3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–11

On the Inconsistency of Bayesian Inference for Misspecified Neural Networks

Yijie Zhang PHILIP.YIJIE.ZHANG@GMAIL.COM
Eric Nalisnick E.T.NALISNICK@UVA.NL
University of Amsterdam

Abstract

Grünwald and Van Ommen (2017) show that Bayesian inference for linear regression can be inconsistent under model misspecification. In this paper, we extend their analysis to Bayesian neural networks (BNNs), investigating if they too can be inconsistent under misspecification. We find that BNNs exhibit the same inconsistency when Hamiltonian Monte Carlo is used for posterior inference. However, variational inference changes this behavior. Surprisingly, we find that variational Bayes leads to BNNs that *are consistent* in the setting studied by Grünwald and Van Ommen (2017). We conjecture that the success of variational Bayes is due to its optimization objective: the evidence lower bound (ELBO) implicitly encourages the posterior approximation to concentrate, mitigating the ill-effects of the misspecification.

1. Introduction

Neural networks (NNs) are well known to exhibit pathological behaviors such as overconfidence (Nguyen et al., 2015). Bayesian inference is often pointed to as a way to ameliorate these pathologies by allowing for principled uncertainty quantification. The only apparent obstacle is performing scalable approximate inference, and consequently this topic has received much attention (Graves, 2011; Blundell et al., 2015; Hernández-Lobato and Adams, 2015; Osawa et al., 2019). However, the story is not so simple. Grünwald and Van Ommen (2017) show that Bayesian inference can perform poorly—specifically, be inconsistent—for lightly misspecified linear models (homoskedastic model, heteroskedastic data). In this paper, we ask the same question of Bayesian NNs: are they also inconsistent in the setting studied by Grünwald and Van Ommen (2017)?

We find that Bayesian NNs are *sometimes* inconsistent in Grünwald and Van Ommen (2017)'s regression setting. Their behavior depends on which posterior inference strategy is employed. Under strong posterior inference algorithms such as Hamiltonian Monte Carlo, BNNs are observed to be similarly inconsistent. However, under variational Bayes (Blei et al., 2017), we observe that BNNs *are consistent* (in this particular setting). We conjecture that the success of variational Bayes is due to the evidence lower bound (ELBO). Our explanation is supported empirically by the fact that the Laplace approximation—which uses the same approximating family but a different optimization objective (MAP vs ELBO)—results in inconsistency.

2. Preliminaries

Below we describe the setting studied by Grünwald and Van Ommen (2017), summarizing their methods and findings. We then extend their work to Bayesian NNs in the next section.

2.1. Simulated Data

Consider a regression problem with features $\mathrm{x} \in \mathbb{R}$ and responses $\mathrm{y} \in \mathbb{R}$. The training data is denoted $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}\} = \{(x_n, y_n)\}_{n=1}^N$, with $(\mathrm{x}_n, \mathrm{y}_n) \sim p^*(\mathrm{x}, \mathrm{y})$ (independently and identically distributed). The ground truth density is denoted $p^*(\mathrm{x}, \mathrm{y}) = p^*(\mathrm{y})\ p^*(\mathrm{x})$. Following Grünwald and Van Ommen (2017), we consider two settings: *homoskedastic* vs *heteroskedastic* response noise. In the homoskedastic setting, let $p^*(\mathrm{x}) = \mathrm{unif}(-1, 1)$ and $p^*(\mathrm{y}) = \mathcal{N}(0, 0.025)$. In the heteroskedastic setting, we first flip a fair coin. If the coin lands heads, then $\mathrm{x} \sim \mathrm{unif}(-1, 1)$ and $\mathrm{y} \sim \mathcal{N}(0, 0.05)$. If the coin lands tails, then we set $(\mathrm{y}, \mathrm{x}) = (0, 0)$. In both settings, y has the same marginal variance, but in the heteroskedastic setting, the presence of the $(0, 0)$ points changes the variance at the origin. As there is no trend between x and y in either setting, the best predictive model in both cases is simply the straight line at $y = 0$. Grünwald and Van Ommen (2017) expand the scalar input x to a vector $\mathbf{x}$ via a 50-dimensional polynomial basis, defining the linear model on that representation of x.

2.2. Model Misspecification

Consider a family of models $\{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Model misspecification refers to when the set of all candidate models (denoted $\boldsymbol{\Theta}$) does not contain the ground truth model, i.e. $p^*(\mathbf{y}|\mathbf{x}) \notin \{p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Grünwald and Van Ommen (2017) consider a form of misspecification where the model assumes constant variance in $\mathbf{x}$, while the variance of the true density $p^*(\mathbf{y}|\mathbf{x})$ is not constant. For the regression setting above, this means that the model assumes homoskedastic response noise but the true noise is heteroskedastic. Grünwald and Van Ommen (2017) analyze this setting for linear models, finding that the Bayesian posterior is not able to concentrate to the best predictive solution $\boldsymbol{\beta}^* = \mathbf{0}$ (the straight line at $y = 0$) despite the Normal prior having its mode at $\boldsymbol{\beta} = \mathbf{0}$.

