Variable Selection with Rigorous Uncertainty Quantification using Bayesian Deep Neural Networks

Jeremiah Zhe Liu*
zhl112@mail.harvard.edu
Harvard University

1 Introduction

The advent of modern data era has given rise to voluminous, high-dimensional datasets in which the outcome has complex, nonlinear dependency on the input features. In this nonlinear, high-dimensional regime, a fundamental task behind many knowledge discovery endeavors is variable selection, i.e., to identify a small subset of features that is the most relevant in explaining the outcome. However, The high-dimension regime brings two challenges: the first challenge is the *curse of dimensionality*, i.e. the exponentially increasing difficulty in estimating the variable importance parameters as the data dimension increases. The second challenge is *multiple comparison*, i.e., the difficulty in constructing a high dimensional decision rule that maintains the correct level of precision (e.g., 1 - false discovery rate). The multiple comparison problem often arise when a multivariate variable-selection decision is made based purely on individual decision rule, ignoring the dependency structure among the decisions across variables [7].

To this end, the main interest of this work is to establish Bayesian Deep Neural Network as an effective tool for tackling both of these challenges. 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, on the other hand, provides a principled framework for uncertainty quantification that naturally handles the multiple comparison problem [14]. By sampling from the joint posterior distribution of variable importance parameters, Bayesian inference provides easy access to not only the model uncertainty about the individual variables, but importantly, a complete picture of the dependency structure among all the input covariates. This information has allowed the variable selection procedures to tailor their decision rule with respect to the correlation structure of the problem, leading to a more precise and informative precision-recall trade-off in high dimension.

Specifically, we propose a simple variable selection method for high-dimensional regression based on credible intervals of a deep Bayesian Rectified Linear Unit (ReLU) network. Consistent with the existing nonlinear variable selection approaches, we measure the importance of an input variable x_p using the empirical norm of its gradient function $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2 = \frac{1}{n}\sum_{i=1}^n |\frac{\partial}{\partial x_p} f(\mathbf{x}_i)|^2$ [40, 30, 42, 19]. We perform variable selection by first computing the $(1-\alpha)$-level simutaneous credible intervals for the joint posterior distribution $\psi(f) = \{\psi_p(f)\}_{p=1}^d$, and then make variable-selection decision by inspecting whether the credible intervals includes 0. It is simple to implement, agnostic to model architecture and is shown to be more effective than a range of existing classic or neural-network-based methods especially in high dimension (see simulation experiment in Section C).

Clearly, the effectiveness of this approach hinges on ReLU networks' ability in learning and quantifying uncertainty about variable importance $\psi_p(f)$ in high dimension. Specifically, two important questions must be answered: (1) *learning accuracy*. i.e., does a ReLU networks' good performance in prediction (i.e. in learning f_0) translates to that in learning the variable importance? (2) *uncertainty quantification*. i.e., does ReLU network properly quantifies its uncertainty about variable importance, such that its 95% credible interval for $\psi_p(f)$ indeed covers the "true" variable importance $\psi_p(f_0)$ for 95% of the time? To this end, we develop a complete set of Bayesian nonparametric theorems results for the deep ReLU network to answer both questions in the affirmative. For learning accuracy, we show that a ReLU network learns the variable importance $\Psi_p(f_0)$ in a rate that is at least as fast

*Now also at Google Research.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

as its rate in learning f_0 (Theorem 1). For uncertainty quantification, we establish a *Bernstein-von Mises (BvM) theorem* to show that the posterior distribution of $\psi(f)$ converge quickly (i.e. in the parametric rate) toward a Gaussian distribution, and the $(1-\alpha)$-level credible interval of the posterior distribution covers the truth $(1-\alpha)\%$ of the time (Theorem 2 and 3). The Bernstein-von Mises (BvM) theorems establish a rigorous frequentist interpretation for a Bayesian ReLU network's credible intervals, and are essential in ensuring the validity of the credible-interval-based variable selection methods in controlling its operating characteristics such as the false discovery rate (FDR). To the authors' knowledge, this is the first semi-parametric BvM result for deep neural network models, and therefore the first Bayesian non-parametric study on the ability of Bayesian neural networks in achieving rigorous uncertainty quantification.

2 Problem Setup

Data For data $\{y_i, \mathbf{x}_i\}_{i=1}^n$ where $y \in \mathbb{R}$ and $\mathbf{x} \in \mathscr{X} = (0,1)^d$ a $d \times 1$ vector of covariates, we consider the nonparametric regression setting where $y_i = f^*(\mathbf{x}_i) + \varepsilon_i$ for $\varepsilon_i \sim N(0, \sigma^2)$. The data dimension d is allowed to be large but assumed to be $o(1)$, i.e. does not increase with the sample size n. Here f^* is an unknown continuous function belonging to certain function class $\mathscr{F}^*$. Recent theoretical development suggests that the model space of a (properly configured) deep ReLU network $\mathscr{F}(L, W, S, B)$ (defined below) achieves excellent approximation performance for a wide class of $f^* \in \mathscr{F}^*$ [43, 33, 24, 35, 17]. Therefore in this work, we focus our analysis on the Bayesian ReLU networks' behavior in learning the optimal $f_0 \in \mathscr{F}(L, W, S, B)$, making an assumption throughout that the ReLU model is properly configured such that $f_0 \in \mathscr{F}$ is either identical to f^* or is sufficiently close to f^* for practical purposes.

Model Denote σ the ReLU activation function. The class of deep ReLU neural networks with depth L and width K can be written as $f(\mathbf{x}) = b_0 + \beta^\top \left[\sigma \mathscr{W}_L \left(\sigma \mathscr{W}_{L-1} \dots \left(\sigma \mathscr{W}_2 (\sigma \mathscr{W}_1 \mathbf{x})\right)\right)\right]$. Following existing approaches in deep learning theory [33, 35], we assume the hidden weights $\mathscr{W}$ satisfy the sparsity constraint $\mathscr{C}_0^S$ and norm constraint $\mathscr{C}_\infty^B$ in the sense that: $\mathscr{C}_0^S = \left\{\mathscr{W} \,\middle|\, \sum_{l=1}^L ||\mathscr{W}_l||_0 \le S\right\}$, $\mathscr{C}_\infty^B = \left\{\mathscr{W} \,\middle|\, \max_l ||\mathscr{W}_l||_\infty \le B,\ B \le 1\right\}$. As a result, we denote the class of ReLU neural networks with depth L, width K, sparsity constraint S and norm constraint B as $\mathscr{F}(L, K, S, B)$:

$$\mathscr{F}(L, K, S, B) = \left\{ f(\mathbf{x}) = b_0 + \beta^\top \left[\circ_{l=1}^L (\sigma \mathscr{W}_l) \circ x \right] \,\middle|\, \mathscr{W} \in \mathscr{C}_0^S, \mathscr{W} \in \mathscr{C}_\infty^B \right\}, \tag{1}$$

and write $\mathscr{F} = \mathscr{F}(L, K, S, B)$ 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 $f \in \mathscr{F}(L, K, S, B)$ in the model space. The prior distribution $\Pi(f)$ is commonly specified through its model weights $\mathscr{W}$ such that the posterior distribution is $\Pi(f, \mathscr{W} | \{y, \mathbf{x}\}_{i=1}^n) \propto \Pi(y | \mathbf{x}, f, \mathscr{W}) \Pi(\mathscr{W})$. Common choices for $\Pi(\mathscr{W})$ include Gaussian [26], Spike and Slab [29], and Horseshoe [16, 23].

3 Learning Variable Importance with Theoretical Guarantee

Throughout the theoretical analysis, we assume for $y_i = f_0(\mathbf{x}_i) + \varepsilon_i$, the true function f_0 has bounded norm $||f_0||_\infty \le C_f$ so the risk minimization problem is well-defined. We also put a weak requirement on the neural network's effective capacity so the total stochasticity in the neural network prior is manageable:

Assumption 1 (Model Size)**.** *The width of the ReLU network model $\mathscr{F}(L, K, S, B)$ does not grow faster than $O(\sqrt{n})$, i.e. $K = o_p(\sqrt{n})$*

This assumption ensures that the posterior estimate for $\psi_p(f)$ is stable in finite sample and converges quickly toward the truth, which is a essential condition for the BvM theorem to hold. Assumption 1 is satisfied by most of the popular architectures in practice [32, 20, 34, 36, 18].

Posterior Consistency and Learning Rates We first investigate a Bayesian ReLU network's ability in accurately learning the variable importance $\Psi(f_0) = ||\frac{\partial}{\partial x_p}(f_0)||_2^2$ in finite sample. It states that, for a ReLU network that learns the true function f_0 in a rate ε_n (in the sense of Definition 1), its posterior distribution for variable importance $\psi_p(f)$ converge consistently to a point mass on the truth $\Psi(f_0)$, and in a speed that is not slower than ε_n.

