JIRSS (2026)
Vol. 25, No. 1, pp 65-90
DOI: 10.22034/jirss.2026.2085011.1182

Bayesian Neural Networks for Nonlinear Regression: Posterior Inference, Uncertainty Quantification and Scalability

Hosseini, F.[1], **Karimi, O.**[1],

[1]Department of Statistics, Semnan University, Semnan, Iran.

Received: 2026-02-09, Accepted: 2026-06-15, Published online: 2026-06-28

Abstract.

Bayesian neural networks provide a probabilistic framework for nonlinear regression by combining the expressive flexibility of neural networks with principled uncertainty quantification. This study presents a comparative analysis of several inference approaches for Bayesian neural-network regression, including Hamiltonian Monte Carlo, the No-U-Turn Sampler and variational inference, with emphasis on predictive uncertainty, posterior calibration and computational efficiency. The results show that different inference strategies often achieve comparable predictive accuracy, whereas substantially larger differences emerge in uncertainty quantification and scalability. Sampling-based approaches provide more reliable posterior characterization and better-calibrated predictive uncertainty, particularly under complex noise structures, but incur substantially higher computational cost. In contrast, variational inference offers competitive predictive performance together with markedly improved computational efficiency. Overall, the findings suggest that the primary practical benefit of Bayesian neural networks lies in reliable and interpretable uncertainty quantification rather than solely improved point prediction. The choice of inference strategy should therefore balance posterior fidelity, uncertainty calibration, and computational scalability according to the requirements of the application.

Keywords. Bayesian neural networks, Hamiltonian Monte Carlo, Variational inference, Uncertainty calibration, Nonlinear regression.

MSC: 62J02, 62F15, 68T07.

Corresponding Author: Fatemeh Hosseini (fatemeh.hoseini@semnan.ac.ir).
Omid Karimi (omid.karimi@semnan.ac.ir).

1 Introduction

Nonlinear relationships are ubiquitous in modern data analysis and arise naturally in a wide range of scientific and applied domains, including data science, machine learning, economics, bioengineering, and complex system modeling (Rasmussen and Williams, 2006; Neal, 1996; Bishop, 2006). Such data are often characterized by intricate functional dependencies, interactions across multiple scales, and latent structures that cannot be adequately captured by simple linear models.

In many practical applications, predictive accuracy alone is insufficient, and reliable uncertainty quantification is equally essential for robust decision-making, particularly in areas such as medicine, finance, and safety-critical systems, where overconfident predictions may lead to substantial risk (Ghahramani, 2015; Kendall and Gal, 2017). For example, underestimation of predictive uncertainty in medical diagnosis systems may result in incorrect clinical decisions, while overconfident risk forecasts in financial systems can lead to substantial economic losses. Similarly, uncertainty miscalibration in autonomous or safety-critical systems may compromise reliability and operational safety. Consequently, uncertainty-aware machine learning models have become increasingly important in modern statistical and artificial intelligence applications.

Classical statistical approaches, including linear regression and generalized linear models, provide interpretable and computationally efficient tools for data analysis (McCullagh and Nelder, 1989). However, their dependence on restrictive assumptions, including linearity and specific error distributions, substantially limits their expressive capacity in settings characterized by complex nonlinear structure or heterogeneous noise. More flexible alternatives, such as generalized additive models (Hastie and Tibshirani, 1990), Gaussian processes (Rasmussen and Williams, 2006), and ensemble-based methods like random forests (Breiman, 2001), have been developed to address these limitations. While successful in many settings, these approaches may face challenges related to scalability, adaptation to high-dimensional inputs, or principled uncertainty quantification in complex models.

Neural networks constitute a highly expressive class of models for nonlinear regression and function approximation (Goodfellow et al., 2016). Through compositions of affine transformations and nonlinear activation functions, they are capable of capturing intricate patterns and high-order interactions in data. Despite their empirical success, conventional neural networks are typically trained via deterministic optimization procedures that yield point estimates of model parameters. As a consequence, uncertainty is not explicitly represented, which can lead to overly confident predictions, particularly in small-sample scenarios or under distributional shift (Gawlikowski et al., 2023).

Bayesian neural networks (BNNs) provide a principled probabilistic extension by placing prior distributions over network parameters and performing inference in a Bayesian framework (Neal, 1996; Gal and Ghahramani, 2016; Blundell et al., 2015). Predictions are obtained by marginalizing over the posterior distribution of the parameters, thereby capturing epistemic uncertainty arising from limited data or model complexity. From a complementary function-space perspective, priors on network parameters induce distributions over nonlinear functions, revealing close connections

between BNNs and Gaussian process models (Neal, 1996; Rasmussen and Williams, 2006). These properties make BNNs particularly appealing in applications where uncertainty quantification is as important as predictive performance.

Despite these conceptual advantages, posterior inference in BNNs remains computationally challenging. The resulting posterior distributions are typically high-dimensional, strongly correlated, and analytically intractable due to the nonlinear dependence of the likelihood on network parameters. As a result, practical inference relies on approximate numerical methods. Gradient-based Markov chain Monte Carlo (MCMC) techniques, such as Hamiltonian Monte Carlo (HMC) (Neal, 1996) and its adaptive extension, the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), provide asymptotically exact inference by efficiently exploring complex posterior landscapes. However, their computational cost can be substantial, particularly for larger networks or repeated analyses.

Variational inference (VI) has emerged as a popular alternative to mitigate these costs (Blei et al., 2017; Zhang et al., 2018). By reformulating posterior inference as an optimization problem, VI offers substantial gains in computational efficiency and scalability. In many empirical studies, variational approximations achieve competitive predictive accuracy, and in simple or well-specified settings their uncertainty estimates may appear comparable to those obtained from sampling-based methods. Nevertheless, it remains unclear how reliable such approximations are when data-generating conditions depart from idealized assumptions, for example in the presence of heavy-tailed noise, heteroscedasticity, weak parameter identifiability, or limited sample sizes.

The primary aim of this work is to critically assess the reliability and calibration of widely used Bayesian inference strategies for neural-network-based nonlinear regression under increasingly realistic modeling conditions, with particular emphasis on their practical suitability for complex regression problems. To this end, we present a unified empirical comparison of HMC, NUTS, and VI within a BNN framework, with the objective of clarifying how trade-offs between posterior fidelity, uncertainty calibration, and computational efficiency manifest across different data-generating scenarios. Our analysis is based on a sequence of controlled simulation studies designed to progressively increase modeling complexity, ranging from well-specified homoscedastic settings to scenarios involving heavy-tailed disturbances, input-dependent heteroscedasticity, and limited sample sizes.

Across these settings, we examine the empirical consequences of inference choices on predictive behavior and uncertainty characterization. While pointwise predictive accuracy is often comparable across methods, the results demonstrate pronounced differences in the calibration and reliability of posterior predictive uncertainty, particularly as data-generating assumptions become increasingly challenging. Taken together, the findings delineate conditions under which approximate Bayesian inference via variational methods provides a computationally efficient yet reliable alternative to sampling-based approaches, as well as scenarios in which fully Bayesian sampling is required to ensure well-calibrated uncertainty.

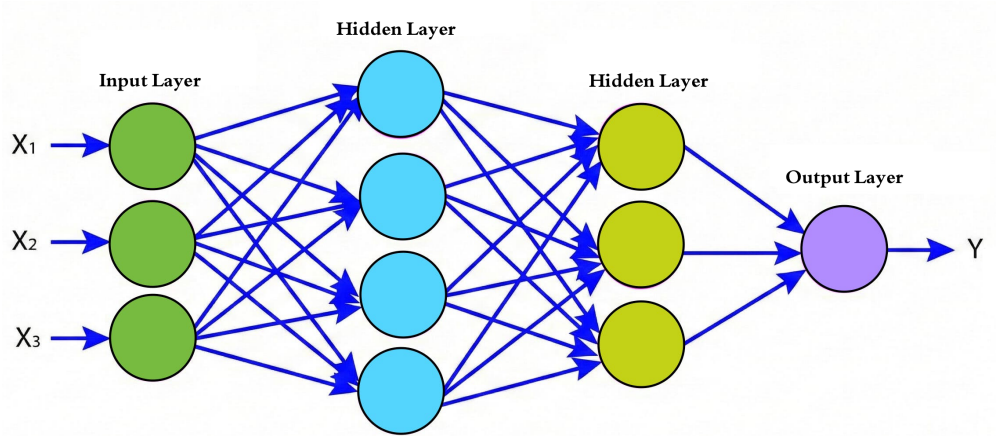


Figure 1: Schematic structure of a feedforward neural network with input, hidden, and output layers.

2 Neural Networks

Artificial neural networks (ANNs) are a class of parametric models designed to represent nonlinear mappings between input and output spaces through compositions of affine transformations and nonlinear activation functions. From a statistical perspective, an ANN defines a highly flexible function class capable of approximating complex relationships in high-dimensional settings, a property formalized by universal approximation results (Bishop, 2006; Hornik, 1991).

