Wide Mean-Field Variational Bayesian Neural Networks Ignore the Data

Beau Coker [1] **Weiwei Pan** [2] **Finale Doshi-Velez** [2]

Abstract

Variational inference enables approximate posterior inference of the highly over-parameterized neural networks that are popular in modern machine learning. Unfortunately, such posteriors are known to exhibit various pathological behaviors. We prove that as the number of hidden units in a single-layer Bayesian neural network tends to infinity, the function-space posterior mean under mean-field variational inference actually converges to zero, completely ignoring the data. This is in contrast to the true posterior, which converges to a Gaussian process. Our work provides insight into the over-regularization of the KL divergence in variational inference.

1. Introduction

Bayesian neural networks (BNNs) provide principled notions of uncertainty in deep learning, but they have not been widely adopted in practice as many properties of this model and its inference remain poorly understood, particularly in the overparameterized, nearly nonparametric regime where modern deep learning take often places. One tool for understanding the behavior of BNNs with a large number of parameters is to take the limit as the number of hidden units (i.e., *width*) goes to infinity. In this case, the prior predictive distribution of a single-layer BNN converges in distribution to the *NNGP*, a Gaussian process with the *neural network kernel* that depends on the prior and architecture of the network (Neal, 1996). Extensions of this result exist also for deep networks (de G. Matthews et al., 2018) and for BNN posteriors (Hron et al., 2020).

In contrast, asymptotic properties of popular variational approximations of BNN posteriors have not been extensively studied. In the finite width regime, variational posterior approximations are known to have deficiencies. For example,

the commonly used mean-field variational posterior, which ignores correlations between parameters, exhibits various pathologies including poor uncertainty estimates between data rich regions (Foong et al., 2019). Moreover, recent works have noted the tendency of mean-field variational BNN posteriors to underfit even with large network architectures (Dusenberry et al., 2020). It is therefore natural to ask if the deficiencies in mean-field variational approximations of finite width BNN posteriors persist in approximations of infinite width BNN posteriors.

In this paper, we show that the answer, unfortunately, is *yes*. For single-layer Bayesian neural networks used for univariate regression, we prove a surprising result: the posterior predictive mean under mean-field variational inference converges to zero (assuming the observed outcomes are centered) as the number of hidden units tends to infinity. That is, the mean-field variational posterior predictive mean completely ignores the data, unlike the true posterior predictive of an infinite width BNN. Furthermore, we provide insight on the cause of this failure — we show that this results from the over-regularizing effect of the KL divergence term (forcing the posterior to match the zero-mean prior) in the variational inference objective.

2. Background

Bayesian neural networks (BNNs) A single-layer feedforward neural network with K hidden units used for univariate regression is given by:

$$f(x, \theta) = \sum_{k=1}^{K} w_k^{(2)} \psi(w_k^{(1)T} x + b_k^{(1)}) \tag{1}$$

where ψ is a nonlinear activation function, $x \in \mathbb{R}^D$ is an input, $w_k^{(1)} \in \mathbb{R}^D$ and $b_k^{(1)} \in \mathbb{R}$ are input-layer weight and bias parameters, respectively, and $w_k^{(2)} \in \mathbb{R}$ is an output-layer weight parameter. For simplicity, we assume the mean has been subtracted from the observed outcomes so that we can omit an output-layer bias parameter. We let $\theta \in \mathbb{R}^{K(D+2)}$ denote the collection of all model parameters. Given independent and identically distributed observations $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, we assume a Gaussian likelihood function $\mathcal{L}(\theta) = \prod_n \mathcal{N}(y_n \mid f(x_n, \theta), \sigma_{\text{noise}}^2)$ and infer the parameters θ by maximizing the likelihood.

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA [2] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. Correspondence to: Beau Coker beaucoker@g.harvard.edu.

A *Bayesian neural network* (BNN) places a prior distribution $p(\theta)$ on the parameters, typically a factorized Gaussian:

$$w_k^{(1)} \sim \mathcal{N}(0, \sigma_{w^{(1)}}^2 I_D) \tag{2}$$

$$b_k^{(1)} \sim \mathcal{N}(0, \sigma_{b^{(1)}}^2) \tag{3}$$

$$w_k^{(2)} \sim \mathcal{N}(0, \sigma_{w^{(2)}}^2), \tag{4}$$

where I_D is a $D \times D$ identity matrix and $k \in \{1, \ldots, K\}$.

Convergence to Gaussian processes Under this prior, the covariance between the function f evaluated at two inputs x and x' is given by:

$$K \sigma_{w^{(2)}}^2 \mathbb{E}[\psi(w_k^{(1)T} x + b_k^{(1)}) \psi(w_k^{(1)T} x' + b_k^{(1)})]$$

Notice by scaling the output-layer prior variance parameter $\sigma_{w^{(2)}}^2$ inversely with the network width K (i.e., setting $\sigma_{w^{(2)}}^2 = \tilde{\sigma}_{w^{(2)}}^2 / K$ for some $\tilde{\sigma}_{w^{(2)}}^2$), the prior covariance is the same for any width. Then, as the number of hidden units tends to infinity, application of the central limit theorem reveals that for any set of inputs x, the prior predictive distribution over the function output $f(x)$ approaches a multivariate Gaussian with the covariance given above. This the definition of a *Gaussian process* (GP). In this case, it is called the neural network Gaussian process (NNGP), since the covariance function of the multivariate Gaussian distribution over $f(x)$ is induced by the neural network.

Note that while scaling the output-layer prior variance parameter by $1/K$ is useful for analyzing the theoretical properties of a BNN, it is an important practical consideration, too. Otherwise the prior predictive variance could become too large as the width increases.

Variational inference Unfortunately, the posterior distribution of a finite-width BNN is not available in closed-form and Markov chain Monte Carlo (MCMC) methods are too slow for all but the smallest networks. Instead, it is common to find the closest distribution q_ϕ in KL divergence to the posterior by maximizing a lower bound on the marginal likelihood called the evidence lower bound (ELBO):

$$\text{ELBO}(\phi) = \mathbb{E}_{\theta \sim q_\phi}[\log \mathcal{L}(\theta)] - \mathbb{KL}[q_\phi || p(\theta)]. \tag{5}$$

The first term in the ELBO is the expected log likelihood, which measures how well the model fits the data, and the second term is the Kullback-Leibler (KL) divergence regularization, which measure how close q_ϕ is to the prior $p(\theta)$.

The *variational* distribution q_ϕ is parameterized by a set of variational parameters ϕ. A common choice for q_ϕ is a product of independent (i.e., *mean-field*) Gaussian distributions, one distribution for each parameter θ_i in the model:

$$q_\phi(\theta) = \prod_{i=1}^{|\theta|} \mathcal{N}(\theta_i \mid \mu_i, \sigma_i^2). \tag{6}$$

Here, $\phi = \{(\mu_i, \sigma_i^2)\}$ are the variational parameters. We call variational inference using Equation 6 *mean-field variational inference* (MFVI).

