Variable Selection with Rigorous Uncertainty Quantification using Deep Bayesian Neural Networks: Posterior Concentration and Bernstein-von Mises Phenomenon

Jeremiah Zhe Liu
jereliu@google.com
Google Research & Harvard University*

Abstract

This work develops a theoretical basis for a deep Bayesian neural network (BNN)'s ability in performing high-dimensional variable selection with rigorous uncertainty quantification. We develop new Bayesian non-parametric theorems to show that a properly configured deep BNN (1) learns the variable importance effectively in high dimensions, and its learning rate can sometimes "break" the curse of dimensionality. (2) BNN's uncertainty quantification for variable importance is rigorous, in the sense that its 95% credible intervals for variable importance indeed covers the truth 95% of the time (i.e. the Bernstein-von Mises (BvM) phenomenon). The theoretical results suggest a simple variable selection algorithm based on the BNN's credible intervals. Extensive simulation confirms the theoretical findings and shows that the proposed algorithm outperforms existing classic and neural-network-based variable selection methods, particularly in high dimensions.

1 Introduction

The advent of the modern data era has given rise to voluminous, high-dimensional data in which the outcome has complex, nonlinear dependencies on input features. In this nonlinear, high-dimensional regime, a fundamental objective is *variable selection*, which refers to the identification of a small subset of features that is relevant in explaining variation in the outcome. However, high dimensionality brings two challenges to variable selection. The first is the *curse of dimensionality*, or the exponentially increasing difficulty in learning the variable importance parameters as the dimension of the input features increases. The second is the impact of *multiple comparisons*, which makes construction of a high dimensional variable-selection decision rule that maintains an appropriate false discovery rate difficult. For example, consider selecting among 100 variables using a univariate variable-selection procedure that has average precision, defined as 1 - false discovery rate (FDR), of 0.95 for selection of a single variable. Then the probability of selecting at least one irrelevant variable out of the 100 is $1 - 0.95^{100} \approx 0.994$ (assuming independence among decisions), leading to a sub-optimal procedure with precision less than 0.006 (Benjamini and Hochberg, 1995). The multiple comparison problem arises when a multivariate variable-selection decision is made based purely on individual decision rules, ignoring the dependency structure among the decisions across variables. This issue arises in a wide variety of application areas, such as genome-wide association studies and portfolio selection, among others (Bhlmann, 2013).

The objective of this work is to develop both theoretical and empirical understanding of the ability of a deep Bayesian neural network (BNN) model in tackling both of these challenges. A deep neural network is known to be an effective model for high-dimensional learning problems, illustrating empirical success in image classification and speech recognition applications. Bayesian inference in neural networks provides a principled framework for uncertainty quantification that naturally handles the multiple comparison problem (Gelman et al., 2012). By sampling from the joint posterior distribution of the variable importance parameters, a deep BNN's posterior distribution provides a complete picture of the dependency structure among the variable importance estimates for all input variables, allowing a variable selection procedure to tailor its decision rule with respect to the correlation structure of the problem.

Specifically, we consider a simple variable selection method for high-dimensional regression based on credible intervals of a deep BNN model. Consistent with

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

the existing nonlinear variable selection literature, we measure the global importance of an input variable x_p using the empirical norm of its gradient function $\psi_p(f) = \|\frac{\partial}{\partial x_p} f\|_n^2 = \frac{1}{n}\sum_{i=1}^n |\frac{\partial}{\partial x_p} f(\mathbf{x}_i)|^2$, where f is the regression function and $p \in \{1, \dots, P\}$ (White and Racine, 2001; Rosasco et al., 2013; Yang et al., 2016; He et al., 2018). We perform variable selection by first computing the $(1-\alpha)$-level simultaneous credible interval for the joint posterior distribution $\psi(f) = \{\psi_p(f)\}_{p=1}^P$, and make variable-selection decisions by inspecting whether the credible interval includes 0 for a given input. Clearly, the validity and effectiveness of this approach hinges critically on a deep BNN's ability to accurately learn and quantify uncertainty about variable importance in high dimensions. Unfortunately, neither property of a deep BNN model is well understood in the literature.

Summary of Contributions In this work, we establish new Bayesian nonparametric theorems for deep BNNs to investigate their ability in learning and quantifying uncertainty of variable importance measures derived from the model. We ask two key questions: (1) *learning accuracy*: does a deep BNN's good performance in prediction (i.e. in learning the true function f_0) translate to its ability to learn the variable importance $\psi_p(f_0)$? (2) *uncertainty quantification*: does a deep BNN properly quantify uncertainty about variable importance, such that a 95% credible interval for variable importance $\psi_p(f)$ covers the "true" value $\psi_p(f_0)$ 95% of the time? Our results show that, for *learning accuracy*, a deep Bayesian neural network learns the variable importance at a rate that is at least as fast as that achieved when learning f_0 (Theorem 1). That is, good performance in prediction translates to good performance in learning variable importance. For *uncertainty quantification*, we establish a *Bernstein-von Mises (BvM) theorem* to show that the posterior distribution of $\psi_p(f)$ converges to a Gaussian distribution, and the $(1-\alpha)$-level credible interval obtained from this distribution covers the true variable importance $\psi_p(f_0)$ $(1-\alpha)$% of the time (Theorem 2 and 3). The BvM theorems establish a rigorous frequentist interpretation for a deep BNN's simultaneous credible intervals, and are essential in ensuring the validity of the credible-interval-based variable selection methods. To the authors' knowledge, this is the first semi-parametric BvM result for the standard deep Bayesian neural network model under the i.i.d. Gaussian prior, and therefore one of the first Bayesian non-parametric studies on the deep BNN' ability to achieve rigorous uncertainty quantification.

Related Work The existing variable selection methods for neural networks fall primarily under the frequentist paradigm (Anders and Korn, 1999; Castellano and Fanelli, 2000; Guyon and Elisseeff, 2003; May et al., 2011). These existing methods include penalized estimation / thresholding of the input weights (Feng and Simon, 2017; Lu et al., 2018; Scardapane et al., 2017), greedy elimination based on the perturbed objective function (LeCun et al., 1990; Ye and Sun, 2018), and re-sampling based hypothesis tests (Giordano et al., 2014; La Rocca and Perna, 2005). For Bayesian inference, the recent work of Liang et al. (2018) proposed Spike-and-Slab priors on the input weights and performing variable selection based on the posterior inclusion probabilities for each variable. Rigorous uncertainty quantification based on these approaches can be difficult, due to either the non-identifiability of the neural network weights, the heavy computation burden of the re-sampling procedure, or the difficulty in developing BvM theorems for the neural network model.

The literature on the theoretical properties of a BNN model (e.g., posterior concentration and Bernstein von-Mises phenonmenon) is relatively sparse. Among known results, Lee (2000) established the posterior consistency of a one-layer BNN for learning continous or square-integrable functions. Rockova and Polson (2018) generalized this result to deep architectures and to more general function spaces (i.e., the β-Hölder space), and Chrief-Abdellatif (2020) generalized it further to the variational posterior that is obtained through optimization. In terms of uncertainty quantification, concurrent with this work, Wang and Rockova (2020) studies the asymptotic normality for the scalar-valued functionals of a special class of deep BNN under the spike-and-slab priors. In contrast, this work considers a more general class of deep BNN with no explicit sparse-inducing constraints. Furthermore, it develops a *multivariate* BvM theorem for the simultaneous credible interval of a vector-valued quadratic functional $\psi(f) = \{\|\frac{\partial}{\partial x_p}(f)\|_n^2\}_{p=1}^P$, and conducts thorough simulation to understand the functional's empirical behavior under practical scenarios.

