Can Bayesian Neural Networks Make Confident Predictions?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bayesian inference promises a framework for principled uncertainty quantification of neural network predictions. Barriers to adoption include the difficulty of fully characterizing posterior distributions on network parameters and the interpretability of posterior predictive distributions. We demonstrate that under a discretized prior for the inner layer weights, we can exactly characterize the posterior predictive distribution as a Gaussian mixture. This setting allows us to define equivalence classes of network parameter values which produce the same training error, and to relate the elements of these classes to the network's scaling regime—defined via ratios of the training sample size, the size of each layer, and the number of final layer parameters. Of particular interest are distinct parameter realizations that map to low training error and yet correspond to distinct modes in the posterior predictive distribution. We identify settings that exhibit such predictive multimodality, and thus provide insight into the accuracy of unimodal posterior approximations. We also characterize the capacity of a model to "learn from data" by evaluating contraction of the posterior predictive in different scaling regimes.

1 Introduction

Uncertainty is key to learning. Questions of how to quantify neural network prediction uncertainty are inextricable from questions of how expressive models learn to generalize [28, 29, 22]. Progress on these questions has been made through analysis of relatively simple networks, including random features models [23] and neural tangent kernels [15], which demonstrate the double descent phenomenon [5, 4, 19, 6, 1]. An array of uncertainty metrics have been proposed for neural networks, as detailed by [11], but most approaches rely on heuristics which make interpretation challenging even in simple networks.

Bayesian neural networks (BNNs) promise a principled framework for obtaining predictive distributions conditioned on training data [20, 18, 2]. Realizing this promise has been complicated by the need to design appropriate prior and likelihood models and to characterize multimodal posterior distributions. Locating all modes via sampling is generally intractable, though mode connectivity and algorithm choice may aid in discovering parameter values that successfully generalize [10, 14, 25, 21]. Many strategies for approximate inference in BNNs have also been developed. The Laplace approximation [9, 17] represents the predictive distribution with a single mode. Variational inference methods [13, 3, 7] are more flexible, but typically capture at most a few posterior modes. Such approaches seem to risk underestimating uncertainty, though the debate about "cold posteriors" has raised the possibility that narrower distributions may produce better generalization [27]. Partially Bayesian networks [16, 26] could offer uncertainty estimates without introducing the challenge of learning distributions over all parameters. Broadly, however, there is a need for tools that provide insight into what these approximations of the Bayesian posterior miss.

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2024. Do not distribute.

In this work, we demonstrate that adopting a discrete prior on the inner layer weights of a BNN is a useful tool for accessing the predictive distribution without exhaustively sampling parameter space. Such priors allow us to identify cases where different posterior modes map to distinct modes in the predictive distribution. Then we can determine when predictions based on a single posterior mode will fail. To the authors' knowledge, this approach to analyzing multimodality is unique, though the different treatment of inner and final layer parameters during inference bears some resemblance to work on subnetwork inference [8, 26], partial Bayesian networks [16, 26], and random features models [23]. Furthermore, characterizing the posterior predictive distribution allows us to identify settings where the predictive uncertainty does *not* contract as the network and training set size grow proportionally. This behavior raises the question of whether overparameterized BNNs can produce "confident predictions," i.e., predictions whose posterior distribution contracts around the truth as the network and data set size grow.

Section 2 outlines our model and approach to inference. Sections 3 and 4 examine the impact of network and training set size on predictive uncertainty for a discretized Gaussian prior and then for a prior which puts mass on optimal parameter values. We conclude with a discussion of the implications of multimodality for approximate inference tools and the role of Bayesian uncertainty in successful generalization.

2 Predictive distribution of Bayesian neural network

We consider an L-layer neural network in a regression setting,

$$\begin{aligned} \hat{y} &= w^\top x_L, && (1)\\ x_\ell &= \sigma(\Theta_{\ell-1}^\top x_{\ell-1}) \; - \; b_\ell, \qquad 1 < \ell \leq L, && (2) \end{aligned}$$

where $x_1 \in \mathbb{R}^d$ is the network input, $\hat{y} \in \mathbb{R}$ is the output, and σ is a nonlinear activation function that operates component-wise. The trainable parameters include the final layer weights $w \in \mathbb{R}^p$ and interior parameters $\Theta := \{\Theta_\ell \in \mathbb{R}^{d_\ell \times d_{\ell+1}},\ b_\ell \in \mathbb{R}^{d_\ell}\}_{\ell=1}^L$. Note that $d_1 = d$ and $d_L = p$. We make the prior assumption that

$$w \sim \mathcal{N}(0, p^{-1}\mathbf{I}_p), \qquad \mathbb{P}(\Theta = \Theta^{(j)}) = \rho_j, \qquad \sum_{j=1}^J \rho_j \; = \; 1, \tag{3}$$

where each $\Theta^{(j)}$ is a fixed realization of the interior parameters. Crucially, the discrete prior on Θ allows us to derive an analytical representation of the Bayesian posterior predictive distribution.

Our training set has the form $\{(x_1^{(i)}, y^{(i)})\}_{i=1}^n$ where we assume that

$$y^{(i)} = g(x_1^{(i)}) \; + \; \varepsilon^{(i)}, \qquad \varepsilon^{(i)} \overset{\text{iid}}{\sim} \; N(0, \gamma^2), \tag{4}$$

for some (unknown) data-generating function $g : \mathbb{R}^d \to \mathbb{R}$. For convenience, we define $X_\ell := [x_\ell^{(1)}, \ldots, x_\ell^{(n)}] \in \mathbb{R}^{d_\ell \times n}$, for any $\ell \in [L]$, and $Y := (y^{(1)}, \ldots, y^{(n)}) \in \mathbb{R}^n$. The training data can thus be written more concisely as (X_1, Y). Let $\widetilde{x}_1 \in \mathbb{R}^d$ be an input value at which we will test our network predictions and let $\widetilde{y} \in \mathbb{R}$ denote the corresponding output. Under our model assumptions, the posterior predictive density for $\widetilde{y}$ is a J-component Gaussian mixture:

$$\pi(\widetilde{y} \,|\, X_1, Y, \widetilde{x}_1) \;\; = \;\; \sum_{j=1}^J \mathbb{P}(\Theta^{(j)} \,|\, X_1, Y)\, \pi(\widetilde{y} \,|\, X_1, Y, \widetilde{x}_1, \Theta^{(j)}), \tag{5}$$

For each j, Bayesian linear regression yields

$$\begin{aligned} \pi(\widetilde{y} \,|\, X_1, Y, \widetilde{x}_1, \Theta^{(j)}) \;\; = \;\; & \mathcal{N}\big(\widetilde{y};\, p^{-1}\widetilde{x}_L^\top X_L (p^{-1} X_L^\top X_L \; + \; \gamma^2 \mathbf{I})^{-1} Y, \\ & \qquad \gamma^2 \mathbf{I} \; + \; \gamma^2 p^{-1} \widetilde{x}_L^\top (p^{-1} X_L X_L^\top \; + \; \gamma^2 \mathbf{I})^{-1} \widetilde{x}_L\big), \end{aligned} \tag{6}$$

where dependence on $\widetilde{x}_1$ in the mean and variance terms above enters via $\widetilde{x}_L$, as described in (2). Note that both X_L and $\widetilde{x}_L$ depend on $\Theta^{(j)}$. By Bayes' rule, the mixture weights are

$$\mathbb{P}(\Theta^{(j)} \big| X_1, Y) \;=\; \frac{\rho_j \, \pi(Y|X_1, \Theta^{(j)})}{\pi(Y|X_1)} \;=\; \left(1 + \sum_{k \neq j} \frac{\rho_k \pi(Y|X_1, \Theta^{(k)})}{\rho_j \pi(Y|X_1, \Theta^{(j)})}\right)^{-1}, \tag{7}$$

where

$$\pi(Y|X_1, \Theta^{(j)}) = \mathcal{N}\big(Y; \mathbf{0},\ p^{-1}X_L^T X_L + \gamma^2 \mathbf{I}\big) =: \mathcal{L}(X_L(\Theta^{(j)})) \tag{8}$$

is the marginal likelihood function for $\Theta^{(j)}$.

Assuming that $\rho_j = 1/J$ for all $j \in [J]$, the j^{th} mode of the posterior predictive will have a large weight only if $\mathcal{L}(X_L(\Theta^{(j)}))$ is large compared with the marginal likelihood of all other candidate Θ values.

3 Multimodality under a discretized Gaussian prior

At this stage, it is not obvious whether multimodal distributions on (w, Θ) map to multimodal distributions in the space of predictions, e.g., the distribution of $\widetilde{y}$ at at given input $\widetilde{x}_1$. [1] As (5) shows, the posterior predictive distribution is the average over predictive distributions obtained by fixing each $\Theta^{(j)}$ and inferring w. Thus, the predictive distribution can be interpreted as an average over random features models, where the weight of the j^{th} model is determined by how compatible $\Theta^{(j)}$ is with Y compared to each $\Theta^{(k \neq j)}$. It is natural to ask for which regimes of n, p, and d it is possible to obtain multiple realizations of Θ that each produce high marginal likelihoods $\mathcal{L}(\Theta)$, but map to *distinct* predictive modes.

