Journal of Machine Learning Research 23 (2022) 1-54 Submitted 2/21; Revised 9/22; Published 9/22

Mitigating the Effects of Non-Identifiability on Inference for Bayesian Neural Networks with Latent Variables

Yaniv Yacoby* YANIVYACOBY@G.HARVARD.EDU
Weiwei Pan* WEIWEIPAN@G.HARVARD.EDU
Finale Doshi-Velez FINALE@SEAS.HARVARD.EDU
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA

Editor: Qiang Liu

Abstract

Bayesian Neural Networks with Latent Variables (BNN+LVs) capture predictive uncertainty by explicitly modeling model uncertainty (via priors on network weights) and environmental stochasticity (via a latent input noise variable). In this work, we first show that BNN+LV suffers from a serious form of non-identifiability: explanatory power can be transferred between the model parameters and latent variables while fitting the data equally well. We demonstrate that as a result, in the limit of infinite data, the posterior mode over the network weights and latent variables is asymptotically biased away from the ground-truth. Due to this asymptotic bias, traditional inference methods may in practice yield parameters that generalize poorly and misestimate uncertainty. Next, we develop a novel inference procedure that explicitly mitigates the effects of likelihood non-identifiability during training and yields high-quality predictions as well as uncertainty estimates. We demonstrate that our inference method improves upon benchmark methods across a range of synthetic and real data-sets.

Keywords: bayesian neural networks, approximate inference, latent variable models, heteroscedastic noise, non-identifiability

1. Introduction

In machine learning fields such as active learning, Bayesian optimization, and reinforcement learning, one often needs to fit functions to labeled data that produce both high-quality predictions as well as appropriate quantification of predictive uncertainty. Often crucial in these applications is the task of identifying sources of predictive uncertainty, especially those that are reducible through additional data collection (Zhang and Lee, 2019; Henaff et al., 2019).

In general, one can divide the sources of uncertainty in a prediction into two categories. *Epistemic* uncertainty, or model uncertainty, comes from having insufficient data to determine the "true" predictor, and can be reduced with more observations. In contrast, *aleatoric* uncertainty comes from the irreducible or unobservable stochasticity in the environment.

∗. Equal contribution

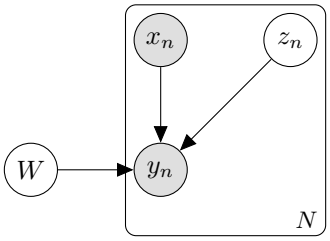


Figure 1: **Bayesian Neural Network with Latent Variables (BNN+LV).** The outputs y are explained using a function, parameterized by W, of the observed inputs x and a latent variable z.

Although in regression one typically assumes a simple fixed form for aleatoric uncertainty (for example, independently and identically sampled Gaussian output noise), in practice, environmental stochasticity can be a complex function of the input; that is, many downstream tasks require us to model heteroscedastic aleatoric uncertainty (Chaudhuri et al., 2017; Nikolov et al., 2018; Griffiths et al., 2019; Kendall and Gal, 2017).

Bayesian Neural Networks with Latent Variables (BNN+LVs) (Wright, 1999; Depeweg et al., 2018) meet the need for scalable predictive models that can capture complex forms of epistemic uncertainty and heteroscedastic aleatoric uncertainty. In particular, a BNN+LV (Figure 1) assumes the following data generation process:

$$W \sim p(W), \quad z_n \sim p(z), \quad \epsilon_n \sim p(\epsilon),$$
$$y_n = f(x_n, z_n; W) + \epsilon_n, \quad n = 1, \ldots, N,$$

where ϵ is a simple form of output noise, W are the parameters of a neural network, and z is an unobserved (latent) input variable associated with *each* observation (x, y), sampled independently of the input x. BNN+LVs explicitly (and separately) model aleatoric and epistemic uncertainties: the distribution over W captures model uncertainty (epistemic uncertainty); together with the output noise ϵ, the stochastic latent input z captures aleatoric uncertainty. Even if the output noise ϵ is assumed to be simple, because the latent input z is passed through the neural network alongside the input x, it can model complex, heteroscedastic noise patterns in data (Depeweg et al., 2018). The presence of both ϵ and z in the model allows us to further explain aleatoric noise as arising from a combination of random observation error and latent stochastic input, which can represent white noise or unobserved but meaningful covariates. Previous work has demonstrated the usefulness of BNN+LV's decomposition of predictive uncertainty (into epistemic and aleatoric uncertainties) in downstream applications like active learning and safe reinforcement learning (Depeweg et al., 2018), where one needs to carefully balance the risks and rewards of gathering new data.

For this model, the goal of inference is to draw samples from the posterior of the weights given the data, $p(W|\mathcal{D})$, in order to compute a Monte Carlo estimate of the posterior predictive distribution,

$$p(y^*|x^*, \mathcal{D}) = \int p(y^*|x^*, W) p(W|\mathcal{D}) dW,$$

which evaluates the probability of a new outcome y^* given a new input x^*. Since $p(W|\mathcal{D})$ requires marginalizing out $z_1, \ldots, z_N$ from $p(W, z_1, \ldots, z_N|\mathcal{D})$ (the posterior over all unknown variables), it is intractable to approximate directly. In practice, one therefore approximates the joint posterior distribution of the weights and latent inputs given the data, $p(W, z_1, \ldots, z_N|\mathcal{D})$, and integrates out the latent inputs $z_1, \ldots, z_N$ empirically (by drawing samples from the joint posterior, but only keeping samples over the weights).

In this work, we show that this approach yields poor quality posterior predictives in practice. We provide a theoretical characterization explaining the problem and present a new inference method to mitigate it. Specifically, we show that in the limit of infinite data, the mode of the *true* joint posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$ is asymptotically biased away from the ground-truth parameters that generated the data. As a result, empirically marginalizing out the latent inputs from an *approximation* of the joint posterior results in an approximation of $p(W|\mathcal{D})$ that explain the data poorly. We then propose a practical approximate inference method, grounded in our theoretical characterization of the asymptotic bias, that forces the inferred posterior to satisfy our modeling assumptions. Our contributions are:

A. (True Posterior Mode) We prove that in the limit of infinite data, the weights W and latent inputs $z_1, \ldots, z_N$ at the mode of BNN+LV joint posterior are asymptotically biased. In the case that z represents meaningful latent covariates, one would hope that $p(W, z_1, \ldots, z_N|\mathcal{D})$ encodes valuable information about these latent covariates. However, in this work, we show that in the limit of infinite data, the BNN+LV joint posterior mode is not located at the ground-truth $W, z_1, \ldots, z_N$ that generated the data. Specifically, we prove that the BNN+LV's likelihood is non-identifiable and use this non-identifiability to show that, for any set of ground-truth $W, z_1, \ldots, z_N$, there exists an alternative set of weights $\widehat{W}, \widehat{z}_1, \ldots \widehat{z}_N$ that is scored as more likely under the posterior. This alternative set of parameters violates our modeling assumption that x and z are independent: $\widehat{z}_1, \ldots \widehat{z}_N$ memorized the inputs, $x_1, \ldots, x_N$, and as a result, $\widehat{W}$ parameterizes a function that explains the observed data poorly and generalizes poorly. This result has two important implications:

- For any downstream application in which we wish to interpret the inferred latent variables $z_1, \ldots, z_N$, it is meaningless to look at the mode of the joint posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$; instead, one must look at the mode of the posterior marginal $p(z_1, \ldots, z_N|\mathcal{D})$ or conditional $p(z_1, \ldots, z_N|\mathcal{D}, W)$ (given a specific W of interest).
- The mode of the joint posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$ is asymptotically biased, and may thus bias imperfect approximations of the joint posterior, $q(W, z_1, \ldots, z_N|\mathcal{D})$, towards approximations that violate our generative modeling assumptions, just like $\widehat{W}, \widehat{z}_1, \ldots \widehat{z}_N$. As a result, empirically marginalizing out $z_1, \ldots, z_N$ from this approximation may yield an approximation $q(W|\mathcal{D})$ of $p(W|\mathcal{D})$ that explains the data poorly (leading to contribution B).

We note that the latter implication is caused by the *approximation* of the joint posterior. Given a tractable method that directly infers $p(W|\mathcal{D})$ without approximation, the resultant posterior predictive may explain the data well. In other words, while the joint posterior mode is asymptotically biased (contribution A), the mode of the marginal posterior $p(W|\mathcal{D})$ may still parameterize the data generating function in the limit of infinite data (i.e. the consistency of the BNN+LV true posterior predictive is an open problem). We emphasize

that this paper focuses on pathologies caused by mean-field variational approximations of the joint posterior $p(W, Z|\mathcal{D})$ (contribution B).

B. (Approximate Posterior) We empirically demonstrate that mean-field variational inference is prone to capturing high mass regions of the joint posterior (near the biased mode), corresponding to posterior predictives that explain the data poorly, generalize poorly and misestimate uncertainty. Specifically, the weights and latent inputs inferred by mean-field variational inference correspond to models that, like the weights and latent inputs of the joint posterior mode, violate our generative modeling assumptions—that the latent variables are independent of the inputs. The resultant approximation of $p(W|\mathcal{D})$ yields a posterior predictive explains the data poorly and generalizes poorly.

C. (Method) We propose a new variational family for BNN+LV that mitigates the effect of the joint posterior bias. Our proposed variational family filters out models that do not satisfy our modeling assumptions—that the latent inputs z are independent of the inputs x – thereby mitigating the effects of the asymptotic bias on mean-field variational inference. Since our proposed variational family is intractable to use directly, we re-formulate variational inference with this family as a constraint optimization problem using a proxy objective. On a range of synthetic and real data-sets, we empirically show that posterior predictives learned via our method, Noise Constrained Approximate Inference (NCAI), perform significantly better: models trained this way consistently recover posterior predictives with properties (generalization and uncertainty estimates) more similar to those of the ground-truth.

2. Related Work

Models for heteroscedastic regression. In the standard Bayesian Neural Network (BNN) model for regression (e.g. MacKay (1992); Neal (2012)), one generally assumes that the irreducible noise (aleatoric uncertainty) in the data is identically and independently distributed. However, many real-world tasks (Kendall and Gal, 2017; Depeweg et al., 2018) require more complex forms of aleatoric uncertainty. In particular, one may need to relax the assumption that the noise is identically distributed—not only may the variance of the noise depend on the input (heteroscedasticity) but the form of the distribution may also change depending on the input. Works that consider more complex noise models take two main forms. The first considers a predictor of the form $y = f(x; W) + \epsilon(x)$, where the output noise ϵ is a stochastic function of the input x (e.g. Kou et al. (2015); Bauza and Rodriguez (2017)). These "output noise" models have a long history in the Gaussian process (GP) literature (e.g. Le et al. (2005); Wang and Neal (2012); Kersting et al. (2007)) and have been formulated more recently for BNNs (e.g. Kendall and Gal (2017); Gal et al. (2016)). Such models are appropriate, for example, when one believes that aleatoric uncertainty is solely rooted in observational error of the output (the input is measured without noise and there are no unobserved explanatory variables), and that this error varies across the input domain. For example, y may represent noisy sensor readings at a region x, and some regions may be more error-prone than others.

However, assuming such an additive noise structure can be restrictive: one must fix a specific family of distributions for the output noise. For example, for the noise distribution, one commonly chooses a zero-mean Gaussian family with input-dependent variance; this choice assumes that the observation noise is symmetrically distributed. In this paper, we focus on an alternative approach to modeling irreducible noise, in which we explicitly consider a latent input variable z for each input x, representing either white noise or meaningful but unobserved covariates, in addition to i.i.d. output noise ϵ: $y = f(x, z; W) + \epsilon$. This "latent input noise" model encapsulates the "output noise" model as a special case and allows us to capture arbitrarily complex noise patterns while assuming simple distributions for z and ϵ. Furthermore, since the "latent input noise" model decomposes the source of noise into observation error and latent stochastic input, this model is more appropriate when, based on domain or task-specific knowledge, one wants to explicitly account for latent factors that affect the output. For example, y may represent patient response to treatment given measured factors, x, including BMI, as well as latent factors, z, such as stress level, which is either not measured or cannot be measured in practice. In this case, x represents stable characteristics of the patient (e.g. BMI does not vary too much from day to day), while z represents transient characteristics of the patient (stress can vary wildly as a function of external factors, e.g. work-stress, family emergencies). Since in this scenario, z cannot be meaningfully predicted from x, we assume that x and z are independent, and that z is randomly sampled every time the patient is observed. During inference, one would ideally infer both the function parameters W as well as the latent factors z that explain the treatment outcome y for a given patient x.

Inference challenges for latent variable models. A related "latent input noise" model is the frequentist mixed effects model. Although the literature in this area is rich, most works only consider cases where the function f is linear. In some of these cases, it has been shown that the MLE of the parameters W of f and the latent variables z is inconsistent due to the "non-vanishing effects" of the random variables z (as the number of observations grows, so does the number of latent variables z that need to be estimated). That is, inferring jointly the model parameters and the latent factors could yield asymptotically biased estimates. This problem is known as the "incidental parameters problem" (Neyman and Scott, 1948; Lancaster, 2000). While for specific models, it is possible to get a consistent MLE estimator for the parameters W by first marginalizing out the latent variable z (Kiefer and Wolfowitz, 1956) there are no works to our knowledge that address the consistency of estimators of W and z in the case that f is an arbitrarily non-linear function.

In the case of Bayesian mixed effects models, there are a number of works that demonstrate the frequentist consistency of the posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$ as both the number of inputs and the number of occurrences of each latent variable approach infinity (Baghishani and Mohammadzadeh, 2012). There are, however, fewer works that examine the consistency of $p(W, z_1, \ldots, z_N|\mathcal{D})$ when the number of occurrences of each latent variable is fixed (Wang and Blei, 2019). In fact, it is generally not known whether or not the joint posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$ or the marginal posterior $p(W|\mathcal{D})$ over W (derived by integrating out z) is consistent for arbitrary non-linear functions f.

Bayesian latent variable models for non-linear regression. For Bayesian "latent input noise" models where the function f is nonlinear, there are a number of works that

place a Gaussian Process prior over f (e.g. Lawrence and Moore (2007); McHutchon and Rasmussen (2011); Damianou et al. (2014)). In contrast, there are only a few works on "latent input noise" models that place a BNN prior over f (Wright, 1999; Depeweg et al., 2018). While GP latent variable models have been shown to successfully capture complex noise patterns in low-dimensional regression data, for high-dimensional data with non-local structure (e.g. images, natural language) it is more natural to apply models like BNN+LV whose priors are over flexible parametric forms of f. However, in this work, we show that the flexibility gained by adding latent input noise variables to BNNs presents new challenges for inference.

Non-identifiability in deep probabilistic models. Due to their complexity, deep probabilistic models are often non-identifiable—that is, there exist several different sets of parameters that all explain the observed data equally well. For example, the weights of a BNN can be permuted while still parameterizing the same function (known as "weight-space symmetry" Pourzanjani et al. (2017)), and the latent space of a Variational Autoencoder (Kingma and Welling, 2013) can be transformed while still explaining the observed data equally well (e.g. Locatello et al. (2019)). In such scenarios, it is has been shown that the undesirable effects of non-identifiability can be mitigated by modifying the model itself (e.g. Pourzanjani et al. (2017); Khemakhem et al. (2020)), or by specifying additional model selection criteria (Zhao et al., 2018).

To our knowledge, we are the first to describe non-trivial likelihood non-identifiability that occurs in BNN+LV and how this non-identifiability impacts inference (particularly variational inference). The non-identifiability we characterize in this paper is different than the previously studied weight-space symmetry of BNNs, and thus presents different challenges for inference; the posterior multi-modality caused by the weight-space symmetry has been empirically shown to slow the convergence for MCMC methods and lead to poor variational approximations (Pourzanjani et al., 2017; Papamarkou et al., 2022). In contrast, the BNN+LV likelihood non-identifiability we characterize is between the weights and latent variables. As we show in this work, it causes the posterior mode of the posterior $p(W, z_1, \ldots, z_N|\mathcal{D})$ to be asymptotically biased. These problems have not been explicitly considered in previous work likely because the impacts of non-identifiability on inference are often attributed to general optimization difficulties. Based on our analysis of non-identifiability in BNN+LV models, we propose modifications to the mean-field variational family that explicitly mitigate the effects of likelihood non-identifiability.

3. Background and Notation

Let $\mathcal{D} = \{(x_1, y_1), \ldots (x_N, y_N)\}$ be a data-set of N observations from the true data distribution $p(y|x)p(x)$, in which each input $x_n \in \mathbb{R}^D$ is a D-dimensional vector and each output $y_n \in \mathbb{R}^L$ is L-dimensional. Let I denote the identity matrix and let capital letters denote sets of variables, e.g. $X = \{x_n\}_{n=1}^N$, $Y = \{y_n\}_{n=1}^N$, $Z = \{z_n\}_{n=1}^N$. See Appendix A for a summary of the notation.

A Bayesian Neural Network (BNN). A BNN assumes a predictor of the form $y = f(x; W) + \epsilon$, where f is a neural network parametrized by W and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I)$ is a normally distributed noise variable. It places a prior $p(W)$ on the network parameters; given

a data-set $\mathcal{D}$, we can apply Bayesian inference to compute the posterior distribution $p(W|\mathcal{D})$ over W or the posterior predictive distribution $p(y^*|x^*, \mathcal{D})$ of over outputs y^* given a new input x^*. As commonly done, we assume $p(W) = \mathcal{N}(0, \sigma_w^2 \cdot I)$.

A Bayesian Neural Network with Latent Variables (BNN+LV). A BNN+LV enables more flexible noise distributions for BNNs by introducing a latent variable $z_n \sim \mathcal{N}(0, \sigma_z^2 \cdot I)$ for each observation (x_n, y_n) (Depeweg et al., 2018). It assumes the following data generation process (Figure 1):

$$\begin{aligned} &W \sim p(W), \quad z_n \sim p(z), \quad \epsilon_n \sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I), \\ &y_n = f(x_n, z_n; W) + \epsilon_n, \quad n = 1, \ldots, N. \end{aligned} \tag{1}$$

In this model, z is independent from x, and can represent either white noise or meaningful latent explanatory variables. When f is non-linear, BNN+LV is able to model heteroscedastic noise by transforming z. Inference for BNN+LVs involves approximating the posterior distribution,

$$p(W, Z|\mathcal{D}) \propto p(W) \cdot \prod_n p(y_n|x_n, z_n, W)p(z_n), \tag{2}$$

(where $Z = \{z_1, \ldots, z_N\}$), over both network weights W and the latent input z_n for each observation x_n. When Z represents meaningful latent variables, we may infer the latent information z_n for each input x_n by marginalizing $p(W, Z|\mathcal{D})$ over W. For a new input x^*, the posterior predictive is given by the expected likelihood under the posterior of W, and the prior of z (Depeweg et al., 2018):

$$\begin{aligned} p(y^*|x^*, \mathcal{D}) &= \int p(y^*|x^*, W)p(W|\mathcal{D})dW \\ &= \iint p(y^*|x^*, z^*, W)p(z^*)dz^*p(W|\mathcal{D})dW. \end{aligned} \tag{3}$$

Note that z^* is sampled from the prior $p(z)$ to compute the posterior predictive for a new input. This is because, in BNN+LVs, the form of environmental stochasticity modeled by the latent input z does not change between train and test time. As an example, suppose that the latent input for our model is $z \sim \mathcal{N}(0, 1)$. Given an observation (x_n, y_n) in the training data, we may infer the values of z_n that is likely to have generated y_n given x_n, i.e. we compute the posterior $p(z_n|x_n, y_n, W)$. Note that the posterior $p(z_n|x_n, y_n, W)$ will generally not be concentrated around $z_n = 0$ like the prior (e.g. if the sampled noise z_n is equal to 2 then the posterior $p(z_n|x_n, y_n, W)$ should concentrate near 2). However, given a new input x^* for which we want to make a prediction (i.e. we are asked to predict y^* rather than being given the target), what we've inferred about the input noise for x_n is irrelevant to the prediction task for x^*, since the latent input z^* for the new input x^* is generated randomly from $p(z)$ and is independent of z_n.

Inference for BNN+LVs. Our inference goal is to draw samples from $p(W|\mathcal{D})$, so we can compute a Monte-Carlo estimate of the posterior predictive (Equation 3). While asymptotically exact, MCMC methods are generally impractical for BNNs with large architectures trained over large data-sets; as such, we focus on variational inference. However, for BNN+LV,

it is intractable to approximate $p(W|\mathcal{D})$ directly using variational inference (by minimizing $D_{\mathrm{KL}}[q_\phi(W|\mathcal{D})||p(W|\mathcal{D})]$), since doing so requires an intractable marginalization of $z_1, \ldots, z_N$ (see Appendix C). Instead, Depeweg et al. (2018) advocate for approximating the posterior over all unobserved variables, $p(W, Z|\mathcal{D})$. One can then easily sample from $p(W|\mathcal{D})$ by sampling from $p(W, Z|\mathcal{D})$ and disregarding the samples over Z.

As commonly done in the BNN literature (e.g. Blundell et al. (2015); Depeweg et al. (2018); Foong et al. (2020)), we approximate the true posterior with a fully factorized Gaussian over network weights and latent variables:

$$\begin{aligned} q_\phi(Z, W|\mathcal{D}) &= q_\phi(Z|\mathcal{D}) \cdot q_\phi(W|\mathcal{D}) \\ &= \prod_n q_\phi(z_n|x_n, y_n) \cdot \prod_i q_\phi(w_i) \\ &= \prod_n \mathcal{N}(z_n|\mu_{z_n}, \sigma^2_{z_n} \cdot I) \cdot \prod_i \mathcal{N}(w_i|\mu_{w_i}, \sigma^2_{w_i}), \end{aligned} \tag{4}$$

where ϕ is the set of variational parameters $\{\mu_{z_n}, \sigma^2_{z_n}\}_{n=1}^N \cup \{\mu_{w_i}, \sigma^2_{w_i}\}_{i=1}^I$, over which we minimize a choice of divergence between the $q_\phi(Z, W|\mathcal{D})$ and the true posterior $p(W, Z|\mathcal{D})$. We choose the commonly used KL-divergence, yielding the following evidence lower bound:

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{q_\phi(Z,W|\mathcal{D})}[p(Y|X, W, Z)] - D_{\mathrm{KL}}[q_\phi(W|\mathcal{D})\|p(W)] - D_{\mathrm{KL}}[q_\phi(Z|\mathcal{D})\|p(Z)]. \tag{5}$$

Maximizing $\mathrm{ELBO}(\phi)$ over ϕ is equivalent to minimizing the KL-divergence of our approximate and true posteriors.

