Overview of Bayesian Deep Learning

Julyan Arbel
✉ julyan.arbel@inria.fr www.julyanarbel.com,

Inria Grenoble Rhône-Alpes
GDR ISIS 2020

A motivating example

Outline

Introduction to Bayesian Deep Learning

Wide limit behavior of Bayesian Neural Networks

Understanding Neural Networks Priors at the Units Level

Posterior inference

Bayesian approach

The distinguishing feature of the Bayesian approach is marginalization instead of optimization.

Prior and Bayes rule are instrumental.

Bayesian model averaging (BMA) We want to obtain a *predictive distribution* for x given data $\mathcal{D}$:

$$p(x|\mathcal{D}) = \int_\Theta \underbrace{p(x|\theta)}_{\text{model}} \underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} \, \mathrm{d}\theta$$

This can also be a *conditional predictive* if we are in a regression or classification problem

$$p(y|x,\mathcal{D}) = \int_{\mathcal{W}} p(y|x,w)p(w|\mathcal{D}) \, \mathrm{d}w$$

Esp. hard with dim of $\mathcal{W}$ being of the order of 10^6.

Uncertainty

Epistemic uncertainty also known as **model uncertainty**, represents uncertainty over which base hypothesis (or parameter) is correct given the amount of available data.

Uncertainty

Epistemic uncertainty also known as **model uncertainty**, represents uncertainty over which base hypothesis (or parameter) is correct given the amount of available data.

Aleatoric uncertainty essentially, noise from the data measurements.

Uncertainty

Epistemic uncertainty also known as **model uncertainty**, represents uncertainty over which base hypothesis (or parameter) is correct given the amount of available data.

Aleatoric uncertainty essentially, noise from the data measurements.

Thus, a Bayesian approach considers epistemic uncertainties in a *principled* way, where these uncertainties are carried over to the posterior distribution on our parameter space.

Link between Bayesian learning and regularized MLE

The Maximum a Posteriori (MAP) is a penalized Maximum Likelihood Estimator

Link between Bayesian learning and regularized MLE

The Maximum a Posteriori (MAP) is a penalized Maximum Likelihood Estimator

It is akin to a Bayesian estimator under the 0-1 loss, but isn't Bayesian: it is obtained by optimization, not marginalization.

Link between Bayesian learning and regularized MLE

The Maximum a Posteriori (MAP) is a penalized Maximum Likelihood Estimator

It is akin to a Bayesian estimator under the 0-1 loss, but isn't Bayesian: it is obtained by optimization, not marginalization.

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

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

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

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

Link between Bayesian learning and regularized MLE

The Maximum a Posteriori (MAP) is a penalized Maximum Likelihood Estimator

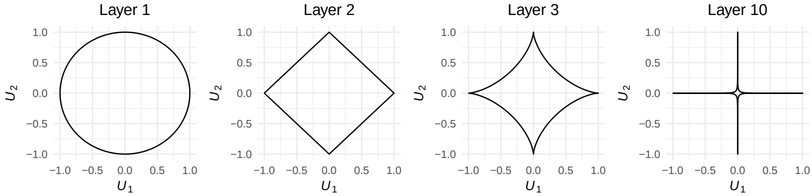
It is akin to a Bayesian estimator under the 0-1 loss, but isn't Bayesian: it is obtained by optimization, not marginalization.

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

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

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

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



Challenges of Bayesian Deep Learning

- Striving to build more interpretable parameter priors
  Vague priors such as Gaussian [Neal, 1995] over parameters are usually the default choice for deep neural networks, and they represent an acceptable description of a priori beliefs.
  Recent works have considered more elaborate priors such as spike and slab [Polson and Ročková, 2018] and horseshoe priors [Ghosh et al., 2019], and more informative parameter priors at the level of function spaces [Vladimirova et al., 2019; Sun et al., 2019; Yang et al., 2019; Louizos et al., 2019; Hafner et al., 2018].

