbol{x} \sim \mathcal{N}(0, \sigma_w^2 \|\boldsymbol{x}\|^2).$$

Then, for normal zero-mean variable g, having variance $\sigma^2 = \sigma_w^2 \|\boldsymbol{x}\|^2$, holds the equality in sub-Gaussian property with variance proxy equals to normal distribution variance and from Lemma 1.1 in the supplementary material:

$$\|g\|_k \asymp \sqrt{k}.$$

As activation function ϕ obeys the extended envelope property, nonlinearity moments are asymptotically equivalent to symmetric variable moments

$$\|\phi(g)\|_k \asymp \|g\|_k \asymp \sqrt{k}.$$

It implies that first hidden layer post-nonlinearities h have sub-Gaussian distribution or sub-Weibull with tail parameter $\theta = 1/2$ (Definition 3.3).

[2]We define the *marginal prior distribution* of a unit as its distribution obtained after all other units distributions are integrated out. *Marginal* is to be understood by opposition to *joint*, or *conditional*.

Inductive step: show that if the statement holds for $\ell - 1$, then it also holds for ℓ.

Suppose the post-nonlinearity of $(\ell - 1)$-th hidden layer satisfies the moment condition. Hidden units satisfy the non-negative covariance theorem (Theorem 3.2):

$$\text{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0, \text{ for any } s, t \in \mathbb{N}.$$

Let the number of hidden units in $(\ell - 1)$-th layer equals to H. Then according to Lemma 2.2 from the supplementary material, under assumption of zero-mean Gaussian weights, pre-nonlinearities of ℓ-th hidden layer $g^{(\ell)} = \sum_{i=1}^{H} W_{m,i}^{(\ell-1)} h_i^{(\ell-1)}$ also satisfy the moment condition, but with $\theta = \ell/2$

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2}.$$

From the extended envelope property (Definition 3.1) post-nonlinearities $h^{(\ell)}$ satisfy the same moment condition as pre-nonlinearities $g^{(\ell)}$. This finishes the proof. □

Remark 3.1. If the activation function ϕ is bounded, such as the sigmoid or tanh, then the units are bounded. As a result, by Hoeffding's Lemma, they have a sub-Gaussian distribution.

Remark 3.2. Normalization techniques, such as batch normalization (Ioffe & Szegedy, 2015) or layer normalization (Ba et al., 2016), significantly reduce the training time in feed-forward neural networks. Normalization operations can be decomposed into a set of elementary operations. According to Proposition 1.4 from the supplementary material, elementary operations do not harm the distribution tail parameter. Therefore, normalization methods do not have an influence on tail behavior.

3.3. Intermediate theorem

This section states with a proof sketch that the covariance between hidden units in the neural network is non-negative.

Theorem 3.2 (Non-negative covariance between hidden units)**.** *Consider the deep neural network described in, and with the assumptions of, Theorem 3.1. The covariance between hidden units of the same layer is non-negative. Moreover, for given ℓ-th hidden layer units $h^{(\ell)}$ and $\tilde{h}^{(\ell)}$, it holds*

$$\textit{Cov}\left[\left(h^{(\ell)}\right)^s, \left(\tilde{h}^{(\ell)}\right)^t\right] \geq 0, \textit{ where } s, t \in \mathbb{N}.$$

For first hidden layer $\ell = 1$ there is equality for all s and t.

Proof. A more detailed proof can be found in the supplementary material in Section 3.

Recall the covariance definition for random variables X and Y

$$\text{Cov}\left[X, Y\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{8}$$

The proof is based on induction with respect to the hidden layer number.

In the proof let us make notation simplifications: $\boldsymbol{w}_m^\ell = \boldsymbol{W}_m^\ell$ and $w_{mi}^\ell = W_{mi}^\ell$ for all $m \in H_\ell$. If the index m is omitted, then $\boldsymbol{w}^\ell$ is some the vectors $\boldsymbol{w}_m^\ell$, w_i^ℓ is i-th element of the vector $\boldsymbol{w}_m^\ell$.