In this section, we consider two layer networks where bias parameters are set to 0 and the remaining components of $\{\Theta^{(j)}\}_{j=1}^J$ are fixed by independently sampling from $\mathcal{N}(0, c/d)$ for some constant c. We set each $\rho_j = 1/J$. Note that this choice of prior may be considered a discretization of a Gaussian prior, a common minimally informative choice for BNNs [2]. As is generally the case for Monte Carlo schemes, it is intractable to fully explore the continuous parameter space represented by a Gaussian prior, but larger J will correspond to greater coverage. For our experiments, we choose $J = 200\,000$. The columns of X and $\widetilde{x}_1$ are drawn from standard Gaussian distributions, and Y and $\widetilde{y}$ are chosen to have unit variance. We consider the rectified linear unit (ReLU) activation function, and set $c = 2\pi/(\pi - 1)$ so that the prior predictive distribution has unit variance.

Figure 1 summarizes the findings of these experiments. The left and center columns show predictive distributions at a given $\widetilde{x}_1$ for select (p, n) pairs and $d = 100$. Each indigo region represents the predictive distribution corresponding to a candidate $\Theta^{(j)}$; darker shades indicate larger weights as given by (7). The black curve marks the full posterior predictive distribution. Clearly, in our setting, Bayesian inference can produce multimodal predictive distributions. Each posterior predictive distribution demonstrates smaller variance than the prior predictive distribution; thus, conditioning on training data has produced a reduction in uncertainty. Appendix A.1 provides examples of the posterior predictive distributions at additional test points and for larger network sizes.

The rightmost column of Figure 1 documents a more extensive exploration of the impact of n, p, and d. For $d \in \{10, 100, 1000\}$ and ratios n/d and p/d ranging from 0.5 to 2, we plot the log of the number of candidates $\Theta^{(j)}$ which produce a Gaussian mixture component with weight larger than 10^{-6}. Note that the total number of trainable parameters in the network we consider is $p(d + 1)$, so each network considered is overparameterized. If we restrict our attention to inference in the final layer weights w, however, only the entries below the right leaning diagonal of each heatmap correspond to overparameterized networks.

Network and training set size clearly influence the number of modes that are significant in the posterior predictive distribution. When $d = 10$ we see that for several of the n and p values considered, up to 98% of the the candidate inner layer parameters make significant contributions to the posterior predictive distribution. By contrast, for larger input dimensions, when n is close to p we often find only one significant mode, leading to a unimodal posterior predictive distribution. These findings are expected if we recall the dependence of (7) on $\mathcal{L}(\Theta)$. If p is sufficiently larger than n, many candidates $\Theta^{(j)}$ will produce large $\mathcal{L}(\Theta)$ due to final layer overparameterization. Since $X_{L-1}^\top \Theta^{(j)}$ is full rank with high probability, if n is much larger than p, it becomes challenging to identify a single candidate $\Theta^{(j)}$ which could reproduce the training data; but many of the available candidates produce similar $\mathcal{L}(\Theta)$ and thus contribute to the posterior predictive. The line $n = p$ represents a

[1] In this paper, we focus only on marginal predictive distributions. It is straightforward, however, to characterize the joint distribution of predictions at any collection of different input values.

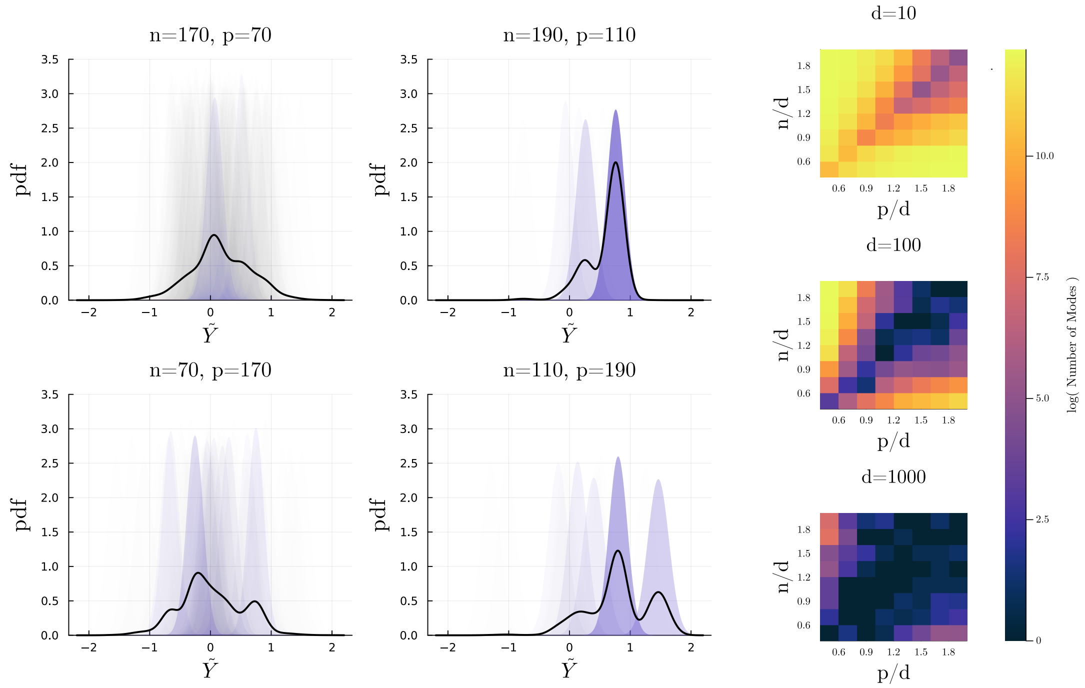

Figure 1: Left and center: posterior predictive distributions for input dimension $d = 100$ at select training set size n and final layer width p, as indicated by each title. The black line shows the pdf which is a mixture of Gaussians. Each shaded distribution is a component of this mixture with transparency corresponding to its weight. Right: Heatmaps depicting the log of the number of component distributions which have weight larger than 10^{-6} for specified network dimensions. Observation noise variance is set to $\gamma^2 = 0.01$ for these results.

phase transition around which one or a few candidates are likely to outperform the others. Appendix A.2 provides more discussion of the impact of this transition from under- to overparameterization (in terms of final layer weights).

It is notable that the region where few candidates produce significant modes becomes larger as the network and training set sizes increase. Among our results, mixtures with a smaller number of component modes tend to have smaller predictive variance, as discussed in Appendix A.2. This empirical observation suggests contraction of the posterior predictive as n increases. However, we also find that the range of $\mathcal{L}(\Theta)$ values widens with n, so we can expect that as n increases, the number of candidates J necessary to adequately cover parameter space will also increase.

These numerical experiments suggest that the full posterior predictive distribution often will not be well represented by an approximation that is based on a single candidate parameter value $\Theta^{(j)}$ producing low training error. Of course, our model for inference does not fully capture the predictive distribution that would be obtained with a continuous prior distribution. It is possible that if we increased J or identified candidate network parameters $\Theta^{(j)}$ with more specific structure, we would find one dominating component of the predictive distribution, or instead see a "filling in" of the predictive distribution. That is, there might be components between existing components that render the continuous predictive distribution unimodal. If such a "filling in" occurs, however, approximations based on one particularly good candidate $\Theta^{(j)}$ would still underestimate the true posterior uncertainty. This possibility opens questions of whether overparameterized BNNs can successfully "forget their priors" to learn from data, and whether a fully Bayesian model of uncertainty is suitable for producing low generalization error. In the next section, we will contrast these initial experiments with predictive distributions found by deliberately constructing inner layer parameter candidates with greater structure.

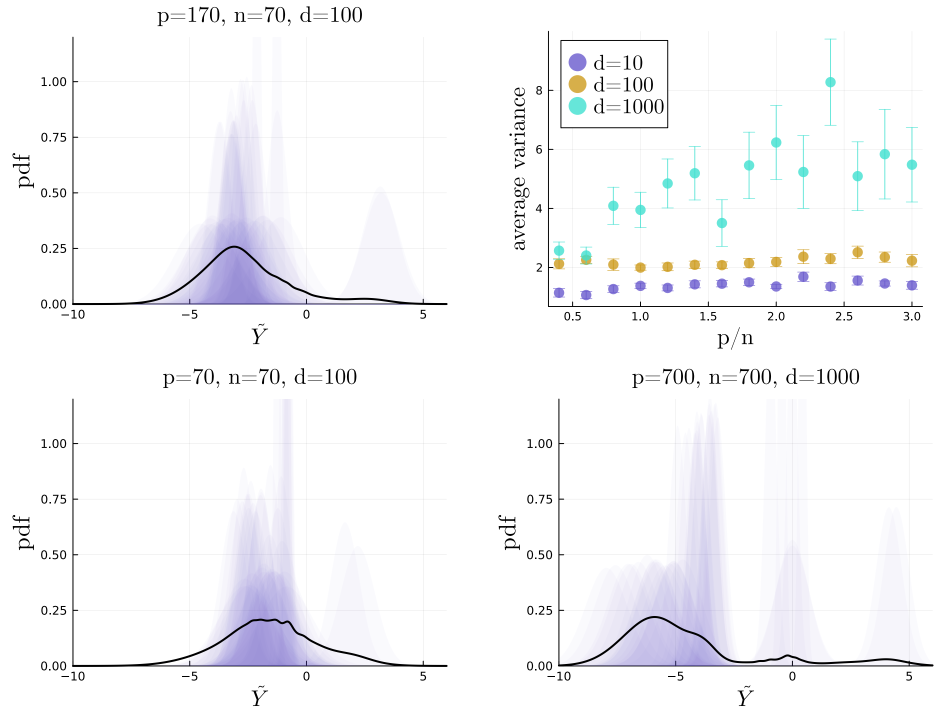

Figure 2: Top left and bottom: Predictive distributions based on candidate parameters constructed to achieve (11). The full distribution is plotted in black and components are shaded according to their weight in indigo. We consider 10 rotations, 10 preimage samples, and 10 column space samples to construct the distribution — a total of 1000 samples. Top right: The scale of predictive distributions for select d and p/n where $n/d = 0.7$. We plot the mean and standard error obtained from 10 realizations of Y for which we find the median predictive variance across 10 realizations of $\widetilde{x}_1$.

