Deep Kernel Posterior Learning under Infinite Variance Prior Weights

Jorge Loría

Department of Computer Science,
Aalto University

joint work with

Anindya Bhadra

Bayesian Neural Networks with Finite Variance Converge to GPs

- A Bayesian neural network (BNN) with priors with finite variance converges to a GP

$$y = f(x) = M^{-1/2} \sum_{j=1}^{M} w_j \psi(x) \xrightarrow{d} GP(0, K).$$

- Generalized to several layers (Lee et al., 2018, ICLR; Garriga-Alonso, et al., 2019, ICLR), with explicit formulas for the kernel.

- Limitation: deterministic kernel, since it follows from CLT. There is no possibility of feature learning.

What occurs with infinite variance?

- First suggested by Neal (1996).
- Der and Lee (2005, NeurIPS) proved the result and obtained the characteristic function.
- First computational method by Loría and Bhadra (2024, UAI) for posterior inference. Limitations: high computational cost ($\mathcal{O}(n^l)$) and only for a single layer.

Kernel learning with infinite variance

- Use α-stable random vectors with characteristic function

$$\phi_{\mathbf{W}}(\mathbf{t}) = \exp(-|\mathbf{t}\Sigma\mathbf{t}^T|^{\alpha/2}).$$

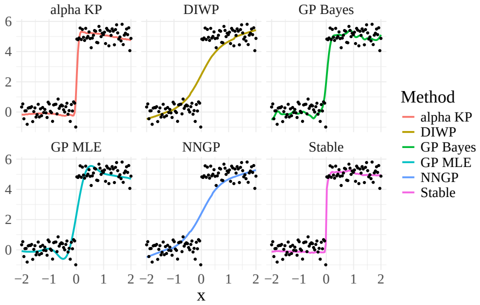
- Decomposed as a Gaussian mixture $\mathbf{W} \overset{d}{=} S^{1/2}\mathbf{G}$, with S is positive $\alpha/2$-stable, and $\mathbf{G} \sim \mathcal{N}(0, \Sigma)$.

- Benefit: exploit the kernel trick to obtain a deep version, with conditionally-known variance but marginally infinite variance.

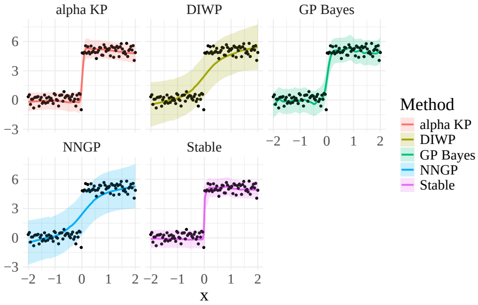
Contributions

- Fast method for posterior inference in the infinite variance setting, and first extension in this regime to deep kernels.
- Feature learning, via the kernel, which is stochastic.
- Numerical demonstrations on synthetic data and UCI data where our method performs better than other GP and kernel methods.

Results in 1D



(a) Function fit for the different methods.



(b) 90% posterior predictive intervals for the Bayesian methods.

Figure: Function fit and uncertainty quantification for the competing methods for a 1-d function with a single jump.

Performance in UCI datasets

Table: Out-of-sample errors in 20 splits. Stable method not available for $l > 2$. Best in **bold**.

Method	Boston ($n = 506, l = 13$) RMSE (SD)	Boston MAE (SD)	Energy ($n = 769, l = 8$) RMSE (SD)	Energy MAE (SD)	Yacht ($n = 308, l = 6$) RMSE (SD)	Yacht MAE (SD)
Dα-KP	2.59 (0.73)	1.78 (0.35)	**0.46** (0.07)	**0.32** (0.05)	**0.31** (0.12)	**0.16** (0.05)
DIWP	2.85 (0.89)	2.01 (0.41)	0.48 (0.06)	0.34 (0.04)	0.60 (0.20)	0.30 (0.10)
NNGP	3.00 (0.87)	2.04 (0.43)	2.18 (0.23)	1.57 (0.18)	3.88 (0.87)	2.58 (0.49)
GP Bayes	**2.58** (0.75)	**1.76** (0.35)	0.68 (0.06)	0.51 (0.04)	0.48 (0.23)	0.22 (0.07)
GP MLE	3.93 (1.02)	2.58 (0.51)	0.49 (0.06)	0.34 (0.04)	0.52 (0.34)	0.25 (0.11)
Stable	–	–	–	–	–	–

Concluding remarks

- Prediction through the latent GP representation is almost as easy as with a GP, with a better performance.
- Neural tangent kernel (NTK) (Jacot, et al., 2018, NeurIPS) uses SGD noise to obtain another GP regime. Non-Gaussian SGD noise (Simsekli, et al., 2019, ICML) could provide analogous non-GP and stable regimes for the NTK.

Main reference:

- **Loría, J**. and Bhadra, A. (2025). Deep Kernel Posterior Learning under Infinite Variance Prior Weights.

Contact information

- jorge.loria@aalto.fi
- loriaj.github.io