Uncertainty decomposition in BNN+LV. Following Depeweg et al. (2018), we quantify the overall uncertainty in the posterior predictive using entropy, $\mathbb{H}\left[p(y_*|x_*)\right]$. We compute the aleatoric uncertainty due to z and ϵ by taking the expectation of $\mathbb{H}\left[p(y_*|W, x_*)\right]$ with respect to W:

$$\mathbb{E}_{q_\phi(W|\mathcal{D})}\left[\mathbb{H}\left[p(y_*|W, x_*)\right]\right]. \tag{6}$$

We then quantify the epistemic uncertainty due to W by computing the difference between total and aleatoric uncertainties:

$$\mathbb{H}\left[p(y_*|x_*)\right] - \mathbb{E}_{q_\phi(W|\mathcal{D})}\left[\mathbb{H}\left[p(y_*|W, x_*)\right]\right]. \tag{7}$$

4. Asymptotic Bias of the BNN+LV Posterior Modes

In this section, we prove that in the limit of infinite data, due to structural non-identifiability in the BNN+LV likelihood, and due to the non-vanishing effect of the prior over the latent inputs, the mode of the BNN+LV joint posterior $p(W, Z|\mathcal{D})$ is asymptotically biased towards parameters that explain the observed data poorly and generalize poorly. We do this by first characterizing a number of non-trivial ways in which network weights W and latent input z in BNN+LV models can be non-identifiable under the likelihood. Then, we prove that due to this non-identifiability, for any ground-truth set of weights W^{true} and corresponding ground-truth latent variables $Z^{\mathrm{true}} = \{z_n^{\mathrm{true}}\}_{n=1}^N$, there exists an alternative sets of weights and latent variables that are scored higher under the posterior as the number

N of observations grows. Furthermore, the functions parametrized by the alternate set of weights explain the observed data poorly and generalize poorly to new data. We note that this is unlike the case of BNNs without latent variables: BNNs are also non-identifiable under the likelihood (e.g. one can permute the weights and retain the same function), but the posterior modes parameterize functions that recover the data generating function as the number of observations increases, and the posterior predictive of BNNs nonetheless concentrates around the ground-truth function, under mild assumptions (Lee, 2000). Our results have two notable consequences:

1. **Interpretation of the Latent Inputs:** For any downstream task in which we wish to interpret the inferred latent inputs Z, one should never summarize $p(W, Z|\mathcal{D})$ with its mode; instead, one may use the mode to summarize $p(Z|\mathcal{D})$ or $p(Z|\mathcal{D}, W)$ (given a specific W of interest).

2. **Approximate Inference:** As we show in Section 5, the asymptotic bias of the mode of the joint posterior $p(W, Z|\mathcal{D})$ exacerbates the bias in our mean-field approximation of the joint posterior, resulting in approximations of $p(W|\mathcal{D})$ that generalize poorly.

For intuition and clarity, we begin by characterizing likelihood non-identifiability and posterior bias for a single node of a BNN+LV and then generalize these results to a 1-layer BNN+LV. We assume the model is well-specified.

4.1 Asymptotic Bias of 1-Node BNN+LV Posterior Modes

Non-Identifiability. Consider univariate output generated by a single hidden-node neural network with LeakyReLU activation. For simplicity, we study a case with zero network biases, unit output weights, and additive input noise:

$$f(x, z; W) = \max\{W(x + z), \alpha W(x + z)\},$$

where $0 < \alpha < 1$. For any non-zero constant C, the pair $\widehat{W}^{(C)} = W/C$, $\widehat{z}^{(C)} = (C-1)x + Cz$ reconstructs the observed data equally well:

$$\max\{W(x + z), \alpha W(x + z)\} = \max\left\{\widehat{W}^{(C)}\left(x + \widehat{z}^{(C)}\right), \alpha\widehat{W}^{(C)}\left(x + \widehat{z}^{(C)}\right)\right\}.$$

Now, suppose that the output is observed with Gaussian noise: $y \sim \mathcal{N}(f(x, z; W), \sigma_\epsilon^2)$. Then the true values of the parameter W and the latent input noise z are equally likely as $\widehat{W}^{(C)}$ and $\widehat{z}^{(C)}$ under the likelihood: $p(y|f(x, z, W), \sigma_\epsilon^2) = p(y|f(x, \widehat{z}^{(C)}, \widehat{W}^{(C)}), \sigma_\epsilon^2)$. We show in Theorem 1 that for this model, *the posterior over the model parameter and latent inputs W, Z is biased away from the ground-truth towards parameters that generalize poorly as the sample size grows, regardless of the choice of W prior.*

Theorem 1 (Asymptotic Bias of the Posterior Mode of 1-Node BNN+LV) *Fix any $W \in \mathbb{R}$ and any bounded prior $p_W(W)$ on W. Suppose that inputs $\{x_1, \ldots, x_N\}$ are sampled i.i.d from p_x with finite first and second moments μ_x, σ_x^2, and that $\{z_1, \ldots, z_N\}$ are sampled i.i.d from $\mathcal{N}(0, \sigma_z^2)$, where $\mu_x^2 + \sigma_x^2 > \sigma_z^2$. There exist a non-zero $\mathcal{C}$ such the probability that the scaled values $(\widehat{W}^{(C)}, \{\widehat{z}_n^{(C)}\}_{n=1}^N)$ are more likely than $\left(W, \{z_n\}_{n=1}^N\right)$ under the posterior approaches 1 as $N \to \infty$, for every $C \in (\mathcal{C}, 1)$.*

The proof of Theorem 1 is in Appendix B.1. The theorem says that in the limit of infinite data, with probability approaching 1, the posterior over W, Z is biased away from the ground-truth model parameters. As such, the function corresponding to these alternative parameters generalizes poorly (that is, since W and $\widehat{W}^{(C)}$ parameterize models with different slopes, $p(y|x, W) \neq p(y|x, \widehat{W}^{(C)})$, and $\widehat{W}^{(C)}$ thereby explains new data poorly). Furthermore, since these results hold for any bounded, data-independent prior over the weights, they suggests that the bias in the joint posterior cannot be removed via model selection (e.g. through a clever selection of priors).

Next, using the exact same mechanism, we characterize non-identifiability and the resultant bias in the posterior for a 1-layer BNN+LV.

4.2 Asymptotic Bias of 1-Layer BNN+LV Posterior Modes

Non-Identifiability. Sources of non-identifiability increase when f is a neural network. Consider a single-output neural network, f, that takes as input x and z (represented as a concatenated vector in $\mathbb{R}^{2D}$) and has a single hidden layer containing H hidden nodes. At the output node, we fix the activation to be the identity. Thus, the activation of the output node is computed as

$$a^{\text{out}} = (a^{\text{hidden}})^\top W^{\text{out}} + b^{\text{out}},$$

where W^{out} is a H-dimensional weight vector, b^{out} is the bias and a^{hidden} is the H-dimensional vector of activations of the hidden nodes. We can further expand a^{hidden} as

$$a^{\text{hidden}} = g\left(W^x x + W^z z + b^{\text{hidden}}\right),$$

where W^x and W^z are weight matrices in $\mathbb{R}^{H\times D}$, b^{hidden} is a H-dimensional bias vector and g is the activation function, applied element-wise. We characterize ways that the models parameters and the latent input variables z are non-identifiable given a set of observed data. For any choice of diagonal matrix $S \in \mathbb{R}^{D\times D}$, vector $U \in \mathbb{R}^D$, and any factorization $W^z = RT$ where T is in $\mathbb{R}^{D\times D}$, we can express a^{hidden} in two *equivalent* ways:

$$a^{\text{hidden}} = g\left(W^x x + W^z z + b^{\text{hidden}}\right) = g\left(\widehat{W}^x x + \widehat{W}^z \widehat{z} + \widehat{b}^{\text{hidden}}\right)$$

by setting:

$$\widehat{W}^x = W^x + W^z S, \quad \widehat{W}^z = R, \quad \widehat{z} = Tz - TSx - U, \quad \widehat{b}^{\text{hidden}} = b + RU. \tag{8}$$

Suppose that the output is observed with Gaussian noise: $y \sim \mathcal{N}(f(x, z; W), \sigma_\epsilon^2)$, where W denotes the set of weights $(W^{\text{out}}, W^x, W^z, b^{\text{out}}, b^{\text{hidden}})$. Then, by only observing outputs generated by this network given observed input, one cannot identify ground-truth parameter and latent variable values under the likelihood: $p(y|f(x, z, W), \sigma_\epsilon^2) = p(y|f(x, \widehat{z}, \widehat{W}), \sigma_\epsilon^2)$. Just like in the case of a network with a single node, we show in Theorem 2 that *the joint posterior is biased away from the ground-truth parameters, regardless of the choice of weight prior.*

Theorem 2 (Asymptotic Bias of the Posterior Mode of 1-Layer BNN+LV) *Fix any set of parameters W and any bounded prior $p_W(W)$ on W. Suppose that $\{x_1, \ldots, x_N\}$*

is sampled i.i.d from p_x, with finite first and second moments, and that $\{z_1, \ldots, z_N\}$ are sampled i.i.d from $\mathcal{N}(0, \sigma_z^2 \cdot I)$. There exists an alternate set of parameters $\left(\widehat{W}, \{\widehat{z}_n\}_{n=1}^N\right)$ such that the probability that these alternative parameters are valued as more likely than $\left(W, \{z_n\}_{n=1}^N\right)$ under the posterior approaches 1 as $N \to \infty$.

The proof for Theorem 2 is in Appendix B.2. As in the 1-node BNN+LV case, the bias in the joint posterior cannot be removed via model selection (e.g. through a clever selection of bounded, data-independent priors over the weights and latent variables).

Note that the form of non-identifiability in 1-layer BNN+LV models is not limited to the cases we describe above. In Appendix B.3, we analyze another form of non-identifiability, in which the latent variable compensates for any scaling of W^{out} by encoding the training outputs y. In practice, we find that poor posterior predictive distributions are always associated with the latent variable encoding for either the input x or the output y, both of which we quantify by measuring the mutual information of the inferred z's and the training data (see Section 7). Lastly, we note that our characterization of non-identifiability can be easily extended to multi-layer networks, which have, at the very least, the types of non-identifiability we describe above at the input layer.

We next demonstrate that the weights and latent inputs most likely under the joint posterior violate our generative modeling assumptions—an insight that we will exploit in Section 6 to develop a new method to mitigate the effects of the asymptotic bias on approximate inference.

4.3 The Joint Posterior Mode Violates Generative Modeling Assumptions

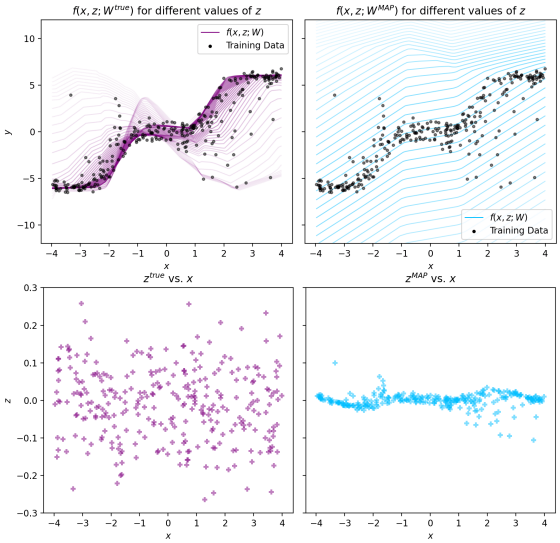
Generally, when assuming a probabilistic model, we hope that in the limit of infinite data, inference recovers model parameters that satisfy our generative modeling assumptions. For example, when assuming a BNN, we hope that in the limit of infinite data, the MLE and MAP over the weights parameterize the true function that generated the data. For the BNN+LV model, however, the parameters given by the MAP of the joint posterior $W^{\text{MAP}}, Z^{\text{MAP}}$ do not satisfy our generative modeling assumptions. To illustrate this, we compare $W^{\text{MAP}}, Z^{\text{MAP}}$ with the ground-truth data generating parameters, $W^{\text{true}}, Z^{\text{true}}$; under our ground-truth data-generating process from Equation 1, Z^{true} satisfies two modeling assumptions:

Assumption 1: z^{true} is independent of x—that is, $p(z, x)$ factorizes.

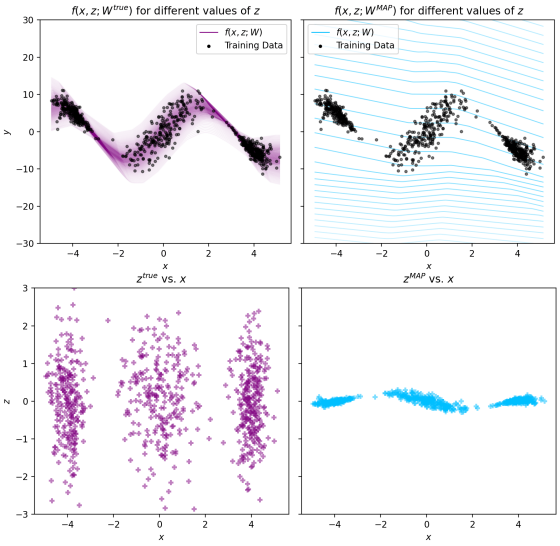
Assumption 2: z^{true} is distributed like the prior $p(z)$.

In contrast, using our characterization of likelihood non-identifiability from Equation 8, in the limit of infinite data, the posterior always prefers an alternative set of parameters $\widehat{W}, \widehat{Z}$ that violate the above modeling assumptions. Specifically in Equation 8, each $\widehat{z}_n$ becomes directly dependent on the inputs x_n, (or indirectly dependent on x_n through y_n—see Appendix B.3), thereby violating assumption 1. Next, since our characterization, $\widehat{z}_n$ is dependent on x_n (which may not be drawn from a Gaussian), $\widehat{z}_1, \ldots, \widehat{z}_N$ may not be distributed like the prior, thereby violating assumption 2.

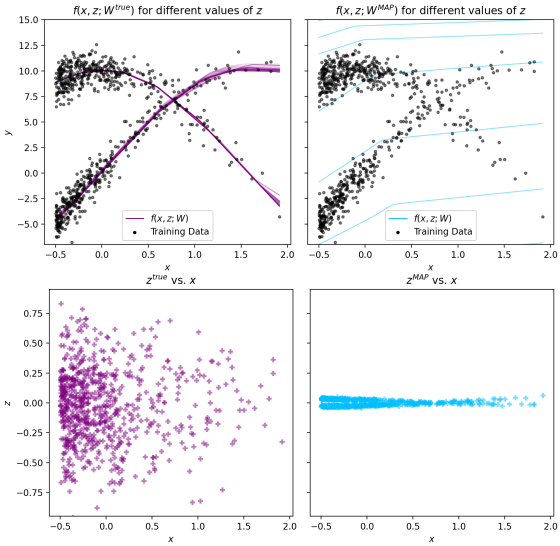
The two assumptions are independent of one another. We note that violating one assumption does not necessarily imply violating the other. Consider for instance $\widehat{z}_1, \ldots, \widehat{z}_N$



(a) **Heavy-Tail:** true (purple) vs. MAP (blue)



(b) **Depeweg:** true (purple) vs. MAP (blue)



(c) **Bimodal:** true (purple) vs. MAP (blue)

	log posterior of $W^{\text{true}}, Z^{\text{true}}$	log posterior of $W^{\text{MAP}}, Z^{\text{MAP}}$
Heavy-Tail	-311.2 ± 12.9	$\mathbf{53.0 \pm 0.4}$
Depeweg	-1719.5 ± 20.7	$\mathbf{-938.7 \pm 0.2}$
Bimodal	-2516.1 ± 22.3	$\mathbf{-1560.2 \pm 0.0}$

(d) Comparison of log-posterior evaluated at the true parameter $W^{\text{true}}, Z^{\text{true}}$ vs. the MAP parameters $W^{\text{MAP}}, Z^{\text{MAP}}$. The posterior is significantly higher at the MAP.

Figure 2: **At the MAP of the joint posterior, W explains the data poorly and Z memorizes the data.** Top-row of (a)-(c): We compare the ground-truth function (purple) vs. function learned via MAP estimate over W, Z (blue) (with $\sigma_w^2 = 10.0$). The functions are plotted for every value of z on $[-3\sigma_z, 3\sigma_z]$ with opacity proportional to $p(z)$. Bottom row of (a)-(c): Z^{true} and Z^{MAP} are plotted vs. X. The MAP estimated parameters (i) have *significantly* higher log-posterior probability than the ground-truth, (ii) Z^{MAP} memorizes the data (violating the assumptions in Section 4.3), and (iii) as a result, the learned functions explain the observed data poorly and over-estimate aleatoric uncertainty.

that are distributed like $p(z)$ (and hence do not violate assumption 2), but that are dependent on $x_1, \ldots, x_N$ (e.g. $\widehat{z}_1, \ldots, \widehat{z}_N$ are sorted such that small $\widehat{z}$'s are paired with small x's and vice versa), thus violating assumption 1. Alternatively, consider $\widehat{z}_1, \ldots, \widehat{z}_N$ that are independent of the x's, but that are not distributed like the $p(z)$ (e.g. $\widehat{z}_1, \ldots, \widehat{z}_N$ are still distributed like a Gaussian but with a different variance), thereby only violating assumption 1. We also note that, while there may be other modeling assumption that the joint posterior mode violates, these are the two assumptions we found to be important empirically and defer investigation of other modeling assumption to future work.

On three synthetic data-sets (for which we know the ground-truth), we next empirically demonstrate that under the joint posterior mode, the weights parameterize a function that explains the observed data poorly, and the latent inputs violate the above modeling assumptions. In Section 5 we demonstrate empirically that, since the mean-field variational posterior is closer to the MAP than to the ground-truth solution, it similarly violates these two modeling assumptions and therefore results in posterior predictives that explain the data poorly and generalize poorly. Then, in Section 6, we use this insight to propose a new inference method that enforces these modeling assumptions explicitly.

4.4 Empirical Demonstration of Asymptotic Bias of Joint Posterior Mode

In Figure 2, we empirically demonstrate that (1) the posterior mode $W^{\text{MAP}}, Z^{\text{MAP}}$, is not located at the ground-truth parameters $W^{\text{true}}, Z^{\text{true}}$, (2) that W^{MAP} parametrizes functions that generalize poorly and misestimate aleatoric uncertainty, by putting mass where there is no data, and (3) that Z^{MAP} violates the two modeling assumptions from Section 4.3. In these experiments, we optimize for the MAP using gradient descent, selecting the best solution across 9 random initializations / 1 initialization at the ground-truth $W^{\text{true}}, Z^{\text{true}}$. In Table 2d, we see that the observed data has a significantly higher log posterior probability under the MAP solution $W^{\text{MAP}}, Z^{\text{MAP}}$ than it does under the ground-truth parameters $W^{\text{true}}, Z^{\text{true}}$, confirming that indeed $W^{\text{true}}, Z^{\text{true}} \neq W^{\text{MAP}}, Z^{\text{MAP}}$. In Figure 2, we visualize the predictive distribution corresponding to W^{MAP} for data drawn from the same distribution as the training data. For each of the three data-sets, we see that the W^{MAP} parametrizes a function that puts mass where there is no data (i.e. over-estimates aleatoric uncertainty), thereby explaining the observed data poorly. Correspondingly, we also see that the Z^{MAP} memorized the data and exhibits a distribution different than the prior, violating the two modeling assumptions from Section 4.3.

We next show that because mean-field variational inference prefers solutions closer to the MAP than to the ground-truth, it suffers from the same issues at the MAP—mean-field VI yields approximations of $p(W|\mathcal{D})$ that explain the observed data poorly because they violate the generative modeling assumptions.

5. Effect of Asymptotic Bias of the BNN+LV Posterior Mode on Variational Inference

Whereas in Section 4 we focused on the asymptotic bias of the BNN+LV joint posterior mode, in this section we unfold the consequence of this asymptotic bias on mean-field variational inference. We show that because solutions returned by mean-field variational

inference are in practice closer to the MAP than they are to the ground-truth parameters, they violate the assumptions made in the generative model. As a result, they correspond to posterior predictives that under-fit the observed data and misestimate uncertainty.

Mean-field VI returns posterior predictives that under-fit the data and violate generative modeling assumptions. We follow the standard practice of: initializing the parameters of the mean-field variational family (Equation 4) using a random initialization (described in Appendix F.1), maximizing the ELBO (in Equation 5), and selecting models with the highest validation log-likelihood over 10 random restarts. In Figure 4 (blue column), we see that traditional inference posterior predictives that under-fit the data and misestimate the noise. Furthermore, the figure shows that the latent variables recovered from mean-field VI exhibit strong dependence on the data (i.e. memorized the data), thereby violating our generative modeling assumptions (Section 4.3). Whereas in Figure 4 we offer qualitative results, in Section 7 we demonstrate that this pathology occurs on a variety of synthetic and real data-sets both qualitatively and quantitatively. We next argue that traditional inference results models that under-fit the data and violate our generative modeling assumptions because, in practice, traditional inference yields posterior distributions closer to the MAP than to the ground-truth.

Mean-field VI may return solutions closer to the MAP than to the ground-truth. To show that mean-field VI typically returns solutions closer to the MAP than to the ground-truth, we show that when initialized at the MAP, mean-field VI yields a higher ELBO than when fixed at the ground-truth, or when initialized at the ground-truth (i.e. the ELBO prefers models initialized at the MAP). Correspondingly, MAP-initialized inference also yields posterior predictives that under-fit the data. This experiment suggests that in practice, the optima of the ELBO commonly found via gradient descent share more characteristics with the problematic MAP of the joint posterior than with the ground-truth data-generating parameters.

