Pathologies of Factorised Gaussian and MC Dropout Posteriors in Bayesian Neural Networks

Andrew Y. K. Foong[1*], **David R. Burt**[1*], **Yingzhen Li**[2], **Richard E. Turner**[1, 2]
[1]University of Cambridge, UK [2]Microsoft Research Cambridge, UK
{ykf21,drb62,ret26}@cam.ac.uk, Yingzhen.Li@microsoft.com

Abstract

Applying Bayesian inference to neural networks often requires approximating the posterior over parameters with simple distributions. The quality of the resulting approximate predictive distribution in function space is poorly understood. We prove that for single hidden layer ReLU networks, there exist simple situations where it is impossible for factorised Gaussian or MC dropout posteriors to give well-calibrated uncertainty estimates. Precisely, they cannot both fit the data confidently and have increased uncertainty in between well-separated clusters of data. This motivates more careful consideration of the consequences of approximate inference in Bayesian neural networks.

1 Introduction

In many domains, quantifying uncertainty is critical for the successful application of machine learning methods. For example, in medical applications, calibrated predictive uncertainty is necessary to determine which patients should be referred to an expert for further tests [6]. In reinforcement learning, good uncertainty estimates are important for balancing exploration and exploitation [4]. Bayesian neural networks (BNNs) hold the promise of being powerful function approximators that return reliable uncertainty estimates. However, the need to resort to approximate inference casts doubt on the quality of their predictive uncertainty and limits their practical utility [27].

Many approximate inference methods (e.g. mean-field variational inference (MFVI) [14; 11; 2], probabilistic backpropagation (PBP) [12], Laplace's approximation [5; 20; 24] and Monte Carlo (MC) dropout [10]) assume a specific parametric form for the approximate posterior. We refer to the set of approximating distributions considered by the method as the *approximating family*. For example, in MFVI and the diagonal Laplace approximation, the approximating family is the set of fully factorised Gaussian distributions over the parameters of the network. In MC dropout, the approximating family is defined by multiplying columns of the weight matrices by independent Bernoulli random variables.

As the approximating family is usually chosen for computational expediency, it is often a crude approximation to the exact posterior in parameter space. It is hoped that the resulting predictive distribution in *function space* still has the qualitative features necessary for the task at hand. However, empirically approximate inference in BNNs frequently fails to represent *'in-between' uncertainty*: that is, increased uncertainty in between well-separated clusters of data [7; 27]. A potential consequence of this pathology is that in medical applications, a BNN will be unjustifiably confident between regions contained in the training data, and hence fail to refer ambiguous cases to an expert. In this work, we explain this behaviour by proving fundamental limitations of the factorised Gaussian and MC dropout approximating families.

[*]Equal contribution.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

2 Theoretical Results

Our main results apply to single-hidden layer ReLU BNNs. They show that there are regions of input space where mean-field Gaussian and MC dropout BNNs are incapable of representing in-between uncertainty. More general statements of these theorems can be found in appendix A.

Theorem 1 (Mean-field Gaussian)**.** *Consider a single-hidden layer ReLU neural network mapping from* $\mathbf{x} \in \mathbb{R}^D$ *to* $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^K$ *with an arbitrary number of hidden units. Suppose we have a fully factorised Gaussian distribution over the weights and biases in the network. Consider any points* $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathbb{R}^D$ *such that* $\mathbf{r} \in \overrightarrow{\mathbf{pq}}$ *and either:*

i. The line segment $\overrightarrow{\mathbf{pq}}$ *contains* $\mathbf{0}$ *and* $\mathbf{r}$ *is closer to* $\mathbf{0}$ *than both* $\mathbf{p}$ *and* $\mathbf{q}$*.*

ii. The line segment $\overrightarrow{\mathbf{pq}}$ *is orthogonal to and intersects the plane* $[\mathbf{x}]_d = 0$*, and* $\mathbf{r}$ *is closer to the plane* $[\mathbf{x}]_d = 0$ *than both* $\mathbf{p}$ *and* $\mathbf{q}$*.*

Then for $1 \leq k \leq K$, $\mathrm{Var}[f_k(\mathbf{r})] \leq \mathrm{Var}[f_k(\mathbf{p})] + \mathrm{Var}[f_k(\mathbf{q})]$.

Theorem 1 states that there are line segments in input space such that the predictive variance on the line is bounded in terms of the variance at the endpoints. It illustrates a limitation of the *approximating family* and is agnostic to inference method or optimisation procedure. This family has been used in the diagonal Laplace approximation [5; 24], variational inference (VI) [14; 11; 2], PBP [12], variational Gaussian dropout [16], stochastic expectation propagation [18], black-box alpha divergence minimisation [13], Rényi divergence VI [19], natural gradient VI [15] and functional variational BNNs [25].[2] The bound in theorem 1 applies to all of these techniques, for any setting of the parameters of the distribution, and therefore for any training dataset. In particular, for VI this bound applies for the *global minimiser* of the KL-divergence between the approximate posterior and the true posterior.

Theorem 2 (MC dropout)**.** *Consider the same network architecture as in theorem 1. Suppose we have an MC dropout distribution over the parameters in the network. Then for any finite set of points* $\mathcal{S} \subset \mathbb{R}^D$ *such that* $\mathbf{0}$ *is in the convex hull of* $\mathcal{S}$*,* $\mathrm{Var}[f_k(\mathbf{0})] \leq \max_{\mathbf{s} \in \mathcal{S}} \{\mathrm{Var}[f_k(\mathbf{s})]\}$ *for* $1 \leq k \leq K$.

Theorem 2 upper bounds the predictive variance at the origin by the variance at points surrounding it. It applies to any MC dropout posterior [10; 22] regardless of training dataset, regularisation or optimisation procedure.

3 Illustrative Examples

As the output variance is a measure of epistemic uncertainty, our theorems imply pathological behaviour when the variance bounded should be high (because there is little data there), and the variances determining the upper bound should be low (because there is a lot of data there).

For theorem 1, case i, consider a dataset where the input density $p(\mathbf{x})$ is essentially bimodal, e.g. if there are two distinct populations in the training set, and the training data is centred at the origin (as is standard practice). Then $\mathbf{p}$ and $\mathbf{q}$ could be taken at locations within the clusters, and $\overrightarrow{\mathbf{pq}}$ would intersect the origin. This is illustrated on a synthetic regression dataset[3] in figure 1. Theorem 1, case ii applies when the training inputs share values along all but one input dimension. This would happen if training inputs are chosen by some experimental procedure where input features are varied one at a time. This is illustrated in appendix B. Theorem 2 is relevant to any centred dataset since the origin will be in the convex hull of the datapoints. The variance at the origin can be upper bounded by the maximum over any subset of datapoints, such that the convex hull of these points still contains the origin. This is illustrated in figure 1.

The approximate posterior obtained by MFVI on a single hidden layer ReLU BNN with 50 hidden units (left) is able to represent uncertainty outside of the region containing data, but not in between the two clusters of data. MC dropout (centre) is similarly unable to show in-between uncertainty, as it is more confident at the midpoint of the data clusters than it is at the clusters themselves. In

[2] Not all these methods necessitate the use of fully factorised Gaussians, but it is a common choice.

[3] While we consider the case of regression, our theorems apply equally to the latent function of a classifier (input to the softmax).

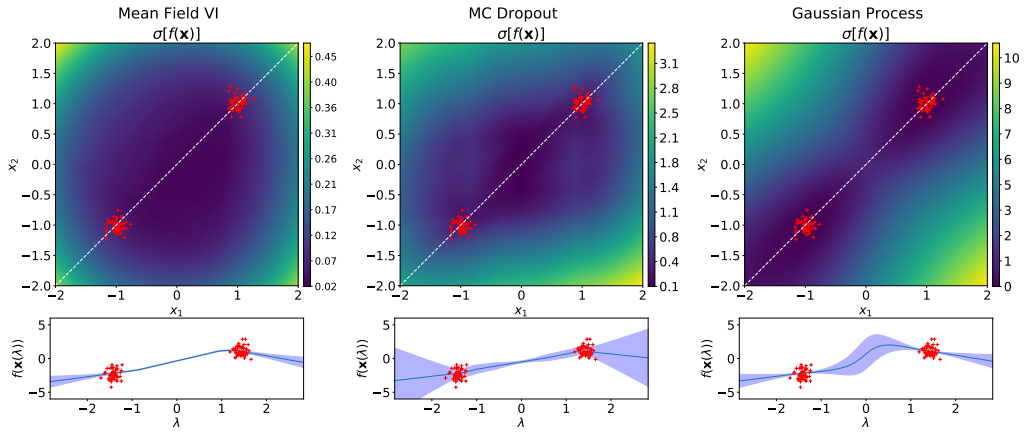

Figure 1: MFVI BNN (left), MC dropout BNN (centre) and GP with a kernel corresponding to the wide limit of the same BNN (right), fit on a regression dataset with 2-dimensional inputs (red crosses). The top plots show the standard deviation of the output in different regions of input space. The bottom plots show the predictive mean with two standard deviation bars along the dashed white line in the top plot. Note MFVI's overconfidence in the region $\lambda \in [-1, 1]$. This behaviour is explained by theorem 1: given the uncertainty is near zero at the data clusters, there is *no* setting of the variational parameters that could have the uncertainty significantly greater than zero in the line segment between them. MC dropout is underconfident at the datapoints and overconfident near the origin. The overconfidence is explained by theorem 2.

contrast, the Gaussian Process (GP) posterior with the equivalent BNN kernel [3] (right) shows increased uncertainty both in between and outside of the observed data. Since GP inference is exact, and the BNN prior approaches the GP prior as the number of hidden units increases [23; 21], we expect the BNN posterior to be qualitatively similar to the GP posterior. The GP posterior shows in-between uncertainty but the BNN approximate posteriors do not, implying that this is a failure of the approximate posterior and not the true BNN posterior.

