Uncertainty Quantification for Sparse Deep Learning

Yuexi Wang and Veronika Ročková
Booth School of Business, University of Chicago

Abstract

Deep learning methods continue to have a decided impact on machine learning, both in theory and in practice. Statistical theoretical developments have been mostly concerned with approximability or rates of estimation when recovering infinite dimensional objects (curves or densities). Despite the impressive array of available theoretical results, the literature has been largely silent about *uncertainty quantification* for deep learning. This paper takes a step forward in this important direction by taking a Bayesian point of view. We study Gaussian approximability of certain aspects of posterior distributions of sparse deep ReLU architectures in non-parametric regression. Building on tools from Bayesian non-parametrics, we provide semi-parametric Bernstein-von Mises theorems for linear and quadratic functionals, which guarantee that implied Bayesian credible regions have valid frequentist coverage. Our results provide new theoretical justifications for (Bayesian) deep learning with ReLU activation functions, highlighting their *inferential potential*.

1 Introduction

Neural networks have emerged as one of the most powerful prediction systems. Their empirical success has been amply documented in many applications including image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012) or game intelligence (Silver et al., 2016). Beyond algorithmic developments, there has been a rapid progress in theoretical understanding of deep learning (Anthony and Bartlett, 2009). The majority of existing *statistical* theory has been concerned with *prediction* aspects, e.g. approximability (Telgarsky, 2016; Yarotsky, 2017; Vitushkin, 1964) or rates of convergence (either from a frequentist point of view (Mhaskar et al., 2017; Poggio et al., 2017; Schmidt-Hieber, 2017) or a Bayesian point of view (Polson and Rockova, 2018)). A distinguishing feature of statistics, that goes beyond mere construction of prediction maps, is providing uncertainty quantification (UQ) for inference (hypothesis testing and confidence assessments). The statistical approach to uncertainty quantification uses observations to construct a random subset (confidence set) which contains the truth with large probability. While computational methods such as Boostrapped DQN (Osband et al., 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) have been proposed to quantify predictive uncertainty, theoretically justifiable developments on UQ for deep learning are more rare.

A structured approach to the problem of uncertainty assessment lies in Bayesian hierarchical modeling. The Bayesian paradigm for deep learning places a probabilistic blanket over architectures/parameters and allows for uncertainty quantification via posterior distributions (Neal, 1993). While exact Bayesian inference is computationally intractable, many approximate methods have been developed including MCMC (Neal, 2012), Variational Bayes (Ullrich et al., 2017), Bayes by Backprop (Blundell et al., 2015), Scalable Data Augmentation (Wang et al., 2019), Monte Carlo Dropout (Gal and Ghahramani, 2016), Hamiltonian methods (Springenberg et al., 2016). The Bayesian inference is fundamentally justified by the Bernstein-von Mises (BvM) theorem. The BvM phenomenon occurs when, as the number of observations increases, the posterior distribution is approximately Gaussian, centered at an efficient estimator of the parameter of interest. Moreover, the posterior credible sets, i.e. regions with prescribed posterior probability, are then also confidence regions with the same asymptotic coverage. While the BvM limit is not unexpected in regular parametric models, infinite-dimensional notions of BvM are far from obvious (see e.g. Castillo and Nickl (2013)).

Proceedings of the 23^{rd}International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

Our paper deals with uncertainty quantification. Our approach is inherently Bayesian and, as such, is conceptually epistemic where uncertainty about the unknown state of nature is expressed through priors and coherently updated with the data. The frequentist notion of uncertainty is primarily aleatoric as it reflects variability in possible realizations of an event that is largely stochastic in nature and is irreducible. The premise of the BvM phenomenon is that these two uncertainties, while qualitatively very different, are not mutually exclusive in the sense that their quantifications can agree. Priors that are not subjective and more automatic do not necessarily adhere to epistemic interpretation and can yield aleatoric measures of quantification. Our work sheds light on the fact that frequentist calibration is an attainable goal of Bayesian statistical procedures, where the BvM phenomenon facilitates communication of uncertainty using the more universally understood frequentist concept (Dawid, 1982).

In this note, we study the *semi-parametric* BvM phenomenon concerning the limiting behavior of the posterior distribution of certain low-dimensional summaries of a regression function. In particular, we assume a non-parametric regression model with fixed covariates and sparse deep ReLU network priors, which have been recently shown to attain the optimal speed of posterior contraction (Polson and Rockova, 2018). Building on Castillo and Rousseau (2015), who laid down the general framework for semi-parametric BvMs, and on Polson and Rockova (2018), we formulate asymptotic normality for linear and quadratic functionals. Related semi-parametric BvM results have been established for density estimation (Rivoirard and Rousseau, 2012), Gaussian process priors (Castillo, 2012b,a), covariance matrix (Gao and Zhou, 2016) and tree/forest priors (Rockova, 2019). Our results provide new frequentist theoretical justifications for Bayesian deep learning inference with certain aspects of a regression function.

Our analysis focuses on sparse deep ReLU networks. Deep networks have been shown to outperform shallow ones in terms of representation power (Telgarsky, 2016), model complexity (Mhaskar et al., 2017) and generalization (Kawaguchi et al., 2017). The ReLU squashing function has been generally preferred due to its expressibility and inherent sparsity. For instance, Yarotsky (2017) provides error bounds for approximating polynomials and smooth functions with deep ReLU networks. Schmidt-Hieber (2017) showed that deep sparse ReLU networks can yield rate-optimal reconstructions of smooth functions and their compositions. Sparse architectures (in addition to ReLU) can reduce the test error. For example, sparsification can be achieved with dropout (Srivastava et al., 2014) which averages over sparse structures by randomly removing nodes and, thereby, alleviates overfitting. More recently, Polson and Rockova (2018) proposed Spike-and-Slab Deep Learning (SS-DL) as a fully Bayesian variant of dropout. Their framework provably does not overfit and achieves an *adaptive* near-minimax-rate optimal posterior concentration. Liu (2019) studies the BvM phenomena for the gradient function of Bayesian deep ReLU network and proposes a variable selection method based on the credible intervals. We continue the theoretical investigation of SS-DL in this paper.