2 Background

Nonparametric Regression For data $\{y_i, \mathbf{x}_i\}_{i=1}^n$ where $y \in \mathbb{R}$ and $\mathbf{x} \in [0,1]^P$ is a $P \times 1$ vector of covariates, we consider the standard nonparametric regression setting where $y_i = f^*(\mathbf{x}_i) + e_i$, for $e_i \sim N(0, s^2)$ with known s. The data dimension P is allowed to be large but assumed to be $o(1)$. That is, the dimension does not increase with the sample size n. The data-generation function f^* is an unknown continuous function belonging to certain function class $\mathcal{F}^*$. Recent theoretical work suggests that the model space of a properly configured deep neural network $\mathcal{F}(L, K, B)$ (defined below) achieves excellent approximation performance for a wide variety of function classes (Yarotsky, 2017; Schmidt-Hieber, 2017; Montanelli and Du, 2019; Suzuki, 2019; Gribonval et al., 2020). Therefore in this work, we focus our analysis on the BNN's behavior in learning the optimal $f_0 \in \mathcal{F}(L, K, B)$, making an assumption throughout that the BNN model is properly configured such that $f_0 \in \mathcal{F}$ is either identical to f^* or is sufficiently close to f^* for practical purposes.

Model Space of a Bayesian Neural Network Denote σ as the Rectified Linear Unit (ReLU) activation function. The class of deep ReLU neural networks with depth L and width K can be written as $f(\mathbf{x}) = b_0 + \beta^\top\big[\sigma\mathcal{W}_L\big(\sigma\mathcal{W}_{L-1}\dots\big(\sigma\mathcal{W}_2(\sigma\mathcal{W}_1\mathbf{x})\big)\big)\big]$. Following existing work in deep learning theory, we assume that the hidden weights $\mathcal{W}$ satisfy the norm constraint $\mathcal{C}^B_\infty$ in the sense that: $\mathcal{C}^B_\infty = \big\{\mathcal{W}\big|\max_l ||\mathcal{W}_l||_\infty \leq B,\ B \leq 1\big\}$ (Schmidt-Hieber, 2017; Suzuki, 2019). As a result, we denote the class of ReLU neural networks with depth L, width K and norm constraint B as $\mathcal{F}(L, K, B)$:

$$\mathcal{F}(L,K,B) = \Big\{f(\mathbf{x}) = b_0 + \beta^\top\big[\circ_{l=1}^{L}(\sigma\mathcal{W}_l)\circ x\big]\Big|\mathcal{W} \in \mathcal{C}^B_\infty\Big\},$$

and for notational simplicity we write $\mathcal{F}(L, K, B)$ as $\mathcal{F}$ when it is clear from the context. The Bayesian approach to neural network learning specifies a prior distribution $\Pi(f)$ that assigns probability to every candidate f in the model space $\mathcal{F}(L, K, B)$. The prior distribution $\Pi(f)$ is commonly specified implicitly through its model weights $\mathcal{W}$, such that the posterior distribution is $\Pi(f|\{y, \mathbf{x}\}) \propto \int \Pi(y|\mathbf{x}, f, \mathcal{W})\Pi(\mathcal{W})d\mathcal{W}$. Common choices for $\Pi(\mathcal{W})$ include Gaussian (Neal, 1996), Spike and Slab (Rockova and Polson, 2018), and Horseshoe priors (Ghosh et al., 2019; Louizos et al., 2018).

Rate of Posterior Concentration The quality of a Bayesian learning procedure is commonly measured by the learning rate of its posterior distribution, as defined by the speed at which the posterior distribution $\Pi_n = \Pi(.|\{y_i, \mathbf{x}_i\}_{i=1}^n)$ shrinks around the truth as $n \to \infty$. Such speed is usually assessed by the radius of a small ball surrounding f_0 that contains the majority of the posterior probability mass. Specifically, we consider the size of a set $A_n = \{f|||f - f_0||_n \leq M\epsilon_n\}$ such that $\Pi_n(A_n) \to 1$. Here, the *concentration rate* ϵ_n describes how fast this small ball A_n concentrates toward f_0 as the sample size increases. We state this notion of posterior concentration formally below (Ghosal and van der Vaart, 2007):

Definition 1 (Posterior Concentration). *For $f^* : \mathbb{R}^P \to \mathbb{R}$ where $P = o(1)$, let $\mathcal{F}(L, K, S, B)$ denote a class of ReLU network with depth L, width K, and norm bound B. Also denote f_0 as the Kullback-Leibler (KL)-projection of f^* to $\mathcal{F}(L, K, B)$, and E_0 the expectation with respect to true data-generation distribution $P_0 = N(f^*, \sigma^2)$. Then we say the posterior distribution f concentrates around f_0 at the rate ϵ_n in P_0^n probability if there exists an $\epsilon_n \to 0$ such that for any $M_n \to \infty$:*

$$E_0\Pi(f : ||f - f_0||_n^2 > M_n\epsilon_n|\{y_i, \mathbf{x}_i\}_{i=1}^n) \to 0 \qquad (1)$$

"Break" the Curse of Dimensionality Clearly, a Bayesian learning procedure with good finite-sample performance should have an ϵ_n that converges quickly to zero. In general, the learning rate ϵ_n depends on the dimension of the input feature P, and the geometry of the "true" function space $f^* \in \mathcal{F}^*$. Under the typical nonparametric learning scenario where $\mathcal{F}^*$ is the space of β-Hölder smooth (i.e., β-times differentiable) functions, the concentration rate ϵ_n is found to be $\epsilon_n = O\big(n^{-2\beta/(2\beta+P)} * (log\,n)^\gamma\big)$ for some $\gamma > 1$(Rockova and Polson, 2018). This exponential dependency of ϵ_n on the dimensionality P is referred to as the ***curse of dimensionality***, which implies that the sample complexity of a neural network explodes exponentially as the data dimension P increases (Bach, 2017). However, recent advances in frequentist learning theory shows that when f^* is sufficiently structured, a neural network model can in fact "break the curse" by adapting to the underlying structure of the data and achieve a learning rate that has no exponential dependency on P (Bach, 2017; Bauer and Kohler, 2019; Suzuki, 2019). To this end, we show that this also holds for Bayesian neural networks in well-specified scenarios, i.e., when $f^* = f_0 \in \mathcal{F}$ such that the target function lies in the model space of the neural network (Proposition 1). We also conduct simulation to study the model behavior under misspecification.

