BML Lecture #7

Bayesian deep learning

Julyan Arbel

Bayesian deep learning

Bayesian neural networks

Introduction

- Bayesian model averaging, Maximum a posteriori = Regularized maximum likehood

Distribution properties

- Wide limits, Gaussian process, sub-Weibull units

Approximations

- Laplace approximation, Variational inference, Monte Carlo dropout

Some recent works

Implementation: Pyro & PyTorch

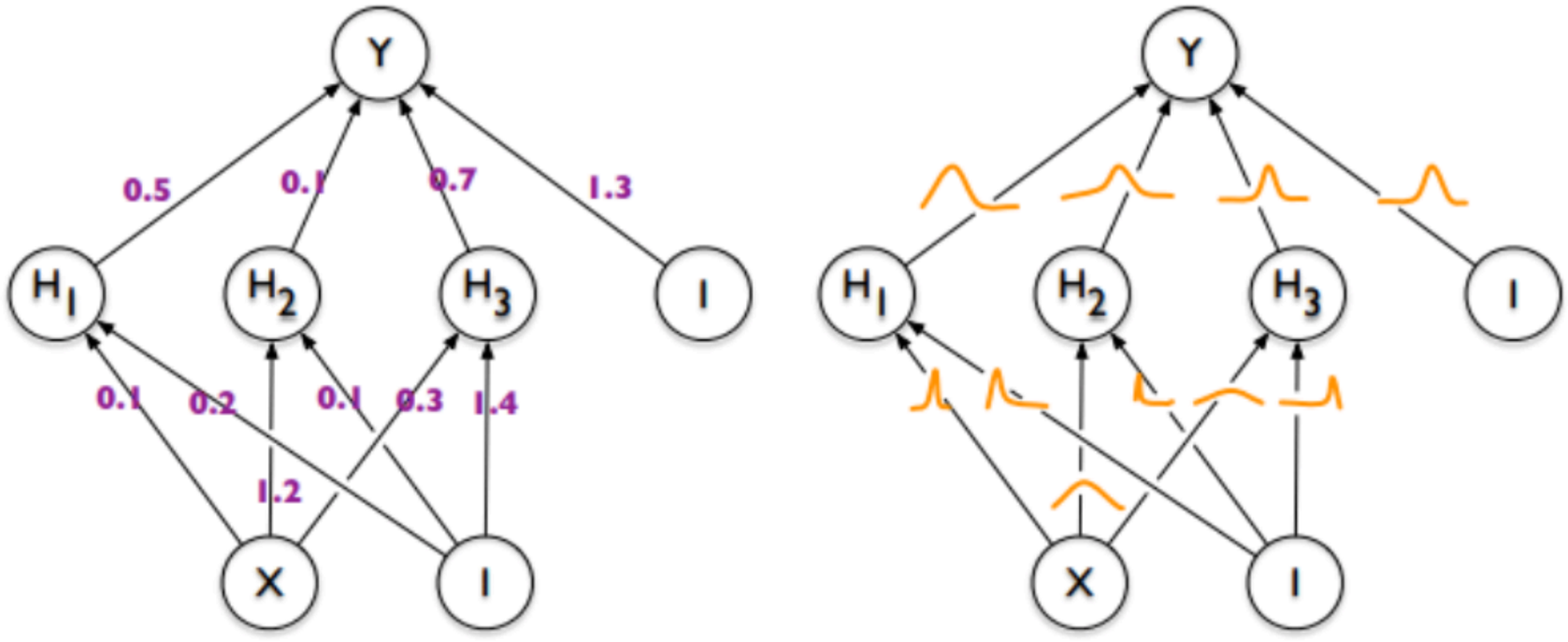
Bayesian deep learning

Bayesian neural networks: objectives

- Learn why Bayesian Neural networks are useful and exciting
- Understand how they're different from normal neural networks
- Appreciate how the uncertainty metrics you can obtain are a major advantage
- Install a library to start work with all things random

Introduction

- Bayesian model averaging
- Maximum a posteriori = Regularized maximum likehood



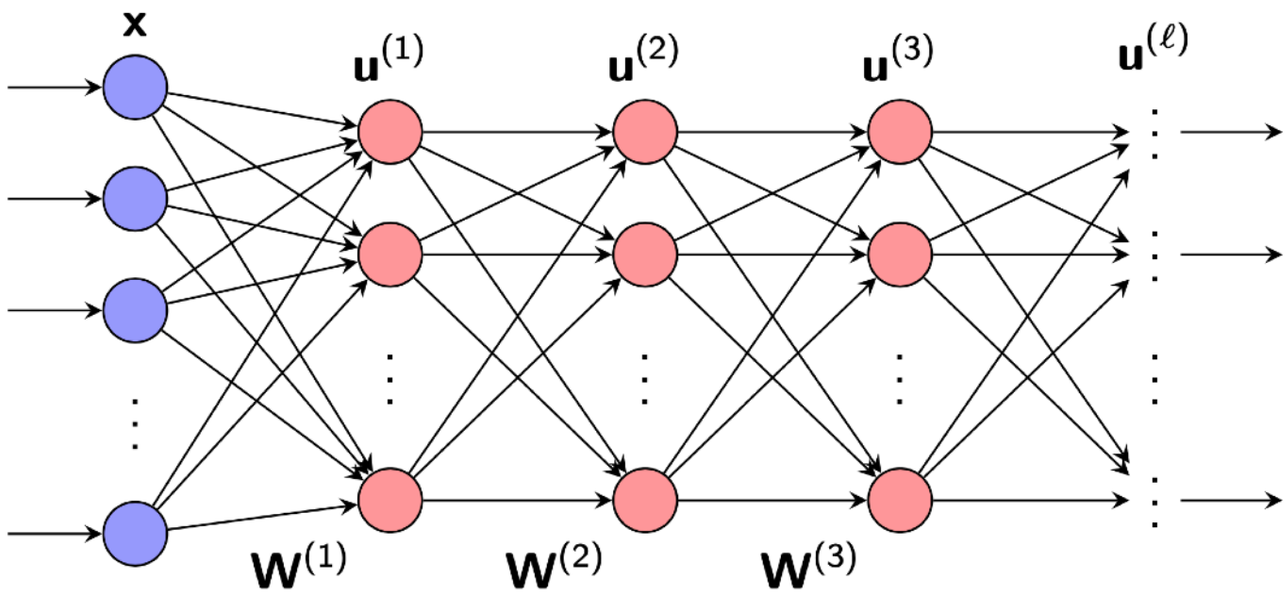
Standard DNN

Bayesian DNN

"Weight Uncertainty in Neural Networks" by Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra

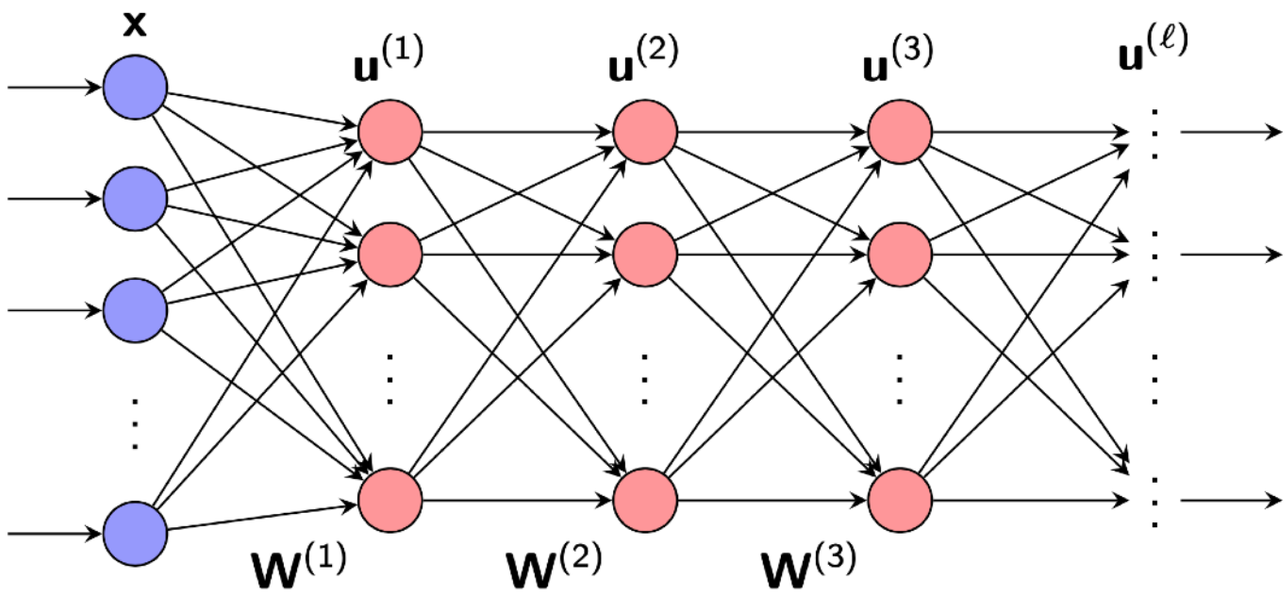
BAYESIAN NEURAL NETWORKS

DEEP NEURAL NETWORK



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

DEEP NEURAL NETWORK



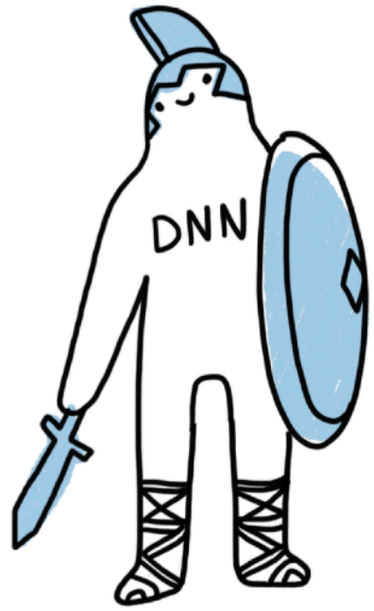
$$L(f(\mathbf{x}), \mathbf{y})$$

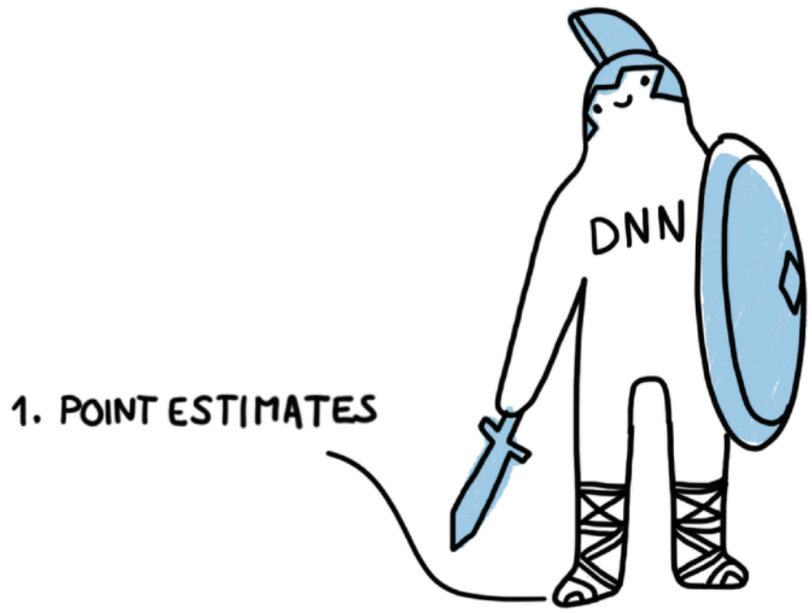
Forward pass:

