Functional Priors for Bayesian Deep Learning


Ba-Hien Tran


Simone Rossi


Dimitrios Milios


Pietro Michiardi


Edwin V. Bonilla


Maurizio Filippone

The ISBA World Meeting, Montreal – 2022

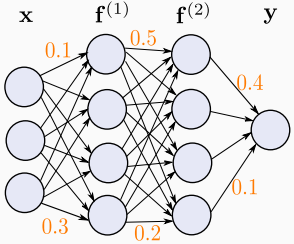
Neural Networks

- Deep neural networks (NNs) are powerful models that have achieved impressive performance.
- But by default, Deep NNs do not provide *uncertainty quantification*.

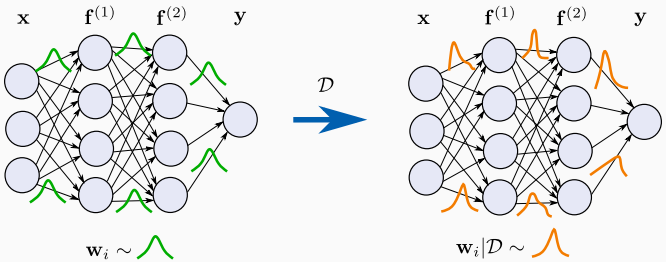
  - Typical Deep neural networks (NNS) solution: a point estimate.
  - Maximum likelihood estimation given a data set $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{w})$$

  - Regularization: L_2 parameter regularization, early stopping, dropout, etc.



Bayesian Neural Networks



Place a prior distribution $p(\mathbf{w})$ over the network's parameters $\mathbf{w}$.

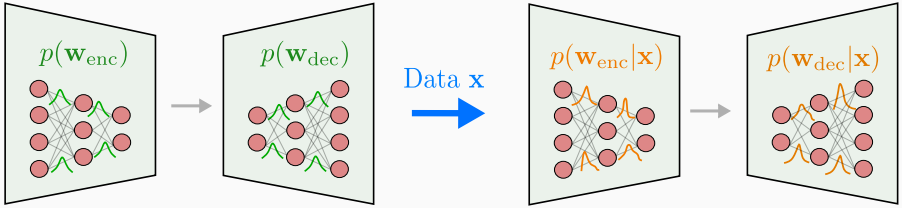
- Compute posterior given a data set $\mathcal{D}$:

$$\underbrace{p(\mathbf{w} \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid \mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

- Posterior predictive:

$$p(\mathbf{y}^* \mid \mathbf{x}^*, \mathcal{D}) = \mathbb{E}_{p(\mathbf{w} \mid \mathcal{D})}[p(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{w})]$$

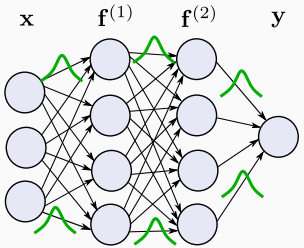
Bayesian Neural Networks for Unsupervised Learning



- Place a prior $p(\mathbf{w})$ over the network's parameters $\mathbf{w} := \{\mathbf{w}_{\text{enc}}, \mathbf{w}_{\text{dec}}\}$
- The target is exactly the input, $\mathbf{y}_n = \mathbf{x}_n$
- Compute posterior given a dataset $\{\mathbf{x}\}$:

$$\underbrace{p(\mathbf{w} \mid \mathbf{x})}_{\text{posterior}} \propto \underbrace{p(\mathbf{x} \mid f(\mathbf{x}; \mathbf{w}))}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

Prior for Bayesian Neural Networks



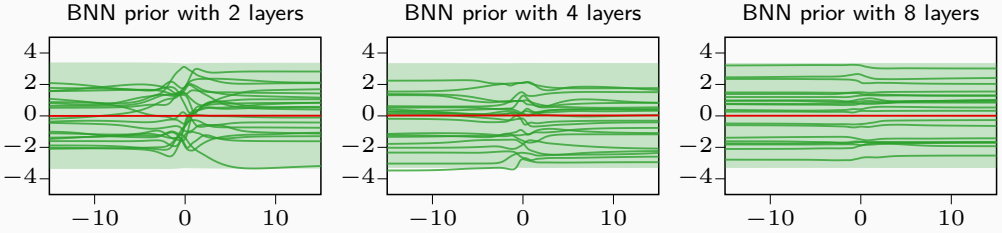
$$\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{I})$$

Specifying a prior for Bayesian neural networks (BNNs) is difficult!