1. First hidden layer. Consider the first hidden layer units $h^{(1)}$ and $\tilde{h}^{(1)}$. The covariance between units is equal to zero and the units are Gaussian, since the weights $\boldsymbol{w}^{(1)}$ and $\tilde{\boldsymbol{w}}^{(1)}$ are from $\mathcal{N}(0, \sigma_w^2)$ and independent. Thus, the first hidden layer units are independent and its covariance (8) is equal to 0. Moreover, since $h^{(1)}$ and $\tilde{h}^{(1)}$ are independent, then $\left(h^{(1)}\right)^s$ and $\left(\tilde{h}^{(1)}\right)^t$ are also independent.

2. Next hidden layers. Assume that the $(\ell-1)$-th hidden layer has $H_{\ell-1}$ hidden units, where $\ell > 1$. Then the ℓ-th hidden layer pre-nonlinearity is equal to

$$g^{(\ell)} = \sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}. \tag{9}$$

We want to prove that the covariance (8) between the ℓ-th hidden layer pre-nonlinearities is non-negative. Let us show firstly the idea of the proof in the case $H_{\ell-1} = 1$ and then briefly describe the proof for any finite $H_{\ell-1} > 1$, $H_{\ell-1} \in \mathbb{N}$.

2.1 One hidden unit. In the case $H_{\ell-1} = 1$, the covariance (8) sign is the same as of the expression

$$\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2(s_1+t_1)}\right] - \mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2s_1}\right]\mathbb{E}\left[\left(h^{(\ell-1)}\right)^{2t_1}\right],$$

since the weighs are zero-mean distributed, its moments are equal to zero with an odd order. According to Jensen's inequality for convex function f, we have $\mathbb{E}[f(x_1, x_2)] \geq f(\mathbb{E}[x_1], \mathbb{E}[x_2])$. Since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \geq 0$ and $x_2 \geq 0$, then, taking $x_1 = \left(h^{(\ell-1)}\right)^{2s_1}$ and $x_2 = \left(h^{(\ell-1)}\right)^{2t_1}$, we have the condition we need (10) being satisfied.

2.1. H hidden units. Now let us consider the covariance between pre-nonlinearities (9) for $H_{\ell-1} = H > 1$. Raise the sum in the brackets to the power

$$\left(\sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)}\right)^s =$$

$$= \sum_{s_H=0}^{s} C_s^{s_H} \left(w_H^{(\ell)} h_H^{(\ell-1)}\right)^{s_H} \left(\sum_{i=1}^{H-1} w_i^{(\ell)} h_i^{(\ell-1)}\right)^{s-s_H}.$$

And the same way for the second bracket $\left(\sum_{i=1}^{H} \tilde{w}_i^{(\ell)} h_i^{(\ell-1)}\right)^t$. Notice that binomial terms will be the same in the minuend and the subtrahend terms of (8). So the covariance in our notations can be written in the form of

$$\operatorname{Cov}\left[\left(\sum_{i=1}^{H_{\ell-1}} w_i^{(\ell)} h_i^{(\ell-1)}\right)^s, \left(\sum_{i=1}^{H_{\ell-1}} \tilde{w}_i^{(\ell)} h_i^{(\ell-1)}\right)^t\right] =$$

$$= \sum\sum C\left(\mathbb{E}\left[AB\right] - \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]\right),$$

where C-terms contain binomial coefficients, A-terms — all possible products of hidden units in $\left(g^{(\ell)}\right)^s$ and B-terms — all possible products of hidden units in $\left(\tilde{g}^{(\ell)}\right)^t$. In order for the covariance to be non-negative, it is sufficient to show that the difference $\mathbb{E}\left[AB\right] - \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]$ is non-negative. Since the weights are Gaussian and independent, we have the following equation, omitting the superscript for simplicity,

$$\mathbb{E}\left[AB\right] = W\tilde{W} \cdot \mathbb{E}\left[\prod_{i=1}^{H} h_i^{s_i+t_i}\right],$$

$$\mathbb{E}\left[A\right]\mathbb{E}\left[B\right] = W\tilde{W} \cdot \mathbb{E}\left[\prod_{i=1}^{H} h_i^{s_i}\right]\mathbb{E}\left[\prod_{i=1}^{H} h_i^{t_i}\right],$$

where $W\tilde{W}$ is the product of weights moments

$$W\tilde{W} = \prod_{i=1}^{H} \mathbb{E}\left[w_i^{s_i}\right]\mathbb{E}\left[\tilde{w}_i^{t_i}\right].$$