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

Back-propagation of the gradients:

$$\boldsymbol{\delta}^{(\ell)} = \frac{\partial L}{\partial \mathbf{h}^{(\ell)}} = \phi'(\mathbf{g}^{(\ell)})\boldsymbol{\delta}^{\ell+1}\mathbf{W}^{\ell+1},$$
$$\frac{\partial L}{\partial \mathbf{W}^{(\ell)}} = \boldsymbol{\delta}^{(\ell)}\phi(\mathbf{g}^{(\ell-1)})\,.$$


DNN


DNN
1. POINT ESTIMATES

DNN

BOXER

PUG

DALMATIAN

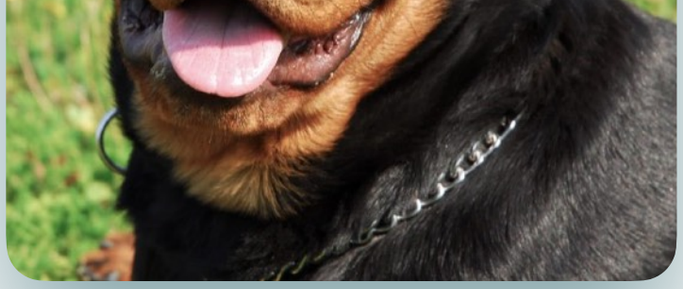
BULLDOG

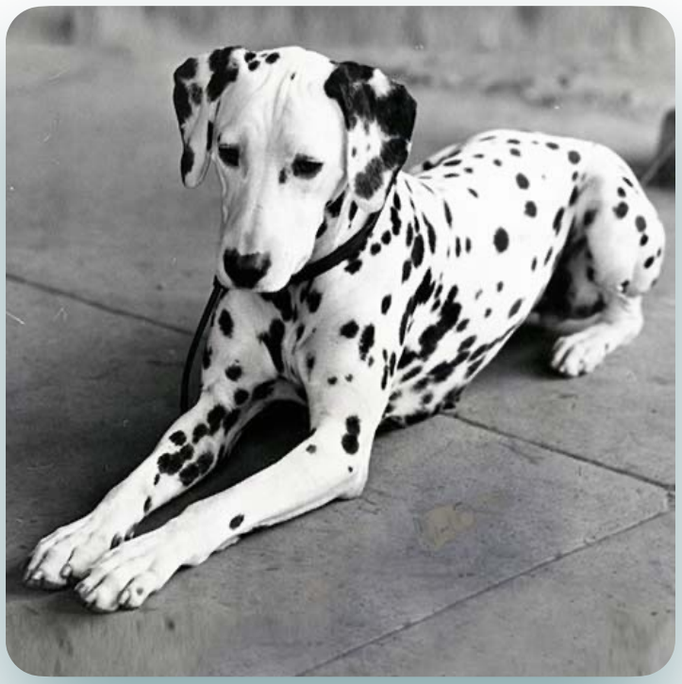
ROTTWEILER

LABRADOR

CHIHUAHUA

informatics mathematics





DNN

BOXER

PUG

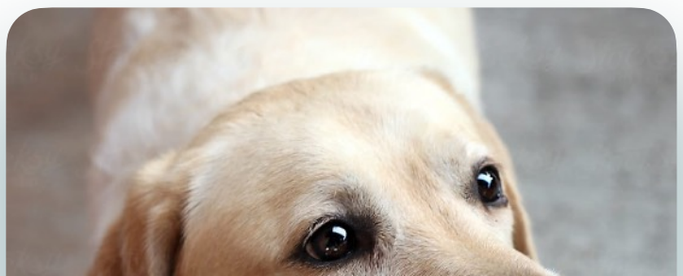
DALMATIAN

BULLDOG

ROTTWEILER

LABRADOR

CHIHUAHUA



informatics mathematics
Inria

DNN

BOXER

PUG

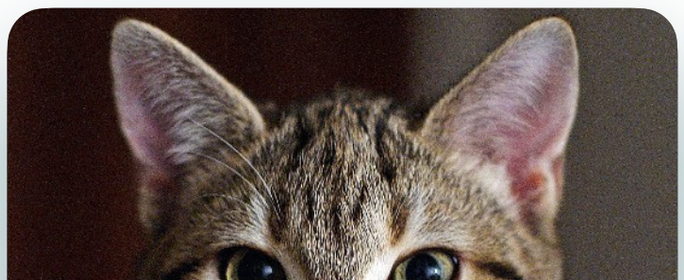
DALMATIAN

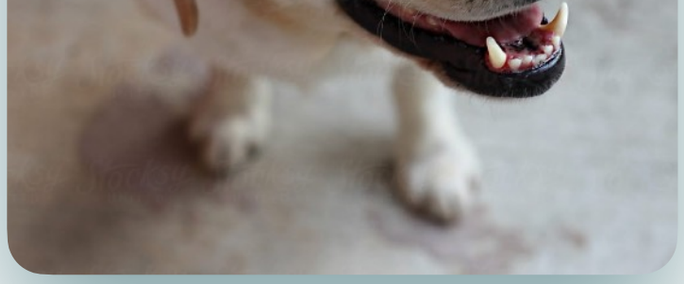
BULLDOG

ROTTWEILER

LABRADOR

CHIHUAHUA





DNN

BOXER

PUG

DALMATIAN

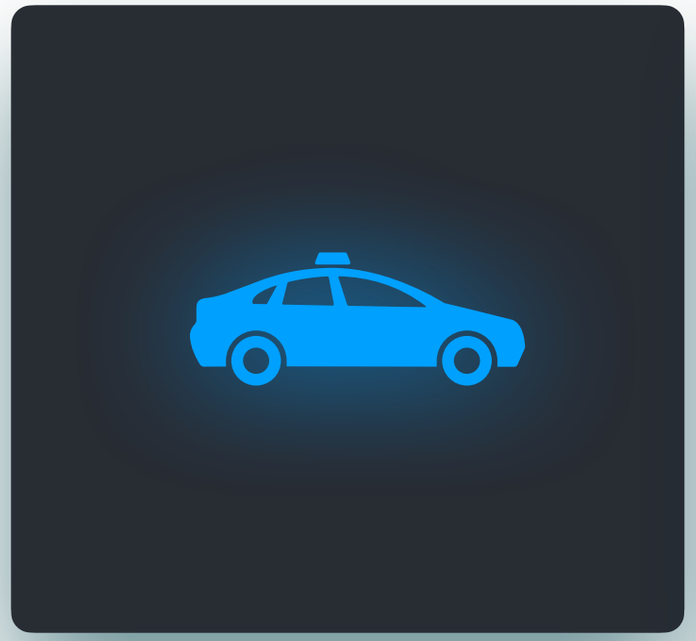
BULLDOG

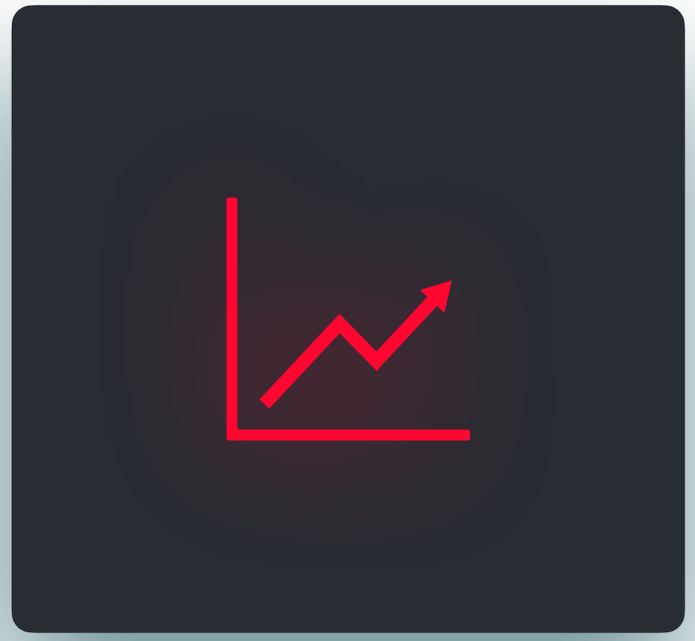
ROTTWEILER

LABRADOR

CHIHUAHUA

informatics mathematics

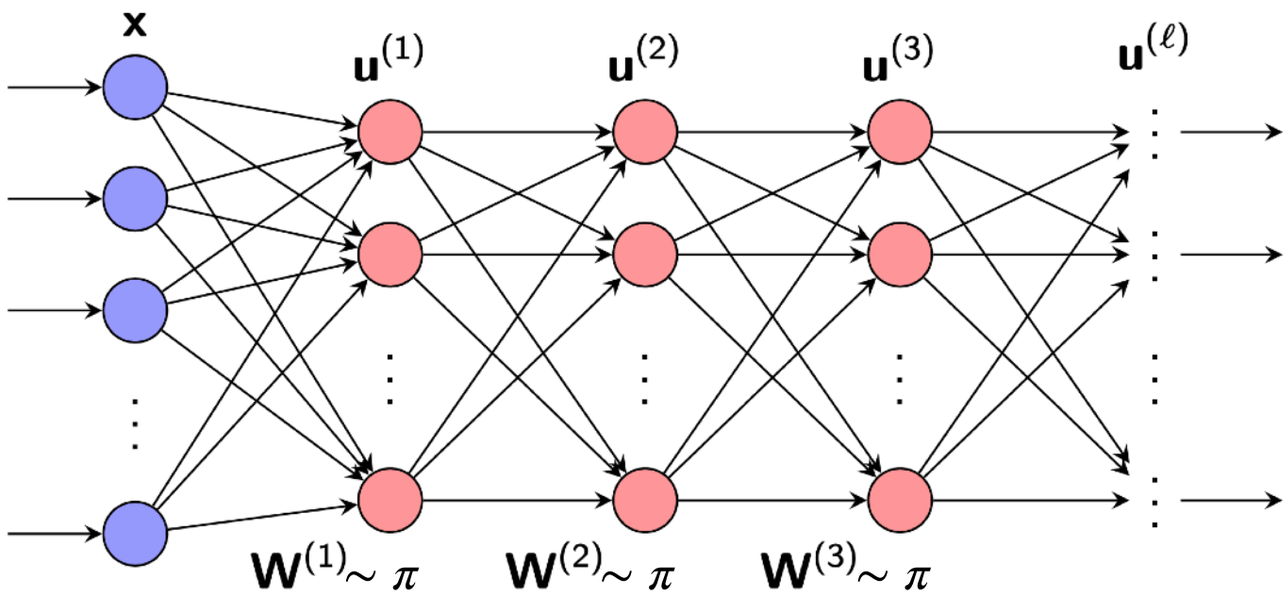




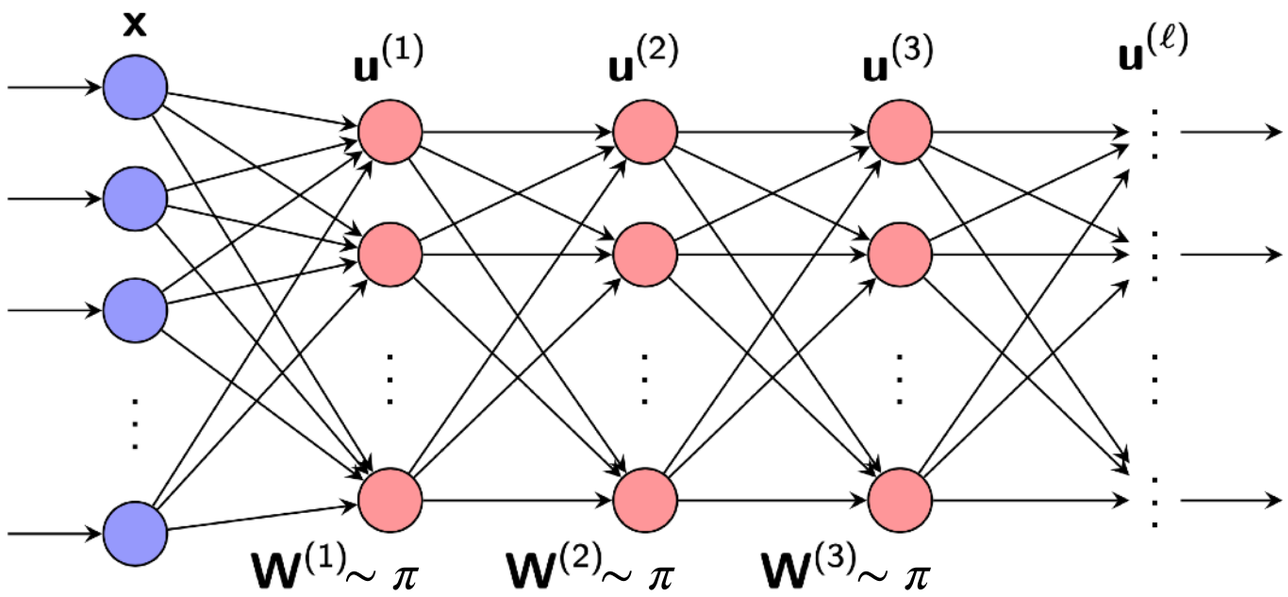
informatics mathematics

Solution: put a prior on weights, $w^{(i)} \sim \pi$