- Neural networks are extremely *high-dimensional* and *unidentifiable*.
  $\longrightarrow$ Reasoning about parameters is very challenging.
- Most work has resorted to priors of convenience.
  $\longrightarrow$ Gaussian priors such as $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 1/D_{l-1})$ are the most popular priors for Bayesian neural network (BNN).

Prior for Bayesian Neural Networks

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



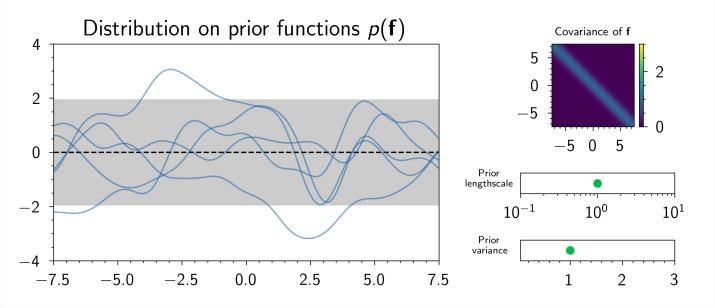
The prior $\mathcal{N}(0, 1)$ is not always problematic, but it can be for deep architectures.

- This is a well-known pathology stemming from increasing model's depth [1].

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

Supervised Learning - Sensible priors through Gaussian Processes

- Gaussian Processes (GPs) are a useful tool for choosing *sensible priors* in *function space*.
- GP is characterized by a mean function $\mu(\cdot)$ and a covariance function $\kappa(\cdot, \cdot)$.

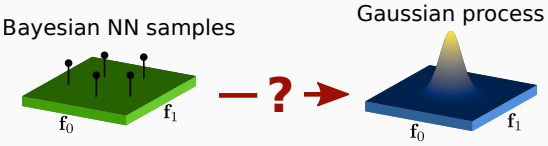


- [**Neal**, **1996**] [2] showed that BNN converge to GP in the limit of infinite network width.

[2] Neal (1996). *Bayesian Learning for Neural Networks*.

Research Question

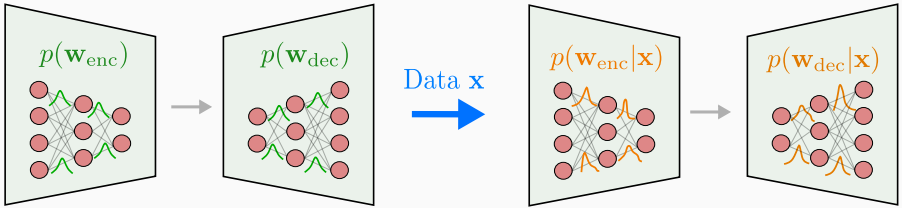
How to impose functional priors on BNNs to enable interpretability, similar to GPs?



This is a challenging task!

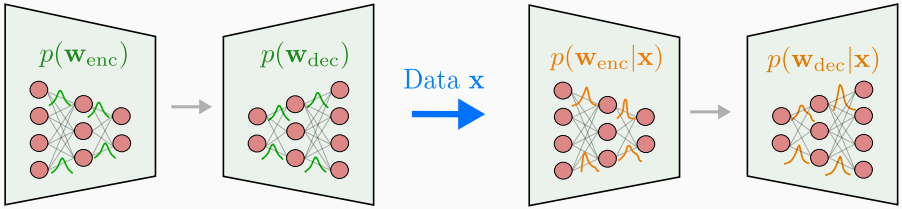
- We aim at matching two stochastic processes $\rightarrow$ infinite-dimensional distributions.
- We don't know closed-form of the density of BNNs.
  - KL minimization is not possible!
  - The **Wasserstein Distance**, instead, is tractable because it's sample-based!

