Bayesian Neural Network Priors Revisited

Anonymous Author(s)
Affiliation
Address
email

Abstract

Isotropic Gaussian priors are the *de facto* standard for modern Bayesian neural network inference. However, there has been recent controversy over the question whether they might be to blame for the undesirable *cold posterior effect*. We study this question empirically and find that for densely connected networks, Gaussian priors are indeed less well suited than more heavy-tailed ones. Conversely, for convolutional architectures, Gaussian priors seem to perform well and thus cannot fully explain the cold posterior effect. These findings coincide with the empirical maximum-likelihood weight distributions discovered by standard gradient-based training.

1 Introduction

In Bayesian deep learning, we are interested in performing inference on the posterior over parameters $p(w|x, y) = p(y|w, x)p(w)/p(y|x)$. In building such a model, choosing a prior $p(w)$ that faithfully reflects our true beliefs is crucial for obtaining a meaningful posterior [Gelman et al., 2013]. Especially when working with heavily overparameterized models, such as neural networks, this becomes challenging due to the opaque interactions between parameters and predictions.

While most practitioners use isotropic Gaussian priors for their Bayesian neural networks (BNNs), there has been some debate over whether these priors are well suited for the task. One side argues that Gaussian priors are ill-suited and are responsible for the unsatisfactory performance of BNNs in practice, as well as for the undesirable *cold posterior effect* [Wenzel et al., 2020]. The other side maintains that any reasonably uninformative prior should be sufficient, when combined with the right neural network architecture and enough data [Wilson and Izmailov, 2020].

In this work, we make an attempt to settle this dispute empirically. We analyse the performance of different BNN priors for different network architectures and compare them to the empirical weight distributions of maximum-likelihood solutions to the learning problems. We conclude that isotropic priors with heavier tails than the Gaussian are better suited for dense NNs (DNNs), but the case of convolutional NNs (CNNs) is still unclear. Thus, we recommend to use a Student-t prior for Bayesian inference in DNNs, while we cannot recommend against using Gaussian priors for CNNs. The cold posterior effect remains elusive.

1.1 Contributions

Our main contributions are

- An analysis of the empirical weight distributions of trained neural networks with different architectures, suggesting that feedforward networks learn more non-Gaussian weights than convolutional networks (Sec. 4.1).

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

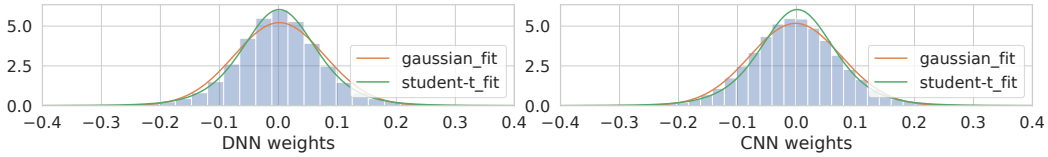


Figure 1: Empirical weight distribution of DNNs and CNNs trained with SGD on MNIST. It can be seen that heavy-tailed distributions yield a better fit than Gaussians for the DNNs, but Gaussians fit better for the CNNs. For the DNN weights, the Shapiro-Wilk test [Shapiro and Wilk, 1965] gives a p-value of 10^{-21}, for the CNN weights of 10^{-13}.

- An empirical study of Bayesian neural network performance with feedforward network architectures, suggesting that heavy-tailed priors can outperform the widely-used Gaussian priors (Sec. 4.2).
- An empirical study of Bayesian neural network performance with convolutional network architectures, suggesting that Gaussian priors perform at least as well as alternative ones and can thus not fully explain the cold posterior effect in such models (Sec. 4.3).

2 Priors for Bayesian Neural Networks

In this study, we contrast the widely used Gaussian priors with more heavy-tailed priors, especially Laplace and Student-t distributions. We chose these distributions mostly based on our observations regarding the empirical weight distributions of trained networks (see Fig. 1 and Sec. 4) and for their ease of implementation and optimization. We now give a quick overview over these different distributions and their most salient properties.

