Journal of Machine Learning Research 26 (2025) 1-60 Submitted 3/24; Published 5/25

Posterior Concentrations of Fully-Connected Bayesian Neural Networks with General Priors on the Weights

Insung Kong GGONG369@SNU.AC.KR
Department of Statistics
Seoul National University
Seoul, 08826, South Korea

Yongdai Kim YDKIM0903@GMAIL.COM
Department of Statistics
Seoul National University
Seoul, 08826, South Korea

Editor: Daniel Roy

Abstract

Bayesian approaches for training deep neural networks (BNNs) have received significant interest and have been effectively utilized in a wide range of applications. Several studies have examined the properties of posterior concentrations in BNNs. However, most of these studies focus solely on BNN models with sparse or heavy-tailed priors. Surprisingly, there are currently no theoretical results for BNNs using Gaussian priors, which are the most commonly used in practice. The lack of theory arises from the absence of approximation results of Deep Neural Networks (DNNs) that are non-sparse and have bounded parameters. In this paper, we present a new approximation theory for non-sparse DNNs with bounded parameters. Additionally, based on the approximation theory, we show that BNNs with non-sparse general priors can achieve near-minimax optimal posterior concentration rates around the true model.

Keywords: Bayesian neural networks, posterior concentration rate, Bayesian nonparametric regression, approximation theory, deep neural networks

1. Introduction

Bayesian Neural Networks (BNNs) (MacKay, 1992; Neal, 2012), a framework for training Deep Neural Networks (DNNs) through Bayesian techniques, have garnered significant attention in the field of machine learning and AI. The distinctive feature of BNNs lies in their ability to combine the flexibility of DNNs and the probabilistic reasoning of Bayesian approaches. This combination results in DNNs that not only yield superior generalization capability across various tasks but also provide improved uncertainty quantification (Wilson and Izmailov, 2020; Izmailov et al., 2021). This attribute is particularly vital in applications where decision-making under uncertainty is crucial. Representative examples of such applications are recommender systems (Wang et al., 2015), computer vision (Kendall and Gal, 2017), active learning (Tran et al., 2019), medicine (Beker et al., 2020) and astrophysics (Cranmer et al., 2021), to name just a few.

The remarkable success of BNNs can be attributed to the inherent capacity of DNNs to automatically learn features from data even if they are parametric models (Wang and Yeung, 2020; Jospin et al., 2022). By leveraging this advantage, flexible BNN models can be devised easily that can handle complex predictive tasks data-adaptively without explicit model specification. This means that even in scenarios where users lack detailed knowledge about the functional relationship between inputs and outputs, BNNs are capable of uncovering intricate patterns and relationships existing in data.

There have been vast amounts of literature that attempt to understanding theoretical properties of BNNs from the nonparametric regression standpoint (Polson and Ročková, 2018; Chérief-Abdellatif, 2020; Bai et al., 2020; Liu, 2021; Sun et al., 2022; Lee and Lee, 2022; Jantre et al., 2023; Kong et al., 2023; Ohn and Lin, 2024). Rather than presuming the true function to be confined within a specific parametric model, these studies make the broader assumption that the true function belongs to a certain functional space, such as the Hölder function class. Notably, Polson and Ročková (2018); Chérief-Abdellatif (2020); Bai et al. (2020); Sun et al. (2022); Lee and Lee (2022); Kong et al. (2023); Ohn and Lin (2024) show that posterior distribution of BNNs concentrate around the true function with near-minimax optimal rates with respect to the sample size when the architecture and the prior on the weights and biases are selected carefully. These results demonstrate that even when the exact form of the data-generating process is unknown or too complex to be captured by traditional parametric models, BNNs possess the capacity for effective generalization, enabling them to learn these underlying patterns efficiently.

However, there is an important limitation in the existing results. That is, the priors on the weights and biases of DNNs do not include those that are commonly used in practice. For example, Polson and Ročková (2018); Chérief-Abdellatif (2020); Bai et al. (2020); Sun et al. (2022); Lee and Lee (2022) consider spike-and-slab priors but significant amounts of additional exploration times for searching a sparsity patterns become a significant obstacle in their practical use. Ohn and Lin (2024) derives the posterior concentration rate of non-sparse BNNs but uses the uniform distribution on the weights whose domain diverges as the sample size increases. It would not be easy to select the optimal size of the domain for given finite data which prevents the Bayesian model of Ohn and Lin (2024) from being popularly used in practice. Kong et al. (2023) considers polynomial-tail priors, but these priors make the calculation of the gradient computationally demanding and thus the development of a computationally efficient MCMC algorithm be difficult.

Surprisingly, there is no theoretical result about BNNs with i.i.d. standard Gaussian priors on the weights and biases, which are most popular priors for BNNs in practice (Fortuin et al., 2022; Jospin et al., 2022). That is, there exists a significant gap between theories and applications. A main reason for the lack of optimal concentration rates of BNNs using Gaussian priors is the absence of an approximation theory of fully-connected DNNs with bounded parameters. Existing approximation theories of DNNs require the weights to be either sparse (Suzuki, 2018; Schmidt-Hieber, 2019; Imaizumi and Fukumizu, 2019; Bauer and Kohler, 2019; Ohn and Kim, 2019; Schmidt-Hieber, 2020; Nakada and Imaizumi, 2020; Kohler et al., 2022; Chen et al., 2022) or unbounded (Kohler and Langer, 2021a; Lu et al., 2021; Jiao et al., 2023).

The aim of this paper is to fill this gap by deriving near-minimax optimal concentration rates of the posterior distributions of BNNs with a class of general priors on the weights including independent Gaussian priors. To achieve this aim, we develop a new technique to approximate the Hölder functions by fully connected DNNs with bounded weights, which offers several advantages: (1) Compared to Schmidt-Hieber (2020), our results allow fully connected DNNs, enabling their application to BNNs without the need for sparse-inducing priors to control model complexity. (2) In contrast to Kohler and Langer (2021a), our results can employ DNNs with bounded parameters, allowing their application to BNNs without resorting to heavy-tailed priors. (3) The ReLU activation function (Nair and Hinton, 2010) is extended to the Leaky-ReLU activation function (Maas et al., 2013), which are known for their optimization merits (Xu et al., 2015).

Based on our new approximation results, we demonstrate that the posterior distributions of fully-connected BNNs with a certain class of priors concentrate around the true function with near-minimax optimal rates. Notably, the assumed conditions on the priors hold for most commonly used prior distributions for BNNs, including the independent Gaussian distribution which has not been covered by existing theories. That is, our results successively fill the important gap between existing theories and applications in BNNs.

2. Preliminaries

2.1 Notation

We write $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$ and $\mathbb{R}^+ := \{x \in \mathbb{R} : x > 0\}$. For an integer $n \in \mathbb{N}$, we denote $[n] := \{1, \dots, n\}$. A capital letter denotes a random variable or matrix interchangeably whenever its meaning is clear. A vector is denoted by a bold letter, and its elements are denoted by regular letters with superscript indices. e.g. $\boldsymbol{x} := (x^{(1)}, \dots, x^{(d)})^\top$. For a d-dimensional vector $\boldsymbol{x} \in \mathbb{R}^d$, we denote $|\boldsymbol{x}|_p := (\sum_{j=1}^d |x^{(j)}|^p)^{1/p}$ for $1 \le p < \infty$, $|\boldsymbol{x}|_0 := \sum_{j=1}^d \mathbb{I}(x^{(j)} \neq 0)$ and $|\boldsymbol{x}|_\infty := \max_{j\in[d]} |x^{(j)}|$. For two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the same dimension, $\max(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is defined elementwise. For a real-valued function $f : \mathcal{X} \to \mathbb{R}$ and $1 \le p < \infty$, we denote $\|f\|_{p,n} := (\sum_{i=1}^n f(\boldsymbol{x}_i)^p/n)^{1/p}$ and $\|f\|_{p,\mathrm{P}_{\boldsymbol{X}}} := \left(\int_{\boldsymbol{X}\in\mathcal{X}} f(\boldsymbol{X})^p d\mathrm{P}_{\boldsymbol{X}}\right)^{1/p}$, where $\mathrm{P}_{\boldsymbol{X}}$ is a probability measure defined on the input space $\mathcal{X}$. We assume $\mathcal{X} \subseteq [-a, a]^d$ for some $a \ge 1$. Also, we define $\|f\|_\infty := \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})|$. For $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, we denote $\partial^{\boldsymbol{\alpha}} := \partial^{\alpha_1} \dots \partial^{\alpha_d}$. We denote $\circ$ as the composition of functions. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we denote $a_n \lesssim b_n$ if there exists a positive sequence $C > 0$ such that $a_n \le Cb_n$ for all $n \in \mathbb{N}$. We denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We use the little o notation, that is, we write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

Let $\beta = q + s$ for some $q \in \mathbb{N}_0$ and $s \in (0, 1]$. We consider the β-Hölder class $\mathcal{H}_d^\beta(K)$ for the class where the true function belongs, which is defined as

$$\mathcal{H}_d^\beta(K) := \left\{f : [-a, a]^d \to \mathbb{R}; \|f\|_{\mathcal{H}^\beta} \le K\right\},$$

where $\|f\|_{\mathcal{H}^\beta}$ denotes the Hölder norm defined by

$$\|f\|_{\mathcal{H}^\beta} := \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1\le q} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\substack{\boldsymbol{\alpha}\in\mathbb{N}_0^d \\ \boldsymbol{\alpha}:|\boldsymbol{\alpha}|_1=q}} \sup_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2\in[-a,a]^d \\ \boldsymbol{x}_1\neq\boldsymbol{x}_2}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_1) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}_2)|}{|\boldsymbol{x}_1 - \boldsymbol{x}_2|_\infty^s}.$$

2.2 Deep Neural Networks

For a depth $L \in \mathbb{N}$ and width $\boldsymbol{r} = (r^{(0)}, r^{(1)}, ..., r^{(L)}, r^{(L+1)})^\top \in \mathbb{N}^{L+2}$ where $r^{(0)} = d$ and $r^{(L+1)} = 1$, Deep Neural Network (DNN) with the $(L, \boldsymbol{r})$ architecture is defined as a DNN model which has L hidden layers and $r^{(l)}$ many neurons at the l-th hidden layer for $l \in [L]$. The output of the DNN model can be written as

$$f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}}(\cdot) := A_{L+1} \circ \boldsymbol{\rho} \circ A_L \cdots \circ \boldsymbol{\rho} \circ A_1(\cdot), \tag{1}$$

where $A_l : \mathbb{R}^{r^{(l-1)}} \mapsto \mathbb{R}^{r^{(l)}}$ for $l \in [L+1]$ is an affine map defined as $A_l(\boldsymbol{x}) := W_l\boldsymbol{x} + \boldsymbol{b}_l$ with $W_l \in \mathbb{R}^{r^{(l)} \times r^{(l-1)}}$ and $\boldsymbol{b}_l \in \mathbb{R}^{r^{(l)}}$ and $\boldsymbol{\rho}$ is an activation function. Here, $\boldsymbol{\theta} := (\boldsymbol{\theta}_w^\top, \boldsymbol{\theta}_b^\top)^\top$ is the concatenation of the parameters of the DNN model, where

$$\begin{aligned}\boldsymbol{\theta}_w &:= (\mathrm{vec}(W_1)^\top, \ldots, \mathrm{vec}(W_{L+1})^\top)^\top,\\ \boldsymbol{\theta}_b &:= (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_{L+1}^\top)^\top\end{aligned}$$

are the concatenation of the weight matrices and bias vectors. We denote J as the dimension of $\boldsymbol{\theta}$, i.e.,

$$J := J(L, \boldsymbol{r}) = \sum_{l=1}^{L+1} (r^{(l-1)} + 1) r^{(l)}.$$

The standard choice for the activation function $\boldsymbol{\rho}$ is the Rectified linear unit (ReLU) activation function (Nair and Hinton, 2010), which is defined as

$$\boldsymbol{\rho}_0(\boldsymbol{x}) = \max\{\boldsymbol{x}, \boldsymbol{0}\}.$$

The ReLU activation function is known to alleviate the vanishing gradient problem compared to the sigmoid or tanh activation functions, enabling efficient gradient propagation and enhancing DNN performance (Goodfellow et al., 2016).

As assumed in other papers for simplicity (Kohler and Langer, 2021a), we consider DNN architectures whose numbers of neurons in each hidden layer are the same. For a given activation function $\boldsymbol{\rho}$, the number of hidden layers $L \in \mathbb{N}$ and the number of neurons in each hidden layer $r \in \mathbb{N}$, we define the set of DNN functions which are parameterized by $\boldsymbol{\theta} \in \mathbb{R}^J$ as below.