Since both the prior and variational distribution are Gaussian, the KL divergence can be calculated in closed-form. For example, the KL divergence between the variational distribution of all K output-layer weights $w^{(2)}$ and a $\mathcal{N}(0, \frac{1}{K} I_K)$ prior is

$$\frac{1}{2} \sum_{k=1}^{K} \left[K \mu_k^2 + K \sigma_k^2 - 1 - \log K \sigma_k^2 \right], \tag{7}$$

where $\{\mu_k\}$ and $\{\sigma_k^2\}$ are the variational parameters. Notice Equation 7 acts like L2 regularization of the mean parameters, which will be key to our proof of Theorem 1.
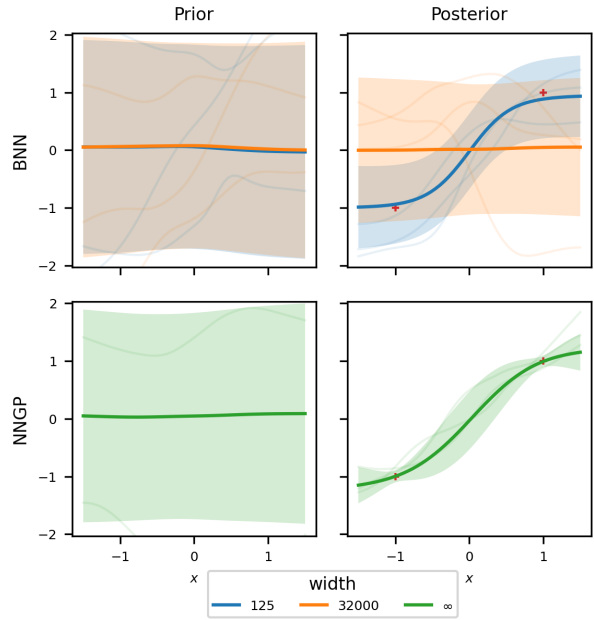

Figure 1. Prior and posterior predictive distributions for mean-field variational BNNs of different widths compared to the NNGP, to which the true posterior of the BNN converges. For a large width, the mean-field variational BNN ignores the data, unlike the NNGP. The shaded regions constitute ± 1 standard deviation around the means (solid lines). All estimates are based on 1000 function samples (a few of which are drawn faintly).

3. Main result

Figure 1 illustrates our main observation. The left column compares the prior predictive distributions of BNNs of various widths (top panel) to the Gaussian process prior to which these BNNs converge (i.e., the NNGP, bottom panel). As expected from the scaling in the prior, the prior predictive mean and variance is the same between all finite and infinite width models. However, the right column reveals significant

differences in the posterior predictive distributions. While the NNGP posterior exhibits a posterior mean that models the data and a posterior variance that expands outside of the data, the MFVI BNN posterior predictives completely ignore the data as the width increases. We use the implementation provided by (Novak et al., 2019) to compute the NNGP.

Like other BNN pathologies observed in practice, it is not immediately obvious whether the behavior is due to the variational objective itself or poor optimization of the ELBO. However, in Theorem 1, we show that the underfitting observed in wide BNNs is due to the choice of the variational objective itself (i.e., the combination of the choice of the variational family, the prior and the divergence measure in variational inference). Specifically, we prove that the MFVI posterior mean converges to zero as the width approaches infinity. Although we assume an error function activation function, we empirically demonstrate that the same conclusions likely hold other activations (we provide results for ReLU and tanh activations in Section 4).

Theorem 1. *Consider a mean-field variational BNN of width K as described by equations 1 and 2-4 and assume an error function activation $\phi = \mathrm{erf}(z) := \int_0^z 2/\sqrt{\pi}\exp(-t^2)\,dt$. Suppose $\hat{\phi}_K$ maximizes the ELBO given in Equation 5. Then for any dataset $\mathcal{D}$ and input x^*, the variational posterior predictive mean $\mathbb{E}_{\theta\sim q_{\hat{\phi}_K}}[f(x^*,\theta)]$ converges to zero as the number of hidden units $K\to\infty$.*

To understand why Theorem 1 holds, think of the single-layer neural network in Equation 1 as a linear model using K nonlinear basis functions $\psi(w_k^{(1)T}x+b_k^{(1)})$. Upon observing data, as the number of basis functions K increases, the prior encourages each of them to be weighted less in the posterior, because of the $1/K$ scaling of the prior variance of the output-layer weights $w^{(2)}$. This tradeoff is what allows the prior predictive variance to converge as $K\to\infty$.

Unfortunately, because of the KL regularization of input-layer parameters $w^{(1)}$ and $b^{(1)}$ in the ELBO, basis functions cannot freely be added to the model without further penalty. Recall that the number of observations is fixed, so there is little improvement to the expected log likelihood term in the ELBO once the neural network has enough hidden units to fit the data. This limits the degree to which the variational distribution over the input-layer parameters can differ from the prior and, in the limit of K, prevents the model from fitting the data.

In contrast, a random features model (e.g., Random Fourier Features (Rahimi & Recht, 2007)), where the input-layer parameters are drawn from a fixed distribution at initialization, has no such issue. Additional basis functions are drawn from the same distribution for all K and can be added to the model without penalty.

The formal proof can be found in the Appendix, but we give a brief sketch here. The proof of Theorem 1 follows in roughly three steps.

- **Step 1: evaluate the ELBO at the prior** Notice the ELBO evaluated at the prior does not depend on the width K. This provides a lower bound on the ELBO that holds for networks of any width.

- **Step 2: bound the KL regularization** Because the ELBO cannot be lower than the bound, the KL divergence cannot be higher than the bound. In particular, this constrains the L2 norm of the optimal variational mean parameters (see Equation 7).

- **Step 3: bound the posterior mean** Using the constraint on variational mean parameters, application a few basic inequalities to the posterior mean reveals convergence to zero.

While Theorem 1 shows that the variational posterior predictive mean converges to the prior predictive mean, we provide empirical evidence that the variational posterior predictive variance converges also to the prior predictive variance. In particular, Figure 6 in the supplementary material shows that the variational posterior variances approaches the prior variances and Figure 2 shows that the variational predictive variance becomes more similar to the prior predictive variance as we increase network width. We thus conjecture that a stronger statement is true: that the mean-field variational posterior distribution converges to the prior. A proof of this is current work. Finally, although Theorem 1 applies to single hidden layer networks, understanding the asymptotic properties of deep networks is also work in progress.
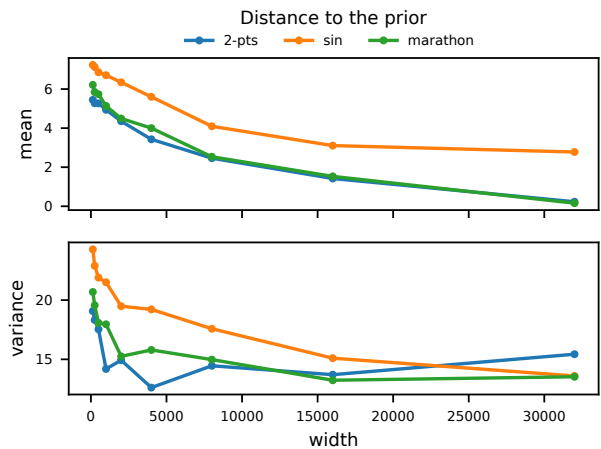

Figure 2. Across datasets, the mean-field variational posterior of a BNN gets closer to the prior as the width increases. *Top:* Euclidean distance between the prior and posterior predictive mean. *Bottom:* Euclidean distance between the prior and posterior predictive variance. We use 50 test points spaced uniformly over a grid centered roughly around the training data.
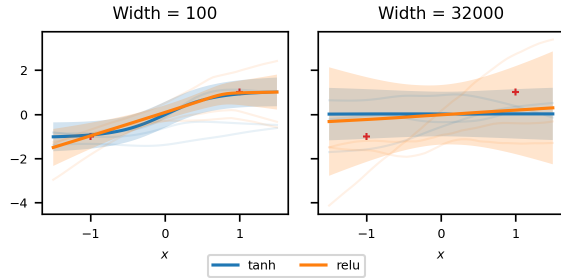
Figure 3. Comparison of the posterior predictives of MFVI BNNs using tanh and ReLU activation functions for small (left panel) and large (right panel) widths. Regardless of the activation, wide MFVI BNNs ignore the data.