Gaussian The Gaussian distribution [Gauss, 1809] is the *de-facto* standard for BNN priors in recent work [Wenzel et al., 2020, Wilson and Izmailov, 2020, Zhang et al., 2019]. Its probability density function (PDF) is

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

It is attractive, because it is the central limit of all finite-variance distributions [Billingsley, 1961] and the maximum entropy distribution for a given mean and scale [Bishop, 2006]. However, its tails are relatively light compared to some of the other distributions that we will consider.

Laplace The Laplace distribution [Laplace, 1774] has heavier tails than the Gaussian and is discontinuous at $x = \mu$. Its PDF is

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

It is often used in the context of lasso regression, where it encourages sparsity in the learned weights [Tibshirani, 1996].

Student-t The Student-t distribution characterizes the mean of a finite number of samples from a Gaussian distribution with respect to the true mean [Student, 1908]. Its PDF is

$$p(x; \mu, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where Γ is the gamma function and ν are the degrees of freedom. The Student-t also arises as the marginal distribution over Gaussians with an inverse-Gamma prior over the variances [Helmert, 1875, Lüroth, 1876]. For $\nu \to \infty$, the Student-t distribution approaches the Gaussian. For any finite ν it has heavier tails than the Gaussian. Its k'th moment is only finite for $\nu > k$. The ν parameter thus offers a convenient way to adjust the heaviness of the tails. Unless otherwise stated, we set $\nu = 3$ in our experiments, such that the distribution has rather heavy tails, while still having a finite mean and variance.

3 Evaluation Metrics

When using BNNs, practitioners might care about different outcomes. In some applications, the predictive accuracy might be the only metric of interest, while in other applications calibrated uncertainty estimates could be crucial. We therefore use a range of different metrics in our experiments in order to highlight the respective strengths and weaknesses of different priors. Moreover, we compare the priors to the empirical weight distributions of conventionally trained networks.

3.1 Empirical weight distributions

When training a standard neural network using stochastic gradient descent (SGD), we achieve a maximum-likelihood (ML) solution to the learning task. Intuitively, the more a BNN prior deviates from the empirical weight distribution reached by this training, the more the posterior will deviate from the ML solution. This can be useful as a regularization in the case where the ML solution is overfitting, but it can also lead to a worse overall performance. Since BNNs are generally rather known for underfitting [Neal, 1995, Wenzel et al., 2020, Dusenberry et al., 2020], we hypothesize that a prior should work better if it allows the posterior to approach the ML solution more closely. In our experiments, we compare the distributions based on the weight values, the weight correlations, and the singular value spectra of the weight matrices.

3.2 Empirical test performance

Test error The test error is probably the most widely used metric in supervised learning. It intuitively measures the performance of the model on a held-out test set and is often seen as an empirical approximation to the true generalization error. While it is often used for model selection, it comes with the risk of overfitting to the used test set [Bishop, 2006] and in the case of BNNs also fails to account for the predictive variance of the posterior.

Test log-likelihood The predictive log-likelihood also requires a test set for its evaluation, but it takes the predictive posterior variance into account. It can thus offer a built-in tradeoff between the mean fit and the quality of the uncertainty estimates. Moreover, it is a proper scoring rule [Gneiting and Raftery, 2007].

3.3 Uncertainty estimates

Uncertainty calibration Bayesian methods are often chosen for their superior uncertainty estimates, so many users of BNNs will not be satisfied with only fitting the posterior mean well. The calibration measures how well the uncertainty estimates of the model correlate with predictive performance. Intuitively, when the model is for instance 70 % certain about a prediction, this prediction should be correct with 70 % probability. Many deep learning models are not well calibrated, because they are often overconfident and assign too low uncertainties to their predictions [Ovadia et al., 2019]. When the models are supposed to be used in safety-critical scenarios, it is often crucial to be able to tell when they encounter an input that they are not certain about [Kendall and Gal, 2017]. For these applications, metrics such as the expected calibration error [Naeini et al., 2015] might be the most important criteria.