Definition 1. *For an activation function $\boldsymbol{\rho}$, depth $L \in \mathbb{R}$ and width $r \in \mathbb{N}$, we define $\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r)$ as the function class of DNNs with the $(L, (d, r, \ldots, r, 1)^\top)$ architecture and the activation function $\boldsymbol{\rho}$. That is,*

$$\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r) := \left\{ f : f = f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}} \textit{ is a DNN with the } (L, (d, r, \ldots, r, 1)^\top) \textit{ architecture} \right\}.$$

Also, for $B \geq 1$, we define $\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r, B)$ as the subset of $\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r)$ consisting of DNNs whose parameter values lie within the absolute bound B. That is,

$$\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r, B) := \Bigg\{ f : f = f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}} \textit{ is a DNN with the } (L, (d, r, \ldots, r, 1)^\top) \\ \textit{architecture, } |\boldsymbol{\theta}|_\infty \leq B \Bigg\}.$$

In addition, for a sparsity $S \in [J]$, we define $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L, r, S, B)$ as the subset of $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L, r, B)$ consisting of DNNs whose number of non-zero parameters is bounded by S. That is,

$$\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L, r, S, B) := \Bigg\{ f : f = f_{\boldsymbol{\theta},\boldsymbol{\rho}}^{\mathrm{DNN}} \text{ is a DNN with the } (L, (d, r, \ldots, r, 1)^\top) \\ \text{architecture, } |\boldsymbol{\theta}|_0 \leq S, |\boldsymbol{\theta}|_\infty \leq B \Bigg\}.$$

For an any activation function $\boldsymbol{\rho}$, depths $L_1 \leq L_2$, widths $r_1 \leq r_2$, sparsity $S_1 \leq S_2$ and $B_1 \leq B_2$, we have $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_2, r_2)$, $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1, B_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_2, r_2, B_2)$ and $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L_1, r_1, S_1, B_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L_2, r_2, S_2, B_2)$ due to the enlarging property of DNNs. Also, by definition, we have $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L_1, r_1, S_1, B_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1, B_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1)$. In addition, if we denote J_1 as the total number of parameters in $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1, B_1)$, we have $\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_1, r_1, B_1) \subseteq \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{SDNN}}(L_2, r_2, J_1, B_1)$.

2.3 Approximation Results for DNNs

Analysis of nonparametric regression using neural networks has been developed over the years. Leshno et al. (1993) and Barron (1993) develop universal approximation properties of shallow neural networks. While shallow neural networks can approximate functions well, Montufar et al. (2014) and Eldan and Shamir (2016) claim that the expressive power of DNNs grows exponentially with the number of layers. Approximation errors of sparse DNNs with the ReLU activation function have been derived for β-times differentiable functions (Yarotsky, 2017) and piecewise smooth functions (Petersen and Voigtlaender, 2018). Schmidt-Hieber (2020) demonstrates that least square estimators based on sparsely connected DNNs with the ReLU activation function and properly chosen architectures achieve near-minimax optimal convergence rates. Similar results for sparse DNNs can be found in Suzuki (2018); Imaizumi and Fukumizu (2019); Bauer and Kohler (2019); Ohn and Kim (2019); Schmidt-Hieber (2019); Nakada and Imaizumi (2020); Kohler et al. (2022); Chen et al. (2022).

These optimal results rely heavily on sparsity constraints on DNNs. This is because the class of fully-connected DNNs is too large to yield optimal results. For instance, Schmidt-Hieber (2020) uses the approximation theorem that there exist positive constants $C_1^{(s)}$, $C_2^{(s)}$, $C_3^{(s)}$ and $c^{(s)}$ such that for every $f_0 \in \mathcal{H}_d^\beta(K)$ and any sufficiently large $M \in \mathbb{N}$, there exists

$$f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_0}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_0}^{\mathrm{SDNN}}\left(\left\lceil C_1^{(s)} \log_2 M \right\rceil, \left\lceil C_2^{(s)} M^{2d} \right\rceil, C_3^{(s)} M^{2d} \log_2 M, 1 \right)$$

with $\|f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_0}^{\mathrm{DNN}} - f_0\|_{\infty,[-a,a]^d} \leq c^{(s)} M^{-2\beta}$. Indeed, their approximation theorem implies that a fully connected DNN in $\mathcal{F}_{\boldsymbol{\rho}_0}^{\mathrm{DNN}}(\lceil C_1^{(s)} \log_2 M \rceil, \lceil C_2^{(s)} M^{2d} \rceil, 1)$ approximates f_0 well, but the complexity of $\mathcal{F}_{\boldsymbol{\rho}_0}^{\mathrm{DNN}}(\lceil C_1^{(s)} \log_2 M \rceil, \lceil C_2^{(s)} M^{2d} \rceil, 1)$ is too large to use, and so they consider sparsity constraints to reduce complexity.

Kohler and Langer (2021a) show that least squares estimators based on fully connected DNNs with the ReLU activation function also achieves near-minimax optimal convergence rates. They use a new approximation theorem that there exist positive constants $C_1^{(k)}$, $C_2^{(k)}$ and $c^{(k)}$ such that for every $f_0 \in \mathcal{H}_d^\beta(K)$ and any sufficiently large $M \in \mathbb{N}$, there exists

$$f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_0}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_0}^{\mathrm{DNN}} \left(\left\lceil C_1^{(k)} \log_2 M \right\rceil , \left\lceil C_2^{(k)} M^d \right\rceil \right)$$

with $\| f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_0}^{\mathrm{DNN}} - f_0 \|_{\infty,[-a,a]^d} \leq c^{(k)} M^{-2\beta}$. A significant advantage of the DNN approximation theorem proposed by Kohler and Langer (2021a) is that it does not require a sparsity condition. However, their DNN approximation has the drawback of unbounded parameter sizes. By following their proof, it can be confirmed that parameters are bounded by M^2, and this bound should diverge to reduce the approximation error. The DNN approximations of Lu et al. (2021) and Jiao et al. (2023) also face the issue of unbounded parameters. This point limits their applications in some areas, such as Bayesian analysis.

2.4 Posterior Concentration Results for BNNs

Posterior concentration, which is an asymptotic behavior of posterior distributions as the sample size increases, is crucial for the frequentist justification of Bayesian methods. Concentration rate is defined by the rate at which there exists a neighborhoods of the true model shrinking meanwhile still capturing most of the posterior mass (Ghosal et al., 2000). One of the key components for deriving the posterior concentration rate is the prior concentration condition, which is the requirement of a sufficient amount of prior mass assigned to a shrinking neighborhood of the true model (Ghosal and van der Vaart, 2017).

For BNNs, the prior concentration condition is usually established by the two steps: firstly choosing a DNN that approximates the true function well and then secondly devising a prior which puts sufficient masses around this DNN. Thus, when the approximation results of DNNs require sparsity assumptions (Suzuki, 2018; Imaizumi and Fukumizu, 2019; Bauer and Kohler, 2019; Ohn and Kim, 2019; Schmidt-Hieber, 2020), sparse-inducing priors (Polson and Ročková, 2018; Chérief-Abdellatif, 2020; Bai et al., 2020; Sun et al., 2022; Lee and Lee, 2022) are inevitably required to make the posterior concentration rate be optimal. Specifically, Polson and Ročková (2018) proves that posterior distributions of BNNs using spike-and-slab priors concentrate around the true function at near-minimax optimal rates on the Hölder spaces. This result has been extended to variational posterior distributions (Chérief-Abdellatif, 2020; Bai et al., 2020), continuous relaxation of spike-and-slab priors (Sun et al., 2022) and the case where the true functions belong to the Besov spaces (Lee and Lee, 2022).

Following the recent results of approximation using non-sparse DNNs (Kohler and Langer, 2021a), efforts have been made to show that BNNs with non-sparse priors can also achieve optimal posterior concentration rates (Kong et al., 2023; Ohn and Lin, 2024). However, to achieve the optimal posterior concentration rates using the approximation results of Kohler and Langer (2021a), heavy-tailed prior distributions should be used. For example, Kong et al. (2023) employs polynomial tail distributions such as the Cauchy distribution as the prior distribution, while Ohn and Lin (2024) employs uniform distributions defined on a diverging range.

So far, no approximation results for non-sparse short-tailed priors to achieve the optimal posterior concentration rates are available. Due to this theoretical shortcoming, there is no result available about the optimal posterior concentration rates of non-sparse BNNs with standard not-heavy-tailed priors such as independent Gaussian distributions.

3. Approximation Using Fully-Connected DNNs with Bounded Parameters

In this section, we develop a new approximation result using fully-connected DNNs with bounded parameters. For activation functions in DNNs, we consider the Leaky-ReLU activation function (Maas et al., 2013), which is defined as

$$\boldsymbol{\rho}_\nu(\boldsymbol{x}) := \max\{\boldsymbol{x}, \nu\boldsymbol{x}\}$$

for $\nu \in [0, 1)$. Note that the Leaky-ReLU activation function includes the ReLU activation function as a special case with $\nu = 0$. The Leaky-ReLU activation function addresses one of the main limitations of the ReLU function: the dying ReLU problem (Douglas and Yu, 2018; Lu et al., 2020). In the original ReLU, negative inputs are zeroed out, potentially leading to dead neurons during training. The Leaky-ReLU addresses this issue by allowing a small, non-zero gradient for negative values. This leads to more consistent trainings of DNNs, and often results in improved performance on various tasks (Xu et al., 2015).

We aim to approximate $f_0 : \mathbb{R}^d \to \mathbb{R}$ that belongs to the β-Hölder class for a pre-specified value $\beta \in (0, \infty)$ by non-sparse DNNs. In the following theorem, we show that any function in the β-Hölder class can be approximated by a fully-connected DNN with bounded parameters.

Theorem 1. *For $\beta \in (0, \infty)$, $K \geq 1$ and $\nu \in [0, 1)$, there exist positive constants C_1, C_2, C_3 and c_1 such that for every $f_0 \in \mathcal{H}_d^\beta(K)$ and every sufficiently large $M \in \mathbb{N}$, there exists $f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(\lceil C_1 \log_2 M \rceil, \lceil C_2 M^d \rceil, C_3)$ with*

$$\left\| f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu} - f_0 \right\|_{\infty, [-a,a]^d} \leq c_1 \frac{1}{M^{2\beta}}.$$

The proof of Theorem 1 is provided in Appendix A and can be seen as a modification of the proof by Kohler and Langer (2021a). Theorem 1 employs DNNs with the number of nodes similar to that in Kohler and Langer (2021a) to achieve a similar error rate. However, while the infinite norm of the parameters in the approximation of Kohler and Langer (2021a) is unbounded, the parameters in our approximating DNNs are bounded by a constant $C_3 > 0$, which is independent of M and plays a crucial role to study posterior concentration rates with non-sparse DNNs and not-heavy-tailed priors. In addition, our theorem demonstrates the result for the Leaky-ReLU activation function using arbitrary $\nu \in [0, 1)$, extending the scope beyond the ReLU activation function used in Kohler and Langer (2021a). The core of our proof lies in using the degree-one homogeneity of the Leaky ReLU activation function to redistribute scale across layers.

4. Posterior Concentration

In this section, we demonstrate the optimal posterior concentration rate of fully-connected BNNs with general priors, including Gaussian priors, based on the result in Theorem 1.

BNNs are defined by a DNN model and priors over its parameters. We consider a DNN model $f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}}$ with the $(L, \boldsymbol{r})$ architecture, where $L \in \mathbb{N}$ and $\boldsymbol{r} = (r^{(0)}, r^{(1)}, ..., r^{(L)}, r^{(L+1)})^\top \in \mathbb{N}^{L+2}$ denote the depth and width of the DNN respectively. We assign a prior distribution Π over the parameters $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(J)})^\top \in \mathbb{R}^J$. Then, the corresponding posterior distribution is given by, for any measurable $A \subset \mathbb{R}^J$,

$$\Pi_n(A \mid \mathcal{D}^{(n)}) = \frac{\int_A \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}^{(n)}) d\Pi(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}^{(n)}) d\Pi(\boldsymbol{\theta})},$$

where $\mathcal{D}^{(n)}$ and $\mathcal{L}(\cdot|\mathcal{D}^{(n)})$ are train data set and corresponding likelihood function, respectively.

For a new test example $\boldsymbol{z} \in \mathcal{X}$, the prediction of a BNN is made based on the predictive distribution:

$$p(y \mid \boldsymbol{z}, \mathcal{D}^{(n)}) = \int_{\boldsymbol{\theta}} p(y \mid \boldsymbol{z}, \boldsymbol{\theta}) \Pi_n(\boldsymbol{\theta} \mid \mathcal{D}^{(n)}) d\boldsymbol{\theta}.$$

Since evaluating this integral directly is challenging, the Monte Carlo method is often employed to approximate it:

$$p(y \mid \boldsymbol{z}, \mathcal{D}^{(n)}) \approx \frac{1}{B} \sum_{b=1}^{B} p(y \mid \boldsymbol{z}, \boldsymbol{\theta}_b),$$

where $\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_B \sim \Pi_n(\boldsymbol{\theta} \mid \mathcal{D}^{(n)})$ are samples drawn from the posterior. These samples are usually generated by stochastic gradient Markov chain Monte Carlo (Welling and Teh, 2011; Chen et al., 2014; Li et al., 2016; Zhang et al., 2020; Heek and Kalchbrenner, 2019; Wilson and Izmailov, 2020) or variational inference (Graves, 2011; Blundell et al., 2015; Louizos and Welling, 2017; Swiatkowski et al., 2020).

4.1 Sufficient Condition for Priors

Since the choice of prior directly influences the resulting posterior distribution much, it has been the subject of considerable discussions (Izmailov et al., 2021; Fortuin, 2022). While the i.i.d. standard Gaussian prior is the most prevalent choice (Izmailov et al., 2021; Fortuin et al., 2022; Jospin et al., 2022), several alternative priors have been proposed, such as Laplace priors (Williams, 1995; Fortuin et al., 2022), radial-directional priors (Oh et al., 2020; Farquhar et al., 2020) and hierarchical priors (Hernández-Lobato and Adams, 2015; Louizos et al., 2017; Wu et al., 2019; Ghosh et al., 2019; Dusenberry et al., 2020; Ober and Aitchison, 2021; Seto et al., 2021), among others.

To demonstrate general theoretical results that encompass these priors, we make only the following very mild assumption about the prior distributions.