2.3. Quantities of Interest

Predictive Performance We are concerned with the large-sample predictive performance of our Bayesian regression model. In turn, two quantities of interest are the *log-risk* and *square-risk* (Grünwald and Van Ommen, 2017). The first is defined as:

$$\mathrm{RISK}^{\log}(\boldsymbol{\theta}|\mathcal{D}) = \mathbb{E}_{(\mathbf{x}^*, \mathbf{y}^*) \sim p^*} \left[-\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}} \left[p_{\boldsymbol{\theta}}(\mathbf{y}^*|\mathbf{x}^*) \right] \right], \tag{1}$$

and the latter is then defined as:

$$\mathrm{RISK}^{\mathrm{sq}}(\boldsymbol{\theta}|\mathcal{D}) = \mathbb{E}_{(\mathbf{x}^*, \mathbf{y}^*) \sim p^*} \left[\mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}} \left[\left(\mathbf{y}^* - f_{\boldsymbol{\theta}}(\mathbf{x}^*) \right)^2 \right] \right], \tag{2}$$

where $f_{\boldsymbol{\theta}}$ is the mean of our parametric density model $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ and $(\mathbf{x}^*, \mathbf{y}^*)$ denotes held-out data sampled from the true generative process.

Hypercompression Let $\tilde{\boldsymbol{\theta}}$ denote the parameter setting that results in the closest model to p^* in terms of Kullback–Leibler (KL) divergence: $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KL}\left(p^*(\mathbf{y}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\right)$. Grünwald and Van Ommen (2017) define *hypercompression* to occur when the following quantity is negative:

$$\underbrace{\mathbb{E}_{(\mathbf{x}^*,\mathbf{y}^*)\sim p^*}\left[-\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}\left[p_{\boldsymbol{\theta}}(\mathbf{y}^*|\mathbf{x}^*)\right]\right]}_{\mathrm{RISK}^{\log}(\boldsymbol{\theta}|\mathcal{D})} - \mathbb{E}_{(\mathbf{x}^*,\mathbf{y}^*)\sim p^*}\left[-\log p_{\tilde{\boldsymbol{\theta}}}(\mathbf{y}^*|\mathbf{x}^*)\right] < 0, \tag{3}$$

meaning that the posterior predictive is a better model of p^* than the single best model $p_{\tilde{\boldsymbol{\theta}}}$. While this may seem to be a good thing—the boons of model uncertainty—Grünwald and Van Ommen (2017) argue that the presence of hypercompression implies posterior inconsistency. To elaborate, hypercompression arises when the posterior is not concentrating around the KL-best model but the posterior predictive has very good log-risk. Since the KL-best model does not have high probability, the good risk values do not stem from modeling the function well, misleading us to think that the model is better than it really is.

Posterior Concentration To measure the degree of posterior concentration, Grünwald and Van Ommen (2017) compare the difference between the *R-log-risk* and *log-risk*:

$$0 \leq \mathcal{C}\left(\boldsymbol{\theta}|\mathcal{D}\right) = \underbrace{\mathbb{E}_{(\mathbf{x}^*,\mathbf{y}^*)\sim p^*}\mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}\left[-\log p_{\boldsymbol{\theta}}(\mathbf{y}^*|\mathbf{x}^*)\right]}_{\mathrm{RISK}^{\text{R-log}}(\boldsymbol{\theta}|\mathcal{D})} - \underbrace{\mathbb{E}_{(\mathbf{x}^*,\mathbf{y}^*)\sim p^*}\left[-\log \mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}\left[p_{\boldsymbol{\theta}}(\mathbf{y}^*|\mathbf{x}^*)\right]\right]}_{\mathrm{RISK}^{\log}(\boldsymbol{\theta}|\mathcal{D})}. \tag{4}$$