Out-of-distribution detection The out-of-distribution (OOD) detection measures how well one can tell in-distribution and out-of-distribution examples apart based on the uncertainties. This is important when we believe that the model might be deployed under some degree of data set shift. In this case, the model should be able to detect these OOD examples and be able to reject them, that is, refuse to make a prediction on them.

3.4 Cold posterior effect

When performing inference in Bayesian models, a tempered version of the posterior can be written as

$$\log p(w|x, y) \propto \frac{1}{T}[\log p(y|w, x) + \log p(w)] \tag{1}$$

for some positive temperature T. With $T = 1$, this reduces to the standard Bayesian posterior.

In their recent work, Wenzel et al. [2020] have drawn attention to the effect that when the posterior is being cooled down in BNNs (i.e., setting $T \ll 1$), the performance of the models often increases. Since previous work has shown that tempered posteriors can be useful under misspecified priors [Grünwald et al., 2017], they conclude that the BNN priors (which were Gaussian in their experiments) must be misspecified. Reversing this argument, we can hypothesize that choosing better priors for BNNs should lead to a less pronounced cold posterior effect, which we can use to evaluate different candidate priors.

4 Experiments

We performed experiments on the MNIST handwritten digit data set [LeCun et al., 1998]. We compare densely connected neural networks (DNNs) and convolutional neural networks (CNNs) on this task.

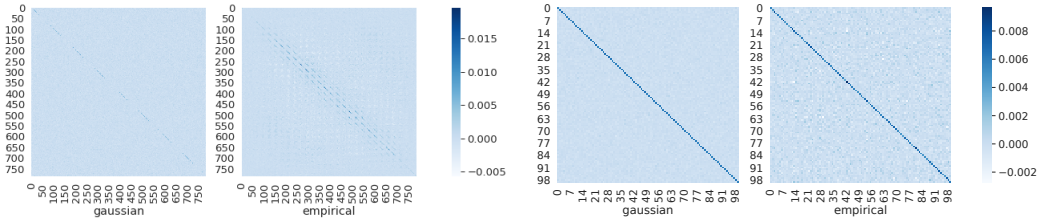
4.1 Weight distributions of DNNs and CNNs

As described in Section 3, we trained the DNNs and CNNs with SGD on the task and analysed the empirical weight distributions. To explicitly contrast the empirical weights with the Gaussian distributional assumption of many prior works, we compare the distributions to the same number of weights sampled from a Gaussian with the same mean and variance as the empirical distribution.

Note that for ease of exposition, the figures in this section are just showing the results for a single layer out of each network. However, the observations readily extend to the other layers as well.

4.1.1 DNN weights

We can see that the weight values of the DNNs follow a more heavy-tailed distribution than a Gaussian (Fig. 1). Moreover, in contrast to the Gaussian, the empirical weight distributions show some significant correlations (Fig. 2). The empirical covariance matrices also have a more heavy-tailed distribution of off-diagonal elements, which further supports this observation (Fig. 3).



(a) Input convariance (b) Output convariance

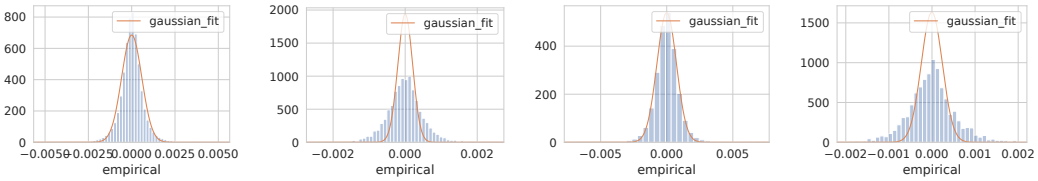
Figure 2: Empirical covariances of the weights of DNNs trained with SGD on MNIST. We see that they contain more systematic correlations than the isotropic Gaussian.

Finally, the singular value spectra for the empirical weight matrices fall off more sharply than for the Gaussian weights (Fig. 4), which follow the ideal singular value distribution more closely [Shen, 2001]. This rapid decay suggests that the weights are well approximated by a low-rank matrix.