Assumption 1. Π *admits a probability density function* $\pi(\cdot)$ *on* $\mathbb{R}^J$ *with respect to Lebesgue measure. Also, for every* $\kappa > 0$*, there exists* $\delta_\kappa > 0$ *(not depending on* J*) such that* $\pi(\boldsymbol{\theta})$ *is lower bounded by* δ_κ^J *on* $\boldsymbol{\theta} \in [-\kappa, \kappa]^J$.

Since a lower bound of a density function usually decreases at an exponential rate as T increases, most prior distributions commonly considered for BNNs satisfy Assumption 1. Independent priors with specific conditions are simple examples, as follows.

Example 1 (Independent prior). *Assume that $\theta^{(1)}, \ldots, \theta^{(J)}$ are independent with the probability density functions $\pi^{(1)}(\cdot), \ldots, \pi^{(J)}(\cdot)$ on $\mathbb{R}$ with respect to Lebesgue measure respectively. Also, for every $\kappa > 0$, there exists $\delta_\kappa > 0$ (not depending on J) such that for every $j \in [J]$, $\pi^{(j)}(\theta)$ is lower bounded by δ_κ on $\theta \in [-\kappa, \kappa]$.*

Specifically, Example 1 includes independent Gaussian and Laplace priors, which have not yet been considered for the study of optimal posterior concentration rates in other papers.

Although independent priors are predominantly used in most BNNs for algorithmic convenience, several studies have explored using hierarchical priors for greater flexibility in the prior structure. Representative examples include zero-mean Gaussian prior with inverse-gamma prior on the prior variance (Hernández-Lobato and Adams, 2015; Wu et al., 2019) and group horseshoe prior (Louizos et al., 2017; Ghosh et al., 2019). Most of hierarchical priors fulfill Assumption 1, as illustrated in the following example.

Example 2 (Hierarchical prior). *Assume that the prior distribution of $\boldsymbol{\theta}$ is defined by a hierarchical structure:*

$$\begin{aligned} \boldsymbol{\psi} &\sim \Xi, \\ \boldsymbol{\theta}|\boldsymbol{\psi} &\sim \Pi_{\boldsymbol{\psi}}, \end{aligned}$$

where $\boldsymbol{\psi} \in \mathbb{R}^S$ is an auxiliary parameter, Ξ is a distribution of $\boldsymbol{\psi}$ and $\Pi_{\boldsymbol{\psi}}$ is a conditional distribution of $\boldsymbol{\theta}$ for given $\boldsymbol{\psi} \in \mathbb{R}^S$. Further assume that there exist a subset $\Psi \subseteq \mathbb{R}^S$ and a positive constant δ_1 such that (1) $\Xi(\boldsymbol{\psi} \in \Psi) \geq \delta_1^J$ and (2) for every $\boldsymbol{\psi} \in \Psi$, $\Pi_{\boldsymbol{\psi}}$ satisfies Assumption 1 with δ_κ not depending on $\boldsymbol{\psi}$. Then, Assumption 1 holds.

Another example satisfying Assumption 1 is the multivariate Gaussian distribution. Examples of the use of multivariate Gaussian priors include continual learning (Nguyen et al., 2018) and transfer learning (Špendl and Pirc, 2022), where they serve as informative priors for transferring information from one domain to another.

Example 3 (Multivariate Gaussian prior). *Assume that there exist positive constants B, $\lambda_{\min}$ and $\lambda_{\max}$ such that Π is a multivariate Gaussian prior with a mean vector in $[-B, B]^J$ and a covariance matrix whose eigenvalues are bounded between $\lambda_{\min}$ and $\lambda_{\max}$. Then, Assumption 1 holds.*

The proofs of the three examples are provided in Appendix B.1. Beyond these examples, most prior distributions commonly considered for BNNs also satisfy Assumption 1. An example of a prior, however, that does not satisfy Assumption 2 is a uniform distribution defined on $[-1, 1]^J$, as its density function has a value 0 for any $\boldsymbol{\theta} \in \mathbb{R}^J \setminus [-1, 1]^J$.

4.2 Results on Nonparametric Gaussian Regression

In nonparametric Gaussian regression problems, we assume that the input vector $\boldsymbol{X} \in \mathcal{X} \subseteq [-a, a]^d$ and the response variable $Y \in \mathbb{R}$ are generated from the model

$$\begin{aligned} \boldsymbol{X} \sim& \mathrm{P}_{\boldsymbol{X}}, \\ Y|\boldsymbol{X} \sim& N(f_0(\boldsymbol{X}), \sigma_0^2), \end{aligned} \tag{2}$$

where $\mathrm{P}_{\boldsymbol{X}}$ is the probability measure defined on $\mathcal{X}$. Here, $f_0 : \mathcal{X} \to \mathbb{R}$ and $\sigma_0^2 > 0$ are the unknown true regression function and variance of the noise, respectively. We assume that the true regression function f_0 satisfies $\|f_0\|_\infty \leq F$ and $f_0 \in \mathcal{H}_d^\beta(K)$ for some $F \geq 1$, $\beta > 0$ and $K \geq 1$. We assume that $\mathcal{D}^{(n)} := \{(\boldsymbol{X}_i, Y_i)\}_{i\in[n]}$ are independent copies.

For Bayesian inference, we consider the probabilistic model

$$Y_i \overset{ind.}{\sim} N\left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i), \sigma^2\right) \tag{3}$$

for a pre-specified $\nu \in [0, 1)$, where T_F is the truncation operator defined as $T_F(x) = \max(-F, \min(x, F))$ and $f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ is the $(L_n, \boldsymbol{r}_n)$ architecture DNN, where L_n and $\boldsymbol{r}_n$ are given by

$$\begin{aligned} L_n &:= \lceil C_1 \log n \rceil, \\ r_n &:= \left\lceil C_2 n^{\frac{d}{2(2\beta+d)}} \right\rceil, \\ \boldsymbol{r}_n &:= (d, r_n, \ldots, r_n, 1)^\top \in \mathbb{N}^{L_n+2} \end{aligned} \tag{4}$$

for constants C_1 and C_2 defined in Theorem 1. Then, the likelihood is expressed as

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2 | \mathcal{D}^{(n)}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (Y_i - T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i))^2}{2\sigma^2} \right).$$

For a given DNN structure, we assign a prior Π over $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$ which satisfies Assumption 1, where J_n is defined by

$$J_n := (d+1)r_n + (L_n - 1)(r_n + 1)r_n + (r_n + 1).$$

Also, we assign a prior Ξ over $\sigma^2 \in \mathbb{R}^+$, which is independent of Π and satisfies the following mild condition.

Assumption 2. *Ξ admits a density with respect to Lebesgue measure, which is continuous and positive on $(0, 2\sigma_0^2)$. Additionally, $\Xi(\sigma^2 > K) \lesssim \frac{1}{K}$ holds for sufficiently large K.*

Assumption 2 holds for most distributions whose support includes σ_0^2. The most commonly used prior for the σ^2 is the inverse-gamma distribution; however, other priors, such as the uniform distribution, can also be employed. The corresponding posterior distribution is given by, for any measurable $A \subset \mathbb{R}^{J_n}$ and $B \subset \mathbb{R}^+$,

$$\Pi_n(A \otimes B \mid \mathcal{D}^{(n)}) = \frac{\int_A \int_B \mathcal{L}(\boldsymbol{\theta}, \sigma^2 | \mathcal{D}^{(n)}) d\Xi(\sigma^2) d\Pi(\boldsymbol{\theta})}{\int \int \mathcal{L}(\boldsymbol{\theta}, \sigma^2 | \mathcal{D}^{(n)}) d\Xi(\sigma^2) d\Pi(\boldsymbol{\theta})}.$$

In the following theorem, we demonstrate that BNNs with general priors (i.e., priors satisfying Assumption 1 and 2) achieve optimal (up to a logarithmic factor) posterior concentration rates around the true regression function.

Theorem 2. *Assume $f_0 \in \mathcal{H}_d^\beta(K)$ and $\|f_0\|_\infty \leq F$ for some $K \geq 1$, $\beta > 0$ and $F \geq 1$. For $\nu \in [0, 1)$, consider the DNN model $f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ with the $(L_n, \boldsymbol{r}_n)$ architecture, where L_n and $\boldsymbol{r}_n$ are given in (4). For any priors Π and Ξ satisfying Assumption 1 and Assumption 2,* *respectively, the posterior distribution of $T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ and σ^2 concentrates around f_0 and σ_0^2 at the rate $\varepsilon_n = n^{-\beta/(2\beta+d)} \log^\gamma(n)$ for $\gamma > 2$, in the sense that*

$$\Pi_n\Big((\boldsymbol{\theta}, \sigma^2) : \ \|T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - f_0\|_{2,\mathrm{P}_X} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \ \Big| \ \mathcal{D}^{(n)}\Big) \overset{\mathbb{P}_0^n}{\to} 0$$

as $n \to \infty$ for any $M_n \to \infty$, where $\mathbb{P}_0^n$ is the probability measure of the training data $\mathcal{D}^{(n)}$.

The proof of Theorem 2 is provided in Appendix B.3. The concentration rate $n^{-\beta/(2\beta+d)}$ is known to be the minimax lower bound when estimating the β-Hölder smooth functions Tsybakov (2009). Our concentration rate is near-optimal up to a logarithmic factor.

Similar concentration rates have been derived in previous works (Polson and Ročková, 2018; Chérief-Abdellatif, 2020; Bai et al., 2020; Sun et al., 2022; Kong et al., 2023; Ohn and Lin, 2024). However, as we mentioned earlier, the prior distributions considered in these studies are not commonly used in practice. Polson and Ročková (2018); Chérief-Abdellatif (2020); Bai et al. (2020); Sun et al. (2022) require sparse-inducing priors, which is computationally demanding due to additional exploration time for searching sparsity patterns. Near-optimal posterior concentration rates for non-sparse BNNs have been obtained by Kong et al. (2023) and Ohn and Lin (2024), but extremely heavy-tailed priors are required which do not even include Gaussian distributions. Theorem 2 stands as the first result to establish the theoretical optimality of BNNs with Gaussian priors.

Remark 1. *To prove Theorem 2, we check the conditions in Ghosal and van der Vaart (2007), which is the standard methodology for demonstrating posterior concentrations in nonparametric regression problems. This technique necessitates showing the prior concentration condition*

$$(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \gtrsim e^{-n\varepsilon_n^2}, \tag{5}$$

where $B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)$ denotes the ε_n-Kullback–Leibler neighbourhood around (f_0, σ_0^2). To establish (5), we must first find a DNN that approximates f_0 and then demonstrate that sufficient prior probability exists around it. While, existing approximation results fall short of enabling the demonstration of (5) for the Gaussian prior, in Section 3 makes it possible. See details in the proof of Lemma B.4 of the Appendix.

4.3 Results on Nonparametric Logistic Regression

In nonparametric logistic regression problems, we assume that the input vector $\boldsymbol{X} \in \mathcal{X} \subseteq [-a, a]^d$ and the response variable $Y \in \{0, 1\}$ are generated from the model

$$\begin{aligned} \boldsymbol{X} \sim& \mathrm{P}_{\boldsymbol{X}}, \\ Y|\boldsymbol{X} \sim& \operatorname{Bernoulli}\left(\phi \circ f_0(\boldsymbol{X})\right), \end{aligned} \tag{6}$$

where $\mathrm{P}_{\boldsymbol{X}}$ is the probability measure defined on $\mathcal{X}$ and $\phi(z) := (1 + \exp(-z))^{-1}$ is the sigmoid function. Here, $f_0 : \mathcal{X} \to \mathbb{R}$ is the logit of the unknown probability function. We assume that the true function f_0 satisfies $\|f_0\|_\infty \le F$ and $f_0 \in \mathcal{H}_d^\beta(K)$ for some $F \ge 1$, $\beta > 0$ and $K \ge 1$. We assume that $\mathcal{D}^{(n)} := \{(\boldsymbol{X}_i, Y_i)\}_{i\in[n]}$ are independent copies.

For Bayesian inference, we consider the probabilistic model

$$Y_i \overset{ind.}{\sim} \text{Bernoulli}\left(\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i)\right) \tag{7}$$

for a pre-specified $\nu \in [0,1)$, where $f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i)$ is the $(L_n, \boldsymbol{r}_n)$ architecture DNN, where L_n and $\boldsymbol{r}_n$ are given by (4). Then, the likelihood is expressed as

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}^{(n)}) = \prod_{i=1}^{n} (\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i))^{Y_i}(1 - \phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i))^{1-Y_i},$$

and the corresponding posterior distribution is given by, for any measurable $A \subset \mathbb{R}^{J_n}$,

$$\Pi_n(A \mid \mathcal{D}^{(n)}) = \frac{\int_A \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}^{(n)}) d\Pi(\boldsymbol{\theta})}{\int \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}^{(n)}) d\Pi(\boldsymbol{\theta})}$$

In the following theorem, we demonstrate that the BNNs with general priors (i.e., priors satisfying Assumption 1) achieve optimal (up to a logarithmic factor) posterior concentration rates around the true conditional class probability.