4 Discussion

The single hidden layer ReLU BNN regression task has been extensively used as a benchmark in the Bayesian deep learning community [12; 18; 19; 26; 15; 25; 10; 22]. Many of these experiments use the mean-field Gaussian and MC dropout approximating families. Since our results indicate all members of these families share a simple pathology, this benchmark is inadequate to evaluate the quality of the inference algorithms. Furthermore, our results demonstrate a case where approximate inference provably leads to extreme overconfidence, even with access to an idealised global optimiser. When designing methods, practitioners should consider whether the approximate posterior is able to represent the type of uncertainty required for the task at hand.

Theorems 1 and 2 only apply to single hidden layer BNNs. In contrast, for 2-hidden layer BNNs there exist mean-field Gaussian parameter distributions that can approximate any continuous predictive mean and variance function (see appendix F). However, [7] observed empirically that MFVI still struggles to represent in-between uncertainty in deeper networks. This illustrates that even though an approximating family contains distributions with desirable properties in function space, this is insufficient to guarantee that an approximate inference method will select those distributions. Theorems 1 and 2 make no assumptions on the method for selecting a distribution from the approximating family. We leave as future work the task of understanding the interaction between the choice of approximating family and the inference algorithm.

Acknowledgements

We would like to thank José Miguel Hernández-Lobato, Sebastian W. Ober and Ross Clarke for helpful discussions. AYKF gratefully acknowledges the Trinity Hall Research Studentship and the George and Lilian Schiff Foundation for funding his studies.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation, 1965.

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[3] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems (NIPS)22*, 2009.

[4] M. Deisenroth and C. E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML)*, pages 465–472, 2011.

[5] J. S. Denker and Y. Lecun. Transforming neural-net output levels to probability distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 1991.

[6] A. Filos, S. Farquhar, A. N. Gomez, T. G. J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, and Y. Gal. Benchmarking Bayesian deep learning with diabetic retinopathy diagnosis. `https://github.com/OATML/bdl-benchmarks`, 2019.

[7] A. Y. K. Foong, Y. Li, J. M. Hernández-Lobato, and R. E. Turner. 'In-between' uncertainty in Bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019.

[8] B. J. Frey and G. E. Hinton. Variational learning in nonlinear Gaussian belief networks. *Neural Computation*, 11(1):193–213, 1999.

[9] Y. Gal. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

[10] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016.

[11] A. Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NIPS) 24*, 2011.

[12] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[13] J. M. Hernández-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner. Black-box alpha divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016.

[14] G. Hinton and D. Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.

[15] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. *Proceedings of The 35th International Conference on Machine Learning (ICML)*, 2018.

[16] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

[17] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[18] Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In *Advances in neural information processing systems*, pages 2323–2331, 2015.

[19] Y. Li and R. E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[20] D. J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[21] A. G. d. G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

[22] J. Mukhoti, P. Stenetorp, and Y. Gal. On the importance of strong baselines in Bayesian deep learning. *arXiv preprint arXiv:1811.09385*, 2018.

[23] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[24] H. Ritter, A. Botev, and D. Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[25] S. Sun, G. Zhang, J. Shi, and R. Grosse. Functional variational Bayesian neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[26] M. B. Tomczak, S. Swaroop, and R. E. Turner. Neural network ensembles and variational inference revisited. In *1st Symposium on Advances in Approximate Bayesian Inference*, pages 1–11, 2018.

[27] J. Yao, W. Pan, S. Ghosh, and F. Doshi-Velez. Quality of uncertainty quantification for Bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

A General Statement of Theorems

In this appendix we provide a more general statement of our theorems. Note that the mean-field Gaussian posterior is a special case of the distribution assumed in theorem 1, and the MC dropout posterior is a special case of that assumed in theorem 2.

Theorem 1 (Mean-field Gaussian)**.** *Consider a single-hidden layer ReLU neural network mapping from $\mathbb{R}^D \rightarrow \mathbb{R}^K$ with $I \in \mathbb{N}$ hidden units. The corresponding mapping is given by $f_k(\mathbf{x}) = \sum_{i=1}^{I} w_{k,i}\phi\left(\sum_{d=1}^{D} u_{i,d}x_d + v_i\right) + b_k$ for $1 \leq k \leq K$, where $\phi(a) = \max(0, a)$. Suppose we have a distribution over network parameters of the form:*

$$q(\mathbf{W}, \mathbf{b}, \mathbf{U}, \mathbf{v}) = \prod_{i=1}^{I} q_i(\mathbf{w}_i|\mathbf{U}, \mathbf{v}) q(\mathbf{b}|\mathbf{U}, \mathbf{v}) \prod_{i=1}^{I}\prod_{d=1}^{D} \mathcal{N}(u_{i,d}; \mu_{u_{i,d}}, \sigma^2_{u_{i,d}}) \prod_{i=1}^{I} \mathcal{N}(v_i; \mu_{v_i}, \sigma^2_{v_i}), \quad (1)$$

where $\mathbf{w}_i = \{w_{k,i}\}_{k=1}^K$ are the weights out of neuron i and $\mathbf{b} = \{b_k\}_{k=1}^K$ are the output biases, and $q_i(\mathbf{w}_i|\mathbf{U}, \mathbf{v})$ and $q(\mathbf{b}|\mathbf{U}, \mathbf{v})$ are arbitrary probability densities with finite first two moments. Consider a line in $\mathbb{R}^D$ parameterized by $[\mathbf{x}(\lambda)]_d = \gamma_d\lambda + c_d$ for $\lambda \in \mathbb{R}$ such that $\gamma_d c_d = 0$ for $1 \leq d \leq D$. Then for any $\lambda_1 \leq 0 \leq \lambda_2$, and any λ_* such that $|\lambda_*| \leq \min(|\lambda_1|, |\lambda_2|)$,

$$\mathrm{Var}[f_k(\mathbf{x}(\lambda_*))] \leq \mathrm{Var}[f_k(\mathbf{x}(\lambda_1))] + \mathrm{Var}[f_k(\mathbf{x}(\lambda_2))] \quad \text{for} \quad 1 \leq k \leq K. \quad (2)$$

In assumption i) in the statement of the theorem in the main body, $c_d = 0$ for $1 \leq d \leq D$. In assumption ii), $\gamma_{d'} = 0$ for $d' \neq d$, and $c_d = 0$.

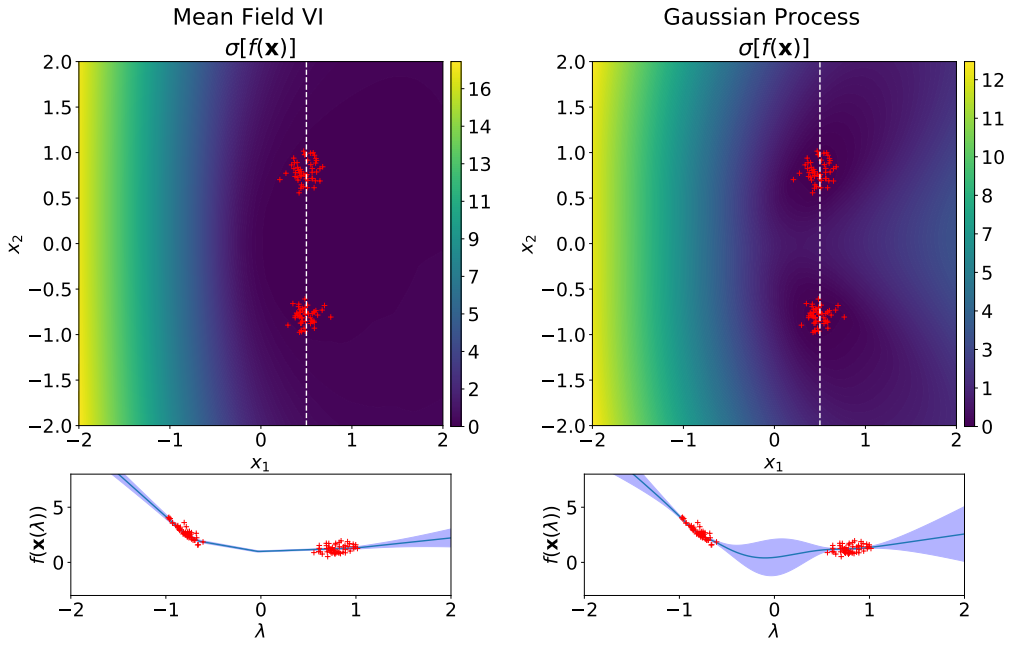

Figure 2: MFVI BNN (left) and GP with a kernel corresponding to the wide limit of the same BNN (right), fit on a regression dataset with 2-dimensional inputs (red crosses). The top plots show the standard deviation of the output in different regions of input space. The bottom plots show the predictive mean with two standard deviation bars along the dashed white line in the top plot. Note MFVI's overconfidence in the region $\lambda \in [-.8, .8]$. This behaviour is explained by theorem 1.

