Bayesian neural networks become heavier-tailed with depth

Mariia Vladimirova

PhD student in Inria Rhone-Alpes,
Master student at MIPT

✉ mariia.vladimirova@inria.fr

Intelligent Data Processing 2018, Gaeta, Italy
October 8, 2018

Problem statement

Goal

Study deep models properties.

Challenges

- Deep learning models are excessively redundant.
- The model interpretability is lost because of its complexity.
- Uncertainty in the model predictions is hard to estimate.

Solution

Investigate models using Bayesian inference by assuming a prior distribution on their parameters.

Bayesian approach

Bayes theorem

$$\pi(\mathbf{w}|\mathcal{D}) = \frac{\pi(\mathbf{w})\pi(\mathcal{D}|\mathbf{w})}{\pi(\mathcal{D})}.$$

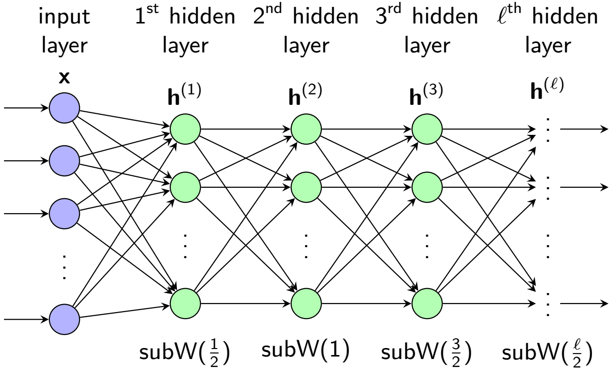
$\pi(\mathbf{w}|\mathcal{D})$ is a posterior of model parameters $\mathbf{w}$ given data $\mathcal{D}$
$\pi(\mathbf{w})$ is a prior distribution
$\pi(\mathcal{D}|\mathbf{w})$ is data likelihood
$\pi(\mathcal{D})$ is a normalization constant, evidence, given by
$\pi(\mathcal{D}) = \int \pi(\mathcal{D}|\mathbf{w})\pi(\mathbf{w})\mathbf{dw}$.

$+$ allows to obtain the uncertainty of model outcomes

$-$ the posterior becomes intractable for large models

$\pm$ the prior distribution choice

Neural network structure



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

$\phi(\cdot)$ — nonlinearity, $\mathbf{g}$ — pre-nonlinearity, $\mathbf{h}$ — post-nonlinearity

Distibution families with respect to tail behavior

$\|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k}$, for all $k \in \mathbb{N}$,
tail parameter $\theta > 0$

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	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^{1/\theta}}$	$\|X\|_k \leq Ck^\theta$

Assumptions on neural network

To prove that Bayesian neural networks become heavier-tailed with depth we make assumptions on:

Parameters. i.i.d with Gaussian prior

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

Nonlinearity. ReLU-like with envelope property: exist $c, m \geq 0$ s.t.

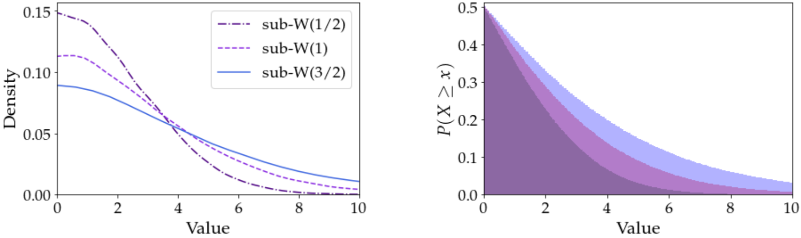
$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \text{for all } u \in \mathbb{R}_+ \text{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \text{for all } u \in \mathbb{R}. \end{aligned}$$

Examples: ReLU, ELU, PReLU etc.

Main theorem

Theorem (Vladimirova, 2018)

Consider a Bayesian neural network with Gaussian parameters and nonlinearity satisfying envelope property. Then a unit of ℓ-th hidden layer $h^{(\ell)}$ follows sub-Weibull distribution with optimal tail parameter $\theta = \ell/2$.



(a) Probability densities

(b) Distribution tails

Marginal distibutions:

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

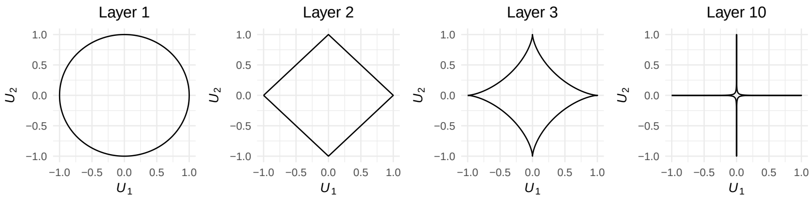
Interpretation: shrinkage effect

Regularized problem:

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

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

Figure: $\mathcal{L}^{2/\ell}$-norm unit balls (in dimension 2) for layers $\ell = 1, 2, 3$ and 10.



MAP on weights $\mathbf{W}$ is weight decay

Maximum A Posteriori (MAP):

$$\pi(\mathbf{W}|\mathcal{D}) \sim \pi(\mathbf{W})\pi(\mathcal{D}|\mathbf{W}) \quad \to \quad \max$$
$$-\log\pi(\mathbf{W}) - \log\pi(\mathcal{D}|\mathbf{W}) \quad \to \quad \min$$

Gaussian prior on the weights:

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

Equivalent to the *weight decay* penalty ($\mathcal{L}^2$):

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

MAP on units $\mathbf{U}$ induces sparsity

Marginal distibutions:

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

Sklar's representation theorem:

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

where C represents the copula of $\mathbf{U}$ (which characterizes all the dependence between the units).

$$\begin{aligned} R(\mathbf{U}) &= -\sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\mathbf{U})), \\ &\approx \sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} |U_m^{(\ell)}|^{2/\ell} - \log C(F(\mathbf{U})), \\ &\approx \|\mathbf{U}^{(1)}\|_2^2 + \|\mathbf{U}_1^{(2)}\|_1 + \cdots + \|\mathbf{U}^{(L)}\|_{2/L}^{2/L} - \log C(F(\mathbf{U})). \end{aligned}$$

MAP on units **U** induces sparsity

Regularizer:

$$R(\mathbf{U}) \approx \|\mathbf{U}^{(1)}\|_2^2 + \|\mathbf{U}_1^{(2)}\|_1 + \cdots + \|\mathbf{U}^{(L)}\|_{2/L}^{2/L} - \log C(F(\mathbf{U})).$$

Comparison of Bayesian neural network shrinkage effect on weights **W** and units **U**:

Layer	Penalty on **W**	Penalty on **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	$\|\mathbf{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$	$\mathcal{L}^{2/\ell}$

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 proved that the marginal prior distribution of the units are *heavier-tailed* as depth increases.

(iii) We offer an interpretation from a *sparsity-inducing viewpoint*.

Future directions:

- a precise description of the copula would provide valuable information about the dependence between the units;
- an interpretation of our result in terms of the full posterior distribution would give an ability to uncertainty;
- Bayesian deep neural networks distributional properties and their sparsifying mechanisms.