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. Bayesian Statistics in the Big Data Era, Nov 2018, Marseille, France. pp.1. hal-01950660

HAL Id: hal-01950660
https://hal.archives-ouvertes.fr/hal-01950660
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.
Introduction

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities. See Vladimirova et al. (2018).

\[
\text{given an input } x \in \mathbb{R}^N, \text{ the } \ell\text{-th hidden layer unit activation functions are defined as }
\]

\[
g^{(\ell)}(x) = W^{(\ell)} h^{(\ell-1)}(x),
\]

\[
h^{(\ell)}(x) = \phi(g^{(\ell)}(x)).
\]

Covariance theorem

The covariance between hidden units of the same layer is non-negative. Moreover, for any \( \ell \)-th hidden layer units \( h^{(\ell)} \) and \( \hat{h}^{(\ell)} \), for \( s, t \in \mathbb{N} \) it holds

\[
\text{Cov} \left[ \left( h^{(\ell)}(x)^s \right), \left( \hat{h}^{(\ell)}(x)^t \right) \right] \geq 0.
\]

Penalized estimation

Regularized problem:

\[
\min_{W} \mathcal{L}(W) + \lambda \mathcal{Z}(W),
\]

where \( \mathcal{L}(W) \) is a loss function, \( \mathcal{Z}(W) \) is a penalty, \( \lambda > 0 \).

Assumptions

- Gaussian prior on weights:

\[
W_{ij} \sim N(0, \sigma_w^2),
\]

- A nonlinearity \( \phi : \mathbb{R} \to \mathbb{R} \) is said to obey the extended envelope property if there exist \( c_1, c_2, d_1, d_2 > 0 \) such that

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

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

Sub-Weibull

A random variable \( X \), such that

\[
P(|X| \geq x) \leq \exp \left( -x^{\gamma}/K \right)
\]

for all \( x \geq 0 \) and for some \( K > 0 \), is called a sub-Weibull random variable with tail parameter \( \theta > 0 \):

\[
X \sim \text{subW}(\theta).
\]

Moment property:

\[
X \sim \text{subW}(\theta)
\]

implies

\[
||X||_k = \mathbb{E}|X|^{\gamma} \leq k^\theta.
\]

Sparsity interpretation

MAP on weights is L2-regularized Gaussian prior

\[
\pi(W) \propto \prod_{\ell=1}^{L} \sum_{i,j} \exp(-\|W_{ij}\|^2/\lambda),
\]

is equivalent to the weight decay penalty with negative log-prior:

\[
\mathcal{L}(W) \approx \sum_{\ell=1}^{L} \sum_{i,j} \|W_{ij}\|^2 = \|W\|^2_{2}\ll_2.
\]

MAP on units induces sparsity

The joint prior distribution for all the units can be expressed by Sklar’s representation theorem as

\[
\pi(U) = \prod_{\ell=1}^{L} \sum_{i,j} \pi(U_{ij}) C(F(U)),
\]

where \( C \) is the copula of \( U \) (characterizes all the dependence between the units), \( F \) is its cumulative distribution function. The penalty is the negative log-prior:

\[
\mathcal{L}(U) \approx \|U(1)\|^2 + \cdots + \|U(L)\|^2\ll_2 - \log C(F(U)).
\]

Layer \( W \)-penalty \( U \)-penalty

\[
\begin{array}{ccc}
1 & \|W(1)\|^2_{2} & \ell^2 \mid \|U(1)\|^2_{2} \\
2 & \|W(2)\|^2_{2} & \ell^2 \mid \|U(2)\|^2_{2} \\
\ell & \|W(\ell)\|^2_{2} & \ell^2 \mid \|U(\ell)\|^2_{2}\ll_2 \ell^2
\end{array}
\]

Conclusion

We prove that the marginal prior unit distributions are heavier-tailed as depth increases. We further interpret this finding, showing that the units tend to be more sparsely represented as layers become deeper. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

References