Bayesian neural network priors at the level of units
Mariia Vladimirova, Julyan Arbel, Pablo Mesejo

To cite this version:
Mariia Vladimirova, Julyan Arbel, Pablo Mesejo. Bayesian neural network priors at the level of units. AABI 2018 - 1st Symposium on Advances in Approximate Bayesian Inference, Dec 2018, Montréal, Canada. pp.1-6. hal-01950659

HAL Id: hal-01950659
https://hal.archives-ouvertes.fr/hal-01950659
Submitted on 11 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Bayesian neural network priors at the level of units

Mariia Vladimirova
Julyan Arbel
Pablo Mesejo

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 Granada, Spain

Abstract

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularization scheme at the level of the units, ranging from convex penalties for the first two layers—$L^2$ regularization for the first and Lasso for the second—to non convex penalties for deeper layers. Thus, although weight decay does not allow for the weights to be set exactly to zero, sparse solutions tend to be selected for the units from the second layer onward. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

Keywords: Bayesian neural network, heavy-tailed prior, sparsity

1. Introduction

Neural networks (NNs), and their deep extensions (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 performances of such neural networks have greatly strengthened the line of research that aims at better understanding the driving mechanisms behind the effectiveness of deep neural networks. One important aspect of this line of research that has recently gained much attention is the study of distributional properties of the NNs through Bayesian inference.

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 the works Neal (1992); MacKay (1992) on Bayesian neural networks which now extends to a broad class of models, including Bayesian Generative adversarial network (Saatci and Wilson, 2017). See Polson and 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 neural networks, the
regularization mechanism is also well appreciated in the literature, since neural networks
traditionally suffer from overparameterization, resulting in overfitting.

Central in the field of regularization techniques is the weight decay penalty (Krogh
and Hertz, 1991), which is equivalent to MAP estimation of a Bayesian neural network
with independent Gaussian priors on the weights. Srivastava et al. (2014) have suggested
dropout as a regularization method in which neurons are randomly turned off. Gal and
Ghahramani (2016) proved that the neural network trained with dropout is equivalent to
a probabilistic model, i.e. a deep Gaussian process (Damianou and Lawrence, 2013). It leads
to the consideration of such neural networks as Bayesian models.

Recent papers study various distributional properties of Bayesian neural networks.
Matthews et al. (2018b), or its extended version Matthews et al. (2018a), and Lee et al.
(2018) showed that deep neural networks tend in distribution to the Gaussian process when
the number of hidden units grows to infinity, under the assumption of Gaussian weights for
properly rescaled prior variances. The work by Bibi et al. (2018) provides the expression
of the first two moments of the output units of a one layer neural network. Obtaining the
moments is a first step to characterizing a whole distribution, however the methodology
of Bibi et al. (2018) is limited to the first two moments and to one layer neural networks.
Vladimirova et al. (2018) investigate the marginal prior distribution of the network units,
stating that as the depth increases, the distribution becomes more heavy-tailed. For the
sake of completeness, we reproduce the result here:

Theorem 1 (Sub-Weibull units, Vladimirova et al., 2018) Consider a feed-forward
Bayesian neural network with Gaussian priors on weights and activation function \( \phi \) satisfying
the extended envelope condition of Definition A.1, Appendix A. Consider the marginal
prior distribution, conditional on the input \( x \), induced by forward propagation on any unit
before or after activation of the \( \ell \)-th hidden layer. It is sub-Weibull with optimal tail pa-
rameter \( \theta = \ell/2 \). That is for any layer \( \ell \) with \( H_\ell \) hidden units, a unit \( U_m^{(\ell)} \) of the layer
\( 1 \leq m \leq H_\ell \)

\[ U_m^{(\ell)} \sim \text{subW}(\ell/2), \]

where a subW distribution is defined in Definition A.2, Appendix A.

We provide in this note an interpretation of Theorem 1 in terms of sparsity-inducing mech-
anism at the level of the units. We stress that the term unit and notation \( U^{(\ell)} \) refer
indistinctly to units before or after activation. The notation \( W^{(\ell)} \) stands for the weight
matrix including the bias vector.