Theorem 1 (Rate of Posterior Concentration for ψ_p)**.** *For $f \in \mathscr{F}(L, K, S, B)$, assuming the posterior distribution $\Pi_n(f)$ contracts around f_0 with rate ε_n, then the posterior distribution for $\psi_p(f) =$*

$||\frac{\partial}{\partial x_p} f||_n^2$ contracts toward $\Psi_p(f_0) = ||\frac{\partial}{\partial x_p} f||_2^2$ at a rate not slower than ε_n, i.e., for any $M_n \to \infty$

$$E_0 \Pi\Big(\sup_{p \in \{1,\dots,P\}} |\psi_p(f) - \Psi_p(f_0)| > M_n \varepsilon_n \Big| \{y_i, \mathbf{x}_i\}_{i=1}^n \Big) \to 0. \tag{2}$$

Proof is in Section D. Theorem 1 confirms two important facts. First, despite the non-identifiablity of the network weights $\mathscr{W}$, a ReLU network can reliably recover the variable importance of the true function $\Psi(f_0)$. Second, a ReLU network learns the variable importance in a speed that is no slower than learning the prediction function f_0, i.e. good performance in prediction translates to good performance in learning variable importance. We validate this conclusion in the experiment (Section 4), and show that the learning speed for $\psi_p(f)$ is in fact much faster than learning f_0. Given the empirical success of deep ReLU networks in high-dimensional prediction, Theorem 1 suggests that ReLU network is an effective tool for learning variable importance in high dimension.

Uncertainty Quantification We first establish a univariate semi-parametric BvM theorem for $\psi_p(f)$ under the deep ReLU network's posterior (Theorem 2), and then extend it to the the multivariate case for $\psi(f) = \{\psi_p(f)\}_{p=1}^P$ to handle the issue of multiple comparison (Theorem 3). Both theorems show that, after proper re-centering, the posterior distributions of the variable importance parameters converge in $O(1/\sqrt{n})$ rate to a (multivariate) Gaussian distribution. More importantly, the credible intervals of this Gaussian distribution achieve the correct frequentist coverage for the true parameter.

Theorem 2 (Bernstein-von Mises (BvM) for ψ_p^c)**.** *For $f \in \mathscr{F}(L, W, S, B)$, assuming the posterior distribution $\Pi_n(f)$ contracts around f_0 at rate ε_n. Denote $D_p : f \to \frac{\partial}{\partial x_p} f$ the differentiation operator, and $H_p = D_p^\top D_p$ the corresponding inner product. For ε the "true" noise such that $y = f_0 + \varepsilon$, define:*

$$\hat{\psi}_p = ||D_p(f_0 + \varepsilon)||_n^2 = \psi_p(f_0) + 2\langle H_p f_0, \varepsilon \rangle_n + \langle H_p \varepsilon, \varepsilon \rangle_n, \tag{3}$$

and its centered version as $\hat{\psi}_p^c = \hat{\psi}_p - \hat{\eta}_n$ where $\hat{\eta}_n = trace(H_p)/n$. Then $\hat{\psi}_p^c$ is an unbiased estimator of $\psi_p(f_0)$, and the posterior distribution for $\psi_p^c(f)$ is asymptotically normal surrounding $\hat{\psi}_p^c$, i.e.,

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

The proof is in Section E.4. Theorem 2 establishes a rigorous theoretical basis for using the ReLU network posterior Π_n to quantify model uncertainty about variable importance. It states that the credible intervals from posterior distribution $\Pi_n\big(\psi_p^c(f)\big)$ achieves correct frequentist coverage (i.e. a 95% credible interval covers the truth 95% of the time). To see why this is the case, notice (4) implies that any $(1-\alpha)$-level credible set $\hat{B}_n$ such that $\Pi_n\big(\psi_p^c(f) \in \hat{B}_n\big) = 1 - \alpha$ will satisfy

$$\Pi_{N(0,1)}\big((\hat{B}_n - \hat{\psi}_p^c)/\sigma_\psi\big) \to 1 - \alpha$$

in probability for $\sigma_\psi^2 = 4||H_p f_0||_n^2/n$, where we have denoted $\Pi_{N(0,1)}$ as the standard Gaussian measure. In other words, the set $\hat{B}_n$ can be written in the form of $\hat{B}_n = \big[\hat{\psi}_p^c - \rho_\alpha * \sigma_\psi, \hat{\psi}_p^c + \rho_\alpha * \sigma_\psi\big]$, which coincides with the $(1-\alpha)$-level confidence intervals of an unbiased and efficient frequentist estimator of $\psi_p(f_0)$, which are known to achieve correct coverage for true parameters [38]. As a result, similar to an efficient frequentist confidence interval, the neural network credible interval (**??**) achieves correct coverage for true parameter $\psi_p^c(f)$ in sufficiently large sample, justifying its ability in achieving rigorous uncertainty quantification.

Notice that Theorem 2 is an univariate result and provides justification only for the univariate confidence intervals. To handle the issue of *multiple comparison* in high dimension, we need to take the statistical dependencies between $\psi_p^c(f)$'s into account. In the Appendix Section B, we extend Theorem 2 to the multivariate case to also ensure the validity of the simutaneous covarage of the ReLU network's credible intervals.

4 Experiment Analysis

We first empirically validate the two core theoretic results of this paper (posterior convergence and Bernstein-von Mises theorem), and then present a comprehensive simulation study to compare the variable-selection effectiveness of the proposed approach against existing classic or neural-network based approaches in Appendix (Section C). To ensure the effectiveness in variable-selection coming strictly from the neural network's ability in uncertainty quantification, we use the standard

independent and identically distributed (i.i.d.) Gaussian priors for model weights, so the model does not have additional sparse-inducing mechanism beyond ReLU. We perform posterior inference using Hamiltonian Monte Carlo (HMC) with an adaptive step size scheme [2].

Learning Accuracy and Convergence Rate We generate data under the Gaussian noise model $y \sim N(f^*, \sigma^2 = 1)$ for data-generation function $f^* : [0,1]^{d^*} \to \mathbb{R}$ with true dimension $d^* = 5$. We vary sample sizes $n \in (100, 2000)$, and vary data dimension between $d \in (25, 200)$. For the neural network model, we consider a 2-layer, 50-hidden-unit feedforward architecture (i.e., $L = 2$ and $K = 50$). We consider three types of data-generating f^*: (1) **linear**, a simple linear model $f^*(\mathbf{x}) = \mathbf{x}^\top \beta$; (2) **neural**, a function $f^* \in \mathscr{F}(L, W, S, B)$, and (3) **complex**, a complex, non-smooth multivariate function that is outside the neural network model's approximation space $\mathscr{F}(L, W, S, B)$[2]. For each setting of (n, d, f^*), we repeat the simulation 20 times and evaluate the neural network's performance in learning f and $\psi_p(f)$ using out-of-sample standardized mean squared error (MSE) against f^* and $\psi(f^*)$:

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

This is essentially the $1 - R^2$ statistic in regression modeling whose value lies within $(0, 1)$, allowing us to directly compare model performance across different data settings. The *std_MSE* for $\psi(f)$ is computed similarly by averaging over all $p \in \{1, \ldots, d\}$.

Figure 1 summarizes the standardized MSEs for learning f^* and $\psi(f^*)$. The first and the second row correspond to *std_MSE*'s for f^* and $\psi(f^*)$, and each column corresponds to a data-generation machanism (**linear**, **neural** and **complex**). We see that the model performance in prediction (i.e., in learning f^*, first row) deteriorates quickly as d increase, showing strong pattern of the *curse of dimensionality*. Comparatively, the model's learning speed for variable importance $\psi^c(f^*)$ (second row) are much faster and less susceptible to the curse of dimensionality. 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 theorem assumption $f^* \in \mathscr{F}$ is violated (e.g. for **complex** f^* on Column 3), the posterior convergence still occur albeit at a slower rate.

Bernstein-von Mises Phenonmenon. We evaluate the model's convergence behavior toward the asymptotic posterior $N(0, \sigma^2_{BvM} = 4||H_p f_0||_n^2)$ using two metrics: (1) the standardized MSE for learning the standard deviation σ_{BvM}, which assesses whether the spread of the posterior distribution is correct. (2) The Cramér von Mises (CvM) statistic (i.e. the empirical L_2 distance) between the standardized posterior sample $\{\psi^c_{std,m}\}_{m=1}^M$ and a Gaussian distribution Φ: $CvM(\psi^c_{std}) = \frac{1}{M}\sum_{m=1}^M \big[\mathbb{F}(\psi^c_{std,m}) - \Phi(\psi^c_{std}, m)\big]^2$, which 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 that can never be zero, 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 summarizes the model's convergence behavior in standard deviation (measured by *std_MSE*, left) and in normality (measured by *CvM*, right). The shaded region in the right figure corresponds to the quantiles of a null CvM distribution. As shown, as the sample size increases, the standardized MSE for $sd(\psi^c)$ converges toward 0, and the CvM statistics enters into the range of the null distribution. The speed of convergence deteriorates as the data dimension increases, although not very dramatically. Above observations indicates that the credible intervals from the variable importance posterior $\Pi_n(\psi^c(f))$ indeed achieve the correct spread and shape in reasonably large sample, i.e. the Bernstein-von Mises phenomenon holds under the neural network model (Theorems 2 - 3). Consequently, the $(1-q)$-level credible intervals from $\Pi_n\Big(\psi^c(f)\Big)$ are valid tools for rigorous uncertainty quantification, i.e. they indeed cover the true parameter $\psi(f^*)$ for $(1-q)\%$ of the time.

Effectiveness in High-dimensional Variable Selection. In Appendix C, we present a comprehensive simulation study to compare the proposed approach (neural variable selection using credible intervals) against nine existing methods based on various models (linear-LASSO, random forest, neural network) and decision rules (LASSO/Spike-and-Slab 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 grow ($n \in (250, 500)$).

