Bayesian Inference for Deep Learning

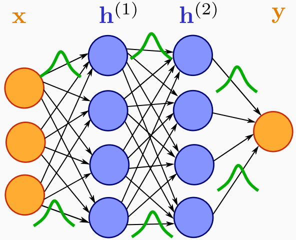
Inference and modern trends for Bayesian Neural Networks: Practical considerations on priors

Simone Rossi and Maurizio Filippone

Data Science Department, EURECOM (France)

Priors for Bayesian Neural Networks

Specifying a prior for Bayesian neural networks is difficult

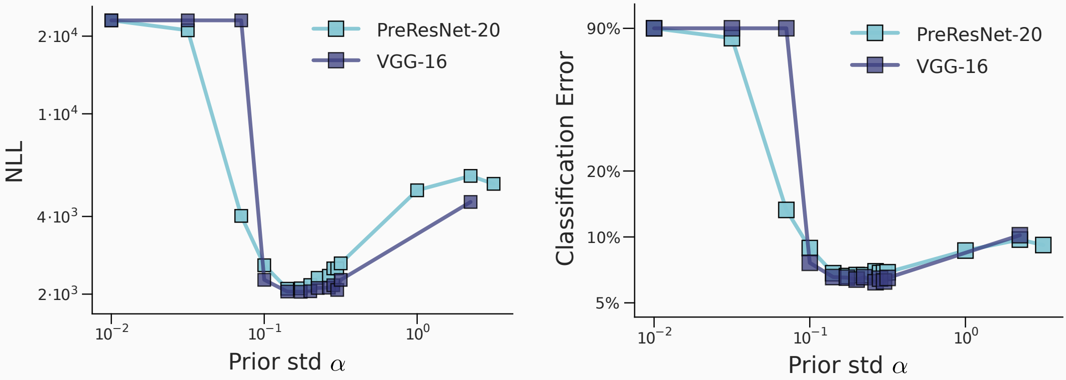


$$W^{(l)}{}_{ij} \sim \mathcal{N}(0, \alpha^2)$$

- Neural networks are extremely *high-dimensional* and *unidentifiable*.
  $\longrightarrow$ Reasoning about parameters is very challenging.
- Most work has resorted to priors of convenience.
  $\longrightarrow \mathcal{N}(0, 1)$ and $\mathcal{N}(0, 1/D^{(l-1)})$ are popular priors for BNN.

For notation, ψ is the set of parameters for the prior (α in this case).

Effect of priors in predictive tasks

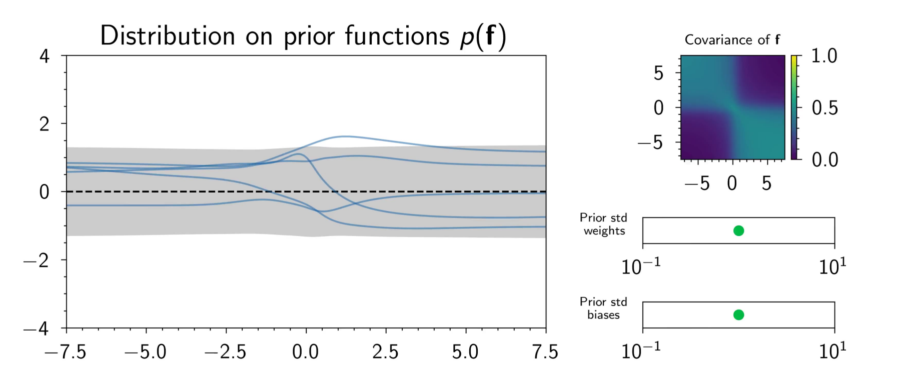


Running a grid search is intractable for Bayesian models.

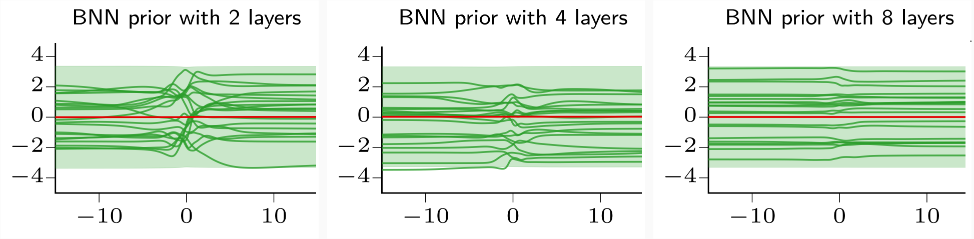
Wilson and Izmailov (2020). *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*.

Prior for Bayesian neural networks

The prior on the parameters of a BNN induces an *unpredictable prior over functions*.