Priors for Bayesian Autoencoders



- We consider the common practice of estimating prior (hyper)parameters via type-II ML
- Equivalent to match functional distribution induced by BAE to the data distribution $\pi(\mathbf{x})$
- A KL divergence-based objective is intractable (unknown form of the functional prior)
- We employ the **Wasserstein Distance** - sample based!

Priors for Bayesian Autoencoders, Details



$$p_\psi(\hat{\mathbf{x}}) = \int f(\mathbf{x};\mathbf{w}) p_\psi(\mathbf{w}) d\mathbf{w} \rightarrow \text{functional prior over BAE output}$$

$$\text{where } \hat{\mathbf{x}} = f(\mathbf{x};\mathbf{w}) \rightarrow \text{functional output of BAE}$$

$$p_\psi(\mathbf{x}) = \int p(\mathbf{x}|\hat{\mathbf{x}}) p_\psi(\hat{\mathbf{x}}) d\hat{\mathbf{x}} \rightarrow \text{marginal likelihood}$$

$$\arg\max_\psi \int \pi(\mathbf{x}) \log p_\psi(\mathbf{x}) = \arg\min_\psi \mathrm{KL}\left[\pi(\mathbf{x}) \parallel p_\psi(\mathbf{x})\right]$$

Wasserstein distance

Definition

Given a measurable space Ω, the Kantorovich dual form of the 1-Wasserstein distance between two Borel's probability measures π and ν in $\mathcal{P}(\Omega)$ is

$$W_1(\pi, \nu) = \sup_{\|\phi\|_L \leq 1} \mathbb{E}_\pi[\phi(\mathbf{x})] - \mathbb{E}_\nu[\phi(\mathbf{x})],$$

where ϕ is a 1-Lipschitz function.

✓ No need to know the closed-form of π and ν as we can estimate expectations with samples.

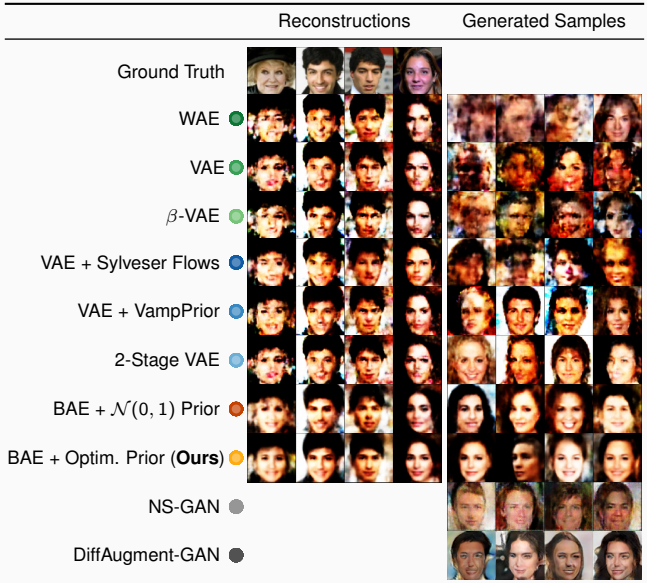
✓ The 1-Lipschitz function ϕ can be parameterized by a neural network.