Solution: put a prior on weights, $w^{(i)} \sim \pi$ ⟶ Bayesian Neural Networks



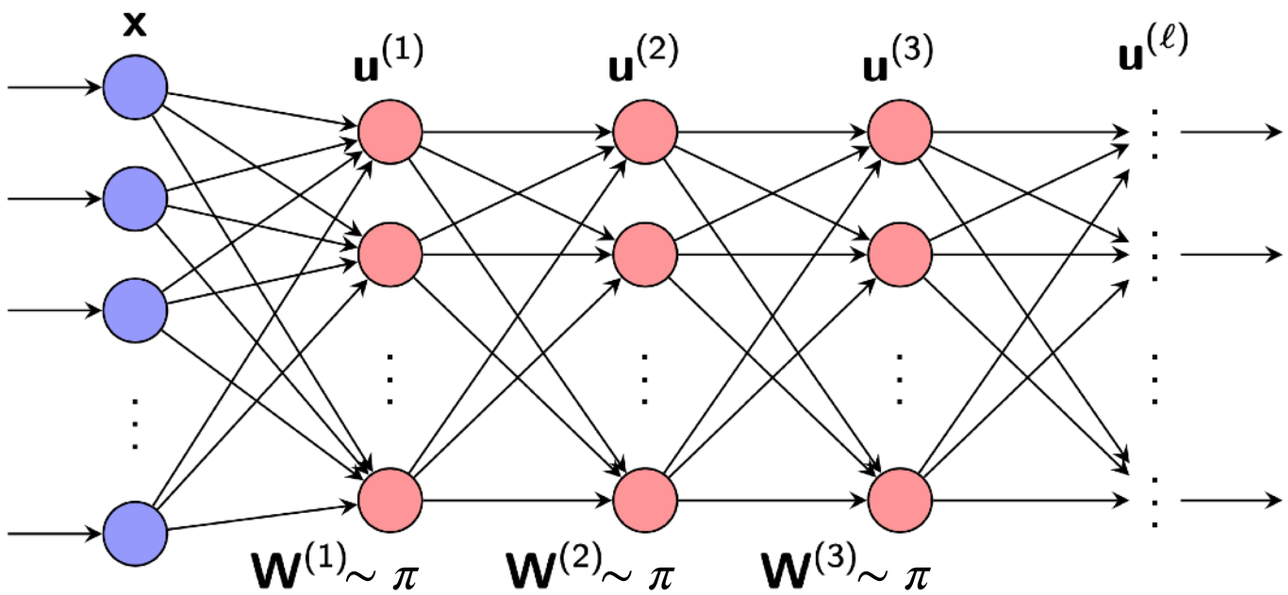
Solution: put a prior on weights, $\mathrm{w}^{(i)} \sim \pi$ ⟶ Bayesian Neural Networks



Posterior predictive

$$\pi(\mathbf{y} \mid \mathbf{x}, \mathscr{D})$$

Solution: put a prior on weights, $\mathrm{w}^{(i)} \sim \pi$ ⟶ Bayesian Neural Networks



Posterior predictive

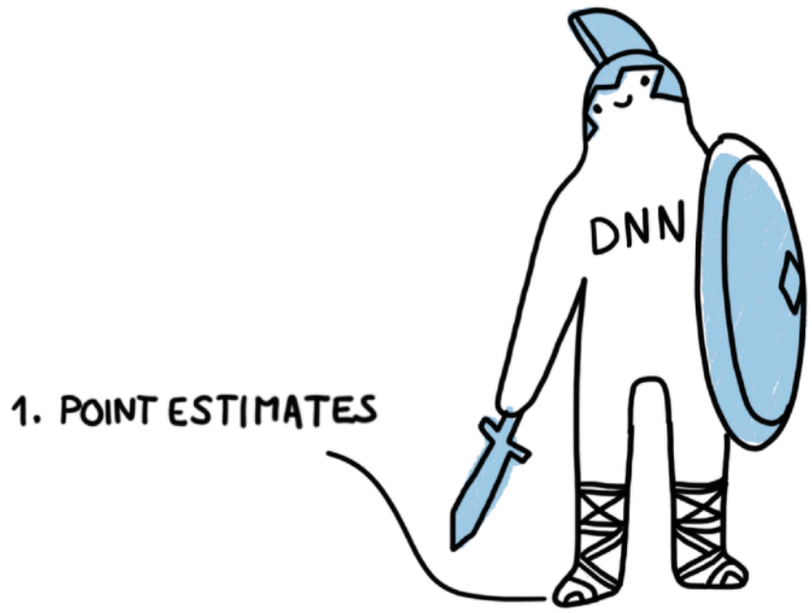
$$\pi(\mathbf{y} \mid \mathbf{x}, \mathscr{D})$$

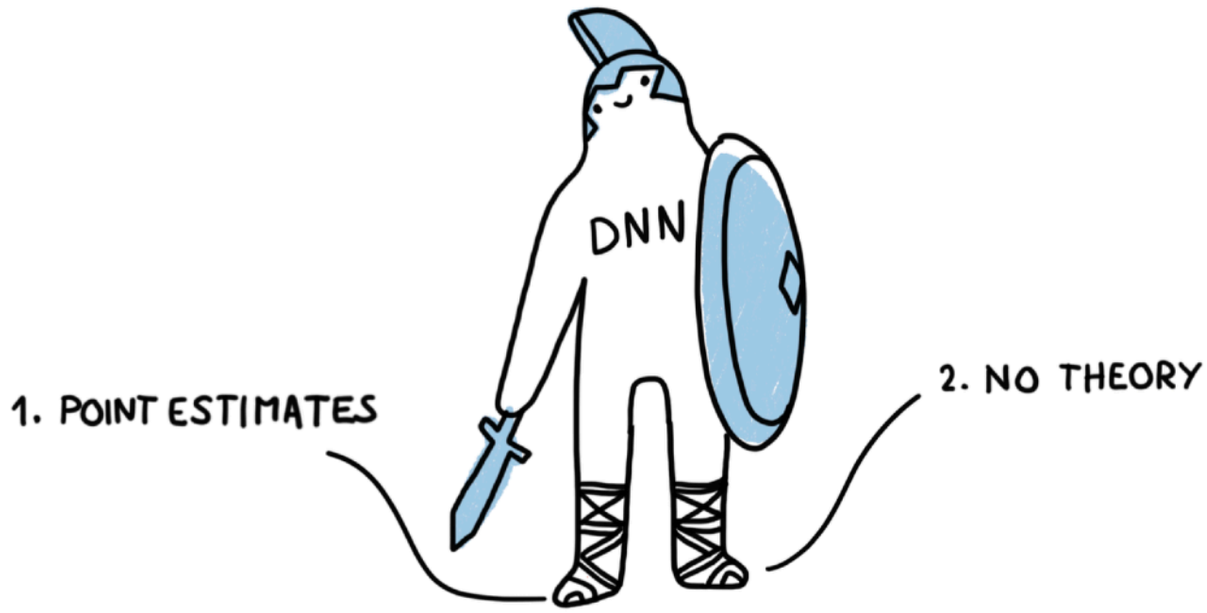
Advantages:

- Allows to model uncertainty
- Represents a standard neural network

Problems:

- Intractable for large datasets
- Have to choose a prior


DNN
1. POINT ESTIMATES


DNN
1. POINT ESTIMATES
2. NO THEORY

Distribution properties

- Wide limits
- Gaussian process limit
- Stable process limit
- Sub-Weibull units

WHAT DO WE WANT?

- To propagate information: avoid activations explosion/ collapse
- To back-propagate information: avoid gradients explosion/collapse

WHAT DO WE WANT?

- To propagate information: avoid activations explosion/ collapse
- To back-propagate information: avoid gradients explosion/collapse

We can try to tackle those problems at initialization!

