Are Bayesian neural networks intrinsically good at out-of-distribution detection?

UDL Workshop 2021

Christian Henning*, **Francesco D'Angelo***, Benjamin F. Grewe

Francesco D'Angelo

* Equal contribution

Out-of-distribution (OOD) detection

We consider a supervised learning problem: $\mathcal{D} \overset{i.i.d.}{\sim} p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y} \mid \boldsymbol{x})$

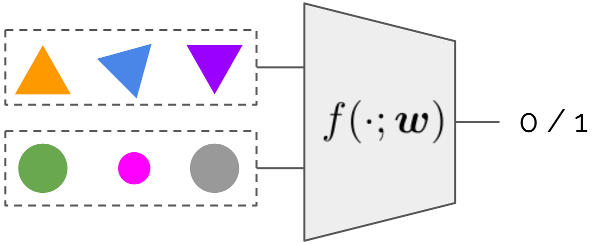
where the goal is to learn the parameters of a neural network $f(\cdot; \boldsymbol{w})$ such that, e.g.:

$$\mathbb{E}_{p(\boldsymbol{x})}\left[\text{KL}\Big(p(\boldsymbol{y} \mid \boldsymbol{x}) \Big|\Big| p(\boldsymbol{y} \mid f(\cdot; \boldsymbol{w}))\Big)\right] \approx 0$$

Data-generating process

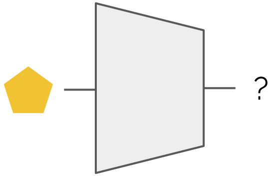
Model

Training



Intuitively, an **OOD point** is an input that is unfamiliar (given the training data), for which we should abstain from making predictions with the learned model.

Deployment



OOD detection via Bayesian Neural Networks (BNN)

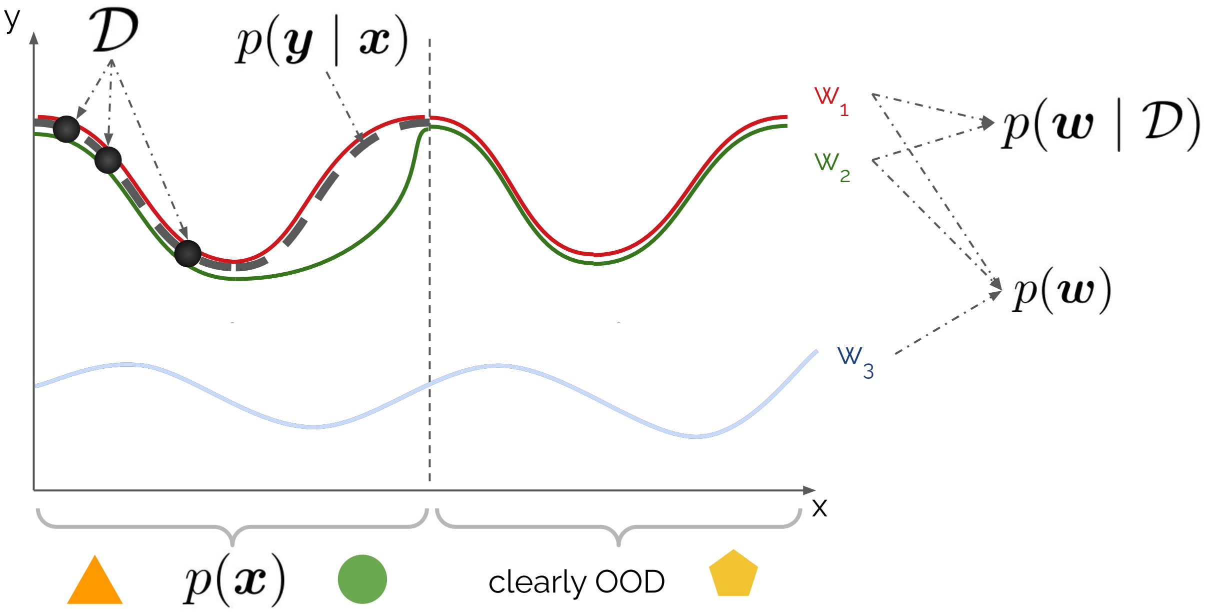
Given a likelihood defined via a neural network and a chosen weight prior $p(\boldsymbol{w})$
BNNs utilize Bayesian statistics to maintain a posterior $p(\boldsymbol{w} \mid \mathcal{D})$

This allows them to capture both **aleatoric** (data-intrinsic) and **epistemic** (limited data availability) uncertainty.

Question: Can we use a BNN's uncertainty to approach the OOD problem?