Theorem 2 (MC dropout)**.** *Consider a single-hidden layer ReLU neural network mapping from* $\mathbb{R}^D \to \mathbb{R}^K$ *with* $I \in \mathbb{N}$ *hidden units. The corresponding mapping is given by* $f_k(\mathbf{x}) = \sum_{i=1}^{I} w_{k,i}\phi\left(\sum_{d=1}^{D} u_{i,d}x_d + v_i\right) + b_k$ *for* $1 \leq k \leq K$*, where* $\phi(a) = \max(0, a)$*. Assume* $\mathbf{v}$ *is set deterministically and*

$$q(\mathbf{W}, \mathbf{b}, \mathbf{U}) = q(\mathbf{U})q(\mathbf{b}|\mathbf{U})\prod_i q_i(\mathbf{w}_i|\mathbf{U}),$$

where $\mathbf{w}_i = \{w_{k,i}\}_{k=1}^K$ *are the weights out of neuron* i*,* $\mathbf{b} = \{b_k\}_{k=1}^K$ *are the output biases and* $q(\mathbf{U})$*,* $q(\mathbf{b}|\mathbf{U})$ *and* $q_i(\mathbf{w}_i|\mathbf{U})$ *are arbitrary probability densities with finite first two moments. Then, for any finite set of points* $\mathcal{S} \subset \mathbb{R}^D$ *such that* $\mathbf{0}$ *is in the convex hull of* $\mathcal{S}$*,*

$$\mathrm{Var}[f_k(\mathbf{0})] \leq \max_{\mathbf{s}\in\mathcal{S}} \{\mathrm{Var}[f_k(\mathbf{s})]\} \quad \text{for} \quad 1 \leq k \leq K. \tag{3}$$

B Additional Figures

Figure 2 shows an illustration of case ii of theorem 1. Figure 3 shows plots of the variance of the output of the MFVI BNNs, along with the bounds implied by theorem 1. We see that if we take the points $\mathbf{p}$ and $\mathbf{q}$ to be the centres of the data clusters, we obtain a bound on the variance between them given by the red line, which explains the extremely small variance obtained by MFVI between them. Note that the variance increases rapidly outside the region $\overrightarrow{\mathbf{pq}}$, where our bounds cease to hold.

Figure 4 shows the variance of the output of the MC dropout BNN, as well as the bound at $\mathbf{0}$ implied by theorem 2.

C Proof of Theorems

In order to prove theorem 1 and theorem 2 we first decompose the variance into a sum of two terms. We prove in lemma 1 that the first term is convex as a function of $\mathbf{x}$. This may be of independent

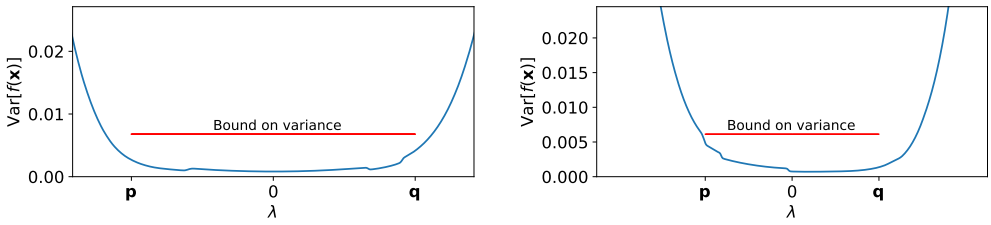

Figure 3: Bounds on the variance for MFVI implied by theorem 1 for the datasets shown in figure 1 (left) and figure 2 (right). $\mathbf{p}$ and $\mathbf{q}$ are the centres of the observed data clusters. Note that while the bound is not saturated in this case, proposition 3 implies that there exist members of the variational family that saturate the bound.

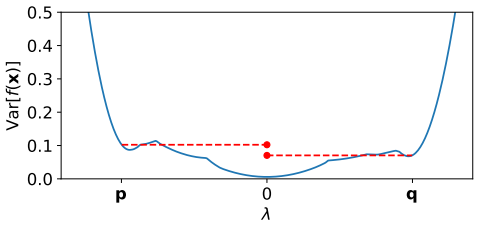

Figure 4: Bound on the variance at the origin for MC dropout implied by theorem 2 for the dataset shown in figure 1. $\mathbf{p}$ and $\mathbf{q}$ are the centres of the observed data clusters. The variance at the origin is upper bounded by the higher of the two red circles.

interest. To prove theorem 1, we note that the second term is a linear combination of the variances of individual neurons. In lemma 2 we show a property of the variance functions of individual neurons, that we leverage in lemma 3 to prove the main result. To prove theorem 2 we note that the second term has a global minimum at $\mathbf{x} = \mathbf{0}$. In the following we will use the notation introduced in appendix A.

C.1 Preliminary Lemmas

Lemma 1 ([7, Appendix B])**.** *Assume a distribution for the output parameters of the form*

$$q(\mathbf{W}, \mathbf{b}|\mathbf{U}, \mathbf{v}) = q(\mathbf{b}|\mathbf{U}, \mathbf{v}) \prod_i q_i(\mathbf{w}_i|\mathbf{U}, \mathbf{v}).$$

Then, $\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]$ *is a convex function of* $\mathbf{x}$*.*

A simplified proof of lemma 1 is in appendix D.1.

Lemma 2. *Consider the variance of a single neuron in the one dimensional case, with activation* $a(x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$*,* $\mu(x) = \mu_u x + \mu_v$ *and* $\sigma^2(x) = \sigma_u^2 x^2 + \sigma_v^2$ *. Let* $\mathcal{T}_1$ *denote the set of functions from* $\mathbb{R} \to [0, \infty)$ *satisfying* $t(x) \geq t(-x)$ *and* $t(x + \delta) \geq t(x)$ *for all* $x, \delta > 0$*. Let* $\mathcal{T}_2$ *denote the set of functions from* $\mathbb{R} \to [0, \infty)$ *satisfying* $t(x) \leq t(-x)$ *and* $t(x + \delta) \leq t(x)$ *for all* $\delta > 0$ *and* $x < 0$*. If* $\mu_u \geq 0$*, then* $\mathrm{Var}[\phi(a(x))] \in \mathcal{T}_1$*. If* $\mu_u \leq 0$*, then* $\mathrm{Var}[\phi(a(x))] \in \mathcal{T}_2$*.*

The proof of lemma 2 is in appendix D.2.

Corollary 1 (Corollary of lemma 2)**.** *Consider a line in* $\mathbb{R}^D$ *parameterized by* $[\mathbf{x}(\lambda)]_d = \gamma_d \lambda + c_d$ *for* $\lambda \in \mathbb{R}$ *such that* $\gamma_d c_d = 0$ *for* $1 \leq d \leq D$*. Let* $a(\mathbf{x}) \coloneqq \sum_{d=1}^D u_d x_d + v$ *with* $\{u_d\}_{d=1}^D$ *and* v *independent and Gaussian distributed. Then,* $\mathrm{Var}[\phi(a(\mathbf{x}(\lambda)))] \in \mathcal{T}_1 \cup \mathcal{T}_2$ *(as a function of* λ*).*