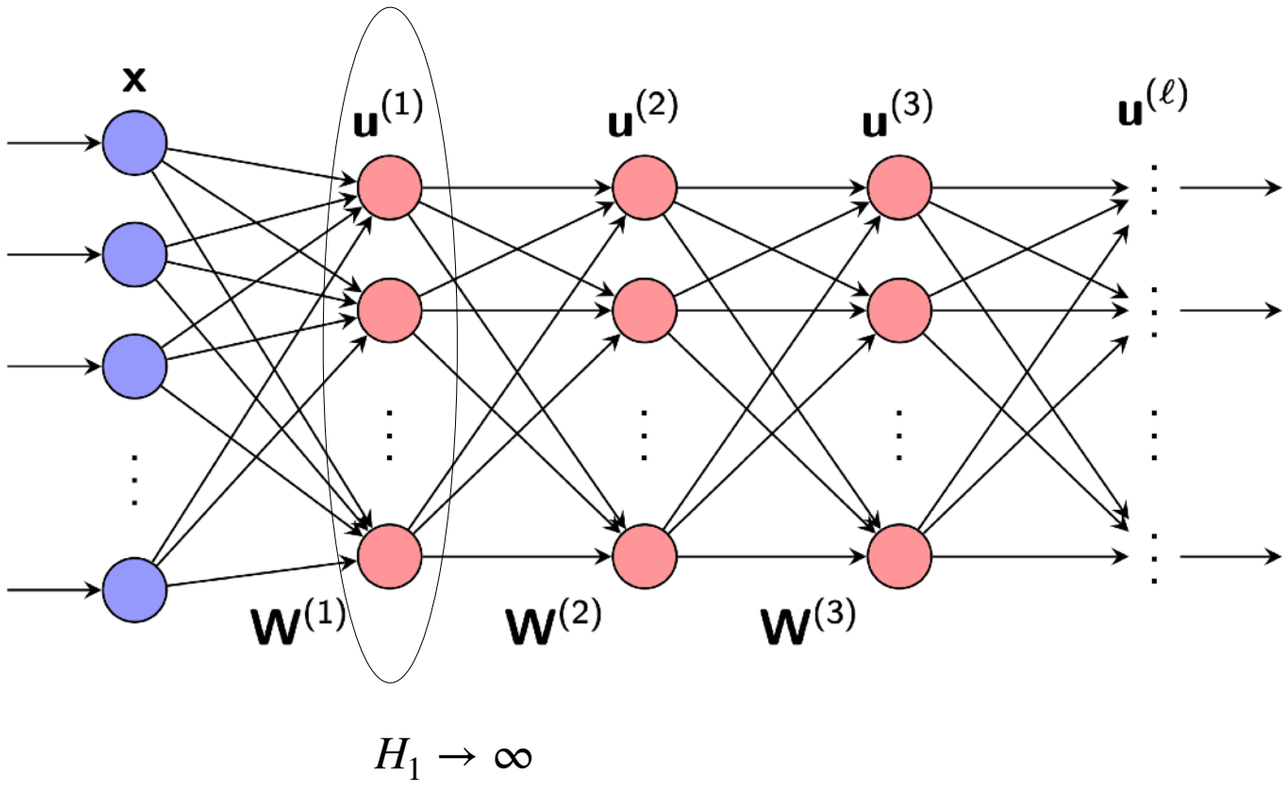
What are the good initializations techniques for $\mathbf{W}^{\ell}$?

Let’s study the behaviour of priors in BNNs and find a good prior distribution!

Get prior -> get initialization!

BNN AND GAUSSIAN PROCESS

WIDE REGIME = NUMBER OF HIDDEN UNITS TENDS TO INFINITY



Marginal prior distribution of $\mathbf{u}^{(1)}$ converges to a Gaussian process (CLT)

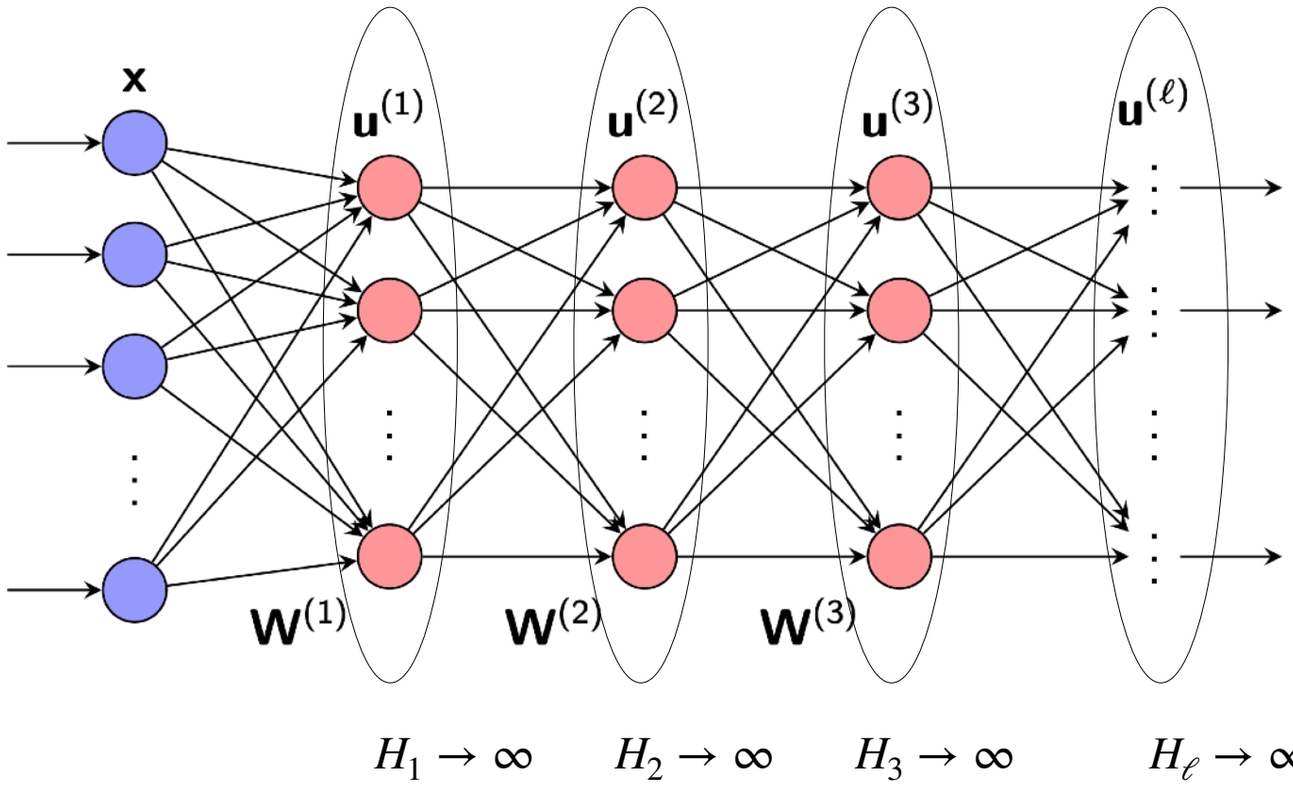
$\Rightarrow$ Distributions $u_i^{(2)}$ are dependent

Similar convergence for $\mathbf{u}^{(2)}$?

[R. Neal, PhD dissertation (1996)]

Gaussian distribution on weights: $\mathbf{W}^{(i)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{H_i}\right)$

WIDE REGIME = NUMBER OF HIDDEN UNITS TENDS TO INFINITY



$\Rightarrow$

Marginal prior distribution of $\mathbf{u}^{(1)}$ converges to a Gaussian process

[R. Neal, PhD dissertation (1996)]

Marginal prior distribution of $\mathbf{u}^{(\ell)}$ converges to a Gaussian process

Matthews et al. (ICLR, 2018) Gaussian process behaviour in wide deep neural networks

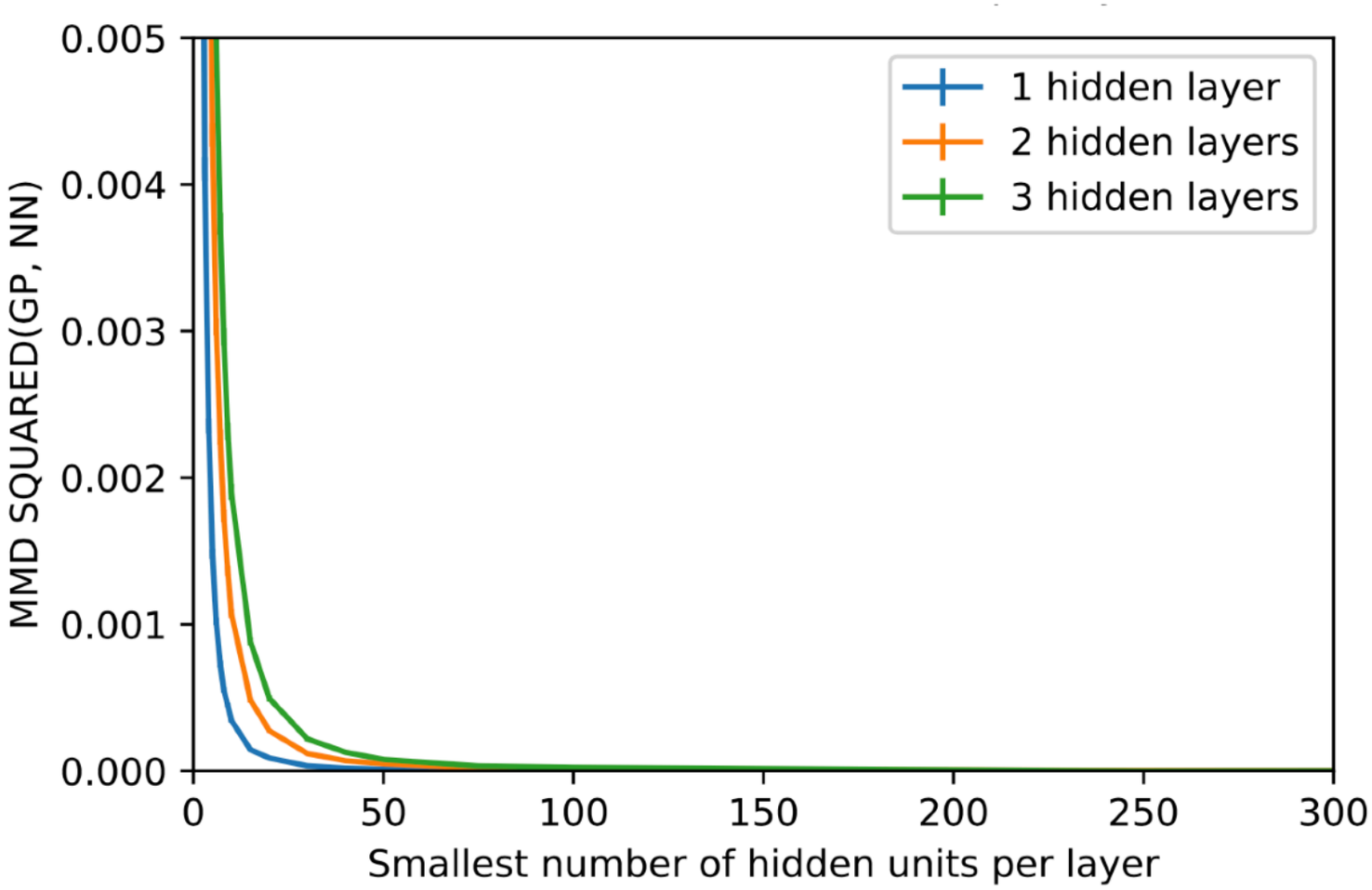
Lee et al. (ICLR, 2018) Deep neural networks as gaussian processes

Gaussian distribution on weights: $\mathbf{W}^{(i)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{H_i}\right)$

BUT

KEEP IN MIND THE REALITY :

DEEPER IN NNs THE DISTANCE FROM GP IS INCREASING



[Matthews et al. (2018) Gaussian process behaviour in wide deep neural networks, ICLR]

Other choices of priors for BNNs

[Favaro et al, 2021. Stable behaviour of infinitely wide deep neural networks. arxiv]

- Considers stable priors St(α,σ) with stability parameter α and scale parameter σ. Very heavy tails, only moments of order less than α exist.
- Convergence in distribution of the infinitely wide limit to a stable process.
- Some open questions:
  - application to Bayesian inference
  - existence of a neural tangent kernel (NTK) in the stable regime
  - application to information propagation