4 Constructing optimal parameters

As discussed above, candidate network parameters produced by drawing a finite set of samples from a Gaussian prior might omit a parameter value that would qualitatively change the behavior of the posterior predictive. To address this limitation, we can use our observation that the weight of each component of the posterior predictive distribution (5) depends on the marginal likelihood of the corresponding candidate (7). Now, we identify a set of candidates $\Theta^{(j)}$ which have high marginal likelihood by construction, and show that a prior which puts mass on these candidates produces a multimodal predictive distribution. A first step toward identifying these candidates is to consider the upper bound

$$\max_{\Theta} \frac{1}{n} \log \mathcal{L}(X_L(\Theta);\ X_1, Y) \ \le\ \max_{X_L} \frac{1}{n} \log \mathcal{L}(X_L;\ X_1, Y). \tag{9}$$

As detailed in Appendix A.3, the matrix X_L^* solving the optimization problem on the right satisfies

$$X_L^{*\top} X_L^* \ =\ YY^\top (1 - \gamma^2 (Y^\top Y)^{-1}), \tag{10}$$

The existence of one or more Θ^* that map to this optimal X_L depends on the choice of activation function. If we consider ReLU activation and assume that bias parameters are 0, then all elements of $X_L(\Theta^*)$ must be nonnegative. If the elements of Y are drawn from a centered distribution with unit variance, the probability that all elements of (10) are nonnegative is vanishingly small. Since $X_L^\top X_L$ is an estimate of the covariance of Y, we conjecture that if we add the constraint that its elements must be nonnegative to the right hand side of (9), we obtain

$$X_L^{*\top} X_L^* \ \approx\ \sigma(YY^\top)(1 - \gamma^2 (Y^\top Y)^{-1}). \tag{11}$$

Note that we require $n \le d_{L-1}$ to be gauranteed a solution Θ^* which maps to the right hand side above. We test (11) empirically in Appendix A.4.

We may now consider an equivalence class $[\Theta]_{\mathcal{L}}$ of network parameters Θ that map to $X_L^{*\top} X_L^*$ as defined in (11). All elements of this class will have *identical* training error and marginal likelihood, $\mathcal{L}(\Theta)$. The parameters of ReLU networks are both permutation and scale invariant [24]; thus, multiple realizations of (Θ, w) map to both identical training error and identical test predictions. But it is also possible to construct $\{\Theta^{(j)}\}_{j=1}^J$ that map to the same training error without relying on permutation and scale invariance. Specifically, we can consider unitary rotations of X_L which preserve the nonegativity of its elements, samples of the preimage of the activation function, and samples of the column space of X_{L-1}. Using these constructions, we will demonstrate that not all elements of $[\Theta]_{\mathcal{L}}$ correspond to identical predictive distributions.

Figure 2 describes posterior predictive distributions for a two layer network with a prior on the inner layer parameters that puts mass on elements of $[\Theta]_{\mathcal{L}}$. For a comparison of this prior to a continuous Gaussian prior, see Appendix A.5. We consider the combinatorial space of parameter candidates created from 10 rotations of X_2^*, 10 samples of the preimage of ReLU, and 10 samples of the column space of X_1. The top left subfigure shows the posterior predictive distribution for the same network size and $\widetilde{x}_1$ considered in the lower left corner of Figure 1. As in Section 3, we find examples of different candidates $\Theta^{(j)}$ which map to distinct predictive distributions. In contrast to Section 3, we also find this behavior for networks where $n = p$, as shown by plots on the bottom row of Figure 1. These results suggest that for many BNNs, the true Bayesian uncertainty of the posterior predictive distribution will be influenced by multiple modes of the posterior, moderated by the prior.

To understand the influence of the prior, we must examine predictive variance. Large variance indicates that the training data is not sufficient to distinguish between the parameter candidates given weight by the prior. Too much prior influence may produce poor generalization. The proportional asymptotics limit—where $n, p, d \to \infty$ while the ratios between these values remain fixed and finite—has been an important setting for examining the generalization of two layer neural networks [19, 12]. Significantly, when we consider parameter candidates from $[\Theta]_{\mathcal{L}}$, the posterior predictive distribution does *not* contract as n, d, and p increase proportionally. One example is shown in the bottom row of Figure 1, and more examples are available in Appendix A.5. The top right plot in Figure 2 summarizes the scale of the posterior predictive distributions for selected network sizes and $n/d = 0.7$. The distributions do not tend to contract as n grows; rather, for large d there may be an increase in variance as the number of last layer parameters increases. The latter behavior may occur because the number of network parameters is $p(d + 1)$, so the degree of overparameterization is increasing with p/n and d. Further exploration of the impact of the prior on predictive variance and generalization will be a focus of our future work.

5 Key implications

We have provided insight into the predictive uncertainty of Bayesian neural networks by choosing a continuous Gaussian prior for the final layer weights and a discrete prior for the interior parameters. The key implications are:

- **Much of the mass of the posterior predictive distribution can be captured without sampling the entire parameter space.** For a given prior, we can construct parameter candidates with high marginal likelihood and prior weight.
- **Unimodal posterior approximations are overconfident.** Multiple posterior modes contribute to the posterior predictive uncertainty of Bayesian neural networks.
- **The posterior predictive distribution does not contract as n, p, and d increase proportionally.** Thus, in overparameterized networks, predictive uncertainty likely reflects an inability to completely forget the prior given the training data—that is, an inability to make confident predictions.

Future work will target each of these implications. More extensive numerical experiments alongside theoretical results will consider different prior assumptions and establish minimum rates at which network size must grow with respect to training set size such that the predictive distribution does not contract. Further, we will characterize equivalence classes of parameters which map to large marginal likelihood for more network and training set sizes. Finally, we will quantify the impact of predictive uncertainty on generalization error.

References

[1] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. *ArXiv*, abs/2008.06786, 2020. URL `https://api.semanticscholar.org/CorpusID:221082525`.

[2] Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: Review and debates. *ArXiv*, abs/2309.16314, 2023. URL `https://api.semanticscholar.org/CorpusID:263134168`.

[3] David Barber and Charles M. Bishop. Ensemble learning in bayesian neural networks. 1998. URL `https://api.semanticscholar.org/CorpusID:14932413`.

[4] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1903070116`.

[6] Lucas Andry Clarte, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. On double-descent in uncertainty quantification in overparametrized models. volume 206, page 7089–7125. PMLR Proceedings of Machine Learning Research, 2023. URL `https://infoscience.epfl.ch/handle/20.500.14299/197303`.

[7] Alp Kucukelbir David M. Blei and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080/01621459.2017.1285773`.

[8] Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antoran, and Jose Miguel Hernandez-Lobato. Bayesian deep learning via subnetwork inference. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2510–2521. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/daxberger21a.html`.

[9] John S. Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'90, page 853–859, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601848.

[10] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8803–8812, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, 56(Suppl 1):1513–1589, jul 2023. ISSN 0269-2821. doi: 10.1007/s10462-023-10562-9. URL `https://doi.org/10.1007/s10462-023-10562-9`.

[12] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124013, dec 2021. doi: 10.1088/1742-5468/ac3ae6. URL `https://dx.doi.org/10.1088/1742-5468/ac3ae6`.

[13] Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13, New York, NY, USA, 1993. Association for Computing Machinery. ISBN 0897916115. doi: 10.1145/168304.168306. URL `https://doi.org/10.1145/168304.168306`.

[14] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[16] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[17] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

[18] David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 05 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.415. URL https://doi.org/10.1162/neco.1992.4.3.415.

[19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. doi: https://doi.org/10.1002/cpa.22008. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008.

[20] Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, New York, 1996. doi: https://doi.org/10.1007/978-1-4612-0745-0.

[21] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2015. URL https://arxiv.org/abs/1412.6614.