Proof. The activation $a(\mathbf{x}(\lambda))$ is a linear combination of Gaussian random variables, and is therefore Gaussian distributed. Moreover the mean is linear in λ. The variance of $a(\mathbf{x}(\lambda))$ is given by:

$$\begin{aligned}
\mathrm{Var}[a(\mathbf{x}(\lambda))] &= \sum_{d=1}^{D} \mathrm{Var}[u_d](\gamma_d\lambda + c_d)^2 + \mathrm{Var}[v] \\
&= \sum_{d=1}^{D} \sigma_{u_d}^2(\gamma_d\lambda + c_d)^2 + \sigma_v^2 \\
&= \lambda^2 \left(\sum_{d=1}^{D} \sigma_{u_d}^2\gamma_d^2\right) + 2\lambda \left(\sum_{d=1}^{D} \sigma_{u_d}^2\gamma_d c_d\right) + \left(\sum_{d=1}^{D} \sigma_{u_d}^2 c_d^2 + \sigma_v^2\right) \\
&= \lambda^2 \left(\sum_{d=1}^{D} \sigma_{u_d}^2\gamma_d^2\right) + \left(\sum_{d=1}^{D} \sigma_{u_d}^2 c_d^2 + \sigma_v^2\right)
\end{aligned}$$

Defining $\sigma_{\tilde{u}}^2 = \sum_{d=1}^{D} \sigma_{u_d}^2\gamma_d^2$ and $\sigma_{\tilde{v}}^2 = \sum_{d=1}^{D} \sigma_{u_d}^2 c_d^2 + \sigma_v^2$, the corollary follows from lemma 2. □

Lemma 3. *Let $\mathcal{C}$ be the set of convex functions from $\mathbb{R} \to [0, \infty)$. Fix any $a < 0 < b$ and c such that $|c| \leq \min(|a|, |b|)$. Then any function f that can be written as a linear combination of functions in $\mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{C}$ with non-negative weights satisfies, $f(c) \leq f(a) + f(b)$.*

The proof of lemma 3 can be found in appendix D.3.

Lemma 4. *Let $f : \mathbb{R}^D \to \mathbb{R}$ be a convex function and consider a finite set of points $\mathcal{S} \subset \mathbb{R}^D$. Then for any point $\mathbf{r}$ in the convex hull of $\mathcal{S}$, $f(\mathbf{r}) \leq \max_{\mathbf{s}\in\mathcal{S}}\{f(\mathbf{s})\}$.*

The proof of lemma 4 can be found in appendix D.4.

C.2 Proof of Theorem 1

Having collected the necessary preliminary lemmas we now prove theorem 1.

Proof of theorem 1. By the law of total variance,

$$\mathrm{Var}[f_k(\mathbf{x})] = \mathbb{E}[\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]] + \mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]].$$

Using lemma 1, $\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]$ is convex as a function of $\mathbf{x}$. As the expectation of a convex function is convex, the first term is a convex function of $\mathbf{x}$. For the second term we have

$$\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}] = \mathbb{E}\left[\sum_{i=1}^{I} w_{k,i}\phi(a_i) + b_k \middle| \mathbf{U}, \mathbf{v}\right] = \sum_{i=1}^{I} \mu_{w_{k,i}}\phi(a_i) + \mu_{b_k},$$

where $\mu_{w_{k,i}} \coloneqq \mathbb{E}[w_{k,i}]$, $\mu_{b_k} \coloneqq \mathbb{E}[b_k]$. In the second line we used linearity of expectation and that conditioned on $(\mathbf{U}, \mathbf{v})$, the a_i are deterministic. Next,

$$\mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]] = \mathrm{Var}\left[\sum_{i=1}^{I} \mu_{w_{k,i}}\phi(a_i) + \mu_{b_k}\right] = \sum_{i=1}^{I} \mu_{w_{k,i}}^2 \mathrm{Var}[\phi(a_i)], \tag{4}$$

since the a_i are independent of each other.

Consider a line in $\mathbb{R}^D$ parameterised by $[\mathbf{x}(\lambda)]_d = \gamma_d\lambda + c_d$ for $\lambda \in \mathbb{R}$ such that $\gamma_d c_d = 0$ for $1 \leq d \leq D$.

By corollary 1, $\mathrm{Var}[\phi(a_i(\mathbf{x}(\lambda)))] \in \mathcal{T}_1 \cup \mathcal{T}_2$ (as a function of λ). Since $\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]$ is convex as a function of $\mathbf{x}$, it is also convex as a function of λ. We have written $\mathrm{Var}[f_k(\mathbf{x}(\lambda))]$ in the form assumed in lemma 3, completing the proof. □

C.3 Proof of Theorem 2

Proof. By the law of total variance,

$$\mathrm{Var}[f_k(\mathbf{x})] = \mathbb{E}[\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}]] + \mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}]].$$

Using lemma 1, $\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}]$ is convex as a function of $\mathbf{x}$. As the expectation of a convex function is convex, the first term is a convex function of $\mathbf{x}$. This implies

$$\mathbb{E}[\mathrm{Var}[f_k(\mathbf{0})|\mathbf{U}]] \leq \max_{\mathbf{s}\in\mathcal{S}} \left\{\mathbb{E}[\mathrm{Var}[f_k(\mathbf{s})|\mathbf{U}]]\right\},$$

by lemma 4. $\mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}]]$ is non-negative everywhere. As the output of the first layer is independent of the matrix $\mathbf{U}$ at $\mathbf{x} = \mathbf{0}$, $\mathbb{E}[f_k(\mathbf{0})|\mathbf{U}]$ is deterministic. So $\mathrm{Var}[\mathbb{E}[f_k(\mathbf{0})|\mathbf{U}]] = 0$, completing the proof. □

D Proof of Lemmas

In this section we prove the preliminary lemmas stated in appendix C.1.

D.1 Proof of Lemma 1

Proof. We assume a distribution for the output weights such that:

$$q(\mathbf{W}, \mathbf{b}|\mathbf{U}, \mathbf{v}) = q(\mathbf{b}|\mathbf{U}, \mathbf{v}) \prod_i q_i(\mathbf{w}_i|\mathbf{U}, \mathbf{v}).$$

Consider the variance of the output under this distribution conditioned on the values of the weights and biases in the input layer:

$$\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}] = \sum_i \mathrm{Var}[w_{k,i}]\phi(a_i)^2 + \mathrm{Var}[b_k]. \tag{5}$$

with $a_i := \sum_{d=1}^{D} u_{i,d}x_d + v_i$. Equation (5) is justified since the weights from different neurons are independent under $q(\mathbf{W}, \mathbf{b}|\mathbf{U}, \mathbf{v})$.

Since $\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]$ is a linear combination of the $\phi(a_i)^2$ with non-negative weights (plus a constant), to prove convexity it suffices to show that each $\phi(a_i)^2$ is convex as a function of $\mathbf{x}$. $\phi(a_i)^2$ is convex as a function of a_i, since it is 0 for $a_i \leq 0$ and a_i^2 for $a_i > 0$. To show that it is convex as a function of $\mathbf{x}$, we write

$$\begin{aligned}
\phi\left(a_i(t\mathbf{x}_1 + (1-t)\mathbf{x}_2)\right)^2 &= \phi\left(\sum_d u_{i,d}\left(t[\mathbf{x}_1]_d + (1-t)[\mathbf{x}_1]_d\right) + v_i\right)^2 \\
&= \phi\left(t\left(\sum_d u_{i,d}[\mathbf{x}_1]_d + v_i\right) + (1-t)\left(\sum_d u_{i,d}[\mathbf{x}_2]_d + v_i\right)\right)^2 \\
&\leq t\phi\left(\sum_d u_{i,d}[\mathbf{x}_1]_d + v_i\right)^2 + (1-t)\phi\left(\sum_d u_{i,d}[\mathbf{x}_2]_d + v_i\right)^2 \\
&= t\phi\left(a_i(\mathbf{x}_1)\right)^2 + (1-t)\phi\left(a_i(\mathbf{x}_2)\right)^2,
\end{aligned}$$

completing the proof. □

D.2 Proof of Lemma 2

For a ReLU network [8]:

$$\mathrm{Var}[\phi(a)] = \mu(x)^2\left(\Phi(r)/r^2 + g(r) - g(r)^2\right) \tag{6}$$

with $\mu(x) = \mathbb{E}\left[a(x)\right] = \mu_u x + \mu_v, \sigma(x)^2 = \sigma_u^2 x^2 + \sigma_v^2, r = r(x) = \frac{\mu(x)}{\sigma(x)}$, $\Phi(r)$ the standard Gaussian CDF and

$$g(r) \coloneqq \frac{N(r)}{r} + \Phi(r),$$

with $N(r)$ the standard Gaussian PDF. We also define

$$f(r) := \Phi(r)/r^2 + g(r) - g(r)^2.$$

In order to prove lemma 2 we use this additional lemma.

Lemma 5. *For $r \neq 0$,*

$$\text{sgn}(f'(r)) = -\text{sgn}(r).$$

We now prove lemma 2 using lemma 5, which is proven in appendix D.5.

Proof. We first show that if $\mu_u > 0$, $\text{Var}[\phi(a(x))]$ is monotonically increasing for $x > 0$ and if $\mu_u < 0$, $\text{Var}[\phi(a(x))]$ is monotonically decreasing for $x < 0$. We show this by determining the sign of the derivative of the variance with respect to x. Using the product and chain rules:

$$\frac{d}{dx}[\text{Var}[\phi(a(x))]] = \frac{d}{dx}[\mu(x)^2 f(r(x))] = \mu(x)\left(2\mu_u f(r(x)) + \mu(x) r'(x) f'(r(x))\right). \tag{7}$$

We now split into cases based on the sign of μ_u and μ_v.