Similar to Rockova (2019), we consider a non-parametric regression model where responses $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)'$ are linked to fixed covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in [0,1]^p$ for $i = 1, \ldots, n$ as follows

$$Y_i = f_0(\boldsymbol{x}_i) + \epsilon_i,\, \epsilon_i \overset{iid}{\sim} N(0,1), \tag{1}$$

where $f_0 \in \mathcal{H}_p^\alpha$ is an α-Hölder smooth function on a unit cube $[0,1]^p$ for some $\alpha > 0$. The true generative model implied by (1) will be denoted by $\mathbb{P}_0^n$. We want to reconstruct f_0 with $f \in \mathcal{F}$, where the model class $\mathcal{F}$ is assigned a prior distribution Π. Our goal is to study the asymptotic behavior of the posterior distribution

$$\Pi\left[\sqrt{n}(\Psi(f) - \hat{\Psi}) \,|\, \boldsymbol{Y}^{(n)}\right],$$

where $\Psi : \mathcal{F} \to \mathbb{R}$ is a measurable function of interest and where $\hat{\Psi}$ is a random centering point (see Theorem 2.1 in Castillo and Rousseau (2015)).

Two functionals are considered in our work. The first one is the linear functional

$$\Psi(f) = \frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{x}_i) f(\boldsymbol{x}_i), \tag{2}$$

with a constant weighting functions $a(\cdot)$. We discuss potential generalizations to the non-constant case later in Section 5 . The second functional of interest is the squared-L^2 norm

$$\Psi(f) = \|f\|_L^2\,, \tag{3}$$

where $\|f\|_L^2 = \frac{1}{n}\sum_{i=1}^{n}[f(\boldsymbol{x}_i)]^2$. Note that $\|\cdot\|_L$ corresponds to the LAN (locally asymptotically normal) norm, which is equivalent to the empirical L^2-norm $\|\cdot\|_n$ in our model. There is extensive literature on minimax estimation of linear and quadratic functionals, initiated in Ibragimov and Khasminskii (1985) and followed by Cai and Low (2005); Efromovich and Low (1996); Collier et al. (2017), to name a few. While the linear functional is useful for inference about the average regression surface, the quadratic functional is useful in many testing problems, including construction of confidence balls (Cai and Low, 2006) and goodness of fit tests (Dümbgen, 1998; Butucea, 2007). We study adaptive estimation of the two functionals from a Bayesian perspective.

First, we give the definition of asymptotic normality.

Definition 1.1. *Denote with β the bounded Lipschitz metric for weak convergence and with τ_n the mapping $\tau_n : f \to \sqrt{n}(\Psi(f) - \Psi_n)$. We say that the posterior distribution of the functional $\Psi(f)$ is asymptotically normal with centering Ψ_n and variance V if*

$$\beta(\Pi[\cdot \mid \boldsymbol{Y}^{(n)}] \circ \tau_n^{-1}, \mathcal{N}(0, V)) \to 0, \tag{4}$$

in $\mathbb{P}_0^n$-probability as $n \to \infty$. We will write this more compactly as $\Pi[\cdot \mid \boldsymbol{Y}^{(n)}] \circ \tau_n^{-1} \rightsquigarrow \mathcal{N}(0, V)$.

Next, we say that the posterior distribution *satisfies the BvM theorem* if (4) holds with $\Psi_n = \hat{\Psi} + o_P(\frac{1}{\sqrt{n}})$ for $\hat{\Psi}$ a linear efficient estimator of $\Psi(f_0)$.

Castillo and Rousseau (2015) provide general conditions on the model and on the function $\Psi(\cdot)$ to guarantee that the BvM phenomenon holds. Our results are built on the first-order approximation technique developed in their work. Essentially, we want to show that the sparse deep learning posterior can approximate both f_0 and the linear expansion term well enough so that the remainder term vanishes when $n \to \infty$.

The rest of our paper is organized as follows. Section 2 defines sparse ReLU networks and reviews the posterior concentration results. Section 3 contains the main results of BvM properties of the two functionals and Section 4 discusses extensions to adaptive priors. Section 5 concludes with a discussion.

2 Deep ReLU Networks

We follow the notation used in Polson and Rockova (2018). We denote with $\mathcal{F}(L, \mathbf{p}, s)$ the class of sparse ReLU networks with $L \in \mathbb{N}$ layers, a vector of $\mathbf{p} = (p_0, \ldots, p_{L+1})' \in \mathbb{N}^{L+2}$ hidden units and sparsity level $s \in \mathbb{N}$, which is the upper bound on the number of nonzero parameters. In our model, we have $p_0 = p$ and $p_{L+1} = 1$. Each function $f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) \in \mathcal{F}(L, \mathbf{p}, s)$ takes the form

$$f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) = W_{L+1}\sigma_{b_L}\left(W_L\sigma_{b_{L-1}} \cdots \sigma_{b_1}(W_1\boldsymbol{x})\right) + b_{L+1} \tag{5}$$

where $b_l \in \mathbb{R}^{p_l}$ are shift vectors and W_l are $p_l \times p_{l-1}$ weight matrices that link neurons between the $(l-1)^{th}$ and l^{th} layers and $\sigma_b(\boldsymbol{x})$ is the squashing function. Throughout this work, we assume the *rectified linear (ReLU)* function $\sigma_b(\boldsymbol{x}) = \max(\boldsymbol{x} + b, 0)$ which applies to vectors elementwise. Note that the top layer shift parameter b_{L+1} is *outside* the ReLU function since the top layer is only a linear function. We denote the sets of all model parameters with

$$\boldsymbol{B} = \{(W_1, b_1), \ldots, (W_L, b_L), (W_{L+1}, b_{L+1})\}. \tag{6}$$

Let $Z_l \in \mathbb{R}^{p_l}$ represent the hidden nodes of the l^{th} layer obtained as

$$Z_l(\boldsymbol{x}) = \sigma_{b_l}(W_l Z_{l-1}(\boldsymbol{x})), \quad \text{for} \quad l = 1 \ldots, L,$$
$$Z_0(\boldsymbol{x}) = \boldsymbol{x}.$$

We use $Z = \{Z_l\}_{l=1}^L$ to represent the collection of all hidden neurons. Their values are completely determined by $\{W_l, b_l\}_{l=1}^L$, independently of the top layer parameters $\{W_{L+1}, b_{L+1}\}$.

