Bayesian neural network priors at the level of units

Mariia Vladimirova, Julyan Arbel, Pablo Mesejo

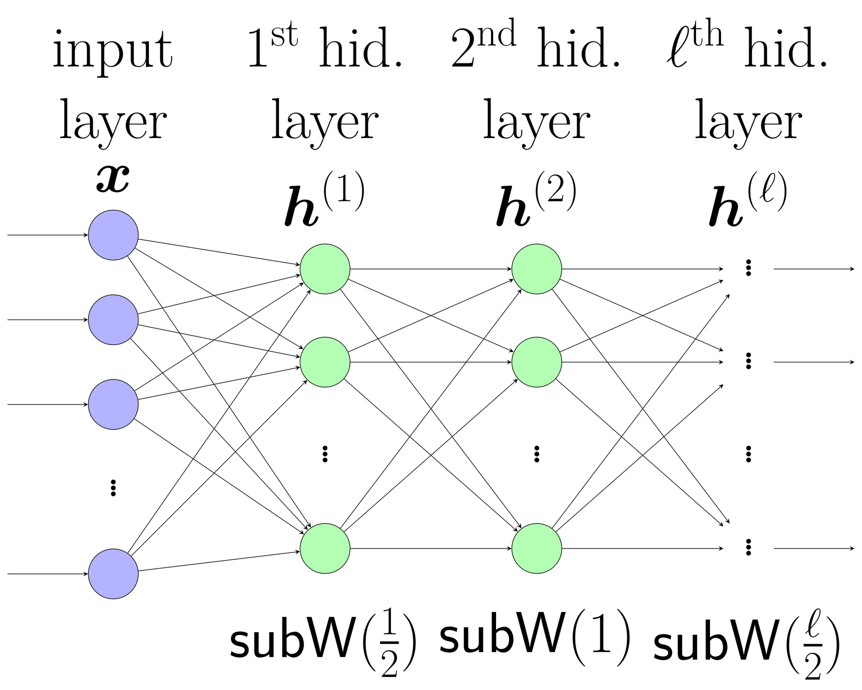
Bayesian neural network priors at the level of units

Mariia Vladimirova[1], Julyan Arbel[1] and Pablo Mesejo[2]

[1]Inria Grenoble Rhône-Alpes, France
[2]University 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 $\boldsymbol{x} \in \mathbb{R}^N$, the ℓ-th hidden layer unit activations are defined as

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

Assumptions

- **Gaussian prior** on *weights*:
$$W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2),$$
- A *nonlinearity* $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the **extended envelope property** if there exist $c_1, c_2, d_2 \geq 0$, $d_1 > 0$ such that
$$|\phi(u)| \geq c_1 + d_1|u| \quad \text{for } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-,$$
$$|\phi(u)| \leq c_2 + d_2|u| \quad \text{for } u \in \mathbb{R}.$$

Sub-Weibull

A random variable X, such that

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

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

$$X \sim \mathsf{subW}(\theta).$$

Moment property:
$X \sim \mathsf{subW}(\theta)$ implies

$$\|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k} \asymp k^\theta,$$

Meaning for all $k \in \mathbb{N}$ and for some constants $d, D > 0$,

$$d < \|X\|_k / k^\theta < D.$$

Covariance theorem

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

$$\mathrm{Cov}\left[\left(h^{(\ell)}\right)^s, \left(\tilde{h}^{(\ell)}\right)^t\right] \geq 0.$$

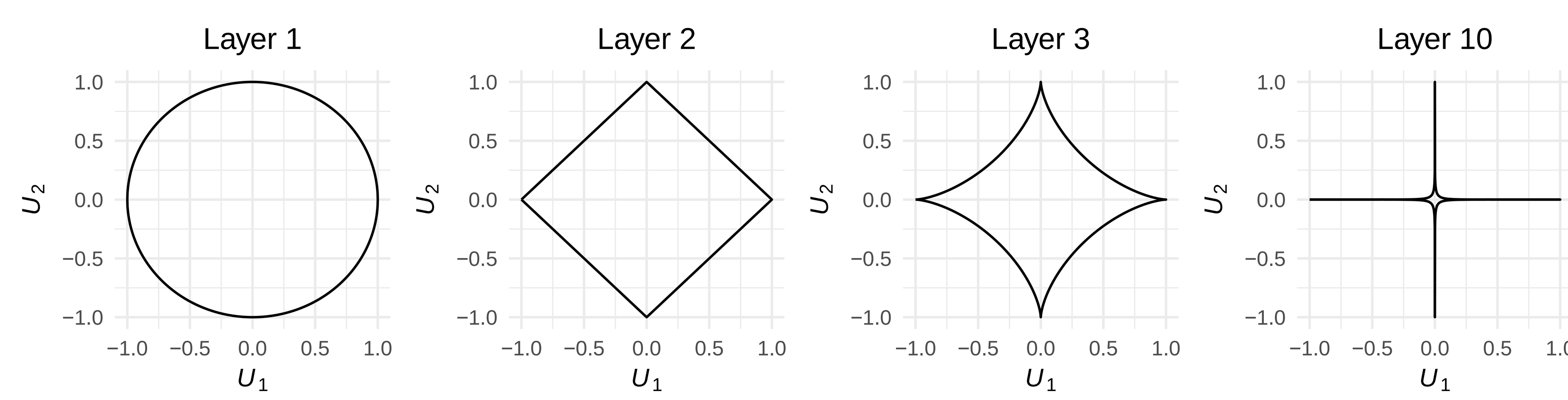
Penalized estimation

Regularized problem:

$$\min_{\boldsymbol{W}} \mathscr{R}(\boldsymbol{W}) + \lambda\mathscr{L}(\boldsymbol{W}), \qquad (1)$$

where $\mathscr{R}(\boldsymbol{W})$ is a **loss function**, $\lambda\mathscr{L}(\boldsymbol{W})$ is a **penalty**, $\lambda > 0$.
For Bayesian models with prior distribution $\pi(\boldsymbol{W})$, the **maximum a posteriori** (MAP) solves (1) with:

$$\mathscr{L}(\boldsymbol{W}) \propto -\log \pi(\boldsymbol{W})$$



Theorem (Vladimirova et al., 2018)

