Wide Mean-Field Bayesian Neural Networks Ignore the Data

Beau Coker*[1], Wessel P. Bruinsma*[23], David R. Burt*[2], Weiwei Pan[1], Finale Doshi-Velez[1]

Correspondence: beaucoker@g.harvard.edu; *equal contribution; [1]Harvard University; [2]University of Cambridge; [3]Invenia Labs

- The optimal mean-field posterior of a BNN with an odd activation **converges to the prior**.
- With a non-odd activation (e.g., ReLU), the posterior **need not** converge to the prior.

Setup

- Bayesian neural network:

$$\mathbf{f}(\mathbf{x}) = \tfrac{1}{\sqrt{M}}\mathbf{W}_{L+1}\phi(\tfrac{1}{\sqrt{M}}\mathbf{W}_L\phi(\cdots\phi(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)\cdots)+\mathbf{b}_L),$$
$$(\mathbf{W}_i,\mathbf{b}_i)_{i=1}^{L+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

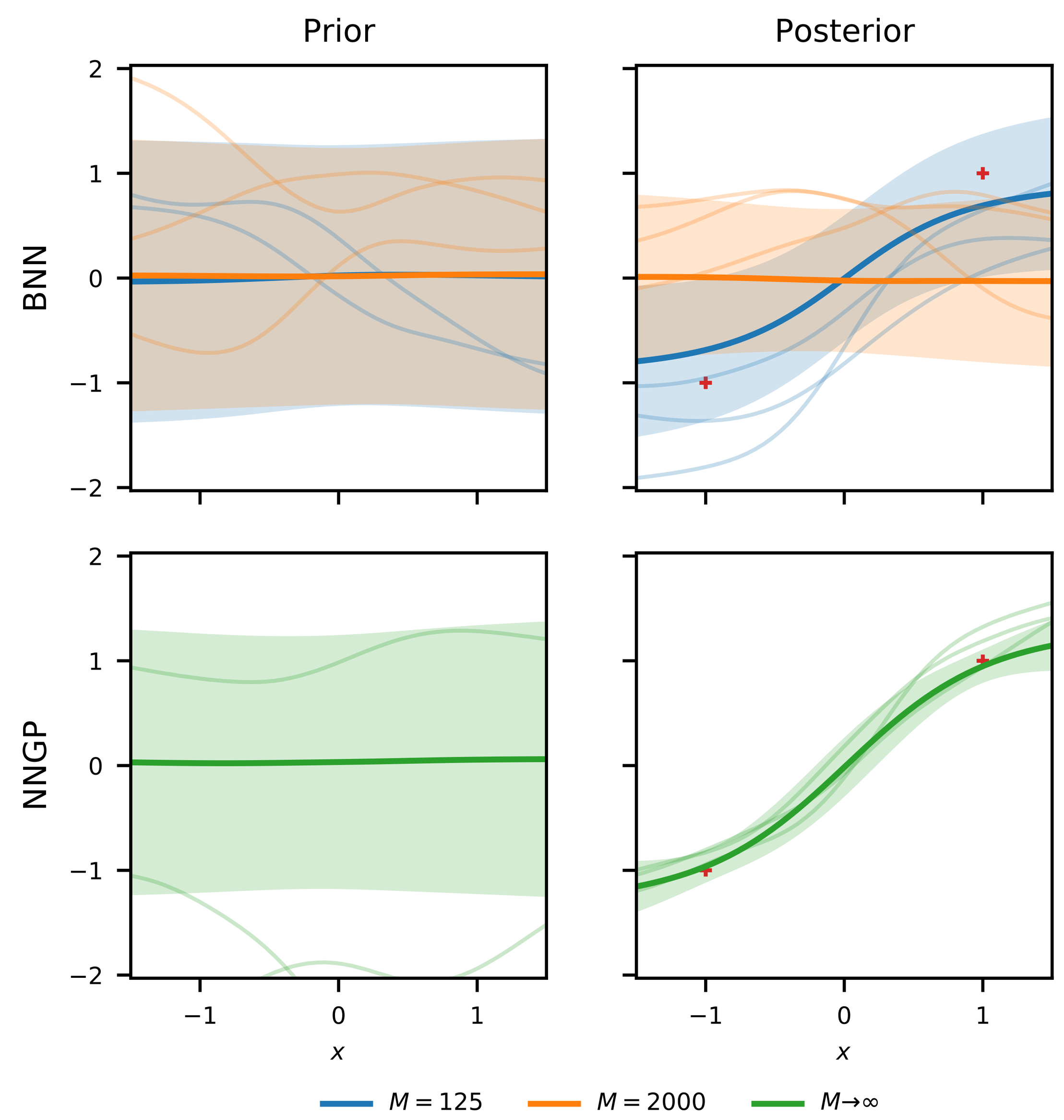
- Variational mean-field inference:

$$Q^* = \operatorname*{arg\,min}_{Q\in\mathcal{Q}_{\text{mean field}}} \mathrm{KL}(Q, P_{|D}) = \operatorname*{arg\,max}_{Q\in\mathcal{Q}_{\text{mean field}}} \mathrm{ELBO}(Q),$$
$$\mathrm{ELBO}(Q) = \mathbb{E}_Q[\log p(\mathbf{y}\,|\,f(\mathbf{X}))] - \mathrm{KL}(Q,P)$$

with $\mathcal{Q}_{\text{mean field}} = \{Q_\theta = \otimes_i Q_{\theta_i}\}$.

