Understanding Priors in Bayesian Neural Networks at the Unit Level

Mariia Vladimirova

Joint work with Julyan Arbel, Jakob Verbeek

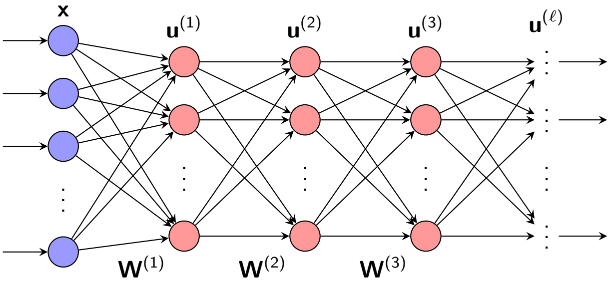
✉ mariia.vladimirova@inria.fr

Machine Learning Reading Group, Grenoble
July 3, 2019

Bayesian neural networks: why?

Neal [1996], MacKay [1992]

- Prior on weights, $w \sim N(\mu, \sigma^2)$
- Allows to model uncertainty
- Represents a standard neural network



Outline

Recent works on distributional properties

Sub-Weibull distributions

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

Regularization interpretation

Wide regime: infinite number of hidden units in the layer

Theorem (Neal [1996])

Consider a Bayesian neural network with (A1) iid Gaussian priors on the weights
(A2) with bounded variances and
(A3) ReLU activation function. Then conditional on input $\mathbf{x}$*,*
the marginal prior distribution of a unit $u^{(2)}$ *of 2-nd hidden layer converges to a Gaussian process in a wide regime.*

Proof.
Components of $\mathbf{u}^{(1)}$ are iid $\Rightarrow$ CLT ☐

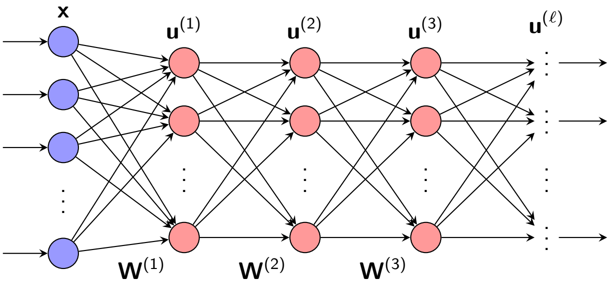
Wide regime: infinite number of hidden units in the layer

Theorem (Neal [1996])

Consider a Bayesian neural network with (A1) iid Gaussian priors on the weights
(A2) with bounded variances and
(A3) ReLU activation function. Then conditional on input $\mathbf{x}$, *the marginal prior distribution of a unit* $u^{(2)}$ *of 2-nd hidden layer converges to a Gaussian process in a wide regime.*

Proof.

Components of $\mathbf{u}^{(1)}$ are iid $\Rightarrow$ CLT $\square$



Wide regime: infinite number of hidden units in the layer

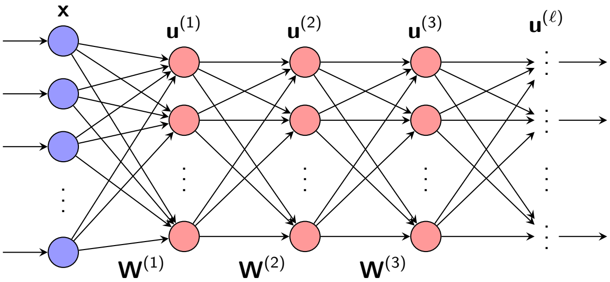
Theorem (Matthews et al. [2018], Lee et al. [2018])

Conditional on input $\mathbf{x}$*, the marginal prior distribution of a unit* $u^{(\ell)}$ *of* ℓ*-nd hidden layer converges to a Gaussian process in a wide regime.*

Wide regime: infinite number of hidden units in the layer

Theorem (Matthews et al. [2018], Lee et al. [2018])

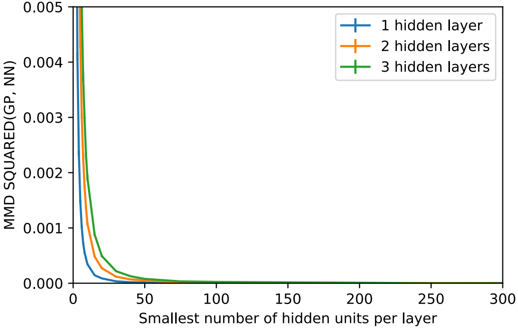
Conditional on input $\mathbf{x}$, *the marginal prior distribution of a unit* $u^{(\ell)}$ *of* ℓ*-nd hidden layer converges to a Gaussian process in a wide regime.*



Wide regime: infinite number of hidden units in the layer

Theorem (Matthews et al. [2018], Lee et al. [2018])

Conditional on input $\mathbf{x}$*, the marginal prior distribution of a unit* $u^{(\ell)}$ *of* ℓ*-nd hidden layer converges to a Gaussian process in a wide regime.*



Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. "Gaussian Process Behaviour in Wide Deep Neural Networks." ICLR (2018).