Theorem 3. *Assume $f_0 \in \mathcal{H}^{\beta}_{d}(K)$ and $\|f_0\|_\infty \le F$ for some $K \ge 1$, $\beta > 0$ and $F > 0$. For $\nu \in [0,1)$, consider the DNN model $f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ with the $(L_n, \boldsymbol{r}_n)$ architecture, where L_n and $\boldsymbol{r}_n$ are given in (4). For any prior Π over $\boldsymbol{\theta}$ satisfying Assumption 1, the posterior distribution of $\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ concentrates around the true conditional class probability at the rate $\varepsilon_n = n^{-\beta/(2\beta+d)} \log^{\gamma}(n)$ for $\gamma > 2$, in the sense that*

$$\Pi_n\Big(\boldsymbol{\theta} : \ \|\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - \phi \circ f_0\|_{2,\text{P}_X} > M_n \varepsilon_n \ \Big| \ \mathcal{D}^{(n)}\Big) \xrightarrow{\mathbb{P}^n_0} 0$$

as $n \to \infty$ for any $M_n \to \infty$, where $\mathbb{P}^n_0$ is the probability measure of the training data $\mathcal{D}^{(n)}$.

4.4 Avoiding the Curse of Dimensionality by Assuming Hierarchical Composition Structure

Due to the inherent hierarchical structure of DNNs, they hold particular advantages when modeling data that also exhibits a hierarchical structure. For example, image data encompasses multiple levels of abstraction, ranging from low-level features such as edges and textures to high-level concepts like objects and scenes, and is hence considered to exhibit a hierarchical structure. Building upon this intuition, Schmidt-Hieber (2020) and Kohler and Langer (2021a) prove that by assuming a hierarchical structure for the true function, DNN models can avoid the curse of dimensionality and achieve faster convergence rates. However, faster concentration rate of hierarchical composition structure functions in BNNs has not yet been investigated. In this subsection, we demonstrate that faster BNN concentration results can be achieved by assuming a similar hierarchical structure.

For a minimum smoothness $\beta_{min} \ge 1$, a maximum smoothness $\beta_{max} \ge \beta_{min}$ and a maximum dimension $d_{max} \in \mathbb{N}$, let

$$\mathcal{P} \subset [\beta_{min}, \beta_{max}] \times \{1, \dots, d_{max}\}$$

be a constraint set consisting of pairs of smoothness and dimension. We assume that the true function f_0 follows a hierarchical composition structure with $\boldsymbol{N} \in \mathbb{N}^q$ for some $q \in \mathbb{N}$ and the constraint set $\mathcal{P}$, which is defined as follows.

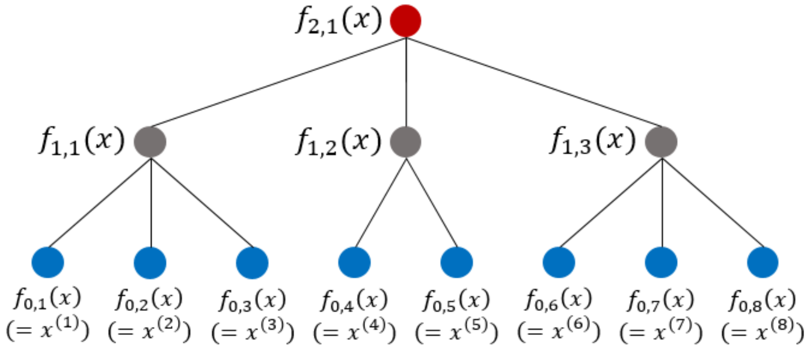
$f_{2,1}(x)$
$f_{1,1}(x)$
$f_{1,2}(x)$
$f_{1,3}(x)$
$f_{0,1}(x)$ $(= x^{(1)})$
$f_{0,2}(x)$ $(= x^{(2)})$
$f_{0,3}(x)$ $(= x^{(3)})$
$f_{0,4}(x)$ $(= x^{(4)})$
$f_{0,5}(x)$ $(= x^{(5)})$
$f_{0,6}(x)$ $(= x^{(6)})$
$f_{0,7}(x)$ $(= x^{(7)})$
$f_{0,8}(x)$ $(= x^{(8)})$

Figure 1: Example of hierarchical composition structure.

Definition 2 (hierarchical composition structure)**.** *We say that a function f follows the hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$ if for $(N_0, \ldots, N_{q-1})^\top := \boldsymbol{N}$ and $N_q := 1$,*

a) For $i \in [N_0]$, there exists $d' \in [d]$ such that

$$f_{0,i}(\boldsymbol{x}) = x^{(d')} \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d.$$

b) For $l \in [q]$ and $i \in [N_l]$, there exists $(\beta_{l,i}, d_{l,i}) \in \mathcal{P}$ such that $\sum_{i=1}^{N_l} d_{l,i} = N_{l-1}$ holds. Also, there exists a C_{Lip}-Lipschitz function $g_{l,i} : \mathbb{R}^{d_{l,i}} \to \mathbb{R}$ with $C_{Lip} \geq 1$ such that $g_{l,i} \in \mathcal{H}_{d_{l,i}}^{\beta_{l,i}}(K)$, $\|g_{l,i}\|_\infty \leq F$ and

$$f_{l,i}(\boldsymbol{x}) = g_{l,i}\left(f_{l-1,\sum_{i'=1}^{i-1} d_{l,i'}+1}(\boldsymbol{x}), \ldots, f_{l-1,\sum_{i'=1}^{i-1} d_{l,i'}+d_{l,i}}(\boldsymbol{x})\right) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d.$$

c) The function f satisfies

$$f(\boldsymbol{x}) = f_{q,1}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d.$$

Figure 1 illustrates a simple example of a hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$, where $\boldsymbol{N} = (8, 3)^\top$ and $\mathcal{P} = \{(4, 2), (4, 3), (5, 3)\}$. In this case, we have $d_{1,1} = 3, d_{1,2} = 2, d_{1,3} = 3$, $d_{2,1} = 3$, $f_{0,i} = x^{(i)}$ for $i \in [8]$ and

$$\begin{aligned}
f_{1,1}(\boldsymbol{x}) &= g_{1,1}(f_{0,1}(\boldsymbol{x}), f_{0,2}(\boldsymbol{x}), f_{0,3}(\boldsymbol{x})), \\
f_{1,2}(\boldsymbol{x}) &= g_{1,2}(f_{0,4}(\boldsymbol{x}), f_{0,5}(\boldsymbol{x})), \\
f_{1,3}(\boldsymbol{x}) &= g_{1,3}(f_{0,6}(\boldsymbol{x}), f_{0,7}(\boldsymbol{x}), f_{0,8}(\boldsymbol{x})), \\
f_{2,1}(\boldsymbol{x}) &= g_{2,1}(f_{1,1}(\boldsymbol{x}), f_{1,2}(\boldsymbol{x}), f_{1,3}(\boldsymbol{x})),
\end{aligned}$$

where $g_{1,1} \in \mathcal{H}_3^4(K)$, $g_{1,2} \in \mathcal{H}_2^4(K)$, $g_{1,3} \in \mathcal{H}_3^5(K)$ and $g_{2,1} \in \mathcal{H}_3^5(K)$. In this example, the actual input dimension of $f_{2,1}$ is 8, but the maximum input dimension of $g_{l,i}$ is 3. The primary advantage of assuming such a hierarchical composition structure is that the dimensions of each g are significantly smaller compared to the overall dimensions, which allows the function to be approximated with a smaller DNN model.

We assume that $\mathcal{D}^{(n)} := \{(\boldsymbol{X}_i, Y_i)\}_{i\in[n]}$ are independent copies generated from (2) for the nonparametric Gaussian regression problem and from (6) for the nonparametric logistic regression problem, where the true function f_0 follows the hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$. For Bayesian inference, we consider the probabilistic model

$$Y_i \overset{ind.}{\sim} N\left(T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i), \sigma^2\right)$$

for nonparametric Gaussian regression problem and

$$Y_i \overset{ind.}{\sim} \text{Bernoulli}\left(\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{X}_i)\right)$$

for nonparametric logistic regression problem, where $\nu \in [0, 1)$ is a pre-specified value and $f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ is the $(L_n, \boldsymbol{r}_n)$ architecture DNN, where L_n and $\boldsymbol{r}_n$ are given by

$$\begin{aligned} L_n &:= \left\lceil \tilde{C}_1 \log_2 n \right\rceil, \\ r_n &:= \left\lceil \tilde{C}_2 \max_{(\beta',d')\in\mathcal{P}} n^{\frac{d'}{2(2\beta'+d')}} \right\rceil, \\ \boldsymbol{r}_n &:= (d, r_n, \dots, r_n, 1)^\top \in \mathbb{N}^{L_n+2}, \end{aligned} \tag{8}$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are constants (depending on β_{min}, β_{max} and d_{max}) defined in Lemma B.8 in Appendix B.5.

The following theorem shows that by assuming hierarchical compositional structure, BNNs with general priors can avoid the curse of dimensionality.

Theorem 4. *Assume that data are generated from (2) for the nonparametric Gaussian regression problem and from (6) for the nonparametric logistic regression problem. Assume that f_0 follows the hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$, defined in Definition 2. For $\nu \in [0, 1)$, consider the DNN model $f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ with the $(L_n, \boldsymbol{r}_n)$ architecture, where L_n and $\boldsymbol{r}_n$ are given in (8). For any priors Π (and Ξ) which satisfy Assumption 1 (and Assumption 2, respectively), the posterior distribution concentrates around the true function at the rate $\varepsilon_n = \max_{(\beta',d')\in\mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}} \log^\gamma(n)$ for $\gamma > 2$, in the sense that for any $M_n \to \infty$,*

$$\Pi_n\Big((\boldsymbol{\theta}, \sigma^2) : \ \|T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - f_0\|_{2,\mathrm{P}_X} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \ \Big|\ \mathcal{D}^{(n)}\Big) \overset{\mathbb{P}_0^n}{\to} 0$$

holds for the nonparametric Gaussian regression problem and

$$\Pi_n\Big(\boldsymbol{\theta} : \ \|\phi \circ T_F \circ f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - \phi \circ f_0\|_{2,\mathrm{P}_X} > M_n\varepsilon_n \ \Big|\ \mathcal{D}^{(n)}\Big) \overset{\mathbb{P}_0^n}{\to} 0$$

holds for the nonparametric logistic regression problem, where $\mathbb{P}_0^n$ is the probability measure of the training data $\mathcal{D}^{(n)}$.

The proof of Theorem 4 is provided in Appendix B.5. Note that the concentration rate $\max_{(\beta',d')\in\mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}}$ is much faster than $n^{-\frac{\beta}{(2\beta+d)}}$ in the case $d_{max} \ll d$ and hence avoids the curse of dimensionality. This rate is known to be the minimax lower bound when estimating hierarchical composition structure functions (Schmidt-Hieber, 2020). Concentration rate in Theorem 4 is near-optimal up to a logarithmic factor.

5. Bayesian Neural Networks Adaptive to Smoothness

Similar to minimax optimal results of least square estimators with DNNs (Schmidt-Hieber, 2020; Kohler and Langer, 2021a), theories in the previous section also face the limitation of requiring knowledge of the smoothness of the true function (or constraint set $\mathcal{P}$ for hierarchical composition structure in Section 4) to choose a network of appropriate size. Since the true smoothness is rarely known, the optimal width is usually determined through the use of a validation data set in practice.

In Bayesian analysis, this issue is often addressed by assigning a prior to the parameter related to smoothness or model complexity. Instead of choosing the width r depending on β as well as n in our previous results in Section 4, we can assign a prior to r. Specifically, we give prior

$$\Gamma(r) \propto e^{-(\log n)^5 r^2}, \tag{9}$$

which is similar to the prior considered in Kong et al. (2023) and Ohn and Lin (2024). For given $r \in \mathbb{N}$, L_n and $\boldsymbol{r}_n$ are given by

$$\begin{aligned} L_n &:= \left\lceil \tilde{C}_1 \log n \right\rceil, \\ \boldsymbol{r}_n &:= (d, r, \ldots, r, 1)^\top \in \mathbb{N}^{L_n+2}, \end{aligned} \tag{10}$$

for the constant $\tilde{C}_1$ used in (8). For a given DNN structure, we assign a prior Π_r over $\boldsymbol{\theta} \in \mathbb{R}^{J_{n,r}}$ which satisfies Assumption 1, where $J_{n,r}$ is defined by

$$J_{n,r} := (d+1)r + (L_n - 1)(r+1)r + (r+1).$$

In addition, for nonparametric Gaussian regression, we assign a prior Ξ over σ^2 which satisfies Assumption 2. For Bayesian inference, we consider the probabilistic model (3) for nonparametric Gaussian regression problem and (7) for nonparametric logistic regression problem.

The following theorem shows that by giving suitable prior on the width, BNNs with general priors achieve optimal (up to a logarithmic factor) posterior concentration rates around the true function, adaptively to the true smoothness.

Theorem 5. *Assume that data are generated from (2) for the nonparametric Gaussian regression problem and from (6) for the nonparametric logistic regression problem. Assume that f_0 follows the hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$, defined in Definition 2. For $\nu \in [0, 1)$, consider the prior (9) over the width r, and consider the DNN model $f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}$ with the $(L_n, \boldsymbol{r}_n)$ architecture, where L_n and $\boldsymbol{r}_n$ are given in (10). For any priors Π_r (and Ξ) which satisfy Assumption 1 (and Assumption 2, respectively), the posterior distribution concentrates around the true function at the rate $\varepsilon_n = \max_{(\beta',d')\in\mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}} \log^\gamma(n)$ for $\gamma > \frac{5}{2}$, in the sense that for any $M_n \to \infty$,*

$$\Pi_n\Big((\boldsymbol{\theta}, \sigma^2) : \ \|T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - f_0\|_{2,\mathrm{P}_X} + |\sigma^2 - \sigma_0^2| > M_n \varepsilon_n \ \Big|\ \mathcal{D}^{(n)} \Big) \xrightarrow{\mathbb{P}_0^n} 0$$

holds for the nonparametric Gaussian regression problem and

$$\Pi_n\Big(\boldsymbol{\theta}:\ \|\phi\circ T_F\circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} - \phi\circ f_0\|_{2,\mathrm{P}_X} > M_n\varepsilon_n \,\Big|\, \mathcal{D}^{(n)}\Big) \overset{\mathbb{P}_0^n}{\to} 0$$

holds for the nonparametric logistic regression problem, where $\mathbb{P}_0^n$ is the probability measure of the training data $\mathcal{D}^{(n)}$.