Case 1: $\mu_u, \mu_v > 0$ We want to show $\text{Var}[\phi(a(x))]$ is monotonically increasing for $x > 0$. We recall equation (7),

$$\begin{aligned}\frac{d}{dx}\text{Var}[\phi(a(x))] &= \mu(x)\left(2\mu_u f(r(x)) + \mu(x) r'(x) f'(r(x))\right) \\ &= \mu(x)\left(2\mu_u f(r(x)) + r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} f'(r(x))\right). \end{aligned} \tag{8}$$

In this region, $r(x) > 0$, and by lemma 5, $f'(r(x)) < 0$. For all $x \geq 0$,

$$\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} \leq \mu_u.$$

It follows that

$$\frac{d}{dx}\text{Var}[\phi(a(x))] \geq \mu(x)\left(2\mu_u f(r(x)) + r(x)\mu_u f'(r(x))\right) \tag{9}$$

$$= 2\mu_u\mu(x)\left(f(r(x)) + \frac{1}{2} r(x) f'(r(x))\right). \tag{10}$$

In order to show this is non-negative for $x > 0$, it suffices to show that $f(r) + \frac{1}{2} r f'(r) \geq 0$ for $r > 0$.

$$\begin{aligned} f(r) + \frac{1}{2} r f'(r) &= \frac{\Phi(r)}{r^2} + g(r) - g(r)^2 + \frac{-\Phi(r) + N(r)^2 + rN(r)\Phi(r)}{r^2} \\ &= g(r)(1 - \Phi(r)) \geq 0. \end{aligned}$$

The last inequality uses that for $r > 0$, $g(r) > 0$.

Case 2: $\mu_v > 0 > \mu_u$. Case 2 proceeds similarly to case 1. We want to show $\text{Var}[\phi(a(x))]$ is monotonically decreasing for $x < 0$. We recall equation (8),

$$\frac{d}{dx}\text{Var}[\phi(a(x))] = \mu(x)\left(2\mu_u f(r(x)) + r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} f'(r(x))\right).$$

In this region, $r(x) > 0$. By lemma 5, $f'(r(x)) < 0$. For all $x \leq 0$,

$$\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} \geq \mu_u.$$

It follows that

$$\frac{d}{dx}\text{Var}[\phi(a(x))] \leq \mu(x)\left(2\mu_u f(r(x)) + r(x)\mu_u f'(r(x))\right) \tag{11}$$

$$= 2\mu_u\mu(x)\left(f(r(x)) + \frac{1}{2} r(x) f'(r(x))\right). \tag{12}$$

As we have already established that $f(r(x)) + \frac{1}{2} r(x) f'(r(x)) \geq 0$ for $r > 0$, $\frac{d}{dx}\text{Var}[\phi(a(x))] \leq 0$ for $x < 0$.

Case 3: $\mu_u, \mu_v < 0$. We want to show the variance is monotonically decreasing for $x < 0$. We recall equation (8)

$$\begin{aligned}\frac{d}{dx}\text{Var}[\phi(a(x))] &= \mu(x)\left(2\mu_u f(r(x)) + r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} f'(r(x))\right) \\ &= 2\mu_u\mu(x)\left(f(r(x)) + \frac{1}{2}r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2} f'(r(x))\right). \end{aligned} \tag{13}$$

Subcase 3.1: $\mu(x) > 0$. If $\mu(x) > 0$,

$$\begin{aligned} \mu(x) = \mu_u x + \mu_v > 0 &\Rightarrow \\ \mu_u x > -\mu_v &\Rightarrow \\ \mu_u x^2 < -\mu_v x &\Rightarrow \\ \mu_u\sigma_u^2 x^2 < -\mu_v\sigma_u^2 x &\Rightarrow \\ \mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2 < \mu_u\sigma_v^2 - \mu_v\sigma_u^2 x &\Rightarrow \\ 1 > \frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2}. \end{aligned}$$

The overall sign in front of the term we just bounded above is positive, so this upper bound can be substituted in to bound the derivative of the variance above, yielding:

$$\frac{d}{dx}\text{Var}[\phi(a(x))] \leq 2\mu_u\mu(x)\left(f(r(x)) + \frac{1}{2}r(x)f'(r(x))\right) \leq 0.$$

As we have already shown that for $r > 0$, $f(r(x)) + \frac{1}{2}r(x)f'(r(x)) > 0$. This proves the subcase.

Subcase 3.2: $\mu(x) < 0$. Similarly, if $\mu_u, \mu_v, x < 0$ and $\mu(x) < 0$, then

$$\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2} > 1.$$

In this case, the overall sign in front of the term we just bounded is negative, so this lower bound can be substituted in to upper bound the derivative of the variance.

$$\frac{d}{dx}\text{Var}[\phi(a(x))] \leq 2\mu(x)\mu_u\left(f(r(x)) + \frac{1}{2}r(x)f'(r(x))\right).$$

We need to show that for $r < 0$, $f(r) + \frac{1}{2}rf'(r) = g(r)(1 - \Phi(r)) < 0$. This is equivalent to showing that $g(r) < 0$ for $r < 0$. This follows from the standard upper bound, $1 - \Phi(a) < \frac{1}{a}N(a)$ for $a > 0$.

Case 4: $\mu_v < 0 < \mu_u$. The proof of case 4 is similar to case 3. We want to show the variance is monotonically increasing for $x > 0$. We recall equation (8)

$$\begin{aligned}\frac{d}{dx}\text{Var}[\phi(a(x))] &= \mu(x)\left(2\mu_u f(r(x)) + r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\sigma_u^2 x^2 + \sigma_v^2} f'(r(x))\right) \\ &= 2\mu_u\mu(x)\left(f(r(x)) + \frac{1}{2}r(x)\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2} f'(r(x))\right). \end{aligned} \tag{14}$$

Subcase 4.1: $\mu(x) > 0$. If $\mu_v < 0$ and $\mu_u, x, \mu(x) > 0$, then

$$\frac{\mu_u\sigma_v^2 - \mu_v\sigma_u^2 x}{\mu_u\sigma_u^2 x^2 + \mu_u\sigma_v^2} < 1.$$

The overall sign in front of the term we just bounded above is negative, so this upper bound can be substituted in to lower bound the derivative of the variance, yielding:

$$\frac{d}{dx}\text{Var}[\phi(a(x))] \geq 2\mu(x)\mu_u\left(f(r(x)) + \frac{1}{2}r(x)f'(r(x))\right).$$

As we have already shown that for $r > 0$, $f(r(x)) + \frac{1}{2}r(x)f'(r(x)) > 0$. This proves the subcase.

Subcase 4.2: $\mu(x) < 0$. If $\mu_v, \mu(x) < 0$ and $\mu_u, x > 0$, then

$$\frac{\mu_u \sigma_v^2 - \mu_v \sigma_u^2 x}{\mu_u \sigma_u^2 x^2 + \mu_u \sigma_v^2} > 1.$$

In this case, the overall sign in front of the term we just bounded is positive, so this lower bound can be substituted in to lower bound the derivative of the variance:

$$\frac{d}{dx}\mathrm{Var}[\phi(a(x))] \geq 2\mu(x)\mu_u \left(f(r(x)) + \frac{1}{2} r(x) f'(r(x)) \right).$$

We need to show that for $r < 0$, $f(r) + \frac{1}{2} r f'(r) = g(r)(1 - \Phi(r)) < 0$. This was done in subcase 3.2.

In case 3 and 4, we excluded the case $\mu(x) = 0$. However, the monotonicity results still hold if $\mu(x) = 0$ by continuity of $\mathrm{Var}[\phi(a(x))]$ as a function of x. This completes the proof that if $\mu_u > 0$, $\mathrm{Var}[\phi(a(x))]$ is monotonically increasing for $x > 0$ and if $\mu_u < 0$, $\mathrm{Var}[\phi(a(x))]$ is monotonically decreasing for $x < 0$.

It remains to show that if $\mu_u > 0, \mathrm{Var}[\phi(a(x))] \geq \mathrm{Var}[\phi(a(-x))]$ for $x > 0$, and if $\mu_u < 0, \mathrm{Var}[\phi(a(x))] \geq \mathrm{Var}[\phi(a(-x))]$ for $x < 0$. From equation (6) we can write

$$\mathrm{Var}[\phi(a)] = \sigma^2(x) \left(\Phi(r) + r h(r) - h(r)^2 \right),$$

where $h(r) \coloneqq N(r) + r\Phi(r)$. Note $h'(r) = \Phi(r)$. We first show that $\alpha(r) \coloneqq \Phi(r) + r h(r) - h(r)^2$ is monotonically increasing.