4. Experiments

We begin by analyzing the rate of convergence of the posterior predictive mean of the mean-field variational BNN to zero for different datasets. The datasets are the "2-points" dataset shown in Figures 1 and 3, a synthetic dataset of noisy observations of a sine wave ($N = 20$), and a real dataset of winning Olympic marathon paces over the period 1896 to 2012 ($N = 27$).[1] We z-score standardize the inputs and outputs of all datasets.

For each dataset and widths ranging from $K = 125$ to $K = 32,000$, the top and bottom panels, respectively, of Figure 2 show the Euclidean distance of the posterior predictive mean and variance under mean-field variational inference to the prior predictive mean and variance. The top panel illustrates the conclusions of Theorem 1: as the width increases, the posterior predictive mean converges to zero. In the bottom panel panel, we see the posterior predictive variance getting closer to the prior, though it is still far away. We will investigate this behavior in future work.

Next we analyze the impact of different activation functions. Although Theorem 1 assumes an erf activation function, Figure 3 provides empirical evidence that the results of Theorem 1 hold for tanh and ReLU activations as well.

4.1. Implementation details

We use a prior variance of 2 for all parameters (i.e., $\sigma^2_{w^{(1)}} = 2$, $\sigma^2_{b^{(1)}} = 2$, and $\tilde{\sigma}^2_{w^{(2)}} = 2$) and a prior observational noise $\sigma^2_{\text{noise}} = .01$. However, we implement the impact of the prior distribution by scaling the parameters by their prior variance in the forward pass and then using a $\mathcal{N}(0, 1)$ distribution as the prior to evaluate the KL divergence term in the ELBO. That is, for any parameter $\theta_0 \in \theta$, where $p(\theta_0) = \mathcal{N}(0, \sigma^2_{\theta_0})$ is its prior distribution, we replace θ_0 with $\sigma_{\theta_0}\theta_0$ in the evaluation of $f(x, \theta)$ and use a $\mathcal{N}(0, 1)$ as the prior. In particular, this means that instead of scaling the output-layer

prior variance $\sigma^2_{w^{(2)}}$ by $1/K$, we scale the function output by $1/K$ and keep $\sigma^2_{w^{(2)}}$ unscaled. This "neural tangent kernel" scaling yields the same prior predictive distribution while enabling a constant learning rate to be used for training networks of different widths (Lee et al., 2019).

We initialize the variational mean and variance parameters from a normal-inverse-gamma family. Specifically, if $q(\theta_0) = \mathcal{N}(\mu, \sigma^2)$ is the variational distribution corresponding to the parameter θ_0, we randomly initialize $\mu \sim \mathcal{N}(0, 1)$ and $\sigma^2 \sim \mathcal{IG}(\nu + 1, \nu)$. It follows from the laws of total expectation and variance that $\mathbb{E}[\theta_0] = 0$ and $\mathbb{V}[\theta_0] = 2$. The hyperparameter ν controls the concentration of σ^2 around its initial mean of one (i.e., $\mathbb{E}[\sigma^2] = 1$ and $\mathbb{V}[\sigma] = 1/(\nu - 1)$). We set ν to the width of the network. We have experimented with other initializations, including very small variances and mean parameters pretrained to maximize the log likelihood, with little impact on the overall results.

5. Discussion

Trippe & Turner (2017) discusses *over-pruning*, which is the phenomenon whereby many of the variational distributions over the output-layer weights concentrate around zero. This is undesireable behavior because the amount of over-pruning increases with the degree of over-parameterization and because over-pruning degrades performance — simpler models that do not permit pruning often perform better. As in our work, the explanation for over-pruning centers around the tension between the likelihood term and the KL divergence term in the ELBO. To reduce the KL divergence, hidden units can be pruned from the model. Relatedly, we prove that as the number hidden units tends to infinity, the KL divergence over-regulates the model by pulling each of the output-layer weights towards zero while limiting the overall probability mass assigned to the input-layer weights (see earlier discussion in Section 3 regarding the intuition of Theorem 1).

Our work also provides theoretical insight into *cold posteriors*, which is the empirical phenomenon that downweighting the importance of the KL divergence in the ELBO yields better model performance (Wenzel et al., 2020). It is possible this practice serves to undo the over-regularization of the KL divergence that we investigate.

A common theme in both of these pathologies is that the effect becomes more prominent as the width of the network increases. Yet, the phenomenon of *double descent* shows that it is in this over-parameterized regime where neural networks perform best (Belkin et al., 2019). Therefore, it is critical to understand the properties of wide variational BNNs — which we prove are considerably different from the true posterior in the mean-field case — if they are to be adopted in practice.

---

[1]Available in the open-source `pods` Python package `https://github.com/sods/ods`.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116.

de G. Matthews, A. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Dusenberry, M., Jerfel, G., Wen, Y., Ma, Y., Snoek, J., Heller, K., Lakshminarayanan, B., and Tran, D. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning (ICML)*, pp. 2782–2792. PMLR, 2020.

Foong, A. Y. K., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. 'In-between' uncertainty in bayesian neural networks. In *Workshop on Uncertainty and Robustness in Deep Learning (ICML)*, 2019.

Hron, J., Bahri, Y., Novak, R., Pennington, J., and Sohl-Dickstein, J. Exact posterior distributions of wide bayesian neural networks. In *Workshop on Uncertainty and Robustness in Deep Learning (ICML)*, 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Neal, R. *Bayesian Learning for Neural Networks*. Springer Verlag, 1996.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations (ICLR)*, 2019.

Pólya, G. Remarks on computing the probability integral in one and two dimensions. In *Berkeley Symposium on Mathematical Statistics and Probability*, pp. 63–78, 1949.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Trippe, B. L. and Turner, R. E. Overpruning in variational bayesian neural networks. In *Advances in Approximate Bayesian Inference (NeurIPS)*, 2017.

Wenzel, F., Roth, K., Veeling, B. S., Swiatkowski, J., Mandt, L. T. S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning (ICML)*, 2020.

Supplementary material

A. Proofs

A.1. Lemmas

Lemma 1. *Let $z \sim \mathcal{N}(\mu, \sigma^2)$. Then*

$$\mathbb{E}[\mathrm{erf}(z)] = \mathrm{erf}\left(\frac{\mu}{\sqrt{1 + 2\sigma^2}}\right). \tag{8}$$

Proof. First we claim

$$\mathbb{E}[\Phi(z)] = \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right), \tag{9}$$

where $\Phi(z) := \int_{-\infty}^{z} \mathcal{N}(t \mid 0, 1)\, dt$ is the cumulative distribution function of a standard normal distribution. To see this, let $x \sim \mathcal{N}(0, 1)$ and notice $P(x \leq z \mid z = t) = P(x \leq t) = \Phi(t)$. By the law of total probability:

$$P(x \leq z) = \int P(x \leq z \mid z = t) p(z = t)\, dt \tag{10}$$

$$= \int \Phi(t)\, \mathcal{N}(z = t \mid \mu, \sigma^2)\, dt \tag{11}$$

$$= \mathbb{E}[\Phi(z)]. \tag{12}$$

Now, since x and z are independent, notice $x - z \sim \mathcal{N}(-\mu, 1 + \sigma^2)$. Therefore,

$$P(x \leq z) = P(x - z \leq 0) = \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right). \tag{13}$$

Equation 9 follows from Equations 12 and 13.