The ℓ-th hidden layer units $U^{(\ell)}$ (pre-activation $g^{(\ell)}$ or post-activation $h^{(\ell)}$) of a feed-forward Bayesian neural network with:

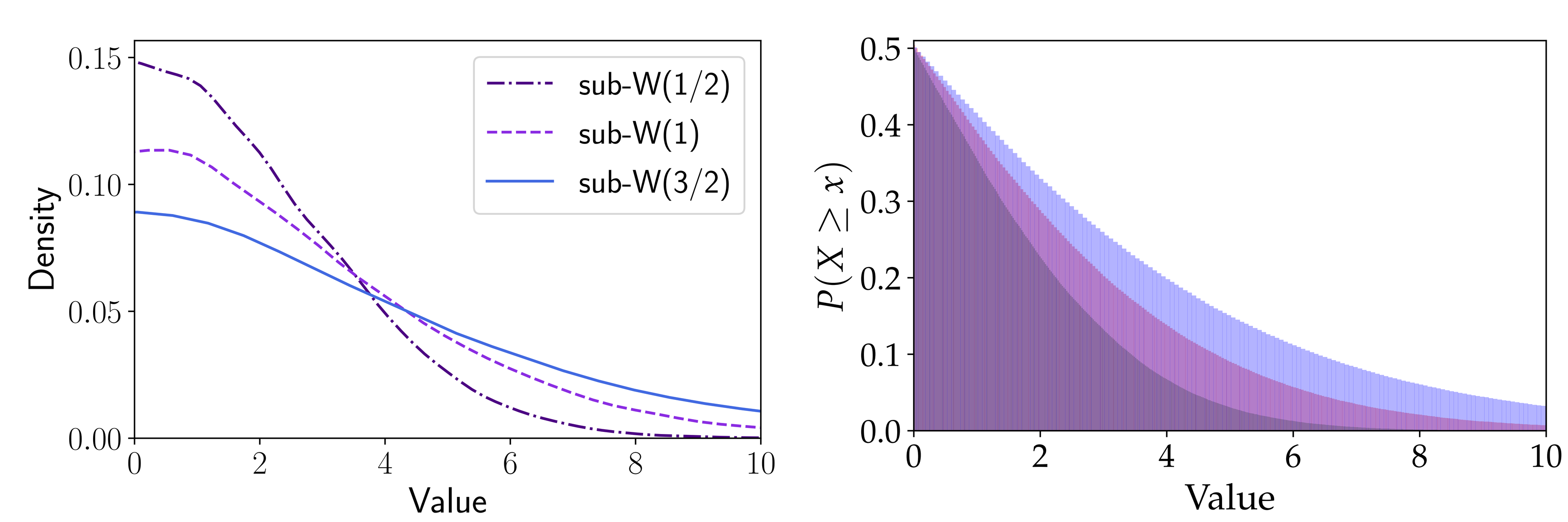
- **Gaussian priors** on weights and
- **extended envelope condition** activation function ϕ

have **sub-Weibull marginal prior distribution** with optimal tail parameter $\theta = \ell/2$, conditional on the input $\boldsymbol{x}$:

$$U^{(\ell)} \sim \mathsf{subW}(\ell/2),$$

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 ℓ:

$$\|h^{(\ell)}\|_k \asymp k^{\ell/2},$$

which is the moment characterization of sub-Weibull variable.

- Extended envelope property implies $\|h^{(\ell)}\|_k \asymp \|g^{(\ell)}\|_k$
- **Base step:** $g \sim N(0, \sigma^2)$, $\|g\|_k \asymp \sqrt{k}$. Thus,
$$\|h\|_k = \|\phi(g)\|_k \asymp \|g\|_k \asymp \sqrt{k}.$$
- **Inductive step:** suppose
$$\|h^{(\ell-1)}\|_k \asymp k^{(\ell-1)/2}.$$
- Lower bound: non-negative covariance theorem:
$$\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0.$$
- Upper bound: Holder's inequality
- $g^{(\ell)} = \sum_{j=1}^{H} W_{i,j}^{(\ell-1)} h_j^{(\ell-1)}$ implies
$$\|h^{(\ell)}\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}.$$

Sparsity interpretation

MAP on weights is L2-reg.
Independent Gaussian prior

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

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

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

MAP on units induces *sparsity*
The joint prior distribution for all the units can be expressed by Sklar's representation theorem as

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

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

$$\mathscr{L}(\boldsymbol{U}) \approx \|\boldsymbol{U}^{(1)}\|_2^2 + \cdots + \|\boldsymbol{U}^{(L)}\|_{2/L}^{2/L} - \log C(F(\boldsymbol{U})).$$

Layer	$\boldsymbol{W}$-penalty	$\boldsymbol{U}$-penalty	
1	$\|\boldsymbol{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(1)}\|_2^2$,	$\mathcal{L}^2$
2	$\|\boldsymbol{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(2)}\|$,	$\mathcal{L}^1$
ℓ	$\|\boldsymbol{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\boldsymbol{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$,	$\mathcal{L}^{2/\ell}$

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

Vladimirova, M., Arbel, J., and Mesejo, P. (2018). Bayesian neural networks increasingly sparsify their units with depth. *arXiv preprint arXiv:1810.05193*.