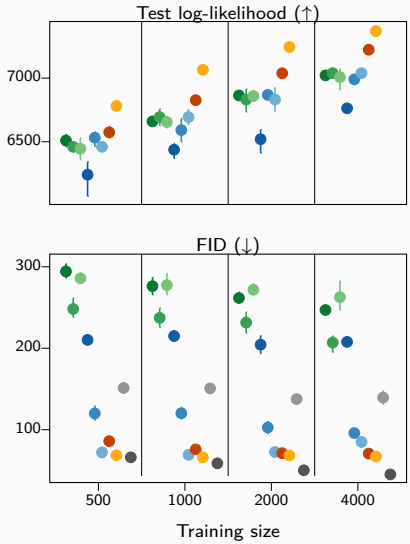
Results

Bayesian Convolutional Neural Networks - CIFAR-10

Architecture	Method	Accuracy - % (↑)	NLL (↓)
VGG16	Deep Ensemble	81.96 ± 0.33	0.7759 ± 0.0033
	Fixed Gauss. prior	81.47 ± 0.33	0.5808 ± 0.0033
	Fixed Gauss. prior + Temp. Scaling	82.25 ± 0.15	0.5398 ± 0.0015
	GPi Gauss. prior (**ours**)	83.34 ± 0.53	0.5176 ± 0.0053
	Fixed Hierar. prior	86.03 ± 0.20	0.4345 ± 0.0020
	GPi Hierar. prior (**ours**)	**87.03** ± **0.07**	**0.4127** ± **0.0007**
PRERESNET20	Deep Ensemble	87.77 ± 0.03	0.3927 ± 0.0003
	Fixed Gauss. prior	85.34 ± 0.13	0.4975 ± 0.0013
	Fixed Gauss. prior + Temp. Scaling	87.70 ± 0.11	0.3956 ± 0.0011
	GPi Gauss. prior (**ours**)	86.86 ± 0.27	0.4286 ± 0.0027
	Fixed Hierar. prior	87.26 ± 0.09	0.4086 ± 0.0009
	GPi Hierar. prior (**ours**)	**88.20** ± **0.07**	**0.3808** ± **0.0007**

Experiments on CelebA Dataset





Conclusions

Conclusions

- Choosing sensible priors for deep models is very important and difficult.
- Proposed a novel objective based on the Wasserstein distance.
  - Impose sensible priors for BNNs in function space.
  - Showed empirical benefits on a variety of neural networks on supervised and unsupervised tasks.
  - Demonstrated that a fully Bayesian treatment of Bayesian deep learning models provides the large performance gains.

References

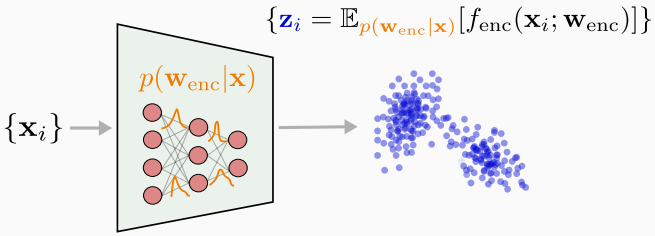
[1] Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. **All You Need is a Good Functional Prior for Bayesian Deep Learning**. *Journal of Machine Learning Research*, 2022.

[2] Ba-Hien Tran, Simone Rossi, Dimitrios Milios, Pietro Michiardi, Edwin V. Bonilla, and Maurizio Filippone. **Model Selection for Bayesian Autoencoders**. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, 2021.

Q&A

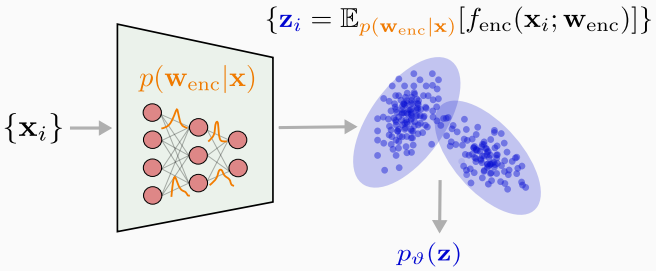
Generative Modeling for Bayesian Autoencoders

Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space



Generative Modeling for Bayesian Autoencoders

Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space



Generative Modeling for Bayesian Autoencoders

Use a Dirichlet process mixture model (Blei and Jordan, 2006) for density estimation in latent space