Measure of Variable Importance $\psi_p(f)$. For a smooth function $f : \mathbb{R}^P \to \mathbb{R}$, the *local importance* of a variable x_p with respect to the outcome $y = f(\mathbf{x})$ at a location $\mathbf{x} \in \mathcal{X}$ is captured by the magnitude of the *weak*[1] partial derivative $\big|\frac{\partial}{\partial x_p}f(\mathbf{x})\big|^2$ (He et al., 2018; Rosasco et al., 2013; Wahba, 1990; Adams and Fournier, 2003). Therefore, a natural measure for the *global importance* of a variable x_p is the integrated gradient norm over the entire feature space $\mathbf{x} \in \mathcal{X}$: $\Psi_p(f) = \big\|\frac{\partial}{\partial x_p}f\big\|_2^2 = \int_{\mathbf{x}\in\mathcal{X}}\big|\frac{\partial}{\partial x_p}f(\mathbf{x})\big|^2 dP(\mathbf{x})$. Given observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$, $\Psi_p(f)$ is approximated as:

$$\psi_p(f) = \big\|\frac{\partial}{\partial x_p}f\big\|_n^2 = \frac{1}{n}\sum_{i=1}^n\big|\frac{\partial}{\partial x_p}f(\mathbf{x}_i)\big|^2. \qquad (2)$$

[1] The notion of *weak* derivative is a mathematical necessity to ensure $\frac{\partial}{\partial x_p}f$ is well-defined, since f involves the ReLU function which is piece-wise linear and not differentiable at 0. However in practice, $\frac{\partial}{\partial x_p}f$ can be computed just as a regular derivative function, since it rarely happens that the pre-activation function is exactly 0.

In practice, $\frac{\partial}{\partial x_p} f(\mathbf{x})$ can be computed easily using standard automatic differentiation tools (Abadi et al., 2016).

3 Learning Variable Importance with Theoretical Guarantee

Throughout this theoretical development, we assume the true function f_0 has bounded norm $||f_0||_\infty \leq C$, so that the risk minimization problem is well-defined. We also put a weak requirement on the neural network's effective capacity so that the total stochasticity in the neural network prior is manageable:

Assumption 1 (Model Size). *The width of the ReLU network model $\mathcal{F}(L, K, B)$ grows slower than $O(\sqrt{n})$, i.e. $K = o(\sqrt{n})$.*

Assumption 1 ensures that the posterior estimate for $\psi_p(f)$ is stable in finite samples so that it converges sufficiently quickly toward the truth, which is a essential condition for the BvM theorem to hold. It also grounds our theoretical analysis to finite-width networks that's used in practice, and makes our result complementary to the recent theoretical literature on Gaussian-process-based analysis of infinite-width neural networks (Jacot et al., 2018; Arora et al., 2019; Du et al., 2019; Lee et al., 2019). Assumption 1 is satisfied by most of the popular architectures in practice. For example, in the ImageNet challenge where there are 1.4×10^7 images, most of the successful architectures, which include AlexNet, VGGNet, ResNet-152 and Inception-v3, have $K = O(10^3)$ nodes in the output layer (Russakovsky et al., 2015; Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; Szegedy et al., 2015; He et al., 2016). Neural networks with fixed architecture also satisfy this requirement, since the growth rate $o(1)$ for these models is also not faster than $\sqrt{n}$.

3.1 Rate of Posterior Concentration

We first investigate a Bayesian ReLU network's ability to accurately learn the variable importance $\Psi_p(f_0) = ||\frac{\partial}{\partial x_p}(f_0)||_2^2$ in finite samples. We show that, for a ReLU network that is able to learn the true function f_0 with rate ϵ_n (in the sense of Definition 1), the *entire* posterior distribution for variable importance $\psi_p(f)$ converges consistently to a point mass at the true $\Psi(f_0)$, at speed not slower than ϵ_n.

Theorem 1 (Rate of Posterior Concentration for ψ_p)**.** *For $f \in \mathcal{F}(L, K, B)$, assuming the posterior distribution $\Pi_n(f)$ concentrates around f_0 with rate ϵ_n, the posterior distribution for $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2$ contracts toward $\Psi_p(f_0) = ||\frac{\partial}{\partial x_p} f_0||_2^2$ at a rate not slower than ϵ_n. That is, for any $M_n \to \infty$*

$$E_0 \Pi_n \Big(\sup_{p \in \{1,\dots,P\}} |\psi_p(f) - \Psi_p(f_0)| > M_n \epsilon_n \Big) \to 0,$$

where $\Pi_n(.) = \Pi(.|\{y_i, \mathbf{x}_i\}_{i=1}^n)$ denotes the posterior distribution.

A proof for this theorem is in Supplementary Section B.1. Theorem 1 confirms two important facts. First, despite the non-identifiablity of the network weights $\mathcal{W}$, a deep BNN can reliably recover the variable importance of the true function $\Psi(f_0)$. Second, a deep BNN learns the variable importance at least as fast as the rate for learning the prediction function f_0. In other words, *good performance in prediction translates to good performance in learning variable importance.* We validate this conclusion in the experiment (Section 4), and show that, interestingly, the learning speed for $\Psi_p(f_0)$ can in fact be much faster than that for learning f_0. Given the empirical success of deep ReLU networks in high-dimensional prediction, Theorem 1 suggests that a ReLU network is an effective tool for learning variable importances in high dimension.

Comment: Possibility for Breaking the Curse of Dimensionality Given the statement of Theorem 1, it is interesting to ask exactly how fast ϵ_n of a BNN can go to zero under various data scenarios. Although it is not the focus of this work to study the convergence rate of a BNN's prediction function, we would like to highlight a possibility result that when learning $f_0 \in \mathcal{F}$, a Bayesian ReLU network with a standard Gaussian prior can in fact "break" the curse of dimensionality and achieve a parametric learning rate of $O(n^{-1/2})$ up to an logarithm factor.

Proposition 1 (Posterior Concentration for $f_0 \in \mathcal{F}$)**.** *For the space of ReLU network $\mathcal{F} = \mathcal{F}(L, K, B)$, assuming*

- *the model architecture satisfies:*

  $$L = O\big(log(N)\big), \quad K = O\big(N\big), \quad S = O\big(N\,log(N)\big),$$

  where $S = \sum_{l=1}^{L} ||\mathcal{W}_l||_0$ is the number of non-zero parameters in the model, and $N \in \mathbb{N}$ is a function of sample size n such that $log(N) \geq \sqrt{log(n)}$.

- *the prior distribution $\Pi(\mathcal{W})$ is an independent and identically distributed (i.i.d.) product of Gaussian distributions,*

then, for $f_0 \in \mathcal{F}$, the posterior distribution $\Pi_n(f) = \Pi(f|\{\mathbf{x}_i, y_i\}_{i=1}^n)$ contracts toward f_0 at a rate of at least $\epsilon_n = O\big((N/n) * log(N)^3\big)$. In particular, if $N = o(\sqrt{n})$ (i.e. Assumption 1), the learning rate is $\epsilon_n = O\big(n^{-1/2} * log(n)^3\big)$.

This result appears to be new to the BNN literature, and we give a full proof in Supplementary Section E. In combination with Theorem 1, this result suggests that when BNN is properly specified for the data (e.g., f^* is a linear function or a complex function with discrete inputs), high-dimension variable selection under a BNN can be rather effective.

The role of modern architecture It is worth noticing that the speed of posterior convergence depends critically on the statistical efficiency of the model architecture. Specifically, Proposition 1 assumes the network's ability in imposing a moderate level of structural sparsity onto its parameters (i.e., the assumption $S = O\big(N\,log(N)\big)$). This is an requirement that is satisfied by the modern architectures such as the Xception and EfficientNet, which uses depthwise separable convolutional layers to achieve state-of-the-art performance with a small parameter count (Chollet, 2017; Tan and Le, 2019). For example, by using depthwise separable convolutional layers, the convolution kernels of EfficientNet are in fact banded Toeplitz matrices with the higher off-diagonal entries set to zero, whose number of parameters is only proportional to the number of the output channels K, making the parameter count for the full network to be roughly on the order of $S = O(L * K) = O\big(N\,log(N)\big)$ (Chollet, 2017; Schmidt-Hieber, 2020). For completeness, in Supplementary section E we also study the case where no sparsity is assumed (Proposition 2). We show that in this case, for the neural network model to achieve a optimal rate of $O(n^{-1/2})$, it in fact needs to be narrower (i.e., $K = O(n^{1/4})$ rather than $O(n^{1/2})$), hence restricting the space of true functions it can reliably approximate in the finite data.