[22] Tomaso Poggio, Qianli Liao, Brando Miranda, Andrzej Banburski, Xavier Boix, and Jack Hidary. Theory iiib: Generalization in deep networks, 2018. URL https://arxiv.org/abs/1806.11379.

[23] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[24] Tommy Rochussen. Structured partial stochasticity in bayesian neural networks, 2024. URL https://arxiv.org/abs/2405.17666.

[25] Simone Rossi, Ankit Singh, and Thomas Hannagan. On permutation symmetries in bayesian neural network posteriors: a variational perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[26] Mrinank Sharma, Sebastian Farquhar, Eric Nalisnick, and Tom Rainforth. Do bayesian neural networks need to be fully stochastic? In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7694–7722. PMLR, 25–27 Apr 2023.

[27] Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/wenzel20a.html.

[28] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL `https://arxiv.org/abs/1611.03530`.

[29] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL `https://doi.org/10.1145/3446776`.

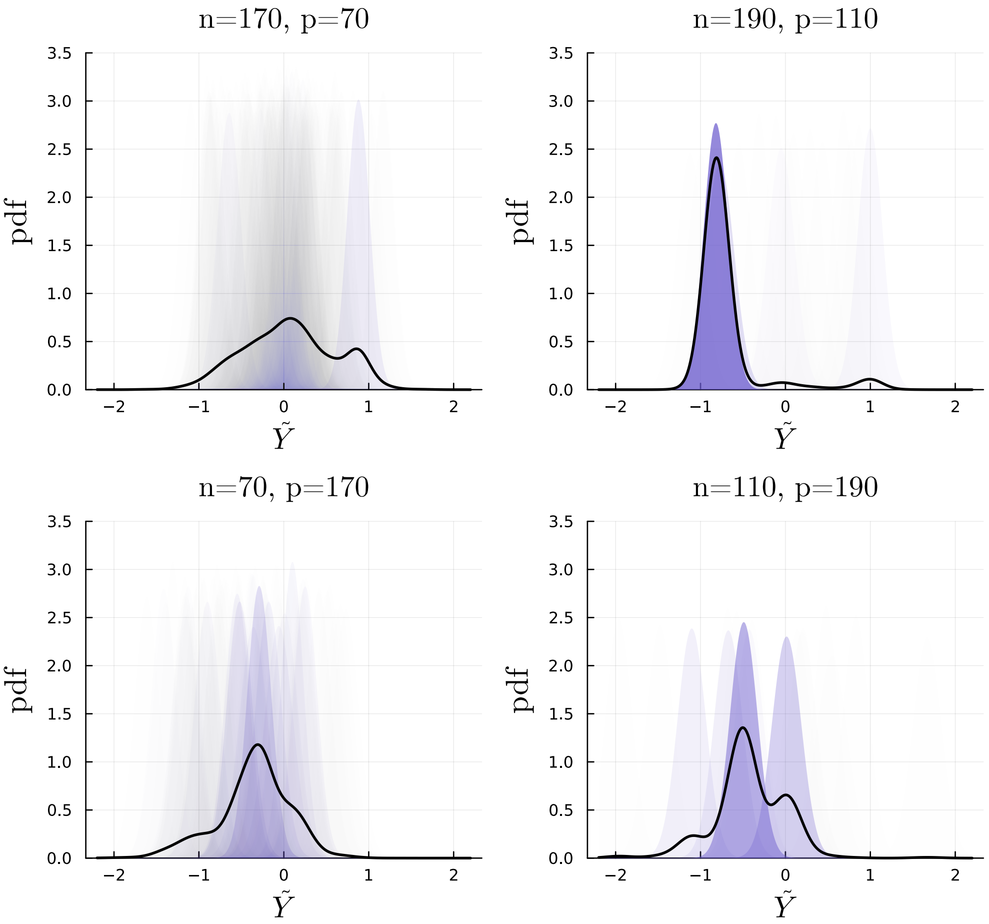

Figure 3: Posterior predictive distributions at test point $\widetilde{x}_1^{(2)}$ for input dimension $d = 100$ at select training set size n and final layer width p, as indicated by each title. The black line shows the pdf which is a mixture of Gaussians. Each shaded distribution is a component of this mixture with transparency corresponding to its weight.

A Appendix / supplemental material

A.1 PDFs under a discretized Gaussian prior

Figure 1 shows the predictive distributions for select network and training set sizes at test location $\widetilde{x}_1^{(1)}$. Here, we provide the pdfs which result from inference with the prior specified in Section 3 for the same ratios n/p at additional locations $\widetilde{x}_1^{(2)}$ and $\widetilde{x}_1^{(3)}$. For all examples, $\gamma^2 = 0.01$. Figures 3 and 4 show results for $d = 100$ while figures 5 and 6 correspond to $d = 1000$. As in Section 3, we see that multiple candidates $\Theta^{(j)}$ contribute to the posterior predictive distributions, leading to multimodality in most cases considered.

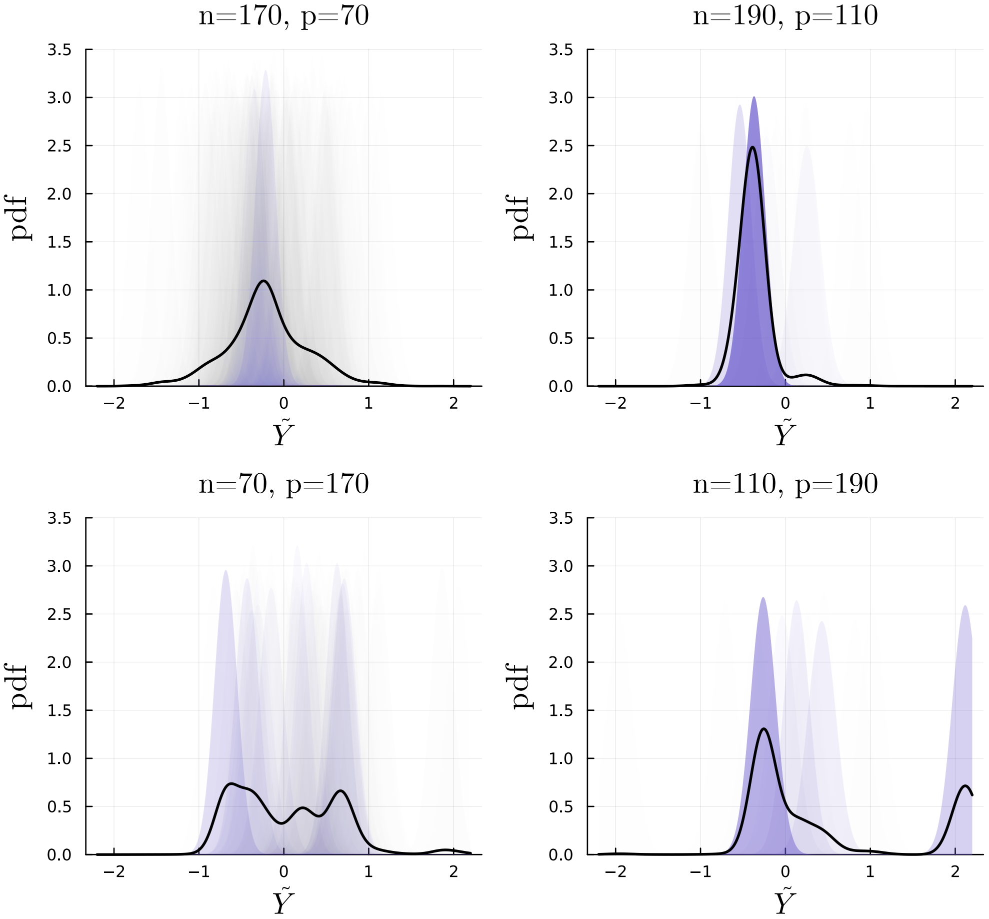

Figure 4: Posterior predictive distributions at test point $\widetilde{x}_1^{(3)}$ for input dimension $d = 100$ at select training set size n and final layer width p, as indicated by each title. The black line shows the pdf which is a mixture of Gaussians. Each shaded distribution is a component of this mixture with transparency corresponding to its weight.