Let $x \in \mathbb{R}^d$ denote a d-dimensional input vector and consider a feedforward neural network composed of L layers. Denote by m_l the number of units in layer l, with $m_0 = d$. The network computes a sequence of intermediate representations according to

$$h^{(l)} = \phi^{(l)}\left(W^{(l)}h^{(l-1)} + b^{(l)}\right), \quad l = 1, \ldots, L,$$

where $h^{(0)} = x$, $W^{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$ is the weight matrix connecting layer $l-1$ to layer l, $b^{(l)} \in \mathbb{R}^{m_l}$ is the corresponding bias vector, and $\phi^{(l)}(\cdot)$ denotes a nonlinear activation function applied element-wise. Common choices for $\phi^{(l)}$ include rectified linear units, hyperbolic tangent functions, and sigmoidal activations. The output of the network is defined as $f(x;\boldsymbol{\theta}) = h^{(L)}$, which represents the model-based prediction of the response variable y given the input x, where

$$\boldsymbol{\theta} = \{W^{(l)}, b^{(l)}\}_{l=1}^{L},$$

collects all trainable parameters of the model. In regression settings, $f(x;\boldsymbol{\theta})$ is interpreted as a parametric approximation to an unknown target function relating the input x to the response variable y. By stacking multiple nonlinear layers, neural networks are able to capture intricate functional dependencies and high-order interactions among covariates, which underlies their widespread use in regression, classification, and representation learning tasks. Fig. 1 illustrates a feedforward architecture consisting of one

input layer, multiple hidden layers, and a single output layer, where each hidden layer performs a nonlinear transformation of its inputs. In conventional neural networks, the parameter vector $\boldsymbol{\theta}$ is treated as fixed but unknown and is typically estimated via deterministic optimization of a loss function, such as the squared error or negative log-likelihood. While this approach often yields accurate point predictions, it does not provide a principled mechanism for quantifying uncertainty in the model parameters or predictions. As a consequence, classical neural networks may produce overconfident predictions and are particularly susceptible to overfitting in small-sample or noisy scenarios, motivating the Bayesian extensions discussed in the following section.

2.1 Bayesian Neural Networks

Bayesian Neural Networks (BNNs) place neural-network modeling within a probabilistic framework by treating all weights and biases as random variables rather than fixed but unknown quantities (Neal, 1996; Blundell et al., 2015; Gal and Ghahramani, 2016). A BNN is specified through a prior distribution $p(\boldsymbol{\theta})$ together with a likelihood function induced by the neural-network output via an observation model. Given observed data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, Bayesian learning corresponds to posterior inference for

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

which represents updated beliefs about the network parameters after observing the data. Predictions in BNNs are obtained through Bayesian marginalization over the posterior distribution of the parameters. For a new input vector x^*, the predictive distribution is given by

$$p(y^* \mid x^*, \mathcal{D}) = \int p(y^* \mid x^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta},$$

yielding predictive distributions rather than deterministic point estimates. This probabilistic formulation enables uncertainty-aware prediction by propagating posterior variability through the neural-network model.

An alternative and complementary interpretation arises in function space: the prior distribution $p(\boldsymbol{\theta})$ induces a distribution over nonlinear functions $f(\cdot; \boldsymbol{\theta})$, thereby defining a prior over mappings from inputs to outputs. This perspective clarifies the close connection between Bayesian neural networks and Gaussian process models, particularly in wide-network limits and related constructions (Neal, 1996; Rasmussen and Williams, 2006). A broader synthesis of Bayesian deep learning methods, including function-space interpretations and practical approximate inference techniques, is provided in Wang and Yeung (2016).

Despite these conceptual advantages, exact posterior inference in Bayesian neural networks is generally infeasible because the posterior distribution is typically high-dimensional, strongly correlated, and analytically intractable. Consequently, practical implementations rely on approximate or numerical inference methods. The choice

of inference strategy therefore plays a central role in posterior approximation quality, uncertainty calibration, and computational efficiency, motivating the comparative analysis of inference approaches considered in this work.

2.2 Epistemic and Aleatoric Uncertainty in BNNs

A principal advantage of Bayesian neural networks lies in their ability to provide principled uncertainty quantification in addition to flexible nonlinear prediction. In many regression problems, particularly those arising in medicine, finance, engineering systems, and scientific modeling, predictive accuracy alone is insufficient for reliable decision-making. Two models with comparable pointwise predictive performance may nevertheless behave substantially differently in terms of predictive confidence, robustness, and uncertainty calibration. Consequently, modern uncertainty-aware machine learning focuses not only on estimation of the conditional mean structure of the response variable, but also on accurate characterization of predictive uncertainty.

From a probabilistic perspective, predictive uncertainty in Bayesian learning is commonly decomposed into two complementary components: epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty, often referred to as model uncertainty or parameter uncertainty, arises from incomplete knowledge of the underlying data-generating mechanism. In Bayesian neural networks, this uncertainty is represented through posterior variability in the network parameters. Since the observed data provide only partial information about the true regression function, uncertainty remains regarding the admissible parameter configurations governing the nonlinear mapping between inputs and outputs. As additional observations become available, posterior concentration typically increases, leading to a reduction in epistemic uncertainty.

In contrast, aleatoric uncertainty reflects the intrinsic stochastic variability of the observations and therefore cannot be eliminated even with arbitrarily large datasets. This component of uncertainty originates from measurement error, latent variability, incomplete covariate information, or fundamentally stochastic phenomena in the data-generating process. In regression models, aleatoric uncertainty is commonly represented through the observation model, for example via the variance structure of the likelihood distribution. When the observational variance depends explicitly on the input variables, the resulting uncertainty structure becomes heteroscedastic, implying that predictive variability changes across different regions of the covariate space.

This probabilistic framework reveals two distinct sources of predictive variability. The conditional distribution $p(y^* \mid x^*, \boldsymbol{\theta})$ captures uncertainty associated with the observation model and therefore reflects aleatoric variability, whereas the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$ represents epistemic uncertainty arising from incomplete knowledge of the model parameters. Unlike deterministic neural networks, which typically produce only point estimates of the regression function, Bayesian neural networks propagate both sources of uncertainty through posterior marginalization, thereby yielding predictive distributions rather than single deterministic predictions.

An important consequence of this formulation is improved uncertainty calibration under complex nonlinear structures. Deterministic neural networks trained via

maximum likelihood or empirical risk minimization frequently exhibit overconfident predictive behavior, particularly in regions characterized by sparse observations, distributional shift, or high observational variability. Because parameter uncertainty is ignored, predictive intervals derived from deterministic models may become unrealistically narrow and fail to accurately represent the true uncertainty associated with the underlying process. Bayesian neural networks address this limitation by explicitly incorporating posterior variability into prediction, thereby producing uncertainty estimates that are generally more robust and interpretable in noisy, weakly identifiable, or limited-sample settings.

These considerations are especially important in high-stakes applications where poorly calibrated predictive uncertainty may lead to unreliable or potentially harmful decisions. In medical diagnosis systems, overconfident predictions may reduce the likelihood of additional clinical evaluation in ambiguous cases, thereby increasing diagnostic risk. Similarly, in financial forecasting and risk analysis, underestimation of predictive uncertainty may result in unstable decision-making under volatile market conditions. Related challenges arise in autonomous systems, engineering reliability analysis, and other safety-critical environments in which uncertainty-aware prediction is essential for robust operational behavior.

The distinction between epistemic and aleatoric uncertainty also provides a useful conceptual framework for interpreting the empirical studies considered in this work. The heavy-tailed regression setting examined in Simulation Study II primarily investigates posterior robustness under non-Gaussian observational variability and limited sample sizes, whereas the heteroscedastic regression setting in Simulation Study III explicitly examines input-dependent aleatoric uncertainty and the ability of Bayesian neural networks to adapt predictive intervals to varying noise levels across the covariate domain.

Consequently, the comparative analysis of HMC, NUTS, and variational inference throughout this study should be interpreted not merely as a comparison of predictive accuracy, but more fundamentally as an investigation of how different Bayesian inference strategies affect the characterization, calibration, and computational tractability of predictive uncertainty in nonlinear regression models.

Figure 2 schematically summarizes the distinction between epistemic and aleatoric uncertainty in Bayesian neural networks. While epistemic uncertainty decreases as additional data become available, aleatoric uncertainty reflects intrinsic observational variability and therefore persists even in large-sample settings. This distinction provides a useful conceptual basis for interpreting the uncertainty-calibration behavior observed in the subsequent simulation studies.