Challenges of Bayesian Deep Learning

- Striving to build more interpretable parameter priors
  Vague priors such as Gaussian [Neal, 1995] over parameters are usually the default choice for deep neural networks, and they represent an acceptable description of a priori beliefs.
  Recent works have considered more elaborate priors such as spike and slab [Polson and Ročková, 2018] and horseshoe priors [Ghosh et al., 2019], and more informative parameter priors at the level of function spaces [Vladimirova et al., 2019; Sun et al., 2019; Yang et al., 2019; Louizos et al., 2019; Hafner et al., 2018].
- Scaling-up algorithms for Bayesian deep learning

Challenges of Bayesian Deep Learning

- Striving to build more interpretable parameter priors
  Vague priors such as Gaussian [Neal, 1995] over parameters are usually the default choice for deep neural networks, and they represent an acceptable description of a priori beliefs.
  Recent works have considered more elaborate priors such as spike and slab [Polson and Ročková, 2018] and horseshoe priors [Ghosh et al., 2019], and more informative parameter priors at the level of function spaces [Vladimirova et al., 2019; Sun et al., 2019; Yang et al., 2019; Louizos et al., 2019; Hafner et al., 2018].
- Scaling-up algorithms for Bayesian deep learning
- Gaining theoretical insight and principled uncertainty quantification for deep learning

Early works

Works by Radford Neal [Neal, 1995] and David MacKay [MacKay, 1992].

Early works

Works by Radford Neal [Neal, 1995] and David MacKay [MacKay, 1992].

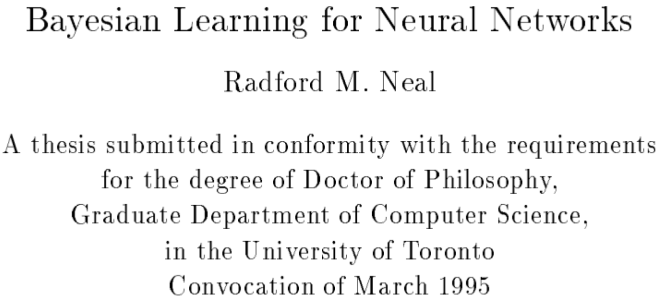
Bayesian Learning for Neural Networks

Radford M. Neal

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy,
Graduate Department of Computer Science,
in the University of Toronto
Convocation of March 1995

Early works

Works by Radford Neal [Neal, 1995] and David MacKay [MacKay, 1992].


Bayesian Learning for Neural Networks

Radford M. Neal

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy,
Graduate Department of Computer Science,
in the University of Toronto
Convocation of March 1995

They have shown in particular the infinite width Gaussian process property of 1 hidden layer neural networks.

Outline

Introduction to Bayesian Deep Learning

Wide limit behavior of Bayesian Neural Networks

Understanding Neural Networks Priors at the Units Level

Posterior inference

Wide regime: infinite number of hidden units in the layer

Theorem (Neal [1995])

Consider a Bayesian neural network with

(A1) iid Gaussian priors on the weights
(A2) with properly scaled 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.*

Wide regime: infinite number of hidden units in the layer

Theorem (Neal [1995])

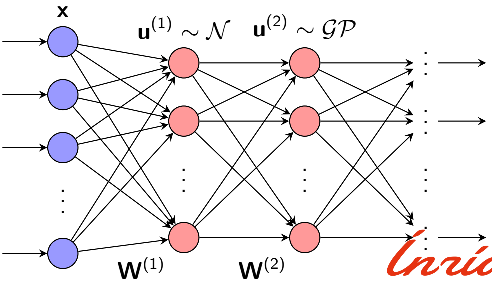
Consider a Bayesian neural network with

(A1) iid Gaussian priors on the weights
(A2) with properly scaled 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 sketch

- $\mathbf{u}^{(1)} \sim \mathcal{N}$.
- Components of $\mathbf{u}^{(1)}$ are iid $\Rightarrow$ CLT.
- $\mathbf{u}^{(2)} \sim \mathcal{GP}$ (from CLT).
- But components of $\mathbf{u}^{(2)}$ are dependent.