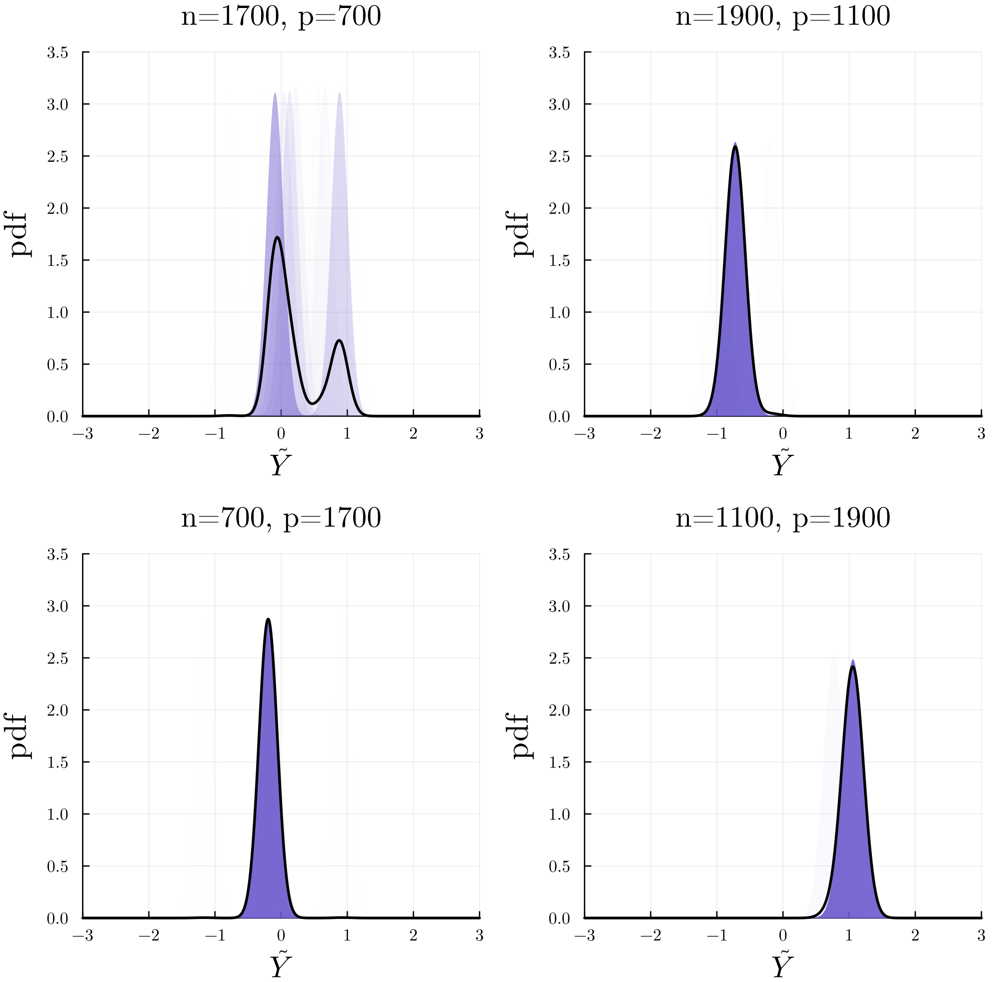

Figure 5: Posterior predictive distributions at test point $\widetilde{x}_1^{(2)}$ for input dimension $d = 1000$ at select training set size n and final layer width p, as indicated by each title. The black line shows the pdf which is a mixture of Gaussians. Each shaded distribution is a component of this mixture with transparency corresponding to its weight.

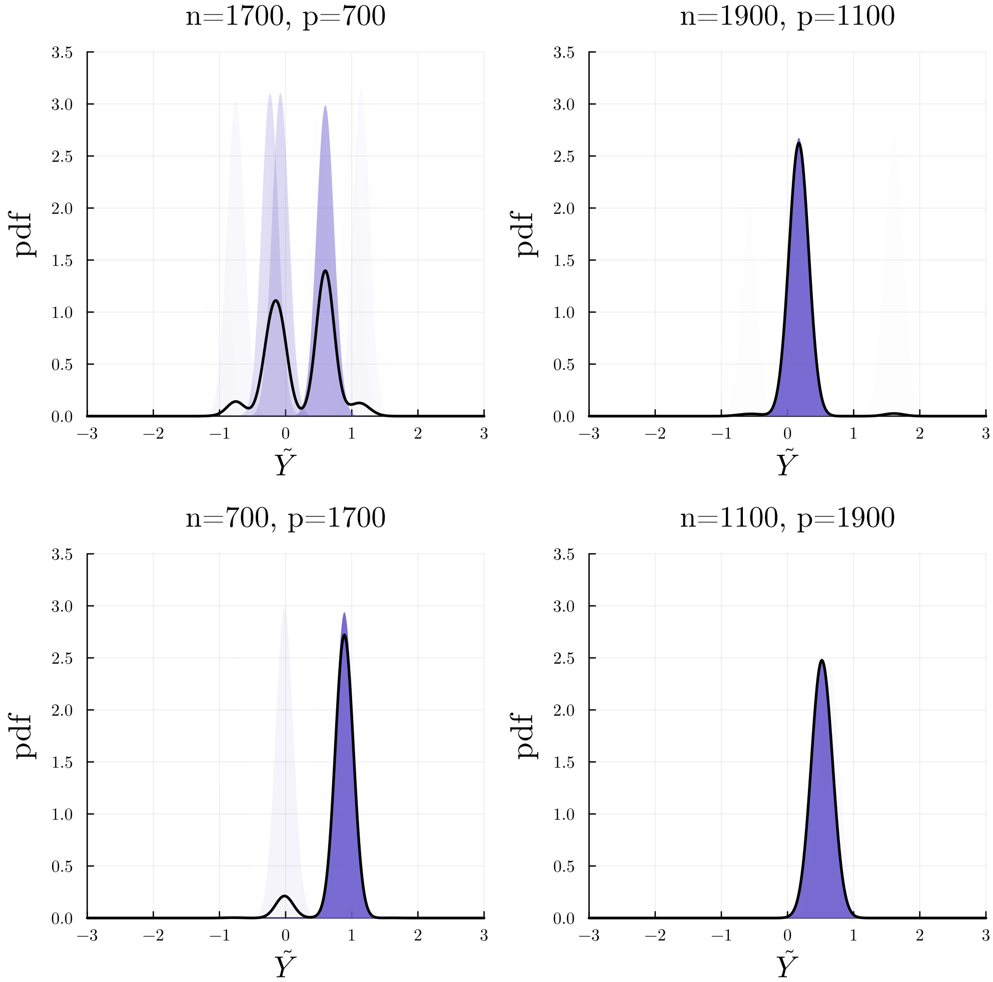

Figure 6: Posterior predictive distributions at test point $\widetilde{x}_1^{(3)}$ for input dimension $d = 1000$ at select training set size n and final layer width p, as indicated by each title. The black line shows the pdf which is a mixture of Gaussians. Each shaded distribution is a component of this mixture with transparency corresponding to its weight.

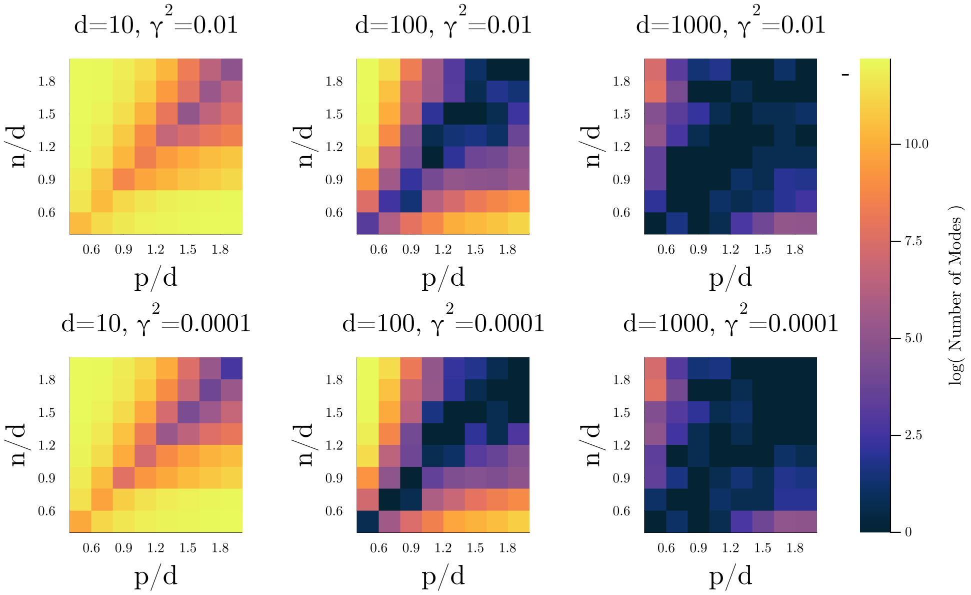

Figure 7: Heatmaps depicting the log of the number of component distributions which have weight larger than 10^{-6} for specified network dimensions. Columns correspond to the input dimension, d: 10, 100, and 1000. Rows correspond to observation variance: 0.01 and 0.0001.

A.2 Identifying modes under a discretized Gaussian prior

This section provides additional results concerning the multimodality and variance of predictive distributions described in Section 3. The first rows of both Figures 7 and 8 match the right column of Figure 1. This set of heatmaps reports the number of modes with weight larger than 10^{-6} found for specified network and training set size at observation noise level $\gamma^2 = 0.01$. They are repeated for the purpose of comparison. In Figure 7, we see that the number of modes located for a specific n, p, d triple is not impacted by reducing observation noise to $\gamma^2 = 0.0001$. In Figure 8, we can see a loose relationship between the number of significant modes and the variance of the predictive distribution. In our numerical experiments, we have found that component distributions of the predictive distribution tend to have similar variance and may have distinct modes. Thus, it is reasonable that finding more significant modes correlates with greater predictive variance, as we see. For this particular example, as we increase the training set size, predictive variance tends to decrease, but this may be an artifact of the prior choice and finite J. Figure 9 demonstrates that the predictive variance can be sensitive to the choice of J.