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

We observe that the proposed method, although simple, over-performs other specially-designed neural-network approaches by a significant margin (Table C). Comparing with other state-of-the-art approaches (e.g. LASSO-Knockoff and random-forest permutation tests), our method remains competitive in low-to-moderate dimension, and out-performs in high dimension due to the Bayesian ReLU network's effectiveness in learning / quantifying uncertainty in variable importance in high-dimension as shown by Theorem 1-3.

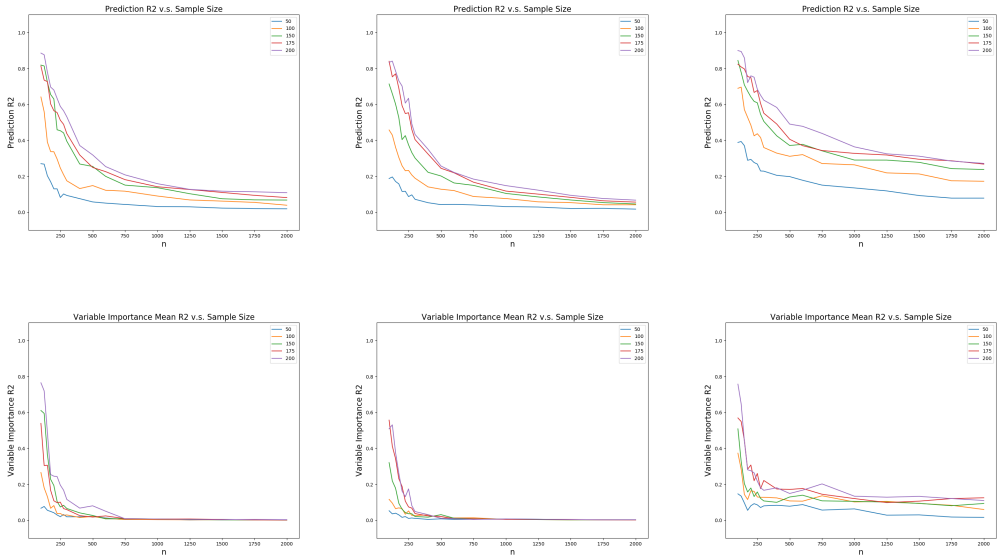

Figure 1: BNN's convergence behavior for learning prediction f^* (first row) and variable importance $\psi(f^*)$ (second row) under sample sizes $n \in (100, 2000)$ for $d \in (50, 200)$, measured by the standardized MSE (i.e. $1 - R^2$). Column 1-3 corresponds to **linear**, **neural**, and **complex**.

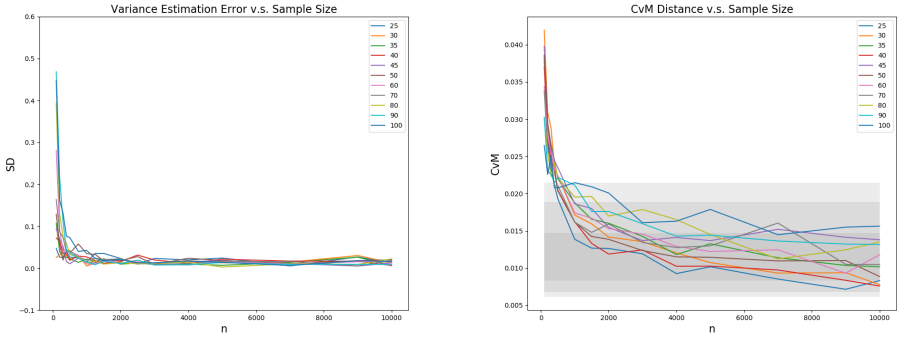

Figure 2: The variable importance posterior's convergence behavior toward the asymptotic standard deviation (left, measured by standardized MSE) and toward normality (right, measured by the CvM distance from a Gaussian distribution) under sample size $n \in (100, 10000)$ and $d \in (25, 100)$. Shaded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

References

[1] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, May 2010.

[2] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, Dec. 2008.

[3] F. Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. *arXiv:1412.8690 [cs, math, stat]*, Dec. 2014. arXiv: 1412.8690.

[4] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, Oct. 2015.

[5] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237, Feb. 2012.

[6] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[7] P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, Sept. 2013.

[8] E. Candès, Y. Fan, L. Janson, and J. 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.

[9] I. Castillo. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1):53–99, Feb. 2012.

[10] I. Castillo and R. Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 41(4):1999–2028, Aug. 2013.

[11] I. Castillo and J. Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43(6):2353–2383, Dec. 2015.

[12] J. Feng and N. Simon. Sparse Input Neural Networks for High-dimensional Nonparametric Regression and Classification. *arXiv:1711.07592 [stat]*, Nov. 2017. arXiv: 1711.07592.

[13] D. Freedman. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, Aug. 1999.

[14] A. Gelman, J. Hill, and M. Yajima. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, Apr. 2012.

[15] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, Feb. 2007.

[16] S. Ghosh and F. Doshi-Velez. Model Selection in Bayesian Neural Networks via Horseshoe Priors. *arXiv:1705.10388 [stat]*, May 2017. arXiv: 1705.10388.

[17] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. *arXiv:1905.01208 [cs, math, stat]*, May 2019. arXiv: 1905.01208.

[18] 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.

[19] X. He, J. Wang, and S. Lv. Scalable kernel-based variable selection with sparsistency. *arXiv:1802.09246 [cs, stat]*, Feb. 2018. arXiv: 1802.09246.

[20] A. Krizhevsky, I. Sutskever, and G. 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.

[21] F. Liang, Q. Li, and L. Zhou. Bayesian Neural Networks for Selection of Drug Sensitive Genes. *Journal of the American Statistical Association*, 113(523):955–972, July 2018.

[22] R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the LASSO. *The Annals of Statistics*, 42, Jan. 2013.

[23] C. Louizos, M. Welling, and D. P. Kingma. Learning Sparse Neural Networks through L_0 Regularization. Feb. 2018.

[24] H. Montanelli and Q. Du. New error bounds for deep networks using sparse grids. *arXiv:1712.08688 [math]*, Dec. 2017. arXiv: 1712.08688.

[25] R. Nakada and M. Imaizumi. Adaptive Approximation and Estimation of Deep Neural Network to Intrinsic Dimensionality. *arXiv:1907.02177 [cs, stat]*, July 2019. arXiv: 1907.02177.

[26] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag, New York, 1996.

[27] A. Rahimi and B. Recht. Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1313–1320. Curran Associates, Inc., 2009.

[28] V. Rivoirard and J. Rousseau. Bernstein–von Mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523, June 2012.

[29] V. Rockova and N. 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.

[30] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. Nonparametric Sparsity and Regularization. *Journal of Machine Learning Research*, 14:1665–1714, 2013.

[31] J. Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application*, 3(1):211–231, June 2016.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.

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

[34] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, Sept. 2014. arXiv: 1409.1556.

[35] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. Sept. 2018.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2015.

[37] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[38] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, June 2000.

[39] A. W. van der Vaart and J. H. van Zanten. Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[40] 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.

[41] P. M. Williams. Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, 7(1):117–143, Jan. 1995.

[42] L. Yang, S. Lv, and J. Wang. Model-free Variable Selection in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, 17(82):1–24, 2016.

[43] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *arXiv:1610.01145 [cs]*, Oct. 2016. arXiv: 1610.01145.

A Theoretical Aspect of Bayesian Learning of Neural Networks

The Bayesian approach to neural network learning specifies a prior distribution $\Pi(f)$ that assigns probability to every candidate $f \in \mathscr{F}(L,K,S,B)$ in the model space. For a neural network model $f(.) = \beta^\top \phi_{\mathscr{W}}(.)$, the prior distribution $\Pi(f)$ is commonly specified through its model weights $\{\beta, \mathscr{W}\}$, i.e., by specifying a hierarchically distribution $\Pi(f,\beta,\mathscr{W})$:

$$\Pi(f,\beta,\mathscr{W}) = \Pi(f|\beta,\mathscr{W})\Pi(\beta)\Pi(\mathscr{W}), \tag{5}$$

where $\Pi(f|\beta,\mathscr{W})$ is the Dirac measure $\mathbb{I}(f = \beta\phi_{\mathscr{W}})$. $\Pi(\beta)$, $\Pi(\mathscr{W})$ are the prior distributions for the output and hidden weights of the neural network model.

It is well-known that for some common choices of $\Pi(\beta)$, $\Pi(f)$ corresponds to a (conditional) Gaussian process (GP) [26]. Specifically, by placing i.i.d. Gaussian prior $N(0,\frac{1}{K})$ on β and $N(0,\sigma_{b_0}^2)$ on b_0, the neural network model $f(.) = \phi_{\mathscr{W}}(.)^\top\beta$ is equal in distribution to Gaussian process with kernel function $k_{\mathscr{W}}(\mathbf{x},\mathbf{x}') = \frac{1}{K}\phi_{\mathscr{W}}(\mathbf{x})^\top\phi_{\mathscr{W}}(\mathbf{x}') + \sigma_{b_0}^2$, i.e., $\Pi(f|\mathscr{W}) = GP(f|0,k_{\mathscr{W}})$.

As a result, under the conditional Gaussian process (GP) representation, the prior distribution for f can be written as:

$$\Pi(f,\mathscr{W}) = \Pi(f|\mathscr{W})\Pi(\mathscr{W}) = GP(f|0,k_{\mathscr{W}})\Pi(\mathscr{W}), \tag{6}$$

where the common choices for $\Pi(\mathscr{W})$ include Gaussian [26], Laplace [41], Spike and Slab [29], and Horseshoe [16, 23].