2.1 Spike-and-Slab Priors

We place a probabilistic structure on $\boldsymbol{B}$ that is slightly different from Polson and Rockova (2018). In particular, we remove the spike-and-slab prior on the top layer L to obtain a fully-connected top layer for each function $f_{\boldsymbol{B}}^{DL}(x)$. Such a relaxation on the top layer facilitates the *change of measure* step in our results. Later we show that having a fully connected top layer *does not* affect the network approximability and the posterior concentration rate.

We convert $\boldsymbol{B}$ into a vector by stacking $\{W_l, b_l\}_{l=1}^{L+1}$ from the bottom to the top and denote $\boldsymbol{B} = (\beta_1, \ldots, \beta_T)'$, where $T = \sum_{l=0}^{L} p_{l+1}(p_l + 1)$ is the number of parameters in a fully connected network with L layers and a vector of $\mathbf{p}$ neurons. Note that $\{\beta_j\}_{j > T-(p_L+1)}$ corresponds to the top layer $\{W_{L+1}, b_{L+1}\}$. Then the priors on $\boldsymbol{B}$ are

$$\pi(\beta_j \mid \gamma_j) = \gamma_j \tilde{\pi}(\beta_j) + (1 - \gamma_j)\delta_0(\beta_j), \tag{7}$$

with

$$\gamma_j = 1 \quad \text{for} \quad j > T - (p_L + 1), \tag{8}$$

where $\tilde{\pi}(\beta)$ is specified as

$$\tilde{\pi}(\beta_j) = \begin{cases} N(0, 1), & j > T - p_L + 1, \\ \text{Uniform}[-1, 1], & j \le T - p_L + 1, \end{cases} \tag{9}$$

i.e., the top layer weights follow standard normal distribution, while the deep weights follow uniform distribution on $[-1, 1]$. $\delta_0(\beta)$ is a dirac spike at zero, and $\gamma_j \in \{0, 1\}$ for whether or not β_j is nonzero. We let $\gamma_j = 1$ for all $j > T - (p_L + 1)$ so that the top layer is fully connected. The vector $\gamma = (\gamma_1, \ldots, \gamma_T)'$ encodes the connectivity pattern below the top layer. We assume that, given the network structure and the sparsity level $s = |\gamma| > p_L$, all architectures are equally likely a priori, i.e.

$$\pi(\gamma \mid \mathbf{p}, s) = \frac{\mathbb{I}(\gamma_j = 1 \text{ for } j > T - p_L - 1)}{\binom{T - p_L - 1}{s - p_L - 1}}. \tag{10}$$

We denote with $\mathcal{V}^{\mathbf{p}, s}$ the set of all combinatorial possibilities of connectivity patterns below the top layer. For a given sparsity level s, we can write

$$\mathcal{F}(L, \mathbf{p}, s) = \bigcup_{\gamma \in \mathcal{V}^{\mathbf{p}, s}} \mathcal{F}(L, \mathbf{p}, \gamma), \tag{11}$$

where each shell $\mathcal{F}(L, \mathbf{p}, \gamma)$ consists of all uniformly bounded functions $f_{\boldsymbol{B}}^{DL}$ with the same connectivity pattern γ, i.e. $\mathcal{F}(L, \mathbf{p}, \gamma) = \{f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) \in \mathcal{F}(L, \mathbf{p}, s) :$ $f_{\boldsymbol{B}}^{DL}(\boldsymbol{x})$ as in (5) with $\boldsymbol{B}$ arising from (7) for a given $\gamma \in \mathcal{V}^{\mathbf{p},s}$and where $\left\|f_{\boldsymbol{B}}^{DL}(\boldsymbol{x})\right\|_\infty < F\}$ for some $F > 0$.

Remark 2.1. *The prior for the deep coefficients β_j in (9) can be replaced by*

$$\tilde{\pi}(\beta_j) = N(0, 1), \forall j = 1, \dots, T. \tag{12}$$

The posterior concentration rate can be also shown to be rate-optimal under this prior. We give the sketch of the proof after Theorem 7.1 in Supplemental Material. Moreover, the BvM property for this prior can be immediately concluded from our proofs of Theorems 3.1-3.3.

2.2 A Connection between Deep ReLUs and Trees

Before proceeding, it will be useful to revisit a connection between networks and trees. Recall that any deep ReLU network function can be written as a sum of local linear functions, i.e.

$$f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{x} \in \Omega_k)(\tilde{\beta}_k' \boldsymbol{x} + \tilde{\alpha}_k), \tag{13}$$

where $\{\Omega_k\}_{k=1}^K$ is a partition of the predictor space made by recursive ReLU layers (see Polson and Sokolov (2017) for illustrations). Both the partition $\{\Omega_k\}_{k=1}^K$ and the coefficients of the local linear functions $\{\tilde{\beta}_k, \tilde{\alpha}_k\}_{k=1}^K$ are determined from $\{W_l, b_l\}_{l=1}^{L+1}$. We have omitted the dependence on $\boldsymbol{B}$ for simplicity of notation.

Balestriero and Baraniuk (2018) view ReLU as Max-Affine Spline Functions (MASO) and describe how the local linear functions and partitions are determined from weights $\boldsymbol{B}$. They point out that the partition by layer l contains up to 2^{p_l} convex conjoint regions. In practice, however, many of them could be empty intersections. Montufar et al. (2014) shows that the number of linear regions K of ReLU networks is upper-bounded by 2^T and lower-bounded by $(\prod_{l=1}^{L-1} \lfloor \frac{p_l}{p} \rfloor^p) \sum_{j=1}^{p} \binom{p_L}{j}$. Hanin and Rolnick (2019) further measure the volume of the boundaries between these regions.