Figure 10 provides context for the number of modes reported in Figure 7. The heatmap shades correspond to the log of the standard deviation of the distribution on $\{n^{-1} \log \mathcal{L}(\Theta^{(j)})\}_{j=1}^{J}$. For a given input dimension, d, and observation noise level, γ^2, the largest standard deviation is found when $n = p$. This effect is likely related to the double descent phenomena: for each X_1, there is one candidate Θ which outperforms all other candidates. As expected, the double descent phenomenon becomes more pronounced as regularization, γ^2, decreases.

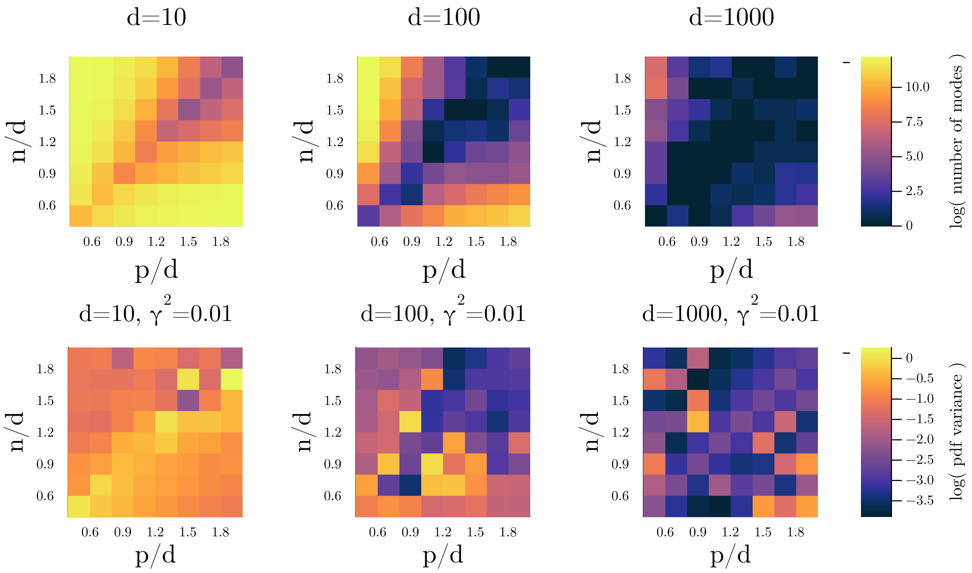

Figure 8: Top: The log of the number of component distributions which have weight larger than 10^{-6} for specified network dimensions. Bottom: The log of the variance of the posterior predictive distribution obtained for each network size. Columns correspond to the input dimension, d: 10, 100, and 1000. All results correspond to observation noise $\gamma^2 = 0.01$.

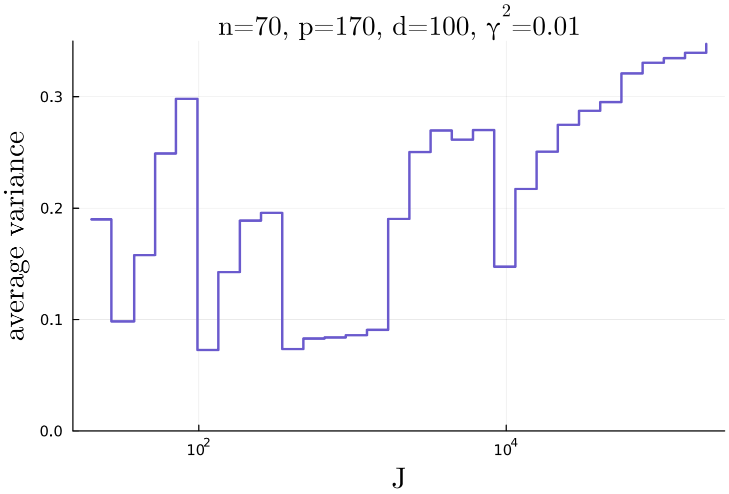

Figure 9: The variance of the posterior predictive distribution averaged over 100 realizations of $\widetilde{x}_1$ plotted against the number of parameter candidates, J. The prior details are specified in Section 3 and the network and training set size are given in the plot title.

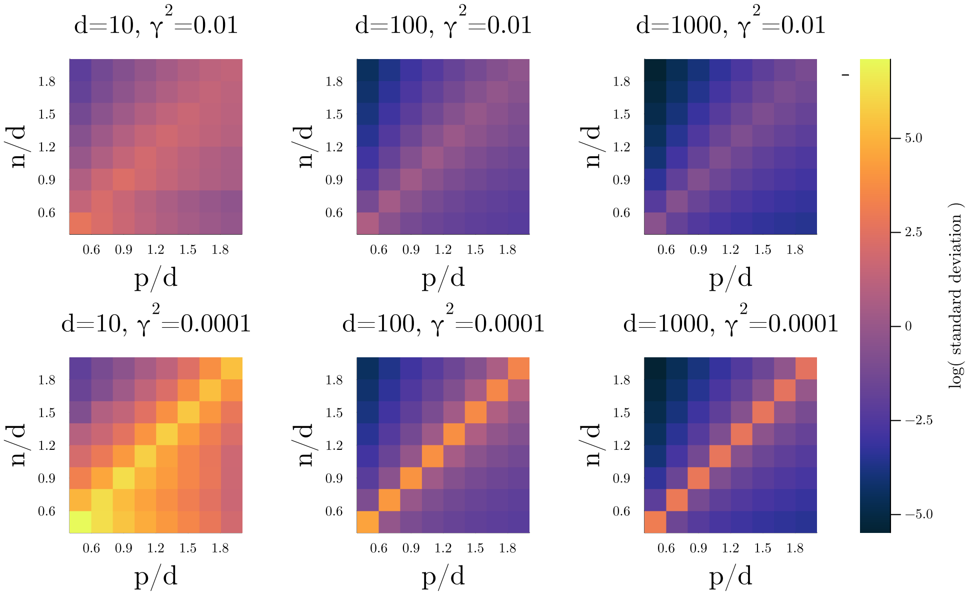

Figure 10: Heatmaps depicting the log of the standard deviation of the distribution of $n^{-1} \log \mathcal{L}(\Theta)$ for candidates $\Theta^{(j)}$ sampled from a Gaussian prior as described in Section 3. Columns correspond to the input dimension, d: 10, 100, and 1000. Rows correspond to observation variance: 0.01 and 0.0001. The diagonal where $n = p$ shows a double descent effect which is stronger for smaller observation noise.

A.3 Optimal parameters

We are interested in Θ which maximizes $\mathcal{L}(\Theta; X, Y)$ as defined in (8). To this end, consider the singular value decomposition

$$U\Lambda^{1/2}Q^\top \;=\; \frac{X_L}{\sqrt{p}}, \tag{12}$$

where $\operatorname{diag}(\Lambda) = [\lambda_1, \ldots, \lambda_{p\vee n}]^\top$ and $Q = [q_1 \ldots q_n] \;\in \mathbb{R}^{n\times n}$. Then,

$$\log \mathcal{L}\left(X_L(\Theta); X, Y\right) = \frac{-1}{2}\sum_{k=1}^{n}\left(\log(2\pi) + \log(\lambda_k + \gamma^2) + \frac{(q_k^\top Y)}{\lambda_k + \gamma^2}\right). \tag{13}$$

Note that because $X_L^\top X_L$ is positive semi-definite, $n^{-1}\log\mathcal{L}\left(X_L\right) \;\leq\; -\log(\gamma)$. We can determine that

$$\begin{aligned}
&\min_{\Theta} \frac{-2}{n}\left(\log\mathcal{L}\left(X_L(\Theta); X, Y\right) + \log(2\pi)\right)\\
\leq\;& \min_{\substack{\Lambda\succeq 0\\ Q^\top Q = QQ^\top = \mathbf{I}_n}} \frac{1}{n}\sum_{k=1}^{n}\left(\log(\lambda_k + \gamma^2) + \frac{(q_k^\top Y)}{\lambda_k + \gamma^2}\right)\\
=\;& \min_{Q^\top Q = QQ^\top = \mathbf{I}_n} \frac{1}{n}\sum_{k=1}^{n}\min_{\lambda_k\geq 0}\left(\log(\lambda_k + \gamma^2) + \frac{(q_k^\top Y)}{\lambda_k + \gamma^2}\right)\\
=\;& \min_{Q^\top Q = QQ^\top = \mathbf{I}_n} \frac{1}{n}\sum_{k=1}^{n}\begin{cases}\log\left(q_k^\top Y\right)^2 + 1 & (q_k^\top Y)^2 \geq \gamma^2,\; k\leq p\\ \log\gamma^2 + \frac{(q_k^\top Y)}{\gamma^2} & \text{otherwise}\end{cases}\\
=\;& \log\gamma^2 \;+ \min_{\substack{\{v_1\geq\cdots\geq v_n\geq 0,\\ \gamma^2\sum_{i=1}^{n} v_i = Y^\top Y\}}} \frac{1}{n}\sum_{k=1}^{n}\begin{cases}\log v_k + 1 & v_k \geq 1,\; k\leq p\\ v_k & \text{otherwise}\end{cases}.
\end{aligned}$$

In the last line, we impose the constraint $v_1 \geq \cdots \geq v_n \geq 0$ to prevent redundant optima. We find that

$$\operatorname*{arg\,min}_{X_L^\top X_L} \frac{1}{n}\sum_{k=1}^{n}\left(\log(\lambda_k + \gamma^2) + \frac{(q_k^\top Y)}{\lambda_k + \gamma^2}\right) \;=\; YY^\top\left(1 - \frac{\gamma^2}{Y^TY}.\right) \tag{14}$$

Provided that $Y^\top Y \geq \gamma^2$, this minimizer is unique. For the results reported in this work, we assume that $Y \in \mathbb{R}^n$ is centered with unit variance. Then, we expect $Y^\top Y \sim \mathcal{O}(n)$.