INITIALIZATION TECHNIQUES

VARIANCE PRESERVATION

Additional assumptions:

- Inputs are random and i.i.d.
- Gradients are i.i.d. and independent from inputs

Ideas:

- Preserve the variance of the pre-activations during propagation
- Preserve the variance of the gradients during back-propagation

VARIANCE PRESERVATION

Additional assumptions:

- Inputs are random and i.i.d.
- Gradients are i.i.d. and independent from inputs

Ideas:

- Preserve the variance of the pre-activations during propagation
- Preserve the variance of the gradients during back-propagation

$$\text{Var}(h^{(\ell)}) = \text{Var}(h^{(\ell+1)})$$

$$\text{Var}\left(\frac{\partial L}{\partial h^{(\ell+1)}}\right) = \text{Var}\left(\frac{\partial L}{\partial h^{(\ell)}}\right) \quad \Rightarrow \quad \begin{aligned} &\text{Var}(\phi(g^{(\ell)})) \approx H_{\ell-1}\text{Var}(\mathbf{W}^{(\ell)}) = 1 \\ &\text{Var}(\phi'(g^{(\ell)})\mathbf{W}^{(\ell)}) \approx H_{\ell}\text{Var}(W^{(\ell)}) = 1 \end{aligned}$$

GLOROT AND HE INITIALIZATIONS

Glorot's initialization

Arithmetic compromise between forward and backward variance preservations

$$\mathbf{W}^{\ell} \sim \mathcal{U}\left(-\frac{\sqrt{6}}{\sqrt{H_{\ell}+H_{\ell-1}}}, \frac{\sqrt{6}}{\sqrt{H_{\ell}+H_{\ell-1}}}\right)$$

$$\text{Var}(\mathbf{W}^{\ell}) = \frac{2}{H_{\ell}+H_{\ell-1}}$$

Glorot and Bengio (2010). Understanding the difficulty of training deep feedforward neural networks

He's initialization

- Preserves the variance in the forward pass ($\sqrt{2}$ for ReLU)
- No variance collapse/explosion in the backward pass

$$\mathbf{W}^{\ell} \sim \mathcal{N}\left(0, \left(\frac{\sqrt{2}}{\sqrt{H_{\ell-1}}}\right)^2\right)$$

He et al. (2015). Delving deep into rectifiers: Surpassing human-level performance on Imagenet classification.

DEEP INFORMATION PROPAGATION

- Propagation of deterministic inputs
- Correlation between two data points across a neural network

Ideas:

- Study the variance of a pre-activation for a given data point a

$$q_{aa}^{\ell} = \mathbb{E}\left[\left(g^{\ell}(a)\right)^2\right]$$

- Study the covariance between pre-activations for given two different data-points a and b

$$q_{ab}^{\ell} = \mathbb{E}\left[g^{\ell}(a), g^{\ell}(b)\right]$$

DEEP INFORMATION PROPAGATION

Assumptions:

- Pre-activations are Gaussian (= close to wide regime)
- Gaussian initialization: $\mathbf{W}^{(\ell)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{H_\ell}\right), \quad \mathbf{b}^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2)$

Reccurence relations of $q_{aa}^{\ell} = \mathbb{E}\left[\left(g^{\ell}(a)\right)^2\right]$ and $q_{ab}^{\ell} = \mathbb{E}\left[g^{\ell}(a), g^{\ell}(b)\right]$:

$$q_{aa}^{\ell} = \sigma_w^2 \int \phi^2\left(\sqrt{q_{aa}^{\ell-1}}\, g\right) \mathfrak{D}g + \sigma_b$$

$\mathfrak{D}g \sim \mathcal{N}(0,1)$ (Gaussian assumption)

Discribes the evolution of a single input a through a neural network!

Poole et al. (2016). Exponential expressivity in deep neural networks through transient chaos, NIPS
Schoenholz et al. (2017). Deep information propagation, ICLR

DEEP INFORMATION PROPAGATION

Assumptions:

- Pre-activations are Gaussian (= close to wide regime)
- Gaussian initialization: $\mathbf{W}^{(\ell)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{H_\ell}\right), \quad \mathbf{b}^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2)$

Reccurence relations of $q_{aa}^{\ell} = \mathbb{E}\left[\left(g^{\ell}(a)\right)^2\right]$ and $q_{ab}^{\ell} = \mathbb{E}\left[g^{\ell}(a), g^{\ell}(b)\right]$:

$$q_{aa}^{\ell} = \sigma_w^2 \int \phi^2\left(\sqrt{q_{aa}^{\ell-1}}\, g\right) \mathfrak{D}g + \sigma_b$$

$\mathfrak{D}g \sim \mathcal{N}(0,1)$ (Gaussian assumption)

Describes the evolution of a single input a through a neural network!

For any choice of σ_w and σ_b with bounded ϕ, there is a fixed point $q^* = \lim_{\ell \to \infty} q_{aa}^{\ell}$

Poole et al. (2016). Exponential expressivity in deep neural networks through transient chaos, NIPS
Schoenholz et al. (2017). Deep information propagation, ICLR

DEEP INFORMATION PROPAGATION

Recurrence relations of $q^{\ell}_{aa} = \mathbb{E}\left[\left(g^{\ell}(a)\right)^2\right]$ and $q^{\ell}_{ab} = \mathbb{E}\left[g^{\ell}(a), g^{\ell}(b)\right]$:

$$q^{\ell}_{aa} = \sigma_w^2 \int \phi^2\left(\sqrt{q^{\ell-1}_{aa}}\, g\right) \mathfrak{D}g + \sigma_b$$

$$q^{\ell}_{ab} = \sigma_w^2 \int \phi(z_1)\phi(z_2)\mathfrak{D}g_1 \mathfrak{D}g_2 + \sigma_b$$

Propagation of two signals through a neural network!

$\mathfrak{D}g \sim \mathcal{N}(0,1)$ Gaussian

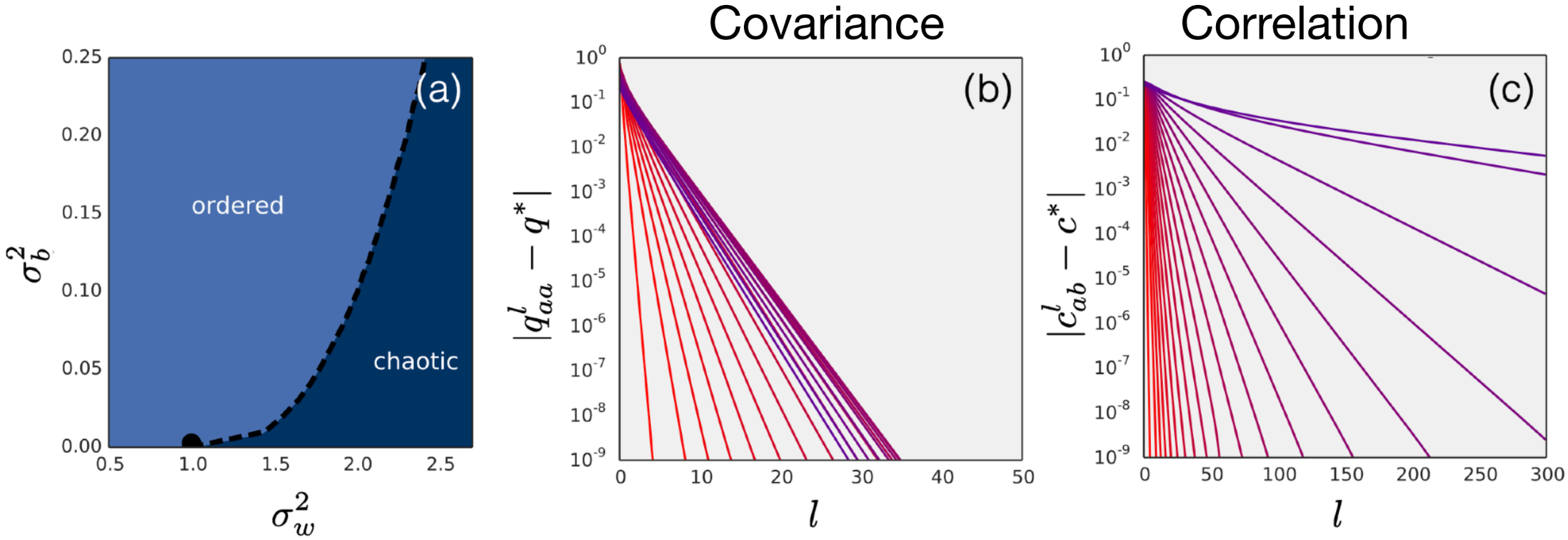
$c^{\ell}_{ab} = \dfrac{q^{\ell}_{ab}}{\sqrt{q^{\ell}_{aa} q^{\ell}_{bb}}}$ distributions correlation

$z_1 = \sqrt{q^{\ell-1}_{aa}}\, g \quad z_2 = \sqrt{q^{\ell-1}_{bb}}\, g \left(c^{\ell}_{ab} g_1 + \sqrt{1 - (c^{\ell}_{ab})^2}\, g_2\right)$

Gaussian approximations to the pre-activations

Poole et al. (2016). Exponential expressivity in deep neural networks through transient chaos, NIPS
Schoenholz et al. (2017). Deep information propagation, ICLR

EDGE OF CHAOS INITIALIZATION



There is a critical line separating an ordered phase (in which c* = 1 and all inputs end up asymptotically correlated) and a chaotic phase (in which c* < 1 and all inputs end up asymptotically decorrelated).

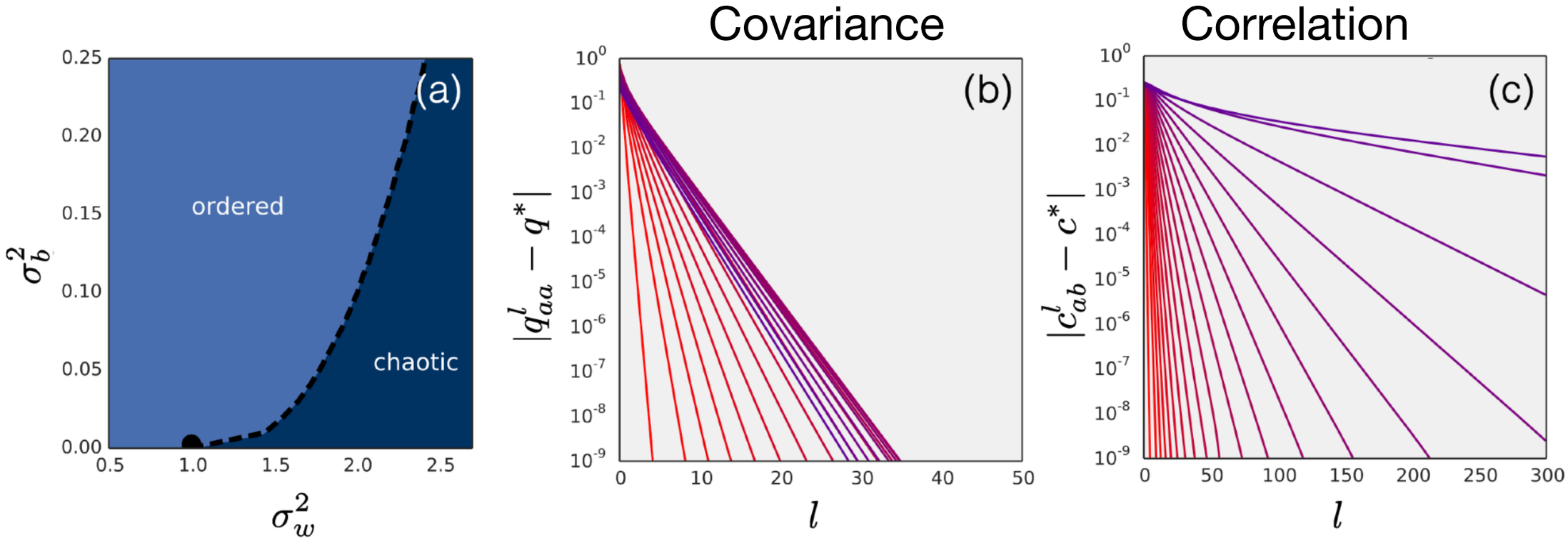
$$q_{aa}^{\ell} = \sigma_w^2 \int \phi^2\left(\sqrt{q_{aa}^{\ell-1}}\, g\right) \mathfrak{D}g + \sigma_b$$

$$q_{ab}^{\ell} = \sigma_w^2 \int \phi(z_1)\phi(z_2) \mathfrak{D}g_1 \mathfrak{D}g_2 + \sigma_b$$

$$c_{ab}^{\ell} = \frac{q_{ab}^{\ell}}{\sqrt{q_{aa}^{\ell} q_{bb}^{\ell}}}$$

Poole et al. (2016). Exponential expressivity in deep neural networks through transient chaos, NIPS