Noting $\mathrm{erf}(z) = 2\Phi(\sqrt{2}z) - 1$ and applying Equation 9 we have the desired result:

$$\mathbb{E}[\mathrm{erf}(z)] = \mathbb{E}[2\Phi(\sqrt{2}z) - 1] \tag{14}$$

$$= 2\mathbb{E}[\Phi(\sqrt{2}z)] - 1 \tag{15}$$

$$= 2\Phi\left(\frac{\sqrt{2}\mu}{\sqrt{1 + 2\sigma^2}}\right) - 1 \tag{16}$$

$$= 2\left(\frac{1}{2}\mathrm{erf}\left(\frac{\mu}{\sqrt{1 + 2\sigma^2}}\right) + \frac{1}{2}\right) - 1 \tag{17}$$

$$= \mathrm{erf}\left(\frac{\mu}{\sqrt{1 + 2\sigma^2}}\right). \tag{18}$$

$\square$

Lemma 2. *For all $z \in \mathbb{R}$,*

$$\mathrm{erf}(z)^2 \leq 1 - \exp(-\frac{4}{\pi}z^2). \tag{19}$$

Proof. Define

$$G(z) = \int_0^z (2\pi)^{-1/2} \exp\left(-\frac{1}{2}t^2\right)\, dt. \tag{20}$$

For any $z \geq 0$, Pólya (1949) proved the following inequality (see Equation 1.5 in the referenced paper):

$$G(z) \leq \frac{1}{2}\left(1 - \exp\left(-\frac{2}{\pi}z^2\right)\right)^{1/2}. \tag{21}$$

Since $\mathrm{erf}(z) := \int_0^z 2/\sqrt{\pi} \exp(-t^2)\, dt$, the change of variables $s = t/\sqrt{2}$ shows $G(z) = \mathrm{erf}(z/\sqrt{2})/2$. Equation 21 is therefore equivalent to

$$\frac{1}{2}\mathrm{erf}(z/\sqrt{2}) \le \frac{1}{2}\left(1 - \exp\left(-\frac{2}{\pi}z^2\right)\right)^{1/2} \tag{22}$$

$$\mathrm{erf}(z/\sqrt{2})^2 \le 1 - \exp\left(-\frac{2}{\pi}z^2\right) \tag{23}$$

$$\mathrm{erf}(z)^2 \le 1 - \exp\left(-\frac{4}{\pi}z^2\right), \tag{24}$$

where the final inequality comes from evaluating at $z \leftarrow \sqrt{2}z$. Note that so far we have assumed $z \ge 0$, but notice each side of the final inequality is the same for $z < 0$ (i.e., $\mathrm{erf}(-z)^2 = \mathrm{erf}(z)^2$ and $1 - \exp\left(-\frac{4}{\pi}(-z)^2\right) = 1 - \exp\left(-\frac{4}{\pi}z^2\right)$), so the final inequality holds for all $z \in \mathbb{R}$. $\qquad\square$

Lemma 3. *Assume* $\sum_{k=1}^{K} \mu_k^2 \le C_0$, *where* $C_0 \in \mathbb{R}$ *and* $\mu_k \in \mathbb{R}$ *for all* $k \in \{1, \dots, K\}$. *Then, for any constant* $C_1 > 0$,

$$\sum_{k=1}^{K} \exp\left(-C_1\mu_k^2\right) \ge K \exp\left(-\frac{1}{K}C_1 C_0\right). \tag{25}$$

Proof. By assumption,

$$\sum_{k=1}^{K} \mu_k^2 \le C_0 \tag{26}$$

$$-\frac{1}{K}\sum_{k=1}^{K} C_1\mu_k^2 \ge -\frac{1}{K}C_1 C_0 \tag{27}$$

$$\exp\left(-\frac{1}{K}\sum_{k=1}^{K} C_1\mu_k^2\right) \ge \exp\left(-\frac{1}{K}C_1 C_0\right). \tag{28}$$

By Jensen's inequality using the convex function $\exp(\cdot)$,

$$\frac{1}{K}\sum_{k=1}^{K} \exp\left(-C_1\mu_k^2\right) \ge \exp\left(\frac{1}{K}\sum_{k=1}^{K}(-C_1\mu_k^2)\right) \tag{29}$$

$$= \exp\left(-\frac{1}{K}\sum_{k=1}^{K} C_1\mu_k^2\right) \tag{30}$$

$$\ge \exp\left(-\frac{1}{K}C_1 C_0\right). \tag{31}$$

Multiplying each side by K gives the desired result. $\qquad\square$

Lemma 4. *Let* $\sum_{k=1}^{K} a_{km}^2 \le C_m$ *for all* $m \in \{1, \dots, M\}$, *where* $C_m \in \mathbb{R}$ *for all* $m \in \{1, \dots, M\}$ *and* $a_{km} \in \mathbb{R}$ *for all* $m \in \{1, \dots, M\}$ *and all* $k \in \{1, \dots, K\}$. *Then*

$$\sum_{k=1}^{K}\left(\sum_{m=1}^{M} a_{km}\right)^2 \le M\sum_{m=1}^{M} C_m. \tag{32}$$

Proof. Let $k \in \{1, \ldots, K\}$. By the Cauchy–Schwarz inequality:

$$\left(\sum_{m=1}^{M} a_{km} \right)^2 = \left(\sum_{m=1}^{M} 1 \cdot a_{km} \right)^2 \tag{33}$$

$$\leq \left(\sum_{m=1}^{M} 1^2 \right) \left(\sum_{m=1}^{M} a_{km}^2 \right) \tag{34}$$

$$= M \sum_{m=1}^{M} a_{km}^2. \tag{35}$$

Therefore,

$$\sum_{k=1}^{K} \left(\sum_{m=1}^{M} a_{km} \right)^2 \leq \sum_{k=1}^{K} \left(M \sum_{m=1}^{M} a_{km}^2 \right) \tag{36}$$

$$= M \sum_{m=1}^{M} \sum_{k=1}^{K} a_{km}^2 \tag{37}$$

$$\leq M \sum_{m=1}^{M} C_m \tag{38}$$

$$\square$$

A.2. Proof of Theorem 1

Proof. For simplicity, assume the prior variance parameters ($\sigma_{w^{(1)}}^2$, $\sigma_{b^{(1)}}^2$, and $\sigma_{w^{(2)}}^2$) and the observational noise variance parameter σ_{noise}^2 are all equal to 1, but the proof generalizes for any positive values of these parameters.

We abuse notation by letting $w_d^{(1)} \in \mathbb{R}^K$ denote the input-layer weight parameters corresponding to *input dimension d* (i.e., and going to all K hidden units), $w_k^{(1)} \in \mathbb{R}^D$ denote the input-layer weight parameters corresponding to *hidden unit k* (i.e., and coming from all D input dimensions), and $w_{kd}^{(1)} \in \mathbb{R}$ denote the single weight parameter corresponding to both input dimension d and hidden unit k. We also let $w^{(2)}$ and $b^{(2)}$ denote all K output-layer weight and bias parameters, respectively.