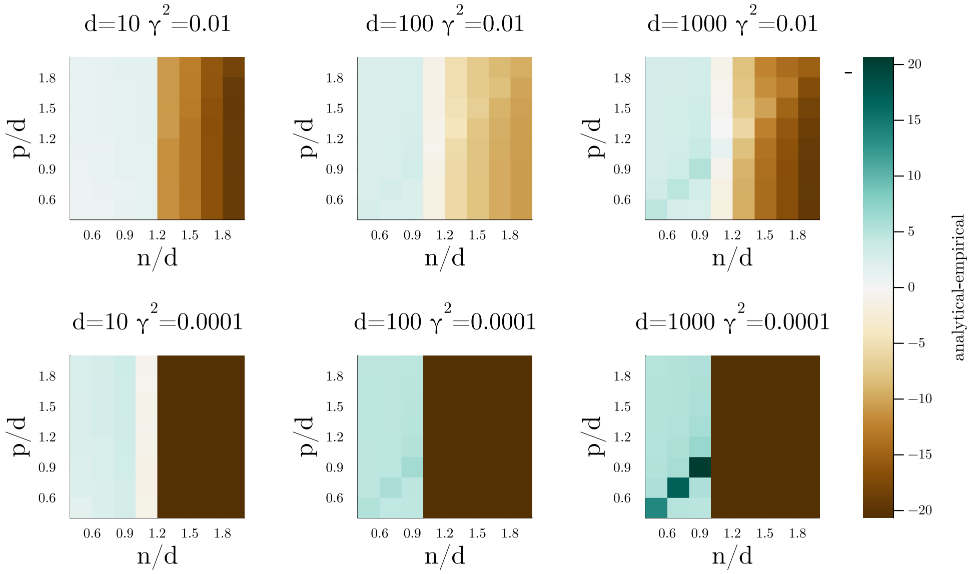

Figure 11: The difference in scaled log marginal likelihood ($n^{-1}\log\mathcal{L}$) based on Θ constructed to satisfy (15) and the best performing Θ with elements sampled iid from a Gaussian prior.

A.4 Optimal parameters for ReLU

In this section, we evaluate the conjecture made by (11). In particular, we compare the largest value of $n^{-1}\log\mathcal{L}(X_L)$ found in Section 3 to the conjectured maximum for a given set of training observations (X_1, Y). Recall that for a two-layer network, we must have $n \le d$ for there to exist some Θ which maps to the conjectured maximizer, X_L^*. Thus, for this section we consider $\mathcal{L}\left(X_L(\Theta)\right)$ such that

$$X_L^\top X_L \;=\; \sigma(P_X YY^\top P_X^\top)\left(1 - \frac{\gamma^2}{Y^\top Y}\right) \tag{15}$$

where P_X projects into the column space of X_1. Thus, when $n \le d$, (15) reduces to (11). Note that we do not necessarily expect $n^{-1}\log\mathcal{L}(X_L)$ under (15) to be optimal when $n > d$.

Figure 11 shows the difference between optimal $n^{-1}\log\mathcal{L}(X_L)$ under (15) and the maximum $n^{-1}\log\mathcal{L}(X_L)$ found empirically in Section 3. We consider $d \in \{10, 100, 1000\}$, $\gamma^2 \in \{0.01, 0.0001\}$, and ratios p/d and n/d ranging from 0.5 to 2. As expected, we find that the conjectured optimum is at least as large as the empirically determined maximum for $n \le d$. In cases where $d < n$, enforcing (15) leads to $n^{-1}\log\mathcal{L}(X_L)$ considerably smaller than our empirically located maximum. It is interesting to note that the distance by which the conjectured maximum outperforms the empirical maximum is exacerbated when d is large and $n = p$.

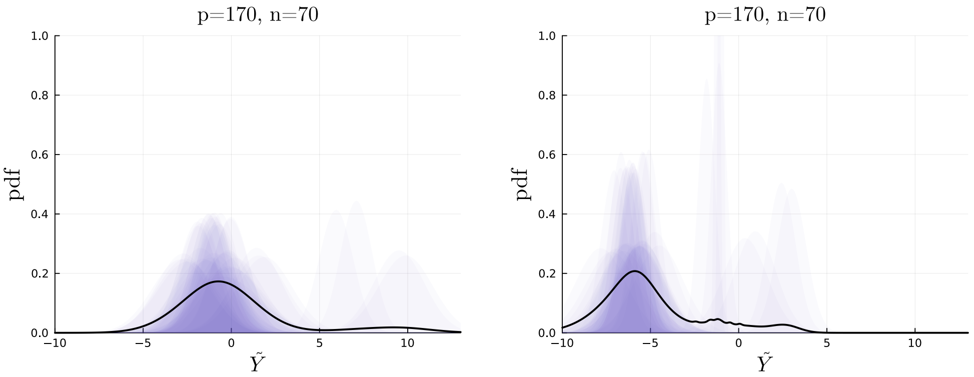

Figure 12: Predictive distribution based on candidate parameters constructed to achieve (11). The full distribution is plotted in black and components are shaded according to their weight in indigo. We consider 10 rotations, 10 preimage samples, and 10 column space samples to construct the distribution. Results are for $n = 70$, $p = 170$, and $d = 100$ at locations $\widetilde{x}_1^{(2)}$ (left) and $\widetilde{x}_1^{(3)}$ (right).

A.5 Predictive distribution for optimal parameters

Here, we summarize additional results from the setting of Section 4. The top left subfigure of Figure 2 shows the predictive distributions for $n = 70$ and $p = 170$ at test location $\widetilde{x}_1^{(1)}$. Here, we provide the predictive distribution at additional locations $\widetilde{x}_1^{(2)}$ and $\widetilde{x}_1^{(3)}$ in Figure 12. For all examples, $\gamma^2 = 0.01$. The candidates $\Theta^{(j)}$ for these plots are constructed from the combination of 10 rotations of X_L, 10 samples of the column space of X_1, and 10 samples of the preimage space. Thus, we have a total of 1000 candidates. To better separate the impact of each approach to constructing candidates, Figure 13 shows predictive pdfs where each column corresponds to a different approach. For instance, in the first column, candidates are constructed based on 10 rotations of X_L, one sample of the column space of X_1, and one sample of the preimage space. Each row corresponds to a different test location: $\widetilde{x}_1^{(2)}$ and $\widetilde{x}_1^{(3)}$. We see that all approaches see to contribute to predictive variance, but rotation and column space samples seem to yield more distinct modes than preimage samples.

Figure 8 provides some evidence that under the setting of Section 3, as n increases, the variance of the predictive distribution decreases, even if p and d increase in proportion to n. The reduction in variance is observed in the region where n is close to p, and occurs in part because we tend to find unimodal predictive distributions in this region when we finitely many sample parameter candidates from a Gaussian distribution. It is possible that this shrinkage is an artifact of the experimental design as it seems unlikely we would see a reduction in uncertainty when the degree of overparameterization, $n/(pd)$, increases. Figure 2 provides some evidence that when the prior puts weight on certain "optimal" parameters, this shrinkage does not occur. Figure 12 provides representative examples of predictive pdfs obtained following the setting of Section 4 when $n = p$. We see that there is no evidence of shrinkage as n increases, and most examples demonstrate multimodality.

Finally, it is worth examining the distribution of the components of the constructed parameter candidates, $\Theta^{(j)}$, and their corresponding final layer weights, w. If these constructed parameters are far outside typically used priors, the predictive modes they produce would not be informative about the behaviors of BNNs in practice. Figure 15 provides a representative comparison between the distribution of constructed "optimal" parameters (indigo) to the distribution of the parameters drawn from the prior distributions considered in Section 3 (gold). We see that the distributions are close though for both Θ and w, the variance of the distributions on the constructed parameters is slightly wider.