The sensibility of $\mathcal{C}(\boldsymbol{\theta}|\mathcal{D})$ follows directly from Jensen's inequality. $\mathcal{C}(\boldsymbol{\theta}|\mathcal{D}) = 0$ if and only if the posterior concentrates to a point mass $\delta(\boldsymbol{\theta}_0)$. Or intuitively, the R-log-risk can be interpreted as measuring the predictive performance under samples from the posterior. On the other hand, the log-risk quantifies the predictive performance of the aggregated model space represented by the posterior predictive distribution. When the posterior is concentrated, then $\mathcal{C}(\boldsymbol{\theta}|\mathcal{D}) \approx 0$ as it should not matter if we make predictions using a particular sample or the aggregate. Note that while $\mathcal{C}(\boldsymbol{\theta}|\mathcal{D})$ can measure concentration, it is ambivalent to *where* the posterior is concentrating. For well-identified models, it would be worthwhile to examine their convergence to $\tilde{\boldsymbol{\theta}}$, but for NNs, $\tilde{\boldsymbol{\theta}}$ is not identifiable. In Appendix Figure 9, we provide convergence results of $\boldsymbol{\theta}$ to $\tilde{\boldsymbol{\theta}}$ for Bayeisan linear models, where $\tilde{\boldsymbol{\theta}}$ is uniquely identifiable.

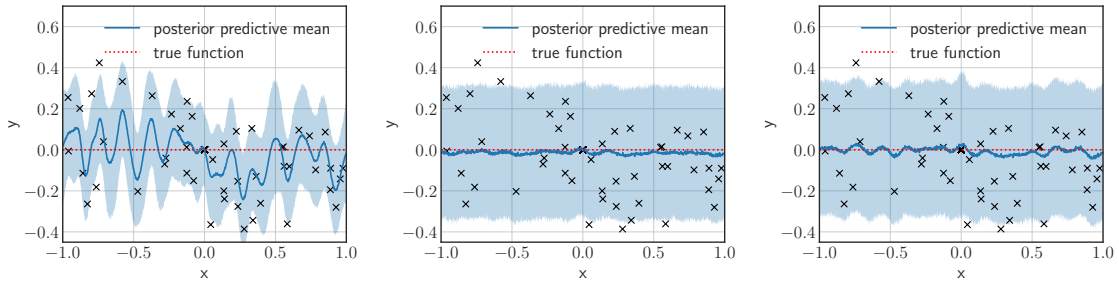
3. Investigation of Misspecified Bayesian Neural Networks

We now turn to Bayesian NNs, examining if they demonstrate the same inconsistency that Grünwald and Van Ommen (2017) identified for linear models. The full implementation details are provided in Appendix C.

Bayesian NN Regression Model Following Grünwald and Van Ommen (2017)'s linear model specification as closely as possible, we study the NN regression model:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}), \qquad \sigma^2 \sim \Gamma^{-1}(\alpha_0, \beta_0), \qquad \mathrm{y} \sim \mathcal{N}\left(f_{\boldsymbol{\beta}}(\mathbf{x}), \sigma^2\right) \tag{5}$$

where $f_{\boldsymbol{\beta}}(\mathbf{x})$ is a feedforward NN with one hidden layer and 20 hidden units with ReLU activations. Following Grünwald and Van Ommen (2017), we set the inverse-gamma hyper-prior's shape as $\alpha = 1.0$ and scale as $\beta = 0.025$. Let $\boldsymbol{\beta}$ denote all of the weights. The full model is denoted with the conditional density $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})$ where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$.



(a) HMC (b) Mean Field VB (c) Full-Covariance VB

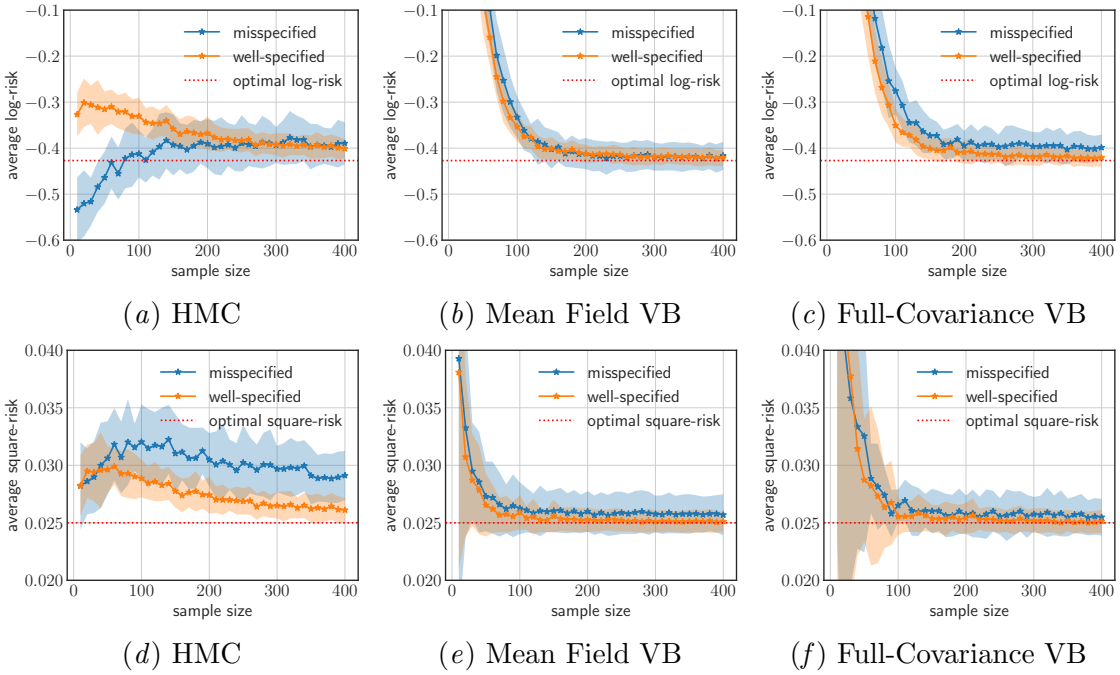
Figure 1: *Misspecified Model Fits* ($N = 100$). The above plots show the posterior predictive distributions for the Bayesian NN when Hamiltonian Monte Carlo (a), mean field (Normal) variational Bayes (b), and full-covariance (Normal) variational Bayes (c) are used for posterior inference. The shaded areas correspond to 95% prediction intervals.