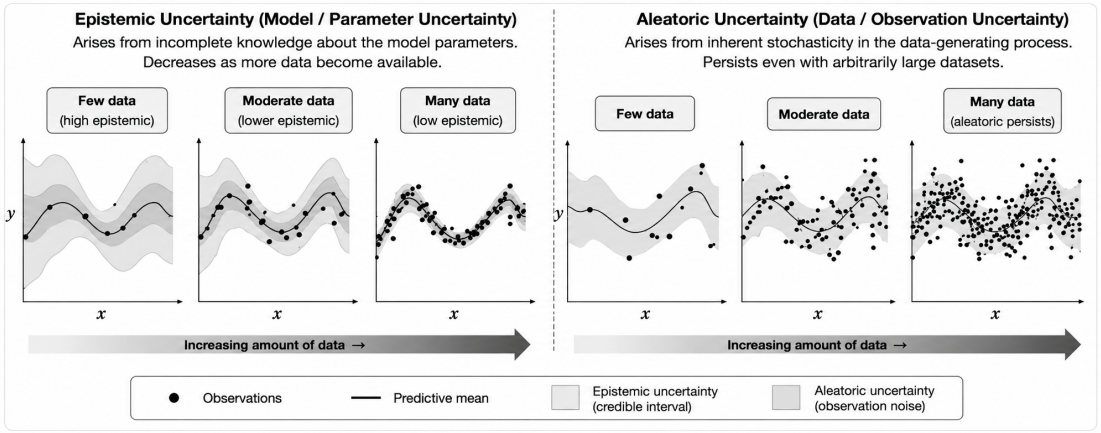


Figure 2: Schematic illustration of epistemic and aleatoric uncertainty in Bayesian neural networks. Epistemic uncertainty decreases as additional observations become available, whereas aleatoric uncertainty reflects intrinsic observational variability and persists even with large datasets.

3 Model Specification

Building on the neural network formulation introduced in the previous section, we model nonlinear relationships between the input variables and the corresponding responses using a BNN regression framework. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote the observed dataset, where $x_i \in \mathbb{R}^d$ denotes the input vector and $y_i \in \mathbb{R}$ is the corresponding response. The observation model is specified as

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \ldots, n,$$

where $f(\cdot; \boldsymbol{\theta})$ denotes the neural network mapping defined in the previous section and $\boldsymbol{\theta}$ collects all network weights and biases. A zero-mean isotropic Gaussian prior is placed on the network parameters, $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$, which provides regularization and stabilizes inference in high-dimensional parameter spaces. Conditional on $\boldsymbol{\theta}$, the likelihood factorizes as

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid x_i, \boldsymbol{\theta}),$$

where $p(y_i \mid x_i, \boldsymbol{\theta}) = \mathcal{N}\big(y_i \mid f(x_i; \boldsymbol{\theta}), \sigma^2\big)$ and $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2. Bayesian inference proceeds by targeting the posterior distribution of the network parameters,

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})},$$

where $p(\mathcal{D})$ denotes the marginal likelihood, which serves as the normalizing constant of the posterior distribution.

In BNNs, the resulting posterior distribution is typically high dimensional, strongly correlated, and analytically intractable due to the nonlinear dependence of the likelihood on the network parameters. As a result, practical inference must rely on approximate strategies. In this work, we consider both sampling-based methods that asymptotically target the exact posterior, namely HMC and its adaptive variant the NUTS, as well as a computationally efficient optimization-based approximation via VI. These approaches are evaluated in controlled nonlinear simulation settings in terms of predictive accuracy, uncertainty coverage, and computational cost.

3.1 Sampling-Based Inference via HMC and NUTS

MCMC methods based on Hamiltonian dynamics can be effective for high-dimensional posteriors because they exploit gradient information to construct proposals that avoid random-walk behavior (Neal, 1996). HMC augments $\boldsymbol{\theta}$ with auxiliary momentum variables $\boldsymbol{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, where $\mathbf{M}$ is a positive-definite mass matrix. Define the potential energy as the negative log of the joint density, ignoring additive constants that are independent of the parameters,

$$U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}) - \log p(\mathcal{D} \mid \boldsymbol{\theta}),$$

and define the kinetic energy as $K(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^\top \mathbf{M}^{-1}\boldsymbol{r}$. The Hamiltonian is then

$$H(\boldsymbol{\theta}, \boldsymbol{r}) = U(\boldsymbol{\theta}) + K(\boldsymbol{r}).$$

Hamiltonian dynamics are simulated approximately using the leapfrog integrator, yielding a proposal $(\boldsymbol{\theta}^*, \boldsymbol{r}^*)$ that is accepted via a Metropolis correction, ensuring that the Markov chain targets $p(\boldsymbol{\theta} \mid \mathcal{D})$.

In our BNN regression model with a Gaussian likelihood and a Gaussian prior, the log-posterior distribution, ignoring additive constants that do not depend on the parameters, is given by

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}) = \sum_{i=1}^{n} \log p(y_i \mid x_i, \boldsymbol{\theta}) - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2,$$

where $\|\boldsymbol{\theta}\|^2 = \boldsymbol{\theta}^\top \boldsymbol{\theta}$ denotes the squared Euclidean norm of the parameter vector, and

$$p(y_i \mid x_i, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i; \boldsymbol{\theta}))^2}{2\sigma^2}\right),$$

denotes the Gaussian likelihood density.

Gradient-based inference methods require evaluating derivatives of the log-posterior with respect to the network parameters $\boldsymbol{\theta}$. Since the neural network mapping $f(\cdot; \boldsymbol{\theta})$ is defined as a composition of differentiable functions, the resulting gradients are well defined and can be evaluated efficiently using automatic differentiation. This approach systematically applies the chain rule over the computational graph of the

Algorithm 1 Hamiltonian Monte Carlo for Bayesian Neural Network Inference

1. **Inputs:** data $\mathcal{D}$; $\log p(\boldsymbol{\theta})$, $\log p(\mathcal{D} \mid \boldsymbol{\theta})$ and gradients; step size ϵ; leapfrog steps L; mass matrix $\mathbf{M}$; number of posterior draws N.
2. Initialize $\boldsymbol{\theta}^{(0)}$.
3. For $n = 1, \ldots, N$:
   (a) Sample momentum $\boldsymbol{r}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$.
   (b) Set $(\boldsymbol{\theta}, \boldsymbol{r}) \leftarrow (\boldsymbol{\theta}^{(n-1)}, \boldsymbol{r}^{(0)})$.
   (c) Compute $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta})$.
   (d) **Leapfrog integration:**
   $$\boldsymbol{r} \leftarrow \boldsymbol{r} - \tfrac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
   For $l = 1, \ldots, L$:
   - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon\, \mathbf{M}^{-1} \boldsymbol{r}$
   - Recompute $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$
   - If $l < L$: $\boldsymbol{r} \leftarrow \boldsymbol{r} - \epsilon\, \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$

   $$\boldsymbol{r} \leftarrow \boldsymbol{r} - \tfrac{\epsilon}{2} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$$
   (e) Negate momentum: $\boldsymbol{r} \leftarrow -\boldsymbol{r}$.
   (f) Compute $H_{\text{old}} = U(\boldsymbol{\theta}^{(n-1)}) + \frac{1}{2} (\boldsymbol{r}^{(0)})^{\top} \mathbf{M}^{-1} \boldsymbol{r}^{(0)}$ and $H_{\text{new}} = U(\boldsymbol{\theta}) + \frac{1}{2} \boldsymbol{r}^{\top} \mathbf{M}^{-1} \boldsymbol{r}$.
   (g) Accept with probability $\alpha = \min\{1, \exp(H_{\text{old}} - H_{\text{new}})\}$ and set
   $$\boldsymbol{\theta}^{(n)} = \begin{cases} \boldsymbol{\theta}, & \text{with probability } \alpha, \\ \boldsymbol{\theta}^{(n-1)}, & \text{otherwise.} \end{cases}$$
4. Adapt ϵ and $\mathbf{M}$ during warm-up (Hoffman and Gelman, 2014).

model, thereby yielding exact gradients up to machine precision without relying on numerical approximation schemes.

In practice, HMC performance is sensitive to the choice of step size ϵ and trajectory length L. To reduce manual tuning, we also employ the NUTS, an adaptive variant of HMC that dynamically selects trajectory lengths by recursively building a binary tree of leapfrog steps and stopping when a U-turn condition is detected (Hoffman and Gelman, 2014). A warm-up phase is used to adapt ϵ, improving robustness and efficiency. All gradient-based inference procedures are implemented using automatic differentiation to compute $\nabla_{\boldsymbol{\theta}} \log p(\mathcal{D} \mid \boldsymbol{\theta})$.