3.2 Uncertainty Quantification

In this section, we show that the deep BNN's posterior distribution for variable importance exhibits the Bernstein-von Mises (BvM) phenomenon. That is, after proper re-centering, $\Pi_n\big(\psi_p(f)\big)$ converges toward a Gaussian distribution whose $(1-q)$-level credible intervals achieve the correct coverage for the true variable importance parameters, i.e., a 95% credible interval indeed covers the true parameter 95% of the time (Castillo and Nickl, 2013) . The BvM theorems provide a rigorous theoretical justification for the BNN's ability to quantify its uncertainty about the importance of input variables.

We first explain why the re-centering is necessary. Notice that under noisy observations, $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2$ is a quadratic statistic that is strictly positive even when $\psi_p(f_0) = 0$. Therefore, the credible interval of un-centered $\psi_p(f)$ will never cover the truth. To this end, it is essential to re-center ψ_p so that it is an unbiased estimate of $\psi(f_0)$:

$$\psi^c_p(f) = \psi_p(f) - \eta_n. \tag{3}$$

Here, $\eta_n = o_p(\sqrt{n})$ is a de-biasing term that estimates the asymptotically vanishing bias $\psi_p(f_0) - E_0(\psi_p(f))$, whose expression we make explicit in the BvM Theorem below.

Theorem 2 (Bernstein-von Mises (BvM) for ψ^c_p)**.** *For $f \in \mathcal{F}(L, W, B)$, assume the posterior distribution $\Pi_n(f)$ contracts around f_0 at rate ϵ_n. Denote $D_p : f \to \frac{\partial}{\partial x_p} f$ to be the weak differentiation operator, and $H_p = D_p^\top D_p$ the corresponding inner product. For $\boldsymbol{\epsilon}$ the "true" noise such that $y = f_0 + \boldsymbol{\epsilon}$, define*

$$\hat{\psi}_p = ||D_p(f_0 + \boldsymbol{\epsilon})||^2_n = \psi_p(f_0) + 2\langle H_p f_0, \boldsymbol{\epsilon}\rangle_n + \langle H_p\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\rangle_n,$$

and its centered version as $\hat{\psi}^c_p = \hat{\psi}_p - \hat{\eta}_n$, where $\hat{\eta}_n = tr(\widehat{H}_p)/n$ for $\widehat{H}_p$ the empirical estimate of H_p. Then, the posterior distribution of the centered Bayesian estimator $\psi^c_p(f) = \psi_p(f) - \eta_n$ is asymptotically normal surrounding $\hat{\psi}^c_p$. That is,

$$\Pi\Big(\sqrt{n}(\psi^c_p(f) - \hat{\psi}^c_p)\Big|\{\mathbf{x}_i, y_i\}_{i=1}^n\Big) \rightsquigarrow N(0, 4||H_p f_0||^2_n).$$

The proof for this result is in Section C.4. Theorem 2 states that the credible intervals from posterior distribution $\Pi_n\big(\psi^c_p(f)\big)$ achieve the correct frequentist coverage in the sense that a 95% credible interval covers the truth 95% of the time. To see why this is the case, notice that a $(1-\alpha)$-level credible set $\hat{B}_n$ under posterior distribution Π_n satisfies $\Pi_n\big(\hat{B}_n\big) = 1 - \alpha$. Also, since $\Pi_n \to N(0, \sigma^2_{\texttt{BvM}})$, $\hat{B}_n$ also satisfies

$$\Pi_{N(0,1)}\big((\hat{B}_n - \hat{\psi}^c_p)/\sigma_{\texttt{BvM}}\big) \to 1 - \alpha \tag{4}$$

in probability for $\sigma^2_{\texttt{BvM}} = 4||H_p f_0||^2_n/n$, where $\Pi_{N(0,1)}$ is the standard Gaussian measure. In other words, the set $\hat{B}_n$ can be written in the form of $\hat{B}_n = \big[\hat{\psi}^c_p - \rho_\alpha\sigma_{\texttt{BvM}}, \hat{\psi}^c_p + \rho_\alpha\sigma_{\texttt{BvM}}\big]$, which matches the $(1-\alpha)$-level confidence intervals of an unbiased frequentist estimator $\hat{\psi}_p(f_0)$, which are known to achieve correct coverage for true parameters[2] (van der Vaart, 2000).

Handling the Issue of Multiple Comparison Notice that Theorem 2 provides justification only for the univariate confidence intervals $\Pi_n(\psi^c_p)$. To handle the issue of *multiple comparisons*, we must take into account the statistical dependencies between all $\{\psi^c_p(f)\}_{p=1}^P$. To this end, we extend Theorem 2 to the multivariate case to verify that the deep BNN's *simultaneous* credible intervals for all $\{\psi^c_p(f)\}_{p=1}^P$ also have the correct coverage.

[2] Here ρ_α is the $1-\frac{\alpha}{2}$ quantile function under a standard Gaussian distribution, e.g., $\rho_\alpha = 1.96$ for 95% credible interval.

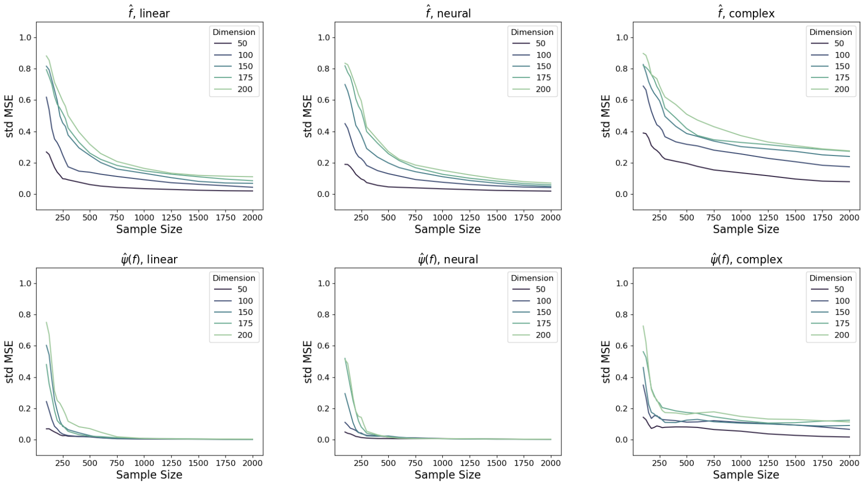

Figure 1: BNN's convergence behavior for learning prediction f^* (first row) and variable importance $\psi(f^*)$ (second row) under sample sizes $n \in (100, 2000)$ for $P \in (50, 200)$, measured by the standardized MSE (i.e. $1 - R^2$). Column 1-3 corresponds to **linear**, **neural**, and **complex**.