The proof of Theorem 5 is provided in Appendix C. Theorem 5 implies that BNNs with a random width achieve near-optimal concentration rates up to a logarithmic factor as long as the prior of the random width is carefully selected. Note that the proposed prior for the width does not utilize $\mathcal{P}$, which means that the BNNs achieve the optimal posterior concentration rates adaptively to the smoothness of the true model.

Posterior inference in such models poses significant challenges, since the dimension of parameters changes as the network structure changes. In such cases, a commonly used method is reversible jump MCMC (Green, 1995), which employs dimension-matching techniques and proposes changes of the dimensionality along with appropriate adjustments in the parameters. Alternatively, one could consider applying masking variables to a sufficiently large DNN, and by using a well-designed proposal distribution for the Metropolis-Hastings algorithm to update the masking variables, faster posterior mixing can be induced (Kong et al., 2023).

6. Discussions

The main contributions of this paper are (1) to provide the new approximation result of DNNs in Theorem 1 and (2) to derive the posterior concentration rates of BNNs with general priors based on Theorem 1. We believe that there are other problems where the new approximation result of DNNs plays a crucial role. For instance, whereas Ohn and Lin (2024) employs a uniform prior over a diverging set and Kong et al. (2023) uses a polynomial-tailed prior, their procedures can also achieve near-minimax optimal rates under a standard Gaussian prior by applying Theorem 1. Another possible example would be asymptotic properties of a certain penalized least square (or maximum likelihood) estimator of DNN. There are some studies of sparse penalties with DNNs (Ohn and Kim, 2022) but no results are available for non-sparse penalties.

We have only considered the posterior concentration rates of BNNs. A more interesting property would be uncertainty quantification. For Bayesian nonparametric regression, Szabó et al. (2015) and Rousseau and Szabo (2020) have studied asymptotic properties in view of uncertainty quantification. It would be expected that similar results hold for BNNs but not yet proved. Our new approximation result could be a good starting point.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00556079) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

Appendix A. Proofs for Section 3

In this section, we prove Theorem 1. In Section A.1, we describe additional notations for the proofs. In Section A.2, we state and prove a re-scaling lemma for Leaky-ReLU DNNs. In Section A.3, we construct auxiliary networks with Leaky-ReLU activation function. Based on these results, we prove Theorem 1 in Section A.4.

Our proof for Theorem 1 closely follows the proof of Kohler and Langer (2021a) but we make the following four modifications: (1) alteration of the activation function, (2) adjustments to the network size accordingly, (3) imposition of the upper bound of the absolute values of parameters in each layer and (4) alteration of the upper bounds to make them similar. For simplicity, we refer to the results in the proof of Kohler and Langer (2021a) unless there is any confusion, and focus on the four modifications.

A.1 Additional notations

For a DNN model f^{DNN}, $L(f^{\mathrm{DNN}})$ denotes the number of hidden layers in f^{DNN}. For $l \in [L(f^{\mathrm{DNN}})+1]$, $W_l(f^{\mathrm{DNN}})$ and $\boldsymbol{b}_l(f^{\mathrm{DNN}})$ denote the weight matrix and bias vector of the l-th layer of f^{DNN}, respectively. We further denote

$$\begin{aligned}\boldsymbol{\theta}_w(f^{\mathrm{DNN}}) :=& (\mathrm{vec}(W_1(f^{\mathrm{DNN}}))^\top, \ldots, \mathrm{vec}(W_{L+1}(f^{\mathrm{DNN}}))^\top)^\top,\\ \boldsymbol{\theta}_b(f^{\mathrm{DNN}}) :=& (\boldsymbol{b}_1(f^{\mathrm{DNN}})^\top, \ldots, \boldsymbol{b}_{L+1}(f^{\mathrm{DNN}})^\top)^\top.\end{aligned}$$

We define the set of Leaky-ReLU DNN functions where the absolute values of the weights and biases are bounded by B_w and B_b, respectively, by $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$ as follows.

Definition 3. *For $L \in \mathbb{N}$, $r \in \mathbb{N}$, $B_w \geq 1$ and $B_b \geq 1$, we define $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}}(L, r, B_w, B_b)$ as the function class of DNNs with the $(L, (d, r, \ldots, r, 1)^\top)$ architecture and the activation function $\boldsymbol{\rho}$ such that the absolute values of the weights are bounded by B_w and the absolute values of the biases are bounded by B_b. That is,*

$$\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b) := \Bigg\{ f : f = f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}} \text{ is a DNN with the } \Big(L, (d, r, \ldots, r, 1)^\top\Big) \text{ architecture,} \\ |\boldsymbol{\theta}_w(f)|_\infty \leq B_w, |\boldsymbol{\theta}_b(f)|_\infty \leq B_b \Bigg\}.$$

A vector function is denoted by a bold letter. e.g. $\boldsymbol{f}(\cdot) := (f^{(1)}(\cdot), \ldots, f^{(k)}(\cdot))^\top$.

A.2 Re-scaling Lemma for Leaky-ReLU DNNs

Lemma A.1. *For $\nu \in [0, 1)$, $L \in \mathbb{N}$ and $\boldsymbol{r} \in \mathbb{N}^{L+2}$, consider a DNN with the $(L, \boldsymbol{r})$ architecture $f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} : \mathbb{R}^d \to \mathbb{R}$ which is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_w^\top, \boldsymbol{\theta}_b^\top)^\top$, where $\boldsymbol{\theta}_w = (\mathrm{vec}(W_1)^\top, \ldots, \mathrm{vec}(W_{L+1})^\top)^\top$ and $\boldsymbol{\theta}_b = (\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_{L+1}^\top)^\top$. For positive constants $\zeta_1, \ldots, \zeta_{L+1}$, we define*

$$\begin{aligned}\tilde{W}_l :=& \zeta_l \cdot W_l,\\ \tilde{\boldsymbol{b}}_l :=& \left(\prod_{l'=1}^{l} \zeta_{l'}\right) \cdot \boldsymbol{b}_l\end{aligned}$$

for $l \in [L+1]$ *and define*

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_w :=& (\text{vec}(\tilde{W}_1)^\top, \ldots, \text{vec}(\tilde{W}_{L+1})^\top)^\top, \\ \tilde{\boldsymbol{\theta}}_b :=& (\tilde{\boldsymbol{b}}_1^\top, \ldots, \tilde{\boldsymbol{b}}_{L+1}^\top)^\top, \\ \tilde{\boldsymbol{\theta}} :=& (\tilde{\boldsymbol{\theta}}_w^\top, \tilde{\boldsymbol{\theta}}_b^\top)^\top.\end{aligned}$$

If $\prod_{l=1}^{L+1} \zeta_l = 1$,

$$f^{\text{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}(\boldsymbol{x}) = f^{\text{DNN}}_{\tilde{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}(\boldsymbol{x})$$

holds for every $\boldsymbol{x} \in \mathbb{R}^d$

Proof. Note that the Leaky-ReLU activation function has the property

$$\boldsymbol{\rho}_\nu(\zeta \boldsymbol{x}) = \max\{\zeta \boldsymbol{x}, \zeta\nu \boldsymbol{x}\} = \zeta \max\{\boldsymbol{x}, \nu \boldsymbol{x}\} = \zeta \boldsymbol{\rho}_\nu(\boldsymbol{x})$$

for every $\zeta > 0$ and vector $\boldsymbol{x}$. For $l \in [L+1]$, we denote $\tilde{A}_l(\boldsymbol{x}) := \tilde{W}_l \boldsymbol{x} + \tilde{\boldsymbol{b}}_l$ as the affine map defined by $\tilde{W}_l$ and $\tilde{\boldsymbol{b}}_l$.

First, we have $\tilde{A}_1(\boldsymbol{x}) = \tilde{W}_1 \boldsymbol{x} + \tilde{\boldsymbol{b}}_1 = \zeta_1 A_1(\boldsymbol{x})$. Assume that for some $l \in [L]$,

$$\tilde{A}_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ \tilde{A}_1(\boldsymbol{x}) = \left(\prod_{l'=1}^{l} \zeta_{l'}\right) \cdot (A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x}))$$

holds. Then, we have

$$\begin{aligned}&\tilde{A}_{l+1} \circ \boldsymbol{\rho}_\nu \circ \tilde{A}_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ \tilde{A}_1(\boldsymbol{x}) \\ &= \tilde{A}_{l+1} \circ \boldsymbol{\rho}_\nu \circ \left(\left(\prod_{l'=1}^{l} \zeta_{l'}\right) \cdot (A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x}))\right) \\ &= \tilde{A}_{l+1} \circ \left(\left(\prod_{l'=1}^{l} \zeta_{l'}\right) \cdot (\boldsymbol{\rho}_\nu \circ A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x}))\right) \\ &= \tilde{W}_{l+1} \left(\left(\prod_{l'=1}^{l} \zeta_{l'}\right) \cdot (\boldsymbol{\rho}_\nu \circ A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x}))\right) + \tilde{\boldsymbol{b}}_{l+1} \\ &= \left(\prod_{l'=1}^{l+1} \zeta_{l'}\right) \cdot (W_{l+1}(\boldsymbol{\rho}_\nu \circ A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x})) + \boldsymbol{b}_{l+1}) \\ &= \left(\prod_{l'=1}^{l+1} \zeta_{l'}\right) \cdot (A_{l+1} \circ \boldsymbol{\rho}_\nu \circ A_l \circ \boldsymbol{\rho}_\nu \cdots \circ \boldsymbol{\rho}_\nu \circ A_1(\boldsymbol{x})).\end{aligned}$$

Hence, by mathematical induction, we have the assertion. ■

Lemma A.1 implies that scale of some layers can be transferred to other layers in Leaky-ReLU DNNs. This is due to the fact that the Leaky-ReLU activation function satisfies the property $\boldsymbol{\rho}_\nu(c\boldsymbol{x}) = c\boldsymbol{\rho}_\nu(\boldsymbol{x})$ for any $\boldsymbol{x}$ and $c > 0$. This lemma is particularly useful when the absolute values of some parameters in the lower layers are large, while those in other layers are not. Figure 2 provides a simple illustration of Lemma A.1 with $\zeta_1 = 2^{-L}$ and $\zeta_2 = \cdots = \zeta_{L+1} = 2$.

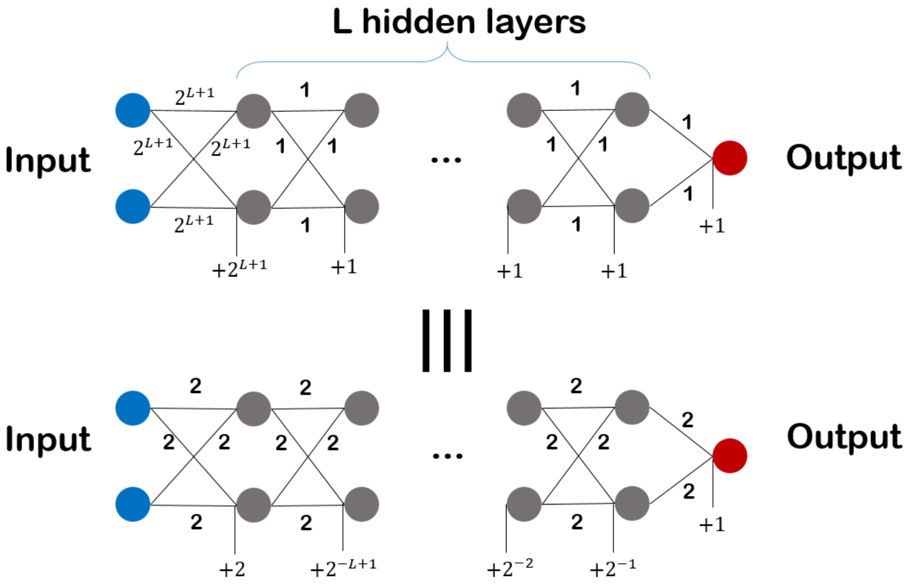

Figure 2: Example of Lemma A.1. A DNN with depth L and width $\boldsymbol{r} = (2,\dots,2,1)^\top$ (above) and its re-scaled DNN (below) using Lemma A.1. We use $\zeta_1 = 2^{-L}$ and $\zeta_2 = \cdots = \zeta_{L+1} = 2$ for re-scaling. The two networks produce a same output for a same input.

