Understanding Priors in Bayesian Neural Networks at the Unit Level

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, Julyan Arbel

Inria, Grenoble, France

✉ mariia.vladimirova@inria.fr

International Conference on Machine Leaning
June 13, 2019

Outline

Sub-Weibull distributions

Main result: Prior on units gets heavier-tailed with depth

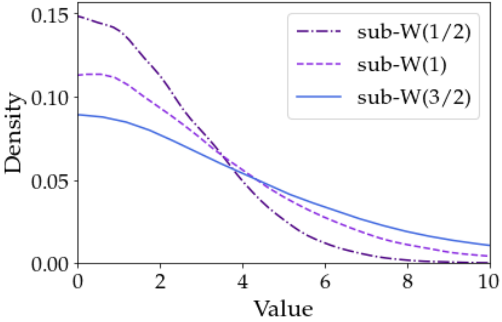
Regularization interpretation

Distribution families with respect to tail behavior

For all $k \in \mathbb{N}$, k-th row moment: $\|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k}$

Distribution	Tail	Moments
Sub-Gaussian	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^2}$	$\|X\|_k \leq C\sqrt{k}$
Sub-Exponential	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x}$	$\|X\|_k \leq Ck$
Sub-Weibull	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^{1/\theta}}$	$\|X\|_k \leq Ck^\theta$

- $\theta > 0$ called tail parameter
- $\|X\|_k \asymp k^\theta \implies X \sim \mathrm{subW}(\theta)$, θ called optimal
- $\mathrm{subW}(1/2) = \mathrm{subG}$, $\mathrm{subW}(1) = \mathrm{subE}$
- Larger θ implies heavier tail



Vladimirova et al. Understanding priors in Bayesian neural networks at the unit level

Outline

Sub-Weibull distributions

Main result: Prior on units gets heavier-tailed with depth

Regularization interpretation

Assumptions on neural network

We analyze Bayesian neural networks which satisfy the following assumptions

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(\mu, \sigma^2)$$

(A2) Nonlinearity. ReLU-like with envelope property: exist $c_1, c_2, d_2 \geq 0$, $d_1 > 0$ s.t.

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned}$$

- Examples: ReLU, ELU, PReLU etc, but no compactly supported like sigmoid and tanh.
- Nonlinearity does not harm the distributional tail:

$$\|\phi(X)\|_k \asymp \|X\|_k, \quad k \in \mathbb{N}$$

Assumptions on neural network

We analyze Bayesian neural networks which satisfy the following assumptions

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(\mu, \sigma^2)$$

(A2) Nonlinearity. ReLU-like with envelope property: exist $c_1, c_2, d_2 \geq 0$, $d_1 > 0$ s.t.

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned}$$

- Examples: ReLU, ELU, PReLU etc, but no compactly supported like sigmoid and tanh.
- Nonlinearity does not harm the distributional tail:

$$\|\phi(X)\|_k \asymp \|X\|_k, \quad k \in \mathbb{N}$$

Assumptions on neural network

We analyze Bayesian neural networks which satisfy the following assumptions

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(\mu, \sigma^2)$$

(A2) Nonlinearity. ReLU-like with envelope property: exist $c_1, c_2, d_2 \geq 0$, $d_1 > 0$ s.t.

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned}$$

- Examples: ReLU, ELU, PReLU etc, but no compactly supported like sigmoid and tanh.
- Nonlinearity does not harm the distributional tail:

$$\|\phi(X)\|_k \asymp \|X\|_k, \quad k \in \mathbb{N}$$

Assumptions on neural network

We analyze Bayesian neural networks which satisfy the following assumptions

(A1) Parameters. The weights w have i.i.d. Gaussian prior

$$w \sim \mathcal{N}(\mu, \sigma^2)$$

(A2) Nonlinearity. ReLU-like with envelope property: exist $c_1, c_2, d_2 \geq 0$, $d_1 > 0$ s.t.