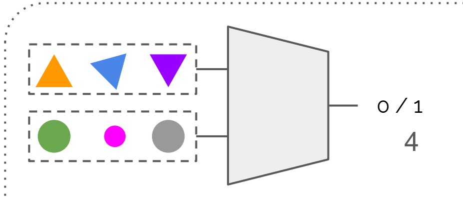
The two ingredients for OOD detection via BNNs

1st ingredient: **epistemic uncertainty**



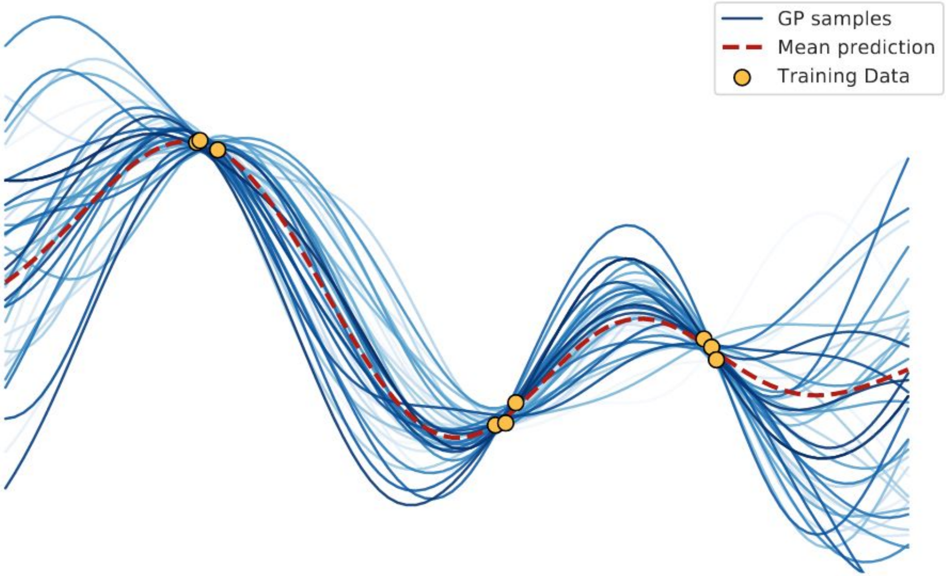
In this toy hypothesis class

➔ **epistemic uncertainty on OOD data vanishes after seeing the data**



The two ingredients for OOD detection via BNNs

2nd ingredient: **rich hypothesis space**



- Neural networks can be universal function approximators

➔ Can we model a **distribution over functions** that **agree** on the seen data but **disagree** everywhere else?

- The **chosen architecture and weight prior** determine the induced **prior in function space**[1]

➔ We don't know how to choose an architecture to allow powerful function approximation

➔ We don't know how much the chosen weight prior restricts the function approximation capabilities of the given architecture

[1] Wilson & Izmailov, "Bayesian Deep Learning and a Probabilistic Perspective of Generalization", 2020.

Are BNNs good at OOD detection?

It seems widely assumed that proper Bayesian inference with neural networks leads to a model that "**knows what it doesn't know**".

➔ For instance, OOD detection is a common benchmark to validate new approximate inference methods, implying that the true posterior is good at OOD

A formal understanding under which conditions BNNs are good at OOD detection is lacking to the best of our knowledge!

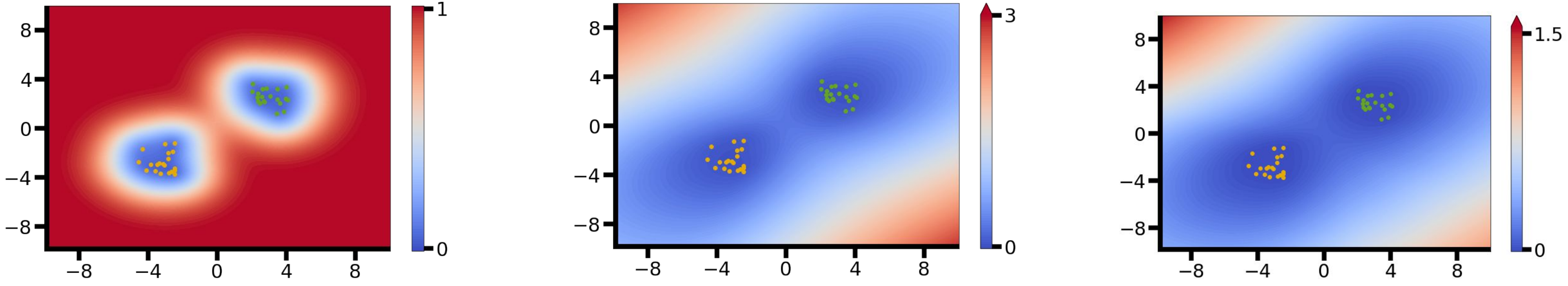
➔ However, such theoretical basis would be desirable for safety-critical applications

Our work aims to create awareness about this problem!

