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.
Bayesian neural network priors at the level of units

Mariia Vladimirova¹, Julyan Arbel¹ and Pablo Mesejo²

¹Inria Grenoble Rhône-Alpes, France
²University of Granada, Spain

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

Notations
Given an input \( \mathbf{x} \in \mathbb{R}^N \), the \( \ell \)-th hidden layer unit activation functions are defined as
\[
g^{(\ell)}(\mathbf{x}) = \mathbf{W}^{(\ell)} h^{(\ell-1)}(\mathbf{x}),
\]
\[
h^{(\ell)}(\mathbf{x}) = \phi(g^{(\ell)}(\mathbf{x})).
\]

Assumptions
- **Gaussian prior** on weights:
  \( W_{i,j} \sim N(0, \sigma^2_w) \),
- **A nonlinearity** \( \phi : \mathbb{R} \to \mathbb{R} \) is said to obey the extended envelope property if there exist \( c_1, c_2, d_1, d_2 \geq 0 \) 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
\[
P(|X| \geq x) \leq \exp\left(-x^\beta/\bar{K}\right)
\]
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:
\[
X \sim \text{subW}(\theta) \quad \Rightarrow \quad \|X\|_k = \left(\mathbb{E}|X|^{\beta}\right)^{\frac{1}{\beta}} \leq k^\theta.
\]

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 \( h^{(\ell)} \), for \( s, t \in \mathbb{N} \) it holds
\[
\text{Cov}\left(h^{(\ell)}(\mathbf{x}), h^{(\ell)}(\mathbf{y})\right) \geq 0.
\]

Penalized estimation
Regularized problem:
\[
\min_{W} \mathcal{L}(W) + \lambda \mathcal{L}(W),
\]
where \( \mathcal{L}(W) \) is a loss function, \( \mathcal{L}(W) \) is a penalty, \( \lambda > 0 \).
For Bayesian models with prior distribution \( P(W) \), the maximum a posteriori (MAP) solves (1) with:
\[
\mathcal{L}(W) \propto -\log P(W)
\]

Sparsity interpretation
MAP on weights is L2-reg.
Independent Gaussian prior
\[
\pi(W) \propto \prod_{l=1}^L \prod_{j=1}^m \exp(-\frac{1}{2}\|W^{(l)}_{i,j}\|^2),
\]
is equivalent to the weight decay penalty with negative log-prior:
\[
\mathcal{L}(W) \propto \sum_{l=1}^L \sum_{j=1}^m (W^{(l)}_{i,j})^2 = \|W^{(l)}\|^2_{l^2},
\]

MAP on units induces sparsity
The joint prior distribution for all the units can be expressed by Sklar’s representation theorem as
\[
\pi(U) = \prod_{l=1}^L \prod_{m=1}^M \exp(-\frac{1}{2}\|U^{(l)}_{i,m}\|^2) C(F(U)),
\]
where \( C \) is the copula of \( U \) (characterizes all the dependence between the units), \( F \) is the cumulative distribution function. The penalty is the negative log-prior:
\[
\mathcal{L}(U) \propto \|U^{(l)}\|^2_{l^2} + \cdots + \|U^{(L)}\|^2_{l^2} - \log C(F(U)).
\]

Prior distributions of layers \( \ell = 1, 2, 3 \)
Illustration of units marginal prior distributions from the first three hidden layers. Neural network parameters: \((N, H_1, H_2, H_3) = (50, 25, 24, 4)\).

Proof sketch
Induction w.r.t. layer depth \( \ell \):
\[
\|h^{(\ell)}\|_k \leq k^\theta,
\]
which is the moment characterization of sub-Weibull variable.
- **Extended envelope property** implies \( \|h^{(\ell)}\|_k \geq \|g^{(\ell)}\|_k \)
- **Base step:** \( g \sim N(0, \sigma^2) \), \( \|g\|_k \geq \sqrt{k} \).
- **Upper bound:** Holder’s inequality implies \( \|h^{(\ell)}\|_k \geq \|g^{(\ell)}\|_k \geq k^\theta \).

References