In this experiment, we initialize the variational parameters using each of the following schemes (after which we optimize the ELBO):

- Ground-Truth (GT): We initialize the variational means to $W^{\text{true}}, Z^{\text{true}}$. Holding the variational means fixed, we optimize the variational variances until convergence.
- MAP: We compute the MAP of $p(W, Z|\mathcal{D})$ via gradient descent, selecting the best of 10 random restarts (1 initialized at $W^{\text{true}}, Z^{\text{true}}$, and 9 with W^{true} and with Z sampled randomly from the prior). We then initialize the variational means to $W^{\text{MAP}}, Z^{\text{MAP}}$. Holding the variational means fixed, we optimize the variational variances until convergence.
- Random: We randomly initialize the variational parameters.
- $\text{NCAI}_{\lambda=0}$: For completeness, we even use the initialization we propose later in this work (in Section 6).

For each of the above initialization schemes, we initialize and run mean-field VI 10 times, selecting the models with the highest ELBO. We also compare the ELBO optimized from each of the above initializations to the ELBO evaluated at the ground-truth initialization

	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5
Lowest ELBO	Fixed @ GT	Fixed @ GT	Fixed @ GT	Fixed @ GT	Fixed @ GT
↓	Random	Random	GT	Random	GT
	GT	GT	Random	GT	Random
	$\text{NCAI}_{\lambda=0}$	MAP	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$
Highest ELBO	MAP	$\text{NCAI}_{\lambda=0}$	MAP	MAP	MAP

(a) Bimodal data-set

	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5
Lowest ELBO	Fixed @ GT	Fixed @ GT	Fixed @ GT	Fixed @ GT	Fixed @ GT
↓	GT	GT	GT	GT	GT
	MAP	MAP	MAP	MAP	MAP
	Random	Random	Random	Random	Random
Highest ELBO	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$	$\text{NCAI}_{\lambda=0}$

(b) Heavy-Tail data-set

	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5
Lowest ELBO	Fixed @ GT	Fixed @ GT	Random	Fixed @ GT	Fixed @ GT
↓	Random	Random	Fixed @ GT	Random	Random
	NCAI	GT	GT	GT	GT
	GT	MAP	MAP	NCAI	MAP
Highest ELBO	MAP	NCAI	NCAI	MAP	NCAI

(c) Depeweg data-set

Table 1: **Mean-field VI returns solutions closer to the MAP than to the ground-truth parameters.** For each of the tables above (each corresponding to a different synthetic data-set, detailed in Appendix G), and for each initialization scheme (Section 5), we initialize and run mean-field VI 10 times, selecting the run with the highest ELBO. We then sort the random initialization from lowest to highest ELBO. These tables show that, both across different data-sets, as well as across different draws from the same data-generating process, MAP initialization results in a higher ELBO than initialization at the ground-truth (GT). Even more problematically, it results in a higher ELBO than the ELBO evaluated directly at the ground-truth (Fixed @ GT).

("Fixed @ GT"). We sort the resultant ELBOs from lowest to highest, and repeat this whole experiment 5 times (for 5 draws of data-sets from each synthetic data-generating process).

Table 1 shows the results. If we were to *only* consider the two ground-truth initializations ("Fixed @ GT" and "GT") and the MAP initialization, we find that we find that *across all* three synthetic data-sets *and across all* 5 draws of each of these data-sets, the MAP initialization yields a higher ELBO (i.e. the relative ordering of blue, light purple and dark purple in *all three* tables is always the same). Correspondingly, inference initialized at the MAP yields posterior predictives that fit the data poorly. For instance, in the Bimodal data-set (Figure 3b), in draws 2, 3 and 4, the posterior predictive places mass where there is no data when $x > 0.5$, in draw 1, the model places a little more mass above the observed data at $x = 0.75$, and in draw 5, the posterior predictive mean is significantly biased relative to the ground-truth. This result helps us explain why in practice, mean-field VI for BNN+LV yields poor quality models: it shows that there exist several different sets of variational

parameters ϕ that are preferred by the ELBO over the ground-truth parameters, but that retain undesirable properties of the joint posterior mode.

Of course, in practice we cannot use the ground-truth initializations, and we do not want to use the MAP initialization (for reasons mentioned here), so we are left with only one option: random initialization. However, as already shown in Figure 4, random initialization also yield poor results in practice. So in practice, what initialization should we use? In Section 6, we propose a novel and practical initialization, $\text{NCAI}_{\lambda=0}$, which we find to be empirically effective. We include this initialization, as well as the random initialization, in Table 1 because these two initializations will help us explore whether these undesirable optima of the ELBO are local or global, discussed next.

Global vs. local optima of the ELBO. While we have shown that MAP-initialized inference yields higher ELBOs than the ELBO at the ground-truth, and returns posterior predictives that explain the observed data poorly, does this phenomenon occur at the local or global optima of the ELBO? That is, even the *best* imperfect approximation of the joint posterior under the ELBO would be biased towards the undesirable posterior mode? We conjecture here that, for data-sets for which the ELBO is a loose bound to the log marginal likelihood, the global optima would exhibit the same properties as the MAP, while for data-sets in which the ELBO is tight, only local optima will exhibit these properties.

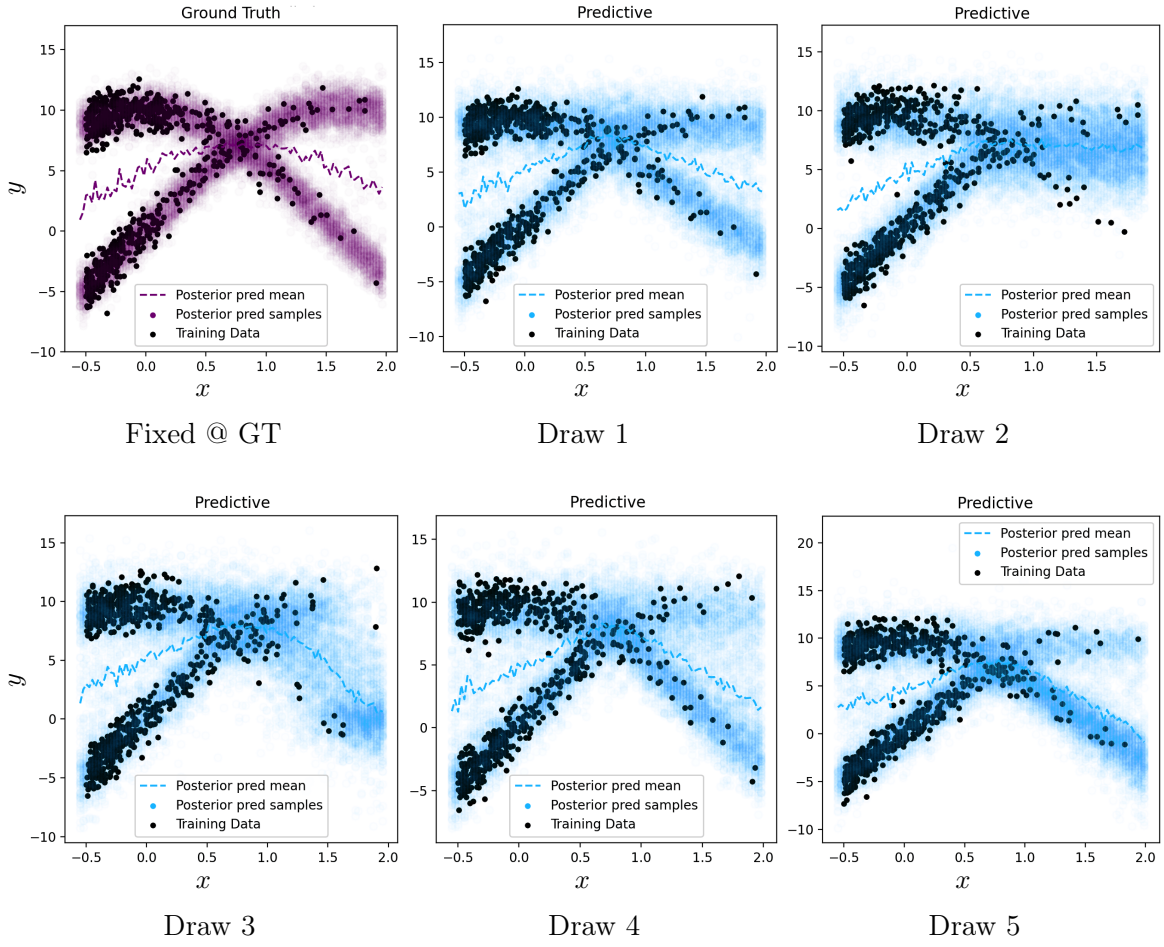
The Bimodal data-set is one for which we expect the ELBO to be loose, and therefore for the global optima of the ELBO to be biased towards the MAP. In this data-set, the ground-truth function uses z as a binary indicator to select which function to generate. As such, the true posterior of z given the ground-truth function $p(Z|\mathcal{D}, W^{\text{true}})$ is highly skewed and poorly approximated by a mean-field Gaussian. This causes the gap between the ELBO and the true marginal likelihood, $\mathbb{E}_{p(Z)}[p(Y|X, W^{\text{true}}, Z)]$, to be particularly high. The results in Table 1 confirm that for the Bimodal data-set, the global optima of the ELBO may be problematic, since the MAP-initialization results in the highest ELBO across all initializations, whereas for the other data-sets, there exist other initializations that yield higher ELBOs, e.g. $\text{NCAI}_{\lambda=0}$ (which, as we show in Section 7, also results in a better fit).

6. Mitigating Inference Challenges Caused by Model Non-Identifiability

In Section 4 we showed that the BNN+LV posterior mode is asymptotically biased towards functions that generalize poorly. We furthermore showed (in Section 4.3) that the parameters preferred by the posterior violate two modeling assumptions—(1) that the latent variable z is independent of the input x and (2) that the latent variable z is drawn from a normal distribution. More surprisingly, in Section 5 we showed that empirically, mean-field VI returns solutions that violate the same modeling assumptions, and as a consequence, that the resultant posterior predictives generalize poorly and misestimate uncertainty. This leads us to hypothesize that when these modeling assumptions are satisfied, the the resultant posterior predictive will generalize well and provide appropriate estimations of uncertainty. In this section, we therefore develop a method to enforce these modeling assumptions *explicitly* during variational inference. We call this method Noise-Constrained Approximate Inference (NCAI), since it restricts the variational family to treat the latent input variable z as identically distributed and independent of x.

Initialization	Draw 1	Draw 2	Draw 3	Draw 4	Draw 5
Fixed @ GT	−3662.098	−3663.944	−3699.185	−3671.299	−3698.769
Ground-Truth	−2117.360	−2102.342	−2151.512	−2127.975	−2152.788
Random	−2150.645	−2120.855	−2146.312	−2139.177	−2150.549
NCAI$_{\lambda=0}$	−2093.275	**−2086.503**	−2115.208	−2104.083	−2113.348
MAP	**−2087.035**	−2088.970	**−2105.432**	**−2083.557**	**−2082.548**

(a) The highest ELBO (of 10 restarts) using different initialization schemes. MAP initialization results in the highest ELBO in 4 of the 5 draws.



(b) Visualization of the posterior predictive, resulting from optimizing the MAP-initialized ELBO. Relative to the ground-truth (purple), posterior predictives learned via MAP-initialization do not explain the observed data well; they place mass where there is no data.

Figure 3: **Mean-field VI returns solutions closer to the MAP than to the ground-truth parameters.** We draw the Bimodal data-set (Depeweg et al. (2018), detailed in Appendix G) 5 times. The table displays the highest ELBO (of 10 restarts) using different initialization schemes, and the figures show the fit of the posterior predictive from MAP-initialized VI. In 4 of 5 draws of the data, the MAP initialization yields the highest ELBO, and corresponds to poor fits on the training data (blue) relative to the ground-truth (purple)—the learned posterior predictives all place mass where there is no data.

Our method consists of two steps: first, an intelligent model-assumption satisfying initialization, and second, variational inference using a variational family constrained to satisfy the modeling assumptions from Section 4.3.

Step 1: Model-Satisfying Initialization. Since local optima are a major concern in BNN+LV inference, we start with settings of the variational parameters ϕ that satisfy the properties implied by our generative model (Equation 1). We initialize the variational means μ_{w_i} of the weights (except for weights associated with the input noise) with those of a deterministic neural network trained on the same data, based on the observation that a neural network is often able to capture the trend of the data (but not the uncertainty). We then initialize the variational means μ_{z_n} of the latent noise to 0, to ensure that in the early stages of optimization, the model is forced to explain the data using W (as opposed to by memorizing it with Z). Lastly, we initialize all variational variances randomly.

Step 2: Inference with Noise-Constrained Variational Family. We further ensure that the two key modeling assumptions—that the noise variables z are drawn *independently* of x and i.i.d from the *prior* $p(z)$—remain satisfied during training by restricting our variational family. Specifically, we construct a variational family by filtering out distributions from the mean-field variational family (Equation 4) that do not obey our modeling assumptions:

$$\mathcal{Q} = \{q_\phi(W, Z|\mathcal{D}) : \underbrace{I_\phi(x; z) = 0}_{\text{assumption 1}} \text{ and } \underbrace{D[q_\phi(z)||p(z)] = 0}_{\text{assumption 2}}\}, \tag{9}$$

Here, $I_\phi(x; z)$ quantifies the statistical dependence between the x's and z's under the posterior,

$$I_\phi(x; z) = D[q_\phi(z|x)p(x)||q_\phi(z)p(x)], \tag{10}$$

where $q_\phi(z|x)$ is the approximate posterior with y marginalized out (since we only have a single y associated with every x, $q_\phi(z|x)$ is approximated with $q_\phi(z|x, y)$.) $D[q_\phi(z)||p(z)]$ quantifies the "distance" between the approximated aggregated posterior $q_\phi(z)$ and the prior $p(z)$. The aggregated posterior is the posterior $q_\phi(z|x, y)$ marginalized over the observed data, approximated as follows (Makhzani et al., 2015):

$$q_\phi(z) = \mathbb{E}_{p(x,y)}\left[q_\phi(z|x, y)\right] \approx \frac{1}{N}\sum_{n=1}^{N} q_\phi(z_n|x_n, y_n), \tag{11}$$

We note that while each posterior $q_\phi(z|x, y)$ can be an arbitrary distribution, the aggregate posterior $q_\phi(z)$ must recover the prior when inference is exact; that is, when $q_\phi(z|x, y) = p(z|x, y)$, we have that $q_\phi(z) = \mathbb{E}_{p(x,y)}[q_\phi(z|x, y)] \approx \mathbb{E}_{p(x,y)}[p(z|x, y)] = p(z)$.

Together, these two constraints explicitly enforce both assumptions, respectively. We emphasize that since both constraints are satisfied by the true posterior (i.e. when $q_\phi(W, Z) = p(W, Z|\mathcal{D})$), these constraints are not at odds with the model—they simply help select a posterior approximation that retains the desired properties of the original model. Moreover, since the two constraints are orthogonal (i.e. satisfying one does not imply satisfying the other—see Section 4.3), both are needed.

Now, using this noise-constrained mean-field variational family, we perform variational inference:

$$\text{argmin}_{q_\phi \in \mathcal{Q}} D_{\text{KL}}[q_\phi(W, Z|\mathcal{D})\|p(W, Z|\mathcal{D})]. \tag{12}$$

As with standard variational inference, once we have the variational approximation that minimizes Equation 12, we use it to compute a Monte-Carlo estimate of the posterior predictive in Equation 3:

$$p(y^*|x^*, \mathcal{D}) = \iint p(y^*|x^*, z^*, W)p(z^*)dz^*p(W|\mathcal{D})dW$$
$$\approx \frac{1}{S}\sum_{s=1}^{S} p(y^*|x^*, z^{(s)^*}, W^{(s)}), \quad z^{(s)^*} \sim p(z^*), \quad W^{(s)}, Z^{(s)} \sim q_\phi(W, Z|\mathcal{D}).$$

Variational Inference with the Noise Constrained Variational Family. Since performing inference over the constrained ϕ-space is challenging, we equivalently re-write Equation 12 as:

$$\begin{aligned} \text{argmin}_\phi D_{\text{KL}}[q_\phi(W, Z|\mathcal{D})||p(W, Z|\mathcal{D})] \quad & \text{s.t} \\ & I_\phi(x; z) = 0, \\ & D[q_\phi(z)||p(z)] = 0. \end{aligned} \tag{13}$$

and solve Equation 13 by gradient descent on the Lagrangian:

$$\mathcal{L}_{\text{NCAI}}(\phi) = -\text{ELBO}(\phi) + \lambda_1 \cdot I_\phi(x; z) + \lambda_2 \cdot D[q_\phi(z)||p(z)]. \tag{14}$$

We emphasize that even though in practice, using our proposed variational family requires solving a constraint optimization problem (similar to posterior regularization (Zhu et al., 2014)), the theoretical justification of our method nevertheless does not deviate from the standard application of variational inference with a specific choice of variational family.

Empirical properties of the constraints. We note that even though the two constraints are theoretically satisfied by the true posterior, in practice, the constraints cannot be minimized to 0 completely. Firstly, even given the best mean-field posterior approximation $\phi^* = \text{argmin}_\phi - \text{ELBO}(\phi)$, the aggregated posterior $q_\phi(z)$ (used in both constraints) may not equal $p(z)$ due to approximation error. Secondly, it is not possible to estimate $I_\phi(x; z)$ without bias: $I_\phi(x; z)$ depends on $q_\phi(z|x)$, which is estimated by integrating out y from $q_\phi(z|x, y)$, and since only one y is observed for every x, $q_\phi(z|x)$ is simply estimated with $q_\phi(z|x, y)$ (see Appendix D.1 for details). While $q_\phi(z|x)$ should approximate $q_\phi(z)$ (i.e. $I_\phi(x; z) = 0$), in general $q_\phi(z|x, y)$ does not approximate $q_\phi(z)$ (i.e. $D[q_\phi(z|x, y)p(x, y)||q_\phi(z)p(x, y)] > 0$). As such, even with no approximation error (i.e. $q_\phi(z|x, y) = p(z|x, y)$), replacing $q_\phi(z|x)$ with $q_\phi(z|x, y)$ in Equation 10 may yield $I_\phi(x; z) > 0$.

These properties of the constraints mean that in practice, we cannot solve the Lagrangian from Equation 14. As such, we instead relax the equality constraints by optimizing Equation 14, for fixed lambdas, selected to maximize validation log-likelihood, as commonly done in the generative modeling literature (Zhao et al., 2018). These properties of the constraints also have two implications on the evaluation of NCAI (Section 7): (1) quantitatively one should not expect NCAI to minimize both constraints all the way to 0, (2) qualitatively the means of $q_\phi(z|x, y)$, while not independent of (x, y), should still exhibit less dependence than simply having memorized the data. In Section 7, we show that using NCAI, the constraints are better satisfied than when using vanilla mean-field VI (MFVI), and as a consequence the learned posterior predictives generalize better.

Differentiable and Easy-to-Optimize Proxies for the Two Constraints. While the KL-divergence in both constraints may seem like a differentiable, easy-to-optimize choice, we find that in practice it is in fact not. In Appendix D, we describe why mutual information in assumption (1) is intractable to compute using KL-divergence; we empirically show how traditional divergences (e.g. Jensen-Shannon, reverse/forward-KL divergences) in assumption (2) can all be trivially minimized by inflating the variational variances; finally, we describe our choices of differential and tractable proxies for these constraints that do not suffer from these issues. To encourage the aggregated posterior to match the prior over z, we propose a proxy based on the Henze-Zirkler statistical test for Gaussianity (Henze and Zirkler, 1990). To minimize $I_\phi(x; z)$, we instead propose a proxy that penalizes the linear correlation between x and z, and that penalizes non-linear correlation between x and z by penalizing the linear correlation between y (which depends non-linearly on x) and z. See Appendix D for the full definitions and justifications of both proxies. While empirically effective (later shown in Section 7), these proxies are nonetheless somewhat heuristic in nature. We therefore consider this method as a demonstration of validity of our theoretical analysis.

7. Experiments

In this section, we demonstrate that on a wide range of data-sets, mean-field variational inference for BNN+LV suffers from the theoretical issues we identify in Section 4. We also demonstrate that NCAI improves the quality of approximate inference for this class of models. That is, we show that latent inputs inferred by NCAI better satisfy modeling assumptions, and as a result, the learned posterior predictives are closer to the ground-truth data-generating models—they generalize better and do not misestimate uncertainty.

7.1 Setup

Data-sets. We consider 5 synthetic data-sets that are frequently used in heteroscedastic regression literature (Goldberg, Williams, Yuan, Depeweg and Heavy Tail), as well as 6 real data-sets with different patterns of epistemic and aleatoric uncertainty (Lidar, Yacht, Energy Efficiency, Airfoil, Abalone, Wine Quality Red), all of which are described in Appendix G. Each data-set is split into 5 random train/validation/test sets. For every split of each data-set, each method is evaluated on the best-learned posterior predictive (according validation log-likelihood) out of 10 random restarts (see Appendix F.1 for details).

Experimental Setup. We use neural networks with LeakyReLU activations with $\alpha = 0.01$ in all experiments. We set the prior variances σ_z^2, σ_w^2 using empirical Bayes and grid-search over remaining hyper-parameters. For optimization, we use Adam (Kingma and Ba, 2014) with a learning rate of 0.01, train for 30,000 epochs (and verify convergence). Lastly, for each method, we select the best hyper-parameters using the average log-likelihood on the validation set. Full details are in Appendix F.

Baselines. We compare NCAI on BNN+LV with unconstrained mean-field VI (Blundell et al., 2015). We also compare selecting constraint strength parameters, λ_1, λ_2, of NCAI through cross-validation (denoted NCAI_λ) against fixing λ_1, λ_2 at zero (denoted $\text{NCAI}_{\lambda=0}$). Finally, we compare the performance of BNN+LV (for all inference methods) with that of a BNN.