Deep ReLU networks are similar to trees/forests methods in the sense that they also partition the predictor space. In fact, any regression tree can be represented by a neural network with a particular activation function, as we illustrate below using an example from Biau et al. (2016).

Example 1 Define an activation function $\tau_b : \mathbb{R} \to \{-1, 1\}$ such that

$$\tau_b(x) = 2\mathbb{I}_{x+b \geq 0} - 1.$$

We can reconstruct a two-dimensional ($p = 2$) example in Figure 1 with a neural network as

$$\begin{aligned} Z_1 &= \tau_{-b_1}(X_1) & Z_2 &= \tau_{-b_2}(X_2), \\ Z_3 &= \tau_{-2}(-Z_1 + Z_2) & Z_4 &= \tau_{-2}(Z_1 + Z_2), \\ Z_5 &= \tau_{-1}(Z_1) & f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) &= \sum_{i=3}^{5} W_i Z_i. \end{aligned}$$

where b_1 and b_2 set the decision boundaries along (X_1, X_2) axes in the tree, and $\{W_i\}_{i=3}^5$ are the jump sizes in each leaf node. A more detailed explanation of the choice of weights can be found in Biau et al. (2016). By analogy, the hierarchical segmentation is determined by the deep layers while the values of the leaf nodes are assigned by the top layer.

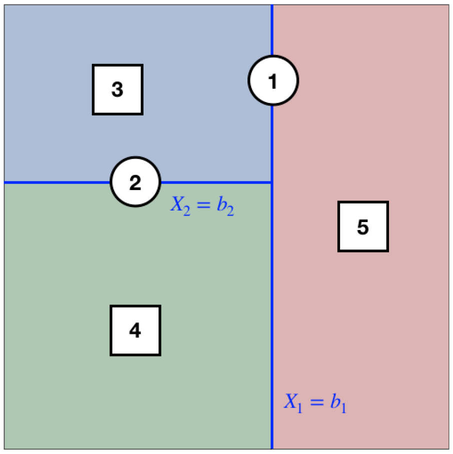

Figure 1: Visualization of Example 1

Deep ReLU networks use a different activation function and thereby place fewer restrictions on the geometry of the partition boundaries (shards as opposed to boxes). There are two aspects that make the analysis of deep ReLU networks more difficult. First, the partitioning lines do not align with coordinate axes when $W_l \neq 0$. Second, the partitioning cells $\{\Omega_k\}_{k=1}^K$ and the local linear coefficients $\{\tilde{\beta}_k, \tilde{\alpha}_k\}_{k=1}^K$ are related as they both depend on the unknown coefficients $\{W_l, b_l\}_{l=1}^L$. In tree models, on the other hand, they are independent parameters.

To illustrate the correspondence between the partitions and local linear functions as well as their relationship to $\boldsymbol{B}$, we consider the following toy example.

Example 2 Consider $L = 1, p = 2$ and $p_1 = 2$. Given the weights and shifts as

$$W_1 = \begin{pmatrix} W_1^1 \\ W_2^1 \end{pmatrix}, b_1 = (b_1^1, b_2^1), W_2 = \begin{pmatrix} W_1^2 \\ W_2^2 \end{pmatrix}, b_2 = b^2,$$

we can write the model as

$$Z_1 = \sigma_{b_1^1}(W_1^1 \boldsymbol{x}), \quad Z_2 = \sigma_{b_2^1}(W_2^1 \boldsymbol{x}),$$
$$f_{\boldsymbol{B}}^{DL}(\boldsymbol{x}) = \sigma_{b^2}(W_1^2 Z_1 + W_2^2 Z_2).$$

Then the corresponding $\{\tilde{\beta}_k, \tilde{\alpha}_k, \Omega_k\}_{k=1}^5$ for each local linear function can be organized as

i	$\tilde{\beta}_i$	$\tilde{\alpha}_i$	Ω_i
1	$W_1^2 W_1^1 + W_2^2 W_2^1$	$W_1^2 b_1^1 + W_2^2 b_2^1 + b^2$	$A_1 \cap A_2 \cap A_3$
2	$W_1^2 W_1^1$	$W_1^2 b_1^1 + b^2$	$A_1 \cap A_2^c \cap A_4$
3	$W_2^2 W_2^1$	$W_2^2 b_2^1 + b^2$	$A_1^c \cap A_2 \cap A_5$
4	0	$\max(b^2, 0)$	$A_1^c \cap A_2^c$
5	0	0	$(\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4)^c$

with

$$A_1 = \{\boldsymbol{x} : W_1^1 \boldsymbol{x} + b_1^1 > 0\}, \quad A_2 = \{\boldsymbol{x} : W_2^1 \boldsymbol{x} + b_2^1 > 0\},$$
$$A_3 = \{\boldsymbol{x} : \tilde{\beta}_1 \boldsymbol{x} + \tilde{\alpha}_1 > 0\}, \quad A_4 = \{\boldsymbol{x} : \tilde{\beta}_2 \boldsymbol{x} + \tilde{\alpha}_2 > 0\},$$
$$A_5 = \{\boldsymbol{x} : \tilde{\beta}_3 \boldsymbol{x} + \tilde{\alpha}_3 > 0\}.$$

Here we use A_i^c to denote the complement of set A_i, i.e., $A_i^c = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \notin A_i\}$. The covariance matrix of $\{\tilde{\beta}_k\}_{k=1}^3$ is

$$\mathsf{Var}\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \end{pmatrix} = \frac{2}{9}\begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

This example is plotted in Figure 2, where the boundaries of the partitions are nested according to $\{\tilde{\beta}_k, \tilde{\alpha}_k\}_{k=1}^5$ and determined by $\{W_l, b_l\}_{l=1}^2$.

