Deep Kernel Posterior Learning under Infinite Variance Prior Weights

Anindya Bhadra
www.stat.purdue.edu/∼bhadra

Purdue University

Overview

- Neal's (1996) foundational PhD thesis established the infinitely wide limit of a shallow Bayesian neural network is a Gaussian process (GP) under finite variance prior weights.
- Proof is an application of the classical central limit theorem (CLT).
- This talk considers infinite variance priors (examples: Cauchy, horseshoe, ...). Classical CLT breaks. Limit is not a GP.
- Questions/Goals: (1) Is this unbounded variance regime interesting? (2) If yes, provide an approach for posterior inference and UQ.
- *Joint work with Jorge Loría (Aalto University).*

Wide limit of a shallow (one hidden layer) BNN

- Define an L layer feedforward deep neural network (DNN) with $L-1$ hidden layers by the recursion:

$$f_j^{(\ell+1)}(\mathbf{x}) = g\left(b_j^{(\ell)} + \sum_{i=1}^{M_\ell} w_{ij}^{(\ell)} f_i^{(\ell)}(\mathbf{x})\right),$$
$$\psi(\mathbf{x}) = \sum_{j=1}^{M_L} w_j^{(L)} f_j^{(L)}(\mathbf{x}),$$

  where $g(\cdot)$ is a nonlinear activation and M_ℓ is width of the ℓ-th layer.

- Consider a shallow net ($L=2$) and let $w_j^{(2)} \overset{ind}{\sim} \mathcal{N}(0, 1/M_2)$,

- By "self-similarity" of normal (or more generally, by classical CLT), Neal (1996) established: the limit is a GP as $M_2 \to \infty$.

Covariance of Neal's GP: kernel methods and deep BNNs

- The kernel of Neal's shallow GP depends on the activation $g(\cdot)$.
- Neal worked out two of these explicitly in his PhD thesis:
  - $g(x) = \mathbf{1}(x > 0)$ leads to exponential (Matérn with $\nu = 1/2$) kernel (very rough GP).
  - $g(x) = \tanh(x)$ leads to squared exponential (Matérn with $\nu \to \infty$) kernel (infinitely smooth GP).
- Cho and Saul (2009, NeurIPS) derived using the kernel trick a recursive kernel formula for deep GPs under ReLU and other activations, of the form: $g_\delta(x) = x^\delta \mathbf{1}(x > 0)$.

Implications and extensions of Neal's (1996) result

- Neal's (1996) result is attractive, because posterior inference and uncertainty quantification are straightforward for a GP, unlike a finite-width BNN.

- Using Cho and Saul (2009), extensions to deep feedforward BNNs are by Lee et al. (2018, ICLR), de G. Matthews et al. (2018, ICLR).

- Also using Cho and Saul (2009), extensions to deep convolutional BNNs are by Garriga-Alonso et al. (2018, ICLR).

- In fact, the "Tensor Program" framework of Yang (2019, NeurIPS) establishes a GP limit under nearly arbitrary architectures.

- All of the above assume finite variance priors.

Infinite variance priors

- The main idea in all of the above is this: if $w_j^{(L)} \overset{ind}{\sim} \mathcal{N}(0, 1/M_L)$, then:

$$\psi(\mathbf{x}) = \sum_{j=1}^{M_L} w_j^{(L)} f_j^{(L)}(\mathbf{x}) \xrightarrow{D} \mathcal{N}(0, \mathbb{E}[f(\mathbf{x})f'(\mathbf{x})]).$$

- If $w_j^{(2)}$ has unbounded variance, the generalized CLT (Gnedenko and Kolmogorov, 1953) establishes an α-stable scaling limit under relatively mild conditions.

- Der and Lee (2005, NeurIPS) used the GCLT to work out an α-stable wide limit for shallow BNNs.

Difficulties with α-stable limits

- Unfortunately, it is much harder to work with α-stable variables/processes for inferential purposes.

- Mean and covariance functions are in general not available.

- Define $\mathbf{Z} \sim S_\alpha(\Sigma)$ an symmetric α-stable with scale matrix $\boldsymbol{\Sigma}$; all we have is the characteristic function:

$$\phi_{\mathbf{Z}}(t) = \mathbb{E}[\exp(\mathrm{i}\mathbf{t}^T\mathbf{Z})] = \exp\{-(\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t})^{\alpha/2}\},$$

  not a closed form density.