The conditional GP representation in (6) is important for analyzing the asymptotic behavior of the Bayesian neural network. It suggests that, if the behavior of the conditional posterior $\Pi_n(f|\mathscr{W})$ does not change drastically under $\Pi_n(\mathscr{W})$, then the asymptotic behavior of $\Pi_n(f,\mathscr{W})$ is analogous to that of a Gaussian process, whose theoretical properties are well-understood in the literature [39, 10]. In Section **??**, we take advantage of this representation to show an BvM phenomenon (i.e. asymptotic normality) for the posterior distribution of variable importance for a wide range of choices for $\Pi(\mathscr{W})$.

Posterior Consistency and Concentration Rates

The quality of a Bayesian learning procedure is commonly measured by the learning rate of its posterior distribution, i.e., the speed at which the posterior distribution Π_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, i.e., a set $A_n = \{f|||f - f_0||_n \leq M\varepsilon_n\}$ such that $\Pi_n(A_n) \to 1$. Here, the *concenration rate* ε_n describes how fast this small ball concentrates toward f_0 as the sample size increases. Clearly, an efficient Bayesian learning procedure that has good finite-sample performance is expected to have an ε_n that converges quickly to zero.

We state the notion of posterior concentration formally as below [15]:

Definition 1 (Posterior Concentration)**.** *For $f^* \in \mathscr{F}^*$ where $d = o(1)$, let $\mathscr{F}(L,K,S,B)$ denote a class of ReLU network with depth L, width W, sparsity bound S and norm bound B. Also denote f_0 the Kullback-Leibler (KL)-projection of f^* to $\mathscr{F}(L,K,S,B)$, and E_0 the expectation with respect to P_0. Then we say the posterior distribution f concentrates around f_0 at the rate ε_n in $\mathbb{P}_0^n$ probability if, for any $M_n \to \infty$, there exists an $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$ and:*

$$E_0\Pi(f : ||f - f_0||_n^2 > M_n\varepsilon_n|\{y_i,\mathbf{x}_i\}_{i=1}^n) \to 0 \tag{7}$$

To ensure the full generality of our theoretical results, in this work we do not assume an specific expression for ε_n (except that it is upper bounded by the optimal parametric rate $O(n^{-1/2})$), requiring only that such rate $\varepsilon_n \to 0$ exists and is attained by the ReLU network configured by the user. In practice, the exact value of the concenration rate ε_n depends on the property of the learning problem. Specifically, ε_n depends on the dimension of the input feature P, and the geometry of the "true" function space $f^* \in \mathscr{F}^*$. For example, under the typical nonparametric assumption that $\mathscr{F}^* = \mathscr{H}^\beta$ is the space of β-Hölder smooth (i.e. β-times differentiable) functions, the concentration rate ε_n is found to be close to minimax optimal up to a logarithm factor, i.e., $\varepsilon_n = O\left(n^{-2\beta/(2\beta+P)} * (log\, n)^\gamma\right)$ for some $\gamma > 1$ [29]. However, recent advances in frequentist learning theory shows that a deep

ReLU network model can surpass the optimal nonparametric rate by adapting to the structure of f^* [3, 35] and to the intrinsic dimension of the input feature space [25], suggesting that the Bayesian concetration rate ε_n can achieve similar adaptivity as well. In particular, if $\mathscr{F}^* = \mathscr{F}(L,K,S,B)$ (i.e. the target function lies in the approximation space of the neural networks), it can be shown that even under the standard i.i.d. Gaussian prior, ε_n can achieve a fast polynomial rate of $O(n^{-\frac{1}{2}})$ (up to a logarithm factor).

B Multivariate Bernstein-von Mises (BvM) Theorem for ψ^c

Theorem 3 (Multivariate Bernstein-von Mises (BvM) for ψ^c)**.** *For $f \in \mathscr{F}(L,W,S,B)$, assuming the posterior distribution $\Pi_n(f)$ contracts around f_0 at rate ε_n. For $\varepsilon = Proj_{\mathscr{F}}(\varepsilon)$, denote $\hat{\psi}^c = [\hat{\psi}^c_1, \ldots, \hat{\psi}^c_P]$ 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 and efficient estimator of $\psi(f_0) = [\psi_1(f_0), \ldots, \psi_P(f_0)]$, and the posterior distribution for $\psi^c(f)$ asymptotically converge toward a multivariate normal distribution surrounding $\hat{\psi}^c$, i.e.

$$\Pi\Big(\sqrt{n}(\psi^c(f) - \hat{\psi}^c)\Big|\{\mathbf{x}_i, y_i\}_{i=1}^n\Big) \rightsquigarrow MVN(0, V_0), \tag{8}$$

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

C Simulation Study for Variable Selection with Bayesian Neural Networks

In this section we empirically study the effectiveness of variable selection using the BNN's posterior credible intervals for variable importance, and compare the performance against other classic or machine-learning-based approaches that are popular in practice. As in Section C, we use i.i.d. Gaussian prior for model weights without any sparse-inducing penalty, so the method's effectiveness in variable selection relies only on the validity of the credible interval in uncertainty quantification.

Model / Metric	**Decision Rule**		
	Thresholding	Hypothesis Test	Knockoff
Linear Model - LASSO	[37]	[4]	[22]
Random Forrest - Impurity	[6]	[8]	[1]
	Group L_1 Thresholding	Spike-and-Slab Probability	Credible Interval
Neural Network - $\mathscr{W}_1$	[12]	[21]	
Neural Network - $\psi^c(f)$			(this work)

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

For the candidate variable selection methods, we notice that a variable selection method is usually consisted of three components: model, measure of variable importance, and the variable-selection decision rule. To this end, we consider nine methods that spans three types of models and three types of decision rule for each model (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 β, 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 [6]. (3) **NNet**, the (deep) neural networks that commonly measure feature importance using the magnitude of the input weights $\mathscr{W}_1$ or the gradient-norm variable importance $\psi^c(f)$. For **LASSO** and **RF**, we consider three types of decision rule: (1) **Heuristic Thresholding**, which selects variable by inspecting if $\hat{\beta}_p$ estimate is 0 or if impurity is greater than 1% of the total impurity summed over all variables; (2) **Knockoff** [8], a nonparametric inference procedure that constructs data-adaptive threshold for variable importance to control for false discovery rate (FDR), and (3) **Hypothesis Test**, which conducts either an asymptotic test on LASSO-regularized $|\beta_p|$ [22] or permutation-based test on impurity [1]. We select the **LASSO** hyper-parameters λ based on 10-fold cross validation, and use 500 regressions trees for **RF**. For **NNet** we also consider three decision rule: the frequentist approach with group-L_1 regularization on input weights $\mathscr{W}_1$ [12], the Bayesian approach with spike-and-slab prior on $\mathscr{W}_1$ [21], and our approach that is based on posterior credible

intervals of $\psi_p^c(f)$. 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, and we use $L = 1, W = 50$ for our credible-interval-based 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 $f^* \in \mathscr{F}(L^* = 1, W^* = 5)$ so the data closely matches the architecture of the regularized **NNets**. Notice this choice puts our method at a slight disadvantage 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 feature $\mathbf{X}_{n\times p}$ with dimension $d \in \{25, 75, 200\}$ which corresponds to low-, moderate, and high-dimensional situtation, and we vary sample size $n \in (250, 500)$. For each simulation setting (n, d), we repeat the experiment 20 times, and summarize each method's variable selection performance using the F_1 score (i.e., 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).

	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

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.

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

Our central conclusion is that **a powerful model along is not sufficient in guaranteeing effective variable selection**. (Recall that a variable-selection method is a combination of *model*, *measure* and *decision rule*). It is important that the variable selection decision is also based on a proper measure of variable importance (e.g., an unbiased and low-variance estimator of the true variable importance), and ideally with a rigorous decision rule that has statistical guarantee in variable selection (e.g., control over FDR or Type-I error). For example, while **NNet-Group** L_1 and **NNet-SpikeSlab** are based on a neural network architecture that closely matches the truth, their measure of variable importance is based on the input weight estimate $\hat{\mathscr{W}}_1$, which is over-parametrized and / or non-identifiable, and has high estimation variance which leads to unstable estimate of variable importance. As a result, the

performance of these two neural-network based methods are worse than **LASSO-knockoff** which is based on a linear model. Comparing between the decision rules, the thresholding-based decision rules (**LASSO-thres** and **RF-thres**) are mostly heuristic and are not optimized for variable selection performance. As a result, they are observed to have worse performance that are the most susceptible to the curse of dimensionality when compared to other methods based on the same model. 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 it produced good results for the low-variance linear model, but comparatively result for the more flexible but high-variance random forest. Finally, the hypothesis tests / credible intervals are model-based procedures whose performance depends on the quality of the model estimate. They are expected to be more powerful when the model is 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 variable-importance measure that has fast rate of convergence, and make its variable selection decision based on a credible-based selection rule that has rigorous statistical guarantee. As a result, even without any sparse-inducing model regularization, **NNet-CI** over-performed its **NNet**-based peers, and is more powerful than other **LASSO** or **RF** based approaches in high dimension.

D Proof for Theorem 1

Proof. Denote $A_n = \{f : ||f - f_0||_n^2 > M_n\varepsilon_n\}$ and $B_n = \{f : |\psi_p(f) - \Psi_p(f_0)| > M_n\varepsilon_n\}$, then showing the statement in (2) is equivalent to showing $\Pi_n(B_n) \to 0$.