4.1.2 CNN weights

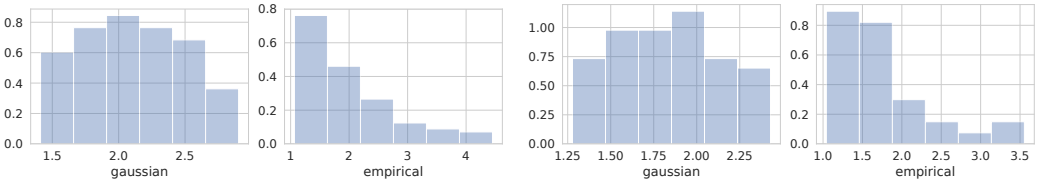
In contrast to the DNN results, when analysing the weight distributions of CNNs, they actually follow empirical marginal distributions that are close to being Gaussian (Fig. 1). However, their weights do show some correlation, though less striking than the correlation for DNNs (Fig. 5). The off-diagonal elements of the empirical covariance matrices seem to also be more heavy-tailed than for isotropic Gaussian weights (Fig. 3).

Finally, the singular value spectra for the empirical CNN weight matrices also decay more quickly than the isotropic Gaussian ones (Fig. 4). This suggests that better priors could be found when explicitly taking the correlation structure into account, but they should still have Gaussian marginals.



(a) DNN, input (b) DNN, output (c) CNN, input (d) CNN, output

Figure 3: Distributions of off-diagonal elements in the empirical covariances of the weights of DNNs and CNNs trained with SGD on MNIST. The empirical distributions are plotted as histograms, while the idealized random Gaussian weights are overlaid in orange. We see that the covariances of the empirical weights are more heavy-tailed than for the Gaussian weights.



(a) DNN weights (b) CNN weights

Figure 4: Distributions of singular values of the weights of DNNs and CNNs trained with SGD on MNIST. We see that the spectra of the empirical weights fall off more sharply than for the Gaussian weights.

We hypothesize that multivariate Gaussians with non-diagonal covariance matrices could be good candidates for CNN priors, but we leave their exploration to future work.

4.2 Bayesian DNN performance with different priors

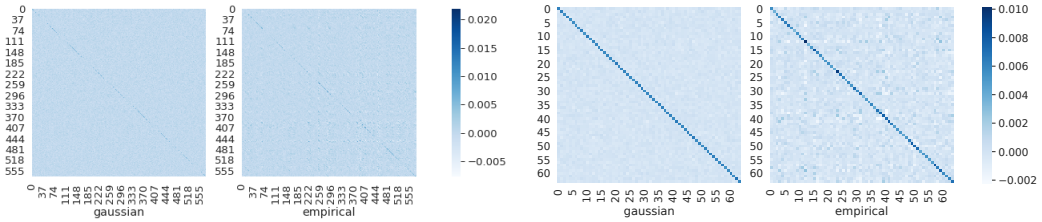
Following our observations from the empirical weight distributions, we can hypothesize that heavy-tailed priors should work better than Gaussian ones for Bayesian DNNs. We tested this by performing BNN inference with the same network architecture as above and different priors (see Sec. 2). We trained the networks on the MNIST training set using cyclical SGLD [Zhang et al., 2019, Wenzel et al., 2020] and report the predictive error and log likelihood on the MNIST test set. We follow Ovadia et al. [2019] in reporting the calibration of the uncertainty estimates on rotated MNIST digits and the OOD detection accuracy on FashionMNIST [Xiao et al., 2017].

We observe that the heavy-tailed priors do indeed outperform the Gaussian one in terms of the different metrics, except for the calibration error (Fig. 6). This suggests that Gaussian priors are indeed too restrictive for feedforward networks on this task and inhibit the posterior from including high-likelihood solutions, such as the ones analysed above (Sec. 4.1). Moreover, we find that the cold posterior effect is less pronounced when using heavy-tailed priors (Fig. 7).