- Some recent attempts: Favaro et al. (2023, Bernoulli), Peluchetti et al. (2020, AISTATS), Lee et al. (2023, JMLR) etc.

- Mostly concerned with the properties of the limiting process, rather than posterior inference, analogous to *kriging* in GP.

A conditional GP representation

- Our essential idea is to write the (marginal) stable process as a (conditional) GP.

- We can exploit West (1987):

$$\mathbf{Z} \sim S_\alpha(\Sigma) \Longleftrightarrow \mathbf{Z} \overset{D}{=} S_+^{1/2}\mathbf{X}, \quad S_+ \sim S^+_{\alpha/2},\ \mathbf{X} \sim \mathcal{N}(0, \mathbf{\Sigma}),\ S_+ \perp \mathbf{X},$$

  where $S^+_{\alpha/2}$ is positive $\alpha/2$-stable.

- Instead doing inference based on $\mathbf{Z}$, do inference on the augmented space ($\mathbf{X}$, S_+), where ($\mathbf{X} \mid S_+$) is a GP with random covariance $S_+\mathbf{\Sigma}$.

A conditional GP representation

- Define: $z_j^{(\ell)}(\mathbf{x}_k) = \frac{1}{M_\ell^{1/2}} \sum_{i=1}^{M_\ell} w_{ij}^{(\ell)} f_i^{(\ell)}(\mathbf{x}_k)$, with $w_{ij}^{(\ell)} = (s_+^{(\ell)})^{1/2} \tilde{w}_{ij}^{(\ell)}$, where $s_+^{(\ell)} \sim S_{\alpha/2}^+$ and $\tilde{w}_{ij}^{(\ell)}$ are (wlog.) zero mean, unit variance.

- Then, marginally $w_{ij}^{(\ell)}$ have infinite variance and $z_j^{(\ell)}$ is α-stable as $M_\ell \to \infty$.

- The marginally stable $z_j^{(\ell)}$ admits the representation:

$$\mathbf{z}_j^{(\ell)} \mid s_+^{(\ell)}, \Sigma^{(\ell)} \sim \mathcal{N}(0, s_+^{(\ell)} \Sigma^{(\ell)}),$$

where $s_+^{(\ell)} \sim S_{\alpha/2}^+$.

Covariance kernel of the conditional GP

- Leads to a Cho and Saul (2009) type recursive expression for the *conditional* covariance kernels (Prop. 1, Loría and Bhadra, 2024+):

$$\Sigma^{(\ell)}_{k,h} = \pi^{-1}\left[\left(1 + s_+^{(\ell-1)}\Sigma^{(\ell-1)}_{k,k}\right)\left(1 + s_+^{(\ell-1)}\Sigma^{(\ell-1)}_{h,h}\right)\right]^{\delta/2} J_\delta(\theta^{(\ell)}_{k,h}),$$

$$\theta^{(\ell)}_{k,h} = \cos^{-1}\left\{\left[1 + s_+^{(\ell-1)}\Sigma^{(\ell-1)}_{k,h}\right]\left[1 + s_+^{(\ell-1)}\Sigma^{(\ell-1)}_{k,k}\right]^{-1/2}\left[1 + s_+^{(\ell-1)}\Sigma^{(\ell-1)}_{h,h}\right]^{-1/2}\right\},$$

  where all the $s_+^{(\ell)}$ are independent $S^+_{\alpha/2}$ random variables.

- Sanity check: The $\alpha \to 2$ limit is Gaussian. In this case, $S^+_{\alpha/2} \to 1$ w.p. 1, and one recovers the Cho and Saul result, with deterministic kernels $\Sigma^{(\ell)}$.

- But for $\alpha < 2$, the covariance kernel $s_+^{(\ell)}\Sigma^{(\ell)}$ is random, although it is positive definite w.p. 1.

Implications of a random kernel on feature learning

- The deterministic kernel under a GP limit can be thought of a degenerate random kernel, that puts all its prior mass on one point.
- Since posterior $\propto$ likelihood $\times$ prior, the posterior is also a degenerate point mass, at the same point.
- A data-dependent learning of the kernel posterior is thus not possible in a GP limit (Aitchison et al., ICML, 2020, 2021).
- However, the posterior of the random kernel $s_+^{(\ell)}\Sigma^{(\ell)}$ is non-degenerate, when $\alpha < 2$. Data-dependent learning is possible.