Specifically, we assume below three facts hold:

Fact 1 $|\psi_p(f) - \psi_p(f_0)| \le ||\frac{\partial}{\partial x_p}f - \frac{\partial}{\partial x_p}f_0||_n^2$

Fact 2 $\sup_{p\in\{1,\dots,P\}} ||\frac{\partial}{\partial x_p}f - \frac{\partial}{\partial x_p}f_0||_n^2 \le C * ||f - f_0||_n^2$ for some constant C.

Fact 3 $\sup_{p\in\{1,\dots,P\}} |\psi_p(f_0) - \Psi_p(f_0)| \lesssim ||f - f_0||_n^2$.

Because if above facts hold, we then have

$$\begin{aligned}
\sup_{p\in\{1,\dots,P\}} |\psi_p(f) - \Psi_p(f_0)| &\le \sup_{p\in\{1,\dots,P\}} |\psi_p(f) - \psi_p(f_0)| + \sup_{p\in\{1,\dots,P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\
&\le \sup_{p\in\{1,\dots,P\}} ||\frac{\partial}{\partial x_p}f - \frac{\partial}{\partial x_p}f_0||_n^2 + \sup_{p\in\{1,\dots,P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\
&\le C * ||f - f_0||_n^2 + \sup_{p\in\{1,\dots,P\}} |\psi_p(f_0) - \Psi_p(f_0)| \\
&\lesssim ||f - f_0||_n^2,
\end{aligned}$$

it then follows that:

$$E_0\Pi_n\Big(\sup_{p\in\{1,\dots,P\}} |\psi_p(f) - \Psi_p(f_0)| \ge M_n\varepsilon_n\Big) \lesssim E_0\Pi_n\Big(||f - f_0||_n^2 \ge M_n'\varepsilon_n\Big) \to 0.$$

We now show Facts 1-3 are true:

- **Fact 1** follows simply from the triangular inequality:

$$\begin{aligned}
|\psi_p(f) - \psi_p(f_0)| &= \Big|||\frac{\partial}{\partial x_p}f||_n^2 - ||\frac{\partial}{\partial x_p}f_0||_n^2\Big| \\
&= \max\Big\{||\frac{\partial}{\partial x_p}f||_n^2 - ||\frac{\partial}{\partial x_p}f_0||_n^2,\ ||\frac{\partial}{\partial x_p}f_0||_n^2 - ||\frac{\partial}{\partial x_p}f||_n^2\Big\} \le ||\frac{\partial}{\partial x_p}f_0 - \frac{\partial}{\partial x_p}f||_n^2.
\end{aligned}$$

- **Fact 2**. First establish some notation. Given data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, denote $\mathbf{f}$ and $\mathbf{f}_0$ the $n \times 1$ vectors with their elements being $f(\mathbf{x}_i)$, $f_0(\mathbf{x}_i)$, respectively. We then have $||f - f_0||_n^2 = ||\mathbf{f} - \mathbf{f}_0||_2^2$. Furthermore,

since $f, f_0 \in \mathscr{F}(L,W,S,B)$, there exisits sets of weight matrices $\{\mathbf{W}_l, \mathbf{S}_l\}_{l=1}^L$, $\{\mathbf{W}_{0,l}, \mathbf{S}_{0,l}\}_{l=1}^L$ and output weights β, β_0 such that $\mathbf{f} = \mathbf{X}\mathbf{W}_1\mathbf{S}_1(\prod_{l=2}^L \mathbf{W}_l\mathbf{S}_l)\beta$ and $\mathbf{f}_0 = \mathbf{X}\mathbf{W}_{0,1}\mathbf{S}_{0,1}(\prod_{l=2}^L \mathbf{W}_{0,l}\mathbf{S}_{0,l})\beta_0$. To keep the notation simple, we write for $\mathbf{f}$ its the input weights as $\mathbf{W}$, and the product of weight matrices after the input layer as $\mathbf{D} = \mathbf{S}_1(\prod_{l=2}^L \mathbf{W}_l\mathbf{S}_l)\beta$, such that $\mathbf{f}$ and $\mathbf{f}_0$ can be written as:

$$\mathbf{f} = \mathbf{X}\mathbf{W}\mathbf{D}, \quad \mathbf{f}_0 = \mathbf{X}\mathbf{W}_0\mathbf{D}_0$$

where recall $\mathbf{X}$ is a $n \times nP$ block diagonal matrix with $1 \times p$ vectors $\mathbf{x}_i$'s on the diagonal. Furthermore, by the definition of gradient functions for ReLU network, we can write the $n \times 1$ vectors of gradient functions as:

$$\partial_p \mathbf{f} = \mathbf{W}_p\mathbf{D}, \quad \partial_p \mathbf{f}_0 = \mathbf{W}_{0,p}\mathbf{D}_0,$$

where $\mathbf{W}_p = \mathbf{I}_n \otimes \mathbf{w}_p, \mathbf{W}_{0,p} = \mathbf{I}_n \otimes \mathbf{w}_{0,p}$ are $n \times nK$ block diagonal matrices. Notice that $||\frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0||_n^2 = ||\partial_p \mathbf{f} - \partial_p \mathbf{f}_0||_2^2$. Finally, we will denote $\mathbb{X} = [\mathbf{X}^\top, \ldots, \mathbf{X}^\top]^\top$ a $nP \times nP$ matrix that is formed by stacking $\mathbf{X}$ for P times.

Consequently:

$$\begin{aligned}
\sup_{p\in\{1,\ldots,P\}} ||\frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0||_n^2 &\leq \sum_{p=1}^P ||\frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0||_n^2 \\
&= \sum_{p=1}^P ||\mathbf{W}_p\mathbf{D} - \mathbf{W}_{0,p}\mathbf{D}_0||_2^2 = ||\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0||_2^2 \\
&\leq ||\mathbb{X}^{-1}\mathbb{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)||_2^2 \leq ||\mathbb{X}^{-1}||_2^2 ||\mathbb{X}\mathbf{W}\mathbf{D} - \mathbb{X}\mathbf{W}_0\mathbf{D}_0||_2^2 \\
&= P * ||\mathbb{X}^{-1}||_2^2 ||\mathbf{X}\mathbf{W}\mathbf{D} - \mathbf{X}\mathbf{W}_0\mathbf{D}_0||_2^2 = P * ||\mathbb{X}^{-1}||_2^2 ||f - f_0||_n^2 \\
&\leq \frac{P}{c_x} ||f - f_0||_n^2.
\end{aligned}$$

The equality on the fourth line follows since $||\mathbb{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)||_2^2 = P * ||\mathbf{X}(\mathbf{W}\mathbf{D} - \mathbf{W}_0\mathbf{D}_0)||_2^2$, due to the fact that $\mathbb{X}$ is formed by stacking the $\mathbf{X}$ matrix P times. Recall that $\mathbf{x}_i = \{x_{i,p}\}_{p=1}^P \in (0,1)^P$, i.e. $x_{i,p}$ is bounded away from zero. We denote this lower bound for $\mathbf{x}$ as $c_x > 0$, such that $x_{i,p} \geq c_x \forall i, p$. Then the inequality on the last line follows since $||\mathbb{X}^{-1}||_2$ is bounded by $\frac{1}{c_x}$. This is because $||\mathbb{X}^{-1}||_2 = \lambda_{max}(\mathbb{X}^{-1}) = 1/\lambda_{min}(\mathbb{X})$, and $\mathbb{X}$ can be re-ordered (through column permutation) into a block matrix with $P \times P$ diagonal matrices $\mathbf{X}_i = diag(x_{i,1}, \ldots, x_{i,p})$, such that: $\mathbb{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n \\ \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n \end{bmatrix}$. Since eigenvalues of $\mathbb{X}$ is invariant under column permutation, and eigenvalue of a block matrix is the eigenvalues of the matrix blocks on the diagonal, we see that $\lambda_{min}(\mathbb{X}) = min_{i,p}(x_{i,p}) \geq c_x$. Consequently, since $\frac{P}{c_x} = O_p(1)$, we have shown that for some constant C:

$$\sup_{p\in\{1,\ldots,P\}} ||\frac{\partial}{\partial x_p} f - \frac{\partial}{\partial x_p} f_0||_n^2 \leq C * ||f - f_0||_n^2$$

- **Fact 3** follows from standard Bernstein-type concentration inequality (see, e.g. Lemma 18 of [30]). Specifically, for $|\frac{\partial}{\partial x_p} f_0(\mathbf{x})|^2$ a random variable with respect to probability measure $P(\mathbf{x})$ that is bounded by L. Given n i.i.d. samples $\{|\frac{\partial}{\partial x_p} f_0(x_i)|^2\}_{i=1}^n$, recall that $\hat{\psi}(f_0) = \frac{1}{n}\sum_{i=1}^n |\frac{\partial}{\partial x_p} f_0(x_i)|^2$ and $\psi(f_0) = E(\frac{\partial}{\partial x_p} f_0)$, then with probability $1 - \eta$:

$$|\hat{\psi}(f_0) - \psi(f_0)| \leq n^{-\frac{1}{2}} * \big(2\sqrt{2} * L * log(2/\eta)\big),$$

that is, $|\hat{\psi}(f_0) - \psi(f_0)| \to 0$ at rate of $O(n^{-\frac{1}{2}})$. Notice that $O(n^{-\frac{1}{2}})$ is the optimal parametric rate that cannot be surpassed by the convergence speed of the ReLU networks (recall the typical convergence rate is $\varepsilon_n \asymp n^{-\frac{\beta}{2\beta+\delta}} * log(n)^\gamma$ for some $\delta > 0$ and $\gamma > 1$). Therefore we have:

$$\sup_{p\in\{1,\ldots,P\}} |\psi_p(f_0) - \Psi_p(f_0)| \lesssim ||f - f_0||_n^2.$$

□

E Proof for Theorem 2

E.1 Background: Semi-parametric BvM Theorem for Smooth Functionals

In this section, we provide background on a general semi-parametric BvM theorem for smooth nonlinear functionals [11]. In nonparametric regression, the regression function $f \in \mathscr{F}$ is infinite-dimensional and the asymptotic distribution of f in this case is in general difficult to characterize [13]. However, in practical applications, we are mostly interested in a finite-dimensional parameter $\psi : \mathscr{F} \to \mathbb{R}^d$ whose asymptotic distribution is easier to reason with. For example, a cumulative distribution function at a fixed point $F(x_0) = \int \mathbb{I}(x < x_0) f(x) dx$ [28].