EDGE OF CHAOS INITIALIZATION



$\mathbf{W}^{(\ell)} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{H_\ell}\right)$, $\mathbf{b}^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2)$, where σ_w and σ_b are chosen from the separating line (Edge of Chaos) Deeper propagates the information!

Schoenholz et al. (2017). Deep information propagation, ICLR

BUT

KEEP IN MIND THE REALITY :

$$q^{\ell}_{aa} = \sigma_w^2 \int \phi^2 \left(\sqrt{q^{\ell-1}_{aa}} g \right) \mathfrak{D}g + \sigma_b$$

$$q^{\ell}_{ab} = \sigma_w^2 \int \phi(z_1)\phi(z_2) \mathfrak{D}g_1 \mathfrak{D}g_2 + \sigma_b$$

Gaussian approximations to the pre-activations

=

Wide Regime!

PRIORS AT THE UNIT LEVEL

SUB-WEIBULL RANDOM VARIABLE

A random variable X, such that

$$\mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K\right)$$

for all $x \geq 0$ and for some $K > 0$, is called a **sub-Weibull** random variable with tail parameter $\theta > 0$:

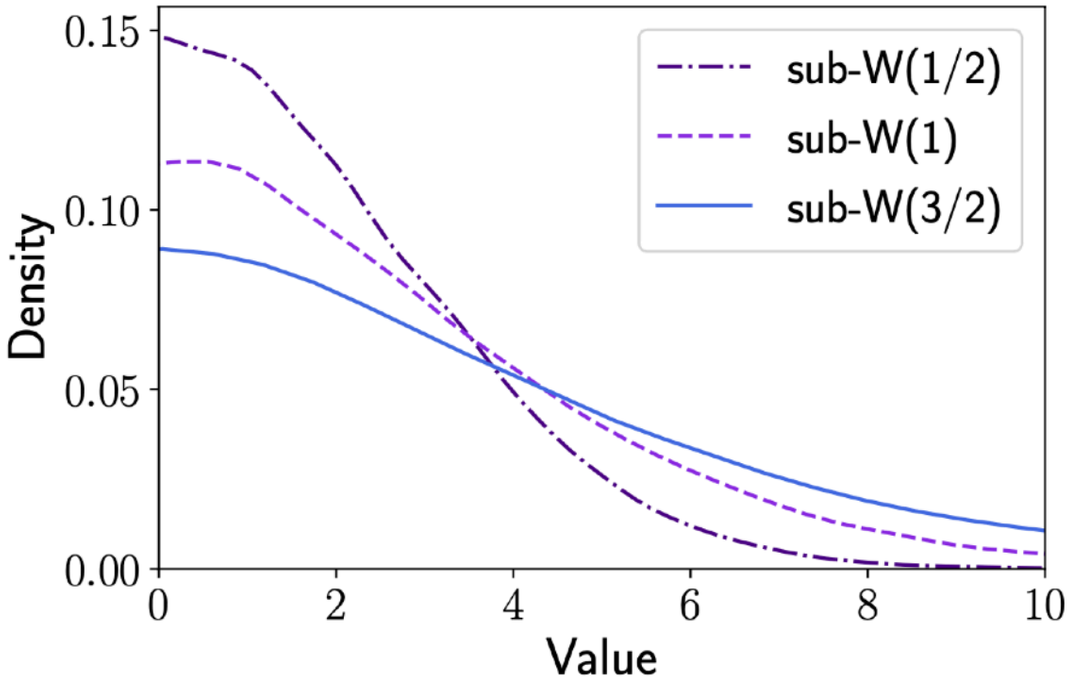
$$X \sim \mathsf{subW}(\theta).$$

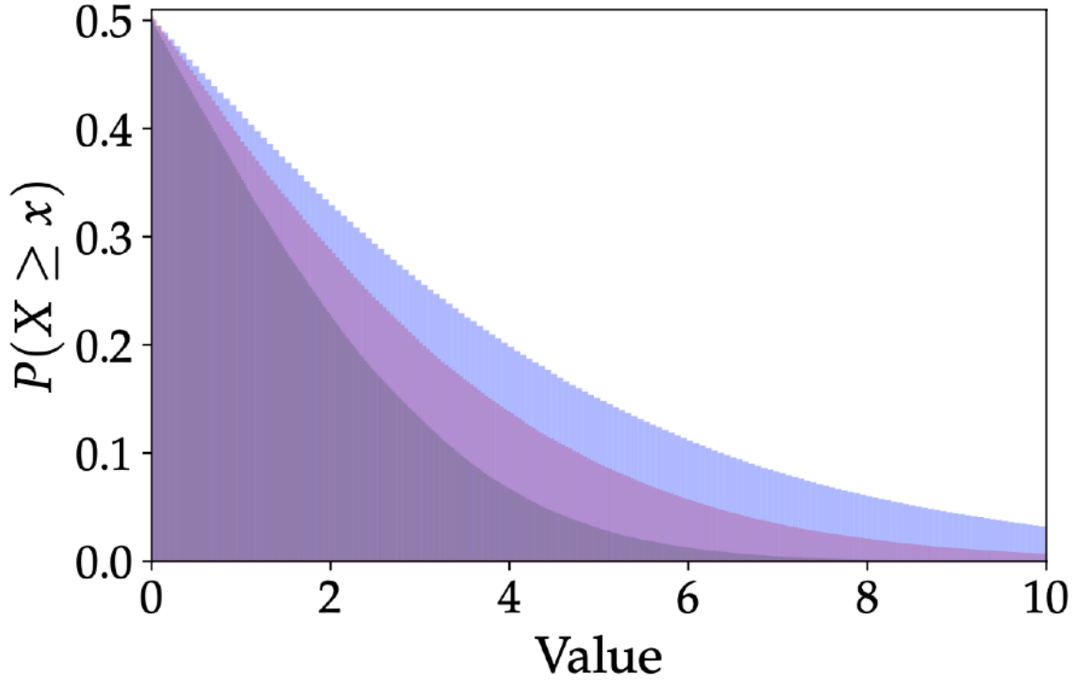
Vladimirova, M., Girard, S., Nguyen, H. and Arbel, J. (2020). Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions, Stat Journal

SUB-WEIBULL RANDOM VARIABLE

A random variable X, such that

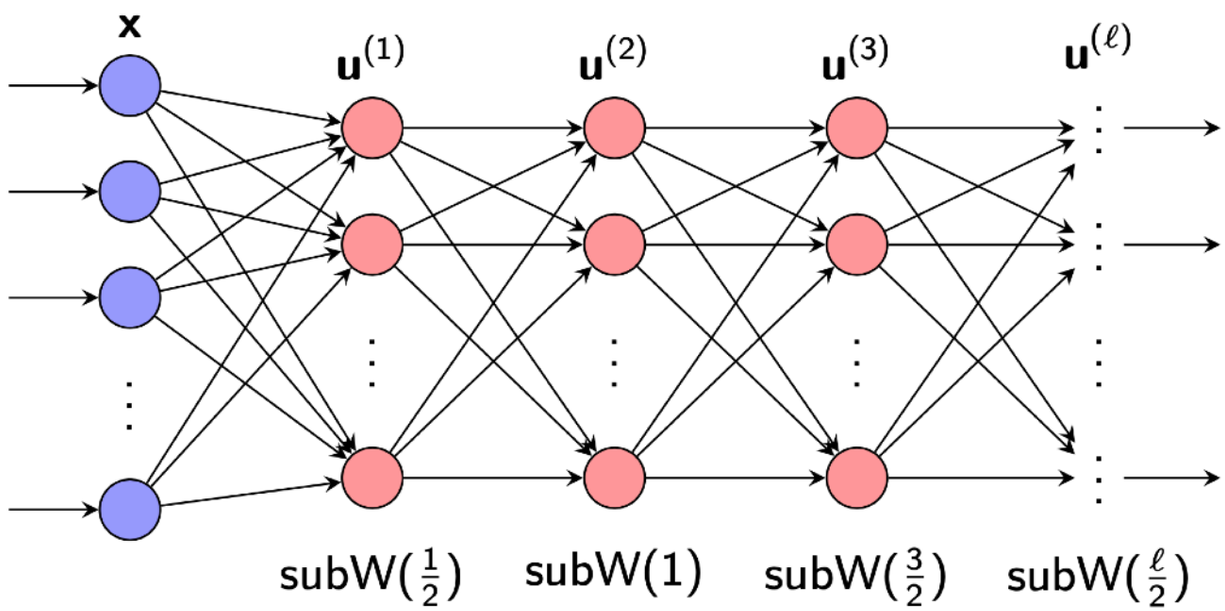
$$\mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K\right)$$





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

OUR RESULT: BNNS BECOME HEAVIER-TAILED WITH DEPTH



$$X \sim \mathsf{subW}(\theta) \quad \Leftrightarrow \quad \mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K\right)$$