Taking derivatives,

$$\alpha'(r) = N(r) + r\Phi(r) + h(r) - 2h(r)\Phi(r) = 2h(r)(1 - \Phi(r)).$$

It suffices to show that $h(r) > 0$. For $r > 0$, $h(r)$ is a sum of positive numbers. For $r < 0$, let $s = -r > 0$. Then

$$h(r) = -\Phi(-s)s + N(-s) = -(1 - \Phi(s))s + N(s).$$

The statement follows from the standard bound $\frac{N(s)}{s} > 1 - \Phi(s)$ for $s > 0$.

We next show that if $\mathrm{sgn}(x) = \mathrm{sgn}(\mu_u)$, $r(x) \geq r(-x)$. Since $\sigma(x) = \sigma(-x)$, it suffices to show that $\mu(x) \geq \mu(-x)$. This follows from

$$\mu(x) = \mu_u x + \mu_v \geq -\mu_u x + \mu_v = \mu(-x),$$

where we have used $\mathrm{sgn}(x) = \mathrm{sgn}(\mu_u)$ in the inequality.

Finally we have that if $\mathrm{sgn}(x) = \mathrm{sgn}(\mu_u)$, $\mathrm{Var}[\phi(a(x))] \geq \mathrm{Var}[\phi(a(-x))]$ since

$$\mathrm{Var}[\phi(a(x))] = \sigma^2(x)\alpha(r(x)) \geq \sigma^2(x)\alpha(r(-x)) = \mathrm{Var}[\phi(a(-x))].$$

In the inequality we have used the monotonicity of $\alpha(r)$ and that $r(x) \geq r(-x)$.

By continuity of the variance in all of its arguments, and that $\mathcal{T}_1$ and $\mathcal{T}_2$ are closed under pointwise limits, the boundary case μ_u, μ_v and x equal to 0 follow. This completes the proof of the lemma. □

D.3 Proof of Lemma 3

Proof. First, note that each of these three sets is closed under positive scaling and addition. We can therefore write f as a sum of three functions, $f(x) = t_1(x) + t_2(x) + s(x)$ with $t_1 \in \mathcal{T}_1, t_2 \in \mathcal{T}_2$ and $s \in \mathcal{C}$. We prove the case when $a < c < 0 < b$.

$$\begin{aligned}
f(c) &= t_1(c) + t_2(c) + s(c) \quad \text{(def.)} \\
&\leq t_1(c) + t_2(a) + s(c) \quad \text{(monotonicity of } \mathcal{T}_2\text{)} \\
&\leq t_1(-c) + t_2(a) + s(c) \quad (c < 0\text{, inequality condition for } \mathcal{T}_1) \\
&\leq t_1(b) + t_2(a) + s(c) \quad \text{(monotonicity of } \mathcal{T}_1, |c| = -c < b) \\
&\leq t_1(b) + t_2(a) + \max(s(a), s(b)) \quad (s \text{ convex}) \\
&\leq t_1(b) + t_2(a) + s(a) + s(b) \\
&\leq t_1(a) + t_1(b) + t_2(a) + t_2(b) + s(a) + s(b) \\
&= f(a) + f(b)
\end{aligned}$$

The case $a < 0 < c < b$ follows from symmetry. □

D.4 Proof of Lemma 4

Proof. Let $\{\mathbf{s}_n\}_{n=1}^N = \mathcal{S}_N \subset \mathbb{R}^D$. We proceed by induction. The lemma is clearly true for $N = 2$. Assume it is true for N. Let $\mathrm{Conv}(\mathcal{S}_{N+1})$ denote the convex hull of $\mathcal{S}_{N+1}$. Consider a point $\mathbf{r}_{N+1} \in \mathrm{Conv}(\mathcal{S}_{N+1})$. Then

$$f(\mathbf{r}_{N+1}) = f\left(\sum_{n=1}^{N+1} \alpha_n \mathbf{s}_n\right) \tag{15}$$

with $\sum_{n=1}^{N+1} \alpha_n = 1$ and $\alpha_n \geq 0$ for $1 \leq n \leq N+1$. We can write

$$f(\mathbf{r}_{N+1}) = f\left(\left(\sum_{n'=1}^{N} \alpha_{n'}\right)\mathbf{t}_N + \alpha_{N+1}\mathbf{s}_{N+1}\right) \tag{16}$$

$$\leq \max\{f(\mathbf{t}_N), f(\mathbf{s}_{N+1})\} \tag{17}$$

where $\mathbf{t}_N \coloneqq \sum_{n=1}^{N} \alpha_n \mathbf{s}_n \Big/ \sum_{n'=1}^{N} \alpha_{n'}$, and we have used the convexity of f. By the induction assumption, $f(\mathbf{t}_N) \leq \max_{\mathbf{s}\in\mathcal{S}_N}\{f(\mathbf{s})\}$, since $\mathbf{t}_N \in \mathrm{Conv}(\mathcal{S}_N)$. Combining this with equation (17) completes the proof. □

D.5 Proof of Lemma 5

In proving lemma 5 we consider the case $r < 0$ and $r > 0$ separately.

Proposition 1. *For* $r < 0$,

$$f'(r) \geq 0.$$

Proof. We begin by calculating $f'(r(x))$,

$$f'(r) = \frac{-2\Phi(r) + 2N(r)^2 + 2N(r)r\Phi(r)}{r^3}. \tag{18}$$

On the interval $r \in (-\infty, 0)$, $f'(r)$ is continuous. Additionally, $f'(-1) \approx .297 > 0$. This implies if $f'(r)$ is negative for some $r < 0$, then there exists an $s < 0$ such that $f'(s) = 0$. Suppose such an s exists, then

$$s = \frac{\Phi(s) - N(s)^2}{\Phi(s)N(s)}$$

We will reach a contradiction by showing that for $r < 0$, $h(r) \coloneqq \Phi(r) - N(r)^2 > 0$, as this contradicts $s < 0$. $h(r)$ is continuously differentiable. Therefore, we can prove $h(r) > 0$ for all $r < 0$ by showing:

1. $\lim_{r\to-\infty} h(r) \geq 0$.
2. $h'(r) > 0$ for all $r < 0$.

We first verify 1:

$$\lim_{r\to-\infty} h(r) = \lim_{r\to-\infty} \Phi(r) - \lim_{r\to-\infty} N(r)^2 = 0 - 0 \geq 0$$

We now verify 2:

$$h'(r) = N(r)(1 + 2rN(r)).$$

It remains to show $rN(r) > -1/2$ for $r < 0$. We find the minimum of $rN(r)$:

$$\frac{d}{dr}[rN(r)] = N(r)(1 - r^2).$$

On the negative real line, this has a unique zero at $r = -1$. This gives, $-N(-1) = -(2\pi e)^{-1/2} > -1/2$. Checking end points of the interval for potential minima: $\lim_{r\to-\infty} rN(r) = 0$ and $\lim_{r\to 0} rN(r) = 0$. Therefore, $rN(r) > -1/2$ implying $h'(r) > 0$ completing the proof.

□

Proposition 2. *For $r > 0$,*

$$f'(r) < 0.$$

Proof. Recall (equation (18)),

$$f'(r) = \frac{-2\Phi(r) + 2N(r)^2 + 2N(r)r\Phi(r)}{r^3}$$

Then for $r > 0$,

$$f'(r) \leq 0 \Leftrightarrow I(r) := -\Phi(r) + N(r)^2 + N(r)r\Phi(r) \leq 0.$$

Rearranging [1, 7.1.13] yields:

$$1 - \frac{2}{r + \sqrt{r^2 + 8/\pi}} N(r) \leq \Phi(r) < 1 - \frac{2}{r + \sqrt{r^2 + 4}} N(r). \tag{19}$$

for $r \geq 0$.

$$\begin{aligned}
I(r) &= -\Phi(r) + N(r)^2 + rN(r)\Phi(r) \\
&\leq -\Phi(r) + N(r)^2 + rN(r)\left(1 - \frac{2}{r + \sqrt{r^2 + 4}} N(r)\right) \\
&\leq -1 + \frac{2}{r + \sqrt{r^2 + 8/\pi}} N(r) + N(r)^2 + rN(r)\left(1 - \frac{2}{r + \sqrt{r^2 + 4}} N(r)\right) \\
&= -1 + \frac{2}{r + \sqrt{r^2 + 8/\pi}} N(r) + N(r)^2 + rN(r) - \frac{2r}{r + \sqrt{r^2 + 4}} N(r)^2 \\
&= -1 + \frac{2}{r + \sqrt{r^2 + 8/\pi}} N(r) + rN(r) + N(r)^2 \left(1 - \frac{2r}{r + \sqrt{r^2 + 4}}\right)
\end{aligned} \tag{20}$$

We now make the use of numerous crude bounds which hold for $r > 0$:

1. $N(r) \leq 1/\sqrt{2\pi}$,
2. $\frac{2}{r + \sqrt{r^2 + 8/\pi}} \leq \frac{2}{\sqrt{8/\pi}} = \sqrt{\pi/2}$,
3. $rN(r) \leq 1/\sqrt{2\pi e}$
4. $\left(1 - \frac{2r}{r + \sqrt{r^2 + 4}}\right) \leq 1$

Plugging these in to equation (20),

$$I(r) \leq -1 + \frac{\sqrt{\pi/2}}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi e}} + \frac{1}{2\pi} = -\frac{1}{2} + \frac{1}{\sqrt{2\pi e}} + \frac{1}{2\pi} \approx -0.098 < 0.$$

□

E Tightness of Bounds

Here we prove that the bound in case i of theorem 1 is tight.