2. Sparsity-inducing prior on the units

Shrinkage is performed by imposing a penalty on the size of the parameters. Denote the
parameters by \( W \), the loss function by \( R(W) \), and the penalty by \( \lambda \mathcal{L}(W) \), where \( \mathcal{L} \)
is some norm on the weight space and \( \lambda \) some positive tuning parameter. Then, the regularized
problem is

\[ \min_W R(W) + \lambda \mathcal{L}(W). \]
The choice of the $L^q$ norm has considerable effects on the problem, as can be sensed geometrically. Consider for instance $L^q$ norms, with $q \geq 0$. For any $q > 1$, the associated $L^q$ norm is differentiable and contours have a round shape without sharp angles. In that case, the penalty effect is to shrink the $W$ coefficients towards 0. The most well-known estimator falling in this class is the ridge regression obtained with $q = 2$, see Figure 1, Layer 1 and 2. In contrast, for any $q \in (0, 1]$, the $L^q$ norm has some non differentiable points along the axis coordinates, see Figure 1, Layer 2, 3 and 10. Such critical points are more likely to be hit by the level curves of the loss function $R(W)$, 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 1: $L^{2/\ell}$-norm unit balls (in dimension 2) for layers $\ell = 1, 2, 3$ and 10.](image)

### 2.1. MAP on weights 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

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

is equivalent to the weight decay penalty—aka ridge penalty in regression problem—with negative log-prior:

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

where products and sums involving $i$ and $j$ above are over $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_{\ell}$, $L$ is a number of neural network layers, $H_0$ and $H_L$ representing respectively the input and output dimensions.
2.2. MAP on units induces sparsity

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 $\ell$-th layer (either before or after activation). As stated in Theorem 1, Vladimirova et al. (2018) show that conditional on the input $x$, 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

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

for some positive constant $K$. The exponent of $u$ in the exponential term above is optimal in the sense that Equation (1) is not satisfied with some parameter $\theta'$ smaller than $\ell/2$. Thus, the marginal density of $U_m^{(\ell)}$ is approximately proportional to

$$\pi_m^{(\ell)}(u) \approx e^{-|u|^{2/\ell}/K}, \quad u \in \mathbb{R},$$

where $K$ can be chosen the same for all $\ell$-th layer units as the smallest constant satisfying equations (1) for all $m$. The joint prior distribution for all the units $U = (U_m^{(\ell)})_{1 \leq \ell \leq L, 1 \leq m \leq H_\ell}$ can be expressed from all the marginal distributions by Sklar’s representation theorem as

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

where $C$ represents the copula of $U$ (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 $\mathcal{L}(U)$,

$$- \sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(U)) \approx \|U^{(1)}\|_2^2 + \cdots + \|U^{(L)}\|_{2/\ell}^2 - \log C(F(U)).$$

The first $L$ terms in the right-hand-side of (3) indicate that some shrinkage operates at the units level: at layer $\ell$, the penalty term $\|U^{(\ell)}\|_{2/\ell}^2$ takes the form of the $\mathcal{L}^{2/\ell}$ norm. Thus, the deeper the layer, the stronger the sparsity at the level of the units.

3. Conclusion

We offer an interpretation in terms of sparsity-inducing mechanism of the heavy-tailed units distribution result derived by Vladimirova et al. (2018) and reproduced here as Theorem 1. Heavy-tailed priors are known to induce a sparse model representation, such as the Lasso in a regression problem. We extrapolate this finding to the setting of a deep Bayesian neural network, showing that the units tend to be more sparsely represented as layers become deeper.
References


### Appendix A. Definitions

**Definition A.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 $c_1, c_2, d_2 \geq 0, d_1 > 0$ such that the following inequalities hold

\[
|\phi(u)| \geq c_1 + d_1 |u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-,
\]

\[
|\phi(u)| \leq c_2 + d_2 |u| \quad \text{for all } u \in \mathbb{R}.
\]

The interpretation of this property is that $\phi$ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$, at least linearly (first line of (4)), and also at most linearly (second line of (4)). 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, SELU (see Klambauer et al. (2017) for details), and others.

**Definition A.2 (Sub-Weibull random variable)** A random variable $X$, that satisfies

\[
P(|X| \geq x) \leq \exp \left( -x^{1/\theta} / K \right) \quad \text{for all } x \geq 0.
\]

for $K > 0$, is called a sub-Weibull random variable with the tail parameter $\theta$, which is denoted by $X \sim \text{subW}(\theta)$.

Informally, the tails of a subW($\theta$) distribution are dominated by (i.e. decay at least as fast as) the tails of a Weibull variable (Rinne, 2008); 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 $\theta$ and equivalent to sub-Gaussian for $\theta = 1/2$ and sub-Exponential for $\theta = 1$. The larger tail parameter $\theta$, the heavier the tails of the sub-Weibull distribution.