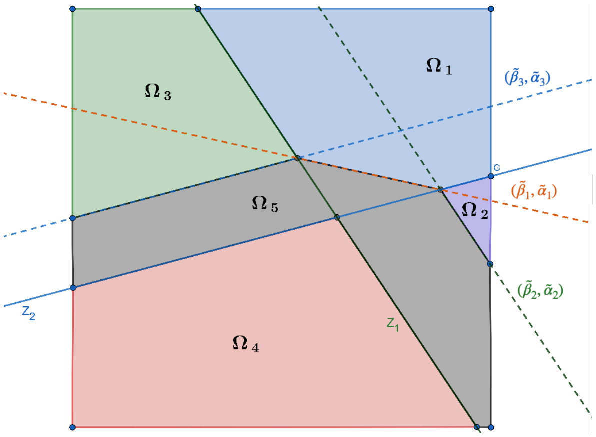
Ω_1
Ω_3
$(\tilde{\beta}_3, \tilde{\alpha}_3)$
G
$(\tilde{\beta}_1, \tilde{\alpha}_1)$
Ω_5
Ω_2
$(\tilde{\beta}_2, \tilde{\alpha}_2)$
Z_2
Z_1
Ω_4

Figure 2: Visualization of Example 2

2.3 Posterior Concentration

One essential prerequisite for our BvM analysis is optimal rate of posterior convergence. Polson and Rockova (2018) (PR18) showed that sparse deep ReLUs attain the near-minimax optimal rate and are *adaptive* to unknown smoothness under suitable priors on the architecture size. Here, we use a modified prior with a fully connected top linear layer (as given by (8)). The posterior concentration result still holds. Indeed, for an arbitrary sparse network, there exists at least one network with a fully connected linear layer that achieves the same approximation error. The approximability of our class of networks is thus the same as the class considered in PR18. We illustrate how such a network can be constructed in the Supplemental Material (Lemma 7.1).

Denoting (L^*, N^*, s^*) as in Theorem 5.1 of PR18 and choosing the parameters of the network as

$$\begin{cases} L = L^* + 1 \asymp \log(n), \\ s = s^* + 24pN^* \lesssim n^{p/(2\alpha+p)}, \end{cases} \tag{14}$$

we define

$$A_n^M = \{f_{\boldsymbol{B}}^{DL} \in \mathcal{F}(L, \mathbf{p}, s) : \left\|f_{\boldsymbol{B}}^{DL} - f_0\right\|_L \leq M\xi_n\} \tag{15}$$

with $\xi_n = n^{-\alpha/(2\alpha+p)} \log^\delta(n)$ for some $M > 0$ and $\delta > 0$. As we formalize in Theorem 7.1 in the Supplement, one can show $\Pi[A_n^{M_n} | \boldsymbol{Y}^{(n)}] = 1 + o_P(1)$ for any $M_n \to \infty$ and uniformly bounded α-Hölder mappings f_0.

Our analyses in Section 3 will be performed locally on sets $A_n^{M_n}$ where the posterior concentrates.

3 Semi-parametric BvM's

Locally on the sets $A_n \equiv A_n^{M_n}$ we will perform expansions of the log-likelihood as well as the functional Ψ. The log-likelihood is denoted with

$$\ell_n(f) = -\frac{n}{2}\log 2\pi - \sum_{i=1}^n \frac{[Y_i - f(\boldsymbol{x}_i)]^2}{2}.$$

and the log-likelihood ratio $\Delta_\ell(f) = \ell(f) - \ell(f_0)$ can be expressed as a sum of a quadratic term and a stochastic term via the LAN expansion as follows

$$\Delta_\ell(f) = -\frac{n}{2}\|f - f_0\|_L^2 + \sqrt{n} W_n(f - f_0)$$

where

$$\begin{aligned} W_n(f - f_0) &= \langle f - f_0, \sqrt{n}\boldsymbol{\epsilon}\rangle_L \\ &= \frac{1}{n}\sum_{i=1}^n \sqrt{n}\epsilon_i[f_0(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)]. \end{aligned}$$

We focus on the first-order approximations of the functionals. For any $f \in A_n$, we write

$$\Psi(f) = \Psi(f_0) + \langle \Psi_0^{(1)}, f - f_0 \rangle_L + r(f, f_0).$$

The first-order term $\Psi_0^{(1)}$ is equal to a for linear functionals (2) and $2f_0$ for the quadratic functional (3). The inner product $\langle \cdot, \cdot \rangle_L$ is defined as $\langle g, h \rangle_L = \frac{1}{n} \sum_{i=1}^n g(\boldsymbol{x}_i) h(\boldsymbol{x}_i)$ for two functions g and h.

Before we dive into the main development, we recall the results in Castillo and Rousseau (2015) which will be leveraged in our analysis.

There are two sufficient conditions for obtaining weak asymptotic normality as defined in (4). The first one is the vanishing remainder

$$\sup_{f \in A_n} \left| t\sqrt{n} r(f, f_0) \right| = o_P(1). \tag{16}$$

The second one is verifying

$$\frac{\int_{A_n} e^{\ell_n(f_t) - \ell_n(f_0)} d\Pi(f)}{\int_{A_n} e^{\ell_n(f) - \ell_n(f_0)} d\Pi(f)} = 1 + o_P(1), \forall t \in \mathbb{R}, \tag{17}$$

where $f_t = f - \frac{t\Psi_0^{(1)}}{\sqrt{n}}$.

The second condition in (17) can be shown with a *change of measure* argument and it guarantees that the posterior has no extra bias term. With these two conditions satisfied, the posterior behavior of $\sqrt{n}(\Psi(f) - \hat{\Psi})$ is asymptotically mean-zero normal with variance $V_0 = \left\| \Psi_0^{(1)} \right\|_L^2$, where

$$\hat{\Psi} = \Psi(f_0) + \frac{W_n(\Psi_0^{(1)})}{\sqrt{n}}$$

is a random centering point.