For $W\tilde{W}$ not equal to zero, all the powers must be even. Now we need to prove

$$\mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2(s_i+t_i)}\right] \geq \mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2s_i}\right]\mathbb{E}\left[\prod_{i=1}^{H/2} h_i^{2t_i}\right]. \tag{10}$$

According to Jensen's inequality for convex functions, since a function $f(x_1, x_2) = x_1 x_2$ is convex for $x_1 \geq 0$ and $x_2 \geq 0$, then, taking $x_1 = \prod_{i=1}^{H/2} h_i^{2s_i}$ and $x_2 = \prod_{i=1}^{H/2} h_i^{2t_i}$, the condition from (10) is satisfied.

3. Post-nonlinearities.

Let show the proof for the ReLU nonlinearity.

The distribution of the ℓ-th hidden layer pre-nonlinearity $g^{(\ell)}$ is the sum of symmetric distributions, which are products of Gaussian variables $w^{(\ell)}$ and the non-negative ReLU output, i.e. the $(\ell-1)$-th hidden layer post-nonlinearity $h^{(\ell-1)}$. Therefore, $g^{(\ell)}$ follows a symmetric distribution and the following inequality

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} gg'\, p(g, g')\, dg\, dg' \geq$$

$$\geq \int_{-\infty}^{+\infty} g\, p(g)\, dg \cdot \int_{-\infty}^{+\infty} g'\, p(g')\, dg'$$

implies the same inequality for a positive part

$$\int_0^{+\infty}\int_0^{+\infty} gg'\,p(g,g')\,dg\,dg' \geq \geq \int_0^{+\infty} g\,p(g)\,dg \cdot \int_0^{+\infty} g'\,p(g')\,dg'.$$

Notice that the equality above is the ReLU function output and for a symmetric distribution we have

$$\int_0^{+\infty} x\,p(x)\,dx = \frac{1}{2}\mathbb{E}\left[|X|\right]. \tag{11}$$

That means if the non-negative covariance is proven for pre-nonlinearities, for post-nonlinearities it is also non-negative. We omit the proof for the other nonlinearities with the extended envelope property, since instead of precise equation (11), the asymptotic equivalence for moments will be used for a positive part and for a negative part — precise expectation expressions which depend on certain nonlinearity. □

3.4. Convolutional neural networks

Convolutional neural networks (Fukushima & Miyake, 1982; LeCun et al., 1998) are a particular kind of neural network for processing data that has a known grid-like topology, which allows to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the neural network. Neurons in such networks are arranged in three dimensions: width, height and depth. There are three main types of layers that can be concatenated in these architectures: convolutional, pooling, and fully-connected layers (exactly as seen in standard NNs). The convolutional layer computes dot products between a region in the inputs and its weights. Therefore, each region can be considered as a particular case of a fully-connected layer. Pooling layers control overfitting and computations in deep architectures. They operate independently on every slice of the input and reduces it spatially. The most commonly functions used in pooling layers are *max pooling* and *average pooling*.

Proposition 3.1. *The operations: 1.* max pooling *and 2.* averaging *do not modify the optimal tail parameter θ of sub-Weibull family. Consequently, the result of Theorem 3.1 carries over to convolutional neural networks.*

The proof can be found in the supplementary material.

