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).

![Diagram of Bayesian neural network layers](image)

Notations

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

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

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

Assumptions

- **Gaussian prior on weights:** $W_{ij} \sim N(0, \sigma^2_w)$.
- **A nonlinearity $\phi: \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property** if there exist $c_1, c_2, d_1, d_2 > 0$, $d_1 > 0$ such that

$|\phi(u)| \leq c_1 + d_1|u|$ for $u \in \mathbb{R}_+$,

$|\phi(u)| \leq c_2 + d_2|u|$ 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 $x \geq 0$, 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

$\|X\|_k = \mathbb{E}[|X|^{\frac{\gamma}{\theta}}] \leq K^{\frac{\theta}{\gamma}}$

Meaning for all $k \in \mathbb{N}$ and for some constants $d, D > 0$, $d < \|X\|_k^{\frac{\theta}{\gamma}} < D$.

Covariance theorem

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

$\text{Cov}\left[h^{(\ell)}(x), h^{(\ell)}(y)\right] \geq 0$.

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, $\lambda > 0$.

For Bayesian models with prior distribution $\pi(W)$, the maximum a posteriori (MAP) solves (1) with:

$\mathcal{Z}(W) \propto -\log \pi(W)$

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 $\ell$:

$\|h^{(\ell)}\|_k \approx k^{\frac{\gamma}{\theta}}$

which is the moment characterization of sub-Weibull variable.

- **Extended envelope property** implies $\|h^{(\ell)}\|_k \propto \|g^{(\ell)}\|_k$
- **Base step:** $g \sim N(0, \sigma^2)$,

$\|g\|_k \approx \sqrt{K}$.

Thus, $\|h\|_k = \|\phi(g)\|_k \approx \|g\|_k \approx \sqrt{K}$.

- **Inductive step:** suppose $\|h^{(\ell-1)}\|_k \approx k^{\frac{\gamma}{\theta}}$.

Lower bound: non-negative covariance theorem:

$\text{Cov}\left[h^{(\ell-1)}(x), h^{(\ell-1)}(y)\right] \geq 0$.

Upper bound: Holder’s inequality:

$g^{(\ell)} = \sum_{j=1}^H W^{(\ell)}_{ij} h^{(\ell-1)}(x)$ implies

$\|h^{(\ell)}\|_k \approx \|g^{(\ell)}\|_k \approx k^{\frac{\gamma}{\theta}}$.

Sparsity interpretation

MAP on weights is L2-regularized Gaussian prior

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

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

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

MAP on units induces sparsity

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

$\pi(U) = \prod_{\ell=1}^L \prod_{i,j} e^{-\frac{1}{2}U_{ij}^2} C(F(U))$.

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