A crucial step is performing the *change of measure* in (17), where we replace f with a shifted function f_t in the integration. This is complicated by the fact that the shifted function f_t does not necessarily have to correspond to a deep ReLU network from the class $\mathcal{F}(L, \mathbf{p}, s)$. In the analysis of trees, for instance, one can condition on the partition parameter and perform the shift of measure on functions supported on the *same* partition, where the shift only affects step heights. For a deep ReLU network, however, partitions Ω_k and local linear coefficients $(\widetilde{b}_k, \widetilde{\alpha}_k)$ in (13) are not independent as they *both* depend on the deep weights $\{W_l, b_l\}_{l=1}^L$. It is thereby not obvious how the shift affects the partitions and the network coefficients. If we want to preserve the partitions of the predictor space, the only "free" parameters left to play with are the top layer weights $\{W_{L+1}, b_{L+1}\}$. Similarly as for trees, we consider conditioning on the *deep* coefficients $\{W_l, b_l\}_{l=1}^L$, which is equivalent to conditioning on γ and $Z = \{Z_l\}_{l=1}^L$, and perform the change of measure only on the top layer. We write the function class conditionally on (γ, Z) as

$$\mathcal{F}(L, \mathbf{p}, \gamma, Z) = \{f \in \mathcal{F}(L, \mathbf{p}, s) : f = W_{L+1} Z_L + b_{L+1} \text{ and } f \text{ has connectivity } \gamma\}. \tag{18}$$

Since the prior of $\{W_l, b_l\}_{l=1}^L$ is continuous, there are infinitely many (γ, Z)-dependent shells $\mathcal{F}(L, \mathbf{p}, \gamma, Z)$ inside $\mathcal{F}(L, \mathbf{p}, s)$. The general scheme of our proof is as follows. First, for each shell $\mathcal{F}(L, \mathbf{p}, \gamma, Z)$, we have a local centering point $\hat{\Psi}_Z^\gamma$ and a local variance V_Z^γ. Moreover, the shifted function f_t inside each shell lives on the *same partition* as f and the change of measure can therefore be performed more easily. Second, we show that $\hat{\Psi}_Z^\gamma$ and V_Z^γ converge *uniformly* to a global centering point $\hat{\Psi}$ and a global variance V_0 for all Z and γ inside A_n. This implies that we recover the global BvM on $\mathcal{F}(L, \mathbf{p}, s)$. The details of the local projections and the proof of all theorems are in Supplemental Material.

3.1 Linear Functionals

To start, we consider the linear functional in (2) where $a(\cdot)$ is a constant function in which case $\Psi(f)$ can be viewed as a constant multiple of the average regression surface evaluated at $\{\boldsymbol{x}_i\}_{i=1}^n$. Let

$$\Psi(f) = \Psi(f_0) + \langle a, f - f_0 \rangle_L, \quad \Psi_0^{(1)} = a,$$
$$\hat{\Psi} = \Psi(f_0) + \frac{W_n(a)}{\sqrt{n}}, \quad V_0 = \|a\|_L^2.$$

Theorem 3.1. *Assume the model (1), where f is endowed with a prior on $F(L, \mathbf{p}, s)$ defined in (7), (8) and (9). Assume that (14) is satisfied and that $f_0 \in \mathcal{H}_p^\alpha$, where $p = \mathcal{O}(1)$ as $n \to \infty$, $\alpha < p$ and $\|f_0\|_\infty \leq F$. When $a(\cdot)$ is constant, we have*

$$\Pi(\sqrt{n}(\Psi(f) - \hat{\Psi}) \mid \mathbf{Y}^{(n)}) \rightsquigarrow N(0, \|a\|_L^2)$$

in $\mathbb{P}_0^n$-probability as $n \to \infty$.

Proof. Reference to a Section 7.4 in Supplemental Material. When $a(\cdot)$ is constant, the shifted functions f_t can be easily constructed by shifting the top intercept $b_{L+1} \to b_{L+1} - \frac{ta}{\sqrt{n}}$. The projection of a is not needed as the remainder term is zero.

Remark 3.1. *When $a(\cdot)$ is not constant, we need the projection of $a(\cdot)$ (conditional on (γ, Z)), denoted by $a_{[Z]}^\gamma$, to be close to a for all Z and γ supported by A_n. In order for the BvM result to hold, we would then require the* no-bias *condition*

$$\langle a - a_{[Z]}^\gamma, f - f_0 \rangle_L = o_P\left(\frac{1}{\sqrt{n}}\right). \tag{19}$$

In order to verify this condition, one could view Z as a collection of random sparse ReLU features and study the approximability of this class. Although there are some studies on the universal approximation error of random ReLU features (Sun et al., 2019; Yehudai and Shamir, 2019), general conditions for the approximation ability of such projections are not yet obvious.

3.2 Squared L^2-norm Functional

We consider the quadratic functional in (3). The estimation of the L^2-norm is closely related to minimax optimal testing of hypothesis under empirical L^2 distance (Collier et al., 2017). This functional could serve as the risk function and has been used in many testing problems (Cai and Low, 2006; Dümbgen, 1998). The next theorem relies on the following notation

$$\Psi(f) = \Psi(f_0) + 2\langle f_0, f - f_0\rangle_L + \|f - f_0\|_L^2, \Psi_0^{(1)} = 2f_0,$$
$$\hat{\Psi} = \Psi(f_0) + \frac{2W_n(f_0)}{\sqrt{n}}, V_0 = 4\|f_0\|_L^2.$$

Theorem 3.2. *Assume the model (1), where f is endowed with a prior on $F(L, \mathbf{p}, s)$ defined in (7), (8) and (9). Assume that (14) is satisfied and that $f_0 \in \mathcal{H}_p^\alpha$, where $p = \mathcal{O}(1)$ as $n \to \infty$, $\alpha \in (\frac{p}{2}, p)$ and $\|f_0\|_\infty \leq F$. Then we have*

$$\Pi(\sqrt{n}(\Psi(f) - \hat{\Psi}) \mid \boldsymbol{Y}^{(n)}) \rightsquigarrow N(0, 4\|f_0\|_L^2)$$

in $\mathbb{P}_0^n$-probability as $n \to \infty$.