Proposition 3. *Let the line segment $\overrightarrow{\mathbf{pq}}$ from case i of theorem 1 be parameterised by λ such that $\mathbf{p} = \mathbf{x}(\lambda_p)$ and $\mathbf{q} = \mathbf{x}(\lambda_q)$ with $\lambda_p < 0 < \lambda_q$. The bound given in theorem 1 is tight in the sense that for all $I \geq 2$ and any $\delta > 0$, for all $\lambda_p < \lambda < \lambda_q$ there exists a distribution of the form assumed in theorem 1 such that $\mathrm{Var}[f_k(\mathbf{p})] + \mathrm{Var}[f_k(\mathbf{q})] - \mathrm{Var}[f_k(\mathbf{x}(\lambda))] < \delta$.*

Proof. Recall

$$\mathrm{Var}[f_k(\mathbf{x})] = \mathbb{E}[\mathrm{Var}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]] + \mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]]$$

where

$$\mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U},\mathbf{v}]] = \mathrm{Var}\left[\sum_{i=1}^{I} \mu_{w_{k,i}}\phi(a_i) + \mu_{b_k}\right] = \sum_{i=1}^{I} \mu_{w_{k,i}}^2 \mathrm{Var}[\phi(a_i)]. \tag{21}$$

We will focus on one term in this sum, and suppress the indices i, k. First consider the case where x is one-dimensional. Recall the variance expression:

$$\mathrm{Var}[\phi(a)] = \sigma^2(x)\left(\Phi(r) + rh(r) - h(r)^2\right).$$

where $\sigma^2(x) = \sigma_u^2 x^2 + \sigma_v^2$, $r = \frac{\mu_u x + \mu_v}{\sqrt{\sigma_u^2 x^2 + \sigma_v^2}}$, and $h(r) \coloneqq N(r) + r\Phi(r)$. Now consider a distribution such that $\sigma_v = \epsilon\tilde{\sigma}_v, \sigma_u = \epsilon^2\tilde{\sigma}_u$ and $\mu_w = \frac{\tilde{\mu}_w}{\epsilon}$ with $\epsilon > 0$. Each term in the sum in equation (21) is of the form:

$$\begin{aligned}
\mu_w^2 \mathrm{Var}[\phi(a)] &= \mu_w^2 \sigma^2(x)\left(\Phi(r) + rh(r) - h(r)^2\right) \\
&= \frac{\tilde{\mu}_w^2}{\epsilon^2}(\sigma_u^2 x^2 + \sigma_v^2)\left(\Phi(r) + rh(r) - h(r)^2\right) \\
&= \tilde{\mu}_w^2(\epsilon^2\tilde{\sigma}_u^2 x^2 + \tilde{\sigma}_v^2)\left(\Phi(r) + rh(r) - h(r)^2\right).
\end{aligned}$$

We also have

$$r(x) = \frac{\mu_u x + \mu_v}{\sqrt{\sigma_u^2 x^2 + \sigma_v^2}} = \frac{\mu_u x + \mu_v}{\epsilon\sqrt{\epsilon^2\tilde{\sigma}_u^2 x^2 + \tilde{\sigma}_v^2}}.$$

Now consider the limit of the variance function as $\epsilon \to 0$. We have

$$\begin{aligned}
\lim_{\epsilon\to 0} \mu_w^2 \mathrm{Var}[\phi(a)] &= \lim_{\epsilon\to 0} \tilde{\mu}_w^2(\epsilon^2\tilde{\sigma}_u^2 x^2 + \tilde{\sigma}_v^2) \lim_{\epsilon\to 0}\left(\Phi(r) + rh(r) - h(r)^2\right) \\
&= \tilde{\mu}_w^2\tilde{\sigma}_v^2 \lim_{\epsilon\to 0}\left(\Phi(r) + rh(r) - h(r)^2\right)
\end{aligned}$$

If $\mu_u > 0$, then for $x > -\frac{\mu_v}{\mu_u}$, $\lim_{\epsilon\to 0} \mu_w^2 \mathrm{Var}[\phi(a(x))] = \tilde{\mu}_w^2\tilde{\sigma}_v^2 \lim_{r\to\infty} \alpha(r)$ where we have defined $\alpha(r) \coloneqq \left(\Phi(r) + rh(r) - h(r)^2\right)$. We have

$$\begin{aligned}
\lim_{r\to\infty} \alpha(r) &= \lim_{r\to\infty}\left(\Phi(r) + rN(r) + r^2\Phi(r)(1-\Phi(r)) - N(r)^2 - 2rN(r)\Phi(r)\right) && (22)\\
&= \lim_{r\to\infty}\left(\Phi(r) + r^2\Phi(r)(1-\Phi(r))\right) && (23)\\
&= \lim_{r\to\infty} \Phi(r) && (24)\\
&= 1. && (25)
\end{aligned}$$

In equation (23) we used that $\lim_{r\to\infty} rN(r) = 0$ since by l'Hôpital's rule we have

$$\lim_{r\to\infty} rN(r) = \lim_{r\to\infty} \frac{r}{\sqrt{2\pi}\exp\left(\frac{1}{2}r^2\right)} = \lim_{r\to\infty} \frac{1}{\sqrt{2\pi}r\exp\left(\frac{1}{2}r^2\right)} = 0.$$

In equation (24) we used the fact that $r^2(1-\Phi(r)) < rN(r)$ for $r > 0$, hence

$$\begin{aligned}
\lim_{r\to\infty} r^2(1-\Phi(r)) &\leq \lim_{r\to\infty} rN(r) \\
&= 0,
\end{aligned}$$

and since $r^2(1-\Phi(r)) > 0$, it follows that $\lim_{r\to\infty} r^2(1-\Phi(r)) = 0$. For $x < -\frac{\mu_v}{\mu_u}$, we have $\lim_{\epsilon\to 0} \mu_w^2 \mathrm{Var}[\phi(a(x))] = \tilde{\mu}_w^2\tilde{\sigma}_v^2 \lim_{r\to-\infty} \alpha(r)$. This is given by

$$\lim_{r\to-\infty} \alpha(r) = \lim_{r\to-\infty}\left(\Phi(r) + rN(r) + r^2\Phi(r)(1-\Phi(r)) - N(r)^2 - 2rN(r)\Phi(r)\right) = 0,$$

where we have used the fact that $\lim_{r\to-\infty} r^2\Phi(r) = 0$, since $r^2\Phi(r) > 0$ and

$$\lim_{r\to-\infty} r^2\Phi(r) = \lim_{r'\to\infty} r'^2\Phi(-r') = \lim_{r'\to\infty} r'^2(1-\Phi(r')) \leq \lim_{r'\to\infty} r'N(r') = 0.$$

In summary, if $\mu_u > 0$, for all $x \neq -\frac{\mu_v}{\mu_u}$ the pointwise limit of $\mu_w^2 \mathrm{Var}[\phi(a(x))]$ as $\epsilon \to 0$ is a Heaviside step function with height $\tilde{\mu}_w^2 \tilde{\sigma}_v^2$, taking the value 0 for all $x < -\frac{\mu_v}{\mu_u}$. Similarly, if $\mu_u < 0$, the pointwise limit of $\mu_w^2 \mathrm{Var}[\phi(a(x))]$ is a step function of the same height but taking the value 0 for all $x > -\frac{\mu_v}{\mu_u}$.

In one dimension we can therefore saturate the bound as follows. Set the number of neurons, $I = 2$. Let the pointwise limit of the contribution of the first neuron to $\mathrm{Var}[\mathbb{E}[f_k(\mathbf{x})|\mathbf{U}, \mathbf{v}]]$ be a step function of height $\tilde{\mu}_{w_1}^2 \tilde{\sigma}_{v_1}^2 \coloneqq V_p$ taking the value 0 for $x < -\frac{\mu_{v_1}}{\mu_{u_1}} \coloneqq \lambda_p$. Let the limit of the contribution of the second neuron be a step function of height $\tilde{\mu}_{w_2}^2 \tilde{\sigma}_{v_2}^2 \coloneqq V_q$ taking the value 0 for $x > -\frac{\mu_{v_2}}{\mu_{u_2}} \coloneqq \lambda_q$. Note that there are enough degrees of freedom to set V_p, V_q, λ_p and λ_q independently. We choose $\lambda_q > \lambda_p$.