Recall we define $\theta = \{(w_k^{(l)}, b_k^{(l)})\}$ as the collection of all model parameters. To avoid complicated subscripts, for any subset $\theta_0 \subset \theta$ of the parameters, define $\phi[\theta_0]$ as the set of variational parameters corresponding to the subset of parameters θ_0. Notice $\phi[\theta_0]$ has twice as many elements as θ_0, since, under the assumed mean-field Gaussian variational distribution, each parameter has a mean and variance variational parameter. So, for example, $\phi[w_d^{(1)}] \in \mathbb{R}^{2K}$ denotes the input-layer variational parameters corresponding to the dth input. Similarly, define $\mu[\theta_0]$ and $\sigma^2[\theta_0]$ as the corresponding variational mean and variance parameters, respectively. We let ϕ denote the set of all variational parameters (i.e., $\phi = \phi[\theta]$).

For any $M \in \{1, 2, \ldots\}$ and any subset $\theta_0 \subset \theta$ of the parameters, define $R_M(\phi[\theta_0])$ as the KL divergence of the variational distribution $q_{\phi[\theta_0]}$ to a $\mathcal{N}(0, \frac{1}{M} I_{|\theta|})$ prior distribution, where $|\theta|$ is the number of elements in θ and $I_{|\theta|}$ is the $|\theta| \times |\theta|$ identity matrix:

$$R_M(\phi[\theta_0]) := \mathbb{KL}\left(q_{\phi[\theta_0]} \,\Big\|\, \mathcal{N}\left(0, \frac{1}{M} I_{|\theta|}\right) \right) \tag{39}$$

$$= \mathbb{KL}\left(\mathcal{N}\left(\mu[\theta_0], \text{diag}(\sigma^2[\theta_0]) \right) \,\Big\|\, \mathcal{N}\left(0, \frac{1}{M} I_{|\theta|}\right) \right) \tag{40}$$

$$= \frac{1}{2} \sum_{i=1}^{|\theta|} \left[M\mu[\theta_{0,i}]^2 + M\sigma^2[\theta_{0,i}] - 1 - \log M\sigma^2[\theta_{0,i}] \right], \tag{41}$$

where $\theta_{0,i}$ denotes the ith element of θ_0.

With this notation, up to an additive constant the negative ELBO for a mean-field variational BNN of width K can be written as:

$$\text{Loss}(\phi) := -\text{ELBO}(\phi) \tag{42}$$

$$= -\mathbb{E}_{\theta \sim q_\phi}[\log \mathcal{L}(\theta)] + \mathbb{KL}[q_\phi || p(\theta)] \tag{43}$$

$$\underbrace{-\frac{1}{2} \sum_{n=1}^{N} \mathbb{E}_{\theta \sim q_\phi} (y_n - f(x_n, \theta))^2}_{:=\text{Error}(\phi)} + \underbrace{\sum_{d=1}^{D} R_1\left(\phi[w_d^{(1)}]\right) + R_1\left(\phi[b^{(1)}]\right) + R_K\left(\phi[w^{(2)}]\right)}_{:=\text{Reg}(\phi)}, \tag{44}$$

where $\text{Error}(\phi)$ and $\text{Reg}(\phi)$, respectively, describe the contribution of fitting the data and the KL regularization to the loss.

For any width $K \in \{1, 2, \dots\}$, let ϕ_K^{prior} be the variational parameters such that the variational distribution $q_{\phi_K^{\text{prior}}}$ is equal to the prior distribution. In other words, the variational parameters where the mean parameters are zero, the input-layer variance parameters are 1, and the output-layer variance parameters are $1/K$. Since $q_{\phi_K^{\text{prior}}}$ is the prior distribution, $\text{Reg}(\phi_K^{\text{prior}}) = 0$.

Step 1: evaluate the ELBO at the prior We will show that the loss evaluated at the prior, $\text{Loss}(\phi_K^{\text{prior}})$, does not depend on K. This will provide a lower bound on the loss evaluated at the optimal parameters, $\text{Loss}(\hat{\phi}_K)$, that holds for any K, which will enable showing the variational distribution needs to stick near the prior.

To see this, first consider the first two moments of the prior predictive, which are easy to compute because the parameters are independent under the prior:

$$\mathbb{E}_{\phi_K^{\text{prior}}}[f(x_n, \theta)] = \mathbb{E}\left[\sum_{k=1}^{K} w_k^{(2)} \psi(w_k^{(1)T} x_n + b_k^{(1)})\right] \tag{45}$$

$$= \sum_{k=1}^{K} \overset{0}{\cancel{\mathbb{E}\left[w_k^{(2)}\right]}} \mathbb{E}\left[\psi(w_k^{(1)T} x_n + b_k^{(1)})\right] \tag{46}$$

$$= 0 \tag{47}$$

and

$$\mathbb{E}_{\phi_K^{\text{prior}}}\left[f(x_n, \theta)^2\right] = \mathbb{V}[f(x_n, \theta)] + \underset{0}{\cancel{\mathbb{E}[f(x_n, \theta)]^2}} \tag{48}$$

$$= \mathbb{V}\left[\sum_{k=1}^{K} w_k^{(2)} \psi(w_k^{(1)T} x_n + b_k^{(1)})\right] \tag{49}$$

$$= \sum_{k=1}^{K} \underbrace{\mathbb{V}\left[w_k^{(2)}\right]}_{1/K} \underbrace{\mathbb{V}\left[\psi(w_k^{(1)T} x_n + b_k^{(1)})\right]}_{:=V(x_n)} \tag{50}$$

$$= K \frac{1}{K} V(x_n) \tag{51}$$

$$= V(x_n), \tag{52}$$

where we define $V(x_n)$ in Equation 50. Since the input-layer parameters of each hidden unit, $w_k^{(1)}$ and $b_k^{(1)}$, have the same distribution under the prior for all hidden units, $V(x_n)$ is the same for all hidden units. Thus we can pull $V(x_n)$ outside of the sum over k. With these two moments computed, we can compute the loss under the prior, which will not depend on K.

$$\text{Loss}(\phi_K^{\text{prior}}) = \text{Error}(\phi_K^{\text{prior}}) + \underbrace{\text{Reg}(\phi_K^{\text{prior}})}_{0} \qquad (53)$$

$$= \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}_{\phi_K^{\text{prior}}}\left[(y_n - f(x_n, \theta))^2\right] \qquad (54)$$

$$= \frac{1}{2}\sum_{n=1}^{N} \mathbb{E}\left[y_n^2 - 2y_n f(x_n, \theta) + f(x_n, \theta)^2\right] \qquad (55)$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(y_n^2 - 2y_n\underbrace{\mathbb{E}\left[f(x_n, \theta)\right]}_{0} + \mathbb{E}\left[f(x_n, \theta)^2\right]\right) \qquad (56)$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(y_n^2 + V(x_n)\right) \qquad (57)$$

$$:= C_X, \qquad (58)$$

where X is the collection of all N training observations. Notice C_X does not depend on K.

Step 2: bound the KL regularization For any width $K = 1, 2, \ldots$, let $[\hat{\mu}_K, \hat{\sigma}_K^2] = \hat{\phi}_K \in \text{argmin}_\phi \text{Loss}(\phi)$ be the variational parameters that minimize the loss. Then the minimum of the loss is bounded above by the loss evaluated at ϕ_K^{prior}, which we showed does not depend on K. In other words,

$$\text{Loss}(\hat{\phi}_K) = \text{Error}(\hat{\phi}_K) + \text{Reg}(\hat{\phi}_K) \qquad (59)$$

$$\leq \text{Loss}(\phi_K^{\text{prior}}) \qquad (60)$$

$$= C_X \qquad (61)$$

To provide further explanation, Equation 60 holds because otherwise $\hat{\phi}_K$ would not be optimal. In other words, a loss of $C_X = \text{Loss}(\phi_K^{\text{prior}})$ could always be achieved for any K by setting $q_{\hat{\phi}_K}$ to the prior (i.e., by setting $\hat{\phi}_K = \phi_K^{\text{prior}}$), so the optimal $\hat{\phi}_K$ cannot achieve a worse loss.