Wide regime: extension to deep networks

Lee et al. [2018]; Matthews et al. [2018]

DEEP NEURAL NETWORKS AS GAUSSIAN PROCESSES

Jaehoon Lee*†, Yasaman Bahri*†, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, Jascha Sohl-Dickstein

Google Brain
{jaehlee, yasamanb, romann, schsam, jpennin, jaschasd}@google.com

GAUSSIAN PROCESS BEHAVIOUR IN WIDE DEEP NEURAL NETWORKS

Alexander G. de G. Matthews
University of Cambridge
am554@cam.ac.uk

Jiri Hron
University of Cambridge
jh2084@cam.ac.uk

Mark Rowland
University of Cambridge
mr504@cam.ac.uk

Richard E. Turner
University of Cambridge
ret26@cam.ac.uk

Zoubin Ghahramani
University of Cambridge, Uber AI Labs
zoubin@eng.cam.ac.uk

Wide regime: useful for developping new theory

Schoenholz et al. [2017]; Hayou et al. [2019]

DEEP INFORMATION PROPAGATION

Samuel S. Schoenholz* Google Brain
Justin Gilmer* Google Brain
Surya Ganguli Stanford University
Jascha Sohl-Dickstein Google Brain

On the Impact of the Activation Function on Deep Neural Networks Training

Soufiane Hayou, Arnaud Doucet, Judith Rousseau *

Department of Statistics
Universiy of Oxford

Gaussian process approximation

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. (2017). Deep information propagation.

In *International Conference on Learning Representations*

- 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

Hayou, S., Doucet, A., and Rousseau, J. (2019). On the impact of the activation function on deep neural networks training.

In *International Conference on Machine Learning*

- 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

Introduction to Bayesian Deep Learning

Wide limit behavior of Bayesian Neural Networks

Understanding Neural Networks Priors at the Units Level

Posterior inference

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

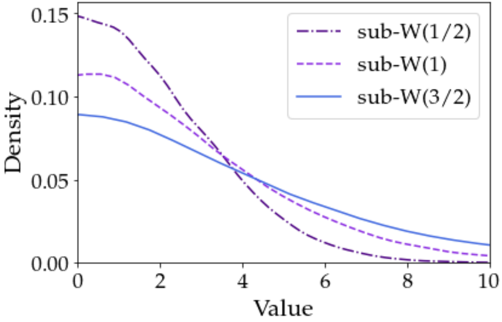
Denoted by $\mathrm{subW}(\theta)$, $\theta > 0$ called tail parameter
$\|X\|_k \asymp k^\theta \implies X \sim \mathrm{subW}(\theta)$, θ called optimal
$\mathrm{subW}(1/2) = \mathrm{subG}$, $\mathrm{subW}(1) = \mathrm{subE}$
$\theta \leq \theta' \implies \mathrm{subW}(\theta) \subset \mathrm{subW}(\theta')$

See Kuchibhotla and Chakrabortty [2018]; Vladimirova et al. [2020] for sub-Weibull

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



Denoted by subW(θ), $\theta > 0$ called tail parameter
$\|X\|_k \asymp k^\theta \implies X \sim \text{subW}(\theta)$, θ called optimal
subW($1/2$) = subG, subW(1) = subE
$\theta \leq \theta' \implies \text{subW}(\theta) \subset \text{subW}(\theta')$

See Kuchibhotla and Chakrabortty [2018]; Vladimirova et al. [2020] for sub-Weibull

Understanding priors: Outline

Sub-Weibull distributions

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

Regularization interpretation

Assumptions on neural network

To prove that Bayesian neural networks become heavier-tailed with depth, following assumptions are required

Assumptions on neural network

To prove that Bayesian neural networks become heavier-tailed with depth, following assumptions are required

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

$$w \sim \mathcal{N}(0, \sigma^2)$$

Assumptions on neural network

To prove that Bayesian neural networks become heavier-tailed with depth, following assumptions are required

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

$$w \sim \mathcal{N}(0, \sigma^2)$$

(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

Theorem (Vladimirova et al, 2019)

Main theorem

Theorem (Vladimirova et al, 2019)

Consider a Bayesian neural network with

(A1) iid Gaussian priors on the weights and
(A2) nonlinearity satisfying envelope property.

Main theorem

Theorem (Vladimirova et al, 2019)

Consider a Bayesian neural network with

(A1) iid 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* $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim \mathit{subW}(\ell/2)$

Main theorem

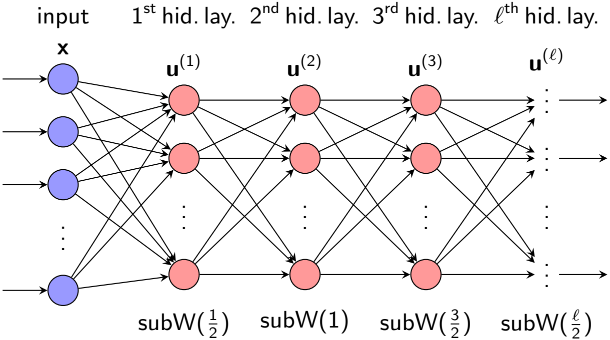
Theorem (Vladimirova et al, 2019)