Proof. Reference to Section 7.5 in Supplemental Material. For this quadratic functional, we use the (γ, Z)-dependent projection $f_{0[Z]}^\gamma$ to approximate $\Psi_0^{(1)} = 2f_0$ so that the *change of measure* can be conducted through $\{W_{L+1}, b_{L+1}\}$. The additional constraint $\alpha > p/2$ is added to obtain $\xi_n^2 = o(\frac{1}{\sqrt{n}})$, which ensures that the remainder term (16) vanishes.

4 Adaptive Priors

The results in previous section are predicated on the assumption that the smoothness α is *known*. This is hardly ever satisfied in practice and the next natural step is to inquire whether similar conclusions can be obtained when α is unknown. Similarly as PR18, instead of the α-dependent choices of the width N and sparsity level s in (14), we deploy the following priors that adapt to smoothness

$$\pi(N) = \frac{\lambda^N}{(e^\lambda - 1)N!}, \text{ for } \lambda \in \mathbb{R}, \tag{20}$$

$$\pi(s) \propto e^{-\lambda_s s}, \text{ for } \lambda_s > 0. \tag{21}$$

The parameter space now consists of shells of sparse ReLU networks with different widths and sparsity levels, i.e.

$$\mathcal{F}(L) = \bigcup_{N=1}^{\infty} \bigcup_{s=0}^{T} \mathcal{F}(L, \mathbf{p}_N^L, s), \tag{22}$$

where $\mathcal{F}(L, \mathbf{p}_N^L, s)$ was defined in (11). An approximating sieve can be constructed that consists of sparse and not so wide networks, i.e.

$$\mathcal{F}_n = \bigcup_{N=1}^{N_n} \bigcup_{s=0}^{s_n} \mathcal{F}(L, \mathbf{p}_N^L, s) \tag{23}$$

with $N_n \asymp n\xi_n^2/\log n$ and $s_n \asymp n\xi_n^2$.

Following the same strategy as in the proof Theorem 6.2 of PR18, we extend the posterior concentration result to the case of adaptive priors (7), (8), (20) and (21) (see Theorem 7.2 in the Supplemental Material). The next step is extending the BvM results from the previous section. The following Theorem shows that one can obtain asymptotic normality of the quadratic and linear functionals without the exact knowledge of α.

Theorem 4.1. *Assume the model (1), where f is endowed with a prior on $F(L)$ defined through (7), (8), (9), (20) and (21) with $L \asymp \log(n)$. Assume that $f_0 \in \mathcal{H}_p^\alpha$, where $p = \mathcal{O}(1)$ as $n \to \infty, \alpha < p$ and $\|f_0\|_\infty \leq F$.*

(i) For the linear functional $\Psi(f)$ in (2) where $a(\cdot)$ is constant, we obtain

$$\Pi(\sqrt{n}(\Psi(f) - \hat{\Psi}) \mid \boldsymbol{Y}^{(n)}) \rightsquigarrow N(0, \|a\|_L^2),$$

where $\hat{\Psi} = \Psi(f_0) + \frac{1}{\sqrt{n}}W_n(a)$.

(ii) For the square L^2-norm functional $\Psi(f)$ in (3), we obtain for $\alpha \in (\frac{p}{2}, p)$

$$\Pi(\sqrt{n}(\Psi(f) - \hat{\Psi}) \mid \boldsymbol{Y}^{(n)}) \rightsquigarrow N(0, 4\|f_0\|_L^2)$$

where $\hat{\Psi} = \Psi(f_0) + \frac{2}{\sqrt{n}}W_n(f_0)$.

Proof. Reference to Section 7.6 in Supplemental Material.

Remark 4.1. *Similar constraints on the smoothness α have been imposed in other related works (Farrell et al., 2018). However, unlike in other developments (Schmidt-Hieber, 2017; Farrell et al., 2018), the convergence rates we build on are* adaptive *in the sense that, beyond the assumption $\alpha < p$, the exact knowledge of α is* not required. *When the imposed smoothness assumptions do not hold, one could still obtain* *asymptotic normality via misspecified BvM-type results (Kleijn and Van der Vaart, 2012) but uncertainty quantification with the implied credible sets would be problematic.*

Remark 4.2. *It is worth noting that our results do not hinge on the assumption that f_0 came from the prior. Instead, f_0 is an arbitrary Hölder smooth function, not necessarily a neural network. While the model is ultimately mis-specified, our results are attainable due to the expressibility of deep ReLU networks where one can approximate f_0 with deep learning mappings with a rapidly vanishing error. The fact that our posterior concentrates around the truth at the optimal rate makes the derivation of BvM and valid inference feasible.*

5 Discussion

In this paper, we obtained asymptotic normality results for linear and squared L^2-norm functionals for deep, sparse ReLU networks. These results can be used as a basis for semi-parametric inference and can be extended in various ways.

First, one could obtain similar formulations for general smooth linear functionals by verifying the *no bias* condition in (19). This relates to the approximation ability of random ReLU features mentioned in Remark 3.1. The ReLU features act similarly as random rotational trees. However, the nested nature of partitions and local linear functions make the analysis difficult. Random features have gained much attention recently. For instance, Rahimi and Recht (2008) show how random features can be connected to kernel methods. Sun et al. (2019) discuss the universal approximation bounds for compositional ReLU features. Huang et al. (2006) and Huang (2014) provide similar results and they propose an implementation of the extreme learning machine implementation, where only the top layer is trained while deep layers are sampled randomly from some distribution. A time-series variant of this algorithm is the Deep Echo State Network (Sun et al., 2017; McDermott and Wikle, 2019).

Another way to obtain BvM for smooth linear functionals would be to construct a less-restrictive projection of the first-order term $\Psi_0^{(1)}$. Schmidt-Hieber (2017) shows that parallelization can be realized using embedding networks. The shifted function f_t could be constructed as an embedding network that simultaneously represents $(f, \Psi_0^{(1)})$. This representation could leverage the approximability of smooth functions a with deep neural networks.