To this end, a series of work by [5, 28, 11] provided general sufficient conditions for BvM theorem in smooth functionals under general models. These results show that, if the functional of interest ψ and the model log likelihood l_n both satisfy certain smoothness conditions, then the marginal posterior of $\psi(f)$ concentrates at the rate $O(n^{-1/2})$, and furthermore, the marginal posterior distribution of $\sqrt{n}(\psi(f) - \hat{\psi})$ converges weakly to a $N(0, V_0)$ under the data-generation distribution P_0, where $\hat{\psi}$ is an efficient estimator of $\psi(f_0)$. Such properties have the implication that it allow the construction of credible regions for which have correct asymptotic frequentist coverage [10].

The main conditions are as below:

1. Locally Asymptotic Normal (LAN) Expansion of Likelihood Function $l_n(f)$:
$$l_n(f) - l_n(f_0) = -\frac{n}{2}||f - f_0||_n^2 + \sqrt{n} W_n(f - f_0). \tag{9}$$

2. Smoothness Expansion of Functional $\psi(f)$:
$$\psi(f) - \psi(f_0) = \langle \psi_1, f - f_0 \rangle_n + \frac{1}{2}\langle \psi_2(f - f_0), f - f_0 \rangle_n. \tag{10}$$

3. Relation between $l_n(f)$ and $\psi(f)$:
For a posterior distribution $\Pi_n(f) = \Pi(f|\{y_i, \mathbf{x}_i\}_{i=1}^n)$ that concentrates around f_0 at rate ε_n, i.e. $\Pi_n(f : ||f - f_0||_n \le \varepsilon_n) = 1 + o_p(1)$, define A_n as the sequence of sets that receive majority of probability mass from Π_n, i.e.
$$\Pi_n(A_n) = \Pi_n(f \in A_n : ||f - f_0||_n \le \varepsilon_n) = 1 + o_p(1).$$
Assume there exists $w_n \in \mathscr{F}$ such that W_n adopts a decomposition
$$W_n(f) = \langle w_n, f \rangle_n + \Delta_n(f),$$
where w_n is the "representor" of W_n such that $\langle w_n, f \rangle_n$ retains majority of information from $W_n(f)$, and $\Delta_n(f)$ is the corresponding residual term. It is required that both of these terms are sufficiently regular in the sense that they satisfy below two conditions:
$$\langle w_n, \psi_2(\psi_1) \rangle_n + \varepsilon_n ||w_n||_n = o_p(\sqrt{n}), \tag{11}$$
$$\sup_{f \in A_n} |\Delta_n(\psi_2(f - f_0))| = o_p(1). \tag{12}$$

Then under some mild additional conditions, BvM is valid in below sense:

Theorem E.1 (Semiparametric BvM Theorem)**.** *Let W_n, w_n and ψ, ψ_1, ψ_2 as defined above. Furthermore, denote*
$$f_t = f - \frac{t}{\sqrt{n}}\Big(\psi_1 + \frac{1}{2}\psi_2(f - f_0)\Big) - \frac{t}{2n}\psi_2(w_n)$$
and
$$\hat{\psi} = \psi(f_0) + \frac{W_n(\psi_1)}{\sqrt{n}} + \frac{1}{2}\frac{\langle w_n, \psi_2(w_n) \rangle_n}{n}, \quad V_{0,n} = \Big|\Big|\psi_1 - \frac{1}{2}\frac{\psi_2(w_n)}{\sqrt{n}}\Big|\Big|_n^2$$
Then the "moment generating function (MGF)" of $\sqrt{n}(\psi(f) - \hat{\psi})$ under posterior distribution Π_n evaluated at the set A_n such that $\Pi_n(A_n) = 1$ can be written as:
$$E_n(e^{t\sqrt{n}(\psi(f) - \hat{\psi})}|A_n) = e^{o_p(1) + t^2 V_{0,n}/2} * \mathscr{I}_n, \quad \textit{where} \quad \mathscr{I}_n = \frac{\int_{A_n} e^{l_n(f_t) - l_n(f_0)} d\Pi(f|\mathscr{W})}{\int_{A_n} e^{l_n(f) - l_n(f_0)} d\Pi(f|\mathscr{W})}$$

Moreover, if $V_{0,n} = ||\psi_1||_n^2 + o_p(1)$ *and* $\mathscr{I}_n = o_p(1)$*, then the posterior distribution* $\sqrt{n}(\psi(f) - \hat{\psi}_p)$ *is asymptotically normal with mean zero and variance* $||\psi_1||_n^2$*, i.e.*

$$\Pi_n\Big(\sqrt{n}(\psi(f) - \hat{\psi}_p)\Big) \rightsquigarrow N(0, ||\psi_1||_n^2) \tag{13}$$

Proof. [11], Theorem 2.1. □

Although the original theorem is stated under the scalar case, the generalization to multivariate case is straightforward, i.e., one only need to generalize V_0 to the corresponding matrix form following the definitition of ψ_1 [11].

E.2 Preliminary I: Notations and Basic Setup

In this section, we set up the basic notations for showing semi-parametric BvM theorems for $\psi_p(f)$ in a nonparametric regression model. We will first verify the model likelihood $l_n(f) = -\frac{1}{2}\sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$ and the functional $\psi(f) = \langle H_p(f), f\rangle_n$ satisfy the three conditions for Theorem E.1, and by doing so, identify the expression for the technical terms W_n, w_n, ψ_1, ψ_2 that are relevant for deriving the asymptotic distribution of $\sqrt{n}(\psi(f) - \hat{\psi})$.

First verify (9) the LAN condition for model likelihood $l_n(f)$ and derive expression for W_n. Under independent Gaussian assumption, the likelihood for nonparametric regression adopts the LAN expansion:

$$l_n(f) - l_n(f_0) = -\frac{n}{2}||f - f_0||_n^2 + \sqrt{n}W_n(f - f_0)$$

where $||f - f_0||_n^2 = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$, and W_n is:

$$W_n(f) = \langle\sqrt{n}\varepsilon, f\rangle_n = \frac{1}{n}\sum_{i=1}^n \sqrt{n}\varepsilon_i * f(\mathbf{x}_i) \tag{14}$$

Now verify the rest of the two conditions, we consider two cases: the univariate case where $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2$, to be used by the univariate BvM Theorem 2, and the multivariate case $\psi(f)_{P\times 1} = [||\frac{\partial}{\partial x_1} f||_n^2, \ldots, ||\frac{\partial}{\partial x_P} f||_n^2]^\top$, to be used by the multivariate BvM Theorem 3..

Now verify (10) the smoothness condition for functional of interest $\psi^c(f)$ and derive expressions for ψ_1, ψ_2. The centered quadratic norm of gradient $\psi^c(f) = \langle H_p(f), f\rangle_n - E(\langle H_p\,\omega, \omega\rangle_n)$ adopts the smoothness expansion:

$$\psi_p^c(f) - \psi_p^c(f_0) = \langle\psi_1, f - f_0\rangle_n + \frac{1}{2}\langle\psi_2(f - f_0), f - f_0\rangle_n,$$

in which ψ_1, ψ_2 take the form:

$$\psi_1 = 2H_p(f_0), \quad \psi_2(f) = 2H_p(f),$$

where $H_p = D_p^\top D_p$ for $D_p : f \to \frac{\partial}{\partial x_p} f$ the differentiation operator and $D_p^\top$ the adjoint of D_p. Given data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, recall the definition of Φ, $\partial_p\Phi$ (Section **??**) and denote $\Phi_{K\times n}^+$ the generalized inverse of Φ, the operator D_p can be evaluated in matrix form as $\mathbf{D}_p = \partial_p\Phi\Phi^+$, and $D_p(f)$ can be evaluated as $\mathbf{D}_p\mathbf{f} = (\partial_p\Phi\Phi^+)\Phi\beta = \partial_p\Phi\beta$ for $f \in \mathscr{F}(L, K, S, B)$. Correspondingly, the operator H_p adopts matrix reprentation $\mathbf{H}_p = \mathbf{D}_p^\top\mathbf{D}_p = (\Phi^+)^\top\partial_p\Phi^\top\partial_p\Phi\Phi^+$, such that $\langle H_p(f), f\rangle = (\mathbf{H}_p\Phi\beta)^\top\Phi\beta = (\partial_p\Phi\beta)^\top(\partial_p\Phi\beta)$.