Because C_X does not depend on K, it follows that for any $K \in \{1, 2, \ldots\}$,

$$\text{Error}(\hat{\phi}_K) + \text{Reg}(\hat{\phi}_K) \leq C_X \qquad (62)$$

$$\text{Reg}(\hat{\phi}_K) \leq C_X - \text{Error}(\hat{\phi}_K) \qquad (63)$$

$$\implies \text{Reg}(\hat{\phi}_K) \leq C_X \qquad (64)$$

$$\sum_{d=1}^{D} R_1\left(\hat{\phi}_K[w_d^{(1)}]\right) + R_1\left(\hat{\phi}_K[b^{(1)}]\right) + R_K\left(\hat{\phi}_K[w^{(2)}]\right) \leq C_X \qquad (65)$$

Therefore, the regularization of the optimal variational parameters $\text{Reg}(\hat{\phi}_K)$ is bounded above by C_X. Furthermore, since each of the regularization terms is non-negative, each is less than the bound:

$$R_1\left(\hat{\phi}_K[w_d^{(1)}]\right) \leq C_X, \quad \forall d \in \{1, \ldots, D\} \qquad (66)$$

$$R_1\left(\hat{\phi}_K[b^{(1)}]\right) \leq C_X \qquad (67)$$

$$R_K\left(\hat{\phi}_K[w^{(2)}]\right) \leq C_X \qquad (68)$$

Additionally, since the contribution of the variance parameters to the KL divergence in Equation 41 is non-negative (i.e. since the function $g(\sigma^2) := M\sigma^2 - 1 - \log M\sigma^2 \geq 0$ for all $M > 0$), the squared mean parameters, summed over all

hidden units, are also bounded (i.e., since $a + b \leq c \implies a \leq c$ if $b \geq 0$):

$$\frac{1}{2} \sum_{k=1}^{K} \left(\hat{\mu}_K[w_{kd}^{(1)}] \right)^2 \leq C_X, \quad \forall d \in \{1, \dots, D\} \tag{69}$$

$$\frac{1}{2} \sum_{k=1}^{K} \left(\hat{\mu}_K[b_k^{(1)}] \right)^2 \leq C_X \tag{70}$$

$$\frac{1}{2} \sum_{k=1}^{K} K \left(\hat{\mu}_K[w_k^{(2)}] \right)^2 \leq C_X \tag{71}$$

Step 3: bound the posterior mean Using the bounds on the optimal variational mean parameters in Equations 69, 70, and 71, we show the absolute value of the posterior mean converges to zero.

$$\left| \mathbb{E}_{\theta \sim q_{\hat{\phi}_K}}[f(x^*, \theta)] \right| = \left| \mathbb{E} \left[\sum_{k=1}^{K} w_k^{(2)} \psi(w_k^{(1)T} x^* + b_k^{(1)}) \right] \right| \tag{72}$$

$$= \left| \sum_{k=1}^{K} \mathbb{E} \left[w_k^{(2)} \right] \mathbb{E} \left[\psi(w_k^{(1)T} x^* + b_k^{(1)}) \right] \right| \tag{73}$$

$$\leq \left(\sum_{k=1}^{K} \mathbb{E} \left[w_k^{(2)} \right]^2 \right)^{1/2} \left(\sum_{k=1}^{K} \mathbb{E} \left[\psi(w_k^{(1)T} x^* + b_k^{(1)}) \right]^2 \right)^{1/2} \tag{74}$$

$$\leq \left(\sum_{k=1}^{K} \left(\hat{\mu}_K[w_k^{(2)}] \right)^2 \right)^{1/2} \left(\sum_{k=1}^{K} \mathbb{E} \left[\psi(w_k^{(1)T} x^* + b_k^{(1)}) \right]^2 \right)^{1/2} \tag{75}$$

$$\leq \left(\frac{2C_X}{K} \right)^{1/2} \left(\sum_{k=1}^{K} \mathbb{E} \left[\psi(w_k^{(1)T} x^* + b_k^{(1)}) \right]^2 \right)^{1/2} \tag{76}$$

where Equation 73 follows because we assume a mean-field posterior and Equation 74 follows from the Cauchy-Schwarz inequality and the last equation follows from Equation 71.

To bound the second term in Equation 76, consider the distribution of the pre-activations $z_k := w_k^{(1)T} x^* + b_k^{(1)}$. Define

$$\hat{\mu}_K[z_k] := \sum_{d=1}^{D} \hat{\mu}_K[w_{kd}^{(1)}] x_d^* + \hat{\mu}_K[b_k^{(1)}] \tag{77}$$

$$\hat{\sigma}_K^2[z_k] := \sum_{d=1}^{D} \hat{\sigma}_K^2[w_{kd}^{(1)}] x_d^{*2} + \hat{\sigma}_K^2[b_k^{(1)}]. \tag{78}$$

Then, since each of the parameters is Gaussian distributed and independent under the mean-field variational posterior distribution, $z_k \sim \mathcal{N}(\hat{\mu}_K[z_k], \hat{\sigma}_K^2[z_k])$.

Next, we use Lemma 4 to bound the sum of the squared means of z_k's so that we can later apply Lemma 3. For $d = 1, \dots, D$, define $a_{kd} = \hat{\mu}_K[w_{kd}^{(1)}] x_d^*$ and for $d = D + 1$ define $a_{kd} = \hat{\mu}_K[b_k^{(1)}]$, so that $\hat{\mu}_K[z_k] = \sum_{d=1}^{D+1} a_{kd}$. Notice for any $d = 1, \dots, D$,

$$\sum_{k=1}^{K} a_{kd}^2 = x_d^{*2} \sum_{k=1}^{K} \left(\hat{\mu}_K[w_{kd}^{(1)}] \right)^2 \leq 2 x_d^{*2} C_X \tag{79}$$

by Equation 69 and for $d = D + 1$:

$$\sum_{k=1}^{K} a_{kd}^2 = \sum_{k=1}^{K} \left(\hat{\mu}_K[b_k^{(1)}] \right)^2 \leq 2 C_X \tag{80}$$

by Equation 70. Therefore, by Lemma 4:

$$\sum_{k=1}^{K} \left(\hat{\mu}_K[z_k]\right)^2 = \sum_{k=1}^{K} \left(\sum_{d=1}^{D+1} a_{kd}\right)^2 \leq (D+1)\left(\sum_{d=1}^{D} 2x_d^{*2}C_X + 2C_X\right) \tag{81}$$

$$= 2(D+1)C_X\left(\sum_{d=1}^{D} x_d^{*2} + 1\right) \tag{82}$$

$$:= C_{X,x^*}. \tag{83}$$

We can now put all the results together to bound the second term in Equation 76.