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. "Deep neural networks as Gaussian processes." ICLR (2018).

Gaussian process approximation

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, Jascha Sohl-Dickstein. "Deep Information Propagation." ICLR (2017).

- Prior on weights, $w \sim N(0, \sigma^2)$ iid
- Initialisation is a crucial step in deep NN
- "Edge of Chaos" initialization can lead to good performances

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. "On the Impact of the Activation Function on Deep Neural Networks Training." ICML (2019).

- Prior on weights, $w \sim N(0, \sigma^2)$ iid
- Gaussian process approximation $u^\ell \approx \mathcal{GP}(0, K^\ell)$ marginally
- "Edge of Chaos" initialization

Results:

- Smooth activation functions (e.g. ELU) are better than ReLU activation, especially if ℓ is large
- "Edge of Chaos" accelerates the training and improves performances

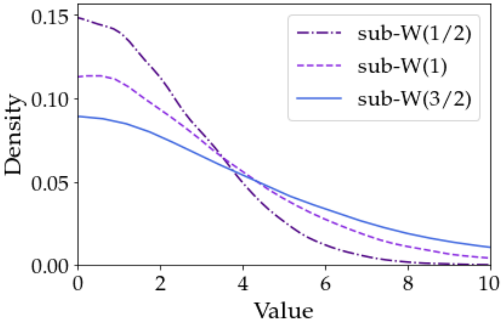
Outline

Distribution families with respect to tail behavior

Vershynin [2018] – subG, subE;
Kuchibhotla and Chakrabortty [2018]; Vladimirova and Arbel [2019] – subW

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	$\overline{F}(x) \leq \mathrm{e}^{-\lambda x^2}$	$\|X\|_k \leq C\sqrt{k}$
$\subset$ 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
- $\mathrm{subW}(1/2) = \mathrm{subG}$, $\mathrm{subW}(1) = \mathrm{subE}$

Outline

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.

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

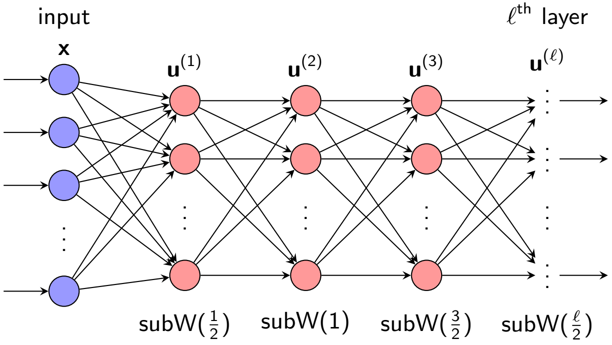
- Examples: ReLU, ELU, PReLU etc, but no compactly supported like sigmoid and tanh.
- Nonlinearity does not harm the distributional tail:

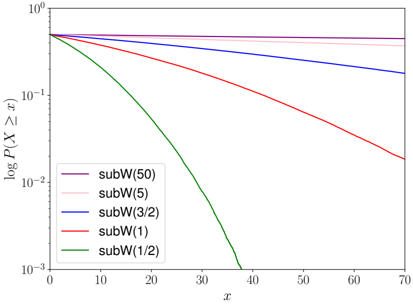
$$\|\phi(X)\|_k \asymp \|X\|_k, \quad k \in \mathbb{N}$$

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 $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim \text{subW}(\ell/2)$





Proof sketch I

Recall. $X \sim \text{subW}(\theta) \iff \exists C > 0, \|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k} \leq Ck^\theta, \quad \text{for all } k \in \mathbb{N}.$

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

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

$$\mathbf{g}^{(\ell)}(\mathbf{x}) = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)}(\mathbf{x}), \quad \mathbf{h}^{(\ell)}(\mathbf{x}) = \phi(\mathbf{g}^{(\ell)}), \quad \ell = \{2, \dots, L\}.$$

Goal. By induction with respect to hidden layer depth ℓ we want to show that

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

Proof sketch I

Recall. $X \sim \text{subW}(\theta) \iff \exists C > 0, \|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k} \leq Ck^\theta$, for all $k \in \mathbb{N}$.

