An Empirical Analysis of the Advantages of Finite- v.s. Infinite-Width Bayesian Neural Networks

Jiayu Yao, Yaniv Yacoby, Beau Coker, Weiwei Pan, Finale Doshi-Velez

Harvard University

Introduction

Motivation: Recent works in Bayesian deep learning note a counter-intuitive phenomenon—that larger architectures (specifically larger width) can hurt model performance [2, 1]. Despite better performance of smaller NNs, posterior inference in finite-width BNNs is hard whereas, in the infinite-width limit, BNNs converge to a well-behaved model, the Neural Network Gaussian Processes (NNGPs), that admit easy theoretical analysis.

Goal: To provide empirical explanations for the counter-intuitive phenomenon, by comparing the inductive biases and frequency spectra of finite- and infinite-width BNNs.

Contributions: Under certain model misspecifications, compared to infinite-width BNNs, we find that finite-width BNNs:

1. Can generalize better;
2. Can generate more diverse datasets;
3. Define coefficient distributions over the frequency spectrum that are heavier-tailed before inference and more similar to the data after inference;
4. Are better at adapting to model mismatch using different frequency components;

Background

We consider single hidden-layer neural networks of width H:

$$f_{\text{NN}}(\mathbf{x}_i) = \frac{1}{\sqrt{H}}\mathbf{w}_1\phi(\mathbf{w}_0\mathbf{x}_i + \mathbf{b}_0) + b_1.$$

- **Bayesian neural networks (BNNs)** place a prior over neural network weights, which implies a prior $p_{\text{BNN}}(f)$ over neural network functions.
- **Gaussian processes (GPs)** place a prior $p_{\text{GP}}(f)$ directly over functions.

As the width tends to infinity, $p_{\text{BNN}}(f) \to p_{\text{NNGP}}(f)$.

We consider BNNs with **erf** and **ReLU** nonlinearity, whose infinite-width counterparts are referred to as *limiting NNGPs* with **Arcsin** and **Arccos** kernels, respectively.

Generalization Performance of Finite- and Infinite-Width BNNs

Can finite-width BNNs outperform infinite-width BNNs?
We draw datasets from a GP that is different from the limiting NNGP and then compute the average difference in two metrics:

$$\Delta\text{NLL}_{\text{BNN}_H} = \frac{1}{S}\sum_{s=1}^{S}[\text{NLL}_{\text{BNN}_H}(\mathcal{D}^{(s)}) - \text{NLL}_{\text{NNGP}}(\mathcal{D}^{(s)})] \quad (1)$$

$$\Delta\text{MSE}_{\text{BNN}_H} = \frac{1}{S}\sum_{s=1}^{S}[\text{MSE}_{\text{BNN}_H}(\mathcal{D}^{(s)}) - \text{MSE}_{\text{NNGP}}(\mathcal{D}^{(s)})] \quad (2)$$

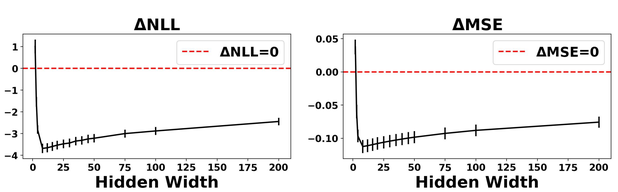

Fig. 1: Test performance of finite-width BNNs: $\mathcal{D}$ ∼GP-RBF($l = 0.5$); BNN with erf, $\sigma_{\mathbf{W}}^2 = \sigma_{\mathbf{b}}^2 = 2.0$

Results: When there is a model mismatch between the data generating model and the limiting NNGP, finite-width BNNs can outperform NNGPs

Quantitatively Comparing Inductive Biases Using the Data Likelihood

Do finite-width BNNs outperform infinite-width BNNs due to their inductive biases (overlap between assumptions of model class and those of the data generating process)?

Commonly, literature uses the Log Marginal Likelihood (LML) that answers "*how likely is the dataset to be generated under the model?*". Unfortunately, the LML is computationally difficult to estimate for BNNs.

We quantify the *overlap* by asking that "*how likely is the data to be generated by the model under the data generating process?*" For datasets sampled from BNNs or limiting NNGPs, we evaluate their LML under the data generative GP, which is easy to compute. We call this the *Log Data Likelihood* (LDL) to avoid confusion:

$$\text{LDL} = -\frac{1}{2}(\mathbf{y} - m(\mathbf{x}))^{\intercal}\Sigma_f^{-1}(\mathbf{y} - m(\mathbf{x})) - \frac{1}{2}\log|\Sigma_f| - \frac{n}{2}\log 2\pi \text{ where } \Sigma_f = K + \sigma_\epsilon^2 I$$

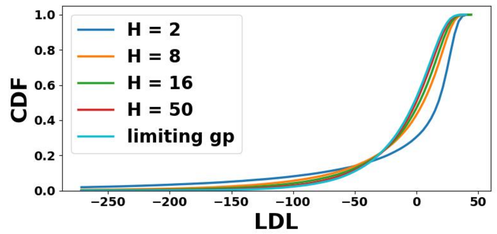

Fig. 2: CDF of LDL with datasets sampled from BNNs or limiting NNGPs and evaluated under GP-RBF($l = 0.5$).