Corollary 3.1. *Consider a convolutional neural network containing convolutional, pooling and fully-connected layers under assumptions from Section 3.1. Then a unit of ℓ-th hidden layer has sub-Weibull distribution with optimal tail parameter $\theta = \ell/2$, where ℓ is the number of convolutional and fully-connected layers.*

Proof. Proposition 3.1 implies that the pooling layer keeps the tail parameter. From discussion at the beginning of the section, the result of Theorem 3.1 is also applied to convolutional neural networks where the depth is considered as the number of convolutional and fully-connected layers. □

4. Regularization scheme on the units

Our main theoretical contribution, Theorem 3.1, characterizes the marginal prior distribution of the network units as follows: when the depth increases, the distribution becomes more heavy-tailed. In this section, we provide an interpretation of the result in terms of regularization at the level of the units. To this end, we first briefly recall shrinkage and penalized estimation methods.

4.1. Short digest on penalized estimation

The notion of penalized estimation is probably best illustrated on the simple linear regression model, where the aim is to improve prediction accuracy by shrinking, or even putting exactly to zero, some coefficients in the regression. Under these circumstances, inference is also more *interpretable* since, by reducing the number of coefficients effectively used in the model, it is possible to grasp its salient features. Shrinking is performed by imposing a penalty on the size of the coefficients, which is equivalent to allowing for a given budget on their size. Denote the regression parameter by $\beta \in \mathbb{R}^p$, the regression sum-of-squares by $R(\beta)$, and the penalty by $\lambda L(\beta)$, where L is some norm on $\mathbb{R}^p$ and λ some positive tuning parameter. Then, the two formulations of the regularized problem

$$\min_{\beta\in\mathbb{R}^p} R(\beta) + \lambda L(\beta), \text{ and}$$
$$\min_{\beta\in\mathbb{R}^p} R(\beta) \text{ subject to } L(\beta) \leq t,$$

are equivalent, with some one-to-one correspondence between λ and t, and are respectively termed the *penalty* and the *constraint* formulation. This latter formulation provides an interesting geometrical intuition of the shrinkage mechanism: the constraint $L(\beta) \leq t$ reads as imposing a total budget of t for the parameter size in terms of the norm L. If the ordinary least squares estimator $\hat{\beta}^{\text{ols}}$ lives in the L-ball with surface $L(\beta) = t$, then there is no effect on the estimation. In contrast, when $\hat{\beta}^{\text{ols}}$ is outside the ball, then the intersection of the lowest level curve of the sum-of-squares $R(\beta)$ with the L-ball defines the penalized estimator.

The choice of the L norm has considerable effects on the problem, as can be sensed geometrically. Consider for instance $\mathcal{L}^q$ norms, with $q \geq 0$. For any $q > 1$, the associated $\mathcal{L}^q$ norm is differentiable and contours have a round shape without sharp angles. In that case, the penalty effect is to

shrink the β coefficients towards 0. The most well-known estimator falling in this class is the *ridge* regression obtained with $q = 2$, see Figure 2 top-left panel. In contrast, for any $q \in (0, 1]$, the $\mathcal{L}^q$ norm has some non differentiable points along the axis coordinates, see Figure 2 top-right and bottom panels. Such critical points are more likely to be hit by the level curves of the sum-of-squares $R(\beta)$, thus setting exactly to zero some of the parameters. A very successful approach in this class is the Lasso obtained with $q = 1$. Note that the problem is computationally much easier in the convex situation which occurs only for $q \geq 1$.