Theorem 3 (Multivariate Bernstein-von Mises (BvM) for ψ^c)**.** *For $f \in \mathcal{F}(L, W, B)$, assuming the posterior distribution $\Pi_n(f)$ contracts around f_0 at rate ϵ_n. Denote $\hat{\psi}^c = [\hat{\psi}^c_1, \dots, \hat{\psi}^c_P]$ for $\hat{\psi}^c_p$ as defined in Theorem 2. Also recall that $P = o(1)$, i.e. the data dimension does not grow with sample size.*

Then $\hat{\psi}^c$ is an unbiased estimator of $\psi(f_0) = [\psi_1(f_0), \dots, \psi_P(f_0)]$, and the posterior distribution for $\psi^c(f)$ asymptotically converge toward a multivariate normal distribution surrounding $\hat{\psi}^c$, i.e.

$$\Pi\Big(\sqrt{n}(\psi^c(f) - \hat{\psi}^c)\Big|\{\mathbf{x}_i, y_i\}_{i=1}^n\Big) \rightsquigarrow MVN(0, V_0),$$

where V_0 is a $P \times P$ matrix such that $(V_0)_{p_1,p_2} = 4\langle H_{p_1} f_0, H_{p_2} f_0 \rangle_n$.

Proof is in Supplementary Section C.5.

4 Experiment Analysis

4.1 Posterior Concentration and Uncertainty Quantification

We first empirically validate the two core theoretic results, posterior convergence and Bernstein-von Mises theorem, of this paper. In all the experiments described here, we use the standard i.i.d. Gaussian priors for model weights, so the model does not have an additional sparse-inducing mechanism beyond ReLU. We perform posterior inference using Hamiltonian Monte Carlo (HMC) with an adaptive step size scheme (Andrieu and Thoms, 2008) on Core i7 CPU with 64G memory and GeForce GTX 1070 GPU.

Learning Accuracy and Convergence Rate We generate data under the Gaussian noise model $y \sim N(f^*, 1)$ for data-generation function f^* with true dimension $P^* = 5$. We vary the dimension of the data between $P \in (25, 200)$, and vary sample sizes $n \in (100, 2000)$. For the neural network model, we consider a 2-layer, 50-hidden-unit feed-forward architecture (i.e., $L = 2$ and $K = 50$) with standard i.i.d. Gaussian priors $N(0, \sigma^2 = 0.1)$ for model weights. We consider three types of data-generating f^*: (1) **linear**: a simple linear model $f^*(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$; (2) **neural**: a function $f^* \in \mathcal{F}(L, W, B)$, and (3) **complex**: a complicated, non-smooth multivariate function[3] that is outside the neural network model's approximation space $\mathcal{F}(L, W, B)$. This latter data-generating model violates the assumption that $f^* \in \mathcal{F}$ in Proposition 1. We repeat the simulation 50 times for each setting, and evaluate the neural network's performance in learning f and $\psi_p(f)$ using out-of-sample standardized mean squared error (MSE), as follows:

$$std\_MSE(f, f^*) = \frac{\frac{1}{n}\sum_{i=1}^n [f(\mathbf{x}_i) - f^*(\mathbf{x}_i)]^2}{\frac{1}{n}\sum_{i=1}^n [f^*(\mathbf{x}_i) - E(f^*(\mathbf{x}_i))]^2}.$$

This is essentially the $1 - R^2$ statistic in regression modeling whose value lies within $(0, 1)$. Use of this statistic allows us to directly compare model performance across different data settings. The std_MSE for $\psi(f) = \{\psi_p(f)\}_{p=1}^P$ is computed similarly by averaging over all $p \in \{1, \dots, P\}$.

Figure 1 summarizes the standardized MSEs for learning f^* and $\psi(f^*)$, where each column corresponds to a data-generation machanism (**linear**, **neural** and **complex**). The first row summarizes the model's convergence behavior in prediction (learning f^*). We see

[3] $f^*(\mathbf{x}) = \frac{sin(max(x_1, x_2)) + arctan(x_2)}{1 + x_1 + x_5} + sin(0.5\, x_3)\big(1 + exp(x_4 - 0.5\, x_3)\big) + x_3^2 + 2\, sin(x_4) + 4\, x_5$, which is non-continuous in terms of x_1, x_2 but infinitely differentiable in terms of x_3, x_4, x_5

Table 1: Summary of variable selection methods included in the empirical study.

Model / Metric	**Decision Rule**		
	Thresholding	Hypothesis Test	Knockoff
Linear Model - LASSO	Tibshirani (1996)	Barber and Cands (2015)	Lockhart et al. (2013)
Random Forrest - Impurity	Breiman (2001)	Cands et al. (2018)	Altmann et al. (2010)
	Group L_1 Thresholding	Spike-and-Slab Probability	Credible Interval
Neural Network - $\mathcal{W}_1$	Feng and Simon (2017)	Liang et al. (2018)	
Neural Network - $\psi^c(f)$			(this work)

that the model's learning speed deteriorates as the data dimension P increases. However, this impact of dimensionality appears to be much smaller in the **linear** and **neural** scenarios, which both satisfy $f^* \in \mathcal{F}$ (Proposition 1). Comparatively, on the second row, the model's learning speed for variable importance are upper bounded by, and in fact *much* faster than, the speed of learning f^*. This verifies our conclusion in Theorem 1 that a model's good behavior in prediction translates to good performance in learning variable importance. We also observe that when the assumption $f^* \in \mathcal{F}$ is violated (e.g. for **complex** f^* in Column 3), the posterior estimate of $\psi_p(f)$ still converges toward $\psi_p(f_0)$, although at a rate that is much slower and is more sensitive to the dimension P of the data.

Bernstein-von Mises Phenonmenon We evaluate the BNN model's convergence behavior toward the asymptotic posterior $N(0, \sigma^2_{\texttt{BvM}} = 4||H_p\, f_0||^2_n)$ using two metrics: (1) the standardized MSE for learning the standard deviation $\sigma_{\texttt{BvM}}$, which assesses whether the *spread* of the posterior distribution is correct. (2) The Cramér von Mises (CvM) statistic as defined as the empirical L_2 distance between the standardized posterior sample $\{\psi^c_{std,m}\}_{m=1}^M$ and a Gaussian distribution Φ. This latter statistic, $CvM(\psi^c_{std}) = \frac{1}{M}\sum_{m=1}^M \big[\mathbb{F}(\psi^c_{std,m}) - \Phi(\psi^c_{std,m})\big]^2$, assesses whether the *shape* of the posterior distribution is sufficiently symmetric and has a Gaussian tail. Notice that since the CvM is a quadratic statistic, it roughly follows a mixture of χ^2 distribution even if true variable importance $\psi(f)$ is zero. Therefore, we compare it against a null distribution of $CvM(\psi^c_{std})$ for which $\psi^c_{std,m}$ is sampled from a Gaussian distribution.

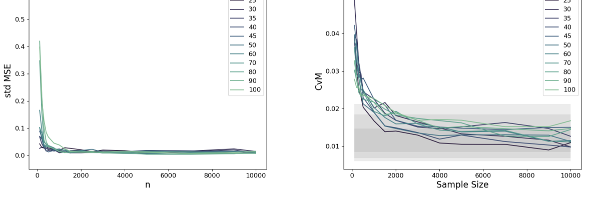

Figure 2: The variable importance posterior's convergence behavior toward the asymptotic standard deviation (left, measured by standardized MSE) and toward normality (right, measured by the CvM distance from a Gaussian distribution) under sample size $n \in (100, 10000)$ and $P \in (25, 100)$. Shaded region in the right figure indicates the $\{5\%, 10\%, 25\%, 75\%, 90\%, 95\%\}$ quantiles of the null CvM distribution.