Then for any $\delta, l > 0$ and any $\lambda_p < \lambda < \lambda_q$, there exists an $\epsilon > 0$ such that

$$\mathrm{Var}[\mathbb{E}[f(\lambda_p - l)|\mathbf{U}, \mathbf{v}]] + \mathrm{Var}[\mathbb{E}[f(\lambda_q + l)|\mathbf{U}, \mathbf{v}]] - \mathrm{Var}[\mathbb{E}[f(\lambda)|\mathbf{U}, \mathbf{v}]] < \delta,$$

since

$$\begin{aligned} \mathrm{Var}[\mathbb{E}[f(\lambda_p - l)|\mathbf{U}, \mathbf{v}]] &\to V_q \\ \mathrm{Var}[\mathbb{E}[f(\lambda_q + l)|\mathbf{U}, \mathbf{v}]] &\to V_p \\ \mathrm{Var}[\mathbb{E}[f(\lambda)|\mathbf{U}, \mathbf{v}]] &\to V_p + V_q \end{aligned}$$

pointwise. Taking $\lambda_p - l$ and $\lambda_q + l$ to be the λ's from the statement of the remark, this proves the remark in the one-dimensional case. Note that in this construction we have not considered the contribution of the first term, $\mathbb{E}[\mathrm{Var}[f(x)|\mathbf{U}, \mathbf{v}]]$. This term can be ignored by setting the $\mathrm{Var}[w_{k,i}]$ sufficiently small. For $I > 2$, neurons not used in this construction can be ignored similarly.

To handle the D-dimensional case, note that each term in the sum in equation (21) is of the form

$$\mu_{w_{k,i}}^2 \mathrm{Var}[\phi(a_i(\mathbf{x}(\lambda)))] = \mu_{w_{k,i}}^2 \mathrm{Var}\left[\phi\left(\sum_{d=1}^{D} u_{i,d}\gamma_d \lambda + v_i\right)\right],$$

so the construction can be reduced to the one-dimensional case upon defining the Gaussian random variable $\tilde{u}_i \coloneqq \sum_{d=1}^{D} u_{i,d}\gamma_d$. Since we can set $\mu_{\tilde{u}_i}, \mu_{v_i}, \sigma_{\tilde{u}_i}^2, \sigma_{v_i}^2$ independently by choosing appropriate values for the $\mu_{u_{i,d}}, \mu_{v_i}, \sigma_{u_{i,d}}^2$ and $\sigma_{v_i}^2$, the same construction can be used as in the one-dimensional case, completing the proof. □

F Deep Networks

In this section we discuss if our results can be extended to deeper networks. In particular, we could ask if mean-field Gaussian posteriors can represent in-between uncertainty with a 2-hidden layer ReLU BNN. We give a construction[4] to show that any mean and variance function can be approximated in this case. Consider a 2-hidden layer BNN with $I \in \mathbb{N}$ hidden units in the first hidden layer, and 2 hidden units in the second hidden layer. The mapping is defined by:

$$\begin{aligned} \mathbf{h}^{(1)}(\mathbf{x}) &= \phi\left(\mathbf{W}^{(0)}\mathbf{x} + \mathbf{b}^{(0)}\right) \\ \mathbf{h}^{(2)}(\mathbf{x}) &= \phi\left(\mathbf{W}^{(1)}\mathbf{h}^{(1)}(\mathbf{x}) + \mathbf{b}^{(1)}\right) \\ f(\mathbf{x}) &= W_1^{(2)} h_1^{(2)}(\mathbf{x}) + W_2^{(2)} h_2^{(2)}(\mathbf{x}) + b^{(2)}, \end{aligned}$$

where $\mathbf{h}^{(1)} \in \mathbb{R}^I$ and $\mathbf{h}^{(2)} = [h_1^{(2)}, h_2^{(2)}]^\mathsf{T} \in \mathbb{R}^2$. Each weight and bias in this network is represented by an independent Gaussian distribution. Consider setting the parameters $(\mathbf{W}^{(0)}, \mathbf{b}^{(0)}, \mathbf{W}^{(1)}, \mathbf{b}^{(1)})$ to be deterministic, by sending their variances to zero. Then the mapping from $\mathbf{x} \to \mathbf{W}^{(1)}\mathbf{h}^{(1)}(\mathbf{x}) + \mathbf{b}^{(1)}$ is a single-hidden layer deterministic MLP. By the universal approximation theorem [17], this mapping can approximate any continuous function for sufficiently large I. Therefore, with $\phi(a) = \max(a, 0)$, the mapping from $\mathbf{x} \to \mathbf{h}^{(2)}(\mathbf{x})$ can approximate any non-negative continuous function. Now consider

[4] We thank Sebastian W. Ober for help with this construction.

setting $W_1^{(2)} = 1$ deterministically, setting the variance $b^{(2)}$ to zero, and letting $W_2^{(2)} \sim \mathcal{N}(0, 1)$. Then for the mean of the output we have

$$\begin{aligned}\mathbb{E}[f(\mathbf{x})] &= \mathbb{E}[W_1^{(2)} h_1^{(2)}(\mathbf{x}) + W_2^{(2)} h_2^{(2)}(\mathbf{x}) + b^{(2)}] \\ &= \mathbb{E}[W_1^{(2)} h_1^{(2)}(\mathbf{x}) + b^{(2)}] \\ &= h_1^{(2)}(\mathbf{x}) + b^{(2)},\end{aligned}$$

and for the variance we have

$$\begin{aligned}\mathrm{Var}[f(\mathbf{x})] &= \mathrm{Var}[W_1^{(2)} h_1^{(2)}(\mathbf{x}) + W_2^{(2)} h_2^{(2)}(\mathbf{x}) + b^{(2)}] \\ &= \mathrm{Var}[W_2^{(2)} h_2^{(2)}(\mathbf{x})] \\ &= \mathrm{Var}[W_2^{(2)}](h_2^{(2)}(\mathbf{x}))^2 \\ &= (h_2^{(2)}(\mathbf{x}))^2.\end{aligned}$$

Since $h_1^{(2)}(\mathbf{x})$ is a universal approximator of non-negative functions, the mean function $\mathbb{E}[f(\mathbf{x})] = h_1^{(2)}(\mathbf{x}) + b^{(2)}$ is a universal function approximator. Since $h_2^{(2)}(\mathbf{x})$ is another universal approximator of non-negative functions, the variance function $\mathrm{Var}[f(\mathbf{x})] = (h_2^{(2)}(\mathbf{x}))^2$ is a universal approximator of non-negative functions. Thus a 2-hidden layer BNN with a mean-field Gaussian posterior can approximate any desired continuous mean and variance function, and hence can represent in-between uncertainty.

Our construction may be considered somewhat pathological, as it sets many of the parameter variances to zero[5]. This would lead to the KL-divergence between the approximate posterior and the prior being infinite if MFVI is used. However, there may be other, less pathological settings of the variational parameters that give similarly flexible mean and variance functions. In practice, this construction simply shows that there exist mean-field Gaussian distributions that can provide in-between uncertainty. It does not tell us that those distributions will be found by, e.g. optimising the ELBO. In fact, experiments have not found that adding depth helps with in-between uncertainty [7]. In future work, we aim to investigate further why MFVI fails to obtain in-between uncertainty even with deep networks.

G Experimental Details

G.1 Data Generation

The input locations of data were generated by sampling 100 total points, 50 each from two distinct Gaussians. In figure 1, one Gaussian was centred at $(-1, -1)$ and the other at $(1, 1)$; both had isotropic variance of 0.01. For figure 2 the Gaussians were centred at $(.5, \pm.8)$ and had isotropic variance of 0.01. The output values were generated by sampling from the Gaussian process prior with kernel resulting from the wide limit of the BNN at these input values.

G.2 Model and Training

Each network contained a single hidden layer with 50 units. The prior standard deviation on biases was set to 1. The prior standard deviation on weights in the top layer was set to $4/\sqrt{50}$ and on the bottom to $4/\sqrt{2}$. The scaling on the prior standard deviation on weights is chosen so that in the wide limit the BNN prior converges to a Gaussian process [23]. The noise variance was set to 0.01. The Gaussian process regression model was fit with the equivalent ReLU kernel [3] using exact inference. All networks were trained for 50,000 epochs using Adam with learning rate 10^{-3}. For MC dropout, we used a dropout probability $p = 0.1$, and the L^2 regularisation constants were chosen such that the 'KL condition' [9, Chapter 3.2.3] holds. We used 500 samples at test time to estimate the mean and standard deviation of the predictive distribution.

[5]This construction can be made rigorous with small strictly positive variances.