Prior for Bayesian Neural Networks



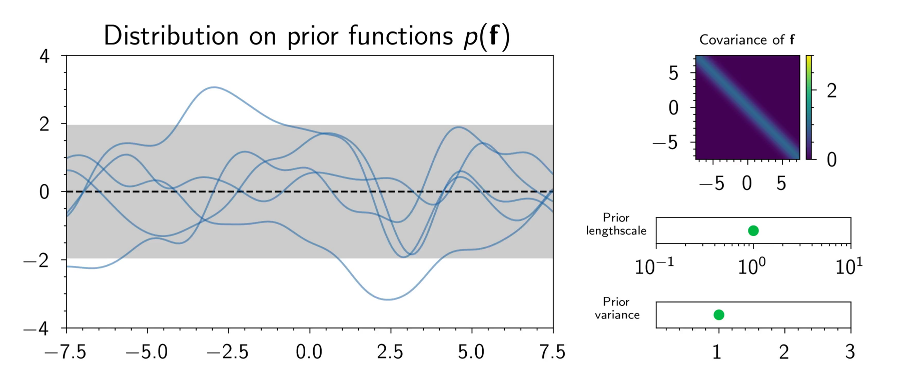
The prior $\mathcal{N}(0, 1)$ is not always problematic, but it can be for deep architectures.

- The sampled functions tend to form straight horizontal lines.
- This is a well-known pathology stemming from increasing model's depth.

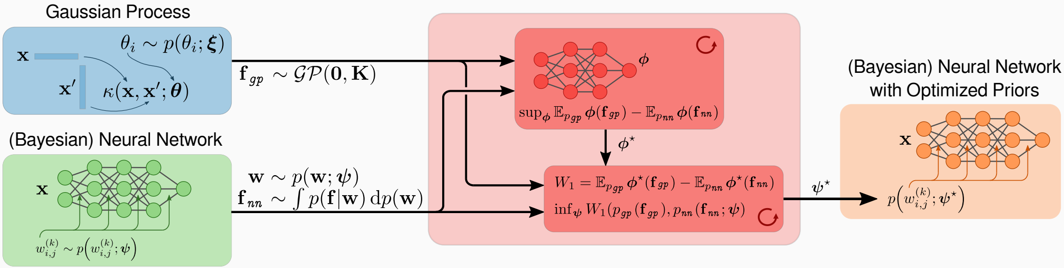
Duvenaud et al. (2014). *Avoiding Pathologies in Very Deep Networks*. AISTATS

Gaussian processes as prior on functions

GP are a useful tool for choosing *sensible priors* on *functions we indent to model*.



Using Gaussian Processes as reference



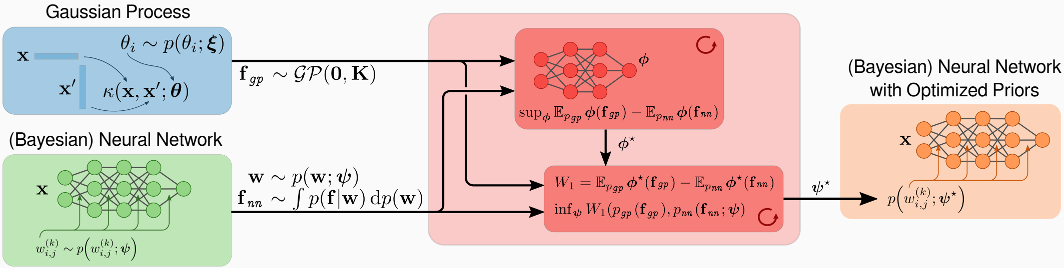
Minimize the Wasserstein distance between samples of $p_{gp}(\mathbf{f})$ and $p_{nn}(\mathbf{f}; \psi)$.

$$\min_{\boldsymbol{\psi}} W_1(p_{gp}, p_{nn}) = \min_{\boldsymbol{\psi}} \max_{\boldsymbol{\phi}} \mathbb{E}_q \left[\underbrace{\mathbb{E}_{p_{gp}}[\phi(\mathbf{f})] - \mathbb{E}_{p_{nn}}[\phi(\mathbf{f})]}_{\mathcal{L}(\boldsymbol{\psi}, \boldsymbol{\phi})} \right],$$

where ϕ is the 1-Lipschitz function parameterized by a neural network.