Finally, for the decomposition $W_n(f) = \langle\omega, f\rangle_n + \Delta_n(f)$, we will define $\omega = P_{\mathscr{F}}(\varepsilon)$ the projection of ε to $\mathscr{F}$, and $\Delta_n(f) = \langle P_{\mathscr{F}}^\perp(\varepsilon), f\rangle$. Given observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$, the projection operator $P_{\mathscr{F}}^\perp$ can be evaluated by computing the projection matrix $\mathbf{P}_{\mathscr{F}} = \mathbf{U}\mathbf{U}^\top$ and $\mathbf{P}_{\mathscr{F}}^\perp = \mathbf{I} - \mathbf{U}\mathbf{U}^\top$ for $\Phi_{n\times K} = \mathbf{U}_{n\times K}\mathbf{D}_{K\times K}\mathbf{V}_{K\times K}^\top$. By noticing that $\mathbf{P}_{\mathscr{F}}$ is a rank K matrix, it is then easy to see that the two conditions (11) and (12) are satisfied since $||\omega||_n = O(K) \lesssim O(\sqrt{n})$ and $P_{\mathscr{F}}^\perp(\varepsilon)$ is orthogonal to $\psi_2(f - f_0) \in \mathscr{F}_{\mathscr{W}}$.

As an aside, we note that due to the existence of the bias term at the output layer, the actual feature matrix is $\Phi_1 = [\mathbf{1}, \Phi]$. However, this does not impact the expression of $\mathbf{D}_p$ or $\mathbf{H}_p$, since $\mathbf{D}_p = \partial_p \Phi_1 \Phi_1^+ = [\mathbf{0}, \partial_p \Phi][(\mathbf{1}^+)^\top, (\Phi^+)^\top]^\top = \partial_p \Phi \Phi^+$ where $\mathbf{1}_{n\times 1}^+$ is a vector that is orthogonal to $\mathbf{1}_{n\times 1}$ and $\Phi_{n\times K}^+$.

E.3 Preliminary II: Proof Strategy and Preliminary Theorems

Recall that under a deep ReLU neural network, the prior distribution adopts a conditional Gaussian process representation (Section **??**):

$$\Pi(f, \mathscr{W}) = \Pi(f|\mathscr{W})\Pi(\mathscr{W}) = GP(f|0, k_{\mathscr{W}})\Pi(\mathscr{W}).$$

This decomposition suggests that a neural network model can be treated as a Gaussian process with an adaptive kernel function $k_{\mathscr{W}}$, whose hyperparameters $\mathscr{W}$ follows a prior distribution $\Pi(\mathscr{W})$.

Consequently, we use a two-step strategy to show BvM phenomenon for ReLU network:

- **Step 1**, fix hidden weight $\mathscr{W}$ and show BvM phenonmenon hold for $GP(f|0, k_{\mathscr{W}})$. This essentially corresponding to performing Bayesian inference for a *randomized neural network* whose hidden weights are sampled *a priori* from certain fixed distribution [27]. Then
- **Step 2**, we show that such BvM phenonmenon for $\Pi_n(f|\mathscr{W})$ still holds under the posterior distribution of hidden weights $\mathscr{W} \sim \Pi_n(\mathscr{W})$.

Theorem E.2 establishes **Step 1**. Notice that in the fixed-$\mathscr{W}$ case, $f \in \mathscr{F}$ follows an exact GP with effective model dimension K (i.e. the rank of the kernel matrix), for whom the BvM phenonmenon are known to hold under suitable regularity conditions [9, 10]. Therefore it is expected that BvM to hold for randomized neural network $f \in \mathscr{F}_{\mathscr{W}}$, provided K does not grow too fast with respect to n (i.e. Assumption (1)) and the functional $\psi_p(f)$ is sufficiently smooth (i.e. satisfying (10)), which is true for $\psi_p(f) = ||\frac{\partial}{\partial x_p} f||_n^2$:

Theorem E.2 (Bernstein-von Mises (BvM) for ψ_p^c, Fixed Hidden Weights)**.** *For $f \in \mathscr{F}_{\mathscr{W}}(L, W, S, B)$ a deep ReLU network with hidden weight fixed to $\mathscr{W}$, denoting $f_{0,\mathscr{W}}$ the projection of f_0 to $\mathscr{F}_{\mathscr{W}}$, and assume the posterior distribution $\Pi_n(f|\mathscr{W})$ contracts around $f_{0,\mathscr{W}}$ at rate ε_n. Denote $D_{\mathscr{W},p} : f \to \frac{\partial}{\partial x_p} f$ the differentiation operator in $\mathscr{F}_{\mathscr{W}}$, and $H_{\mathscr{W},p} = D_{\mathscr{W},p}^\top D_{\mathscr{W},p}$ the corresponding self-adjoint operator. For $\omega_{\mathscr{W}} = Proj_{\mathscr{F}_{\mathscr{W}}}(\varepsilon)$ the projection of ε to $\mathscr{F}_{\mathscr{W}}$, define:*

$$\hat{\psi}_{\mathscr{W},p} = ||D_{\mathscr{W},p}(f_0 + \omega_{\mathscr{W}})||_n^2 = \psi_{\mathscr{W},p}(f_{0,\mathscr{W}}) + 2\langle H_{\mathscr{W},p} f_{0,\mathscr{W}}, \omega_{\mathscr{W}}\rangle_n + \langle H_{\mathscr{W},p}\omega_{\mathscr{W}}, \omega_{\mathscr{W}}\rangle_n, \tag{15}$$

Define $\hat{\psi}_{\mathscr{W},p}^c = \hat{\psi}_{\mathscr{W},p} - \eta_{\mathscr{W},n}$ where $\eta_{\mathscr{W},n} = E_0\big(\langle H_{\mathscr{W},p}\omega, \omega\rangle_n\big)$. Then $\hat{\psi}_{\mathscr{W},p}^c$ is an unbiased and efficient estimator of $\psi_{\mathscr{W},p}(f_0)$, and the posterior distribution for $\psi_{\mathscr{W},p}^c(f)$ is asymptotically normal surrounding $\hat{\psi}_{\mathscr{W},p}^c$, i.e.

$$\Pi\Big(\sqrt{n}(\psi_{\mathscr{W},p}^c(f) - \hat{\psi}_{\mathscr{W},p}^c)\Big|\{\mathbf{x}_i, y_i\}_{i=1}^n, \mathscr{W}\Big) \rightsquigarrow N(0, 4||H_{\mathscr{W},p} f_{0,\mathscr{W}}||_n^2), \tag{16}$$

The proof is delayed to full manuscript. It should be stressed that both operators $D_{\mathscr{W},p}$ and $H_{\mathscr{W},p}$ are defined strictly with respect to $\mathscr{F}_{\mathscr{W}}$. Such that given data, the operator $D_{\mathscr{W},p}$ is evaluated in matrix form as $\mathbf{D}_{\mathscr{W},p} = \partial_p \Phi_{\mathscr{W}} \Phi_{\mathscr{W}}^+$, and $H_{\mathscr{W},p}$ is evaluated as $\mathbf{H}_{\mathscr{W},p} = (\Phi_{\mathscr{W}}^+)^\top \partial_p \Phi_{\mathscr{W}}^\top \partial_p \Phi_{\mathscr{W}} \Phi_{\mathscr{W}}^+$. In comparison, the original D_p and H_p defined Section E.2 are with respect to the optimal solution $f_0 \in \mathscr{F}$.

E.4 Proof for Theorem 2

We now prove Theorem 2, which establishes **Step 2** of the proof strategy outlined in Section E.3.

Our goal is to show that the BvM phenomenon in Theorem E.2 still holds under the *adaptive* case (i.e. $\mathscr{W}$ is not fixed but follows the posterior distribution $\Pi_n(\mathscr{W})$), and furthermore, the whole posterior distribution of $\sqrt{n}(\psi_p^c(f) - \hat{\psi}_p^c(f))$ converges to $N(0, 4||H_p f_0||_n^2)$ where H_p is defined with respect to the optimal solution $f_0 \in \mathscr{F}$.

Proof. Our goal is to show:

$$\Pi\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_p^c)\Big|\{\mathbf{x}_i,y_i\}_{i=1}^n\Big)\rightsquigarrow N(0,4||H_pf_0||_n^2).$$

First notice that by Theorem E.2, the asymptotic distribution of the marginal posterior distribution can be represented as a mixture of Gaussian:

$$\begin{aligned}\Pi_n\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_p^c)\le z\Big)&=\int_{\mathscr{W}}\Pi_n\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_p^c)\le z|\mathscr{W}\Big)d\Pi_n(\mathscr{W})\\&=\int_{\mathscr{W}}\Pi_n\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_{\mathscr{W},p}^c)\le z+\sqrt{n}(\hat{\psi}_p^c-\hat{\psi}_{\mathscr{W},p}^c)|\mathscr{W}\Big)d\Pi_n(\mathscr{W})\\&=\int_{\mathscr{W}}\Phi\Big(\big(z+\sqrt{n}(\hat{\psi}_p^c-\hat{\psi}_{\mathscr{W},p}^c)\big)/\sqrt{V_{\mathscr{W},0}}\,\Big|\mathscr{W}\Big)d\Pi_n(\mathscr{W})\end{aligned}\tag{17}$$

where the last line follows from Theorem E.2, where $V_{\mathscr{W},0}=4||H_{\mathscr{W},p}f_{\mathscr{W},0}||_n^2$ and Φ is the standard Gaussian cumulative distribution function (CDF).