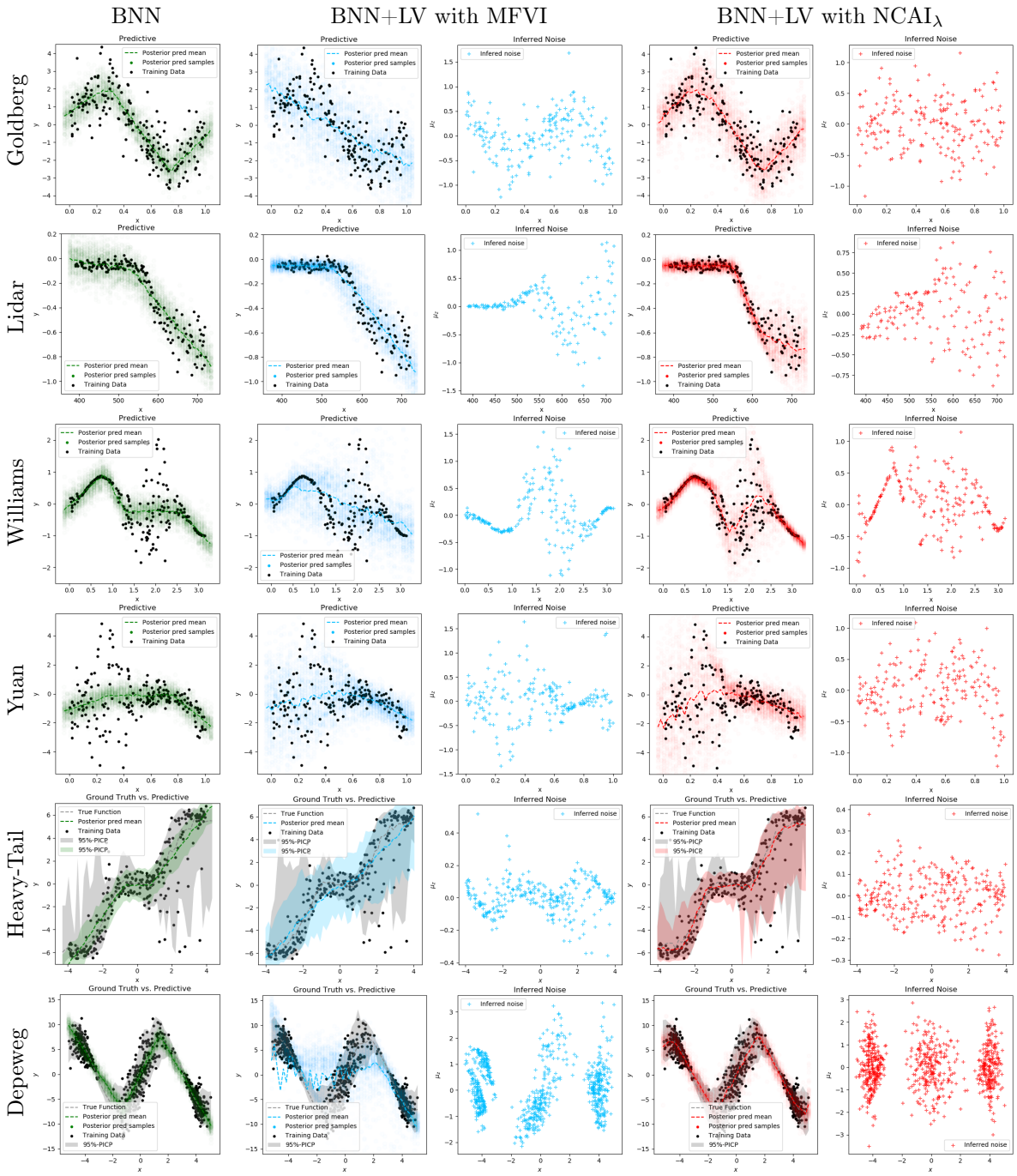


Figure 4: **Comparison of posterior predictives.** BNN (green) captures trend but underestimates variance; BNN+LV with mean-field VI (blue) captures more variance, but infers z's that are dependent on the data, shown in scatter plot. BNN+LV with NCAI$_\lambda$ (red) best captures heteroscedasticity and infers z's that best resemble white noise (shown in scatter plot). For the bottom two rows, since the true function is known, it is visualized in gray.

Mutual Information between x and z (Synthetic Data)

Inference	Heavy Tail	Goldberg	Williams	Yuan	Depeweg
MFVI	0.243 ± 0.079	0.229 ± 0.113	0.982 ± 0.121	$\mathbf{0.24 \pm 0.129}$	0.428 ± 0.04
$\mathbf{NCAI}_{\lambda=0}$	0.051 ± 0.049	$\mathbf{0.02 \pm 0.024}$	$\mathbf{0.519 \pm 0.091}$	0.283 ± 0.112	$\mathbf{0.032 \pm 0.017}$
$\mathbf{NCAI}_{\lambda}$	$\mathbf{0.036 \pm 0.04}$	0.046 ± 0.067	$\mathbf{0.519 \pm 0.091}$	0.283 ± 0.112	$\mathbf{0.032 \pm 0.017}$

Table 2: Comparison of *mutual information* between z and x on synthetic data-sets ($\pm$ std). Across all but one of the data-sets, NCAI_{λ} training infers z's that have the least mutual information in comparison mean-field VI (MFVI). Additional evaluations of model assumption satisfaction are in Appendix H.

Henze-Zirkler Test-Statistic (Synthetic Data)

Inference	Heavy Tail	Goldberg	Williams	Yuan	Depeweg
MFVI	4.701 ± 5.439	0.918 ± 0.41	$\mathbf{6.445 \pm 2.818}$	$\mathbf{5.252 \pm 5.607}$	6.408 ± 2.439
$\mathbf{NCAI}_{\lambda=0}$	7.137 ± 5.436	0.621 ± 0.234	7.248 ± 2.598	8.091 ± 5.185	$\mathbf{0.792 \pm 0.357}$
$\mathbf{NCAI}_{\lambda}$	$\mathbf{0.027 \pm 0.011}$	$\mathbf{0.026 \pm 0.038}$	7.248 ± 2.598	8.091 ± 5.185	$\mathbf{0.792 \pm 0.357}$

Table 3: Comparison of model assumption satisfaction (in terms of the HZ metric) on synthetic data-sets ($\pm$ std). Across all but one data-sets, NCAI_{λ} training infers z's that are more Gaussian (lowest HZ) relative to mean-field VI (MFVI).

Test Log-Likelihood (Synthetic Data)

Model	Inference	Heavy Tail	Goldberg	Williams	Yuan	Depeweg
BNN	**MFVI**	-2.47 ± 0.083	-1.055 ± 0.08	-1.591 ± 0.417	-2.846 ± 0.346	-2.306 ± 0.059
BNN+LV	**MFVI**	-1.867 ± 0.078	-1.026 ± 0.056	-1.033 ± 0.156	-1.278 ± 0.164	-2.342 ± 0.048
BNN+LV	$\mathbf{NCAI}_{\lambda=0}$	-1.481 ± 0.018	$\mathbf{-0.962 \pm 0.040}$	$\mathbf{-0.414 \pm 0.184}$	$\mathbf{-1.211 \pm 0.083}$	$\mathbf{-1.973 \pm 0.049}$
BNN+LV	$\mathbf{NCAI}_{\lambda}$	$\mathbf{-1.426 \pm 0.042}$	-0.963 ± 0.041	$\mathbf{-0.414 \pm 0.184}$	$\mathbf{-1.211 \pm 0.083}$	$\mathbf{-1.973 \pm 0.049}$

Table 4: Comparison of *test log-likelihood* on synthetic data-sets ($\pm$ std). For all data-sets, BNN+LV trained with NCAI outperforms BNN+LV and BNN trained with mean-field VI (MFVI).

Evaluation. We evaluate the learned posterior predictives for quality of fit using test average log-likelihood, RMSE, and calibration of the posterior predictives (using the 95% Prediction Interval Coverage Probability (PICP), and the 95% Mean Prediction Interval Width (MPIW)). We also check whether they satisfy the two generative modeling assumptions from Section 4.3. Specifically, we check whether z is independent of x under the posterior using mutual information (estimated via a non-parametric nearest neighbor method (Kraskov et al., 2004)), and we check that z is Gaussian under the posterior predictive using the Henze-Zirkler test-statistic for normality (as well as using Jensen-Shannon, and forward/reverse KL divergences between the aggregated posterior and the prior). We note that due to the reasons mentioned in Appendix D—that traditional variances are trivially minimized by inflating the variational variances—we primarily focus our analysis on the Henze-Zirkler metric during evaluation, even though it is also used in our objective (Equation 27). Details about evaluation metrics are found in Appendix F.2.

Test Log-Likelihood (Real Data)

Model	Inference	Abalone	Airfoil	Energy	Lidar	Wine	Yacht
BNN	**MFVI**	-1.248 ± 0.153	-0.995 ± 0.143	$\mathbf{1.281 \pm 0.171}$	-0.31 ± 0.069	-1.143 ± 0.027	0.818 ± 0.187
BNN+LV	**MFVI**	-0.843 ± 0.071	-0.512 ± 0.083	0.573 ± 0.288	0.129 ± 0.131	-1.709 ± 0.22	0.638 ± 0.121
BNN+LV	**NCAI**$_{\lambda=0}$	$\mathbf{-0.831 \pm 0.086}$	$\mathbf{-0.462 \pm 0.056}$	0.862 ± 0.138	$\mathbf{0.269 \pm 0.107}$	-1.147 ± 0.025	$\mathbf{0.832 \pm 0.077}$
BNN+LV	**NCAI**$_{\lambda}$	$\mathbf{-0.831 \pm 0.086}$	$\mathbf{-0.462 \pm 0.056}$	0.898 ± 0.452	0.263 ± 0.11	$\mathbf{-0.849 \pm 0.038}$	$\mathbf{0.832 \pm 0.077}$

Table 5: Comparison of *test log-likelihood* on real data-sets ($\pm$ std). For all but one of the data-sets, BNN+LV trained with NCAI outperforms BNN+LV and BNN trained with mean-field VI (MFVI).

7.2 Results

The BNN+LV joint posterior prefers models that do not explain the observed data well. In Figure 2, we compare the ground-truth function and the function learned via the MAP estimate over W, Z. We show that in comparison to the true parameters $W^{\text{true}}, Z^{\text{true}}$, the MAP-inferred parameters $W^{\text{MAP}}, Z^{\text{MAP}}$ are (1) scored *significantly* higher under the log-posterior, (2) have Z^{MAP} that violate our modeling assumptions (Section 4.3), and (3) have W^{MAP} that generalize poorly and misestimates aleatoric uncertainty.

NCAI better satisfies generative modeling assumptions than mean-field VI. As shown in Tables 2 and 10, across all synthetic and real data-sets, NCAI learns posterior predictives for which z and x are significantly less dependent under the posterior than those learned via mean-field VI, thereby satisfying assumption 1. As shown in Tables 3 and 11, across all synthetic and real data-sets, NCAI learns posterior predictives for which the aggregated posterior better matches the prior, thereby satisfying assumption 2. Furthermore, on synthetic data, Figure 4 confirms qualitatively from scatter plots of variational posterior mean of z vs. x that NCAI better satisfies modeling assumptions. Additional evaluation metrics of model assumption satisfaction can be found in Appendix H (described in Appendix F.2).

Learned posterior predictives perform better when generative model assumptions are satisfied. On all 1-dimensional data-sets, Figure 4 shows a qualitative comparison of the posterior predictive distributions of BNN+LV trained with NCAI$_\lambda$ compared with benchmarks. We see that, as expected, BNNs underestimate the posterior predictive uncertainty, whereas BNN+LV with mean-field VI improves upon the BNN in terms of log-likelihood by expanding posterior predictive uncertainty nearly symmetrically about the predictive mean. The predictive distribution obtained by BNN+LV trained with NCAI, however, captures the asymmetry of the observed heteroscedasticity. Furthermore, its predictive mean better captures the overall trend and its predictive uncertainty is better-calibrated.

Across all synthetic and real data-sets (apart from Energy Efficiency), when modeling assumptions are satisfied (i.e. when NCAI is used), the learned posterior predictives have higher average log-likelihood (Tables 4, 5 for synthetic and real, respectively). On Energy Efficiency, the BNN performs best in terms of test log-likelihood, but drastically underestimates the uncertainty in the data; specifically, the 95%-PICP and MPIW show that BNN has a small predictive interval width that only covers about 80% of the data, whereas NCAI overs about 94% of the data (see Table 14 details). This is because the BNN, when properly trained, is able to capture the trends in the data but tends to underestimate the

variance (log-likelihood and calibration)—this tendency is especially apparent in the presence of heteroscedastic noise.

Selecting between NCAI$_{\lambda=0}$ and NCAI$_{\lambda>0}$. Generally, we observe that on data-sets in which the noise is roughly symmetric around the posterior predictive mean (as in the Goldberg, Yuan, Williams, Lidar, and Depeweg data-sets) NCAI$_{\lambda=0}$ and NCAI$_{\lambda>0}$ perform comparably well on average test log-likelihood. However, when the noise is skewed around the posterior predictive mean (like in the HeavyTail data-set), we find that NCAI$_{\lambda>0}$ outperforms NCAI$_{\lambda=0}$. This is because NCAI$_{\lambda=0}$ first fits the variational parameters of the weights to best capture as best as possible, often fitting a function that represents the mean. After the warm-start, when training with respect to the variational parameters of the z's, the uncertainty is increased about the mean to best capture the data, often in a way that does not significantly alter the parameters of the weights, thereby resulting in a posterior predictive with symmetric noise.

7.3 Application: Uncertainty Decomposition

By explicitly modeling sources of epistemic uncertainty, W, and aleatoric noise, z and ϵ, the BNN+LV model can decompose the uncertainty in its posterior predictive distribution. This decomposition can improve performance on downstream tasks that rely on exploiting uncertainty in data. For example, Depeweg et al. (2018) shows that accurate decomposition improves active-learning with BNN+LV in the presence of complex noise; the authors also formulate a new 'risk-sensitive criterion' for safe model-based RL based on the decomposition of predictive uncertainties in BNN+LV.

Following Depeweg et al. (2018), we quantify the uncertainty in the posterior predictive using entropy (see Equations 6 and 7). Using Hamiltonian Monte Carlo (HMC) (Neal et al., 2011) as the "gold-standard" approximation of the true posterior, we compare the uncertainty decomposition learned by BNN+LV and NCAI with that learned by HMC. Figure 6 shows that like HMC, NCAI has appropriately high total and aleatoric uncertainties at x's for which there is a high variance in y, as well as high epistemic uncertainty for x's near the boundary of the data. In contrast, BNN+LV trained via mean-field VI does not. This is evidence that our method learns a decomposition closer to that given by the "ground-truth" posterior predictive.

We see that BNN+LV likelihood non-identifiability negatively impacts the accuracy of uncertainty decompositions: BNN+LV trained with mean-field VI, while able to reconstruct training data well, produces inaccurate uncertainty decompositions. In contrast, NCAI consistently produces aleatoric and epistemic uncertainties that align well with those produced by HMC (more details in Appendix E).

8. Discussion

Non-identifiability negatively impacts inference in theory and practice. In Section 4 we show that BNN+LV models are meaningfully non-identifiable under the likelihood and, as a consequence, the posterior mode is asymptotically biased towards models that both explain the data poorly and generalize poorly, regardless of the choices of priors. We argue that approximations of the joint posterior via mean-field VI are negatively impacted by this

asymptotic bias, resulting in posterior predictives that will similarly explain the data poorly, generalize poorly and misestimate predictive uncertainty. In Section 7 we demonstrate the negative effect on mean-field VI using a variety of synthetic and real data-sets.

Enforcing modeling assumptions explicitly during training mitigates the effects of non-identifiability. In Section 4.3, we show that model parameters that are scored as likely under the posterior often violate the generative modeling assumption—that the latent variable is generated i.i.d from the prior. Based on this analysis, we develop a two-step method, NCAI, that explicitly enforces these assumptions during training. We demonstrate on both synthetic and observed data that in enforcing these assumptions explicitly, we recover posterior predictives that generalize better and do not misestimate predictive uncertainty.

Can one alter the original BNN+LV model to correct for the asymptotic bias of the joint posterior? The insights from this work suggest that this is likely not possible. As N increases, any non-degenerate prior over z (that is independent of x under the generative process) will have a non-vanishing effect on the joint posterior mode. One may be tempted to construct a prior over z that weakens as N increases; however, when this prior is sufficiently weak, this reduces to the "incidental parameters problem" (Neyman and Scott, 1948; Lancaster, 2000). This work therefore highlights the more general challenge of approximate inference for Bayesian models that have both global parameters and local latent variables.

Limitations and future work. In this work, we prove that the joint posterior $p(W, Z|\mathcal{D})$ is biased towards models that generalize poorly; however, it is not currently known whether the posterior predictive of BNN+LVs (computed using the marginal of the weights $p(W|\mathcal{D})$) is consistent. We hope to study this in future work. Experimentally, we show that vanilla VI suffers from both local and global optima problems that cause the learned models to generalize poorly. In future work, we hope to extend our analysis to other inference methods, such as MCMC-based methods. Although in this work we proposed a new training framework, NCAI, which out-performs naive VI, our framework still has several limitations. Firstly, our training objective can be challenging to optimize: our proposed intelligent initialization alone does not always recover good quality models, and our proposed constraints that are trained jointly with the ELBO require additional optimization tricks to avoid local optima. Secondly, as we show in Section 6, some of the constraints cannot be estimated without bias, and more generally, the constraints cannot be optimized directly, requiring tractable proxies. In future work, we hope to both explore proxies for our constraints that are more amenable to easy optimization. As such, we regard the specific instantiation of NCAI in this paper as an extension of our theoretical analysis—as a link between the asymptotic bias of the posterior mode and the solutions returned by mean-field VI—as opposed to a method to be readily deployed in safety-critical applications. We expect that the analysis provided in this paper is useful in diagnosing poor performance of other similar models.

9. Conclusion

In this paper we identify a key issue with a promising class of flexible latent variable models for Bayesian regression—that model non-identifiability can bias the posterior mode towards model parameters that generalize poorly. By analyzing the sources of non-identifiability in BNN+LV models, we propose an approximate inference framework, NCAI, that explicitly

enforces model assumptions during training. On synthetic and real data-sets with complex patterns and sources of uncertainty, we demonstrate that NCAI better recovers posterior predictives that generalize well and accurately estimate uncertainty relative to baselines.

Acknowledgments

YY acknowledges support from NIH 5T32LM012411-04 and from IBM Research. WP acknowledges support from Harvard's IACS. We thank Melanie F. Pradier, Beau Coker and Jiayu Yao for helpful feedback and discussions.

Appendix

Table of Contents

Appendix A. Notation

Symbol	Definition
N	Number of observations (or training points)
(x_n, y_n)	The nth observed input and output, where $x_n \in \mathbb{R}^D, y_n \in \mathbb{R}^L$.
(x^*, y^*)	A point from the test-set.
z_n	The latent code corresponding to the nth observation. Since the latent variables are sampled i.i.d from the prior, $p(z_n) = p(z)$.
z^*	The latent code corresponding to x^*. Again, $p(z^*) = p(z)$.
$\mathcal{D}$	$\{(x_n, y_n)\}_{n=1}^N$
X	$\{x_n\}_{n=1}^N$
Y	$\{y_n\}_{n=1}^N$
Z	$\{z_n\}_{n=1}^N$
$p(y\|x)p(x)$ or $p(y\|x; W^{\text{true}})p(x)$	The ground-truth data-generating distribution (that generated $\mathcal{D}$).
$f(\cdot, \cdot; W)$	A neural network f parameterized by weights W.
W	The set of all neural network weights w_i.
$p(W, Z\|\mathcal{D})$	The BNN+LV joint posterior (Equation 2).
$p(W\|\mathcal{D})$	The marginal posterior of the weights, $\int_Z p(W, Z\|\mathcal{D})dZ$ (intractable).
$q_\phi(Z, W\|\mathcal{D}) = q_\phi(Z\|\mathcal{D})q_\phi(W\|\mathcal{D})$	The mean-field variational approximation of the joint posterior (Equation 4), parameterized by ϕ.
$Z^{\text{MAP}}, W^{\text{MAP}}$	The MAP of the true joint posterior: $Z^{\text{MAP}}, W^{\text{MAP}} = \text{argmax}_{Z,W} p(Z, W\|\mathcal{D})$
$Z^{\text{true}}, W^{\text{true}}$	The ground-truth data-generating weights W^{true} and latent codes $Z^{\text{true}} = \{z_n^{\text{true}}\}_{n=1}^N$ that produced the observed data.
$\widehat{Z}, \widehat{W}$	The alternative set of latent variables $\widehat{Z} = \{\widehat{z}_n\}_{n=1}^N$ and weights, defined in Equation 8, that have a higher probability under the joint posterior.
$q_\phi(z)$	The aggregated posterior (Equation 11).
$q_\phi(z\|x)$	The approximate posterior marginalized over y: $\int_y q_\phi(z\|x, y)p(y\|x)dy$. Since we only observe one y for every x, we cannot obtain unbiased estimates of this distribution in practice.

Table 6: **Notation**

Appendix B. Asymptotic Bias of the BNN+LV Joint Posterior Mode

B.1 Asymptotic Bias of 1-Node BNN+LV Posterior Mode

Consider univariate output generated by a single hidden-node neural network with LeakyReLU activation:

$$\begin{aligned} z &\sim \mathcal{N}(0, \sigma_z^2) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ y &= \max\{W(x+z), \alpha W(x+z)\} + \epsilon \end{aligned} \tag{15}$$

where α is a fixed constant in $(0, 1)$.

Theorem 1 (Asymptotic Bias of the Posterior Mode of 1-Node BNN+LV) *Fix any $W \in \mathbb{R}$ and any bounded prior $p_W(W)$ on W. Suppose that inputs $\{x_1, \ldots, x_N\}$ are sampled i.i.d from p_x with finite first and second moments μ_x, σ_x^2, and that $\{z_1, \ldots, z_N\}$ are sampled i.i.d from $\mathcal{N}(0, \sigma_z^2)$, where $\mu_x^2 + \sigma_x^2 > \sigma_z^2$. There exist a non-zero $\mathcal{C}$ such the probability that the scaled values $(\widehat{W}^{(C)}, \{\widehat{z}_n^{(C)}\}_{n=1}^N)$ are more likely than $\left(W, \{z_n\}_{n=1}^N\right)$ under the posterior approaches 1 as $N \to \infty$, for every $C \in (\mathcal{C}, 1)$.*

Proof

We assume the model in Equation 15. We denote the prior on W as p_W and suppose that it is bounded, we denote the prior on z as p_z, and we suppose that p_x, the distribution over the input x, has bounded first and second moments μ_x and σ_x^2. For any non-zero constant $0 < C < 1$, we define

$$\widehat{W}^{(C)} = W/C, \quad \widehat{z}_n^{(C)} = (C-1)x_n + Cz_n,$$

and we define $D_N^{(C)}$ to be the difference between the scaled parameters $(\widehat{W}^{(C)}, \widehat{z}_1^{(C)}, \ldots \widehat{z}_N^{(C)})$ and the ground-truth parameters $(W, z_1, \ldots z_N)$ under the joint posterior $\log p(W, Z|\mathcal{D})$.