Posterior inference and prediction

- Suppose one observes $(\mathbf{y}, \mathbf{x}) = \{y_k, \mathbf{x}_k\}_{k=1}^n$ from the model:

$$y_k = \psi(\mathbf{x}_k) + \varepsilon_k, \ \varepsilon_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2),$$

  and goal is to find the posterior predictive of $(\mathbf{y}^* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}^*)$.

- We have (Prop. 2, Loría and Bhadra, 2024+):

$$\mathbf{y}^* \mid \mathbf{y}, \mathbf{x}, \mathbf{x}^*, \{s_+^{(\ell)}\}_{\ell=2}^L \sim \mathcal{N}_m(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*),$$

- A kriging-like result, except $\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*$ depend on the stable $s_+^{(\ell)}$.

- We propose from the prior of $s_+^{(\ell)} \sim S_{\alpha/2}^+$ to implement an independent samples Metropolis, so the above likelihood is what needs to be evaluated in MCMC.

Results: conditional mutual information

- In GP, one would look at decay of correlation over distance.

- But covariance does not exist for stable processes. Need an alternative.

- Cover and Thomas define conditional mutual information (CMI) as:

$$I(Y_1; Y_2 \mid S) = \int_{\mathcal{S}} D_{KL}[p(Y_1, Y_2 \mid s) \mid\mid p(Y_1 \mid s)p(Y_2 \mid s)]p(s)ds,$$

- Which in our case becomes:

$$I(Y_1; Y_2 \mid S) = -(1/2) \int_{\mathcal{S}} \log(1 - \rho^2_{Y_1, Y_2}(s))p(s)ds.$$

Results: conditional mutual information

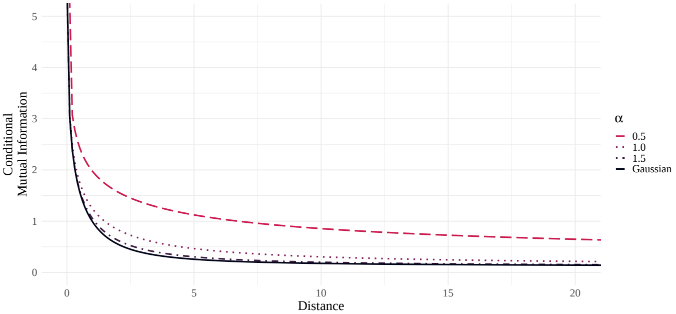


Figure: Decay of the conditional mutual information for the deep α-kernel process as a function of the distance between the inputs with $L = 2, \delta = 1$. The limiting Gaussian case ($\alpha = 2$) is also included.

Results: function fit and UQ

- The true function in 1-d is:

$$f(\xi) = 5 \times \mathbf{1}_{\{\xi>0\}}$$

and generate observations as $y(\xi) = f(\xi) + \varepsilon;\ \varepsilon \sim \mathcal{N}(0, 0.5^2)$.

- The true function in 2-d is:

$$f(\xi_1, \xi_2) = 5 \times \mathbf{1}_{\{\xi_1>0\}} + 5 \times \mathbf{1}_{\{\xi_1>0\}}$$

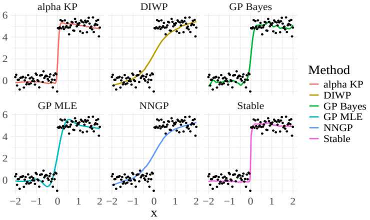
and generate $y(\xi_1, \xi_2) = f(\xi_1, \xi_2) + \varepsilon;\ \varepsilon \sim \mathcal{N}(0, 0.5^2)$,

- The true function in 10-d is:

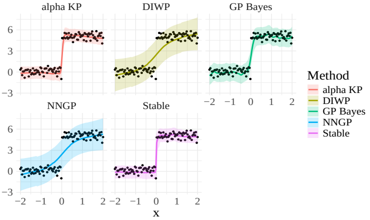
$$\begin{aligned} f(\boldsymbol{\xi}) =& 6\mathrm{sign}(\xi_1) + 8\mathrm{sign}(\xi_2 + \xi_3) + 6\mathrm{sign}(\xi_4 + \xi_5) + 6\mathrm{sign}(\xi_6 + \xi_7) \\ &+ 6\mathrm{sign}(\xi_8 + \xi_9) + 6\mathrm{sign}(\xi_{10}) \end{aligned}$$

and generate $y(\boldsymbol{\xi}) = f(\boldsymbol{\xi}) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, 0.5^2)$.