Notations. $\phi(\cdot)$ — nonlinearity, **g** — pre-nonlinearity, **h** — post-nonlinearity

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

$$\mathbf{g}^{(\ell)}(\mathbf{x}) = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)}(\mathbf{x}), \quad \mathbf{h}^{(\ell)}(\mathbf{x}) = \phi(\mathbf{g}^{(\ell)}), \quad \ell = \{2, \ldots, L\}.$$

Goal. By induction with respect to hidden layer depth ℓ we want to show that

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

Proof sketch I

Recall. $X \sim \mathrm{subW}(\theta) \iff \exists C > 0, \|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k} \leq Ck^\theta, \quad \text{for all } k \in \mathbb{N}.$

Notations. $\phi(\cdot)$ — nonlinearity, **g** — pre-nonlinearity, **h** — post-nonlinearity

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

$$\mathbf{g}^{(\ell)}(\mathbf{x}) = \mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)}(\mathbf{x}), \quad \mathbf{h}^{(\ell)}(\mathbf{x}) = \phi(\mathbf{g}^{(\ell)}), \quad \ell = \{2, \ldots, L\}.$$

Goal. By induction with respect to hidden layer depth ℓ we want to show that

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

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\| \sum_{i=1}^{H_1} w_i^{(1)} x_i \right\|_k \asymp k^{1/2}$$

   From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\| \sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)} \right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

   2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s, t$, $\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$
   2.2 Upper bound for $(\star)$ by Hölder's inequality

   From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\| \sum_{i=1}^{H_1} w_i^{(1)} x_i \right\|_k \asymp k^{1/2}$$

   From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\| \sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)} \right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

   2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s, t$, $\mathrm{Cov}[(h^{(\ell-1)})^s, (\tilde{h}^{(\ell-1)})^t] \geq 0$
   2.2 Upper bound for $(\star)$ by Hölder's inequality

   From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\|\sum_{i=1}^{H_1} w_i^{(1)} x_i\right\|_k \asymp k^{1/2}$$

From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\|\sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)}\right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s,t$, $\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$
2.2 Upper bound for $(\star)$ by Hölder's inequality

From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\| \sum_{i=1}^{H_1} w_i^{(1)} x_i \right\|_k \asymp k^{1/2}$$

From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\| \sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)} \right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s, t$, $\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$

2.2 Upper bound for $(\star)$ by Hölder's inequality

From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\| \sum_{i=1}^{H_1} w_i^{(1)} x_i \right\|_k \asymp k^{1/2}$$

From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\| \sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)} \right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s, t$, $\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$
2.2 Upper bound for $(\star)$ by Hölder's inequality

From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Proof sketch II

1. Base step: weights $w_i^{(1)}$ are iid Gaussian $\Rightarrow \|w\|_k \asymp k^{1/2}$; for 1st layer

$$\|g^{(1)}\|_k = \left\| \sum_{i=1}^{H_1} w_i^{(1)} x_i \right\|_k \asymp k^{1/2}$$

From nonlinearity ϕ assumption

$$\|h^{(1)}\|_k = \|\phi(g^{(1)})\|_k \asymp \|g^{(1)}\|_k \asymp k^{1/2}$$

2. Induction step: if $g^{(\ell-1)}, h^{(\ell-1)} \sim subW((\ell-1)/2)$, then for ℓ-th layer

$$\|g^{(\ell)}\|_k = \left\| \sum_{i=1}^{H} w_i^{(\ell)} h_i^{(\ell-1)} \right\|_k \overset{(\star)}{\asymp} k^{1/2} \cdot k^{(\ell-1)/2} = k^{\ell/2}$$

2.1 Lower bound for $(\star)$ by positive covariance result: $\forall s, t$, $\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$
2.2 Upper bound for $(\star)$ by Hölder's inequality

From nonlinearity ϕ assumption

$$\|h^{(\ell)}\|_k = \|\phi(g^{(\ell)})\|_k \asymp \|g^{(\ell)}\|_k \asymp k^{\ell/2}$$

Outline

Interpretation: shrinkage effect

Maximum a Posteriori (MAP) is a Regularized problem

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

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

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

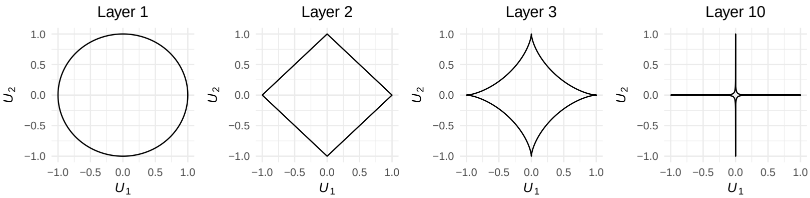
$L(\mathbf{W})$ is a loss function,
$R(\mathbf{W})$ is a norm on $\mathbb{R}^p$, regularizer.

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 ($\mathcal{L}^2$):

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



Interpretation: shrinkage effect

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).

Regularizer:

$$R(\mathbf{U}) = -\sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\mathbf{U})),$$

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

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

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 offered an interpretation from a regularization viewpoint.

Future directions:

- prove the Gaussian process limit of sub-Weibull distributions in the wide regime using Kuchibhotla and Chakrabortty [2018];
- investigate if the described regularization mechanism induces sparsity at the unit level.

References

Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julyan Arbel. Understanding priors in Bayesian neural networks at the unit level. *International Conference on Machine Learning*, 2019

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*, 2018.

David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *Proceedings of the 6th International Conference on Learning Representations.*, 2018.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Mariia Vladimirova and Julyan Arbel. Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *arXiv preprint arXiv:1905.04955*, 2019.