$$\sum_{k=1}^{K} \mathbb{E}_{\hat{\phi}_K}\left[\psi(w_k^{(1)T}x + b_k^{(1)})\right]^2 = \sum_{k=1}^{K} \mathbb{E}\left[\psi(z_k)\right]^2 \tag{84}$$

$$= \sum_{k=1}^{K} \mathrm{erf}\left(\frac{\hat{\mu}_K[z_k]}{\sqrt{1 + 2\hat{\sigma}_K^2[z_k]}}\right)^2 \tag{85}$$

$$\leq \sum_{k=1}^{K} \left(1 - \exp\left(-\frac{4}{\pi}\frac{\hat{\mu}_K[z_k]^2}{1 + 2\hat{\sigma}_K^2[z_k]}\right)\right) \tag{86}$$

$$\leq K - \sum_{k=1}^{K} \exp\left(-\frac{4}{\pi}\frac{\hat{\mu}_K[z_k]^2}{1 + 2\hat{\sigma}_K^2[z_k]}\right) \tag{87}$$

$$\leq K - \sum_{k=1}^{K} \exp\left(-\frac{4}{\pi}\hat{\mu}_K[z_k]^2\right) \tag{88}$$

$$\leq K - K\exp\left(-\frac{1}{K}\frac{4}{\pi}C_{X,x^*}\right) \tag{89}$$

$$= K\left(1 - \exp\left(-\frac{1}{K}\frac{4}{\pi}C_{X,x^*}\right)\right), \tag{90}$$

where Equation 85 follows from Lemma 1, Equation 86 follows from Lemma 2, and Equation 89 follows from Lemma 3 (using the bound in Equation 83). Equation 88 follows because $\forall a, b \in \mathbb{R}$, one can show $\exp(-a^2/(1+b^2)) \geq \exp(-a^2)$.

Now plug into Equation 76:

$$\left|\mathbb{E}_{\theta \sim q_{\hat{\phi}_K}}[f(x^*, \theta)]\right| \leq \left(\frac{2C_X}{K}\right)^{1/2}\left(K\left(1 - \exp\left(-\frac{1}{K}\frac{4}{\pi}C_{X,x^*}\right)\right)\right)^{1/2} \tag{91}$$

$$= (2C_X)^{1/2}\left(1 - \exp\left(-\frac{1}{K}\frac{4}{\pi}C_{X,x^*}\right)\right)^{1/2} \tag{92}$$

$$\to 0 \text{ as } K \to \infty. \tag{93}$$

Lastly, since the absolute value of the posterior mean converges to zero, so does posterior mean without the absolute value.

$$\square$$

A.3. Additional experiments

First we illustrate that regardless of width, the BNN priors in function space are approximately the same and resembles that of the NNGP. Figure 4 shows the prior of BNNs of increasing width (rows) as compared to the NNGP prior (bottom row) based on 1,000 function samples from the prior. The prior settings are the same as in the main paper. The horizontal axis is the 1-dimensional input, x. As expected, all of the prior predictive distributions (left column) have the same mean and variance. However, this is only the first two moments of the distribution. To measure the smoothness of the function-space prior, in the right column we show a histogram of the x locations at which a function sample crosses the line $y = 0$, called

an *upcrossing*. The number of upcrossings is a measure of the inverse lengthscale (i.e., inverse smoothness). All models exhibit approximately the same number and distribution of upcrossings.
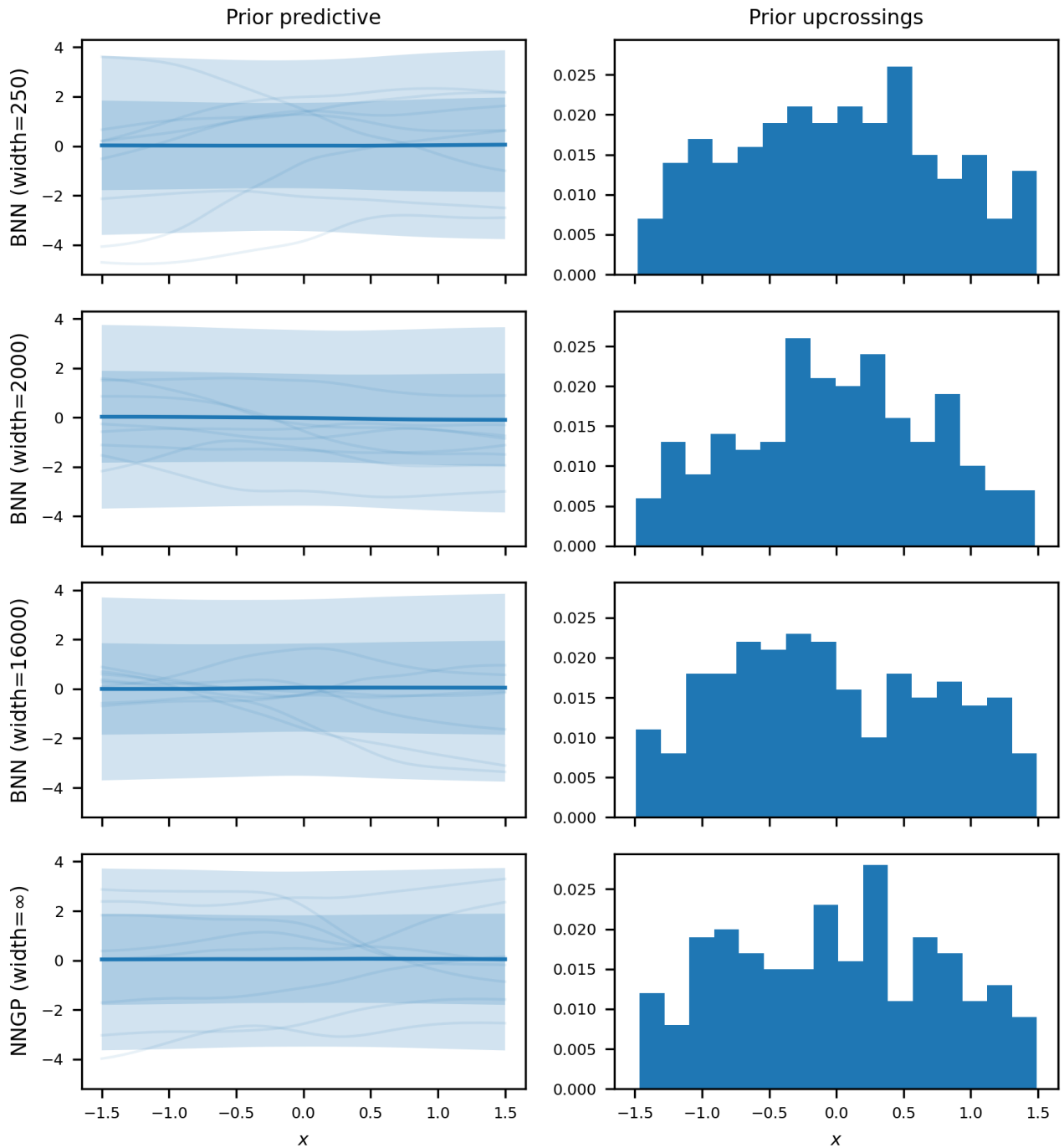

Figure 4. Prior predictive distributions and histogram of upcrossings of $y = 0$ based on 1000 prior function samples. The shaded regions constitute ± 1 (darker shade) and ± 2 (lighter shade) standard deviations around the means (solid lines), with a few samples drawn faintly.