4.3 Bayesian CNN performance with different priors

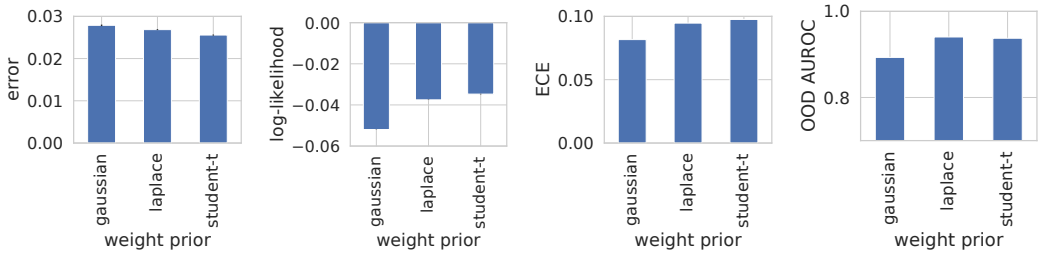
We repeated the same experiment for Bayesian CNNs. Following our observations from the empirical weights, in this case we would not expect the heavy-tailed priors to outperform the Gaussian one. Indeed, this is also what we observe (Fig. 8). However, interestingly, we see that the cold posterior effect is still less pronounced with heavy-tailed priors (Fig. 9).

Moreover, we observe in this experiment as well as the previous one that the cold posterior effect as described by Wenzel et al. [2020] is most pronounced with respect to predictive error and log likelihood. When comparing the uncertainty measures (calibration and OOD detection), we actually observe the opposite effect, namely that the true Bayesian posterior ($T = 1$) performs better then the cold ones. It will be an interesting avenue for future work to investigate the connection between different performance measures and the cold posterior effect in more depth.



(a) Input covariance (b) Output covariance

Figure 5: Empirical covariances of the weights of CNNs trained with SGD on MNIST. We see that they also contain more systematic correlations than the isotropic Gaussian.



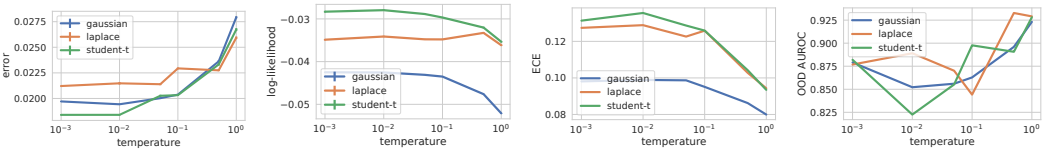
(a) Error (b) Log likelihood (c) Calibration (d) OOD detection

Figure 6: Performance of Bayesian DNNs with different priors on MNIST in terms of different metrics. We see that the heavy-tailed priors perform better than the Gaussian ones, except in terms of expected calibration.

5 Related Work

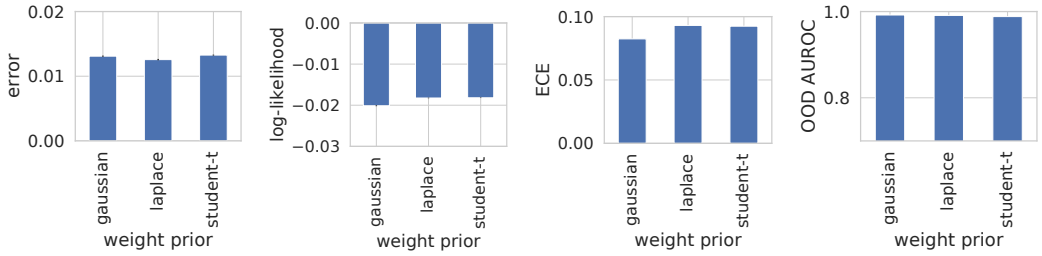
Since the inception of Bayesian neural networks, scholars have thought about choosing good priors for them, including hierarchical [MacKay, 1992] and heavy-tailed ones [Neal, 1995]. In the context of infinite-width limits of such networks [Lee et al., 2017, Matthews et al., 2018, Garriga-Alonso et al., 2018, Yang, 2019, Tsuchida et al., 2019] it has also been shown that heavy-tailed priors can lead to more interesting properties than finite-variance ones [Neal, 1995, Peluchetti et al., 2020]. Nonetheless, most state-of-the-art BNN methods these days still use simple isotropic Gaussian priors [Osawa et al., 2019, Zhang et al., 2019, Maddox et al., 2019, Wilson and Izmailov, 2020, Dusenberry et al., 2020]. It has been hypothesized that this could be one of the reasons why BNNs are still not convincingly outperforming standard neural networks on many tasks [Wenzel et al., 2020].