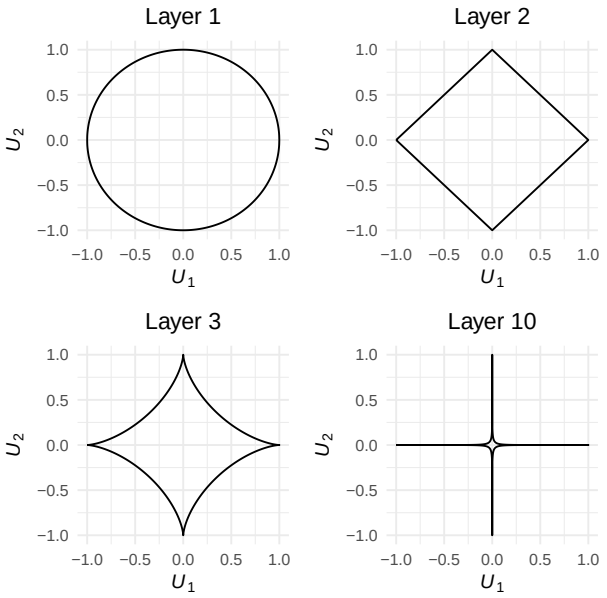

Figure 2. $\mathcal{L}^{2/\ell}$-norm unit balls (in dimension 2) for layers $\ell = 1, 2, 3$ and 10.

4.2. MAP on weights W is weight decay

These penalized methods have a simple Bayesian counterpart in the form of the maximum a posteriori (MAP) estimator. In this context, the objective function R is the negative log-likelihood, while the penalty L is the negative log-prior. The objective function takes on the form of sum-of-squared errors for regression under Gaussian errors, and of cross-entropy for classification.

For neural networks, it is well-known that an independent Gaussian prior on the weights

$$\pi(\boldsymbol{W}) \propto \prod_{\ell=1}^{L} \prod_{i,j} \mathrm{e}^{-\frac{1}{2}(W_{i,j}^{(\ell)})^2}, \tag{12}$$

is equivalent to the weight decay penalty, also known as ridge regression:

$$L(\boldsymbol{W}) = \sum_{\ell=1}^{L} \sum_{i,j} (W_{i,j}^{(\ell)})^2 = \|\boldsymbol{W}\|_2^2, \tag{13}$$

where products in (12) and sums in (13) involving i and j above are over $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$, H_0 and H_L representing respectively the input and output dimensions.

4.3. MAP on units U

Now moving the point of view from *weights* to *units* leads to a radically different shrinkage effect. Let $U_m^{(\ell)}$ denote the m-th unit of the ℓ-th layer (either pre- or post-nonlinearity). We prove in Theorem 3.1 that conditional on the input $\boldsymbol{x}$, a Gaussian prior on the weights translates into some prior on the units $U_m^{(\ell)}$ that is marginally sub-Weibull with optimal tail index $\theta = \ell/2$. This means that the tails of $U_m^{(\ell)}$ satisfy

$$\mathbb{P}(|U_m^{(\ell)}| \geq u) \leq \exp\left(-u^{2/\ell}/K_1\right) \quad \text{for all } u \geq 0, \tag{14}$$

for some positive constant K_1. The exponent of u in the exponential term above is optimal in the sense that Equation (14) is not satisfied with some parameter θ' smaller than $\ell/2$. Thus, the marginal density of $U_m^{(\ell)}$ on $\mathbb{R}$ is approximately proportional to

$$\pi_m^{(\ell)}(u) \approx \mathrm{e}^{-|u|^{2/\ell}/K_1}. \tag{15}$$

The joint prior distribution for all the units $\boldsymbol{U} = (U_m^{(\ell)})_{1 \leq \ell \leq L, 1 \leq m \leq H_\ell}$ can be expressed from all the marginal distributions by Sklar's representation theorem (Sklar, 1959) as

$$\pi(\boldsymbol{U}) = \prod_{\ell=1}^{L} \prod_{m=1}^{H_\ell} \pi_m^{(\ell)}(U_m^{(\ell)})\, C(F(\boldsymbol{U})), \tag{16}$$

where C represents the copula of $\boldsymbol{U}$ (which characterizes all the dependence between the units) while F denotes its cumulative distribution function. The penalty incurred by such a prior distribution is obtained as the negative log-prior,

$$\begin{aligned} L(\boldsymbol{U}) &= -\sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\boldsymbol{U})), \\ &\overset{\text{(a)}}{\approx} \sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} |U_m^{(\ell)}|^{2/\ell} - \log C(F(\boldsymbol{U})), \\ &\approx \|\boldsymbol{U}^{(1)}\|_2^2 + \|\boldsymbol{U}_1^{(2)}\|_1 + \cdots + \|\boldsymbol{U}^{(L)}\|_{2/L}^{2/L} \\ &\quad - \log C(F(\boldsymbol{U})), \end{aligned} \tag{17}$$

where (a) comes from (15). The first L terms in (17) indicate that some shrinkage operates at every layer of the network, with a penalty term that approximately takes the form of the $\mathcal{L}^{2/\ell}$ norm at layer ℓ. Thus, the deeper the layer, the stronger the regularization induced at the level of the units, as summarized in Table 1.