What happens with mean-field variational inference in wide networks ($M \to \infty$)?

MFVI with Odd Activations Converges to the Prior



Theorem 1. *For a Gaussian likelihood, the optimal mean field solution Q^* converges to the prior as $M \to \infty$:*

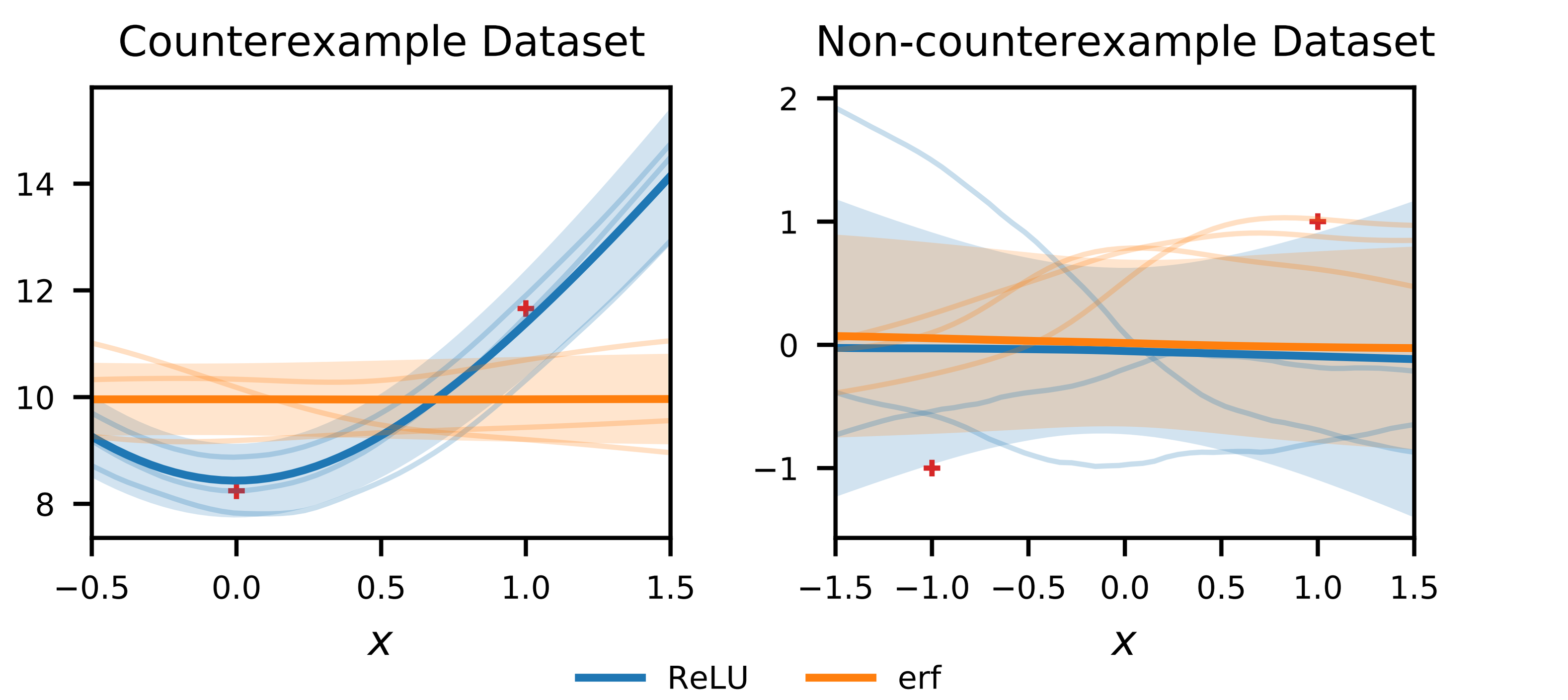
$$Q_f^* \Rightarrow P_f \quad \text{as} \quad M \to \infty.$$