Consider a Bayesian neural network with

(A1) iid 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* $\theta = \ell/2$: $\pi^{(\ell)}(u) \sim subW(\ell/2)$



Main theorem

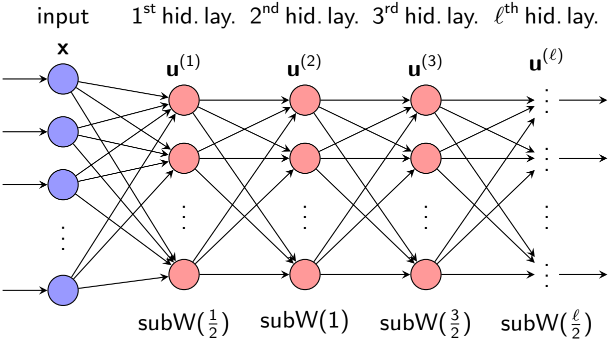
Theorem (Vladimirova et al, 2019)

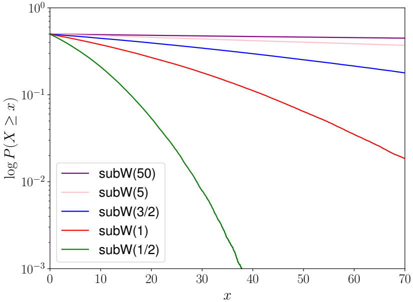
Consider a Bayesian neural network with

(A1) iid 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. $X \sim \mathrm{subW}(\theta) \iff \exists C > 0, \|X\|_k = \left(\mathbb{E}|X|^k\right)^{1/k} \leq Ck^{\theta}, \quad \text{for all } k \in \mathbb{N}.$

Proof sketch I

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

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

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

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

Proof sketch I

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

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

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

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

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

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

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

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

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

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

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: $\forall s, t$, $\text{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0$

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

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

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

Understanding priors: Outline

Sub-Weibull distributions

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

Regularization interpretation

Interpretation: shrinkage effect

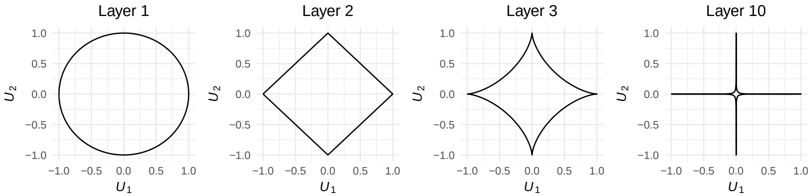
Maximum a Posteriori (MAP) is a Regularized problem

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

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

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

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



MAP on weights **W** is weight decay

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 $\mathbf{U}$: regularization scheme

Marginal distributions:

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

MAP on units $\mathbf{U}$: regularization scheme

Marginal distributions:

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)

MAP on units $\mathbf{U}$: regularization scheme

Marginal distributions:

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)

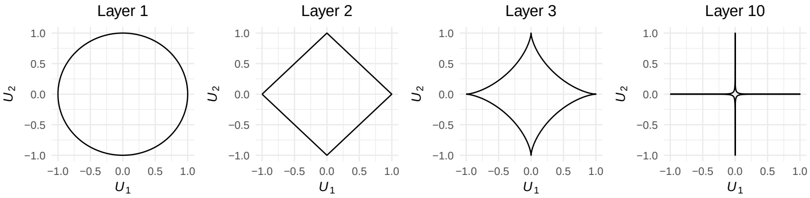
$$\begin{aligned} R(\mathbf{U}) &= -\sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} \log \pi_m^{(\ell)}(U_m^{(\ell)}) - \log C(F(\mathbf{U})), \\ &\approx \sum_{\ell=1}^{L} \sum_{m=1}^{H_\ell} |U_m^{(\ell)}|^{2/\ell} - \log C(F(\mathbf{U})), \\ &\approx \|\mathbf{U}^{(1)}\|_2^2 + \|\mathbf{U}_1^{(2)}\|_1 + \cdots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

MAP on units U: regularization scheme

Regularizer:

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

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.