Figure 2 summarizes the posterior distribution's convergence behavior in standard deviation (measured by std_MSE, top) and in normality (measured by CvM, bottom). The shaded region in the lower figure corresponds to the quantiles of a null CvM distribution. The figure shows that, as the sample size increases, the standardized MSE between $sd(\psi^c)$ and σ_{BvM} converges toward 0, and the CvM statistics enters into the range of the null distribution. The speed of convergence deteriorates as the dimension of the data increases, although not dramatically. These observations indicate that the credible intervals from the variable importance posterior $\Pi_n(\psi^c(f))$ indeed achieve the correct spread and shape in reasonably large samples, i.e. the Bernstein-von Mises phenomenon holds under the neural network model.

4.2 Effectiveness in High-dimensional Variable Selection

Finally, we study the effectiveness of the proposed variable selection approach (neural variable selection using credible intervals) by comparing it against nine existing methods based on various models (linear-LASSO, random forest, neural network) and decision rules (heuristic thresholding, hypothesis testing, Knockoff). We consider both low- and high-dimension situations ($d \in \{25, 75, 200\}$) and observe how the performance of each variable selection method changes as the sample size grows.

For the candidate variable selection methods, we notice that a variable selection method usually consists of three components: model, measure of variable importance, and the variable-selection decision rule. To this end, we consider nine methods that span three types of models and three types of decision rules (See Table 1 for a summary). The models we consider are (1) **LASSO**, the classic linear model $y = \sum_{p=1}^P x_p\beta_p$ with LASSO penalty on regression coefficients $\boldsymbol{\beta}$, whose variable importance is measured by the magnitude of β_p. (2) **RF**, the random forest model that measures variable importance using *impurity*, i.e., the decrease in regression error due to inclusion of a variable x_p (Breiman, 2001). (3) **NNet**, the (deep) neural networks that measure feature importance using either the magnitude of the input weights $\mathcal{W}_1$ or, in our case, the integrated gradient norm $\psi^c(f)$. For **LASSO** and **RF**, we consider three types of deci-

Table 2: F_1 score for classic and machine-learning based variable selection methods (summarized in Table 1) under low-dimension (d=25), moderate-dimension (d=75) and high-dimension data (d=200). Boldface indicates the best-performing decision rules in each dimension-model combination.

	Model	Rule	n=250	n=300	n=350	n=400	n=450	n=500
d=25	LASSO	thres	0.65 ± 0.11	0.64 ± 0.06	0.63 ± 0.08	0.76 ± 0.11	0.72 ± 0.09	0.73 ± 0.06
		knockoff	0.99 ± 0.02	0.99 ± 0.04	0.94 ± 0.09	0.98 ± 0.04	0.99 ± 0.03	0.99 ± 0.04
		test	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.89 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	RF	**thres**	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		knockoff	0.62 ± 0.48	1.00 ± 0.02	0.96 ± 0.16	0.90 ± 0.30	0.94 ± 0.19	0.99 ± 0.03
		test	0.91 ± 0.05	0.98 ± 0.05	1.00 ± 0.00	0.98 ± 0.05	0.98 ± 0.05	0.98 ± 0.05
	NNet	**Group L_1**	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		SpikeSlab	0.68 ± 0.05	0.68 ± 0.05	0.70 ± 0.06	0.69 ± 0.07	0.71 ± 0.08	0.72 ± 0.13
		CI (ours)	0.90 ± 0.04	0.97 ± 0.05	0.98 ± 0.04	0.97 ± 0.05	0.93 ± 0.06	1.00 ± 0.00
			n=250	n=300	n=350	n=400	n=450	n=500
d=75	LASSO	thres	0.32 ± 0.04	0.31 ± 0.03	0.31 ± 0.06	0.46 ± 0.11	0.56 ± 0.00	0.53 ± 0.11
		knockoff	0.93 ± 0.14	0.90 ± 0.14	0.89 ± 0.15	0.94 ± 0.08	0.94 ± 0.11	0.98 ± 0.04
		test	0.75 ± 0.03	0.83 ± 0.07	0.91 ± 0.00	0.66 ± 0.33	0.71 ± 0.00	0.89 ± 0.00
	RF	thres	0.66 ± 0.10	0.67 ± 0.06	0.72 ± 0.10	0.68 ± 0.06	0.80 ± 0.04	0.86 ± 0.04
		knockoff	0.79 ± 0.37	0.93 ± 0.14	0.93 ± 0.17	0.92 ± 0.18	0.95 ± 0.09	0.98 ± 0.05
		test	0.89 ± 0.12	0.93 ± 0.07	0.86 ± 0.04	0.88 ± 0.07	0.90 ± 0.09	0.95 ± 0.05
	NNet	Group L_1	0.77 ± 0.00	0.67 ± 0.27	0.68 ± 0.23	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00
		SpikeSlab	0.63 ± 0.09	0.66 ± 0.06	0.65 ± 0.08	0.65 ± 0.06	0.67 ± 0.07	0.68 ± 0.10
		CI (ours)	0.98 ± 0.04	0.97 ± 0.04	0.91 ± 0.07	0.97 ± 0.04	0.98 ± 0.05	1.00 ± 0.00
			n=250	n=300	n=350	n=400	n=450	n=500
d=200	LASSO	thres	0.29 ± 0.05	0.32 ± 0.01	0.28 ± 0.05	0.38 ± 0.10	0.42 ± 0.08	0.35 ± 0.06
		knockoff	0.31 ± 0.42	0.68 ± 0.38	0.88 ± 0.21	0.89 ± 0.11	0.90 ± 0.09	0.87 ± 0.18
		test	0.21 ± 0.04	0.25 ± 0.03	0.04 ± 0.00	0.49 ± 0.02	0.27 ± 0.13	0.61 ± 0.04
	RF	thres	0.37 ± 0.02	0.42 ± 0.01	0.43 ± 0.06	0.52 ± 0.02	0.54 ± 0.05	0.59 ± 0.05
		knockoff	0.12 ± 0.25	0.29 ± 0.39	0.38 ± 0.42	0.70 ± 0.42	0.80 ± 0.39	0.44 ± 0.49
		test	0.79 ± 0.10	0.81 ± 0.13	0.79 ± 0.07	0.87 ± 0.11	0.83 ± 0.09	0.70 ± 0.08
	NNet	Group L_1	0.67 ± 0.00	0.67 ± 0.00	0.67 ± 0.00	0.67 ± 0.00	0.67 ± 0.00	0.67 ± 0.00
		SpikeSlab	0.45 ± 0.26	0.53 ± 0.17	0.57 ± 0.14	0.60 ± 0.14	0.57 ± 0.12	0.57 ± 0.11
		CI (ours)	0.84 ± 0.10	0.76 ± 0.08	0.84 ± 0.08	0.93 ± 0.07	0.98 ± 0.04	0.92 ± 0.08

sion rule: (1) **Heuristic Thresholding**, which selects a variable by inspecting if the estimate of $\hat{\beta}_p$ is 0 or if the impurity for that variable is greater than 1% of the total impurity summed over all variables (Ye and Sun, 2018); (2) **Knockoff**, a nonparametric inference procedure that controls the FDR by constructing a data-adaptive threshold for variable importance (Cands et al., 2018), and (3) **Hypothesis Test**, which conducts either an asymptotic test on a LASSO-regularized $|\beta_p|$ estimate (Lockhart et al., 2013) or permutation-based test based on random forest impurity (Altmann et al., 2010), For both of these, we perform the standard Bonferroni correction. We select the **LASSO** hyper-parameters λ based on 10-fold cross validation, and use 500 regression trees for **RF**. For **NNet**, we also consider three decision rules: the frequentist approach with group-L_1 regularization on input weights $\mathcal{W}_1$ (Feng and Simon, 2017), a Bayesian approach with spike-and-slab priors on $\mathcal{W}_1$ (Liang et al., 2018), and our approach that is based on 95% posterior credible intervals of $\psi^c_p(f)$. Regarding the **NNet** architecture, we use $L = 1, W = 5$ for the LASSO- and Spike-and-slab-regularized networks as suggested by the original authors(Feng and Simon, 2017; Liang et al., 2018). We use $L = 1, W = 50$ for our approach since it is an architecture that is more common in practice.