For HMC and NUTS, we run an initial warm-up phase to stabilize the sampler and improve efficiency. During warm-up, the step size ϵ is adapted, and the mass matrix $\mathbf{M}$

Algorithm 2 No-U-Turn Sampler for Bayesian Neural Network Inference

1. **Inputs:** data $\mathcal{D}$; $U(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$; initial step size ϵ; mass matrix $\mathbf{M}$; number of posterior draws N.
2. Initialize $\boldsymbol{\theta}^{(0)}$.
3. For $n = 1, \ldots, N$:
   (a) Sample momentum $\boldsymbol{r}^{(0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$.
   (b) Sample slice variable:
   $$u \sim \text{Uniform}\Big(0, \exp\Big(- U(\boldsymbol{\theta}^{(n-1)}) - \tfrac{1}{2}(\boldsymbol{r}^{(0)})^\top \mathbf{M}^{-1} \boldsymbol{r}^{(0)}\Big)\Big).$$
   (c) Initialize: $\boldsymbol{\theta}^- = \boldsymbol{\theta}^+ = \boldsymbol{\theta}^{(n-1)}$, $\boldsymbol{r}^- = \boldsymbol{r}^+ = \boldsymbol{r}^{(0)}$, $j = 0$, $\boldsymbol{\theta}^{\text{cand}} = \boldsymbol{\theta}^{(n-1)}$, and $s = 1$.
   (d) While $s = 1$:
      i. Choose direction $v \in \{-1, +1\}$ uniformly.
      ii. Expand the tree by 2^j leapfrog steps in direction v, returning updated endpoints $(\boldsymbol{\theta}^-, \boldsymbol{r}^-)$, $(\boldsymbol{\theta}^+, \boldsymbol{r}^+)$, a candidate $\boldsymbol{\theta}'$, and an updated stop flag s based on the slice condition and divergence checks.
      iii. Update the candidate (e.g., proportional to the number of valid states in the new subtree): $\boldsymbol{\theta}^{\text{cand}} \leftarrow \boldsymbol{\theta}'$.
      iv. Check the no-U-turn stopping rule:
      $$s \leftarrow s \cdot \mathbb{I}\Big((\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)^\top \boldsymbol{r}^- \geq 0\Big) \cdot \mathbb{I}\Big((\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)^\top \boldsymbol{r}^+ \geq 0\Big).$$
      v. Set $j \leftarrow j + 1$.
   (e) Set $\boldsymbol{\theta}^{(n)} \leftarrow \boldsymbol{\theta}^{\text{cand}}$.
4. Adapt ϵ and $\mathbf{M}$ during warm-up (Hoffman and Gelman, 2014).

may be updated when supported by the software backend. After warm-up, samples are collected from the stationary phase. To assess sampling quality and convergence, we report standard MCMC diagnostics including the potential scale reduction factor ($\widehat{R}$) and effective sample size (ESS). We also monitor divergent transitions and empirical acceptance behavior as indicators of sampler stability. Unless otherwise stated, posterior summaries and uncertainty quantification are computed from post-warm-up draws after thinning to reduce autocorrelation. Algorithms 1 and 2 summarize the sampling-based inference methods used in this study. HMC constructs efficient proposals using Hamiltonian dynamics and gradient information, while NUTS extends HMC by adaptively selecting the trajectory length to avoid manual tuning. Both methods employ a warm-up phase to improve efficiency and stability.

3.2 Optimization-Based Inference via Variational Methods

Variational inference (VI) provides an optimization-based framework for approximate Bayesian inference in high-dimensional models where exact posterior computation is analytically intractable (Blei et al., 2017; Zhang et al., 2018). Rather than generating samples from the posterior distribution, VI approximates the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ by introducing a tractable variational distribution $q(\boldsymbol{\theta}; \boldsymbol{\phi})$ parameterized by variational parameters $\boldsymbol{\phi}$. The objective is to determine the member of the variational family that is closest to the true posterior distribution in the Kullback–Leibler sense.

This approximation is obtained by maximizing the evidence lower bound (ELBO),

$$\mathrm{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\phi})}[\log p(\mathcal{D}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta}; \boldsymbol{\phi})]\,,$$

which satisfies

$$\log p(\mathcal{D}) = \mathrm{ELBO}(\boldsymbol{\phi}) + D_{\mathrm{KL}}(q(\boldsymbol{\theta}; \boldsymbol{\phi}) \,\|\, p(\boldsymbol{\theta} \mid \mathcal{D}))\,.$$

Since the Kullback–Leibler divergence is nonnegative, maximizing the ELBO is equivalent to minimizing the discrepancy between the variational approximation and the true posterior distribution.

Using the decomposition $p(\mathcal{D}, \boldsymbol{\theta}) = p(\mathcal{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$, the ELBO can be written as

$$\mathrm{ELBO}(\boldsymbol{\phi}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\phi})}[\log p(\mathcal{D} \mid \boldsymbol{\theta})] - D_{\mathrm{KL}}(q(\boldsymbol{\theta}; \boldsymbol{\phi}) \,\|\, p(\boldsymbol{\theta}))\,,$$

where

$$D_{\mathrm{KL}}(q(\boldsymbol{\theta}; \boldsymbol{\phi}) \,\|\, p(\boldsymbol{\theta})) = \int q(\boldsymbol{\theta}; \boldsymbol{\phi}) \log\left(\frac{q(\boldsymbol{\theta}; \boldsymbol{\phi})}{p(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}$$

, denotes the Kullback–Leibler divergence between the variational distribution and the prior distribution.

In this work, we adopt a mean-field Gaussian variational family in which the variational distribution factorizes across parameters as

$$q(\boldsymbol{\theta}; \boldsymbol{\phi}) = \prod_{j=1}^{p} q(\theta_j; \phi_j),$$

with each component modeled as an independent univariate Gaussian distribution,

$$q(\theta_j; \phi_j) = \mathcal{N}(\mu_j, \sigma_j^2).$$

The variational parameters are given by $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{\rho})$, where $\boldsymbol{\mu}$ denotes the vector of variational means and the standard deviations are parameterized through

$$\boldsymbol{\sigma} = \log(1 + \exp(\boldsymbol{\rho})),$$

ensuring positivity of the variance parameters.

The expectation appearing in the ELBO is approximated using Monte Carlo integration together with the reparameterization trick, yielding low-variance stochastic

Algorithm 3 Variational Inference for Bayesian Neural Network Inference

1. **Inputs:** data $\mathcal{D}$; prior $p(\boldsymbol{\theta})$; likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$; learning rate η; iterations T.
2. Choose $q(\boldsymbol{\theta}; \boldsymbol{\phi})$ mean-field Gaussian with $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{\rho})$, $\boldsymbol{\sigma} = \log(1 + \exp(\boldsymbol{\rho}))$.
3. Initialize $\boldsymbol{\mu}^{(0)}, \boldsymbol{\rho}^{(0)}$.
4. For $t = 1, \ldots, T$:
   (a) Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and reparameterize $\boldsymbol{\theta}^{(t)} = \boldsymbol{\mu}^{(t-1)} + \boldsymbol{\sigma}^{(t-1)} \odot \boldsymbol{\epsilon}$.
   (b) Compute a stochastic ELBO estimate:
   $$\widehat{\mathrm{ELBO}}(\boldsymbol{\phi}) = \log p(\mathcal{D} \mid \boldsymbol{\theta}^{(t)}) + \log p(\boldsymbol{\theta}^{(t)}) - \log q(\boldsymbol{\theta}^{(t)}; \boldsymbol{\phi}).$$
   (c) Update $\boldsymbol{\phi}$ using Adam:
   $$\boldsymbol{\phi}^{(t)} \leftarrow \boldsymbol{\phi}^{(t-1)} + \eta \cdot \mathrm{Adam}\Big(\nabla_{\boldsymbol{\phi}} \widehat{\mathrm{ELBO}}(\boldsymbol{\phi})\Big).$$
5. **Output:** variational parameters $\boldsymbol{\phi}^{(T)}$.

gradient estimators. Optimization of the ELBO is performed using the Adam optimizer (Kingma and Ba, 2014), and convergence is monitored through the ELBO trajectory and stability across repeated runs.

Algorithm 3 summarizes the variational inference procedure used for approximate posterior inference in Bayesian neural networks. The algorithm employs the reparameterization trick to obtain low-variance stochastic gradient estimates of the ELBO, enabling efficient gradient-based optimization in high-dimensional parameter spaces.

4 Simulation Study I: Larger-Scale Nonlinear Regression

To evaluate the behavior of Bayesian inference strategies in a more realistic Bayesian neural-network setting, we consider a larger-scale nonlinear regression problem involving multiple covariates and a deeper neural network architecture. This experiment replaces the introductory low-dimensional nonlinear example and is designed to jointly assess predictive accuracy, uncertainty calibration, posterior diagnostic behavior, sensitivity to initialization, and computational scalability.

