Paper Review

Bayesian neural networks become heavier-tailed with depth

Mariia Vladimirova, Julyan Arbel, Pablo Mesejo

Marat Khusainov

September 2023

Outline

Motivation & Problem statement

- What are hidden units prior distributions in Bayesian neural networks under assumption of independent Gaussian weights?

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

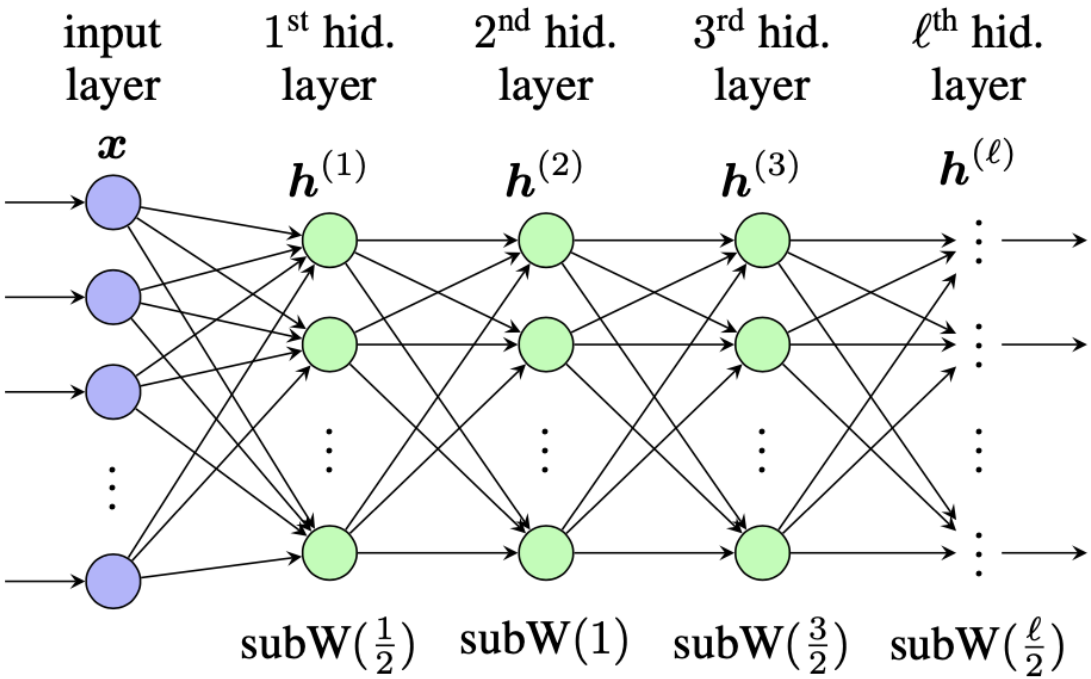


Figure: Neural network architecture and characterization of the ℓ-layer units prior distribution as sub-Weibull distribution with tail parameter $\ell/2$.

Definitions & Assumption on neural network

Definition (Sub-Weibull random variable)

A random variable X, that satisfies

$$\mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K\right) \qquad \text{for all } x \geq 0.$$

for $K > 0$, is called a sub-Weibull random variable with the tail parameter $\theta > 0$, which is denoted by $X \sim \text{subW}(\theta)$.

Let all weights (including biases) be independent and have zero-mean normal distribution $W_{i,j}^{(\ell)} \sim \mathcal{N}\left(0, \sigma_w^2\right)$, for all $1 \leq \ell \leq L, 1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$.