For any set of parameters $(W, z_1, \ldots z_N)$, we show that we can find alternate parameters $(\widehat{W}^{(C)}, \widehat{z}_1^{(C)}, \ldots \widehat{z}_N^{(C)})$, that are scored as more likely under the log-posterior $\log p(W, \{z_n\}|\mathcal{D})$. To do this, we first show that the alternative parameters are scored as equally likely under the log-likelihood, and then we show that the alternate parameters are scored as more likely under the log-prior.

Since we have that

$$\max\left\{W^{\text{true}}(x + z_n^{\text{true}}), \alpha W^{\text{true}}(x + z_n^{\text{true}})\right\} = \max\left\{\widehat{W}^{(C)}(x + \widehat{z}_n^{(C)}), \alpha\widehat{W}^{(C)}(x + \widehat{z}_n^{(C)})\right\},$$

we see that the alternate parameters are as likely as the original parameters under the likelihood; that is,

$$\prod_n p(y_n|x_n, z_n^{\text{true}}, W^{\text{true}}) = \prod_n p\left(y_n|x_n, \widehat{z}_n^{(C)}, \widehat{W}^{(C)}\right).$$

Since the likelihood under both sets of parameters is equal, the difference between the log-posterior, $D_N^{(C)}$, is simply the difference of the two sets of parameters under the log-prior:

$$D_N^{(C)} = \log p_W(\widehat{W}^{(C)}) - \log p_W(W) + \sum_{n=1}^{N} \log p_z(\widehat{z}_n^{(C)}) - \log p_z(z_n).$$

We now show that as $N \to \infty$, the probability that $(\widehat{W}^{(C)}, \widehat{z}_1^{(C)}, \dots \widehat{z}_N^{(C)})$ is valued higher than $(W, z_1, \dots z_N)$ in the posterior approaches 1. That is,

$$\lim_{N\to\infty} \Pr\left[D_N^{(C)} > 0\right] = \lim_{N\to\infty} \mathbb{E}_{p_x p_z}\left[\mathbb{I}(D_N^{(C)} > 0)\right] = 1.$$

Since we are only interested in the sign of $D_N^{(C)}$, we demonstrate, equivalently, that

$$\lim_{N\to\infty} \mathbb{E}_{p_x p_z}\left[\mathbb{I}\left(\frac{1}{N} \cdot D_N^{(C)} > 0\right)\right] = 1.$$

Expanding $\frac{1}{N} \cdot D_N^{(C)}$, we obtain

$$\frac{1}{N} \cdot D_N^{(C)} = \underbrace{\frac{1}{N} \cdot \left(\log p_W(\widehat{W}^{(C)}) - \log p_W(W)\right)}_{L_N^{(C)}} + \underbrace{\frac{1}{N} \sum_{n=1}^{N} \log p_z(\widehat{z}_n^{(C)}) - \log p_z(z_n)}_{R_N^{(C)}}.$$

We note that as $N \to \infty$, the term $L_N^{(C)}$ approaches 0. Thus, it suffices to analyze the sign of $R_N^{(C)}$; that is, we compute $\Pr[R_N^{(C)} > 0]$.

Let $\mathbb{E}\left[R_N^{(C)}\right]$ and $\mathbb{V}\left[R_N^{(C)}\right]$ denote the mean and variance of the random variable $R_N^{(C)}$. We next show that $\mathbb{E}\left[R_N^{(C)}\right]$ is positive, allowing us to use Cantelli's Inequality to bound the probability that $R_N^{(C)}$ is negative under p_x and p_z as follows:

$$\begin{aligned}
\Pr\left[R_N^{(C)} < 0\right] &= \Pr\left[R_N^{(C)} - \mathbb{E}\left[R_N^{(C)}\right] < -\mathbb{E}\left[R_N^{(C)}\right]\right] \\
&= 1 - \Pr\left[R_N^{(C)} - \mathbb{E}\left[R_N^{(C)}\right] \geq -\mathbb{E}\left[R_N^{(C)}\right]\right] \\
&\leq \frac{\mathbb{V}\left[R_N^{(C)}\right]}{\mathbb{V}\left[R_N^{(C)}\right] + \mathbb{E}\left[R_N^{(C)}\right]^2}.
\end{aligned}$$

As $N \to \infty$, we show that this upper bound on $\Pr\left[R_N^{(C)} < 0\right]$ approaches 0.

We start by expanding $R_N^{(C)}$ as follows:

$$\begin{aligned}
R_N^{(C)} &= \frac{1}{N}\sum_{n=1}^{N} \log p_z(\widehat{z}_n^{(C)}) - \log p_z(z_n) \\
&= -\frac{1}{2\sigma_z^2}\cdot\frac{1}{N}\sum_{n=1}^{N}(\widehat{z}_n^{(C)})^2 - z_n^2 \\
&= \frac{1}{2\sigma_z^2}\cdot\frac{1}{N}\sum_{n=1}^{N} z_n^2 - (\widehat{z}_n^{(C)})^2 \\
2\sigma_z^2\cdot R_N^{(C)} &= \frac{1}{N}\sum_{n=1}^{N} z_n^2 - (\widehat{z}_n^{(C)})^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} z_n^2 - ((C-1)x_n + Cz_n)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(2C - C^2 - 1)\cdot x_n^2 + (1-C^2)\cdot z_n^2 + (2C-2C^2)\cdot x_n z_n \\
&= (2C-C^2-1)\cdot\frac{1}{N}\sum_{n=1}^{N} x_n^2 + (1-C^2)\cdot\frac{1}{N}\sum_{n=1}^{N} z_n^2 + (2C-2C^2)\cdot\frac{1}{N}\sum_{n=1}^{N} x_n z_n
\end{aligned}$$

The variance of $R_N^{(C)}$ can be computed as follows:

$$\mathbb{V}\left[R_N^{(C)}\right] = \frac{\mathbb{V}\left[R_1^{(C)}\right]}{N}.$$

Therefore, as N increases, the variance approaches 0 at a rate of $1/N$.

The mean of $R_N^{(C)}$ can be written as:

$$\begin{aligned}
2\sigma_z^2\cdot\mathbb{E}\left[R_N^{(C)}\right] &= (2C-C^2-1)\cdot\mathbb{E}_{p_x}\left[x^2\right] + (1-C^2)\cdot\mathbb{E}_{p_z}\left[z^2\right] + (2C-2C^2)\cdot\mathbb{E}_{p_x p_z}\left[xz\right] \\
&= (2C-C^2-1)\cdot(\sigma_x^2+\mu_x^2) + (1-C^2)\cdot\sigma_z^2 + (2C-2C^2)\cdot\mu_x\cdot 0 \\
&= (2C-C^2-1)\cdot(\sigma_x^2+\mu_x^2) + (1-C^2)\cdot\sigma_z^2
\end{aligned}$$

Thus, the mean is positive when

$$0 < \frac{\sigma_x^2+\mu_x^2-\sigma_z^2}{\sigma_x^2+\mu_x^2+\sigma_z^2} < C < 1,$$

which is satisfied whenever $\sigma_x^2+\mu_x^2 > \sigma_z^2$.

Now, by Cantelli's Inequality, we have that:

$$\Pr\left[R_N^{(C)} < 0\right] \le \frac{\mathbb{V}\left[R_N^{(C)}\right]}{\mathbb{V}\left[R_N^{(C)}\right] + \mathbb{E}\left[R_N^{(C)}\right]^2} = \frac{\mathbb{V}\left[R_1^{(C)}\right]}{\mathbb{V}\left[R_1^{(C)}\right] + N\cdot\mathbb{E}\left[R_N^{(C)}\right]^2},$$

which shows that $\Pr\left[R_N^{(C)} < 0\right]$ approaches 0 as $N \to \infty$.

■

B.2 Asymptotic Bias of 1-Layer BNN+LV Posterior Mode

Let W be the set of network weights $\{W^x, W^z, W^{\text{out}}, b^{\text{hidden}}, b^{\text{out}}\}$. Consider a data-set generated from the following model:

$$\begin{aligned} W &\sim p(W) \\ z &\sim \mathcal{N}(0, \sigma_z^2 \cdot I) \\ \epsilon &\sim N(0, \sigma_\epsilon^2) \\ a^{\text{hidden}} &= g\left(W^x x + W^z z + b^{\text{hidden}}\right), \\ y &= (a^{\text{hidden}})^\top W^{\text{out}} + b^{\text{out}} + \epsilon \end{aligned} \tag{16}$$

Theorem 2 (Asymptotic Bias of the Posterior Mode of 1-Layer BNN+LV) *Fix any set of parameters W and any bounded prior $p_W(W)$ on W. Suppose that $\{x_1, \ldots, x_N\}$ is sampled i.i.d from p_x, with finite first and second moments, and that $\{z_1, \ldots, z_N\}$ are sampled i.i.d from $\mathcal{N}(0, \sigma_z^2 \cdot I)$. There exists an alternate set of parameters $\left(\widehat{W}, \{\widehat{z}_n\}_{n=1}^N\right)$ such that the probability that these alternative parameters are valued as more likely than $\left(W, \{z_n\}_{n=1}^N\right)$ under the posterior approaches 1 as $N \to \infty$.*

Proof

The proof follows in the same fashion as the one for Theorem 1. We assume the model in Equation 16. For a given W, let $\widehat{W}$ denote the set $\{\widehat{W}^x, \widehat{W}^z, W^{\text{out}}, \widehat{b}^{\text{hidden}}, b^{\text{out}}\}$, where

$$\widehat{W}^x = W^x + W^z S, \quad \widehat{W}^z = R, \quad \widehat{z} = Tz - TSx - U, \quad \widehat{b}^{\text{hidden}} = b + RU,$$

for any choice of diagonal matrix $S \in \mathbb{R}^{D \times D}$, vector $U \in \mathbb{R}^D$, and any factorization $W^z = RT$ where $T \in \mathbb{R}^{D \times D}$.

For any set of parameters $(W, z_1, \ldots z_N)$, we show that we can find alternate parameters $(\widehat{W}, \widehat{z}_1 \ldots \widehat{z}_N)$, that are scored as more likely under the log-posterior $\log p(W, \{z_n\}|\mathcal{D})$. We define D_N to be the difference between the scaled parameters $(\widehat{W}, \widehat{z}_1, \ldots \widehat{z}_N)$ and the ground-truth parameters $(W, z_1, \ldots z_N)$ under the joint posterior $\log p(W, Z|\mathcal{D})$. Since $f(x_n, z_n; W) = f(x_n, \widehat{z}_n; \widehat{W})$ by construction, the alternate parameters are equally as likely as the original parameters under the likelihood: $p(y_n|x_n, z_n, W) = p(y_n|x_n, \widehat{z}_n, \widehat{W})$. As such, D_N is simply the difference of the two sets of parameters under the log-prior:

$$D_N = \log p_W(\widehat{W}) - \log p_W(W) + \sum_{n=1}^N \log p_z(\widehat{z}_n) - \log p_z(z_n).$$

We now show that as $N \to \infty$, the probability that $(\widehat{W}, \widehat{z}_1, \ldots \widehat{z}_N)$ is valued higher than $(W, z_1, \ldots z_N)$ in the posterior approaches 1. That is,

$$\lim_{N \to \infty} \Pr\left[D_N > 0\right] = \lim_{N \to \infty} \mathbb{E}_{p_x p_z}\left[\mathbb{I}(D_N > 0)\right] = 1.$$

Since we are only interested in the sign of D_N, we demonstrate, equivalently, that

$$\lim_{N\to\infty} \mathbb{E}_{p_x p_z}\left[\mathbb{I}\left(\frac{1}{N}\cdot D_N > 0\right)\right] = 1.$$

Expanding $\frac{1}{N}\cdot D_N$, we obtain

$$\frac{1}{N}\cdot D_N = \underbrace{\frac{1}{N}\cdot\left(\log p_W(\widehat{W}) - \log p_W(W)\right)}_{L_N} + \underbrace{\frac{1}{N}\sum_{n=1}^{N} \log p_z(\widehat{z}_n) - \log p_z(z_n)}_{R_N}.$$

We note that as $N\to\infty$, the term L_N approaches 0. Thus, it suffices to analyze the sign of R_N; that is, we compute $\Pr[R_N > 0]$.

Let $\text{diag}(t)$ represent a diagonal matrix with diagonal elements $t_1,\ldots,t_D$. For simplicity, we set $U=0$, $S=I$, $R=W^z T^{-1}$ and $T=\text{diag}(t)$ (where $t_d\neq 0$), giving us:

$$\widehat{W}^x = W^x + W^z, \quad \widehat{W}^z = W^z T^{-1}, \quad \widehat{z} = T\cdot z - T\cdot x, \quad \widehat{b}^{\text{hidden}} = b.$$

Following the proof for Theorem 1, it suffices to show that the mean of R_N is positive and that the its variance approaches 0 at a rate of $1/N$ in order to prove that $\Pr[R_N < 0]$ approaches 0 as $N\to\infty$.

For this model, R_N is defined as:

$$\begin{aligned}
R_N &= \frac{1}{N}\sum_{n=1}^{N} \log p_z(\widehat{z}_n) - \log p_z(z_n) \\
2\sigma_z^2\cdot R_N &= \frac{1}{N}\sum_{n=1}^{N} z_n^\intercal z_n - (Tz_n - Tx_n)^\intercal (Tz_n - Tx_n) \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} -t_d^2\cdot x_{n,d}^2 + 2t_d^2\cdot x_{n,d}\cdot z_{n,d} - t_d^2 z_{n,d}^2 + z_{n,d}^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} -t_d^2\cdot x_{n,d}^2 + 2t_d^2\cdot x_{n,d}\cdot z_{n,d} + (1-t_d^2)\cdot z_{n,d}^2
\end{aligned}$$

The mean of R_N can be computed as follows:

$$\begin{aligned}
\mathbb{E}[R_N] &= \sum_{d=1}^{D} -t_d^2\cdot \mathbb{E}_{p_{x_d}}\left[x^2\right] + 2t_d^2\cdot \mathbb{E}_{p_{x_d}}[x]\cdot \mathbb{E}_{p_{z_d}}[z] + (1-t_d^2)\cdot \mathbb{E}_{p_{z_d}}\left[z^2\right] \\
&= \sum_{d=1}^{D} -t_d^2\cdot \mathbb{E}_{p_{x_d}}\left[x^2\right] + 2t_d^2\cdot \mathbb{E}_{p_{x_d}}[x]\cdot 0 + (1-t_d^2)\cdot \mathbb{E}_{p_{z_d}}\left[z^2\right] \\
&= \sum_{d=1}^{D} -t_d^2\cdot (\mu_{x_d}^2 + \sigma_{x_d}^2) + (1-t_d^2)\cdot \sigma_{z_d}^2
\end{aligned}$$

where p_{x_d}, p_{z_d} are the dth marginals of p_x, p_z, and $\mu_{x_d}, \sigma^2_{x_d}$ are the mean and variance of the dth marginal of x. As such, for any t_d that satisfies the following conditions,

$$0 < |t_d| < \sqrt{\frac{\sigma_z^2}{\mu_{x_d}^2 + \sigma_{x_d}^2 + \sigma_z^2}},$$

$\mathbb{E}[R_N]$ is positive. Just as in the proof for the 1-Node BNN+LV, the variance of R_N can be computed as follows:

$$\mathbb{V}[R_N] = \frac{\mathbb{V}[R_1]}{N}.$$

Therefore, as N increases, the variance approaches 0 at a rate of $1/N$. Now, by Cantelli's Inequality, we have that:

$$\Pr[R_N < 0] \leq \frac{\mathbb{V}[R_N]}{\mathbb{V}[R_N] + \mathbb{E}[R_N]^2} = \frac{\mathbb{V}[R_1]}{\mathbb{V}[R_1] + N \cdot \mathbb{E}[R_N]^2},$$

which shows that $\Pr\left[R_N^{(C)} < 0\right]$ approaches 0 as $N \to \infty$.

■

B.3 Additional Types of Non-Identifiability of 1-Layer BNN+LV Models

Assume that the activation function g is invertible. Let $\{(x_n, y_n)\}_{n=1}^N$ be a set of observed data generated by the model parameters W. Define $\widehat{b}^{\text{hidden}}, b^{\text{out}}$ to be zero, $\widehat{W}^x$ to be the $H \times D$ zero matrix and $\widehat{W}^{\text{out}}$ to be the $1 \times H$ matrix of consisting of $\frac{1}{DH}$ in all entries. Finally, let $\widehat{W}^z$ be a $H \times D$ matrix of 1's and let $\widehat{z}_n = g^{-1}(y_n)$.

Then, we have that $y_n = f(x_n, \widehat{z}_n; \widehat{W})$. That is, the alternate set of model parameters $\widehat{W}^x$ reconstructs the observed data perfectly. We note that in this case, the latent noise variable z is a function of the observed output y and is hence dependent on the input x.

Appendix C. Intractability of Marginalizing out Z

Instead of approximating the joint posterior $p(W, Z|\mathcal{D})$, one might argue that it is better to directly approximate the posterior marginal $p(W|\mathcal{D})$, since only the posterior marginal is used in the posterior predictive (Equation 3). This allows us to bypass the aforementioned local optima caused by approximating $p(W, Z|\mathcal{D})$. However, marginalizing out Z may be intractable:

$$\text{argmin}_\phi D_{\text{KL}}[q_\phi(W|\mathcal{D})||p(W|\mathcal{D})] = \text{argmin}_\phi \mathbb{E}_{q_\phi(W|\mathcal{D})}[\log \underbrace{\mathbb{E}_{p(Z)}[p(Y|X,W,Z)]}_{\text{intractable}}] - D_{\text{KL}}[q_\phi(W|\mathcal{D})||p(W)] \tag{17}$$

In order to tractably approximate the above expectation over Z, it is natural to apply a variational lower bound, as commonly does in the Variational Autoencoder literature (Kingma and Welling, 2013):

$$\log \mathbb{E}_{p(Z)}[p(Y|X,W,Z)] \geq \mathbb{E}_{q_\phi(Z|\mathcal{D})}[\log p(Y|X,W,Z)] - D_{\text{KL}}[q_\phi(Z|\mathcal{D})||p(Z)]. \tag{18}$$

By plugging Equation 18 into Equation 17, we get the original mean-field ELBO in Equation 5. This derivation of the ELBO illuminates an additional issue: since the marginal likelihood, $\mathbb{E}_{p(Z)}[p(Y|X,W,Z)]$, is lower-bounded, it is no longer highest at the ground-truth weights (or at weights that parameterize equivalent functions). As such, the $q_\phi(W|\mathcal{D})$ that maximizes the ELBO may not concentrate around functions that explain the data well. In fact, the harder $p(Z|\mathcal{D})$ is to approximate using $q_\phi(Z|\mathcal{D})$, the larger the gap is between the bound and the true marginal likelihood, and the more $q_\phi(W|\mathcal{D})$ will concentrate around functions that explain the data poorly. We note that this issue occurs at the *global* optima of the ELBO. To alleviate this issue, in which the global optima of the ELBO prefer functions that explain the data poorly, one can use a more flexible variational family, leading to a tighter variational bound. The challenge here, though, is that it is unknown on what types of data-sets it is important to model the posterior dependencies between the weights W and latent inputs Z, which result in computationally expensive estimates of the ELBO.

Appendix D. Choosing Differentiable Forms of the NCAI Objective

Tractable training with NCAI depends on instantiating a differentiable form of the training Equation 14. In the following, we consider computationally efficient proxies for the two constraints in the NCAI objective.

D.1 Defining $I_\phi(x;z)$

Limitations of directly using $I_\phi(x;z)$. If we use KL-divergence, we get the following estimator:

$$I_\phi(x;z) = D_{\mathrm{KL}}[q_\phi(z|x)p(x)\|q_\phi(z)p(x)] \tag{19}$$

$$= \mathbb{E}_{q_\phi(z|x)p(x)}\left[\log \frac{q_\phi(z|x)p(x)}{q_\phi(z)p(x)}\right] \tag{20}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_\phi(z_n|x_n)}\left[\log \frac{q_\phi(z_n|x_n)}{\frac{1}{N}\sum_{n=1}^{N} q_\phi(z_n|x_n,y_n)}\right] \tag{21}$$

where

$$q_\phi(z_n|x_n) = \mathbb{E}_{p(y_n|x_n)}\left[q_\phi(z_n|x_n,y_n)\right] \tag{22}$$

The above estimator is intractable to compute for two reasons:

1. Estimating $q_\phi(z_n|x_n)$ requires samples from $p(y_n|x_n)$ (the posterior predictive). To sample from $p(y_n|x_n)$ one already needs to have a good estimate of the posterior; however, given a good estimate of $p(y_n|x_n)$, there would be no need to estimate $I_\phi(x;z)$ in the first place. Alternatively, to sample from $p(y_n|x_n)$, one needs multiple observations of y for the same x, which we do not have. As such we are forced to estimate $q_\phi(z_n|x_n)$ with the single observation we have (x_n,y_n), giving us $q_\phi(z_n|x_n,y_n)$ as the approximation. As a result, it is not possible to estimate $I_\phi(x;z)$ without bias.

2. The nested expectations above require too many samples for a low-variance estimator of the gradient of $I_\phi(x; z)$.

Defining a computationally efficient, differentiable proxy. Instead of using a traditional divergence to estimate $I_\phi(x; z)$, we choose to minimize a proxy. Minimizing a proxy, however, also comes with challenges. There generally do not exist differentiable analytic forms for common measures of dependence such as mutual information, thus necessitating the use of upper/lower-bounds (Dieng et al., 2019). Moreover, in using lower bounds (e.g. MINE (Belghazi et al., 2018)), we found that even if the bound is tight for fixed variational parameters, these estimators need to be re-trained even when the variational means are adjusted by a gradient step. We found this is prohibitively expensive to compute in practice.