While many interesting distributions have been proposed as variational posteriors for BNNs [Louizos and Welling, 2017, Farquhar et al., 2019, Swiatkowski et al., 2020, Dusenberry et al., 2020], these approaches have still relied on simple Gaussian priors. Although a few different priors have been proposed for BNNs, these were mostly designed for specific tasks [Atanov et al., 2018, Ghosh and Doshi-Velez, 2017, Overweg et al., 2019, Cui et al., 2020, Hafner et al., 2020] or relied heavily on



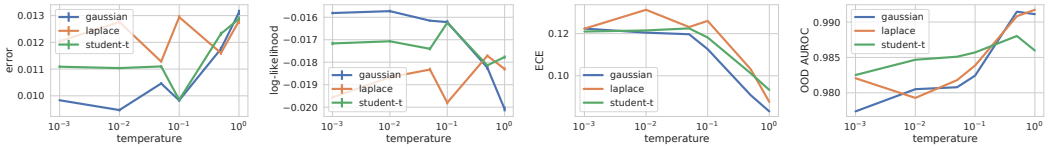
(a) Error (b) Log likelihood (c) Calibration (d) OOD detection

Figure 7: Cold posterior effect of Bayesian DNNs with different priors on MNIST in terms of different metrics. We see that the heavy-tailed priors generally lead to a less pronounced cold posterior effect than the Gaussian ones.



(a) Error (b) Log likelihood (c) Calibration (d) OOD detection

Figure 8: Performance of Bayesian CNNs with different priors on MNIST in terms of different metrics. We see that the Gaussian prior performs as well as the other ones.



(a) Error (b) Log likelihood (c) Calibration (d) OOD detection

Figure 9: Cold posterior effect of Bayesian CNNs with different priors on MNIST in terms of different metrics. We see that the heavy-tailed priors generally lead to a less pronounced cold posterior effect than the Gaussian ones.

non-standard inference methods [Sun et al., 2019, Ma et al., 2019, Karaletsos and Bui, 2020, Pearce et al., 2020].

In this work we explicitly study the question of whether non-Gaussian priors are generally useful (or even necessary) for Bayesian neural networks. We make an attempt to answer this question for different neural network architectures and also compare the priors to the empirical distributions of weights in networks trained via maximum likelihood. While the empirical weight distributions of standard neural networks have been studied before [Bellido and Fiesler, 1993, Go and Lee, 1999], these works have not systematically compared different architectures and did not use their insights to inform Bayesian prior choices. Moreover, we used reliable Markov Chain Monte Carlo (MCMC) BNN inference [Neal, 1992, Zhang et al., 2019, Wenzel et al., 2020] in our experiments to be able to make claims about the true posteriors of the models. This is in contrast to many of the works mentioned above, which used variational inference approaches.

6 Conclusion

We have shown that especially in densely connected BNNs, heavy-tailed non-Gaussian priors can yield a better performance across many metrics and also fit the empricial weight distributions of maximum-likelihood solutions better. We would thus recommend to use them instead of the widely-used Gaussian priors. However, in convolutional BNNs, Gaussian priors seem to be well adjusted to the empirical marginal distributions and perform at least as well as more heavy-tailed ones.

Extending our analyses to non-isotropic priors that match the empirical covariances of the maximum-likelihood weights more closely will be an exciting avenue for future work. Moreover, it will be interesting to relate these findings more thoroughly to the cold posterior effect with respect to different performance metrics.

References

Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitry Vetrov, and Max Welling. The deep weight prior. *arXiv preprint arXiv:1810.06943*, 2018.