A.3 Auxiliary networks with the Leaky-ReLU activation function

Lemma A.2. *For $\nu \in [0,1)$ and $k \in \mathbb{N}$,*

a) There exists a neural network $\boldsymbol{f}_{id}(\cdot) : \mathbb{R}^k \to \mathbb{R}^k$ such that for every $\boldsymbol{x} \in \mathbb{R}^k$

$$\boldsymbol{f}_{id}(\boldsymbol{x}) = \boldsymbol{x},$$

and for every neural network $\boldsymbol{g}(\cdot) : \mathbb{R}^d \to \mathbb{R}^k$ in $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$,

$$\boldsymbol{f}_{id}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}\left(L+1, \max(r, 2k), B_w, B_b\right).$$

b) There exists a neural network $\boldsymbol{f}_{\rho_0}(\cdot) : \mathbb{R}^k \to \mathbb{R}^k$ such that for every $\boldsymbol{x} \in \mathbb{R}^k$

$$\boldsymbol{f}_{\rho_0}(\boldsymbol{x}) = \boldsymbol{\rho}_0(\boldsymbol{x}),$$

and for every neural network $\boldsymbol{g}(\cdot) : \mathbb{R}^d \to \mathbb{R}^k$ in $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$,

$$\boldsymbol{f}_{\rho_0}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}\left(L+1, \max(r, 2k), \max\left(B_w, \frac{1}{1-\nu^2}\right), B_b\right).$$

Proof. a) For $\boldsymbol{x} = (x^{(1)}, \dots, x^{(k)})^\top$,

$$\boldsymbol{f}_{id}(\boldsymbol{x}) := \frac{1}{1+\nu}\boldsymbol{\rho}_\nu(\boldsymbol{x}) - \frac{1}{1+\nu}\boldsymbol{\rho}_\nu(-\boldsymbol{x})$$

satisfies $f_{id}(\boldsymbol{x})^{(i)} = \frac{x^{(i)}}{1+\nu} + \frac{\nu x^{(i)}}{1+\nu} = x^{(i)}$ for $x^{(i)} \geq 0$ and $f_{id}(\boldsymbol{x})^{(i)} = \frac{\nu x^{(i)}}{1+\nu} + \frac{x^{(i)}}{1+\nu} = x^{(i)}$ for $x^{(i)} < 0$. Also, $\boldsymbol{f}_{id}(\boldsymbol{g}(\cdot))$ requires r neurons in each of the 1st through Lth hidden layers and $2k$ neurons in the $(L+1)$th hidden layer. Hence, we get the assertion.

b) For $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(k)})^\top$,

$$\boldsymbol{f}_{\rho_0}(\boldsymbol{x}) := \frac{1}{1-\nu^2}\boldsymbol{\rho}_\nu(\boldsymbol{x}) + \frac{\nu}{1-\nu^2}\boldsymbol{\rho}_\nu(-\boldsymbol{x})$$

satisfies $f_{\rho_0}(\boldsymbol{x})^{(i)} = \frac{x^{(i)}}{1-\nu^2} - \frac{\nu^2 x^{(i)}}{1-\nu^2} = x^{(i)}$ for $x^{(i)} \geq 0$ and $f_{\rho_0}(\boldsymbol{x})^{(i)} = \frac{\nu x^{(i)}}{1-\nu^2} - \frac{\nu x^{(i)}}{1-\nu^2} = 0$ for $x^{(i)} < 0$. Also, $\boldsymbol{f}_{\rho_0}(\boldsymbol{g}(\cdot))$ requires r neurons in each of the 1st through Lth hidden layers and $2k$ neurons in the $(L+1)$th hidden layer. Hence, we get the assertion.

■

For $L \in \mathbb{N}$, we denote $\boldsymbol{f}_{id}^L(\cdot) := \boldsymbol{f}_{id} \circ \boldsymbol{f}_{id} \circ \cdots \circ \boldsymbol{f}_{id}(\cdot) \in \tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, 2k, 1, 1)$, where k is the input dimension of $\boldsymbol{f}_{id}^L$.

Lemma A.3. *Let $\nu \in [0,1)$, $R \in \mathbb{N}$, $\boldsymbol{b}_1, \boldsymbol{b}_2 \in [-a,a]^d$ with $b_2^{(i)} - b_1^{(i)} \geq \frac{2}{R}$ for all $i \in [d]$ and*

$$K_{1/R} = \left\{\boldsymbol{x} \in \mathbb{R}^d : \forall i \in [d], x^{(i)} \notin [b_1^{(i)}, b_1^{(i)} + 1/R) \cup (b_2^{(i)} - 1/R, b_2^{(i)})\right\}.$$

a) There exists a neural network $f_{ind,[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that for $\boldsymbol{x} \in K_{1/R}$

$$f_{ind,[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x}) = \mathbb{I}_{[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x}), \tag{A.1}$$

and for $\boldsymbol{x} \in \mathbb{R}^d$

$$\left|f_{ind,[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x}) - \mathbb{I}_{[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x})\right| \leq 1, \tag{A.2}$$

and for every neural network $\boldsymbol{g}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ in $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$

$$f_{ind,[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}\left(L+2, \max(r, 4d), \max\left(B_w, \frac{R}{1-\nu^2}\right), \max(B_b, 1+a)\right).$$

b) Let $|s| \leq R$. Then there exists the network $f_{test}(\cdot) : \mathbb{R}^{3d+1} \to \mathbb{R}$ such that for $\boldsymbol{x} \in K_{1/R}$

$$f_{test}(\boldsymbol{x}, \boldsymbol{b}_1, \boldsymbol{b}_2, s) = s \cdot \mathbb{I}_{[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x}), \tag{A.3}$$

and for $\boldsymbol{x} \in \mathbb{R}^d$

$$\left|f_{test}(\boldsymbol{x}, \boldsymbol{b}_1, \boldsymbol{b}_2, s) - s \cdot \mathbb{I}_{[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x})\right| \leq |s|, \tag{A.4}$$

and for $\boldsymbol{g}_1(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ in $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r_1, B_w, B_b)$, $\boldsymbol{g}_2(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ in $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r_2, B_w, B_b)$ and $\boldsymbol{g}_3(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ in $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r_3, B_w, B_b)$, we have

$$\begin{aligned} f_{test}(\boldsymbol{g}_1(\cdot), \boldsymbol{g}_2(\cdot), \boldsymbol{g}_3(\cdot), s) \in \tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}\Bigg(& L+2, \max(r_1+r_2+r_3, 8d+4), \\ & \max\left(B_w, \frac{R^2}{1-\nu^2}\right), B_b\Bigg). \end{aligned} \tag{A.5}$$

Proof. For the network f_{ρ_0} defined on Lemma A.2 with $k = 1$, we define

$$\begin{aligned} f_{ind,[\boldsymbol{b}_1,\boldsymbol{b}_2)}(\boldsymbol{x}) &:= f_{\rho_0}\Big(1 - R \cdot \sum_{i=1}^{d} \left(f_{\rho_0}\left(b_1^{(i)} + 1/R - x^{(i)}\right) + f_{\rho_0}\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big) \\ &= \rho_0\Big(1 - R \cdot \sum_{i=1}^{d} \left(\rho_0\left(b_1^{(i)} + 1/R - x^{(i)}\right) + \rho_0\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big), \end{aligned}$$

and

$$\begin{aligned} f_{test}(\boldsymbol{x}, \boldsymbol{b}_1, \boldsymbol{b}_2, s) &:= f_{\rho_0}\Big(f_{id}(s) - R^2 \cdot \sum_{i=1}^{d} \left(f_{\rho_0}\left(b_1^{(i)} + 1/R - x^{(i)}\right) + f_{\rho_0}\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big) \\ &\quad - f_{\rho_0}\Big(-f_{id}(s) - R^2 \cdot \sum_{i=1}^{d} \left(f_{\rho_0}\left(b_1^{(i)} + 1/R - x^{(i)}\right) + f_{\rho_0}\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big) \\ &= \rho_0\Big(f_{id}(s) - R^2 \cdot \sum_{i=1}^{d} \left(\rho_0\left(b_1^{(i)} + 1/R - x^{(i)}\right) + \rho_0\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big) \\ &\quad - \rho_0\Big(-f_{id}(s) - R^2 \cdot \sum_{i=1}^{d} \left(\rho_0\left(b_1^{(i)} + 1/R - x^{(i)}\right) + \rho_0\left(x^{(i)} - b_2^{(i)} + 1/R\right)\right)\Big). \end{aligned}$$

Then, (A.1), (A.2), (A.3) and (A.4) hold by Lemma 6 of Kohler and Langer (2021b), and (A.5) holds by Lemma A.2. ■

Lemma A.4. *Let $\nu \in [0, 1)$ and sufficiently large $R \in \mathbb{N}$ be given.*

a) There exists a neural network $f_{mult}(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ such that

$$|f_{mult}(\boldsymbol{x}) - x^{(1)}x^{(2)}| \leq c_3 \cdot 4^{-R}$$

for $\boldsymbol{x} \in [-a, a]^2$, and for every neural network $\boldsymbol{g}(\cdot) : \mathbb{R}^d \to [-a, a]^2$ in the class $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$,

$$f_{mult}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L + R, \max(r, 24), \max(B_w, c_4), \max(B_b, c_5)),$$

where c_3, c_4 and c_5 are constants not depending on R.

b) There exists a neural network $f_{mult,d}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that

$$\left| f_{mult,d}(\boldsymbol{x}) - \prod_{i=1}^{d} x^{(i)} \right| \leq c_6 \cdot 4^{-R}$$

for $\boldsymbol{x} \in [-a, a]^d$, and for every neural network $\boldsymbol{g}(\cdot) : \mathbb{R}^d \to [-a, a]^d$ in the class $\tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$,

$$f_{mult,d}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L + R\lceil\log_2(d)\rceil, \max(r, 24d), \max(B_w, c_4), \max(B_b, c_5)),$$

where c_6 is a constant not depending on R.

c) For $N \in \mathbb{N}$, let $m_1, \ldots, m_{\binom{d+N}{d}}$ denote all monomials of the form

$$\prod_{k=1}^{d} \left(x^{(k)}\right)^{\alpha_k}$$

for some $\alpha_1, \ldots, \alpha_d \in \mathbb{N}_0$ such that $\alpha_1 + \cdots + \alpha_d \leq N$. For $u_1, \ldots, u_{\binom{d+N}{d}} \in [-1, 1]$, define

$$p\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) := \sum_{i=1}^{\binom{d+N}{d}} u_i \cdot y_i \cdot m_i(\boldsymbol{x}).$$

Then, there exists a neural network $f_p(\cdot, \ldots, \cdot) : \mathbb{R}^{[d+\binom{d+N}{d}]} \to \mathbb{R}$ such that

$$\left| f_p\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) - p\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) \right| \leq c_7 \cdot 4^{-R}$$

for $\boldsymbol{x} \in [-a, a]^d$ and $y_i \in [-a, a]$, and for every neural network $\boldsymbol{g}_0(\cdot) : \mathbb{R}^d \to [-a, a]^d$ in the class $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r_0, B_w, B_b)$ and $g_i(\cdot) : \mathbb{R}^d \to [-a, a]$ in the class $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r_i, B_w, B_b)$ for $i \in [\binom{d+N}{d}]$, we have

$$\begin{aligned} & f_p\left(\boldsymbol{g}_0(\cdot), g_1(\cdot), \ldots, g_{\binom{d+N}{d}}(\cdot)\right) \\ & \in \tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}\Bigg(L + R\lceil \log_2(N+1) \rceil, \max\left(\sum_{i=0}^{\binom{d+N}{d}} r_i,\ 24(N+1)\binom{d+N}{d}\right), \\ & \qquad\qquad\qquad\qquad \max(B_w, c_4), \max(B_b, c_5)\Bigg), \end{aligned}$$

where c_7 is a constant not depending on R.

Proof. a) We define $f_\wedge(\cdot) : \mathbb{R} \to \mathbb{R}$ as

$$f_\wedge(x) := 2 f_{\rho_0}(x) - 4 f_{\rho_0}(x - 0.5) + 2 f_{\rho_0}(x - 1),$$

where f_{ρ_0} is defined on Lemma A.2. Then, we can check that $f_\wedge(\cdot)$ satisfies $f_\wedge(x) = 2x \cdot \mathbb{I}(0 \leq x < 0.5) + 2(1 - x) \cdot \mathbb{I}(0.5 \leq x < 1)$ for every x. Also, for any $g(\cdot) : \mathbb{R} \to \mathbb{R}$ in $\tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L, r, B_w, B_b)$, we have $f_\wedge(g(\cdot)) \in \tilde{\mathcal{F}}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}(L + 1, \max(r, 6), \max(B_w, \frac{4}{1-\nu^2}), B_b)$.

Now, for $\hat{f}_{1,0}(x) = \hat{f}_{2,0}(x) = x$ and $\hat{f}_{3,0}(x) = 0$, we define

$$\begin{aligned} \hat{f}_{1,l}(x) &= f_{id}(\hat{f}_{1,l-1}(x)) \\ \hat{f}_{2,l}(x) &= f_\wedge(\hat{f}_{2,l-1}(x)) \\ \hat{f}_{3,l}(x) &= f_{id}(\hat{f}_{3,l-1}(x)) - f_\wedge(\hat{f}_{2,l-1}(x))/2^{2l} \end{aligned}$$

for $l \in [R - 1]$ and

$$f_{sq_{[0,1]}}(x) = f_{id}(\hat{f}_{1,R-1}(x)) - f_\wedge(\hat{f}_{2,R-1}(x))/2^{2R} + f_{id}(\hat{f}_{3,R-1}(x)).$$

By the proof of Lemma 20 of Kohler and Langer (2021c), we obtain

$$|f_{sq_{[0,1]}}(x) - x^2| \leq 2^{-2R-2}$$

for $x \in [0,1]$. Note that $f_{sq_{[0,1]}}(g(\cdot))$ requires r neurons in each of the 1st through Lth hidden layers and $2+6+2$ neurons in each of the $(L+1)$th through $(L+R)$th hidden layers. Hence, we have $f_{sq_{[0,1]}}(g(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L+R, \max(r,10), \max(B_w, \frac{4}{1-\nu^2}), B_b)$.