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

OUR RESULT: BNNS BECOME HEAVIER-TAILED WITH DEPTH

Main Theorem

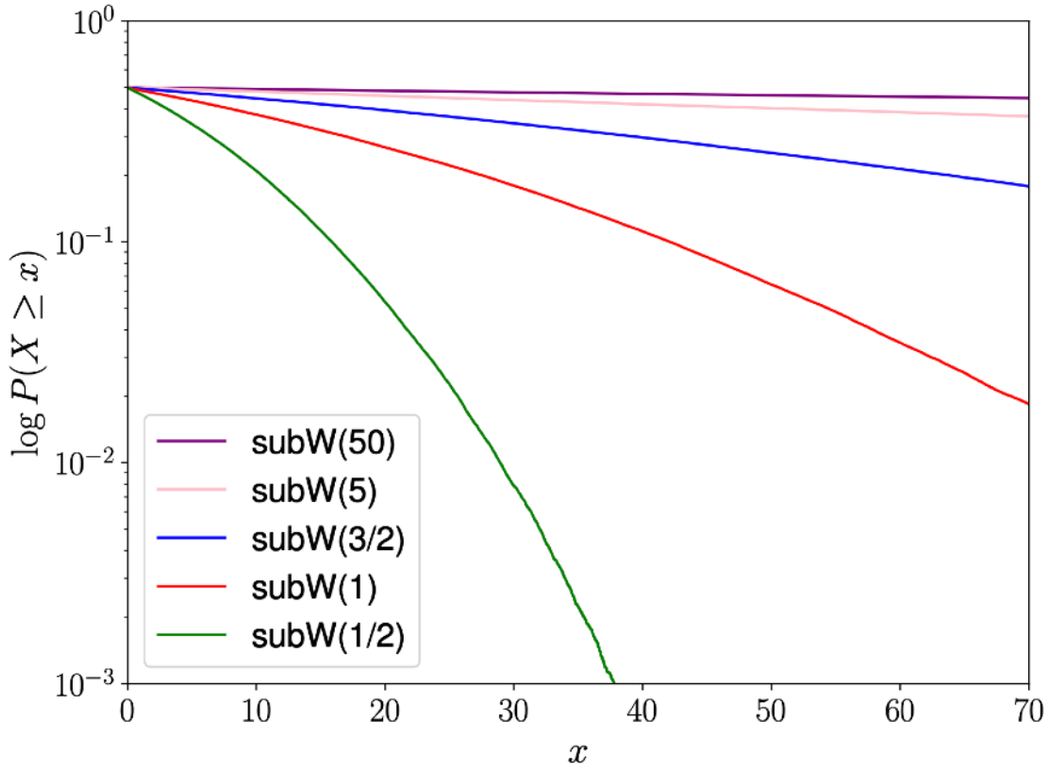
The ℓ-th hidden layer units $U^{(\ell)}$ (pre-activation $g^{(\ell)}$ or post-activation $h^{(\ell)}$) of a feed-forward Bayesian neural network with:

- **Gaussian priors** on weights and
- **extended envelope condition** activation function ϕ

have **sub-Weibull marginal prior distribution** with tail parameter $\theta = \ell/2$, conditional on the input $\boldsymbol{x}$:

$$U^{(\ell)} \sim \mathsf{subW}(\ell/2),$$

Priors for 100-layer BNN



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

POSSIBLE REGULARIZATION INTERPRETATION

POSSIBLE REGULARIZATION INTERPRETATION

Maximum a Posterior (MAP)
Is a Regularized Problem

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

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

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

Loss function

Regularizer

POSSIBLE REGULARIZATION INTERPRETATION

Maximum a Posterior (MAP)
Is a Regularized Problem

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

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

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

Loss function — Regularizer

Example:

Gaussian Prior $\Leftrightarrow$ Weight Decay (L2) Regularization

$$\pi(\mathbf{W}) = \prod_{\ell=1}^{L} \prod_{i,j} \mathrm{e}^{-\frac{1}{2}(W_{i,j}^{(\ell)})^2} \qquad R(\mathbf{W}) = \sum_{\ell=1}^{L} \sum_{i,j} (W_{i,j}^{(\ell)})^2 = \|\mathbf{W}\|_2^2$$

POSSIBLE REGULARIZATION INTERPRETATION

Maximum a Posterior (MAP)
Is a Regularized Problem

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

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

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

Loss function

Regularizer

Gaussian
Prior on
Weights

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

$$\pi(w) \approx e^{-w^2}$$

Sub-Weibull
Prior on Units

$$\mathbb{P}(|U_m^{(\ell)}| \geq u) \leq \exp\left(-u^{2/\ell}/K_1\right) \quad \text{for all } u \geq 0$$

$$\pi_m^{(\ell)}(u) \approx \mathrm{e}^{-|u|^{2/\ell}/K_1}$$

POSSIBLE REGULARIZATION INTERPRETATION

Maximum a Posterior (MAP)
Is a Regularized Problem

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

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

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

Loss function — Regularizer

Gaussian
Prior on
Weights

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

$$\pi(w) \approx e^{-w^2}$$

Sub-Weibull
Prior on Units

$$\mathbb{P}(|U_m^{(\ell)}| \geq u) \leq \exp\left(-u^{2/\ell}/K_1\right) \quad \text{for all } u \geq 0$$

$$\pi_m^{(\ell)}(u) \approx \mathrm{e}^{-|u|^{2/\ell}/K_1}$$

Sklar's theorem: $$\pi(\boldsymbol{U}) = \prod_{\ell=1}^{L} \prod_{m=1}^{H_\ell} \pi_m^{(\ell)}(U_m^{(\ell)})\, C(F(\boldsymbol{U})),$$

where C represents the copula of $\mathbf{U}$
(all the dependences between the units)

POSSIBLE REGULARIZATION INTERPRETATION

Maximum a Posterior (MAP)
Is a Regularized Problem

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

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

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

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

BUT

POSSIBLE REGULARIZATION INTERPRETATION

Weight distribution
$\pi(w) \approx e^{-w^2}$

$\Rightarrow$

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

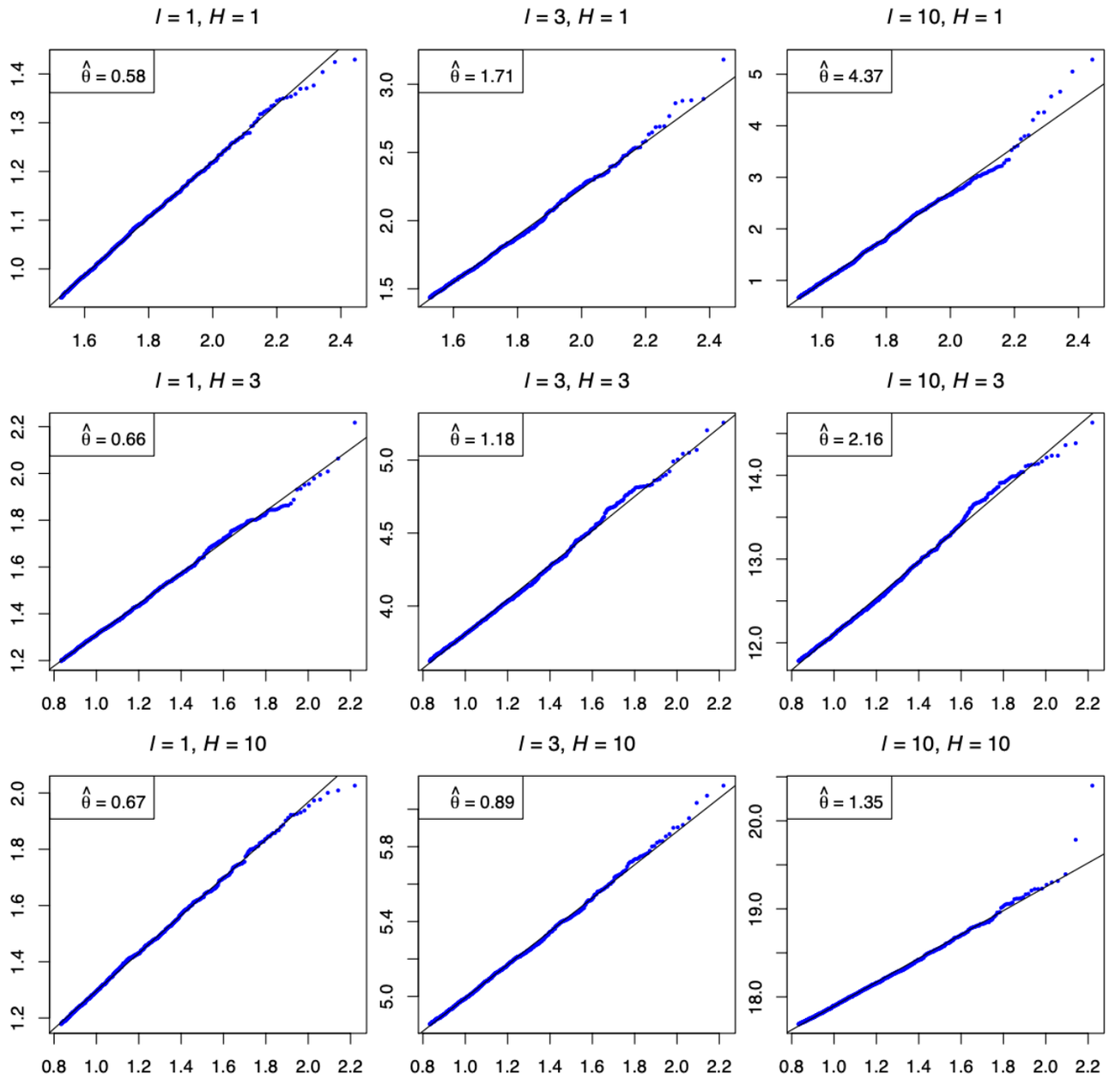
Layer	Penalty on **W**	Penalty on **U**	
1	$\|\mathbf{W}^{(1)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(1)}\|_2^2$	$\mathcal{L}^2$ (weight decay)
2	$\|\mathbf{W}^{(2)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(2)}\|$	$\mathcal{L}^1$ (Lasso)
ℓ	$\|\mathbf{W}^{(\ell)}\|_2^2$, $\mathcal{L}^2$	$\|\mathbf{U}^{(\ell)}\|_{2/\ell}^{2/\ell}$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

Two approximations:

- Unit distribution by its tails,
- Joint distribution by Sklar's theorem

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

TAIL COEFFICIENT ESTIMATION



Depth L equals 1, 3 and 10

Width H equals 1, 3 and 10 (for all considered layers L)

Theoretical:

$\theta = \ell/2$ for finite H

$\theta = 1/2$ for infinite H

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

Approximations

- Laplace approximation (MacKay, 1992, Neur. Comp.)
- Variational inference (Hinton and van Kamp, 1993, Barber & Bishop, 1998, NIPS, Blundell et al, 2015, ICML)
- Monte Carlo dropout (Gal & Ghahramani, 2016, ICML)

Bayesian NN under a Laplace approximation

Chap 5.7 of Bishop

Univariate regression $t \in \mathbb{R}$

Goussian $p(t \mid x)$ to be Goussian

the mean $= y(x, w)$ output of NN

input, weight

variance β^{-1} β precision

Data $\mathcal{D} = \{(x_n, t_n), n = 1 \dots N\}$

$$\begin{cases} \text{Model} & p(t \mid x) = \mathcal{N}(t \mid y(x, w), \beta^{-1}) \\ \text{Prior} & p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I) \end{cases}$$

$$\text{Posterior} \quad p(w \mid \mathcal{D}, \alpha, \beta) \propto \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, w), \beta^{-1}) \, p(w \mid \alpha)$$

w_{MAP} : numerically find local optimum.

$$\log p(w \mid \mathcal{D}, \alpha, \beta) = \underbrace{-\frac{\alpha}{2} w^T w}_{\text{penalty to the MLE.}} - \frac{\beta}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^2 + c^t$$

Laplace approx. of the posterior:

$$q(w \mid \mathcal{D}, \alpha, \beta) = N(w \mid w_{MAP}, A^{-1})$$

$$A = -\nabla\nabla \log p = \alpha I + \beta H \text{, with } H = \text{Hess}(SSE)$$

approx. predictive

$$p(t \mid x, \mathcal{D}) = \int_w \underbrace{p(t \mid x, w)}_{N(t \mid y(x,w), \beta^{-1})} q(w \mid \mathcal{D}, \alpha, \beta) \, dw \quad (*)$$

Taylor approximation for NN :

$$y(x,w) \approx y(x, w_{MAP}) + g^T (w - w_{MAP})$$

$$g = \nabla_w y(x,w) \big|_{w=w_{MAP}}$$

$$p(t \mid x, w, \beta) \approx N\left(t \mid y(x, w_{MAP}) + g^T (w - w_{MAP}), \beta^{-1}\right)$$

plug this in (*) :

Plus use $\begin{cases} p(x) = N(x \mid \mu, \Lambda^{-1}) \\ p(y \mid x) = N(y \mid Ax + b, L^{-1}) \end{cases}$

$$\Rightarrow p(y) = N(y \mid A\mu + b, \underbrace{L^{-1} + A\Lambda^{-1}A^T})$$

$$p(t \mid x, \mathcal{D}, \alpha, \beta) \approx N(t \mid y(x, w_{MAP}), \sigma^2(x))$$

$$\sigma^2(x) = \underbrace{\beta^{-1}}_{\text{aleatoric}} + \underbrace{g^T A^{-1} g}_{\text{x-dependent : epistemic.}}$$

α, β hyperparameters

Marginal likelihood $p(\mathcal{D} \mid \alpha, \beta)$, aka evidence

$\rightsquigarrow$ w is integrated out.