Tran et al. (2020). *All You Need is a Good Functional Prior for Bayesian Deep Learning*

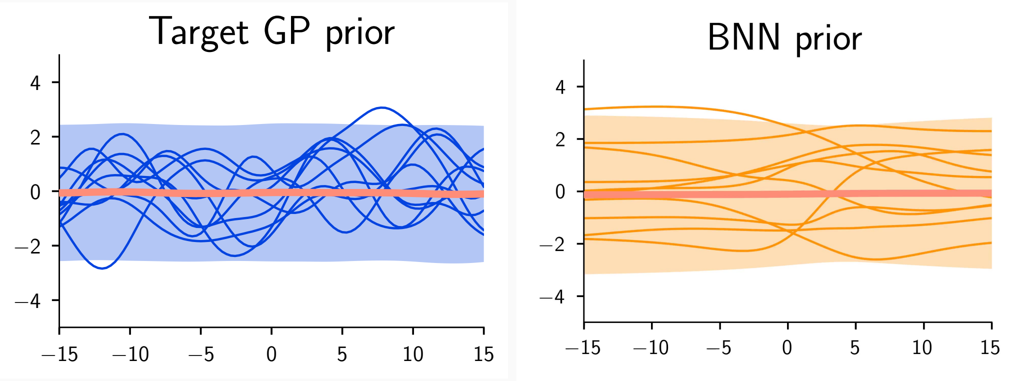
Using Gaussian Processes as reference



- The objective is *fully sampled-based*
  - $\longrightarrow$ Not necessary to know the closed-form of the marginal density $p_{nn}(\mathbf{f}; \psi)$.
  - $\longrightarrow$ Can consider any stochastic process as a target prior over functions.
- The objective can be optimized with gradient descent algorithms with back-propagation.

Tran et al. (2020). *All You Need is a Good Functional Prior for Bayesian Deep Learning*

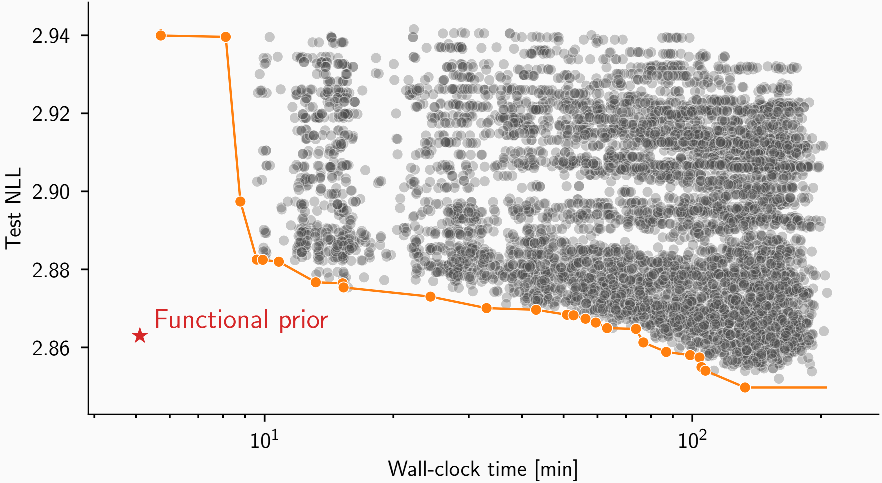
“Learning” priors by matching stochastic processes



The flexibility of the scheme allows for using more complex prior distributions, like *normalizing flows*.

Grid-search? Functional prior!

Cross-validation with 64 parallel workers



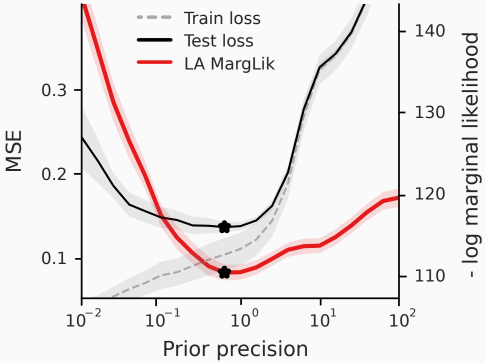
Prior matters in practice (CIFAR10)

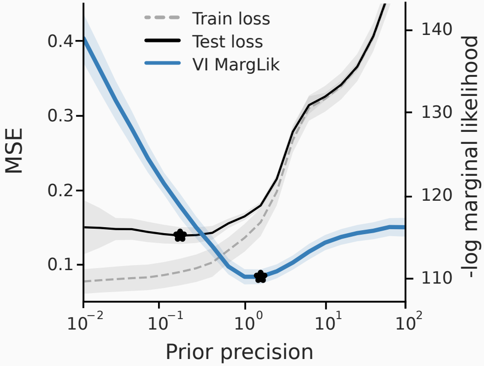
Architecture	Method	Accuracy (↑)	NLL (↓)
VGG	Gauss. prior	81.25%	0.5826
	GPi Gauss. prior	82.94%	0.5292
	GPi Hierarchical prior	**87.11%**	**0.406**
PreResNet	Gauss. prior	85.45%	0.4915
	GPi Gauss. prior	86.41%	0.4513
	GPi hierarchical prior	**88.31%**	**0.3796**

Tran et al. (2020). *All You Need is a Good Functional Prior for Bayesian Deep Learning*

Empirical Bayes with approximate inference

Use Laplace approximation and Variational Inference as proxy to marginal likelihood optimization.





Khan et al. (2019). *Approximate Inference Turns Deep Networks into Gaussian Processes*. NeurIPS