To sum up, our semi-parametric BvM results certify that (semi-parametric) inference with Bayesian deep learning is valid and that meaningful uncertainty quantification is attainable. Possible applications of our results include casual inference, whereby embedding our model within a missing data framework (Ray and van der Vaart, 2018), the average functional can be used for average treatment effect estimation. In this vein, our results are relevant for the development/understanding of the widely sought after machine learning methods for causal inference (Athey and Wager, 2017). In particular, an extension of our work along these lines will constitute a fully-Bayesian variant of the doubly-robust plug-in approach of Farrell et al. (2018). In addition, the main theorems (Theorem 3.1-3) provide foundations for testing hypotheses such as exceedance of a level $\sum_{i=1}^n f_0(x_i) > c$. Lastly, an important future direction will be quantifying uncertainty about the *entire function* f_0 (not only its functionals), which was recently formalized for Bayesian CART by Castillo and Rockova (2019).

Our work is primarily concerned with theoretical frequentist study of the posterior distribution. Investigating practical usefulness and computation of our priors is an important future direction. There are various ways to approximate aspects of deep learning posterior distributions under spike-and-slab prior, see Polson and Rockova (2018) for a discussion on possible implementations. In addition, Deng et al. (2019) proposed an adaptive empirical Bayesian method for sparse deep learning with a self-adaptive spike-and-slab prior.

Acknowledgements

The authors gratefully acknowledge support from the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

References

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Athey, S. and Wager, S. (2017). Efficient policy learning. *arXiv*.

Balestriero, R. and Baraniuk, R. G. (2018). A spline theory of deep learning. In *International Conference on Machine Learning*, pages 374–383.

Biau, G., Scornet, E., and Welbl, J. (2016). Neural random forests. *Sankhya A*, pages 1–40.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *International Conference on Machine Learning*, volume 37, pages 1613–1622. JMLR. org.

Butucea, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930.

Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343.

Cai, T. T. and Low, M. G. (2006). Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228.

Castillo, I. (2012a). Semiparametric Bernstein-von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A*, 74(2):194–221.

Castillo, I. (2012b). A semiparametric Bernstein-von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1-2):53–99.

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

Castillo, I. and Rockova, V. (2019). Multiscale analysis of Bayesian CART. *Submitted*, pages 1–75.

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

Collier, O., Comminges, L., and Tsybakov, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Deng, W., Zhang, X., Liang, F., and Lin, G. (2019). An adaptive empirical Bayesian method for sparse deep learning. In *Advances in Neural Information Processing Systems*, pages 5564–5574.

Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics*, 26(1):288–314.

Edmunds, D. E. and Triebel, H. (2008). *Function spaces, entropy numbers, differential operators*, volume 120. Cambridge University Press.

Efromovich, S. and Low, M. G. (1996). On optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 24(3):1106–1125.

Farrell, M. H., Liang, T., and Misra, S. (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059.

Gao, C. and Zhou, H. H. (2016). Bernstein-von Mises theorems for functionals of the covariance matrix. *Electronic Journal of Statistics*, 10(2):1751–1806.

Ghosal, S. and Van Der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223.

Hanin, B. and Rolnick, D. (2019). Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, pages 2596–2604.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29.

Huang, G.-B. (2014). An insight into extreme learning machines: random neurons, random features and kernels. *Cognitive Computation*, 6(3):376–390.

Huang, G.-B., Chen, L., and Siew, C. K. (2006). Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17(4):879–892.

Ibragimov, I. A. and Khasminskii, R. Z. (1985). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2017). Generalization in deep learning. *arXiv*.

Kleijn, B. J. K. and Van der Vaart, A. W. (2012). The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing systems*, pages 1097–1105.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

Liu, J. Z. (2019). Variable selection with rigorous uncertainty quantification using deep Bayesian neural networks: Posterior concentration and Bernstein-von Mises phenomenon. *arXiv*.

McDermott, P. L. and Wikle, C. K. (2019). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, 30(3):e2553.

Mhaskar, H., Liao, Q., and Poggio, T. (2017). When and why are deep networks better than shallow ones? In *AAAI*, pages 2343–2349.

Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2924–2932.

Neal, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems*, pages 475–482.

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

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, pages 4026–4034.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519.

Polson, N. and Rockova, V. (2018). Posterior concentration for sparse deep learning. In *Advances in Neural Information Processing Systems*, pages 938–949.

Polson, N. G. and Sokolov, V. (2017). Deep learning: a Bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184.

Ray, K. and van der Vaart, A. (2018). Semiparametric Bayesian causal inference using Gaussian process priors. *arXiv*.

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

Rockova, V. (2019). On semi-parametric Bernstein-von Mises theorems for BART. *arXiv*.

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. *arXiv*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 4134–4142.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sun, X., Li, T., Li, Q., Huang, Y., and Li, Y. (2017). Deep belief echo-state network and its application to time series prediction. *Knowledge-Based Systems*, 130:17–29.

Sun, Y., Gilbert, A., and Tewari, A. (2019). On the approximation capabilities of ReLU neural networks and random ReLU features. *arXiv*.

Telgarsky, M. (2016). Benefits of depth in neural networks. In *Conference on Learning Theory*, pages 1517–1539.

Ullrich, K., Meeds, E., and Welling, M. (2017). Soft weight-sharing for neural network compression. In *International Conference on Learning Representation*.

Vitushkin, A. G. (1964). A proof of the existence of analytic functions of several variables not representable by linear superpositions of continuously differentiable functions of fewer variables. In *Doklady Akademii Nauk*, volume 156, pages 1258–1261. Russian Academy of Sciences.

Wang, Y., Polson, N. G., and Sokolov, V. O. (2019). Scalable data augmentation for deep learning. *arXiv*.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114.

Yehudai, G. and Shamir, O. (2019). On the power and limitations of random features for understanding neural networks. *arXiv*.