Main references:

- Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding Priors in Bayesian Neural Networks at the Unit Level.
  ICML https://arxiv.org/abs/1810.05193

- Vladimirova, M., Girard, S., Nguyen, H. D., and Arbel, J. (2020). Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions.
  Submitted https://arxiv.org/abs/1905.04955

Outline

Introduction to Bayesian Deep Learning

Wide limit behavior of Bayesian Neural Networks

Understanding Neural Networks Priors at the Units Level

Posterior inference

Scaling-up algorithms for Bayesian deep learning

How to deal with the dealing with the huge dimensionality of Bayesian model averaging? There are a variety of [scalable] approximate inference techniques available:

- Hamiltonian Monte Carlo (not scalable) [Neal, 1995]
- mean-field variational inference [Hinton and Van Camp, 1993; Blundell et al., 2015]
- Monte Carlo dropout [Gal and Ghahramani, 2016]
- exploring the link between deep networks and Gaussian processes [Lee et al., 2018; Matthews et al., 2018; Khan et al., 2019],
- iterative learning from small mini-batches [Welling and Teh, 2011],
- using weight-perturbation approaches [Khan et al., 2018],
- investigating the information contained in the stochastic gradient descent trajectory [Maddox et al., 2019],
- exploiting properties of the loss landscape [Garipov et al., 2018], by focusing on subspaces of low dimensionality that capture a large amount of the variability of the posterior distribution [Izmailov et al., 2019],
- applying non-linear transformations for dimensionality reduction [Pradier et al., 2018]

Advertising: two-year postdoc joint at Oxford and Grenoble

- On the themes of the presentation: challenges of **Bayesian Deep Learnings**
- With Judith Rousseau and myself
- Funded by Judith's ERC "General Theory for Big Bayes" and Grenoble's IDEX
- Starting date between now and end of 2020
- Write to us if interested: julyan.arbel@inria.fr judith.rousseau@stats.ox.ac.uk

References

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *International Conference on Machine Learning*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). Model selection in bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research*, 20(182):1–46.

Hafner, D., Tran, D., Lillicrap, T., Irpan, A., and Davidson, J. (2018). Reliable uncertainty estimates in deep neural networks using noise contrastive priors.

Hayou, S., Doucet, A., and Rousseau, J. (2019). On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning*.

Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13.

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. (2019). Subspace inference for bayesian deep learning. *arXiv preprint arXiv:1907.07504*.

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and scalable bayesian deep learning by weight-perturbation in adam. *arXiv preprint arXiv:1806.04854*.

Khan, M. E. E., Immer, A., Abedi, E., and Korzepa, M. (2019). Approximate inference turns deep networks into gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3088–3098.

Kuchibhotla, A. K. and Chakrabortty, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International Conference on Machine Learning*.

Louizos, C., Shi, X., Schutte, K., and Welling, M. (2019). The functional neural process. In *Advances in Neural Information Processing Systems*, pages 8743–8754.

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

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13132–13143.

Matthews, A., Rowland, M., Hron, J., Turner, R., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, volume 1804.11271.

Neal, R. (1995). *Bayesian learning for neural networks*. Springer.

Polson, N. G. and Ročková, V. (2018). Posterior concentration for sparse deep learning. In *Advances in Neural Information Processing Systems*, pages 930–941.

Pradier, M. F., Pan, W., Yao, J., Ghosh, S., and Doshi-Velez, F. (2018). Latent projection bnns: Avoiding weight-space pathologies by learning latent representations of neural network weights. *arXiv preprint arXiv:1811.07006*.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. (2017). Deep information propagation. In *International Conference on Learning Representations*.

Sun, S., Zhang, G., Shi, J., and Grosse, R. (2019). Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*.

Vladimirova, M., Girard, S., Nguyen, H. D., and Arbel, J. (2020). Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Submitted*.

Vladimirova, M., Verbeek, J., Mesejo, P., and Arbel, J. (2019). Understanding Priors in Bayesian Neural Networks at the Unit Level. *ICML*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Yang, W., Lorch, L., Graule, M. A., Srinivasan, S., Suresh, A., Yao, J., Pradier, M. F., and Doshi-Velez, F. (2019). Output-constrained bayesian neural networks. *arXiv preprint arXiv:1905.06287*.