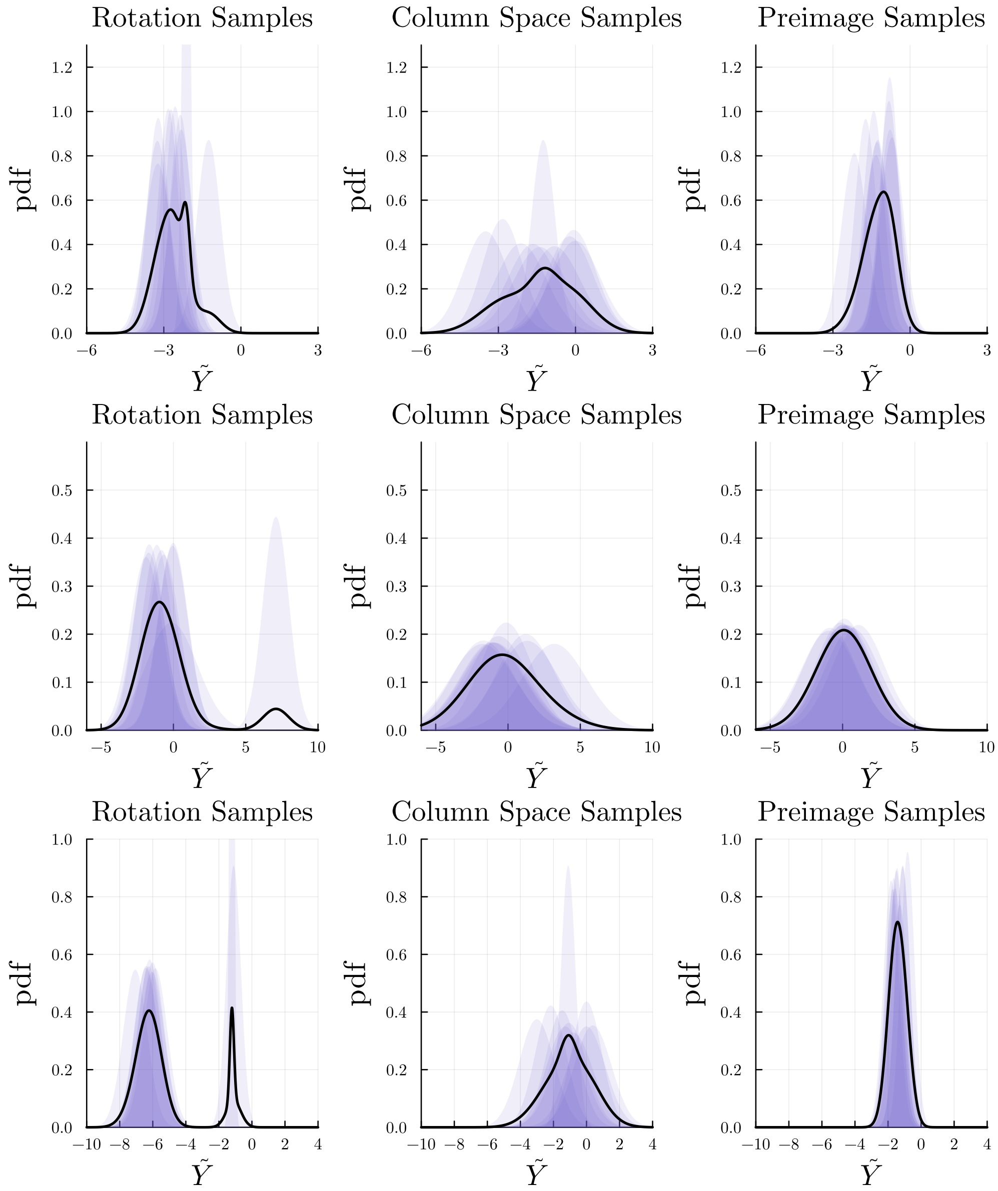

Figure 13: Predictive distributions based on 10 parameter candidates differentiated by the method of their construction: rotation (left), sampling the column space (center), sampling the preimage (right). For all results, $n = 70$, $p = 170$, and $d = 100$. Each row corresponds to a different test location: $\widetilde{x}_1^{(1)}$, $\widetilde{x}_1^{(2)}$, and $\widetilde{x}_1^{(3)}$.

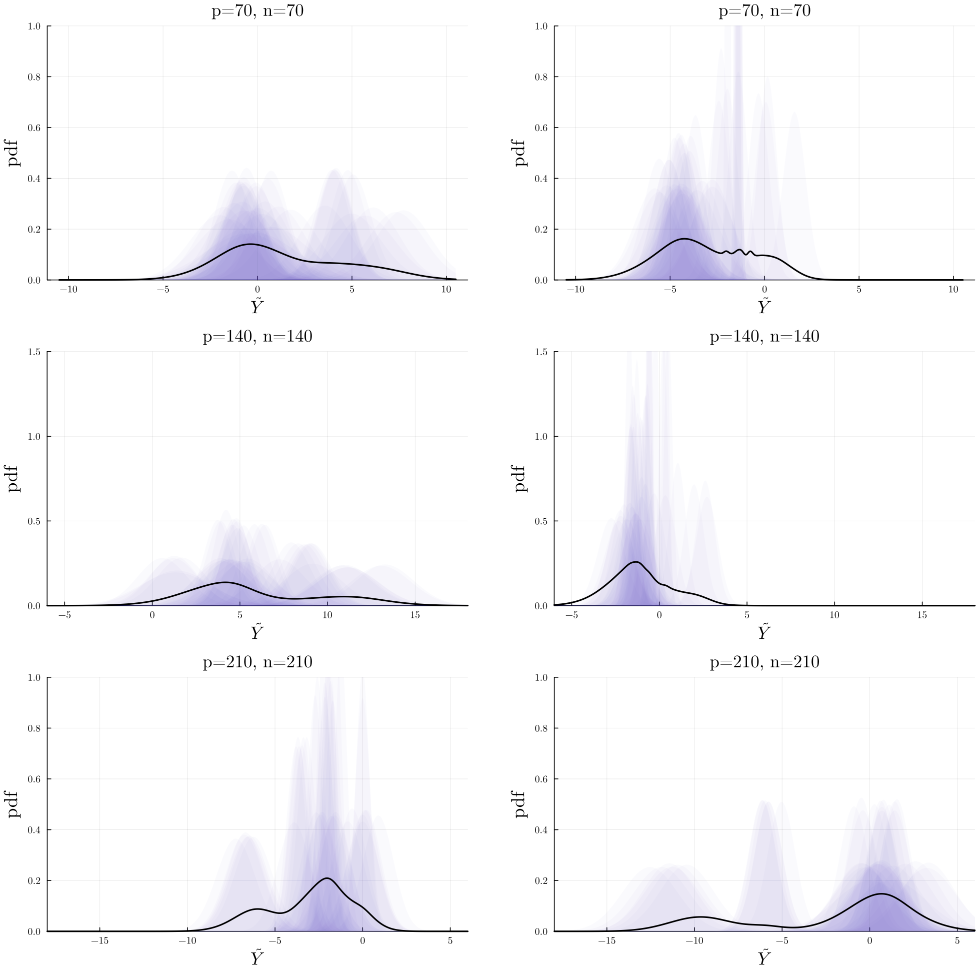

Figure 14: Predictive distributions for select n and p based on candidate parameters constructed to achieve (11). The full distribution is plotted in black and components are shaded according to their weight in indigo. We consider 10 rotations, 10 preimage samples, and 10 column space samples to construct the distribution. Each row corresponds to a different input dimension; from top to bottom, we consider $d \in \{100, 200, 300\}$. Each column corresponds to a test location: $\widetilde{x}_1^{(2)}$ (left) and $\widetilde{x}_1^{(3)}$ (right).

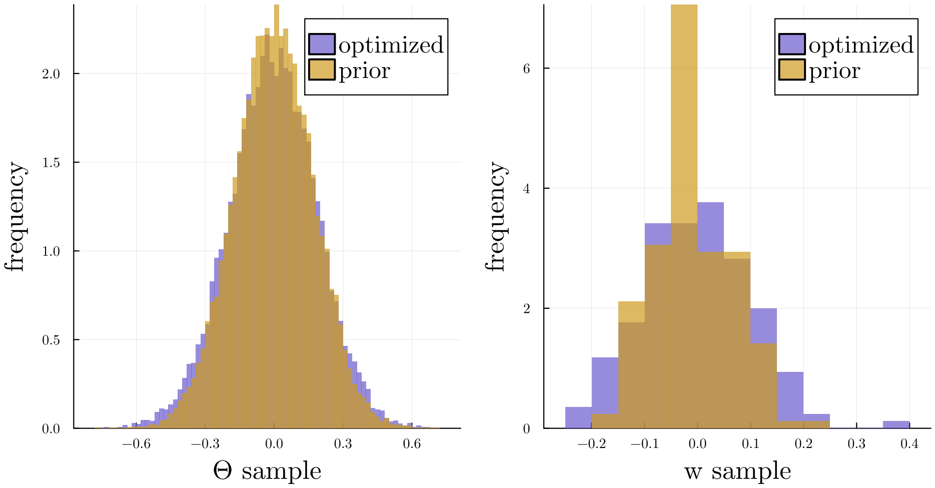

Figure 15: Comparison between the distribution of representative parameters constructed as described in Section 4 (indigo) and parameters sampled from the prior used in Section 3 (gold). The left plot shows interior parameters, Θ, while the right plot shows final layer parameters, w.