Definition (Extended envelope property for nonlinearities)

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 the following inequalities hold

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

Main theorem

Theorem (Sub-Weibull units)

Consider a feed-forward Bayesian neural network with Gaussian priors with nonlinearity ϕ satisfying the extended envelope condition. Then conditional on the input $\boldsymbol{x}$, the marginal prior distribution induced by forward propagation on any unit (pre- or post-nonlinearity) of the ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,

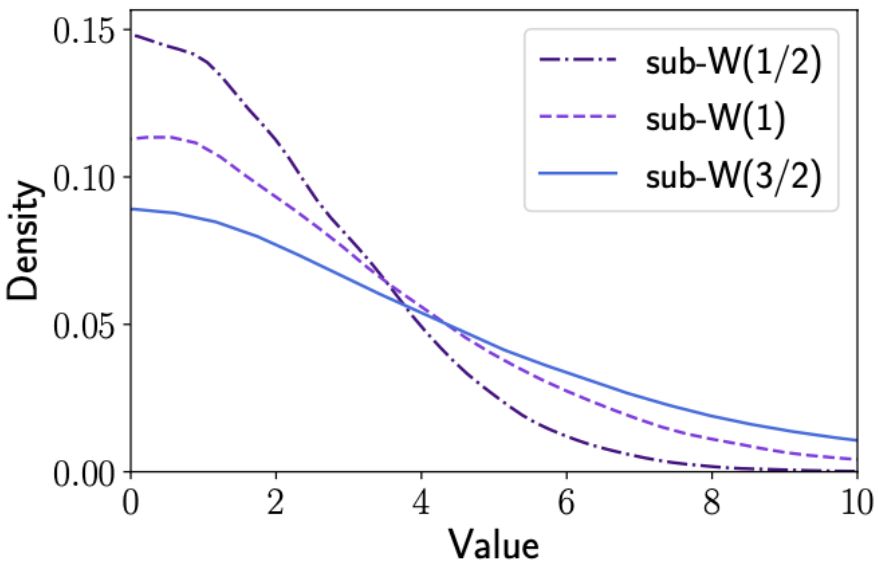
$$U_m^{(\ell)} \sim \text{subW}(\ell/2)$$

where $U_m^{(\ell)}$ is either a pre-nonlinearity $g_m^{(\ell)}$ or a post-nonlinearity $h_m^{(\ell)}$.

Experiment

Densities are obtained as kernel density estimators from a sample of size 10^5 from the prior on the pre-nonlinearities, which is itself obtained by sampling 10^5 sets of weights $\boldsymbol{W}$ from the Gaussian prior and forward propagation.

- 3 hidden layers of NN have $H_1 = 25, H_2 = 24$ and $H_3 = 4$ hidden units
- The nonlinearity ϕ is the ReLU function
- $\boldsymbol{x} \in \mathbb{R}^{50}$ is sampled from a standard normal distribution.



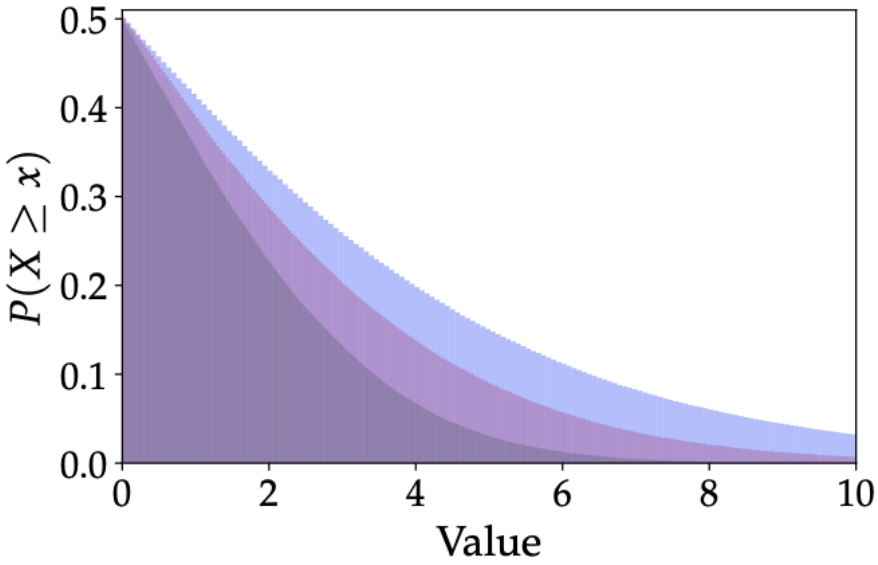


Figure: Illustration of the first three layers hidden units marginal prior distributions.