Our approach to illustrate the problem

- For certain weight priors, **neural networks converge to Gaussian processes (GP) in the infinite-width limit**[2]
- **Bayesian inference can be exact** in this limit, which allows us to study the OOD capabilities of the true posterior
- We can use **HMC on finite-width networks** to verify whether the **OOD behavior is consistent** with the infinite-width case

Plots show epistemic uncertainty



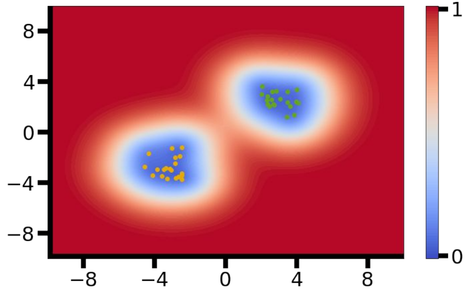
GP with RBF Kernel | 2-layer ReLU (width: ∞) | 2-layer ReLU (width: 20)

[2] R. Neal, "Bayesian Learning for Neural Networks", Springer, 1996.

Why does the RBF kernel perform best?

The analytic expression of the posterior variance for a **GP with RBF kernel** is reminiscent of:

$$\text{const.} - p(\boldsymbol{x})$$

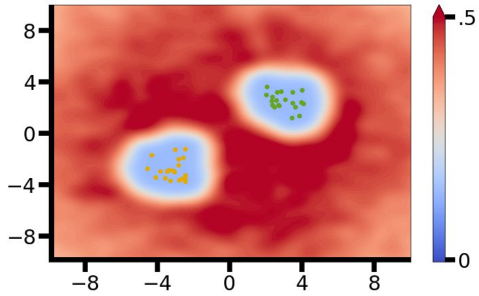


GP with RBF Kernel

Can we obtain a similar behavior with BNNs?

The kernel induced by an **infinite-width RBF network** has promising properties[3]:

$$k(\boldsymbol{x}, \boldsymbol{x}') \propto \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2\sigma_m^2}\right)\exp\left(-\frac{\|\boldsymbol{x}'\|^2}{2\sigma_m^2}\right)\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma_s^2}\right)$$

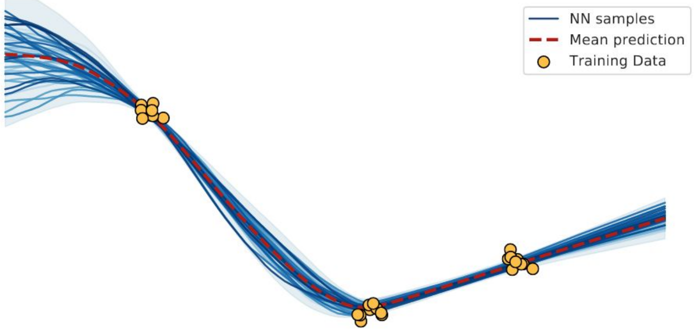


RBF network
(width: 500)

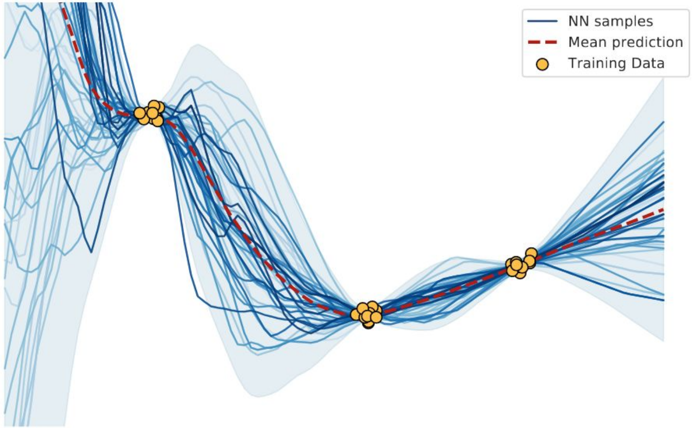
[3] C. Williams, "Computing with Infinite Networks", NIPS, 1996.

On the importance of the weight prior

- The infinite-width limit makes strong assumptions about the weight prior!
- But the induced prior in function space determines the OOD capabilities



width-aware prior: $p(\boldsymbol{w}) = \mathcal{N}\left(\mathbf{0}, \frac{1}{100} I\right)$



standard prior: $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, I)$

2-layer ReLU network (width: 100)

Summary & Conclusions

- The expected advantage of BNNs for OOD detection is not reflected in the experience researchers made in the past few years
  - We argue that this cannot be solely explained by the use of **approximate inference**
  - Instead, we hypothesize that the **function space priors** induced by common **architectures** and/or **weight priors** are not suitable for OOD detection
- Our paper provides **insights** into this problematic and discusses **possible future avenues** to enhance the OOD capabilities of BNNs
- To "**know what you don't know**" should be a requirement when deploying AI, which calls for a thorough understanding under which conditions the use of BNNs for OOD detection is justified

Thank you