Now, with the additional network $f_{tran} : [-2a, 2a] \to [0,1]$ such that

$$f_{tran}(x) := \frac{x}{4a} + \frac{1}{2},$$

we define

$$f_{sq}(x) := 16a^2 f_{sq_{[0,1]}}(f_{tran}(x)) - 4a f^R_{id}(x) - 4a^2.$$

Then, for $x \in [-2a, 2a]$, we have

$$\begin{aligned}
|f_{sq}(x) - x^2| =& |(16a^2 f_{sq_{[0,1]}}(f_{tran}(x)) - 4ax - 4a^2) - (4a f_{tran}(x) - 2a)^2| \\
=& |(16a^2 f_{sq_{[0,1]}}(f_{tran}(x)) - 16a^2 f_{tran}(x) + 4a^2) - (4a f_{tran}(x) - 2a)^2| \\
\leq& 16a^2 \cdot |f_{sq_{[0,1]}}(f_{tran}(x)) - (f_{tran}(x))^2| \\
\leq& 16a^2 \cdot 4^{-R-1} \\
=& 4a^2 \cdot 4^{-R}.
\end{aligned}$$

and

$$f_{sq}(g(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L+R, \max(r,12), \max(B_w, c_4), \max(B_b, c_5)),$$

where $c_4 = \max(16a^2, \frac{4}{1-\nu^2})$ and $c_5 = 4a^2$.

Now, we define $f_{mult}(\cdot,\cdot) : [-a,a]^2 \to \mathbb{R}$ as

$$f_{mult}(x,y) := \frac{1}{4} f_{sq}(x+y) - \frac{1}{4} f_{sq}(x-y).$$

Since $|x+y| \leq 2a$ and $|x-y| \leq 2a$, we have

$$\begin{aligned}
|f_{mult}(x,y) - xy| =& |f_{mult}(x,y) - \frac{1}{4}((x+y)^2 - (x-y)^2)| \\
\leq& \frac{1}{4}|f_{sq}(x+y) - (x+y)^2| + \frac{1}{4}|f_{sq}(x-y) - (x-y)^2| \\
\leq& 2a^2 \cdot 4^{-R}
\end{aligned}$$

for any $x, y \in [-a,a]$. Also, we have

$$f_{mult}(g(\cdot)^{(1)}, g(\cdot)^{(2)}) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L+R, \max(r,24), \max(B_w, c_4), \max(B_b, c_5)).$$

b) For $q = \lceil \log_2(d) \rceil$, we consider $\tilde{\boldsymbol{x}} := (x^{(1)}, \ldots, x^{(d)}, 1, \ldots, 1)^\top \in [-a,a]^{2^q}$. In the first R layers, we compute

$$f_{mult}(\tilde{x}^{(1)}, \tilde{x}^{(2)}), \ldots, f_{mult}(\tilde{x}^{(2^q-1)}, \tilde{x}^{(2^q)}),$$

which can be formulated by R hidden layers and $24 \cdot 2^{q-1} \leq 24d$ neurons in each of the hidden layers. We define $f_{mult,d} : \mathbb{R}^d \to \mathbb{R}$ as the deep neural network which iteratively pairing those outputs and applying f_{mult}. Then, we have

$$f_{mult,d}(\boldsymbol{g}(\cdot)) \in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L + R\lceil \log_2(d) \rceil, \max(r, 24d), \max(B_w, c_4), \max(B_b, c_5)).$$

By Lemma 8 of Kohler and Langer (2021b), we obtain the other assertion.

c) Using $f_{mult,d}$, we can construct a neural network $f_{m_i}(\cdot, \ldots, \cdot) : [-a, a]^{[d+\binom{d+N}{d}]} \to \mathbb{R}$ for $i \in [\binom{d+N}{d}]$ such that

$$\left| f_{m_i}\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) - y_i \cdot m_i(\boldsymbol{x}) \right| \leq c_8 4^{-R}$$

and

$$\begin{aligned} &f_{m_i}\left(\boldsymbol{g}_0(\cdot), g_1(\cdot), \ldots, g_{\binom{d+N}{d}}(\cdot)\right) \\ &\in \tilde{\mathcal{F}}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}\left(L + R\lceil \log_2(N+1) \rceil, \max\Bigg(\sum_{i=0}^{\binom{d+N}{d}} r_i, 24(N+1)\Bigg), \max(B_w, c_4), \max(B_b, c_5)\right), \end{aligned} \tag{A.6}$$

where c_8 is a constant not depending on R.

Now, we define a network $f_p(\cdot, \ldots, \cdot)$ as

$$f_p(\cdot, \ldots, \cdot) := \sum_{i=1}^{\binom{d+N}{d}} u_i \cdot f_{m_i}(\cdot, \ldots, \cdot),$$

which satisfies

$$\begin{aligned} &\left| f_p\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) - p\left(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) \right| \\ \leq &\sum_{i=1}^{\binom{d+N}{d}} |u_i| \cdot \left| f_{m_i}(\boldsymbol{x}, y_1, \ldots, y_{\binom{d+N}{d}}) - y_i \cdot m_i(\boldsymbol{x}) \right| \\ \leq & c_8 \cdot \binom{d+N}{d} \cdot \max_i |u_i| \cdot 4^{-R}. \end{aligned}$$

Also, we have the other assertion by (A.6).

■

A.4 Proof for Theorem 1

We follow the notations and partitions of Kohler and Langer (2021b). For a half-open cube $C \subset [-a, a]^d$ with length $s > 0$, which is defined by

$$C = \left\{ \boldsymbol{x} : C^{(j)}_{left} \leq x^{(j)} < C^{(j)}_{left} + s,\ j \in [d] \right\},$$

we denote the "bottom left" corner of C by $\boldsymbol{C}_{left}$. Also, for a half-open cube $C \subset [-a,a]^d$ with length s and $0 < \delta < \frac{s}{2}$, we denote $C_\delta^0 \subset C$ as the half-open cube which contains all $\boldsymbol{x} \in C$ that lie with a distance of at least δ to the boundaries of C. i.e.,

$$C_\delta^0 = \Big\{ \boldsymbol{x} : C_{left}^{(j)} + \delta \le x^{(j)} < C_{left}^{(j)} + s - \delta,\ j \in [d] \Big\}.$$

We partition $[-a,a)^d$ into M^d and M^{2d} equal-sizes half-open cubes (i.e., length of $2a/M$ and $2a/M^2$). We denote these partitions as

$$\mathcal{P}_1 := \{C_{i,1}\}_{i \in [M^d]}$$

and

$$\mathcal{P}_2 := \{C_{j,2}\}_{j \in [M^{2d}]}.$$

Furthermore, let

$$\boldsymbol{v}_1, \dots, \boldsymbol{v}_{M^d} \in \left\{ 0, \frac{2a}{M^2}, \dots, \frac{2a(M-1)}{M^2} \right\}^d$$

be the d-dimensional M^d different vectors. We denote the half-open cubes $\tilde{C}_{1,i}, \dots, \tilde{C}_{M^d,i}$ as the half-open cubes of $\mathcal{P}_2$ that are contained in $C_{i,1}$ and ordered in such a way that

$$(\tilde{\boldsymbol{C}}_{k,i})_{left} = (\boldsymbol{C}_{i,1})_{left} + \boldsymbol{v}_k$$

holds for all $k \in [M^d]$ and $i \in [M^d]$. Then, we have

$$\mathcal{P}_2 = \{C_{j,2}\}_{j \in [M^{2d}]} = \{\tilde{C}_{k,i}\}_{k \in [M^d]\}, i \in [M^d]}.$$

For a partition $\mathcal{P}$ ($=\mathcal{P}_1$ or $\mathcal{P}_2$) and $\boldsymbol{x} \in [-a,a)^d$, we denote $C_{\mathcal{P}}(\boldsymbol{x})$ as the half-open cube of $\mathcal{P}$ which includes $\boldsymbol{x}$.

Lemma A.5. *For $\beta > 0$ and $K \ge 1$, let $f \in \mathcal{H}_d^\beta(K)$ be the β-Hölder class function defined on $[-a,a]^d$. For any $\nu \in [0,1)$ and sufficiently large $M \in \mathbb{N}$, there exists a neural network $f_{net,\mathcal{P}_2}(\cdot) : [-a,a]^d \to \mathbb{R}$ such that*

$$f_{net,\mathcal{P}_2}(\cdot) \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \Big(\lceil c_9 \log_2 M \rceil, c_{10} M^d, c_{11} \Big)$$

and

$$|f_{net,P_2}(\boldsymbol{x}) - f(\boldsymbol{x})| \le c_{12} \frac{1}{M^{2\beta}}$$

holds for all $\boldsymbol{x} \in \bigcup_{j \in [M^{2d}]} (C_{j,2})^0_{1/M^{2\beta+2}}$ and

$$|f_{net,P_2}(\boldsymbol{x})| \le c_{13}$$

holds for all $\boldsymbol{x} \in [-a,a]^d$, where $c_9, c_{10}, c_{11}, c_{12}$ and c_{13} are constants not depending on M.

Proof. We construct the networks

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{1,1} &= \left(\hat{\phi}_{1,1}^{(1)}, \dots, \hat{\phi}_{1,1}^{(d)} \right) = f_{id}^2(\boldsymbol{x}), \\ \hat{\boldsymbol{\phi}}_{2,1} &= \left(\hat{\phi}_{2,1}^{(1)}, \dots, \hat{\phi}_{2,1}^{(d)} \right) = \sum_{i \in [M^d]} (\boldsymbol{C}_{i,1})_{left} \cdot f_{ind, C_{i,1}}(\boldsymbol{x}) \end{aligned} \tag{A.7}$$

using Lemma A.2 and Lemma A.3, where R in Lemma A.3 is chosen as M. Also, for $j \in [M^d]$ and $\boldsymbol{l} \in \{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{\binom{d+\lfloor\beta\rfloor}{d}}\} := \{\boldsymbol{l} \in \mathbb{N}_0^d,\ \|\boldsymbol{l}\|_1 \leq \lfloor\beta\rfloor\}$, we construct the network

$$\hat{\phi}_{3,1}^{(\boldsymbol{l},j)} = \sum_{i\in[M^d]} (\partial^{\boldsymbol{l}} f)\left((\tilde{\boldsymbol{C}}_{j,i})_{left}\right) \cdot f_{ind,C_{i,1}}(\boldsymbol{x})$$

using Lemma A.3, where R in Lemma A.3 is chosen as $R = M$. Then,

$$\left(\hat{\boldsymbol{\phi}}_{1,1}, \hat{\boldsymbol{\phi}}_{2,1}, \hat{\phi}_{3,1}^{(\boldsymbol{l}_1,1)}, \ldots, \hat{\phi}_{3,1}^{\left(\boldsymbol{l}_{\binom{d+\lfloor\beta\rfloor}{d}}, M^d\right)}\right) \tag{A.8}$$

requires 2 hidden layers, $2d + M^d \cdot 4d$ neurons in each of the hidden layers, the absolute values of weights are bounded by $\frac{M}{1-\nu^2}$ and the absolute values of biases are bounded by $a+1$. In other words,

$$(A.8) \in \tilde{\mathcal{F}}_{\boldsymbol{\rho}_\nu}^{\text{DNN}}\left(2, 2d + 4dM^d, \frac{M}{1-\nu^2}, a+1\right).$$

Next, on the top of the (A.8), we construct the networks

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{1,2} &= \left(\hat{\phi}_{1,2}^{(1)}, \ldots, \hat{\phi}_{1,2}^{(d)}\right) = f_{id}^2(\hat{\boldsymbol{\phi}}_{1,1}), \\ \hat{\boldsymbol{\phi}}_{2,2} &= \left(\hat{\phi}_{2,2}^{(1)}, \ldots, \hat{\phi}_{2,2}^{(d)}\right), \end{aligned} \tag{A.9}$$

where

$$\hat{\phi}_{2,2}^{(i)} = \sum_{j=1}^{M^d} f_{test}\left(\hat{\boldsymbol{\phi}}_{1,1}, \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_j, \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_j + \frac{2a}{M^2} \cdot \boldsymbol{1}, \hat{\phi}_{2,1}^{(i)} + v_j^{(i)}\right)$$

is constructed using Lemma A.3, where R in Lemma A.3 is chosen as M. Also, for $\boldsymbol{l} \in \{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{\binom{d+\lfloor\beta\rfloor}{d}}\}$, we construct the networks

$$\hat{\phi}_{3,2}^{(\boldsymbol{l})} = \sum_{j=1}^{M^d} f_{test}\left(\hat{\boldsymbol{\phi}}_{1,1}, \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_j, \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_j + \frac{2a}{M^2} \cdot \boldsymbol{1}, \hat{\phi}_{3,1}^{(\boldsymbol{l},j)}\right)$$

using Lemma A.3, where R in Lemma A.3 is chosen as $R = M$. Then,

$$\left(\hat{\boldsymbol{\phi}}_{1,2}, \hat{\boldsymbol{\phi}}_{2,2}, \hat{\phi}_{3,2}^{(\boldsymbol{l}_1)}, \ldots, \hat{\phi}_{3,2}^{\left(\boldsymbol{l}_{\binom{d+\lfloor\beta\rfloor}{d}}\right)}\right) \tag{A.10}$$

requires 2+2 hidden layers, $\max\left(2d + 4dM^d, 2d + d \cdot M^d \cdot (8d+4) + \binom{d+\lfloor\beta\rfloor}{d} \cdot M^d \cdot (8d+4)\right)$ neurons in each of the hidden layers, the absolute values of weights are bounded by $\frac{M^2}{1-\nu^2}$ and the absolute values of biases are bounded by $a+1$. In other words,

$$(A.10) \in \tilde{\mathcal{F}}_{\boldsymbol{\rho}_\nu}^{\text{DNN}}\left(4, 2d + (8d+4)M^d \cdot \left(\binom{d+\lfloor\beta\rfloor}{d} + d\right), \frac{M^2}{1-\nu^2}, a+1\right).$$

Note that by Lemma A.2 and Lemma A.3, we have $\hat{\boldsymbol{\phi}}_{1,2}(\boldsymbol{x}) = \boldsymbol{x}$ for $\boldsymbol{x} \in [-a,a]^d$ and $\hat{\boldsymbol{\phi}}_{2,2}(\boldsymbol{x}) = (C_{\mathcal{P}_2}(\boldsymbol{x}))_{left}$ for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]}(C_{i,2})^0_{1/M^{2p+2}}$.