As such, as a proxy for $I_\phi(x; z)$, we penalize the correlation between x and the variational means of z as well as the correlation between y and the variational means of z:

$$\lambda_3 \,\mathrm{PC}(\{x_n\}, \{\mu_{z_n}\}) + \lambda_4 \,\mathrm{PC}(\{y_n\}, \{\mu_{z_n}\}) \tag{23}$$

where $\mathrm{PC}(\cdot, \{\mu_{z_n}\})$ is a measure of the average Pearson correlation between pairs of dimensions in $\{x_n\}$ or $\{y_n\}$ and the latent noise means $\{\mu_{z_n}\}$:

$$\mathrm{PC}(\{a_n\}, \{b_n\}) = \frac{1}{D} \sum_{d=1}^{D} \frac{\mathrm{Cov}(\{a_n^{(d)}\}, \{b_n^{(d)}\})}{\sqrt{\mathbb{V}[\{a_n^{(d)}\}]\mathbb{V}[\{b_n^{(d)}\}]}}. \tag{24}$$

Minimizing the correlation between x and z reduces the linear dependence between the input and the inferred noise, following the form of non-identifiability identified in Equation 8; minimizing the correlation between z and y reduces the non-linear dependence between the input and the inferred noise, following the form of non-identifiability identified in Appendix B.3. The fact that both our constraint and the exact form of non-identifiability we characterize in Equation 8 are both linear explains why this simple constraint is empirically effective.

D.2 Defining $D[q_\phi(z)||p(z)]$

Limitations of directly using $D[q_\phi(z)||p(z)]$. Minimizing $D[q_\phi(z)||p(z)]$ with respect to the variational parameters is challenging for several reasons: (1) common choices for divergences may be biased (He et al., 2019), (2) traditional divergences such as forward/reverse-KL, Jensen-Shannon and MMD (Gretton et al., 2012) are expensive to compute, and lastly (3) as we show in this section, these divergences can be artificially reduced by trivially inflating the variational variances.

Defining a computationally efficient, differentiable proxy. For these reasons, as a proxy for $D[q_\phi(z)||p(z)]$, we choose to penalize the Henze-Zirkler (HZ) differentiable non-parametric test-statistic for normality (Henze and Zirkler, 1990) applied to the set of latent

noise means, $\{\mu_{z_n}\}_{n=1}^N$:

$$\begin{aligned}\mathrm{HZ}(\{a_n\}_{n=1}^N) = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \exp\left(-\frac{\beta^2}{2}(a_i-a_j)^\intercal\Sigma^{-1}(a_i-a_j)\right) \\ - 2(1+\beta^2)^{\frac{-k}{2}}\sum_{i=1}^N \exp\left(\frac{-\beta^2}{2(1+\beta^2)}(a_i-\mu)^\intercal\Sigma^{-1}(a_i-\mu)\right) \\ + N(1+2\beta^2)^{\frac{-k}{2}}\end{aligned} \tag{25}$$

where μ, Σ are the sample mean and covariance of $\{a_n\}_{n=1}^N$, and $\beta = \frac{1}{\sqrt{2}}\left(\frac{N(2k+1)}{4}\right)^{-k-4}$. This encourages the aggregated posterior $q_\phi(z)$ to be Gaussian. We additionally penalize the ℓ_2 penalty of the off-diagonal terms of the empirical covariance of the variational mean of the latent noise:

$$\lambda_1 \mathrm{HZ}(\{\mu_{z_n}\}_{n=1}^N) + \lambda_2 \left\|\text{offdiag}\ \widehat{\Sigma}\left(\{\mu_{z_n}\}_{n=1}^N\right)\right\|_2 \tag{26}$$

Lastly, we note that traditional divergences can be trivially minimized by setting each $q_\phi(z_n|x_n, y_n) = p(z)$ (a phenomenon known as posterior collapse (He et al., 2019)). Our HZ based constraint cannot be fooled in this way.

To demonstrate that our proposed instantiation cannot be trivially minimized by inflating the variational variances, we performed the following experiment: (1) we initialized the variational means uniformly inside a triangle, and the variational variances randomly, (2) we minimized $D[q_\phi(z)||p(z)]$ using one of the aforementioned divergences. Figure 5 shows $q_\phi(z)$ and the distribution of the posterior means before and after the optimization. As the figure shows, only our Henze-Zirkler penalty optimizes the means to be Gaussian; in contrast, the remaining divergences were trivially minimized by keeping the means near their initialized positions and inflating the variational variances. This suggests that a poor initialization of the posterior means μ_{z_n} cannot be corrected without our Henze-Zirkler penalty.

D.3 Defining the NCAI Objective

Finally, we incorporate the ELBO and the differentiable forms of the constraints (as exponentially smoothed penalties) into Equation 14:

$$\begin{aligned}\mathcal{L}_{\mathrm{NCAI}}(\phi) = &-\mathrm{ELBO}(\phi) \\ &+ \lambda_1 N \exp\left(\frac{\mathrm{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\}}{\epsilon_T}\right) + \lambda_2 N\|\text{offdiag}\ \widehat{\Sigma}(\{\mu_{z_1},\ldots,\mu_{z_N}\})\|_2 \\ &+ \lambda_3 N \exp\left(\frac{\mathrm{PC}(\{x_n\},\{\mu_{z_n}\})}{\epsilon_x}\right)\cdot\exp\left(\frac{\mathrm{PC}(\{y_n\},\{\mu_{z_n}\})}{\epsilon_y}\right)\end{aligned} \tag{27}$$

where $\epsilon_T, \epsilon_x, \epsilon_y$ control the growth rate of the exponential penalties. We minimize the negative ELBO following Bayes by Backprop (BBB) (Blundell et al., 2015): back-propagating through $\mathbb{E}_{q_\phi(Z,W|\mathcal{D})}[\cdot]$ in the ELBO using the reparameterization trick (Kingma and Welling, 2013), computing the KL-divergence terms and constraints using closed-form expressions.

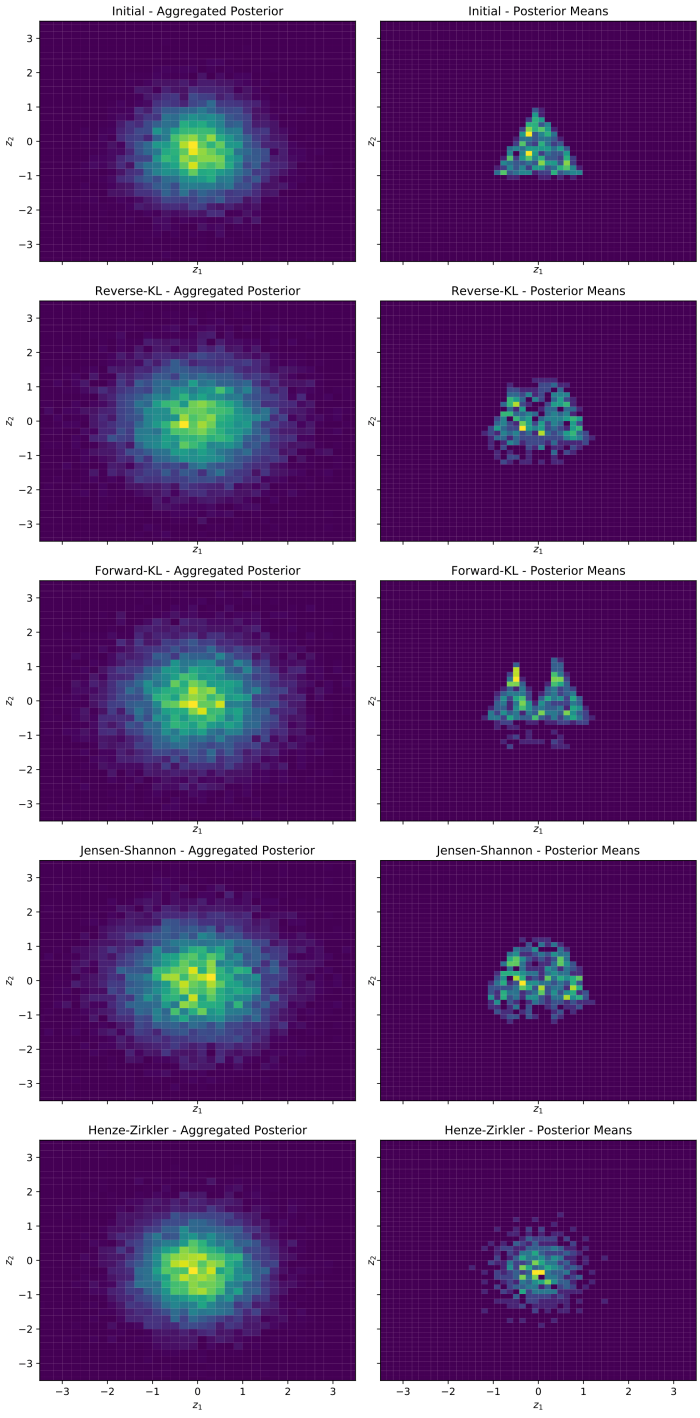
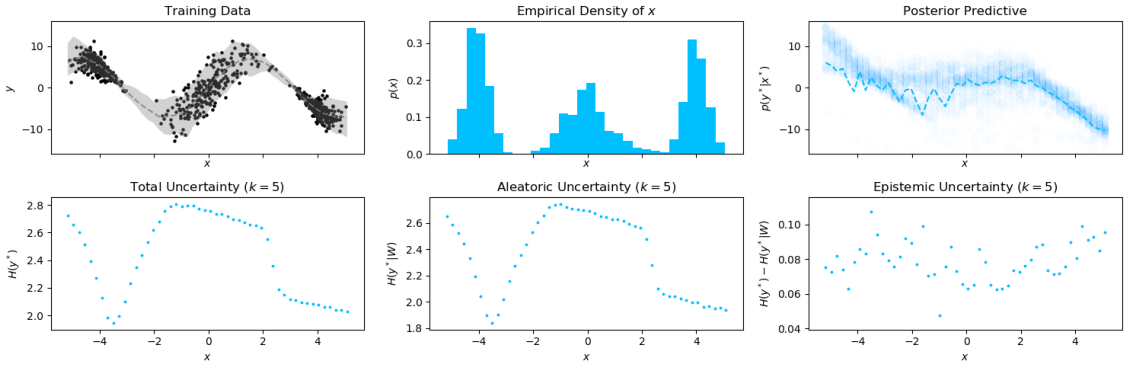


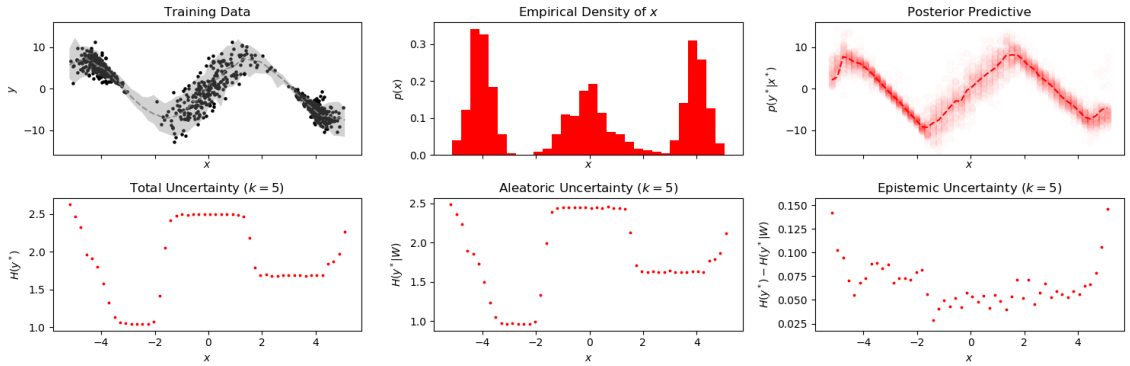
Figure 5: Left column: aggregated posterior $q_\phi(z)$. Right column: posterior means $\{\mu_{z_1}\}_{n=1}^N$. The top row shows the initial variational parameters—the means of the posterior are uniformly distributed in a triangle, while the aggregated posterior looks Gaussian. Subsequent rows show the result of minimizing $D[q_\phi(z)||p(z)]$ for different divergences. Only our Henze-Zirkler optimizes the means to be Gaussian; in contrast, the remaining divergences can be trivially minimized by keeping the means near their initialized positions and inflating the variational variances.

Appendix E. Uncertainty Decomposition

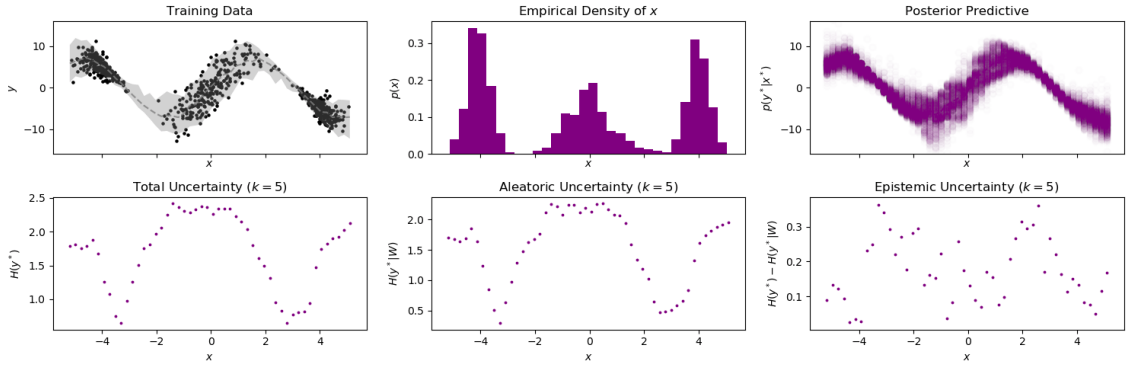
In addition to the quantitative results in the paper, showing that the uncertainty decomposition learned by NCAI is quantitatively closer to that produced by HMC than the decomposition learned by BNN+LV, we also show qualitatively that our uncertainty decomposition is closer to that of HMC in Figure 6. The figure shows that like HMC, NCAI has appropriately high total and aleatoric uncertainties at x's for which there is a high variance in y, as well as high epistemic uncertainty for x's near the boundary of the data (near $x = -5$ and $x = 5$). In contrast, BNN+LV trained via mean-field VI does not. In comparison to HMC, however, both NCAI and BNN+LV tend to underestimate the epistemic uncertainty, which should be high where $p(x)$ is low, signifying uncertainty over the model parameters due to lack of data. This is because mean-field variational inference is generally unable to capture "in-between uncertainty" in BNNs (Foong et al., 2020).



(a) Uncertainty decomposition given by BNN+LV with mean-field VI.



(b) Uncertainty decomposition given by BNN+LV with NCAI.



(c) Uncertainty decomposition given by HMC.

Figure 6: **Comparison of uncertainty decompositions.** We expect the epistemic noise to be roughly inversely proportional to the empirical density of x, e.g. epistemic uncertainty is higher where x is sparsely sampled. We expect the aleatoric uncertainty to match the observed level of noise in the data. We see that BNN+LV trained with mean-field VI is unable to detect the highly noisy regions in data while BNN+LV with NCAI training learns aleatoric uncertainty that matches these regions very well. The uncertainties are estimated using a nearest neighbor based entropy estimator with $k = 5$ nearest neighbors. HMC was trained with $\sigma_z^2 = 1.0, \sigma_\epsilon^2$, as in the ground-truth, and with $\sigma_w^2 = 10.0$.

Data-set	Size	Dimensionality	Hidden Nodes
Lidar	221	1	10
Yacht	309	6	5
Energy Efficiency	768	8	10
Airfoil	Subsampled to 1000	5	30
Abalone	Subsampled to 1000	10 (1-hot for categorical)	10
Wine Quality Red	1600	1	20

Table 7: Experimental details for the real data-sets

Appendix F. Experimental Setup

F.1 Experimental Details

Architecture. The network architectures we use for all experiments are summarized in Tables 7 and 8. We note that we have purposefully selected lower capacity architectures to encourage the non-identifiability described in the paper to occur in practice. We also note that the non-identifiability occurs even when the ground-truth network capacity is known, as in the case of the HeavyTail and Depeweg data-sets, which have been generated using a neural network. As such, even when the data was generated by the same generative process as the one assumed by the model, the problem of non-identifiability still occurs.

Train/Validation/Test Data-splits. We split each data-set into train/validation/test set 6 times. We use the first data-split to select hyper-parameters by selecting the hyper-parameters that yield the best average log-likelihood performance on the validation set across 10 random restarts. After having selected the hyper-parameter for each method, we select between $\text{NCAI}_{\lambda=0}$ and NCAI_{λ} by picking the approach that yielded the best log-likelihood performance on the validation set across the 10 random restarts. Now, using the selected hyper-parameters and form of NCAI, we train our models on the remaining 5 data-splits, averaging the best-of-10 random restarts (selected by validation log-likelihood) across the data-splits.

For Abalone, Airfoil, Boston Housing, Energy Efficiency, Lidar, Wine Quality Red, Yacht, Goldberg, Williams, Yuan, we split the data into a 70% training set, 20% validation set and 10% test set.

Prior Hyper-parameters Selection. Since hyper-parameter search is expensive to compute, we choose hyper-parameters of the priors $p(z), p(W)$ using empirical Bayes (MAP Type II). That is, we place Inverse Gamma priors ($\alpha = 3.0, \beta = 0.5$) on the variances of the network weights and the latent variables; then we approximate the negative ELBO with the MAP estimates of the variances, making the assumption that these terms dominate the respective integrals in which they appear:

$$\begin{aligned} -\text{ELBO}(\phi) \approx &- \mathbb{E}_{q_\phi(Z,W|\mathcal{D})}[\log p(Y|X,W,Z)] \\ &+ D_{\text{KL}}[q_\phi(W|\mathcal{D}) \| p(W|s_w^*)p(s_w^*)] \\ &+ D_{\text{KL}}[q_\phi(Z|\mathcal{D}) \| p(Z|s_z^*)p(s_z^*)] \end{aligned} \tag{28}$$

where we define:

$$s_z^* = \underset{s_z}{\operatorname{argmin}}\, D_{\mathrm{KL}}[q_\phi(Z|\mathcal{D})\|p(Z|s_z)p(s_z)] = \frac{2\beta + \frac{1}{N}\sum_n \left[\operatorname{tr}\left(\Sigma_{q_n}\right) + \mu_{q_n}^T \mu_{q_n}\right]}{K + 2\alpha - 2}, \tag{29}$$

$$s_w^* = \underset{s_w}{\operatorname{argmin}}\, D_{\mathrm{KL}}[q_\phi(W|\mathcal{D})\|p(W|s_w)p(s_w)] = \frac{2\beta + \operatorname{tr}\left(\Sigma_q\right) + \mu_q^T \mu_q}{H + 2\alpha - 2}, \tag{30}$$

with K and H as the dimensionality of z_n and W, respectively. The optimal variances, s_z^*, s_w^*, have analytic solutions:

$$s_z^* = \frac{2\beta_z + \frac{1}{N}\sum_n \left[\operatorname{tr}\left(\Sigma_{q_n}\right) + \mu_{q_n}^\intercal \mu_{q_n}\right]}{K + 2\alpha_z - 2}, \tag{31}$$

$$s_w^* = \frac{2\beta_W + \operatorname{tr}\left(\Sigma_q\right) + \mu_q^\intercal \mu_q}{K + 2\alpha_W - 2}, \tag{32}$$

where the β's and α's are the parameters of the respective Inverse Gamma distributions. In training, we update the objective as well as the optimal hyper-parameters via coordinate descent. That is, we iteratively compute s_z^*, s_w^* in closed-form given the current ϕ, and then optimize ϕ while holding s_z^*, s_w^* fixed.

Remaining Hyperparameter Selection. For the data-sets Abalone, Airfoil, Boston Housing, Energy Efficiency, Lidar, Wine Quality Red, Yacht, Goldberg, Williams, Yuan, we used grid-searched over the following parameters:

- BNN: $\sigma_\epsilon^2 = \{1.0, 0.1, 0.01\}$
- BNN+LV: $\sigma_\epsilon^2 = \{0.1, 0.01\}$
- NCAI: $\sigma_\epsilon^2 = \{0.1, 0.01\}$, $\lambda_2 = \{10.0\}$, $\epsilon_T = \{0.01, 0.0003\}$, $\epsilon_y = \{0.1, 0.5\}$, $\epsilon_x = \{0.5, 1.0\}$

For HeavyTails we grid-searched over the following parameters:

- BNN: $\sigma_\epsilon^2 = \{1.0, 0.1, 0.5\}$
- BNN+LV: $\sigma_\epsilon^2 = \{0.1\}, \sigma_z^2 = \{0.01\}$
- NCAI: $\sigma_\epsilon^2 = \{0.1\}$, $\lambda_2 = \{10.0\}$, $\epsilon_T = \{0.01, 0.0003\}$, $\epsilon_y = \{0.1, 0.5\}$, $\epsilon_x = \{0.5, 1.0\}$

For Depeweg we grid-searched over the following parameters:

- BNN: $\sigma_\epsilon^2 = \{1.0, 0.1, 0.5\}$
- BNN+LV: $\sigma_\epsilon^2 = \{0.1\}, \sigma_z^2 = \{1.0\}$
- NCAI:
  $\sigma_\epsilon^2 = \{0.1\}$, $\lambda_2 = \{10.0\}$, $\epsilon_T = \{0.01, 0.0003\}$, $\epsilon_y = \{0.1, 0.5\}$, $\epsilon_x = \{0.5, 1.0\}$

Data-set	Hidden Nodes	Layers
Williams	20	2
Yuan	20	1
Goldberg	20	1
HeavyTail	50	1
Depeweg	50	1
Bimodal	50	2

Table 8: Experimental details for the synthetic data-sets

Random Initialization. For all models without a specialized random initialization, we initialize all variational parameters initialized to samples drawn from xavier-normal with a gain of 1.0. For runs in which the prior variances are estimated via empirical Bayes MAP (described above), we initialize the prior variances to samples from the hyper-prior.

F.2 Evaluation Metrics

Quality of Fit. We measure the *training reconstruction MSE*, the ability of the model to reconstruct the training targets with the inferred weights and latent variables:

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\phi(z_n|x_n,y_n)q_\phi(W|\mathcal{D})}\left[\|y_n - f(z_n,W)\|_2^2\right]. \tag{33}$$

At test time, we measure the quality of the posterior predictive distribution of the model by computing the *average log-likelihood*:

$$\frac{1}{N}\sum_{n=1}^{N}\log\mathbb{E}_{p(z_n)q_\phi(W|\mathcal{D})}\left[p(y_n|x_n,W,z_n)\right]. \tag{34}$$

We also compute the predictive quality of the model by computing the *predictive MSE*:

$$\frac{1}{N}\sum_{n=1}^{N}\left[\|y_n - \mathbb{E}_{p(z_n)q_\phi(W|\mathcal{D})}[f(z_n,W)]\|_2^2\right]. \tag{35}$$

Note that the difference between the reconstruction MSE (Equation 33) and the predictive MSE (Equation 35) is that in the latter we sample the latent variables from the prior distributions rather than the learned posterior distributions.