Data and Misspecification We use the same data simulation process as described in Section 2.1.[1] The Bayesian NN described in Equation 5 is correctly specified in the case of homoskedastic noise and misspecified in the case of heteroskedastic noise. Yet, even in the misspecified case there exists the best available conditional density $p_{\tilde{\theta}}(\mathbf{y}|\mathbf{x}) = \mathcal{N}(0, 0.025)$ that is closest to the true density $p^*(\mathbf{y}|\mathbf{x})$ in terms of KL divergence. While we know the best conditional, the corresponding parameters $\tilde{\boldsymbol{\theta}}$ are not identifiable as many configurations can lead to this conditional (e.g. all-zero first layer, all-zero second layer, all-ReLU units evaluate to their inactive regime, etc.).

3.1. Hamiltonian Monte Carlo

We first examine our Bayesian NN's performance when the posterior is obtained by Hamiltonian Monte Carlo (HMC). Firstly, Figure 1a shows the posterior predictive distribution for the misspecified setting. We see that the predictive mean is wiggly and similar to the fit of the Bayesian linear model shown in Figure 5a. Clearly, the posterior did not concentrate on the best predictive function, which confirms that the inconsistency identified by Grünwald and Van Ommen (2017) can occur for Bayesian NNs as well. Figures 2a and 2d show the predictive performance in terms of log-risk and square-risk respectively, comparing the well-specified setting (orange) to the misspecified setting (blue). While the misspecified model has a better or at least competitive log-risk, its square-risk is clearly inferior, suggesting that hypercompression has likely occurred. Figure 3a confirms that hypercompression indeed takes place for $N < 200$, suggesting that Grünwald and Van Ommen (2017)'s explanation for linear models extends to NNs as well. Lastly, Figure 4a shows that the misspecified model's posterior is clearly less concentrated than the corresponding well-specified model.

1. We do make one slight change from Grünwald and Van Ommen (2017): instead of a 50-dimensional polynomial basis, we instead use a 101-dimensional Fourier basis, following Heide et al. (2020). This choice was just to improve visualization of the model fits and changes the analysis in no other way.



(a) HMC (b) Mean Field VB (c) Full-Covariance VB

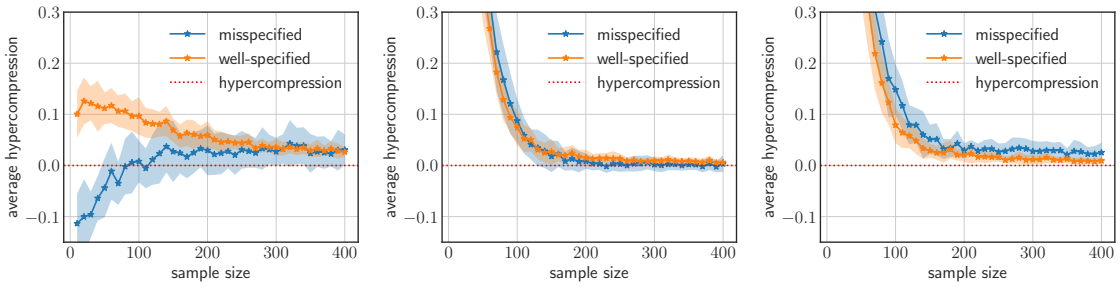
(d) HMC (e) Mean Field VB (f) Full-Covariance VB

Figure 2: *Predictive Performance.* The above plots show the average log-risk (a-c) and square-risk (d-f) for Bayesian NNs fit using Hamiltonian Monte Carlo (a, d), mean field (Normal) variational Bayes (b, e), and full-covariance (Normal) varitaional Bayes. The shaded areas correspond to ± 1 standard deviation across 30 random seeds.