Clearly, for BvM to hold in the case of (20), it is sufficient to show below two conditions [11]:

$$|V_{\mathscr{W},0}-V_0|=o_p(1),\quad \sqrt{n}|\hat{\psi}_p^c-\hat{\psi}_{\mathscr{W},p}^c|=o_p(1).\tag{18}$$

The first condition follows from the continuous mapping theorem for $V(H_pf_0)=2||H_pf_0||_n^2$, along with the fact that :

$$\begin{aligned}||H_{\mathscr{W},p}f_{\mathscr{W},0}-H_pf_0||_n&\le||(H_{\mathscr{W},p}-H_p)f_0||_n+||H_{\mathscr{W},p}(f_{\mathscr{W},0}-f_0)||_n\\&=O(||H_{\mathscr{W},p}-H_p||_n)+O(||f_{\mathscr{W},0}-f_0||_n),\\&=O(\frac{1}{\sqrt{n}}||H_{\mathscr{W},p}-H_p||_F)+o_p(1)\\&=O(\frac{K}{\sqrt{n}})+o_p(1)=o_p(1),\end{aligned}$$

where the first equality follows from the boundedness of $|f_0|_\infty$ and $|H_{\mathscr{W},p}|_\infty$ (by assumption in main article and also Proposition **??**), the second equality follows from the definition of $||.||_n$ for matrix and the fact about posterior concentration of $||f-f_0||_n^2$ in the statement of Theorem 2. The last line follows since $||\mathbf{H}_p||_F=O(K)$ by Proposition (**??**) and the assumption that $K=o_p(\sqrt{n})$ (Assumption (1) in the main article).

The second condition in (18) is the important *no-bias* condition which ensures that under $\mathscr{W}\sim\Pi_n(\mathscr{W})$, all the conditional posterior $\psi_p^c|\mathscr{W}$ converges toward the same target $\hat{\psi}_p^c$ [11, 31]. Recall that $\hat{\psi}_p^c=\psi_p(f_0)+2\langle H_pf_0,\omega\rangle_n+\langle H_p\omega,\omega\rangle_n-E(\langle H_p\omega,\omega\rangle_n)$, then the second condition can be written as:

$$\begin{aligned}\sqrt{n}|\hat{\psi}_p^c-\hat{\psi}_{\mathscr{W},p}^c|\le&\sqrt{n}|\psi_p(f_0)-\psi_{\mathscr{W},p}(f_{\mathscr{W},0})|+2\sqrt{n}|\langle H_pf_0,\omega\rangle_n-\langle H_{\mathscr{W},p}f_{\mathscr{W},0},\omega_{\mathscr{W}}\rangle_n|+\\&\sqrt{n}|\langle H_{\mathscr{W},p}\omega_{\mathscr{W}},\omega_{\mathscr{W}}\rangle_n-\langle H_p\omega,\omega\rangle_n|+\sqrt{n}|E(\langle H_{\mathscr{W},p}\omega_{\mathscr{W}},\omega_{\mathscr{W}}\rangle_n)-E(\langle H_p\omega,\omega\rangle_n)|,\end{aligned}\tag{19}$$

where all four terms are $o_p(1)$ since they are all $O(K/\sqrt{n})$ and that the model dimension K is not too large (i.e. $K=o_p(n^{1/2})$). We delay the detailed arguments to the end of the proof.

Consequently, since both conditions in (18) are satisfied, the expression in (20) converge uniquely to a normal distribution under the posterior distribution $\Pi_n(\mathscr{W})$, i.e.,

$$\begin{aligned}\Pi_n\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_p^c)\le z\Big)&=\int_{\mathscr{W}}\Phi\Big(\big(z+\sqrt{n}(\hat{\psi}_p^c-\hat{\psi}_{\mathscr{W},p}^c)\big)/\sqrt{V_{\mathscr{W},0}}\,\Big|\mathscr{W}\Big)d\Pi_n(\mathscr{W})\\&=\int_{\mathscr{W}}\Phi\Big(\big(z+o_p(1)\big)/\sqrt{V_0+o_p(1)}\,\Big|\mathscr{W}\Big)d\Pi_n(\mathscr{W})\\&=\Phi\big(z/\sqrt{V_0}\big),\qquad\text{where}\qquad V_0=4||H_0f_0||_n^2\end{aligned}\tag{20}$$

which implies the statement of interest:

$$\Pi\Big(\sqrt{n}(\psi_p^c(f)-\hat{\psi}_p^c)\Big|\{\mathbf{x}_i,y_i\}_{i=1}^n\Big)\rightsquigarrow N(0,4||H_pf_0||_n^2).$$

We are only left to show that all four terms in the expression (19) are $o_p(1)$. Specifically, recall that $H_p = D_p^\top D_p$ such that $\langle H_p a, b\rangle_n = \langle D_p a, D_p b\rangle_n$ for any $a, b \in \mathscr{F}$, then:

- **First Term**: Recall $\psi_p(f_0) = \langle H_p f_0, f_0\rangle_n = ||D_p f_0||_n^2$, then the first term can be expressed as:

$$\begin{aligned}
\sqrt{n}|\psi_p(f_0) - \psi_{\mathscr{W},p}(f_{\mathscr{W},0})| &= \sqrt{n}\Big|||D_p f_0||_n^2 - ||D_{\mathscr{W},p} f_{\mathscr{W},0}||_n^2\Big| \\
&\leq \sqrt{n}\Big(||D_p||_n^2||f_0||_n^2 + ||D_{\mathscr{W},p}||_n^2||f_{\mathscr{W},0}||_n^2\Big) \\
&= \sqrt{n}\Big(O_p(||D_p||_n^2) + O_p(||D_{\mathscr{W},p}||_n^2)\Big) = \frac{1}{\sqrt{n}}\Big(O_p(||\mathbf{D}_p||_F^2) + O_p(||\mathbf{D}_{\mathscr{W},p}||_n^2)\Big) \\
&= O(\frac{K}{\sqrt{n}}) = o_p(1)
\end{aligned}$$

  where on the third line, the first equality follows since f_0 and $f_{\mathscr{W},0}$ are both bounded, the second equality follows by the definition of the matrix Euclidean norm $||\mathbf{M}||_n^2 = \frac{1}{n}\sum_{i,j}\mathbf{M}_{i,j}^2 = \frac{1}{n}||\mathbf{M}||_F^2$. On the last line, the first eqality follows by $||\mathbf{D}_{\mathscr{W},p}||_F^2 = tr(\mathbf{H}_{\mathscr{W},p}) = O(K)$ due to Proposition **??**, and the second equality follows by Assumption $K = o_p(n^{1/2})$.

- **Second Term**: Similarly, the second term can be expressed as:

$$\begin{aligned}
\sqrt{n}|\langle H_p f_0, \omega\rangle_n - \langle H_{\mathscr{W},p} f_{\mathscr{W},0}, \omega_{\mathscr{W}}\rangle_n| &= \sqrt{n}\big|\langle D_p f_0, D_p\omega\rangle_n - \langle D_{\mathscr{W},p} f_{\mathscr{W},0}, D_{\mathscr{W},p}\omega_{\mathscr{W}}\rangle_n\big| \\
&\leq \sqrt{n}\Big(||D_p||_n^2||f_0||_n||\omega||_n + ||D_{\mathscr{W},p}||_n^2||f_{\mathscr{W},0}||_n||\omega_{\mathscr{W}}||_n\Big) \\
&= \sqrt{n}\Big(O_p(||D_p||_n^2) + O_p(||D_{\mathscr{W},p}||_n^2)\Big) \\
&= O_p(\frac{K}{\sqrt{n}}) = o_p(1)
\end{aligned}$$

  where the equality on the third line follows from the fact that $f_{0,\mathscr{W}}$ is bounded and $\omega = P_{\mathscr{F}}\varepsilon$ is a random variable with bounded variance. The rest of the equalities follow similarly as those in the First term.

- **Third and Fourth Terms** are similar to the first term except for f_0 is replaced by ω. As a result:

$$\begin{aligned}
\sqrt{n}|\langle H_{\mathscr{W},p}\omega_{\mathscr{W}}, \omega_{\mathscr{W}}\rangle_n - \langle H_p\omega, \omega\rangle_n| &= \sqrt{n}\Big|||D_{\mathscr{W},p}\omega_{\mathscr{W}}||_n^2 - ||D_p\omega||_n^2\Big| \\
&= \sqrt{n}\Big(O_p(||D_p||_n^2) + O_p(||D_{\mathscr{W},p}||_n^2)\Big) \\
&= O_p(\frac{K}{\sqrt{n}}) = o_p(1) \\
\sqrt{n}|E(\langle H_{\mathscr{W},p}\omega_{\mathscr{W}}, \omega_{\mathscr{W}}\rangle_n) - E(\langle H_p\omega, \omega\rangle_n)| &= \sqrt{n}O_p\Big(|\langle H_{\mathscr{W},p}\omega_{\mathscr{W}}, \omega_{\mathscr{W}}\rangle_n - \langle H_p\omega, \omega\rangle_n|\Big) \\
&= O_p(\frac{K}{\sqrt{n}}) = o_p(1)
\end{aligned}$$

□