Input vectors are generated independently as

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})^\top \sim \mathrm{Uniform}(-2, 2)^5, \qquad i = 1, \ldots, n,$$

with sample size fixed at $n = 1000$. The response variable is generated according to

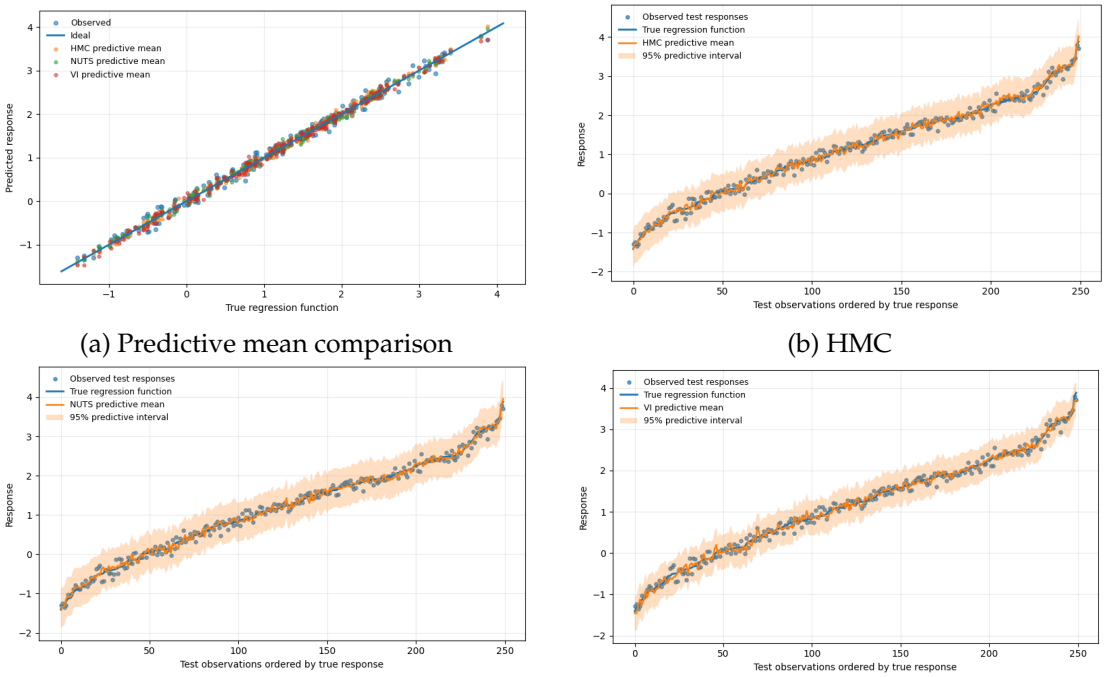
$$y_i = \sin(x_{i1}) + 0.5x_{i2}^2 - 0.3x_{i3}x_{i4} + \exp(-x_{i5}^2) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 0.1^2).$$

This data-generating mechanism combines nonlinear main effects with interaction structure and therefore provides a substantially more demanding benchmark than a simple univariate nonlinear regression setting.

A Bayesian neural network with architecture $5 \to 20 \to 20 \to 1$ is fitted to the simulated data. The hidden layers employ nonlinear activation functions, while the output layer represents the conditional mean of the Gaussian regression model. Independent zero-mean Gaussian priors are assigned to all network weights and biases. Posterior inference is performed using HMC, NUTS, and mean-field variational inference. For HMC and NUTS, warm-up adaptation is used for step-size and mass-matrix tuning, whereas VI is optimized using the Adam optimizer together with the reparameterization trick. To reduce dependence on a single random realization, the simulation was repeated over 50 independent replications. Reported results correspond to averages across replications. Additional sensitivity analyses were conducted by varying random parameter initializations and hidden-layer widths. The resulting qualitative conclusions remained stable across these alternative configurations. Model performance is evaluated using mean squared prediction error (MSE), empirical coverage of 95% posterior predictive intervals, average predictive interval width, computational runtime, and uncertainty calibration diagnostics. For sampling-based methods, convergence diagnostics including the potential scale reduction factor $\hat{R}$ and effective sample size (ESS) are additionally monitored.

Figure 3 summarizes the posterior predictive behavior of the different Bayesian inference strategies. Since the regression problem is multivariate, test observations are ordered according to the true regression function to facilitate one-dimensional visualization. Panel (a) compares posterior predictive means against the true regression function for all methods simultaneously. Predicted responses remain highly concentrated around the identity line, indicating that HMC, NUTS, and VI all achieve accurate point prediction performance in this larger-scale nonlinear setting. Panels (b)–(d) present posterior predictive means together with 95% posterior predictive intervals obtained from HMC, NUTS, and variational inference, respectively. All three methods successfully recover the dominant nonlinear structure of the data-generating mechanism and produce highly similar predictive means. The primary differences arise in uncertainty quantification. HMC and NUTS generate slightly wider predictive intervals, particularly in regions where the regression surface changes more rapidly, whereas VI produces somewhat narrower intervals, reflecting mild posterior variance underestimation associated with the mean-field approximation.

Figure 4 presents additional posterior predictive diagnostics. Panel (a) reports calibration curves comparing nominal and empirical predictive coverage levels. The dashed diagonal line represents ideal calibration. HMC and NUTS remain very close to the ideal diagonal across the considered confidence levels, indicating well-calibrated posterior predictive uncertainty. VI follows the same overall trend but exhibits mild undercoverage at higher nominal confidence levels, consistent with the tendency of mean-field variational approximations to underestimate posterior variance. Panel (b) presents posterior predictive checks based on replicated draws from the posterior predictive distribution. The replicated responses reproduce the overall shape, spread,



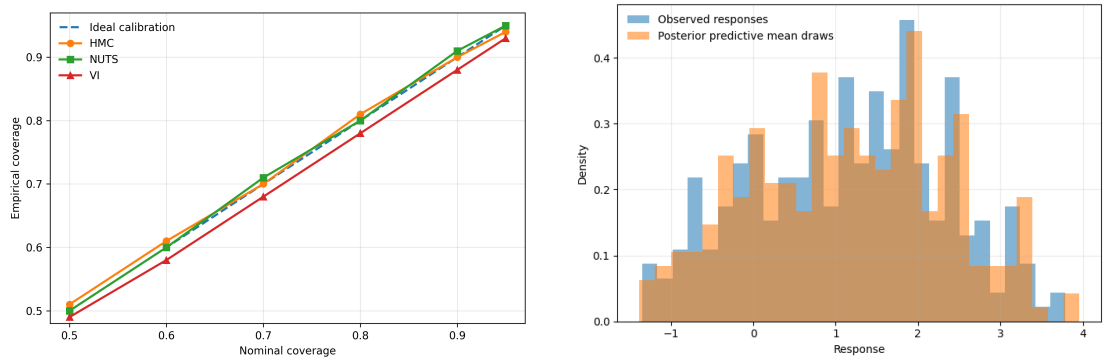
(a) Predictive mean comparison

(b) HMC

(c) NUTS

(d) Variational inference

Figure 3: Posterior predictive results for Simulation Study I under different Bayesian inference strategies. Panel (a) compares predictive means against the true regression function, while panels (b)–(d) show posterior predictive means and 95% predictive intervals obtained from HMC, NUTS, and variational inference.



(a) Calibration curves

(b) Posterior predictive check

Figure 4: Additional posterior predictive diagnostics for Simulation Study I: (a) calibration curves and (b) posterior predictive checks.

and central tendency of the observed distribution reasonably well, indicating that the Bayesian neural-network specification provides an adequate probabilistic representa-

tion of the simulated nonlinear regression process. The numerical values underlying the calibration analysis are summarized in Table 1. At the nominal 95% level, HMC and NUTS achieve empirical coverage values of 0.94 and 0.95, respectively, whereas VI attains 0.93. Thus, although VI remains reasonably calibrated overall, its predictive intervals are slightly more optimistic than those produced by sampling-based posterior inference.

Table 1: Nominal and empirical coverage levels used for calibration analysis.

Method	50%	60%	70%	80%	90%	95%
HMC	0.51	0.61	0.70	0.81	0.90	0.94
NUTS	0.50	0.60	0.71	0.80	0.91	0.95
VI	0.49	0.58	0.68	0.78	0.88	0.93

Quantitative performance measures are summarized in Table 2. All three methods achieve low MSE values, indicating broadly comparable point prediction accuracy. More substantial differences arise in uncertainty quantification and computational efficiency. HMC and NUTS attain coverage probabilities closest to the nominal 95% level and produce slightly wider predictive intervals, whereas VI achieves substantially shorter runtime together with competitive predictive accuracy. However, its narrower predictive intervals lead to mild undercoverage relative to the sampling-based approaches.

Table 2: Performance comparison for larger-scale nonlinear regression.

Method	MSE	Coverage (95%)	Avg. Interval Width	Runtime (s)	ESS
HMC	0.0124	0.94	0.4821	684.35	1248
NUTS	0.0118	0.95	0.4967	812.60	1395
VI	0.0131	0.93	0.4518	96.42	—

For HMC and NUTS, convergence diagnostics indicated stable posterior sampling, with $\hat{R} < 1.01$ across monitored parameters and sufficiently large effective sample sizes. No substantial pathological sampling behavior was observed after warm-up adaptation. Sensitivity analyses over repeated initializations and alternative hidden-layer widths yielded the same qualitative ranking of the methods, indicating that the conclusions are not driven by a particular initialization or network specification.