Results: Finite- and infinite-width BNNs have distinct inductive biases. BNNs with larger widths generate less diverse datasets since the cdf of wider BNNs is higher in high LDL region (i.e. they generate limited datasets with high LDL).

Qualitatively Comparing Inductive Biases in Function Space

Do finite-BNNs outperform infinite-width BNNs due to their spectral properties?

Discrete cosine transform (DCT) Given a function $\mathbf{f}^{\intercal} = [f_0, \ldots, f_{N-1}]^{\intercal}$, DCT is a linear and invertible function, $\mathbf{T}_{\text{DCT}}$, that expresses a function as a weighted sum of cosines of different frequencies:

$$\mathbf{a} = \mathbf{T}_{\text{DCT}}\mathbf{f}, \quad \text{where} \quad \mathbf{T}_{\text{DCT}_{ij}} = \frac{\sqrt{2}}{\sqrt{2^{\mathbb{I}[i=0]}N}}\cos\left(\frac{\pi i(2j+1)}{2N}\right) \quad (3)$$

where $\mathbf{a} \in \mathbb{R}^N$ are the DCT coefficients that define the weights on cosines of different frequencies—i.e., a_0 is the weight on the lowest frequency and a_{N-1}, the highest.

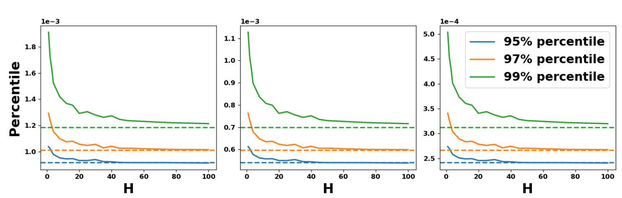

Fig. 3: The distribution of DCT coefficients $\mathbf{a}$ under the prior predictive distribution: solid line—$f \sim p_{BNN}(f)$; dashed line—$f \sim p_{NNGP}(f)$.

Results: finite-width BNNs define a spectrum with more mass on large coefficient values across different frequencies

Low-pass filtering. Given a function $\mathbf{f}$, we first compute its DCT coefficients $\mathbf{a}$, and reconstruct that function via $\mathbf{f} = \mathbf{T}_{\text{DCT}}^{\intercal}\mathbf{a}$. We then remove high-frequency components from the function by setting the corresponding DCT coefficients to 0. More removed frequencies implies smoother draws from the functional prior.

Low-pass filtered BNNs. We filter out high frequencies from the model ("LPF-BNN") by setting $a_i = 0$ for $i > tN$ and training on data from a GP that is not the limiting NNGP (GP-RBF($l = 0.5$)).

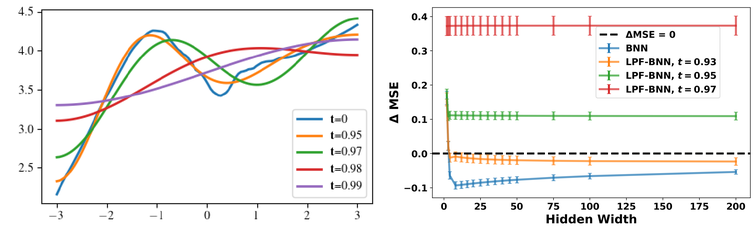

Fig. 4: Left: function draws from LPF-BNNs; Right: test performance of LPF-BNNs.

Results: Removing high-frequency components hurts the generalization performance of finite- width BNNs.

Low-pass filtered datasets. We filter out high frequencies from the data and examine the distribution of the DCT coefficients under the posterior.

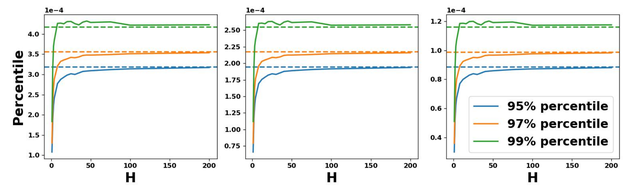

Fig. 5: Test performance of finite-width BNNs on low-pass filtered datasets.

Results: Finite-width BNNs are better at adapting to the designed model mismatch. For finite-width BNNs, the distributions of DCT coefficients corresponding to the high-frequency components put more mass around 0, which resembles the true data generating process ($a_i = 0$ for $i > tN$).

Future Work

- The optimal width given different dataset size.
- Non-Gaussian models (e.g., finite-width BNNs, student's-T processes) or non-Gaussian approximation of the true posterior.

Reference

[1] Jaehoon Lee et al. "Deep neural networks as gaussian processes". In: *arXiv preprint arXiv:1711.00165* (2017).

[2] Geoff Pleiss and John P Cunningham. "The limitations of large width in neural networks: A deep Gaussian process perspective". In: *Advances in Neural Information Processing Systems* 34 (2021).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2007076. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.