Posterior Predictive Calibration. We measure the quality of the model's predictive uncertainty by computing the percentage of observations for which the ground-truth y lies within a 95% predictive-interval (PI) of the learned posterior predictive—this quantity is called the Prediction Interval Coverage Probability (PICP). We measure the tightness of the model's predictive uncertainty by computing the 95% Mean Prediction Interval Width (MPIW).

Satisfaction of Model Assumptions. We estimate the mutual information between x and z by computing the Kraskov nearest-neighbor based estimator (Kraskov et al., 2004) (with 5 nearest neighbors) on the x's and the means of the z's: $\hat{I}(x;\mu_z)$. We use the

variational means μ_{z_n} instead of samples from the aggregated posterior $z \sim q_\phi(z)$, since otherwise dependence can be artificially reduced with large variational variances $\sigma^2_{z_n}$.

For the univariate case, when $D = K = 1$, we use the Kolmogorov-Smirnov (KS) two-sample test statistic (kst, 2008) to evaluate divergence between $q_\phi(z)$ and $p(z)$. When computing the test statistic, we represent $q_\phi(z)$ using μ_{z_n}'s and $p(z)$ using its samples. This is because the $\sigma^2_{z_n}$'s are large, the distance between $q_\phi(z)$ and $p(z)$ more difficult to detect. A lower KS test-statistic indicates that $q_\phi(z)$ and $p(z)$ are more similar. We compute the Jensen-Shannon divergence between $q_\phi(z)$ and $p(z)$ in multivariate cases.

Appendix G. Data-sets

Synthetic Data Generated with Ground-Truth. We generate three synthetic data-sets with corresponding ground-truth in order to guarantee that our generative process matches our data. We generate these data-sets by training a neural network to map the x_n's and ground-truth z_n's to the noise-less y_n's (i.e. before adding ϵ), specified by some ground-truth function f_{GT}. We then re-generate the y_n's from the learned neural network and treat that network as the ground-truth function. We specify the parameters of the generative process in Equation 1 used to generate each of these data-sets below:

1. **Heavy-Tail**:

$$\begin{aligned} f_{\text{GT}}(x, z) &= 6 \cdot \tanh\left(0.1x^3(z+1)^6 - 10 \cdot xz^2 + z\right) \\ \sigma^2_\epsilon &= 0.1 \\ \sigma^2_z &= 0.01 \end{aligned}$$

   For this data-set, we sampled 300 training, validation and test x's uniformly on $[-4, 4]$.

2. **Depeweg** (Depeweg et al., 2018):

$$\begin{aligned} f_{\text{GT}}(x, z) &= 7\sin(x) + 3|\cos(x/2)|z \\ \sigma^2_\epsilon &= 0.1 \\ \sigma^2_z &= 1.0 \end{aligned}$$

   For this data-set, we sampled 750 training x's, 250 validation and 250 test x's from a uniform mixture of the following gaussians: $\mathcal{N}(0, -4.0, 0.16), \mathcal{N}(0, 0, 0.81), \mathcal{N}(0, 4.0, 0.16)$. (Note: $\sigma^2_\epsilon = 1.0$ in the original data-set from Depeweg et al. (2018)).

3. **Bimodal** (Depeweg et al., 2018):

$$\begin{aligned} f_{\text{GT}}(x, z) &= \mathbb{I}(z > 0) \cdot 10\sin(x) + \mathbb{I}(z \leq 0) \cdot 10\cos(x) \\ \sigma^2_\epsilon &= 1.0 \\ \sigma^2_z &= 0.1 \end{aligned}$$

   For this data-set, we sampled 750 training x's, 250 validation and 250 test x's from an exponential distribution with $\lambda = 2$, shifted to the left by 0.5 and clipped to the range of $[-0.5, 2.0]$.

Synthetic Data without Ground-Truth. We also consider 3 synthetic data-sets, most of which have been widely used to evaluate heteroscedastic regression models (Goldberg et al., 1998; Wright, 1999; Kersting et al., 2007; Wang and Neal, 2012):

1. **Goldberg** (Goldberg et al., 1998): targets are given by $y = 2\sin(2\pi x) + \epsilon(x)$, where $\epsilon(x) \sim \mathcal{N}(0, x + 0.5)$. Evaluated on 200 training input, 200 validation and 200 test inputs uniformly sampled from $[0, 1]$.

2. **Yuan** (Yuan and Wahba, 2004): targets are given by $y = 2[\exp\{-30(x - 0.25)^2 + \sin(\pi x^2)\}] - 2 + \epsilon(x)$, where $\epsilon(x) \sim \mathcal{N}(0, \exp\{\sin(2\pi x)\})$. Evaluated on 200 training input, 200 validation and 200 test inputs uniformly sampled from $[0, 1]$.

3. **Williams** (Williams, 1996): the targets are given by $y = \sin(2.5x) \cdot \sin(1.5x) + \epsilon(x)$, where $\epsilon(x) \sim \mathcal{N}(0, 0.01 + 0.25(1 - \sin(2.5x))^2)$. Evaluated on 200 training input, 200 validation and 200 test inputs uniformly sampled from $[0, 1]$.

Real Data. We use 6 UCI data-sets (Dua and Graff, 2017) and a data-set commonly used in the heteroscedastic literature, Lidar (Sigrist et al., 1994)—see Table 7 for details.

Appendix H. Additional Quantitative Results and Metrics

Table 9 shows the RMSE for all models on synthetic data—NCAI consistently achieves lower RMSE than mean-field VI. Tables 10 and 11 show the mutual information between x and z under the posterior on real data-sets and the Henze-Zirkler test-statistic—on most data-sets NCAI better satisfies modeling assumptions than mean-field VI. Lastly, Tables 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22) provide the complete results for all experiments, including evaluation metrics from Appendix F.2 omitted from the main text for brevity.

RMSE on Test Data for Synthetic Data-sets

Model	Inference	Heavy Tail	Goldberg	Williams	Yuan	Depeweg
BNN	**MFVI**	1.831 ± 0.074	0.335 ± 0.025	1.017 ± 0.06	0.607 ± 0.035	1.953 ± 0.071
BNN+LV	**MFVI**	1.882 ± 0.088	0.376 ± 0.032	1.118 ± 0.096	0.622 ± 0.039	3.523 ± 0.501
BNN+LV	**NCAI**$_{\lambda=0}$	1.787 ± 0.094	0.339 ± 0.026	0.978 ± 0.083	0.619 ± 0.039	1.932 ± 0.059
BNN+LV	**NCAI**$_{\lambda}$	1.79 ± 0.09	0.337 ± 0.025	0.978 ± 0.083	0.619 ± 0.039	1.932 ± 0.059

Table 9: Comparison of RMSE on synthetic data-sets ($\pm$ std). Across all data-sets BNN+LV trained with NCAI$_{\lambda}$ training yields comparable if not better generalization than BNN+LV and BNN trained with mean-field variational inference (MFVI).

Mutual Information between x and z (Real Data)

Inference	Abalone	Airfoil	Energy	Wine	Lidar	Yacht
MFVI	0.152 ± 0.015	0.485 ± 0.054	0.139 ± 0.086	0.045 ± 0.012	0.667 ± 0.061	0.077 ± 0.012
$\mathbf{NCAI}_{\lambda=0}$	0.149 ± 0.078	0.29 ± 0.021	0.162 ± 0.063	0.047 ± 0.011	0.373 ± 0.037	0.087 ± 0.012
$\mathbf{NCAI}_{\lambda}$	0.149 ± 0.078	0.29 ± 0.021	0.226 ± 0.041	0.029 ± 0.008	0.842 ± 0.06	0.087 ± 0.012

Table 10: Comparison of *mutual information* between z and x on real data-sets ($\pm$ std). Across nearly all data-sets, NCAI training infers z's that have the least mutual information in comparison mean-field VI (MFVI).

Henze-Zirkler Test-Statistic (Real Data)

Inference	Abalone	Airfoil	Energy	Wine	Lidar	Yacht
MFVI	26.148 ± 4.394	28.108 ± 3.205	16.976 ± 3.519	52.566 ± 2.633	5.09 ± 0.991	27.059 ± 4.144
$\mathbf{NCAI}_{\lambda=0}$	17.975 ± 6.725	49.122 ± 11.426	19.071 ± 5.414	53.201 ± 1.247	7.804 ± 1.727	51.283 ± 9.548
$\mathbf{NCAI}_{\lambda}$	17.975 ± 6.725	49.122 ± 11.426	1.186 ± 0.558	1.641 ± 0.242	0.005 ± 0.001	51.283 ± 9.548

Table 11: Comparison of model assumption satisfaction (in terms of the HZ metric) on real data-sets ($\pm$ std). Across nearly all data-sets, NCAI training infers z's that are more Gaussian (lowest HZ) relative to mean-field VI (MFVI).

	NCAI_{λ}	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.007 ± 0.004	0.021 ± 0.019	N/A	0.009 ± 0.003
$D_{\text{KL}}(p(z)\|\|q(z))$	0.015 ± 0.012	0.05 ± 0.046	N/A	0.015 ± 0.006
$\hat{I}(x;\mu_z)$	0.146 ± 0.131	0.149 ± 0.078	N/A	0.152 ± 0.015
$\hat{I}(x;z)$	0.015 ± 0.006	0.012 ± 0.005	N/A	0.011 ± 0.002
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.839 ± 0.15	17.975 ± 6.725	N/A	26.148 ± 4.394
s_w^*	0.2 ± 0.018	1.272 ± 1.043	0.948 ± 0.68	0.202 ± 0.02
s_y^*	0.01 ± 0.0	0.01 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s_z^*	0.252 ± 0.004	0.248 ± 0.001	N/A	0.249 ± 0.0
95%-MPIW Test (Unnorm)	7.564 ± 0.459	7.027 ± 0.357	4.112 ± 0.103	7.127 ± 0.311
95%-MPIW Train (Unnorm)	7.716 ± 0.448	7.096 ± 0.446	4.111 ± 0.099	7.248 ± 0.215
95%-PICP Test	94.3 ± 2.515	92.3 ± 0.975	76.9 ± 4.292	94.4 ± 2.329
95%-PICP Train	94.771 ± 0.559	93.343 ± 1.743	77.771 ± 1.476	94.4 ± 0.509
$\text{PC}(x,\mu_z)$	0.001 ± 0.0	0.006 ± 0.002	N/A	0.005 ± 0.002
$\text{PC}(y,\mu_z)$	0.032 ± 0.003	0.063 ± 0.023	N/A	0.115 ± 0.016
Post-Pred Avg-LL Test	-0.837 ± 0.105	-0.831 ± 0.086	-1.248 ± 0.153	-0.843 ± 0.071
Post-Pred Avg-LL Train	-0.798 ± 0.051	-0.799 ± 0.064	-1.086 ± 0.099	-0.832 ± 0.022
RMSE Test (Unnorm)	0.208 ± 0.012	0.205 ± 0.013	0.194 ± 0.011	0.204 ± 0.012
RMSE Train (Unnorm)	0.208 ± 0.012	0.205 ± 0.013	0.194 ± 0.011	0.204 ± 0.011
Recon MSE	0.011 ± 0.0	0.012 ± 0.0	N/A	0.165 ± 0.002
Hyperparams	$\sigma_\epsilon^2 = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.54$, $\epsilon_y = 1.0$	$\sigma_\epsilon^2 = 0.01$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.1$

Table 12: Experiment Evaluation Summary for Abalone ($\pm$ std).

	NCAI$_\lambda$	NCAI$_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	-0.0 ± 0.003	-0.001 ± 0.003	N/A	0.001 ± 0.004
$D_{\text{KL}}(p(z)\|\|q(z))$	0.002 ± 0.002	-0.0 ± 0.002	N/A	0.005 ± 0.003
$\hat{I}(x;\mu_z)$	0.262 ± 0.03	0.29 ± 0.021	N/A	0.485 ± 0.054
$\hat{I}(x;z)$	0.107 ± 0.005	0.105 ± 0.008	N/A	0.174 ± 0.025
HZ($\{\mu_{z_1}, \ldots, \mu_{z_N}\}$)	0.25 ± 0.11	49.122 ± 11.426	N/A	28.108 ± 3.205
s^*_w	0.548 ± 0.084	0.509 ± 0.076	0.762 ± 0.182	0.105 ± 0.032
s^*_y	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s^*_z	0.25 ± 0.001	0.247 ± 0.0	N/A	0.25 ± 0.0
95%-MPIW Test (Unnorm)	10.426 ± 0.81	10.283 ± 0.423	8.999 ± 0.164	17.023 ± 1.972
95%-MPIW Train (Unnorm)	10.294 ± 0.732	10.15 ± 0.41	8.985 ± 0.148	16.974 ± 1.925
95%-PICP Test	94.7 ± 2.797	95.5 ± 1.969	91.9 ± 1.475	92.9 ± 1.245
95%-PICP Train	97.429 ± 0.598	97.286 ± 0.769	93.2 ± 2.077	94.686 ± 0.584
PC(x, μ_z)	0.001 ± 0.0	0.005 ± 0.002	N/A	0.004 ± 0.001
PC(y, μ_z)	0.002 ± 0.001	0.03 ± 0.01	N/A	0.162 ± 0.069
Post-Pred Avg-LL Test	-0.48 ± 0.076	-0.462 ± 0.056	-0.512 ± 0.083	-0.995 ± 0.143
Post-Pred Avg-LL Train	-0.407 ± 0.043	-0.401 ± 0.026	-0.422 ± 0.043	-0.972 ± 0.137
RMSE Test (Unnorm)	0.05 ± 0.005	0.05 ± 0.003	0.05 ± 0.003	0.094 ± 0.009
RMSE Train (Unnorm)	0.05 ± 0.005	0.05 ± 0.003	0.05 ± 0.003	0.094 ± 0.008
Recon MSE	0.177 ± 0.01	0.174 ± 0.006	N/A	0.139 ± 0.013
Hyperparams	$\sigma^2_\epsilon = 0.1$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.5$, $\epsilon_y = 1.0$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.1$

Table 13: Experiment Evaluation Summary for Airfoil($\pm$ std).

	NCAI$_\lambda$	NCAI$_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.01 ± 0.005	0.011 ± 0.009	N/A	0.032 ± 0.012
$D_{\text{KL}}(p(z)\|\|q(z))$	0.021 ± 0.01	0.025 ± 0.01	N/A	0.066 ± 0.029
$\hat{I}(x;\mu_z)$	0.226 ± 0.041	0.162 ± 0.063	N/A	0.139 ± 0.086
$\hat{I}(x;z)$	0.14 ± 0.006	0.122 ± 0.011	N/A	0.127 ± 0.02
HZ($\{\mu_{z_1}, \ldots, \mu_{z_N}\}$)	1.186 ± 0.558	19.071 ± 5.414	N/A	16.976 ± 3.519
s^*_w	0.496 ± 0.403	2.91 ± 2.843	1.87 ± 1.005	0.921 ± 1.313
s^*_y	0.01 ± 0.0	0.01 ± 0.0	0.01 ± 0.0	0.01 ± 0.0
s^*_z	0.251 ± 0.003	0.245 ± 0.001	N/A	0.247 ± 0.0
95%-MPIW Test (Unnorm)	11.828 ± 2.308	10.534 ± 1.751	5.704 ± 0.16	15.159 ± 5.563
95%-MPIW Train (Unnorm)	11.925 ± 2.154	10.329 ± 1.331	5.71 ± 0.146	13.813 ± 2.45
95%-PICP Test	94.51 ± 2.384	93.464 ± 1.533	81.438 ± 2.758	94.51 ± 2.981
95%-PICP Train	95.139 ± 2.296	94.36 ± 1.556	84.527 ± 4.621	94.731 ± 2.809
PC(x, μ_z)	0.002 ± 0.001	0.005 ± 0.004	N/A	0.008 ± 0.004
PC(y, μ_z)	0.003 ± 0.001	0.014 ± 0.006	N/A	0.022 ± 0.006
Post-Pred Avg-LL Test	0.898 ± 0.452	0.862 ± 0.138	1.281 ± 0.171	0.573 ± 0.288
Post-Pred Avg-LL Train	0.953 ± 0.393	0.941 ± 0.108	1.443 ± 0.16	0.657 ± 0.205
RMSE Test (Unnorm)	0.035 ± 0.007	0.029 ± 0.005	0.016 ± 0.002	0.041 ± 0.011
RMSE Train (Unnorm)	0.035 ± 0.007	0.029 ± 0.005	0.016 ± 0.002	0.041 ± 0.011
Recon MSE	0.028 ± 0.001	0.031 ± 0.003	N/A	0.028 ± 0.002
Hyperparams	$\sigma^2_\epsilon = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.0003$, $\epsilon_x = 0.1$, $\epsilon_y = 1.0$	$\sigma^2_\epsilon = 0.01$	$\sigma^2_\epsilon = 0.01$	$\sigma^2_\epsilon = 0.01$

Table 14: Experiment Evaluation Summary for Energy Efficiency($\pm$ std).

	NCAI_λ	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.014 ± 0.004	0.019 ± 0.006	N/A	0.037 ± 0.022
$D_{\text{KL}}(p(z)\|\|q(z))$	0.032 ± 0.009	0.041 ± 0.015	N/A	0.082 ± 0.058
$\hat{I}(x;\mu_z)$	0.036 ± 0.04	0.051 ± 0.049	N/A	0.243 ± 0.079
$\hat{I}(x;z)$	0.018 ± 0.025	0.023 ± 0.03	N/A	0.214 ± 0.052
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.027 ± 0.011	7.137 ± 5.436	N/A	4.701 ± 5.439
s_w^*	2.643 ± 0.226	2.355 ± 0.28	0.12 ± 0.007	1.246 ± 0.149
s_y^*	0.1 ± 0.0	0.1 ± 0.0	0.5 ± 0.0	0.1 ± 0.0
s_z^*	0.01 ± 0.0	0.01 ± 0.0	N/A	0.01 ± 0.0
95%-MPIW Test (Unnorm)	5.428 ± 0.403	5.418 ± 0.729	2.979 ± 0.016	6.789 ± 0.408
95%-MPIW Train (Unnorm)	5.371 ± 0.38	5.368 ± 0.7	2.98 ± 0.005	6.67 ± 0.342
95%-PICP Test	94.933 ± 0.723	93.333 ± 1.054	74.2 ± 2.834	94.733 ± 0.641
95%-PICP Train	95.4 ± 0.894	93.467 ± 2.116	73.867 ± 3.288	94.867 ± 1.095
KS Test-Stat	0.023 ± 0.004	0.051 ± 0.028	N/A	0.058 ± 0.035
$\text{PC}(x,\mu_z)$	0.001 ± 0.0	0.0 ± 0.0	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.111 ± 0.017	0.068 ± 0.086	N/A	0.126 ± 0.107
Post-Pred Avg-LL Test	-1.426 ± 0.042	-1.481 ± 0.018	-2.47 ± 0.083	-1.867 ± 0.078
Post-Pred Avg-LL Train	-1.399 ± 0.058	-1.429 ± 0.066	-2.6 ± 0.143	-1.894 ± 0.078
RMSE Test (Unnorm)	1.79 ± 0.09	1.787 ± 0.094	1.831 ± 0.074	1.882 ± 0.088
RMSE Train (Unnorm)	1.789 ± 0.09	1.787 ± 0.094	1.831 ± 0.074	1.883 ± 0.087
Recon MSE	0.142 ± 0.003	0.16 ± 0.017	N/A	0.13 ± 0.007
Hyperparams	$\sigma_z^2 = 0.01$, $\sigma_\epsilon^2 = 0.1$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.1$, $\epsilon_y = 1.0$	$\sigma_z^2 = 0.01$,$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.5$	$\sigma_z^2 = 0.01$,$\sigma_\epsilon^2 = 0.1$

Table 15: Experiment Evaluation Summary for Heavy-Tail ($\pm$ std).

	NCAI_λ	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.002 ± 0.002	0.001 ± 0.002	N/A	0.003 ± 0.003
$D_{\text{KL}}(p(z)\|\|q(z))$	0.001 ± 0.001	0.002 ± 0.002	N/A	0.007 ± 0.002
$\hat{I}(x;\mu_z)$	0.046 ± 0.067	0.02 ± 0.024	N/A	0.229 ± 0.113
$\hat{I}(x;z)$	-0.006 ± 0.008	-0.011 ± 0.007	N/A	0.076 ± 0.101
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.026 ± 0.038	0.621 ± 0.234	N/A	0.918 ± 0.41
s_w^*	0.456 ± 0.093	0.463 ± 0.087	0.627 ± 0.039	0.416 ± 0.152
s_y^*	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s_z^*	0.249 ± 0.002	0.247 ± 0.0	N/A	0.248 ± 0.001
95%-MPIW Test (Unnorm)	3.969 ± 0.184	4.091 ± 0.21	2.265 ± 0.102	4.226 ± 0.676
95%-MPIW Train (Unnorm)	3.948 ± 0.171	4.099 ± 0.176	2.264 ± 0.096	4.298 ± 0.674
95%-PICP Test	91.8 ± 2.308	92.7 ± 1.924	73.4 ± 3.681	91.8 ± 2.49
95%-PICP Train	93.9 ± 0.894	94.1 ± 0.822	75.5 ± 2.598	93.5 ± 1.173
KS Test-Stat	0.016 ± 0.004	0.019 ± 0.005	N/A	0.025 ± 0.003
$\text{PC}(x,\mu_z)$	0.001 ± 0.001	0.0 ± 0.0	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.247 ± 0.148	0.403 ± 0.04	N/A	0.193 ± 0.227
Post-Pred Avg-LL Test	-0.963 ± 0.041	-0.962 ± 0.04	-1.055 ± 0.08	-1.026 ± 0.056
Post-Pred Avg-LL Train	-0.885 ± 0.03	-0.884 ± 0.033	-0.95 ± 0.07	-0.981 ± 0.107
RMSE Test (Unnorm)	0.337 ± 0.025	0.339 ± 0.026	0.335 ± 0.025	0.376 ± 0.032
RMSE Train (Unnorm)	0.337 ± 0.026	0.339 ± 0.026	0.335 ± 0.025	0.376 ± 0.031
Recon MSE	0.16 ± 0.008	0.151 ± 0.007	N/A	0.156 ± 0.013
Hyperparams	$\sigma_\epsilon^2 = 0.1$, $\lambda_2 = 10$, $\epsilon_T = 0.0003$, $\epsilon_x = 0.1$, $\epsilon_y = 1.0$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.1$

Table 16: Experiment Evaluation Summary for Goldberg($\pm$ std).