Results: visualizations of function fit and UQ in 1-d



(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.

- GP tends to oversmooth the jump discontinuity. Stable captures jumps better.
- Agapiou and Castillo (2024, AoS) give theoretical support for this behavior.

Results: effect of α in 1-d

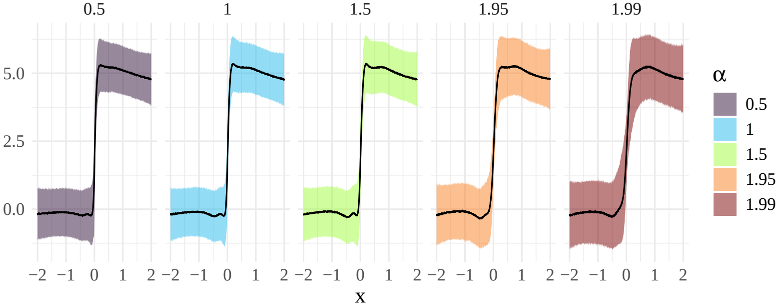


Figure: Comparison of predictions (solid lines) and 25th to 75th percentile posterior predictive intervals (shaded regions) in one dimension for different values of α for the DαKP.

Results: Prediction RMSE and MAE in 1,2 and 10-d

Table: Out-of-sample errors in numerical examples, in twenty different splits. Best in **bold**. Stable not available for more than 2 dimensions.

	One Dimension		Two Dimensions		Ten Dimensions	
Method	RMSE (SD)	MAE (SD)	RMSE (SD)	MAE (SD)	RMSE (SD)	MAE (SD)
Dα-KP	0.57 (0.05)	0.45 (0.04)	0.86 (0.09)	0.67 (0.07)	**8.08** (0.38)	**6.48** (0.33)
DIWP	1.08 (0.04)	0.84 (0.03)	1.69 (0.05)	1.36 (0.05)	8.92 (0.38)	7.21 (0.32)
GP Bayes	0.69 (0.05)	0.52 (0.04)	0.90 (0.06)	0.70 (0.06)	10.39 (0.63)	8.32 (0.52)
GP MLE	0.77 (0.06)	0.57 (0.04)	1.19 (0.08)	0.92 (0.07)	8.32 (0.38)	6.68 (0.34)
NNGP	1.08 (0.04)	0.84 (0.03)	1.69 (0.05)	1.36 (0.04)	8.92 (0.39)	7.21 (0.34)
Stable	**0.52** (0.03)	**0.42** (0.03)	**0.57** (0.08)	**0.45** (0.04)	–	–

Additional results

- The paper contains additional results on predictive performance in some benchmark UCI data sets.
- Evidence of non-Gaussian feature learning, timing, MCMC mixing, coverage of the posterior credible intervals are all available.

Concluding remarks

- The conditional GP representation makes inference and prediction almost as easy as GPs.
- However, there are important distinctions with a GP regime in terms of representation learning, and function fit.
- Another GP regime is the neural tangent kernel or NTK (Jacot et al., 2018, NeurIPS), which arises due to Gaussian SGD noise.
- Non-Gaussian SGD noise (Simsekli et al., 2019, ICML) should give rise to analogous non-GP stable regime for the NTK.

Main references

- Loría, J. and **Bhadra, A.** (2024+). Deep Kernel Posterior Learning under Infinite Variance Prior Weights. (*submitted*). [arXiv:2410.01284]

- Loría, J. and **Bhadra, A.** (2024). Posterior Inference on Shallow Infinitely Wide Bayesian Neural Networks under Weights with Unbounded Variance. *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024)*, PMLR 244, 2331–2349.

- Neal, R. M. (1996). Priors for infinite networks. In *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, pp. 29–53. Springer New York.

- Cho, Y. and Saul, L. (2009). Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*.