Layer	Penalty on $\boldsymbol{W}$	Approximate penalty on $\boldsymbol{U}$	
1	$\|\boldsymbol{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\boldsymbol{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\boldsymbol{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

Table 1. Comparison of Bayesian neural network penalties on weights $\boldsymbol{W}$ and units $\boldsymbol{U}$.

5. Experiments

We illustrate the result of Theorem 3.1 on a 100 layers MLP. The hidden layers of neural network have $H_1 = 1000$, $H_2 = 990$, $H_3 = 980$, ..., $H_\ell = 1000 - 10(\ell - 1)$, ..., $H_{100} = 10$ hidden units, respectively. The input $\boldsymbol{x}$ is a vector of features from $\mathbb{R}^{10^4}$. Figure 3 represents the tails of first three, 10th and 100th hidden layers pre-nonlinearity marginal distributions in logarithmic scale. Units of one layer have the same sub-Weibull distribution since they share the same input and prior on the corresponding weights. The curves are obtained as histograms from a sample of size 10^5 from the prior on the pre-nonlinearities, which is itself obtained by sampling 10^5 sets of weights $\boldsymbol{W}$ from the Gaussian prior (2) and forward propagation via (1). The input vector $\boldsymbol{x}$ is sampled with independent features from a standard normal distribution once for all at the start. The nonlinearity ϕ is the ReLU function. Being a linear combination involving symmetric weights $\boldsymbol{W}$, pre-nonlinearities $\boldsymbol{g}$ also have a symmetric distribution, thus we visualize only their distribution on $\mathbb{R}_+$.

Figure 3 corroborates our main result. On the one hand, the prior distribution of the first hidden units is Gaussian (green curve), which corresponds to a **subW**$(1/2)$ distribution. On the other hand, deeper layers are characterized by heavier-tailed distributions. The deepest considered layer (100th, violet curve) has an extremely flat distribution, which corresponds to a **subW**(50) distribution.

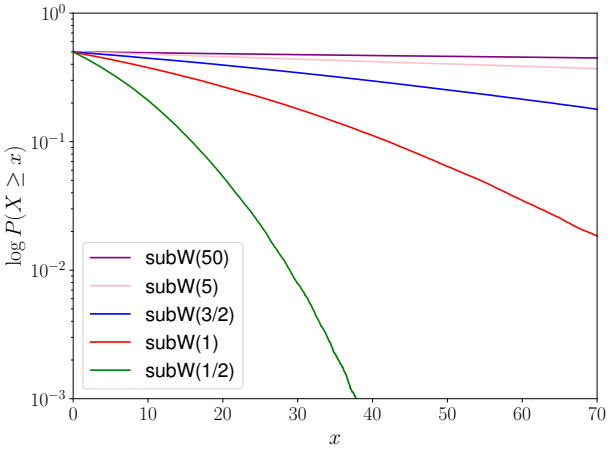

Figure 3. Illustration of layers $\ell = 1, 2, 3, 10$ and 100 hidden units (pre-nonlinearities) marginal prior distributions. They correspond respectively to **subW**$(1/2)$, **subW**(1), **subW**$(3/2)$, **subW**(5) and **subW**(50).