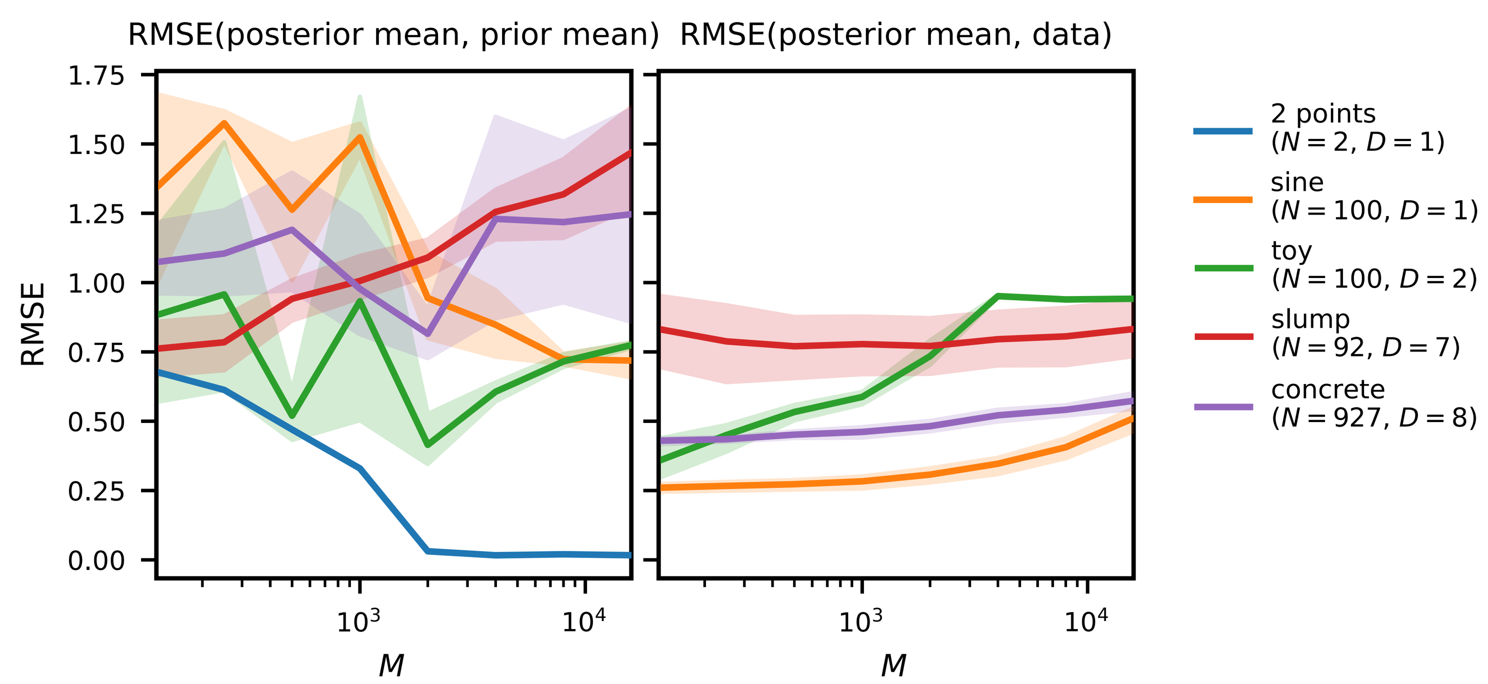
Theorem 2. *There exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that*

$$|\mathbb{E}_Q[f(\mathbf{x})] - \mathbb{E}_P[f(\mathbf{x})]| \le c_1 c_2^{L-1}\frac{1+\frac{1}{\sqrt{D_i}}\|\mathbf{x}\|_2}{\sqrt{M}}\mathrm{KL}(Q,P)\left(\mathrm{KL}(Q,P)^{\frac{L-1}{2}} \vee 1\right),$$
$$|\mathbb{E}_Q[f^2(\mathbf{x})] - \mathbb{E}_P[f^2(\mathbf{x})]| \le c_3 c_4^{L-1}\frac{1+\frac{1}{D_i}\|\mathbf{x}\|_2^2}{\sqrt{M}}\mathrm{KL}(Q,P)^{\frac{1}{2}}\left(\mathrm{KL}(Q,P)^{L+\frac{1}{2}} \vee 1\right).$$

- $\mathrm{KL}(Q^*, P) = O(1)$ for most commonly used likelihoods.

The Case of Non-Odd Activations





- Counterexample: **For non-odd activation functions (like ReLU), MFVI posterior need not converge to prior!**
- Non-counterexample: However, ReLU networks appear to converge to the prior on a different dataset.
- Empirically, across many datasets, we see under-fitting of wide networks with non-odd activations, but not necessarily convergence to the prior.

Discussion

- Should mean-field VI be abandoned for BNNs?

⇒ *We recommend using great care.*

- Does using a ReLU activation solve all of the issues with MFVI?

⇒ *Wide networks still underfit, even if this can't always be attributed to convergence to the prior.*

- Can the dependence of Theorem 2 on depth (L) be improved? In particular, should we expect the optimal MFVI posterior in deeper networks to converge more or less quickly to the prior as width (M) increases?

Links

Paper: https://arxiv.org/abs/2202.11670
Code: https://github.com/dtak/wide-bnns-public