3.2. Variational Bayes

Having confirmed that inconsistency can occur for HMC-obtained posteriors, we next turn to variational inference (Blei et al., 2017). As MCMC is usually impractical for NNs, this setting is closer to typical Bayesian NN implementations. We consider the most popular form of variational inference (VI): *variational Bayes* (VB). We implemented VB using two different variational families. The first uses a *mean field* factorization, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}\{\boldsymbol{\Sigma}\})$, and the second uses a full-covariance matrix, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Figures 1b and 1c show the model fit under the mean field and full-covariance posteriors respectively. Unlike the HMC fit, the VB posteriors do not seem to suffer from any inconsistency as their mean is very near the $y = 0$ line and the 95% prediction interval captures the response noise quite well. Examining the predictive quantities in Figures 2b, 2c (log-risk) and 2e, 2f (square-risk), the risks of the specified and misspecified models nearly match in all four plots. Moreover, the square-risk of both models comes quite close to the optimal value—unlike the HMC-misspecified's square-risk. The hypercompression plots in Figures 3b and 3c are positive for all sample sizes, meaning that hypercompression never occurs (compare to HMC's in 3a). Lastly, Appendix Figures 4b and 4c show the posteriors are much more concentrated than for HMC.



(a) HMC (b) Mean Field VB (c) Full-Covariance VB

Figure 3: *Hypercompression.* The above plots show the hypercompression inequality (Equation 3) for the three inferences strategies. HMC exhibits hypercompression (negative values) where as VB does not (positive values). The shaded areas correspond to ±1 standard deviation across 30 random seeds.

4. Why is Variational Bayes Consistent Under Misspecification?

Comparison to Laplace Approximation Our first thought was to examine the ELBO, as perhaps it introduces some beneficial inductive bias that leads to consistency. We tested this hypothesis by fitting the Bayesian NN using the Laplace approximation (MacKay, 1992). As Laplace's method also uses a Normal approximation of the posterior, this experiment keeps the variational family the same while changing only the optimization objective (ELBO vs MAP). Appendix Figure 5b shows the fit of the misspecified NN. We see that the line is not similar to the optimal $y = 0$ line, showing signs of inconsistency and behavior that is dissimilar to VB's. Hence, our hypothesis is supported.

ELBO: Concentration and R-log-risk Inspecting the analytical form of the ELBO, we notice that the expected likelihood $\mathbb{E}_q[\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})]$ looks similar to the R-log-risk defined in Equation 4. It is not the same—primarily because the ELBO is evaluated on training data, not held-out data. Yet, the expectation is similar in its placement outside of the log. Also, by Jensen's inequality we know that $\mathbb{E}_q\left[-\log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\right] \geq -\log \mathbb{E}_q\left[p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\right]$. Thus, we conjecture that the ELBO is optimizing the concentration of q, implicitly decreasing $\mathcal{C}(\boldsymbol{\theta}|\mathcal{D})$ (Equation 4). This does not speak directly to whether or not hypercompression (Equation 3) occurs, but it does seem less likely for more concentrated posteriors. That is because the posterior predictive will likely be close to the in-family, KL-optimal conditional $p_{\tilde{\boldsymbol{\theta}}}$.

5. Conclusions

Inspired by Grünwald and Van Ommen (2017)'s analysis of Bayesian linear models, we performed a similar inspection of Bayesian NNs. We found that while the same inconsistency likely occurs for the true posterior, approximate inference can mitigate the inconsistency. In fact, variational Bayes can recover the best predictive function while HMC and the Laplace approximation could not. We suspect that the ELBO may be the cause of VB's success, as the Laplace approximation was observed to be inconsistent despite having the same approximating family. For future work, we plan to analyze other variational inference strategies, deep NN architectures, and the choice of prior.

References

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *arXiv preprint arXiv:1810.09538*, 2018.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Alex Graves. Practical Variational Inference for Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2348–2356, 2011.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Rianne Heide, Alisa Kirichenko, Peter Grunwald, and Nishant Mehta. Safe-Bayesian Generalized Linear Regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2623–2633. PMLR, 2020.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Diederik Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations*, 2014.