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned}$$

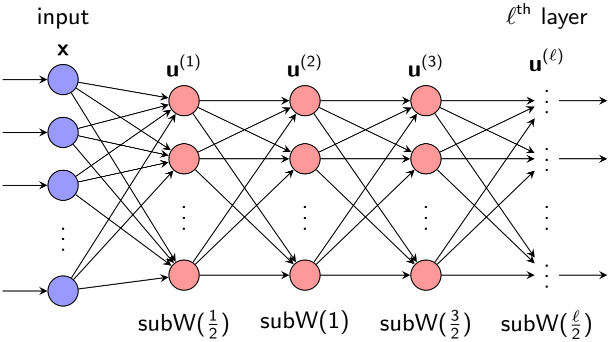
- Examples: ReLU, ELU, PReLU etc, but no compactly supported like sigmoid and tanh.
- Nonlinearity does not harm the distributional tail:

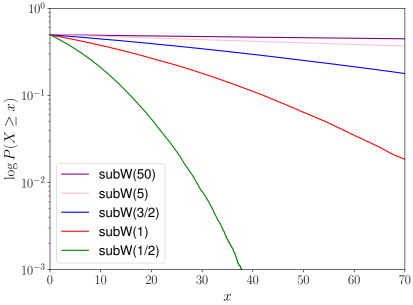
$$\|\phi(X)\|_k \asymp \|X\|_k, \quad k \in \mathbb{N}$$

Main theorem

Consider a Bayesian neural network with (A1) i.i.d. Gaussian priors on the weights and
(A2) nonlinearity satisfying envelope property.

Then conditional on input $\mathbf{x}$, the marginal prior distribution of a unit $u^{(\ell)}$ of ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim \text{subW}(\ell/2)$





Vladimirova et al. Understanding priors in Bayesian neural networks at the unit level

Outline

Sub-Weibull distributions

Main result: Prior on units gets heavier-tailed with depth

Regularization interpretation

Interpretation: shrinkage effect

Maximum a Posteriori (MAP) is a Regularized problem

$$\max_{\mathbf{W}} \pi(\mathbf{W}|\mathcal{D}) \propto \mathcal{L}(\mathcal{D}|\mathbf{W})\pi(\mathbf{W})$$

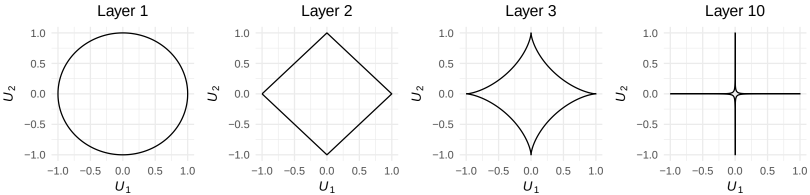
$$\min_{\mathbf{W}} -\log \mathcal{L}(\mathcal{D}|\mathbf{W}) - \log \pi(\mathbf{W})$$

$$\min_{\mathbf{W}} L(\mathbf{W}) + \lambda R(\mathbf{W})$$

$L(\mathbf{W})$ is a loss function,
$R(\mathbf{W})$ is a norm on $\mathbb{R}^p$, regularizer.

Weight distribution $\pi(w) \approx e^{-w^2}$ $\Rightarrow$ ℓ-th layer unit distribution $\pi^{(\ell)}(u) \approx e^{-u^{2/\ell}}$

Layer	Penalty on $\mathbf{W}$	Penalty on $\mathbf{U}$	
1	$\|\mathbf{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\mathbf{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\mathbf{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$



Vladimirova et al. Understanding priors in Bayesian neural networks at the unit level

Conclusion

(i) We define the notion of sub-Weibull distributions, which are characterized by tails lighter than (or equally light as) Weibull distributions.

(ii) We prove that the marginal prior distribution of the units are heavier-tailed as depth increases.

(iii) We offer an interpretation from a regularization viewpoint.

Future directions:

- Prove the Gaussian process limit of sub-Weibull distributions in the wide regime;
- Investigate if the described regularization mechanism induces sparsity at the unit level.