6. Conclusion and future work

Despite the ubiquity of deep learning throughout science, medicine and engineering, the underlying theory has not kept pace with applications for deep learning. In this paper, we have extended the state of knowledge on Bayesian neural networks by providing a characterization of the marginal prior distribution of the units. Matthews et al. (2018a) and Lee et al. (2018) proved that unit distributions have a Gaussian process limit in the wide regime, i.e. when the number of hidden units tends to infinity. We showed that they are heavier-tailed as depth increases, and discussed this result in terms of a regularizing mechanism at the level of the units. We anticipate that the Gaussian process limit of sub-Weibull distributions in a given layer for increasing width could be also recovered through a modification of the Central Limit Theorem for heavy-tailed distributions, see Kuchibhotla & Chakrabortty (2018).

Since initialization and learning dynamics are key in modern machine learning in order to properly tune deep learning algorithms, a good implementation practice requires a proper understanding of the prior distribution at play and of the regularization it incurs.

We hope that our results will open avenues for further research. Firstly, Theorem 3.1 regards the *marginal* prior distribution of the units, while a full characterization of the joint distribution of all units $\boldsymbol{U}$ remains an open question. More specifically, a precise description of the copula defined in Equation (16) would provide valuable information about the dependence between the units, and also about the precise geometrical structure of the balls induced by that penalty. Secondly, the interpretation of our result (Section 4) is concerned with the maximum a posteriori of the units, which is a point estimator. One of the benefits of the Bayesian approach to neural networks lies in its ability to provide a principled approach to uncertainty quantification, so that an interpretation of our result in terms of the full posterior distribution would be very appealing. Lastly, the practical potentialities of our results are many: to better comprehend the regularizing mechanisms in deep neural networks will contribute to design and understand strategies to avoid overfitting and improve generalization.

Acknowledgements

We would like to thank Stéphane Girard for fruitful discussions on Weibull-like distributions and Cédric Févotte for pointing out the potential relationship of our heavy-tail result with sparsity-inducing priors.

References

Ba, J., Kiros, J., and Hinton, G. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bibi, A., Alfadly, M., and Ghanem, B. Analytic expressions for probabilistic moments of PL-DNN with Gaussian input. In *CVPR*, 2018.

Cho, Y. and Saul, L. Kernel methods for deep learning. In *NeurIPS*, 2009.

Damianou, A. and Lawrence, N. Deep Gaussian processes. In *AISTATS*, 2013.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *AISTATS*, 2014.

Fukushima, K. and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*. Springer, 1982.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Graves, A., Mohamed, A., and Hinton, G. Speech recognition with deep recurrent neural networks. In *ICASSP*, pp. 6645–6649, 2013.

Hayou, S., Doucet, A., and Rousseau, J. On the impact of the activation function on deep neural networks training. *arXiv preprint arXiv:1902.06853*, 2019.

Hron, J., Matthews, A., and Ghahramani, Z. Variational Bayesian dropout: pitfalls and fixes. In *ICML*, 2018.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Kingma, D., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *NeurIPS*, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Krogh, A. and Hertz, J. A simple weight decay can improve generalization. In *NeurIPS*, 1991.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as Gaussian processes. In *ICML*, 2018.

Liang, F., Li, Q., and Zhou, L. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972, 2018.

MacKay, D. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Matthews, A., Rowland, M., Hron, J., Turner, R., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv*, 1804.11271, 2018a.

Matthews, A., Rowland, M., Hron, J., Turner, R., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *ICLR*, 2018b.

Molchanov, D., Ashukha, A., and Vetrov, D. Variational dropout sparsifies deep neural networks. In *ICML*, 2017.

Neal, R. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG-TR-92-1, University of Toronto, 1992.

Neal, R. *Bayesian learning for neural networks*. Springer, 1996.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Hron, J., Abolafia, D., Pennington, J., and Sohl-Dickstein, J. Bayesian deep convolutional networks with many channels are Gaussian processes. In *ICLR*, 2019.

Polson, N. G. and Sokolov, V. Deep learning: A Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Saatci, Y. and Wilson, A. Bayesian GAN. In *NeurIPS*, 2017.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Vladimirova, M. and Arbel, J. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *arXiv preprint arXiv:1905.04955*, 2019.