This experiment demonstrates that all inference strategies achieve competitive predictive accuracy in larger-scale nonlinear regression, although sampling-based approaches provide slightly better uncertainty calibration at substantially higher computational cost.

5 Simulation Study II: Heavy-Tailed Noise

The first simulation study demonstrated that, in simple and well-specified nonlinear regression settings, different Bayesian inference strategies can yield nearly indistinguishable predictive performance and uncertainty estimates. While such results are reassuring, they may conceal important differences that emerge under more realistic data-generating mechanisms. In practice, observational noise is often heavy-tailed and data may contain atypical or extreme observations, conditions under which posterior inference becomes more sensitive to both modeling assumptions and algorithmic choices.

To investigate inference behavior in such settings, we consider a nonlinear regression problem similar in structure to Simulation Study I but with heavy-tailed observation noise. Covariates are sampled uniformly from the interval $[-3, 3]$, and responses are generated according to

$$y_i = \sin(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim t_\nu(0, \sigma),$$

where $t_\nu(0, \sigma)$ denotes a Student-t distribution with $\nu = 3$ degrees of freedom and scale parameter $\sigma = 0.1$. To mimic occasional extreme deviations commonly encountered in practice, 5% of observations are further contaminated by inflating the noise scale. The sample size is fixed at $n = 50$, representing a small-sample regime in which epistemic uncertainty is non-negligible and uncertainty calibration becomes practically relevant. We fit a Bayesian nonlinear regression model with the same prior specification and inference settings as Simulation Study I,

$$f(x; \boldsymbol{\theta}) = \theta_0 + \theta_1 \sin(\theta_2 x),$$

using identical prior specifications so that observed differences can be attributed to the inference strategy rather than to model structure. Posterior inference is performed using a tuned HMC sampler and mean-field VI. Inference performance is evaluated using MSE for point predictions, empirical coverage of 95% posterior predictive intervals, the average width of these intervals computed on a dense grid over $[-3, 3]$, and computational runtime.

Fig. 5 displays posterior predictive means and 95% predictive intervals for both methods. Both HMC and VI accurately recover the underlying sinusoidal regression function, and differences in point predictions are visually negligible. The primary distinction lies in uncertainty quantification and computational cost rather than in predictive accuracy.

Quantitative results are summarized in Table 3. Both methods achieve nearly identical MSE values and the same empirical coverage level of 0.92, slightly below the nominal 95% level, which is expected given the small sample size and the presence of heavy-tailed noise with outlier contamination. The average predictive interval width is comparable across methods, with VI producing a marginally wider interval on average. In contrast, computational efficiency differs substantially: mean-field VI achieves similar posterior predictive summaries while reducing runtime by approximately a factor of four relative to HMC.

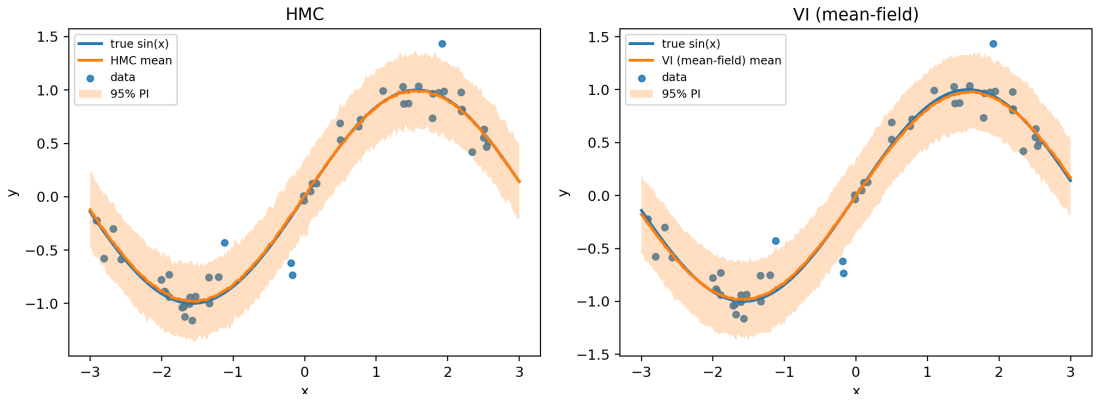


Figure 5: Posterior predictive results for Simulation Study II under heavy-tailed noise with outliers.

Overall, this simulation highlights an important practical message. When a robust likelihood model is employed to accommodate heavy-tailed noise, approximate inference via variational methods can deliver uncertainty estimates that are comparable to those obtained by sampling-based approaches, while offering a substantial reduction in computational cost. Under such conditions, differences among inference strategies are driven less by predictive accuracy and more by the trade-off between computational efficiency and algorithmic robustness.

Table 3: Performance comparison under heavy-tailed noise (Simulation Study II, $n = 50$, $\nu = 3$, 5% outliers).

Method	MSE	Coverage (95%)	Avg. Interval Width	Runtime (s)
HMC (tuned)	0.03094	0.92	0.69793	115.05
VI (mean-field)	0.03048	0.92	0.71754	25.90

6 Simulation Study III: Heteroscedastic Regression

The previous simulation studies focused on settings with homoscedastic Gaussian noise and heavy-tailed disturbances. While these scenarios already illustrate meaningful differences between inference strategies, many real-world regression problems exhibit an additional layer of complexity: the variability of the observations depends explicitly on the input.

Such heteroscedastic structures are common in practice and present challenges for models that rely solely on deterministic parameter estimates or simplified uncertainty assumptions. To investigate this setting, we consider a nonlinear regression problem with input-dependent noise. Covariates are sampled uniformly from the interval

$[-3, 3]$, and responses are generated according to

$$y_i = \sin(x_i) + \sigma(x_i)\varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1), \quad \sigma(x) = 0.10 + 0.15|x|.$$

This construction induces heteroscedasticity, with noise variance increasing smoothly as $|x|$ grows. Consequently, predictive uncertainty is expected to remain relatively narrow near the center of the domain and gradually widen toward the boundary regions where observational variability becomes larger.

We compare two modeling approaches. The first is a classical neural network trained via maximum likelihood estimation (ANN-MLE), in which network parameters are treated as fixed and uncertainty is captured solely through the estimated noise model. The second is a BNN with variational inference (BNN-VI), which explicitly models uncertainty in the network parameters and propagates this uncertainty to predictions. The network architecture is fixed as $1 \to 20 \to 20 \to 2$, where the two output units correspond to the conditional mean $\mu(x)$ and standard deviation $\sigma(x)$. This design enables direct modeling of input-dependent variance in both approaches. Predictive means and 95% predictive intervals are evaluated on a dense grid of 300 points over $[-3, 3]$.

Fig. 6 visualizes the resulting predictive behavior. Both models accurately recover the underlying sinusoidal regression structure, indicating that the nonlinear mean function is well captured under heteroscedastic noise. Differences between the methods arise primarily in predictive uncertainty characterization. The ANN-MLE model produces relatively stable predictive intervals across the covariate domain, whereas the BNN-VI approach yields uncertainty bands that adapt more flexibly to changes in the input-dependent noise structure. In particular, predictive intervals become gradually wider in regions associated with larger observational variability, reflecting the heteroscedastic nature of the data-generating mechanism.

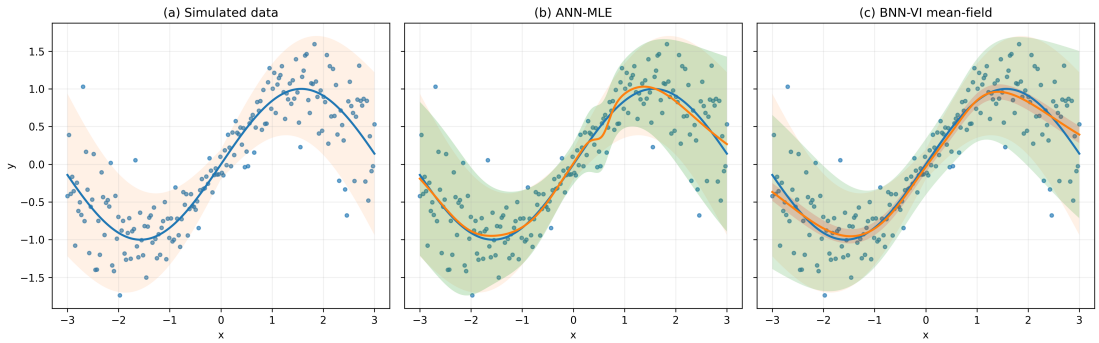


Figure 6: Posterior predictive behavior in Simulation Study III under heteroscedastic nonlinear regression. While both ANN-MLE and BNN-VI capture the nonlinear regression pattern, the Bayesian model provides smoother and input-adaptive predictive uncertainty estimates, particularly in regions with increased observational variability.