We generate data by sampling the true function from the neural network model $f^* \in \mathcal{F}(L^* = 1, W^* = 5)$. Notice that this choice puts our method at a disadvantage compared to other **NNets** methods, since our network width $W = 50 > W^*$. We fix the number of data-generating covariates to be $d^* = 5$, and perform variable selection on input features $\mathbf{X}_{n\times P}$ with dimension $P \in \{25, 75, 200\}$ which corresponds to low-, moderate-, and high-dimensional situations. We vary sample size $n \in (250, 500)$. For each simulation setting (n, P), we repeat the experiment 20 times, and summarize each method's variable selection performance using the F_1 score, defined as the geometric mean of variable selection precision $prec = |\hat{S} \cap S|/|\hat{S}|$ and recall $recl = |\hat{S} \cap S|/|S|$ for S the set of data-generating variables and $\hat{S}$ the set of model-selected variables.

Table 2 summarizes the performance as quantified by the $F1$ score of the variable-selection methods in low-, medium- and high-dimension situations. In general, we observe that across all methods, **LASSO-knockoff**, **RF-test** and our proposed **NNet-CI** tend to have good performance, with **NNet-CI** being more effective in higher dimensions (d=200).

Our central conclusion is that **a powerful model alone is not sufficient to guarantee effective variable selection**. A good measure of variable importance, in terms of an unbiased and low-variance estimator of the true variable importance, and also a rigorous decision rule that has performance guarantee in terms of control over FDR or Type-I error are equally important. For example, although based on a neural network that closely matches the truth, **NNet-Group L_1** and **NNet-SpikeSlab** measures variable importance using the input weight $\widehat{\mathcal{W}}_1$, which is an unstable estimate of variable importance due to over-parametrization and/or non-identifiablity. As a result, the performance of these two models are worse than the linear-model based **LASSO-knockoff**. Comparing between the decision rules, the heuristic thresholding rules (**LASSO-thres** and **RF-thres**) are mostly not optimized for variable selection performance. As a result, they tend to be susceptible to the multiple comparison problem and their performance deteriorates quickly as the dimension increases. The Knockoff-based methods (**LASSO-knockoff** and **RF-knockoff**) are nonparametric procedures that are robust to model misspecification but tend to have weak power when the model variance is high. As a result, the Knockoff approach produced good results for the low-variance linear-LASSO model, but comparatively worse result for the more flexible but high-variance random forest model. Finally, the hypothesis tests / credible intervals are model-based procedures whose performance depends on the quality of the model. Hypothesis tests are expected to be more powerful when the model yields an unbiased and low-variance estimate of f^* (i.e. **RF-test** and **NNet-CI**), but has no performance guarantee when the model is misspecified (i.e. **LASSO**). In summary, we find that the **NNet-CI** method combines a powerful model that is effective in high dimension with a good variable-importance measure that has fast rate of convergence and also a credible-based selection rule that has a rigorous statistical guarantee. As a result, even without any sparse-inducing model regularization, **NNet-CI** out-performed its **NNet**-based peers, and is more powerful than other **LASSO**- or **RF**-based approaches in high dimensions.

5 Discussion and Future Directions

In this work, we investigate the theoretical basis underlying the deep BNN's ability to achieve rigorous uncertainty quantification in variable selection. Using the square integrated gradient $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2$ as the measure of variable importance, we established two new Bayesian nonparametric results on the BNN's ability to learn and quantify uncertainty about variable importance. Our results suggest that the neural network can learn variable importance effectively in high dimensions (Theorem 1), in a speed that in some cases "breaks" the curse of dimensionality (Proposition 1). Moreover, it can generate rigorous and calibrated uncertainty estimates in the sense that its $(1-q)$-level credible intervals for variable importance cover the true parameter $(1-q)\%$ of the time (Theorem 2 and 3). The simulation experiments confirmed these theoretical findings, and revealed the interesting fact that BNN can learn variable importance $\psi_p(f)$ at a rate much faster than learning predictions for f^* (Figure 1). The comparative study illustrates the effectiveness of the proposed approach for the purpose of variable selection in high dimensions, which is a scenario where the existing methods experience difficulties due to model misspecification, the curse of dimensionality, or the issue of multiple comparison.

Discussion: learning variable importance under mis-specification The theoretical results developed in this work assumes a well-specified scenario where the model's prediction function f is guaranteed to converge toward the true function f^* as $n \rightarrow \infty$. However, it is important to ask if learning variable importance is still possible under different types of model mis-specification. To this end, we note that if the mis-specification is mild (e.g., f^* does not belong to $\mathcal{F}$ but to the β-Hölder space that contains $\mathcal{F}$), the recent literature suggests that the posterior concentration of the prediction function f and the variable importance $\psi_p(f)$ is still likely, although at a much slower rate (Schmidt-Hieber, 2017; Rockova and Polson, 2018). This observation is empirically validated by the experiment in Section 4, where the model's learning speed is indeed evidently slower under the mis-specified scenario (Figure 1). However, the situation becomes more complex when the misspecification is severe, where the posterior convergence of the prediction function f does not hold even under infinite data. In this case, the model's variable importance estimate $\psi_p(f)$ does not converge toward the truth $\psi_p(f^*)$ unless we impose additional assumption on the true function f^*. For example, if f^* is a generalized additive function $f = \sum_{p=1}^P h_p^*(x_p)$ with $D_p f^* = \frac{\partial}{\partial_p} h_p^*$, then the model can correctly learn the variable importance $\psi_p(f^*) = ||D_p f^*||_2^2 = ||\frac{\partial}{\partial_p} h_p^*||_2^2$ as long as it can correctly specify the marginal prediction function h_p^*. On the other hand, if f^* adopts a tensor product form $f^* = \prod_{p=1}^P h_p(x_p)$ with $D_p f^* = \frac{\partial}{\partial_p} h_p^* * \big[\prod_{p' \neq p} h_{p'}^*(x_{p'})\big]$, then posterior convergence is not likely without additional assumptions on $||\prod_{p' \neq p} h_{p'}^*(x_{p'})||_2^2$. We leave a full theoretical discussion of this topic for future work.

Future work Consistent with the classic Bayesian nonparametric and deep learning literature (Castillo and Rousseau, 2015; Rokov and Saha, 2019; Barron, 1993; Barron and Klusowski, 2018), this work assumes the noise distribution $\boldsymbol{\epsilon}$ is known. Furthermore, computing the exact credible intervals under a BNN model requires the use of Markov Chain Monte Carlo (MCMC) procedures, which can be infeasible for large datasets. Therefore two important future directions of this work are to investigate the BNN's ability to learn variable importance under distributional misspecification, and to identify posterior inference methods (e.g., particle filter (Dai et al., 2016) that scale to large datasets while also achieve rigorous uncertainty quantification.