	NCAI_{λ}	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.012 ± 0.005	0.015 ± 0.005	N/A	0.006 ± 0.005
$D_{\text{KL}}(p(z)\|\|q(z))$	0.025 ± 0.011	0.031 ± 0.009	N/A	0.016 ± 0.008
$\hat{I}(x;\mu_z)$	0.614 ± 0.075	0.519 ± 0.091	N/A	0.982 ± 0.121
$\hat{I}(x;z)$	0.155 ± 0.048	0.059 ± 0.02	N/A	0.235 ± 0.035
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.015 ± 0.019	7.248 ± 2.598	N/A	6.445 ± 2.818
s^*_w	2.927 ± 1.612	2.368 ± 1.55	0.75 ± 0.065	0.997 ± 0.943
s^*_y	0.01 ± 0.0	0.01 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s^*_z	0.247 ± 0.002	0.246 ± 0.001	N/A	0.247 ± 0.001
95%-MPIW Test (Unnorm)	1.468 ± 0.207	1.314 ± 0.126	0.89 ± 0.036	1.881 ± 0.171
95%-MPIW Train (Unnorm)	1.479 ± 0.249	1.31 ± 0.162	0.889 ± 0.033	1.908 ± 0.165
95%-PICP Test	95.4 ± 1.517	92.9 ± 1.294	77.2 ± 3.439	92.9 ± 3.008
95%-PICP Train	96.9 ± 0.894	95.0 ± 0.5	78.8 ± 3.978	95.1 ± 0.418
KS Test-Stat	0.03 ± 0.008	0.034 ± 0.008	N/A	0.033 ± 0.011
$\text{PC}(x,\mu_z)$	0.005 ± 0.002	0.001 ± 0.001	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.018 ± 0.007	0.41 ± 0.113	N/A	0.572 ± 0.068
Post-Pred Avg-LL Test	-0.489 ± 0.154	-0.414 ± 0.184	-1.591 ± 0.417	-1.033 ± 0.156
Post-Pred Avg-LL Train	-0.228 ± 0.125	-0.195 ± 0.108	-1.357 ± 0.119	-0.965 ± 0.078
RMSE Test (Unnorm)	0.987 ± 0.103	0.978 ± 0.083	1.017 ± 0.06	1.118 ± 0.096
RMSE Train (Unnorm)	0.988 ± 0.101	0.979 ± 0.083	1.017 ± 0.06	1.117 ± 0.096
Recon MSE	0.017 ± 0.001	0.017 ± 0.001	N/A	0.145 ± 0.004
Hyperparams	$\sigma^2_\epsilon = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.5$, $\epsilon_y = 0.5$	$\sigma^2_\epsilon = 0.01$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.1$

Table 17: Experiment Evaluation Summary for Williams($\pm$ std).

	NCAI_{λ}	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.005 ± 0.003	0.006 ± 0.003	N/A	0.008 ± 0.003
$D_{\text{KL}}(p(z)\|\|q(z))$	0.01 ± 0.005	0.013 ± 0.006	N/A	0.012 ± 0.004
$\hat{I}(x;\mu_z)$	0.254 ± 0.057	0.283 ± 0.112	N/A	0.24 ± 0.129
$\hat{I}(x;z)$	0.128 ± 0.025	0.006 ± 0.03	N/A	0.028 ± 0.017
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.004 ± 0.004	8.091 ± 5.185	N/A	5.252 ± 5.607
s^*_w	0.418 ± 0.083	0.304 ± 0.028	0.251 ± 0.137	0.311 ± 0.031
s^*_y	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s^*_z	0.248 ± 0.001	0.248 ± 0.0	N/A	0.249 ± 0.0
95%-MPIW Test (Unnorm)	6.862 ± 1.262	5.243 ± 0.573	2.007 ± 0.155	5.275 ± 0.569
95%-MPIW Train (Unnorm)	6.346 ± 0.821	4.906 ± 0.354	2.006 ± 0.153	4.957 ± 0.36
95%-PICP Test	95.5 ± 2.151	94.5 ± 2.0	63.4 ± 1.981	93.6 ± 2.485
95%-PICP Train	97.4 ± 1.14	95.5 ± 0.707	69.6 ± 4.519	94.9 ± 0.822
KS Test-Stat	0.031 ± 0.012	0.027 ± 0.006	N/A	0.029 ± 0.009
$\text{PC}(x,\mu_z)$	0.0 ± 0.0	0.0 ± 0.0	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.109 ± 0.036	0.879 ± 0.028	N/A	0.813 ± 0.159
Post-Pred Avg-LL Test	-1.285 ± 0.066	-1.211 ± 0.083	-2.846 ± 0.346	-1.278 ± 0.164
Post-Pred Avg-LL Train	-1.111 ± 0.065	-1.04 ± 0.057	-2.347 ± 0.154	-1.079 ± 0.065
RMSE Test (Unnorm)	0.635 ± 0.042	0.619 ± 0.039	0.607 ± 0.035	0.622 ± 0.039
RMSE Train (Unnorm)	0.635 ± 0.042	0.62 ± 0.039	0.607 ± 0.035	0.622 ± 0.039
Recon MSE	0.145 ± 0.008	0.159 ± 0.007	N/A	0.153 ± 0.006
Hyperparams	$\sigma^2_\epsilon = 0.1$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.1$, $\epsilon_y = 1.0$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.1$

Table 18: Experiment Evaluation Summary for Yuan($\pm$ std).

	NCAI$_\lambda$	NCAI$_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.012 ± 0.012	0.002 ± 0.002	N/A	0.001 ± 0.002
$D_{\text{KL}}(p(z)\|\|q(z))$	0.042 ± 0.012	0.003 ± 0.003	N/A	0.003 ± 0.003
$\hat{I}(x;\mu_z)$	0.029 ± 0.008	0.047 ± 0.011	N/A	0.045 ± 0.012
$\hat{I}(x;z)$	0.013 ± 0.007	0.022 ± 0.003	N/A	0.02 ± 0.004
HZ($\{\mu_{z_1},\ldots,\mu_{z_N}\}$)	1.641 ± 0.242	53.201 ± 1.247	N/A	52.566 ± 2.633
s_w^*	0.209 ± 0.017	0.15 ± 0.002	0.197 ± 0.183	0.147 ± 0.005
s_y^*	0.01 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
s_z^*	0.257 ± 0.001	0.251 ± 0.0	N/A	0.251 ± 0.0
95%-MPIW Test (Unnorm)	2.4 ± 0.121	2.45 ± 0.04	1.028 ± 0.014	2.466 ± 0.026
95%-MPIW Train (Unnorm)	2.396 ± 0.106	2.439 ± 0.04	1.027 ± 0.013	2.463 ± 0.034
95%-PICP Test	94.796 ± 1.185	94.279 ± 1.479	61.191 ± 2.176	94.734 ± 1.426
95%-PICP Train	94.897 ± 1.145	94.893 ± 0.286	63.265 ± 1.179	94.969 ± 0.257
PC(x,μ_z)	0.001 ± 0.0	0.004 ± 0.001	N/A	0.003 ± 0.0
PC(y,μ_z)	0.049 ± 0.01	0.142 ± 0.041	N/A	0.154 ± 0.055
Post-Pred Avg-LL Test	-0.849 ± 0.038	-1.147 ± 0.025	-1.709 ± 0.22	-1.143 ± 0.027
Post-Pred Avg-LL Train	-0.805 ± 0.033	-1.119 ± 0.013	-1.479 ± 0.056	-1.123 ± 0.015
RMSE Test (Unnorm)	0.983 ± 0.023	0.976 ± 0.016	0.92 ± 0.022	0.981 ± 0.017
RMSE Train (Unnorm)	0.983 ± 0.023	0.976 ± 0.017	0.92 ± 0.022	0.981 ± 0.017
Recon MSE	0.011 ± 0.001	0.114 ± 0.001	N/A	0.113 ± 0.001
Hyperparams	$\sigma_\epsilon^2 = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.1$, $\epsilon_y = 0.5$	$\sigma_\epsilon^2 = 0.01$	$\sigma_\epsilon^2 = 0.1$	$\sigma_\epsilon^2 = 0.1$

Table 19: Experiment Evaluation Summary for Wine Quality Red($\pm$ std).

	NCAI$_\lambda$	NCAI$_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.006 ± 0.008	-0.0 ± 0.002	N/A	0.005 ± 0.004
$D_{\text{KL}}(p(z)\|\|q(z))$	0.003 ± 0.006	0.001 ± 0.005	N/A	0.002 ± 0.004
$\hat{I}(x;\mu_z)$	0.086 ± 0.013	0.087 ± 0.012	N/A	0.077 ± 0.012
$\hat{I}(x;z)$	0.108 ± 0.002	0.108 ± 0.004	N/A	0.107 ± 0.001
HZ($\{\mu_{z_1},\ldots,\mu_{z_N}\}$)	5.42 ± 0.747	51.283 ± 9.548	N/A	27.059 ± 4.144
s_w^*	1.094 ± 0.903	1.137 ± 0.84	1.395 ± 1.155	0.384 ± 0.026
s_y^*	0.01 ± 0.0	0.01 ± 0.0	0.01 ± 0.0	0.01 ± 0.0
s_z^*	0.251 ± 0.001	0.246 ± 0.001	N/A	0.247 ± 0.0
95%-MPIW Test (Unnorm)	6.734 ± 0.212	6.712 ± 0.235	6.163 ± 0.178	8.095 ± 0.762
95%-MPIW Train (Unnorm)	6.724 ± 0.223	6.703 ± 0.217	6.163 ± 0.193	8.118 ± 0.601
95%-PICP Test	99.016 ± 2.199	98.689 ± 2.933	97.377 ± 5.865	98.033 ± 2.137
95%-PICP Train	99.444 ± 0.387	99.444 ± 0.387	97.593 ± 2.544	98.611 ± 0.655
PC(x,μ_z)	0.007 ± 0.003	0.007 ± 0.002	N/A	0.007 ± 0.003
PC(y,μ_z)	0.005 ± 0.002	0.022 ± 0.01	N/A	0.017 ± 0.01
Post-Pred Avg-LL Test	0.836 ± 0.074	0.832 ± 0.077	0.818 ± 0.187	0.638 ± 0.121
Post-Pred Avg-LL Train	0.865 ± 0.025	0.872 ± 0.024	0.868 ± 0.074	0.678 ± 0.047
RMSE Test (Unnorm)	0.005 ± 0.001	0.005 ± 0.001	0.005 ± 0.001	0.008 ± 0.001
RMSE Train (Unnorm)	0.005 ± 0.001	0.005 ± 0.001	0.005 ± 0.001	0.008 ± 0.001
Recon MSE	0.014 ± 0.0	0.014 ± 0.0	N/A	0.014 ± 0.001
Hyperparams	$\sigma_\epsilon^2 = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.5$, $\epsilon_y = 0.5$	$\sigma_\epsilon^2 = 0.01$	$\sigma_\epsilon^2 = 0.01$	$\sigma_\epsilon^2 = 0.01$

Table 20: Experiment Evaluation Summary for Yacht($\pm$ std).

	NCAI_{λ}	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.004 ± 0.002	0.004 ± 0.001	N/A	0.006 ± 0.002
$D_{\text{KL}}(p(z)\|\|q(z))$	0.008 ± 0.003	0.008 ± 0.004	N/A	0.01 ± 0.003
$\hat{I}(x;\mu_z)$	0.842 ± 0.06	0.373 ± 0.037	N/A	0.667 ± 0.061
$\hat{I}(x;z)$	0.035 ± 0.012	-0.015 ± 0.007	N/A	0.146 ± 0.026
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.005 ± 0.001	7.804 ± 1.727	N/A	5.09 ± 0.991
s^*_w	0.488 ± 0.026	0.444 ± 0.019	0.28 ± 0.132	0.231 ± 0.002
s^*_y	0.01 ± 0.0	0.01 ± 0.0	0.1 ± 0.0	0.01 ± 0.0
s^*_z	0.247 ± 0.0	0.247 ± 0.0	N/A	0.248 ± 0.0
95%-MPIW Test (Unnorm)	0.258 ± 0.026	0.247 ± 0.022	0.366 ± 0.005	0.313 ± 0.03
95%-MPIW Train (Unnorm)	0.293 ± 0.011	0.277 ± 0.012	0.366 ± 0.005	0.336 ± 0.014
95%-PICP Test	96.818 ± 2.591	96.364 ± 2.033	96.364 ± 3.447	95.455 ± 2.784
95%-PICP Train	95.871 ± 0.736	94.968 ± 0.957	93.161 ± 1.08	93.29 ± 0.866
KS Test-Stat	0.028 ± 0.005	0.025 ± 0.005	N/A	0.024 ± 0.009
$\text{PC}(x,\mu_z)$	0.0 ± 0.0	0.0 ± 0.0	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.007 ± 0.003	0.084 ± 0.009	N/A	0.121 ± 0.008
Post-Pred Avg-LL Test	0.263 ± 0.11	0.269 ± 0.107	-0.31 ± 0.069	0.129 ± 0.131
Post-Pred Avg-LL Train	0.155 ± 0.043	0.159 ± 0.046	-0.386 ± 0.035	-0.021 ± 0.053
RMSE Test (Unnorm)	0.995 ± 0.059	0.988 ± 0.061	1.143 ± 0.087	1.231 ± 0.057
RMSE Train (Unnorm)	0.994 ± 0.059	0.988 ± 0.062	1.143 ± 0.087	1.231 ± 0.056
Recon MSE	0.017 ± 0.0	0.017 ± 0.0	N/A	0.015 ± 0.001
Hyperparams	$\sigma^2_\epsilon = 0.01$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.5$, $\epsilon_y = 0.5$	$\sigma^2_\epsilon = 0.01$	$\sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 0.01$

Table 21: Experiment Evaluation Summary for Lidar($\pm$ std).

	NCAI_{λ}	$\text{NCAI}_{\lambda=0}$	BNN	BNN+LV
$D_{\text{JS}}(q(z)\|\|p(z))$	0.003 ± 0.001	0.005 ± 0.001	N/A	0.031 ± 0.005
$D_{\text{KL}}(p(z)\|\|q(z))$	0.008 ± 0.002	0.009 ± 0.002	N/A	0.357 ± 0.088
$\hat{I}(x;\mu_z)$	0.057 ± 0.017	0.032 ± 0.017	N/A	0.428 ± 0.04
$\hat{I}(x;z)$	0.047 ± 0.015	0.024 ± 0.014	N/A	0.387 ± 0.045
$\text{HZ}(\{\mu_{z_1},\ldots,\mu_{z_N}\})$	0.015 ± 0.004	0.792 ± 0.357	N/A	6.408 ± 2.439
s^*_w	34.575 ± 17.89	22.305 ± 7.342	1.805 ± 0.094	12.39 ± 4.903
s^*_y	0.1 ± 0.0	0.1 ± 0.0	1.0 ± 0.0	0.1 ± 0.0
s^*_z	1.0 ± 0.0	1.0 ± 0.0	N/A	1.0 ± 0.0
95%-MPIW Test (Unnorm)	7.375 ± 0.263	7.145 ± 0.16	4.011 ± 0.006	22.165 ± 10.073
95%-MPIW Train (Unnorm)	7.433 ± 0.299	7.114 ± 0.217	4.011 ± 0.001	22.267 ± 10.346
95%-PICP Test	93.84 ± 1.78	93.44 ± 1.757	73.68 ± 1.842	96.0 ± 1.095
95%-PICP Train	95.493 ± 0.256	95.227 ± 0.289	75.493 ± 1.489	96.773 ± 0.446
KS Test-Stat	0.014 ± 0.001	0.02 ± 0.002	N/A	0.044 ± 0.007
$\text{PC}(x,\mu_z)$	0.003 ± 0.001	0.0 ± 0.0	N/A	0.0 ± 0.0
$\text{PC}(y,\mu_z)$	0.035 ± 0.009	0.138 ± 0.014	N/A	0.161 ± 0.039
Post-Pred Avg-LL Test	-1.979 ± 0.04	-1.973 ± 0.049	-2.306 ± 0.059	-2.342 ± 0.048
Post-Pred Avg-LL Train	-1.92 ± 0.021	-1.895 ± 0.018	-2.217 ± 0.069	-2.229 ± 0.04
RMSE Test (Unnorm)	1.985 ± 0.051	1.932 ± 0.059	1.953 ± 0.071	3.523 ± 0.501
RMSE Train (Unnorm)	1.985 ± 0.051	1.933 ± 0.059	1.953 ± 0.071	3.521 ± 0.501
Recon MSE	0.124 ± 0.002	0.122 ± 0.001	N/A	0.123 ± 0.005
Hyperparams	$\sigma^2_z = 1.0$, $\sigma^2_\epsilon = 0.1$, $\lambda_2 = 10$, $\epsilon_T = 0.01$, $\epsilon_x = 0.5$, $\epsilon_y = 1.0$	$\sigma^2_z = 1.0, \sigma^2_\epsilon = 0.1$	$\sigma^2_\epsilon = 1.0$	$\sigma^2_z = 1.0$,$\sigma^2_\epsilon = 0.1$

Table 22: Experiment Evaluation Summary for Depeweg($\pm$ std).

References

Kolmogorov–Smirnov Test, pages 283–287. Springer New York, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_214. URL https://doi.org/10.1007/978-0-387-32833-1_214.

Hossein Baghishani and Mohsen Mohammadzadeh. Asymptotic normality of posterior distributions for generalized linear mixed models. *Journal of Multivariate Analysis*, 111: 66–77, 2012.

Maria Bauza and Alberto Rodriguez. A probabilistic data-driven model for planar pushing. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 3008–3015. IEEE, 2017.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. Active heteroscedastic regression. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 694–702, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/chaudhuri17a.html.

Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. Variational inference for uncertainty on the inputs of gaussian process models. *arXiv preprint arXiv:1409.2287*, 2014.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1192–1201, 2018.

Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405. PMLR, 2019.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

Yarin Gal et al. Uncertainty in deep learning. 2016.

Paul W Goldberg, Christopher KI Williams, and Christopher M Bishop. Regression with input-dependent noise: A gaussian process treatment. In *Advances in neural information processing systems*, pages 493–499, 1998.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773, 2012.

Ryan-Rhys Griffiths, AA Aldrick, GO Miguel, and AA Lee. Heteroscedastic bayesian optimisation in scientific discovery. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2019.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. *arXiv e-prints*, art. arXiv:1901.05534, Jan 2019.

Mikael Henaff, Alfredo Canziani, and Yann LeCun. Model-Predictive Policy Learning with Uncertainty Regularization for Driving in Dense Traffic. *arXiv e-prints*, art. arXiv:1901.02705, Jan 2019.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19(10):3595–3617, 1990. doi: 10.1080/03610929008830400. URL https://doi.org/10.1080/03610929008830400.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.pdf.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th international conference on Machine learning*, pages 393–400. ACM, 2007.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2013. arXiv: 1312.6114.

Peng Kou, Deliang Liang, Lin Gao, and Jianyong Lou. Probabilistic electricity price forecasting with variational heteroscedastic gaussian process and active learning. *Energy Conversion and Management*, 89:298–308, 2015.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. 69:066138, June 2004. doi: 10.1103/PhysRevE.69.066138.

Tony Lancaster. The incidental parameter problem since 1948. *Journal of econometrics*, 95 (2):391–413, 2000.

Neil D Lawrence and Andrew J Moore. Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488. ACM, 2007.

Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496. ACM, 2005.

H. K. Lee. Consistency of posterior distributions for neural networks. *Neural Networks: The Official Journal of the International Neural Network Society*, 13(6):629–642, July 2000. ISSN 0893-6080.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

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

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial Autoencoders. *arXiv e-prints*, art. arXiv:1511.05644, Nov 2015.

Andrew McHutchon and Carl E Rasmussen. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, pages 1341–1349, 2011.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-Directed Exploration for Deep Reinforcement Learning. *arXiv e-prints*, art. arXiv:1812.07544, Dec 2018.

Theodore Papamarkou, Jacob Hinkle, M. Todd Young, and David Womble. Challenges in Markov Chain Monte Carlo for Bayesian Neural Networks. *Statistical Science*, 37(3):425 – 442, 2022. doi: 10.1214/21-STS840. URL `https://doi.org/10.1214/21-STS840`.

Arya A Pourzanjani, Richard M Jiang, and Linda R Petzold. Improving the identifiability of neural networks for bayesian inference. In *NIPS Workshop on Bayesian Deep Learning*, volume 4, page 31, 2017.

M.W. Sigrist, S.M. W, J.D. Winefordner, and I.M. Kolthoff. *Air Monitoring by Spectroscopic Techniques*. Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications. Wiley, 1994. ISBN 9780471558750. URL `https://books.google.com/books?id=_vkyfs5_OAoC`.

Chunyi Wang and Radford M Neal. Gaussian process regression with heteroscedastic or non-gaussian residuals. *arXiv preprint arXiv:1212.6246*, 2012.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Peter M Williams. Using neural networks to model conditional multivariate densities. *Neural Computation*, 8(4):843–854, 1996.

WA Wright. Bayesian approach to neural-network modeling with input uncertainty. *IEEE Transactions on Neural Networks*, 10(6):1261–1270, 1999.

Ming Yuan and Grace Wahba. Doubly penalized likelihood estimator in heteroscedastic regression. *Statistics & probability letters*, 69(1):11–20, 2004.

Yao Zhang and Alpha A. Lee. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science*, 10(35):8154–8163, 2019. ISSN 2041-6520, 2041-6539. doi: 10.1039/C9SC00616H. URL `http://xlink.rsc.org/?DOI=C9SC00616H`.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, 2018.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1): 1799–1847, 2014.