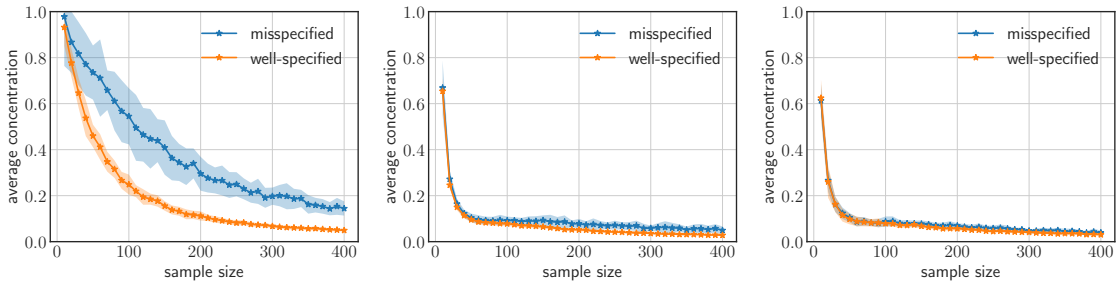
David JC MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

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.

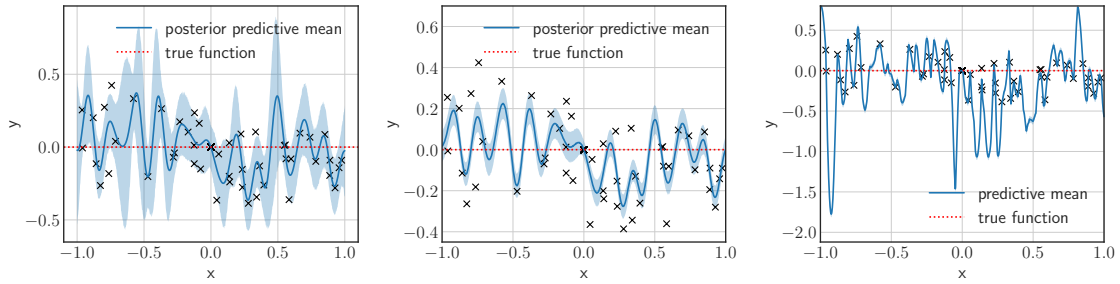
Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

Appendix A. Additional Results for Bayesian NNs



(a) HMC (b) Mean Field VB (c) Full-Covariance VB

Figure 4: *Posterior Concentration.* The above plots characterize the posterior concentration using $\mathcal{C}(\theta|\mathcal{D})$, defined in Equation 4. Lower values imply a higher degree of concentration. Subfigures (a-c) show the concentration of each of the three posterior inference strategies. The shaded areas correspond to ± 1 standard deviation across 30 random seeds.