Next, using the priors in 4 we infer the mean-field variational posteriors based on the three datasets (Figure 5). As expected from 1, as the width increases (rows) the mean-field variational posteriors begin to ignore the data as they converge to zero. On the other hand, the NNGP (bottom row) fits the data nicely. Recall that while the true BNN posterior converges to the NNGP posterior as the width K tends to infinity, the mean-field variational posterior does not.
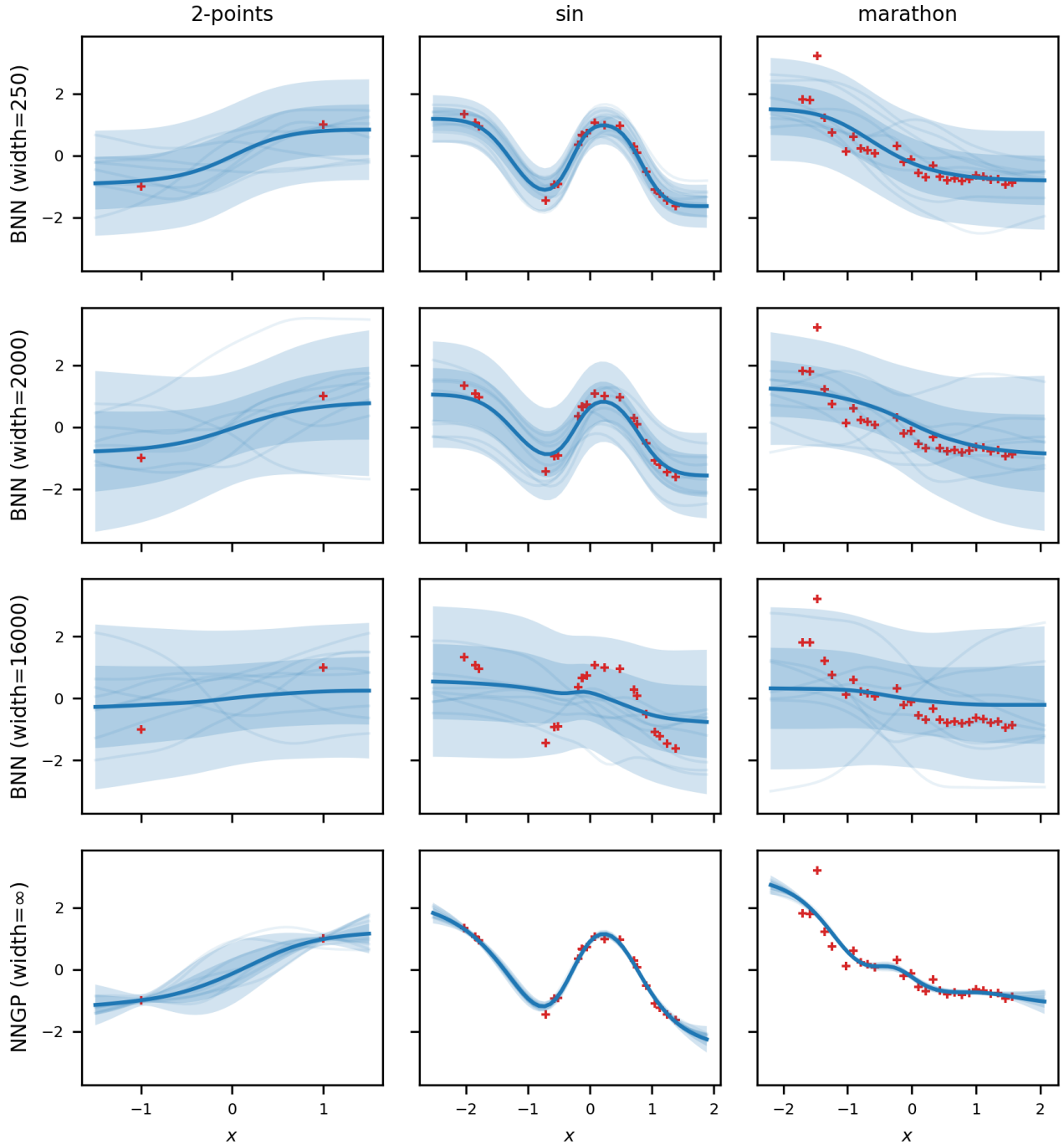
Figure 5. Posterior predictive distributions for mean-field variational BNNs of different widths and trained on different datasets. For a large width, the mean-field variational BNN ignores the data. The shaded regions constitute ±1 (darker shade) and ±2 (lighter shade) standard deviations around the means (solid lines). All estimates are based on 1000 function samples (a few of which are drawn faintly).

Figure 6 shows the distribution of the variational parameters after optimizing the ELBO. We use the "2 points" dataset. To be clear, this is a kernel density estimate of the trained variational parameters themselves across hidden units of the network, not the variational distributions that the variational parameters define. Going from left-to-right, as the network width K increases, the variational parameters move closer to the $\mathcal{N}(0, 1)$ prior. Note that the prior variance of the output-layer weights is still effectively scaled by $1/K$ but in our implementation the scaling is performed in the function evaluation, which enables all parameters to have a $\mathcal{N}(0, 1)$ prior.
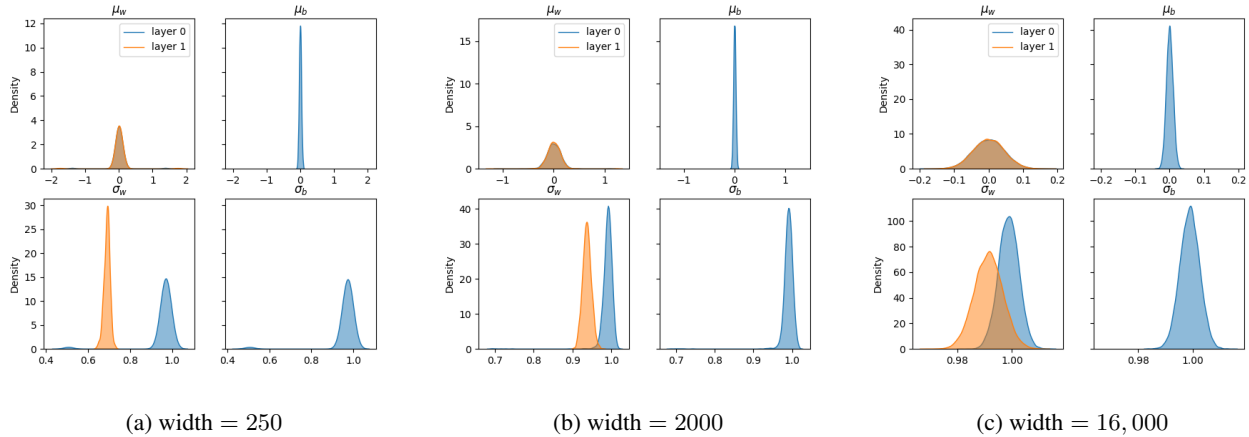
(a) width = 250

(b) width = 2000

(c) width = 16, 000

Figure 6. Kernel density estimates of the distribution of trained variational parameters using the "2-points" dataset. We break it down by layer, parameter type (weight or bias), and variational parameter type (mean or variance). In our implementation, since we scale by the prior parameters in the function evaluation (i.e., we use a "neural tangent kernel" scaling) all variational parameters have a $\mathcal{N}(0, 1)$ prior distribution. Notice the trained variational variance parameters (bottom panels) shift closer the prior value of 1 as the network size increases. For the largest network, all variational parameters are near their prior values.

In all variational inference experiments, we use 20,000 epochs of full-batch gradient descent with a learning rate of 0.001 and a momentum of 0.9 optimization. We use gradient clipping and cosine annealing of the learning rate, with warm restarts every 500 epochs (Loshchilov & Hutter, 2017). To evaluate the ELBO, we use the analytical form of the KL divergence and the reparameterization trick (Kingma & Welling, 2014) with 64 samples to approximate the expected log likelihood term.