Quantitative results are reported in Table 4. While both methods achieve comparable MSE, the BNN attains empirical coverage closer to the nominal 95% level. This

improvement in uncertainty calibration comes at a higher computational cost, reflecting the additional complexity of Bayesian inference. Overall, the heteroscedastic regression experiment demonstrates that Bayesian neural networks provide more adaptive and better-calibrated predictive uncertainty under input-dependent noise structures, although this improvement is accompanied by increased computational cost relative to deterministic neural networks.

Table 4: Performance comparison under heteroscedastic noise.

Method	MSE	Coverage (95%)	Avg. Interval Width	Runtime (s)
ANN-MLE	0.0169	0.95	1.334	12.24
BNN-VI	0.0203	0.97	1.338	36.02

7 Real-World Regression Experiment: Energy Efficiency Dataset

To complement the controlled simulation experiments and evaluate practical performance under real observational data, we additionally consider a real-world nonlinear regression problem based on the Energy Efficiency dataset from the UCI Machine Learning Repository. This dataset is frequently used as a benchmark for nonlinear regression and uncertainty-aware prediction because it exhibits moderate dimensionality, nonlinear feature-response relationships, and heterogeneous variability patterns while remaining computationally tractable for Bayesian neural-network inference.

The dataset contains building design characteristics together with corresponding heating and cooling energy demands. Following common practice in the literature, the heating-load variable is used as the response. The predictor variables consist of relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. Let $\mathbf{x}_i \in \mathbb{R}^8$ denote the covariate vector associated with the ith building configuration, and let y_i denote the corresponding heating-load measurement. This dataset is particularly suitable for the present study for several reasons. First, the relationship between architectural design variables and heating demand is strongly nonlinear, making simple linear regression assumptions inadequate. Second, the dataset contains interacting geometric effects and moderate heterogeneity in response variability, which provides a meaningful setting for evaluating posterior uncertainty quantification. Third, the relatively small-to-moderate dataset size allows direct comparison between sampling-based Bayesian inference methods and variational approximations without making computation prohibitively expensive. Consequently, the dataset provides an appropriate intermediate benchmark between controlled synthetic simulations and large-scale industrial prediction tasks.

Prior to model fitting, all predictor variables are standardized to have zero mean and unit variance. The data are randomly divided into training and testing subsets using an 80/20 split. To reduce dependence on a single partition, the experiment is

repeated over 30 independent train-test splits, and the reported results correspond to averages across replications.

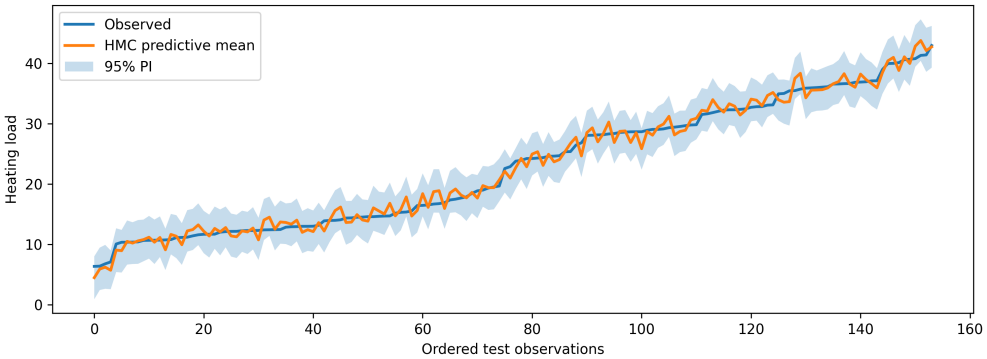
A Bayesian neural network with architecture $8 \to 20 \to 20 \to 1$ is fitted to the training data. The two hidden layers employ nonlinear activation functions, while the output layer represents the conditional mean of a Gaussian regression likelihood. Independent zero-mean Gaussian priors are assigned to all network weights and biases. Posterior inference is performed using HMC, NUTS, and mean-field variational inference under computational settings comparable to those used in the simulation studies.

The selected architecture balances representational flexibility and computational tractability. Two hidden layers allow the model to capture nonlinear interactions among building characteristics, while the moderate hidden-layer width avoids excessive parameter dimensionality that could make sampling-based posterior inference prohibitively expensive. Additional sensitivity checks using different random initializations and alternative hidden-layer widths produced qualitatively similar conclusions. Model performance is evaluated using mean squared prediction error (MSE), empirical coverage of posterior predictive intervals, average predictive interval width, computational runtime, calibration diagnostics, and posterior predictive checks. For HMC and NUTS, convergence diagnostics including the potential scale reduction factor $\hat{R}$ and effective sample size (ESS) are also monitored.

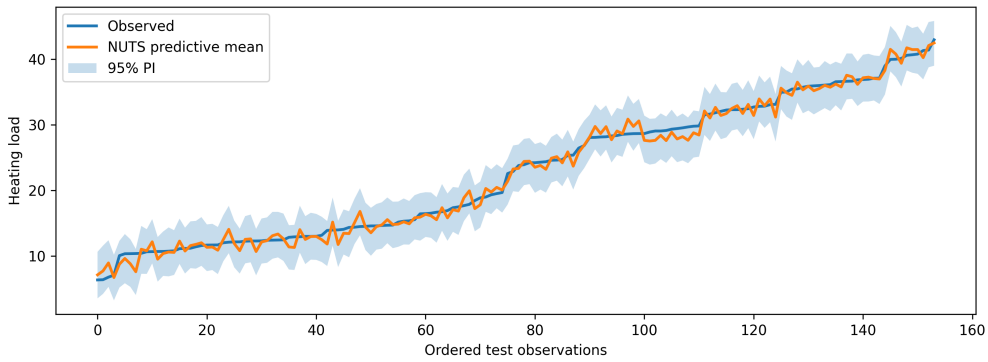
Figures 7a–7c present posterior predictive summaries for the Energy Efficiency dataset under different Bayesian inference strategies. Since the regression problem is multivariate, observations are ordered according to the observed heating-load values to facilitate one-dimensional visualization. All three methods successfully capture the dominant nonlinear relationship between building characteristics and heating load. The posterior predictive means closely follow the observed response pattern, indicating satisfactory predictive performance. The primary differences arise in uncertainty quantification. HMC and NUTS produce slightly wider predictive intervals and more conservative uncertainty estimates, whereas VI generates somewhat narrower intervals, consistent with mild posterior variance underestimation under the mean-field approximation.

Figure 8 presents additional predictive diagnostics for the Energy Efficiency dataset. Panel (a) compares posterior predictive means against the observed heating-load responses for all inference strategies simultaneously. Predicted responses remain highly concentrated around the diagonal line, indicating competitive point prediction accuracy across all methods. Panel (b) reports calibration curves comparing nominal and empirical predictive coverage levels. HMC and NUTS remain close to ideal calibration across most confidence levels, whereas VI exhibits mild undercoverage at higher nominal levels.

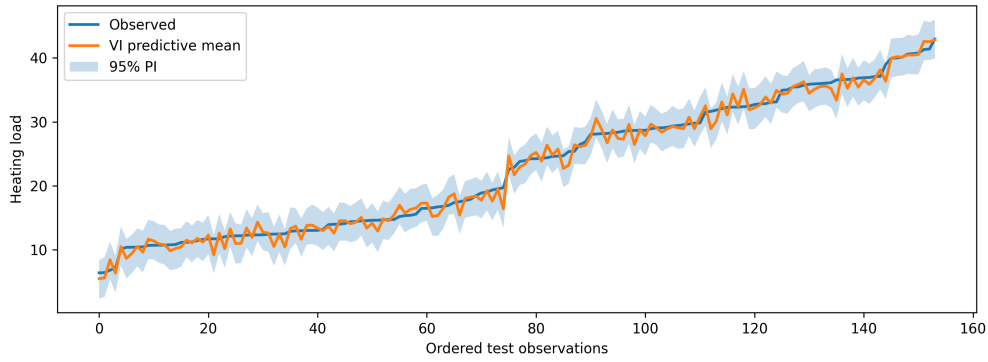
The numerical values underlying the calibration analysis are summarized in Table 5. At the nominal 95% level, HMC and NUTS achieve empirical coverage values of 0.94 and 0.95, respectively, whereas VI achieves 0.92. Thus, VI remains reasonably calibrated but produces slightly more optimistic predictive intervals than the sampling-based methods.