I Bellido and Emile Fiesler. Do backpropagation trained neural networks have normal weight distributions? In *International Conference on Artificial Neural Networks*, pages 772–775. Springer, 1993.

Patrick Billingsley. The Lindeberg-Lévy theorem for martingales. *Proceedings of the American Mathematical Society*, 12(5):788–792, 1961.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Tianyu Cui, A. Havulinna, P. Marttinen, and S. Kaski. Informative gaussian scale mixture priors for bayesian neural networks. *arXiv preprint arXiv:2002.10243*, 2020.

Michael W Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-an Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. *arXiv preprint arXiv:2005.07186*, 2020.

Sebastian Farquhar, Michael Osborne, and Yarin Gal. Radial bayesian neural networks: Robust variational inference in big models. *arXiv preprint arXiv:1907.00865*, 2019.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Carl Friedrich Gauss. *Theoria motvs corporvm coelestivm in sectionibvs conicis solem ambientivm*. Sumtibus F. Perthes et IH Besser, 1809.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Soumya Ghosh and Finale Doshi-Velez. Model selection in bayesian neural networks via horseshoe priors. *arXiv preprint arXiv:1705.10388*, 2017.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Jinwook Go and Chulhee Lee. Analyzing weight distribution of neural networks. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 2, pages 1154–1157. IEEE, 1999.

Peter Grünwald, Thijs Van Ommen, et al. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. In *Uncertainty in Artificial Intelligence*, pages 905–914. PMLR, 2020.

Friedrich Robert Helmert. Über die berechnung des wahrscheinlichen fehlers aus einer endlichen anzahl wahrer beobachtungsfehler. *Z. Math. U. Physik*, 20(1875):300–303, 1875.

Theofanis Karaletsos and Thang D Bui. Hierarchical gaussian process priors for bayesian neural network weights. *arXiv preprint arXiv:2002.04033*, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

Pierre Simon Laplace. Mémoire sur la probabilité de causes par les évenements. *Memoire de l'Academie Royale des Sciences*, 1774.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. *arXiv preprint arXiv:1703.01961*, 2017.

J Lüroth. Vergleichung von zwei werthen des wahrscheinlichen fehlers. *Astronomische Nachrichten*, 87:209, 1876.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, pages 4222–4233, 2019.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13153–13164, 2019.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 2901. NIH Public Access, 2015.

Radford M Neal. Bayesian training of backpropagation networks by the hybrid monte carlo method. Technical report, Citeseer, 1992.

Radford M Neal. Bayesian leaning for neural networks, 1995.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. In *Advances in neural information processing systems*, pages 4287–4299, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13991–14002, 2019.

Hiske Overweg, Anna-Lena Popkes, Ari Ercole, Yingzhen Li, José Miguel Hernández-Lobato, Yordan Zaykov, and Cheng Zhang. Interpretable outcome prediction with sparse bayesian neural networks in intensive care. *arXiv preprint arXiv:1905.02599*, 2019.

Tim Pearce, Russell Tsuchida, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Expressive priors in bayesian neural networks: Kernel combinations and periodic functions. In *Uncertainty in Artificial Intelligence*, pages 134–144. PMLR, 2020.

Stefano Peluchetti, Stefano Favaro, and Sandra Fortini. Stable behaviour of infinitely wide deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 1137–1146. PMLR, 2020.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Jianhong Shen. On the singular values of gaussian random matrices. *Linear Algebra and its Applications*, 326(1-3):1–14, 2001.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779*, 2019.

Jakub Swiatkowski, Kevin Roth, Bastiaan S Veeling, Linh Tran, Joshua V Dillon, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in bayesian neural networks. *arXiv preprint arXiv:2002.02655*, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Russell Tsuchida, Fred Roosta, and Marcus Gallagher. Richer priors for infinitely wide multi-layer perceptrons. *arXiv preprint arXiv:1911.12927*, 2019.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świ atkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *Advances in Neural Information Processing Systems*, pages 9951–9960, 2019.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.