Acknowledgements

I would like to thank Dr. Rajarshi Mukherjee and my doctoral advisor Dr. Brent Coull at Harvard Biostatistics for the helpful discussions and generous support. Especially to Rajarshi for going through the proof in detail and Brent for spending time helping me with manuscript editing during the holidays. This work was not possible without their help.

The first part of this work (Theorems 1 and 2) was done at Harvard University and supported by USEPA grant RD-83587201. Its contents are solely the responsibility of the grantee and does not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces, Volume 140*. Academic Press, Amsterdam, 2 edition edition, July 2003. ISBN 978-0-12-044143-3.

Andr Altmann, Laura Toloi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, May 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq134.

Ulrich Anders and Olaf Korn. Model selection in neural networks. *Neural Networks*, 12(2):309–323, March 1999. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00117-8.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18 (4):343–373, December 2008. ISSN 1573-1375. doi: 10.1007/s11222-008-9110-y.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alch-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8141–8150. Curran Associates, Inc., 2019.

Francis Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. ISSN 1533-7928.

Rina Foygel Barber and Emmanuel J. Cands. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, October 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1337.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. ISSN 0018-9448. doi: 10.1109/18.256500.

Andrew R. Barron and Jason M. Klusowski. Approximation and Estimation for High-Dimensional Deep Learning Networks. *arXiv:1809.03090 [cs, stat]*, September 2018. arXiv: 1809.03090.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47 (4):2261–2285, August 2019. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1747. Publisher: Institute of Mathematical Statistics.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300, 1995. ISSN 0035-9246.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Peter Bhlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, September 2013. ISSN 1350-7265. doi: 10.3150/12-BEJSP11. Publisher: Bernoulli Society for Mathematical Statistics and Probability.

Emmanuel Cands, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3), June 2018. doi: 10.1111/rssb.12265.

Giovanna Castellano and Anna Maria Fanelli. Variable selection using neural-network models. *Neurocomputing*, 31(1):1–13, March 2000. ISSN 0925-2312. doi: 10.1016/S0925-2312(99)00146-0.

Ismal Castillo and Richard Nickl. Nonparametric Bernsteinvon Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028, August 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1133.

Ismal Castillo and Judith Rousseau. A Bernsteinvon Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, December 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1336.

F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, July 2017. doi: 10.1109/CVPR.2017.195. ISSN: 1063-6919.

Badr-Eddine Chrief-Abdellatif. Convergence Rates of Variational Inference in Sparse Deep Learning. *Proceedings of the International Conference on Machine Learning*, 1, 2020.

Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable Bayesian Inference via Particle Mirror Descent. In *Artificial Intelligence and Statistics*, pages 985–994, May 2016.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, May 2019. ISSN: 2640-3498.

Jean Feng and Noah Simon. Sparse Input Neural Networks for High-dimensional Nonparametric Regression and Classification. *arXiv:1711.07592 [stat]*, November 2017. arXiv: 1711.07592.

Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, April 2012. ISSN 1934-5747. doi: 10.1080/19345747.2011.618213.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, February 2007. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053606000001172.

Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. Model Selection in Bayesian Neural Networks via Horseshoe Priors. *Journal of Machine Learning Research*, 20(182):1–46, 2019. ISSN 1533-7928.

Francesco Giordano, Michele La Rocca, and Cira Perna. Input Variable Selection in Neural Network Models. *Communications in Statistics - Theory and Methods*, 43(4):735–750, February 2014. ISSN 0361-0926. doi: 10.1080/03610926.2013.804567.

Rmi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks, July 2020.

Isabelle Guyon and Andr Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182, 2003. ISSN ISSN 1533-7928.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

Xin He, Junhui Wang, and Shaogao Lv. Scalable kernel-based variable selection with sparsistency. *arXiv:1802.09246 [cs, stat]*, February 2018. arXiv: 1802.09246.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

Michele La Rocca and Cira Perna. Variable selection in neural network regression models with dependent data: a subsampling approach. *Computational Statistics & Data Analysis*, 48(2):415–429, February 2005. ISSN 0167-9473. doi: 10.1016/j.csda.2004.01.004.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal Brain Damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 598–605. Morgan-Kaufmann, 1990.

Herbert Lee. Consistency of posterior distributions for neural networks. *Neural networks : the official journal of the International Neural Network Society*, 13:629–42, August 2000. doi: 10.1016/S0893-6080(00)00045-9.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alch-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8570–8581. Curran Associates, Inc., 2019.

Faming Liang, Qizhai Li, and Lei Zhou. Bayesian Neural Networks for Selection of Drug Sensitive Genes. *Journal of the American Statistical Association*, 113 (523):955–972, July 2018. ISSN 0162-1459. doi: 10.1080/01621459.2017.1409122.

Richard Lockhart, Jonathan Taylor, Ryan Tibshirani, and Robert Tibshirani. A significance test for the

LASSO. *The Annals of Statistics*, 42, January 2013. doi: 10.1214/13-AOS1175.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning Sparse Neural Networks through L_0 Regularization. In *International Conference on Learning Representations*, 2018.

Yang Lu, Yingying Fan, Jinchi Lv, and William Stafford Noble. DeepPINK: reproducible feature selection in deep neural networks. pages 8676–8686, 2018.

Robert May, Graeme Dandy, and Holger Maier. Review of Input Variable Selection Methods for Artificial Neural Networks. *Artificial Neural Networks - Methodological Advances and Biomedical Applications*, April 2011. doi: 10.5772/16004.

Hadrien Montanelli and Qiang Du. New Error Bounds for Deep ReLU Networks Using Sparse Grids. *SIAM Journal on Mathematics of Data Science*, 1(1):78–92, January 2019. doi: 10.1137/18M1189336. Publisher: Society for Industrial and Applied Mathematics.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, New York, 1996. ISBN 978-0-387-94724-2.

Veronika Rockova and Nicholas Polson. Posterior Concentration for Sparse Deep Learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 930–941. Curran Associates, Inc., 2018.

Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric Sparsity and Regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013.

Veronika Rokov and Enakshi Saha. On Theory for BART. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848. PMLR, April 2019. ISSN: 2640-3498.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252, December 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, June 2017. ISSN 0925-2312. doi: 10.1016/j.neucom.2017.02.029.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv:1708.06633 [cs, math, stat]*, August 2017. arXiv: 1708.06633.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, August 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1875. Publisher: Institute of Mathematical Statistics.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2015. doi: 10.1109/CVPR.2016.308.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, May 2019. ISSN: 2640-3498.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, June 2000. ISBN 978-0-521-78450-4.

Grace Wahba. *Spline Models for Observational Data*. SIAM, September 1990. ISBN 978-0-89871-244-5. Google-Books-ID: ScRQJEETs0EC.

Yuexi Wang and Veronika Rockova. Uncertainty Quantification for Sparse Deep Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 298–308. PMLR, June 2020. ISSN: 2640-3498.

H. White and J. Racine. Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks*, 12(4):657–673, July 2001. ISSN 1045-9227. doi: 10.1109/72.935080.

Lei Yang, Shaogao Lv, and Junhui Wang. Model-free Variable Selection in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, 17 (82):1–24, 2016. ISSN 1533-7928.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, October 2017. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.07.002.

Mao Ye and Yan Sun. Variable Selection via Penalized Neural Network: a Drop-Out-One Loss Approach. In *International Conference on Machine Learning*, pages 5620–5629, July 2018.