(a) HMC



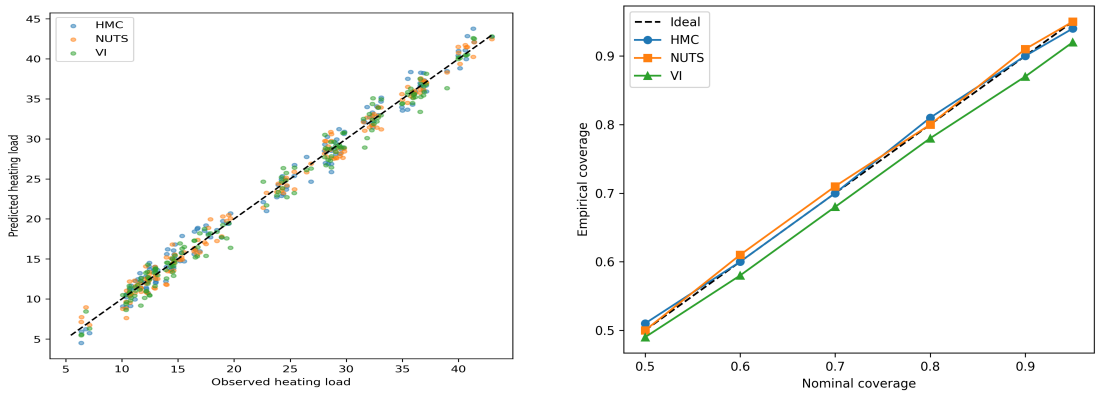
(b) NUTS



(c) Variational inference

Figure 7: Posterior predictive summaries for the Energy Efficiency dataset under different Bayesian inference strategies. The solid curves denote posterior predictive means, while shaded regions represent 95% posterior predictive intervals.

Posterior predictive checks are shown in Figure 9. The replicated posterior predictive responses reproduce the overall location, spread, and shape of the observed heating-load distribution reasonably well. No substantial systematic discrepancy is visible, suggesting that the Bayesian neural-network specification provides an ade-



(a) Predictive mean comparison (b) Calibration curves

Figure 8: Additional predictive diagnostics for the Energy Efficiency dataset: (a) predictive mean comparison and (b) calibration curves.

Table 5: Nominal and empirical coverage levels for the Energy Efficiency dataset.

Method	50%	60%	70%	80%	90%	95%
HMC	0.51	0.60	0.70	0.81	0.90	0.94
NUTS	0.50	0.61	0.71	0.80	0.91	0.95
VI	0.49	0.58	0.68	0.78	0.87	0.92

quate probabilistic representation of the dominant nonlinear energy-demand structure in the data.

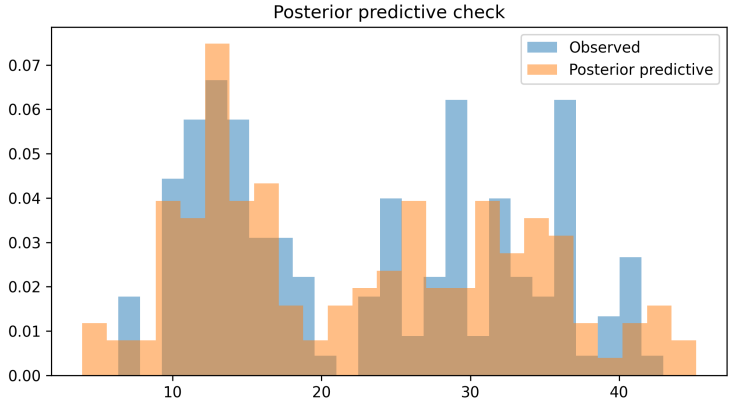


Figure 9: Posterior predictive check for the Energy Efficiency dataset.

Quantitative results are reported in Table 6. All three methods achieve competitive predictive accuracy, with relatively small differences in MSE. However, uncertainty-related metrics reveal clearer differences. HMC and NUTS attain coverage probabilities

closest to the nominal 95% level and produce slightly wider predictive intervals. VI achieves substantially shorter runtime and competitive MSE, but its narrower intervals lead to mild undercoverage.

Table 6: Performance comparison for the Energy Efficiency dataset.

Method	MSE	Coverage (95%)	Avg. Interval Width	Runtime (s)
HMC	1.284	0.94	6.82	742.15
NUTS	1.217	0.95	7.04	891.37
VI	1.336	0.92	6.11	118.54

Overall, the real-world experiment confirms that the main differences among inference strategies arise primarily in uncertainty calibration and computational cost rather than point prediction accuracy. HMC and NUTS provide more reliable predictive coverage, whereas VI offers a faster approximation with mild undercoverage.

Discussion and Conclusion

The results of this study highlight that the primary advantage of Bayesian neural networks in nonlinear regression lies not necessarily in improved point prediction accuracy, but rather in their ability to provide calibrated and interpretable predictive uncertainty under increasingly complex modeling conditions. Across the simulation studies and the real-world regression experiment, HMC, NUTS and variational inference generally achieved comparable predictive mean performance, whereas substantially larger differences emerged in uncertainty quantification and computational efficiency. In relatively well-specified settings with sufficient data, all inference strategies produced similar predictive behavior and near-nominal uncertainty coverage. However, as the data-generating process became more challenging through heavy-tailed noise, heteroscedasticity, limited sample size, and increased model complexity, the differences among inference methods became more pronounced. Sampling-based approaches such as HMC and NUTS consistently produced more reliable posterior uncertainty estimates and better predictive calibration, although at substantially higher computational cost. Variational inference provided a computationally efficient alternative and remained competitive in terms of predictive accuracy across all considered settings. In many experiments, VI achieved uncertainty estimates reasonably close to those obtained by sampling-based inference while requiring substantially shorter runtime. Nevertheless, the results also confirmed the well-known tendency of mean-field variational approximations to underestimate posterior variance, leading to mildly overconfident predictive intervals in more complex scenarios.

The heteroscedastic regression experiment further emphasized the importance of explicitly modeling predictive uncertainty in nonlinear regression problems. While deterministic neural networks and Bayesian neural networks produced similar predictive means, the Bayesian approach more effectively adapted predictive intervals to

input-dependent variability. This finding illustrates that reliable uncertainty quantification may become substantially more important than marginal improvements in point prediction, particularly in applications involving risk-sensitive decision-making.

The real-world Energy Efficiency experiment supported the conclusions obtained from the simulation studies. In practical nonlinear regression settings, all methods achieved satisfactory predictive performance, but differences in posterior calibration and computational scalability remained evident. These findings suggest that the practical choice of Bayesian inference strategy should depend on the required balance between computational efficiency and posterior fidelity.

The practical implications of these results are especially relevant in high-stakes domains such as medicine, finance, engineering reliability, and safety-critical systems, where poorly calibrated uncertainty may lead to overconfident and potentially harmful decisions. In such settings, sampling-based methods remain preferable when highly reliable posterior characterization is essential. In contrast, variational inference may provide an attractive compromise in larger-scale applications where computational scalability is a primary concern and approximate uncertainty estimation is acceptable.

Several limitations should also be acknowledged. First, although the revised experiments include larger neural-network architectures and a real-world dataset, the considered models remain moderate in scale relative to modern deep learning systems. Second, the variational analysis was restricted to a mean-field Gaussian approximation, which may not adequately capture complex posterior dependence structures or multimodality. More expressive variational families, including structured approximations and normalizing-flow-based approaches, may substantially improve uncertainty calibration. Finally, the present work focused primarily on regression settings, and future research may extend the analysis to classification problems, high-dimensional deep architectures, and hybrid inference strategies combining variational initialization with sampling-based posterior refinement.

References

Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.

Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American Statistical Association. 2017;112(518):859–877. https://doi.org/10.1080/01621459.2017.1285773.

Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D.: Weight Uncertainty in Neural Networks; 2015. https://arxiv.org/abs/1505.05424.

Breiman L. Random forests. Machine Learning. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Gal Y, Ghahramani Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning; 2016. https://arxiv.org/abs/1506.02142.

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep learning. Machine Learning. 2023;112:745–811. https://doi.org/10.1007/s10994-021-05946-3.

Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521(7553):452–459. https://doi.org/10.1038/nature14541.

Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

Hastie T, Tibshirani R. Generalized Additive Models. Chapman and Hall/CRC; 1990.

Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research. 2014;15(1):1593–1623. https://doi.org/10.5555/2627435.2638586.

Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Networks. 1991;4(2):251–257. https://doi.org/10.1016/0893-6080(91)90009-T.

Kendall A, Gal Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?; 2017. https://arxiv.org/abs/1703.04977.

Kingma DP, Ba J.: Adam: A Method for Stochastic Optimization; 2014. https://arxiv.org/abs/1412.6980.

McCullagh P, Nelder JA. Generalized Linear Models. 2 ed. Chapman and Hall; 1989.

Neal RM. Bayesian Learning for Neural Networks. Lecture Notes in Statistics, Springer; 1996.

Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.

Wang H, Yeung DY. Towards Bayesian deep learning: A framework and some existing methods. IEEE Transactions on Knowledge and Data Engineering. 2016;28(12):3395–3408. https://doi.org/10.1109/TKDE.2016.2606428.

Zhang C, Butepage J, Kjellstrom H, Mandt S. Advances in variational inference. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;41(8):2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774.