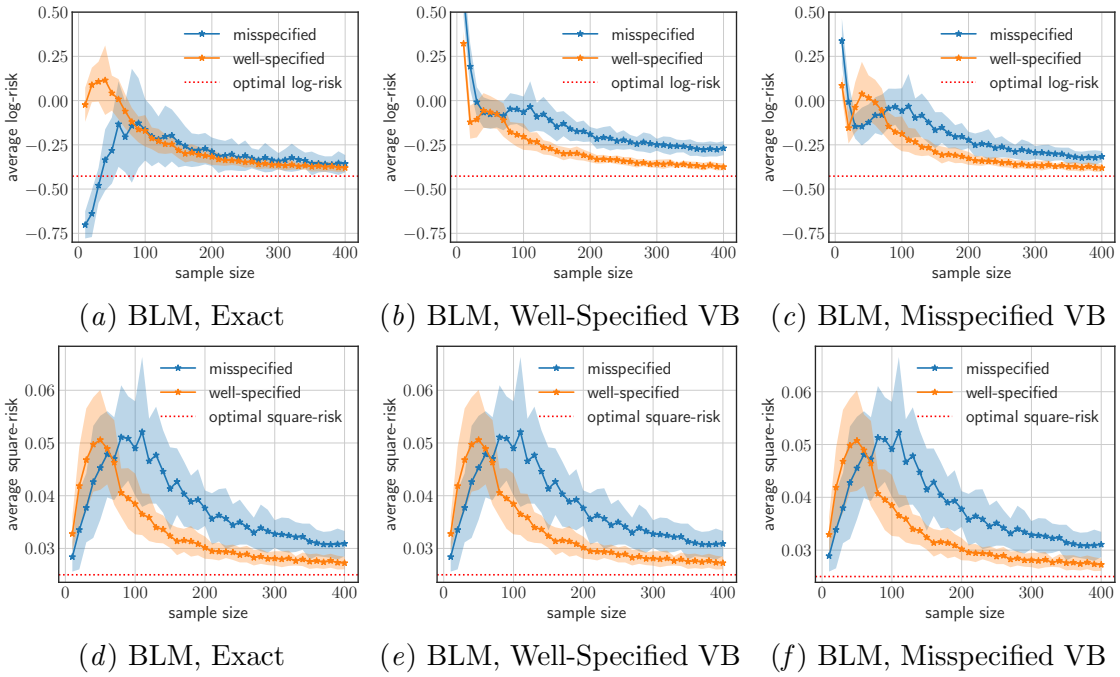


(a) BLM, Exact (b) Bayes NN, Laplace (c) NN, Max. Likelihood

Figure 5: *Misspecified Model Fits* ($N = 100$). The above plots show the fits of three misspecified models. Subfigure (a) shows the misspecified Bayesian linear model (BLM), recreating the results of Grünwald and Van Ommen (2017). Subfigure (b) shows the posterior predictive distribution for a misspecified Bayesian NN when the Laplace approximation is used for posterior inference. Subfigure (c) shows the fit for a NN with maximum likelihood estimation (no regularization). The shaded areas correspond to 95% prediction intervals.

Appendix B. Results for Bayesian Linear Models

To further understand the success of variational Bayes, we provide comprehensive results and analysis for Bayesian linear models. In particular, there are two reasons to investigate variational Bayes for Bayesian linear models. Firstly, since with Bayesian linear models, the exact posteriors are obtainable for variational Bayes, it shows us evidence whether or not the behaviors of variational Bayes for NNs are due to inaccurate posterior approximation. Secondly, as discussed in Section 2.3, unlike with NNs, with linear models, there exists a unique parameter setting $\tilde{\theta} = (\tilde{\beta} = \mathbf{0}, \tilde{\sigma}^2 = 0.025)$ that results in the KL-optimal model



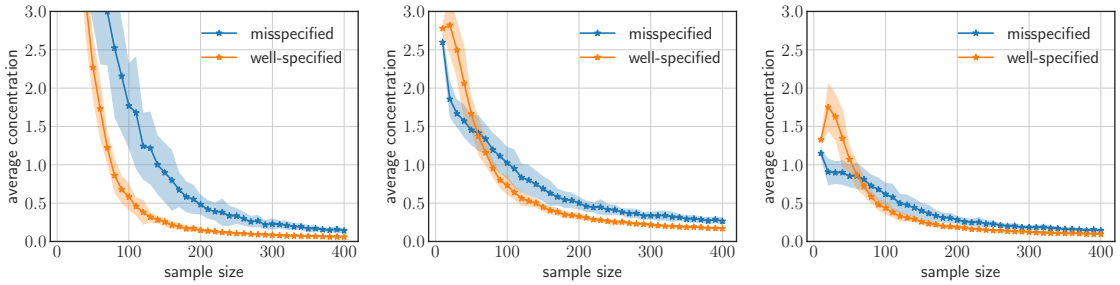
(a) BLM, Exact (b) BLM, Well-Specified VB (c) BLM, Misspecified VB

(d) BLM, Exact (e) BLM, Well-Specified VB (f) BLM, Misspecified VB

Figure 6: *Predictive Performance.* The above plots show the average log-risk (a-c) and square-risk (d-f) for Bayesian linear model (BLM) fit using exact posterior (a, d), well-specified full-covariance (Normal) variational Bayes (b, e), and misspecified mean field (Normal) varitaional Bayes. The shaded areas correspond to ± 1 standard deviation across 30 random seeds.

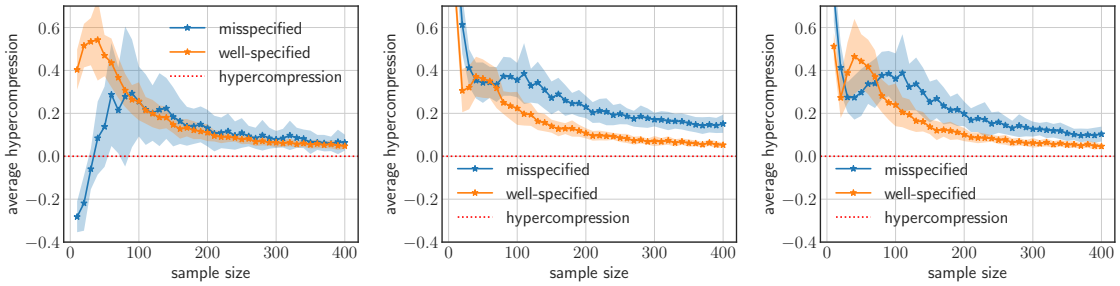
to the true generative process $p^*(\mathrm{x}, \mathrm{y})$. This property allows us to validate the consistency of VB posteriors: if the VB posteriors are consistent, they must not only concentrate, but also concentrate to $\tilde{\boldsymbol{\theta}}$. To measure this, we use the expected squared $\ell 2$-distance: $\mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}\left[||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2^2\right]$.