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)}{\operatorname{argmax}}\, p(\mathcal{D} \mid \alpha, \beta).$$

$$p(\mathcal{D} \mid \alpha, \beta) = \int p(\mathcal{D} \mid w, \beta)\, p(w \mid \alpha)\, dw$$

Laplace approximation: z_0 argmax, $A = -\nabla\nabla \log f(z)\big|_{z=z_0}$

$$\int f(z)\, dz = f(z_0) \int \exp\left(-\frac{1}{2}(z - z_0)^T A (z - z_0)\right) dz$$

$$= f(z_0)\, \frac{(2\pi)^{M/2}}{|A|^{1/2}} \qquad M = \#(z)$$

$$\Rightarrow \log p(\mathcal{D} \mid \alpha, \beta) = \underbrace{-E(w_{MAP})}_{\text{regularized SSE}} - \frac{1}{2}\log|A| + \frac{W}{2}\log\alpha + \frac{N}{2}\log\beta + \underbrace{\frac{\alpha}{2} w_{MAP}^T w_{MAP} + \frac{\beta}{2}\sum_{n=1}^{N}\left(y(x_n, w_{MAP}) - t_n\right)^2}_{\text{pen.}}$$

optimization in (α, β) is done by analogy with linear regression. → Bishop 5.7.

Other techniques for practical BNN:

variational inference: with various assumptions on approx family.

VI (Blundell, 2015)

parameters are weights $w = (w_1, \dots, w_{\bar{w}})$

Approximate family : Gaussian : $N(\mu, \sigma^2)$

w_i approximated by $\theta_i = (\mu_i, \sigma_i^2)$, $i = 1, \dots, \bar{w}$

$$q_\theta(w) = \prod_{i=1}^{\bar{w}} N(w_i \mid \mu_i, \sigma_i^2)$$

$$Q = \{ q_\theta(w), \ \mu_i \in \mathbb{R}, \ \sigma_i > 0 \}$$

$$KL\left(q_\theta(w) \,\|\, p(w \mid X, Y) \right) \qquad X, Y \begin{cases} X = (X_i) \\ Y = (Y_i) \end{cases} \quad i = 1 \dots N$$

$$= \int q_\theta(w) \log \left(\frac{q_\theta(w)}{p(w \mid X, Y)} \right) dw$$

$$= \int q_\theta(w) \log \left(\frac{q_\theta(w)\, p(Y \mid X)}{p(w)\, p(Y \mid w, X)} \right) dw$$

Some recent works

Other choices of priors for BNNs

[Fortuin et al, 2021. Bayesian Neural Network Priors Revisited. arxiv]

- Questions the use of iid Gaussian priors for weights in Bayesian Neural Networks
- Shows suboptimality of Gaussian priors essentially through the lens of the cold posterior effect
- Compares w/ other priors including Laplace, Student-t and Multivariate Gaussian with Matern covariance

Bayesian Neural Network Priors Revisited

When performing inference in Bayesian models, we can temper the posterior by a positive temperature T, giving

$$\log p(w|x,y)^{\frac{1}{T}} = \frac{1}{T}[\log p(y|w,x) + \log p(w)] + Z(T)$$

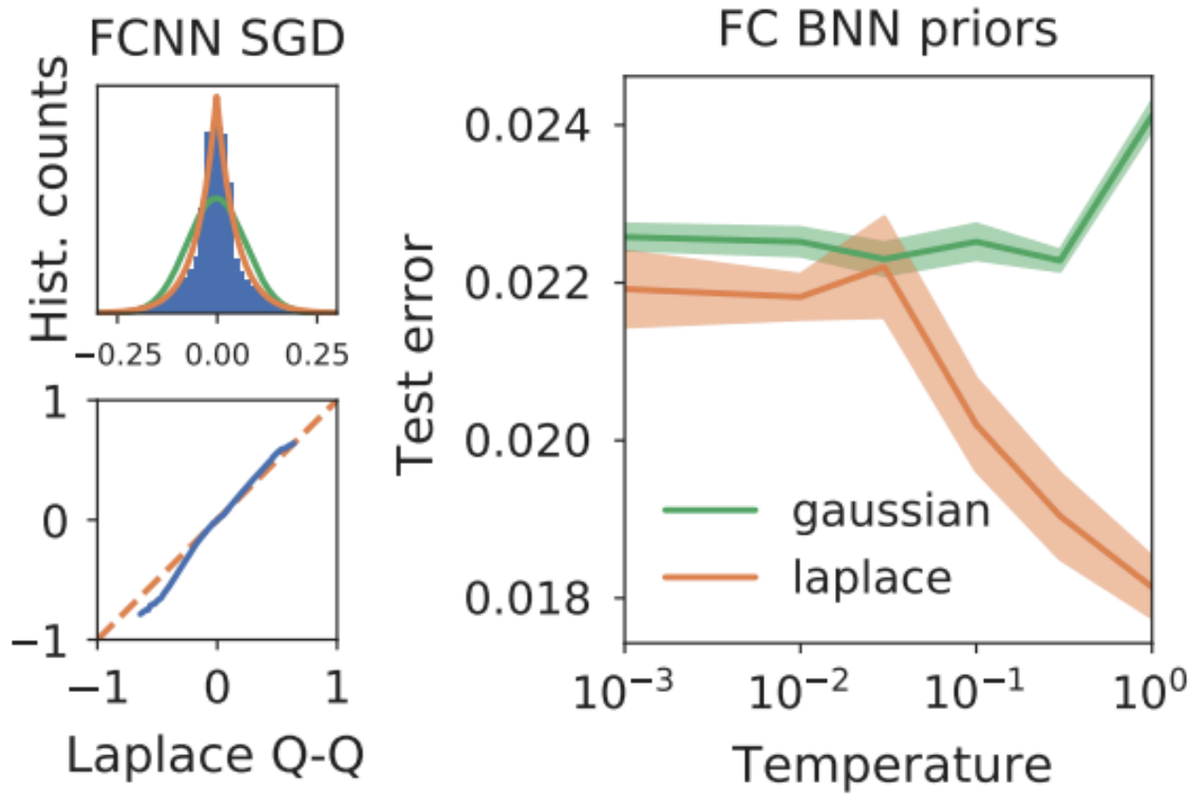


Figure 1: The weights of SGD-trained neural networks on MNIST follow a distribution that is more heavy-tailed than a Gaussian, and better approximated by a Laplace. When using a Laplace prior instead of a Gaussian prior for a BNN on the same task, the performance is improved and the cold posterior effect inverted, such that now the true Bayes posterior ($T = 1$) performs best.

[Fortuin et al, 2021. Bayesian Neural Network Priors Revisited. arxiv]

Bayesian Neural Network Priors Revisited

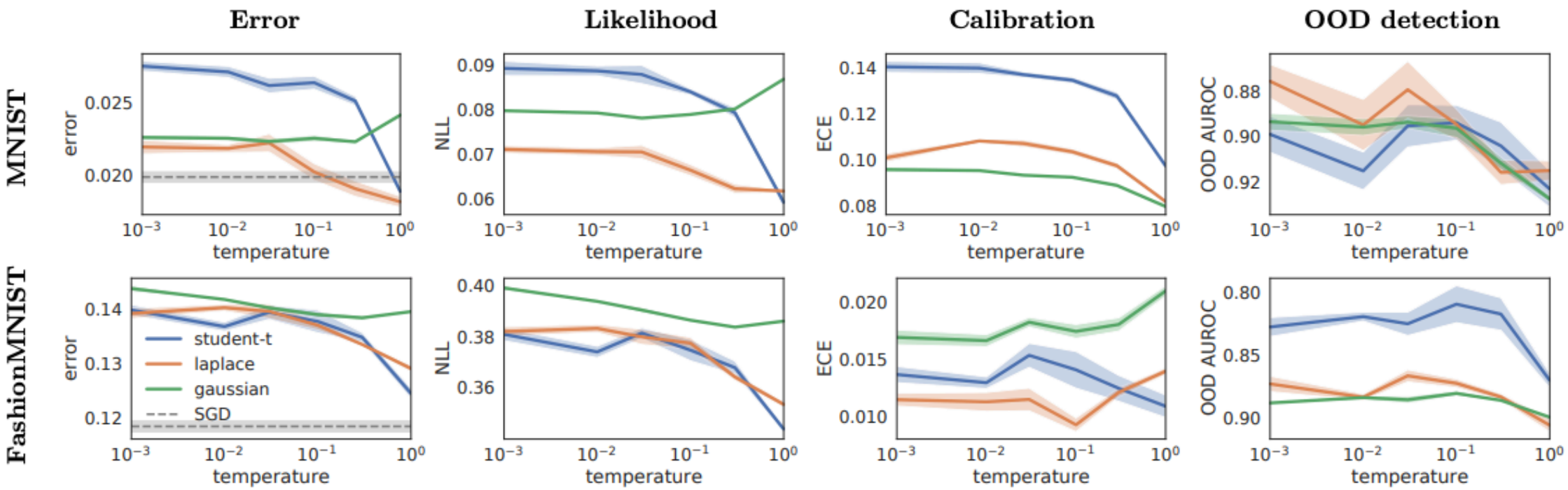


Figure 6: Performances of fully connected BNNs with different priors on MNIST and FashionMNIST (see Sec. 4.2). The heavy-tailed priors generally perform better, especially at higher temperatures, and lead to a less pronounced cold posterior effect. Note the reversed y-axis for OOD detection on the right to ensure that lower values are better in all plots.

[Fortuin et al, 2021. Bayesian Neural Network Priors Revisited. arxiv]

Bayesian deep learning

Pyro example on Bayesian neural networks

- Demonstrates how to use NUTS to do inference on a simple (small) Bayesian neural network with two hidden layers.