We further distinguish between two different settings: Bayesian linear models with well-specified full covariance (Normal) VB posteriors and with misspecified mean field (Normal) VB posteriors. The former is implemented by approximating $p(\boldsymbol{\beta}, \sigma^2|\mathcal{D})$ with $q(\boldsymbol{\beta})q(\sigma^2)$, where $q(\boldsymbol{\beta})$ is a multivariate Gaussian distribution and $q(\sigma^2)$ is an inverse-gamma distribution. The latter is implemented by firstly running the former. While keeping the mean of $q(\boldsymbol{\beta})$ the same, we replace the covariance matrix of $q(\boldsymbol{\beta})$ with one that is diagonal and has a smaller trace. These two settings are designed to mimic the behaviors of VB with NNs: while the first setting investigates the behaviors of VB when the posteriors are perfectly approachable, the second setting could be similar to the situation with NNs, where the approximate posteriors are more concentrated as they try to cover modes of the possibly multi-modal true posteriors.



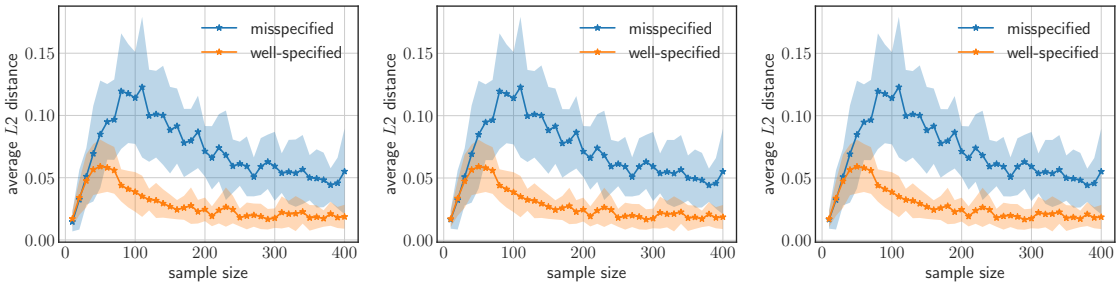
(*a*) BLM, Exact (*b*) BLM, Well-Specified VB (*c*) BLM, Misspecified VB

Figure 7: *Posterior Concentration.* The above plots characterize the posterior concentration using $\mathcal{C}(\theta|\mathcal{D})$ defined in Equation 4. Lower values imply a higher degree of concentration. Subfigures (a-c) show the concentration of exact posterior of BLM, well-specified VB posterior of BLM and misspecified VB posterior of BLM. The shaded areas correspond to ± 1 standard deviation across 30 random seeds.



(*a*) BLM, Exact (*b*) BLM, Well-Specified VB (*c*) BLM, Misspecified VB

Figure 8: *Hypercompression.* The above plots show the hypercompression inequality (Equation 3) for exact posterior of BLM, well-specified full-covariance VB posterior of BLM and misspecified mean field VB posterior of BLM. Exact posterior exhibits hypercompression (negative values) where as VB does not (positive values). The shaded areas correspond to ± 1 standard deviation across 30 random seeds.



(*a*) BLM, Exact (*b*) BLM, Well-Specified VB (*c*) BLM, Misspecified VB

Figure 9: *Expected ℓ2-distance.* The above plots show the expected ℓ2-distance for exact posterior of BLM, well-specified full-covariance VB posterior of BLM and misspecified mean field VB posterior of BLM. The shaded areas correspond to ± 1 standard deviation across 30 random seeds.

Appendix C. Implementation Details

In the experiments, we use NumPyro (Phan et al., 2019; Bingham et al., 2018) for HMC, mean field VB, full-covariance VB and Laplace approximation. For HMC, we use 1000 warm-up steps. For all three variational inference strategies (mean field VB, full-covariace VB and Laplace appproximation), we use Normal distribution to approximate the joint posterior, making the appropriate transformation (log) to $\mathbb{R}$ for the variance σ^2. Specially, for Laplace approximation, we use NumPyro to calculate the MAP solution but as the Hessian is usually not invertible, we approximate the Hessian using the Gaussian-Newton matrix, which is guaranteed to be positive semidefinite. For all three variational inference strategies, we use Adam (Kingma and Ba, 2014) to optimize ELBO with a learning rate of 0.005 and a step size of 20000. For all the model fits plot, we use a fixed sample size of 100. To compute all the quantities of interest numerically, we sample 1000 samples from the true generative process $p^*(\mathrm{x}, \mathrm{y})$ together with 2000 samples from the posteriors for each inference strategy.