Then, on the top of the (A.10), we construct the network

$$\tilde{f}_{net,\mathcal{P}_2}(x) = f_p\left(\hat{\boldsymbol{\phi}}_{1,2} - \hat{\boldsymbol{\phi}}_{2,2}, \hat{\phi}_{3,2}^{(\boldsymbol{l}_1)}, \dots, \hat{\phi}_{3,2}^{\left(\boldsymbol{l}_{\binom{d+\lfloor\beta\rfloor}{d}}\right)}\right)$$

using Lemma A.4, where the coefficients $u_1, \dots, u_{\binom{d+N}{d}}$ in Lemma A.4 are chosen as $u_i = \frac{1}{\boldsymbol{l}_i!}$ and N and R in Lemma A.4 are chosen as $N = \max(1, \lfloor\beta\rfloor)$ and $R = \lceil \log_2\left(M^\beta\right)\rceil$, respectively. By Lemma 3 of Kohler and Langer (2021b), for all $\boldsymbol{x} \in \bigcup_{j\in[M^{2d}]}(C_{j,2})^0_{1/M^{2\beta+2}}$

$$|\tilde{f}_{net,P_2}(\boldsymbol{x}) - f(\boldsymbol{x})| \le c_{14} \cdot (\max(2a, K))^{4(\lfloor\beta\rfloor+1)} \cdot \frac{1}{M^{2\beta}}, \tag{A.11}$$

where c_{14} is a constant not depending on M and for all $\boldsymbol{x} \in [-a,a]^d$

$$|\tilde{f}_{net,P_2}(\boldsymbol{x})| \le 1 + e^{2ad}K. \tag{A.12}$$

Note that $\tilde{f}_{net,\mathcal{P}_2}(\cdot)$ requires $4 + \max(1, \lceil\log_2(\lfloor\beta\rfloor + 1)\rceil) \cdot \lceil\log_2\left(M^\beta\right)\rceil$ hidden layers and $\max\left(2d + (8d+4)M^d \cdot \left(\binom{d+\lfloor\beta\rfloor}{d} + d\right), 24(\lfloor\beta\rfloor+1)\cdot\binom{d+\lfloor\beta\rfloor}{d}\right)$ neurons in each of the hidden layers.

Also, we define $U_l(\tilde{f}_{net,\mathcal{P}_2}) \in \mathbb{R}^+$ for $l \in [L(\tilde{f}_{net,\mathcal{P}_2}) + 1]$ as

$$\begin{aligned} U_l(\tilde{f}_{net,\mathcal{P}_2}) &:= \frac{M^2}{1-\nu^2} && l \in \{1,2,3,4,5\}, \\ U_l(\tilde{f}_{net,\mathcal{P}_2}) &:= c_4 && l \in \{6, \dots, L(\tilde{f}_{net,\mathcal{P}_2}) + 1\}, \end{aligned}$$

where c_4 is a constant defined on Lemma A.4. Then, $U_l(\tilde{f}_{net,\mathcal{P}_2})$ satisfies

$$\max(1, |\operatorname{vec}(W_l(\tilde{f}_{net,\mathcal{P}_2}))|_\infty) \le U_l(\tilde{f}_{net,\mathcal{P}_2}).$$

Now we choose ζ_l as

$$\zeta_l := \frac{1}{U_l(\tilde{f}_{net,\mathcal{P}_2})}\left(c_4^{L(\tilde{f}_{net,\mathcal{P}_2})+1-5}\left(\frac{M^2}{1-\nu^2}\right)^5\right)^{\frac{1}{L(\tilde{f}_{net,\mathcal{P}_2})+1}}$$

and re-scale the parameters of $\tilde{f}_{net,\mathcal{P}_2}$ using Lemma A.1. We denote this DNN model as $f_{net,\mathcal{P}_2}$. Since $\prod_{l=1}^{L(\tilde{f}_{net,\mathcal{P}_2})+1}\zeta_l = 1$, (A.11) and (A.12) also hold for $f_{net,\mathcal{P}_2}$ by Lemma A.1. Also note that

$$\begin{aligned} \left(c_4^{L(\tilde{f}_{net,\mathcal{P}_2})-4}\left(\frac{M^2}{1-\nu^2}\right)^5\right)^{\frac{1}{L(\tilde{f}_{net,\mathcal{P}_2})+1}} \le& \frac{c_4}{1-\nu^2}\left(M^{10}\right)^{\frac{1}{\beta\log_2(M)}} \\ =& \frac{c_4}{1-\nu^2}2^{10/\beta} =: c_{15}, \end{aligned}$$

which means $|\boldsymbol{\theta}_w(f_{net,\mathcal{P}_2})|_\infty \le c_{15}$ holds. Finally, since $\prod_{l'=1}^{l}\zeta_{l'} \le 1$ for every $l \in [L(f_{net,\mathcal{P}_2}) + 1]$, we have $|\boldsymbol{\theta}_b(f_{net,\mathcal{P}_2})|_\infty \le a+1$. By choosing $c_{11} = \max(c_{15}, a+1)$, we get the last assertion. ■

However, Lemma A.5 only holds for $\boldsymbol{x} \in \bigcup_{j\in[M^{2d}]} (C_{j,2})^0_{1/M^{2\beta+2}}$. To approximate $f(\boldsymbol{x})$ on every $\boldsymbol{x} \in [-a,a)^d$, we need additional networks. The strategy is to approximate $w_{\mathcal{P}_2}(\boldsymbol{x})f(\boldsymbol{x})$ rather than $f(\boldsymbol{x})$, where $w_{\mathcal{P}_2}(\boldsymbol{x})$ is defined by

$$w_{\mathcal{P}_2}(\boldsymbol{x}) = \prod_{j=1}^{d} \max\left(0, 1 - \frac{M^2}{a} \cdot \left|(C_{\mathcal{P}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{a}{M^2} - x^{(j)}\right|\right).$$

Note that $w_{\mathcal{P}_2}(\boldsymbol{x})$ takes maximum value 1 at the center of $C_{\mathcal{P}_2}(\boldsymbol{x})$ and gradually decreases linearly to 0 towards the edge of $C_{\mathcal{P}_2}(\boldsymbol{x})$. Also,

$$w_{\mathcal{P}_2}(\boldsymbol{x}) \le \frac{2}{M^\beta} \tag{A.13}$$

holds for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} C_{i,2} \setminus (C_{i,2})^0_{2/M^{2\beta+2}}$.

We aim to approximate $w_{\mathcal{P}_2}(\boldsymbol{x})f(\boldsymbol{x})$ for every $\boldsymbol{x} \in [-a,a)^d$ by following three steps. The first step is to construct a network $f_{check,\mathcal{P}_2}$, which ascertains whether $\boldsymbol{x}$ falls within the boundaries of $\mathcal{P}_2$ or not.

Lemma A.6. *For any $\nu \in [0,1)$ and sufficiently large $M \in \mathbb{N}$, there exists a neural network $f_{check,\mathcal{P}_2}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that*

$$f_{check,\mathcal{P}_2}(\cdot) \in \mathcal{F}^{\text{DNN}}_{\boldsymbol{\rho}_\nu}\left(L(f_{net,\mathcal{P}_2}), c_{16}M^d, c_{17}\right)$$

and

$$f_{check,\mathcal{P}_2}(\boldsymbol{x}) = 1 \tag{A.14}$$

for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} C_{i,2} \setminus (C_{i,2})^0_{1/M^{2\beta+2}}$,

$$f_{check,\mathcal{P}_2}(\boldsymbol{x}) = 0 \tag{A.15}$$

for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} (C_{i,2})^0_{2/M^{2\beta+2}}$ and

$$|f_{check,\mathcal{P}_2}(\boldsymbol{x})| \le 1 \tag{A.16}$$

for $\boldsymbol{x} \in [-a,a)^d$, where c_{16} and c_{17} are constants not depending on M.

Proof. We construct the network

$$\begin{aligned}\bar{f}_{check,\mathcal{P}_2}(\boldsymbol{x}) = 1 - f_{\rho_0}\Bigg(&- f^2_{id}\Big(1 - \sum_{k=1}^{M^d} f_{ind,(C_{k,1})^0_{1/M^{2\beta+2}}}(\boldsymbol{x})\Big) \\ &+ \sum_{k=1}^{M^d} f_{test}\Big(f^2_{id}(\boldsymbol{x}), \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_k + \frac{1}{M^{2\beta+2}} \cdot \mathbf{1}, \hat{\boldsymbol{\phi}}_{2,1} + \boldsymbol{v}_k + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2\beta+2}} \cdot \mathbf{1}, 1\Big)\Bigg)\end{aligned}$$

using Lemma A.2, Lemma A.3 and (A.7), where R in Lemma A.3 is chosen as M. Further, we construct the network

$$\tilde{f}_{check,\mathcal{P}_2}(\boldsymbol{x}) = f_{id}^{L(f_{net,\mathcal{P}_2})-5}\left(\bar{f}_{check,\mathcal{P}_2}(\boldsymbol{x})\right).$$

Then, by Lemma 10 of Kohler and Langer (2021b), (A.14), (A.15) and (A.16) hold for $\tilde{f}_{check,\mathcal{P}_2}$. Note that $\tilde{f}_{check,\mathcal{P}_2}(\cdot)$ requires $L(f_{net,\mathcal{P}_2})$ hidden layers and $\max(4dM^d + 2d + 4dM^d, 2 + (8d+4)M^d)$ neurons in each of the hidden layers.

Also, we define $U_l(\tilde{f}_{check,\mathcal{P}_2}) \in \mathbb{R}^+$ for $l \in [L(\tilde{f}_{check,\mathcal{P}_2}) + 1]$ as

$$\begin{aligned} U_l(\tilde{f}_{check,\mathcal{P}_2}) &:= \frac{M^2}{1-\nu^2} \qquad && l \in \{1, \dots, 6\}, \\ U_l(\tilde{f}_{check,\mathcal{P}_2}) &:= 1 && l \in \{7, \dots, L(f_{net,\mathcal{P}_2}) + 1\}, \end{aligned}$$

which satisfies $\max(1, |\operatorname{vec}(W_l(\tilde{f}_{check,\mathcal{P}_2}))|_\infty) \leq U_l(\tilde{f}_{check,\mathcal{P}_2})$. Now we choose ζ_l as

$$\zeta_l := \frac{1}{U_l(\tilde{f}_{check,\mathcal{P}_2})} \left(\left(\frac{M^2}{1-\nu^2} \right)^6 \right)^{\frac{1}{L(\tilde{f}_{check,\mathcal{P}_2})+1}}$$

and re-scale the parameters of $\tilde{f}_{check,\mathcal{P}_2}$ using Lemma A.1. We denote this DNN model as $f_{check,\mathcal{P}_2}$. Since $\prod_{l=1}^{L(\tilde{f}_{check,\mathcal{P}_2})+1} \zeta_l = 1$, (A.14), (A.15) and (A.16) also holds for $f_{check,\mathcal{P}_2}$ by Lemma A.1. Also note that

$$\begin{aligned} \left(\left(\frac{M^2}{1-\nu^2} \right)^6 \right)^{\frac{1}{L(\tilde{f}_{check,\mathcal{P}_2})+1}} &\leq \frac{1}{1-\nu^2} (M^{12})^{\frac{1}{\beta \log_2(M)}} \\ &\leq \frac{1}{1-\nu^2} 2^{12/\beta} =: c_{18}, \end{aligned}$$

which means $|\boldsymbol{\theta}_w(f_{check,\mathcal{P}_2})|_\infty \leq c_{18}$ holds. Finally, since $\prod_{l'=1}^{l} \zeta_{l'} \leq 1$ for every $l \in [L(f_{check,\mathcal{P}_2}) + 1]$, we have $|\boldsymbol{\theta}_b(f_{check,\mathcal{P}_2})|_\infty \leq a + 1$. By choosing $c_{17} = \max(c_{18}, a+1)$, we get the last assertion. ■

The second step is to construct a network $f_{w_{\mathcal{P}_2}}(\cdot)$, which approximates $w_{\mathcal{P}_2}(\cdot)$ on $\boldsymbol{x} \in \bigcup_{j \in [M^{2d}]} (C_{j,2})^0_{1/M^{2\beta+2}}$.

Lemma A.7. *For any $\nu \in [0,1)$ and sufficiently large $M \in \mathbb{N}$, there exists a neural network $f_{w_{\mathcal{P}_2}}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that*

$$f_{w_{\mathcal{P}_2}}(\cdot) \in \mathcal{F}^{\mathrm{DNN}}_{\rho_\nu} \left(\lceil c_{19} \log_2 M \rceil, c_{20} M^d, c_{21} \right)$$

and

$$|f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x})| \leq c_{22} \frac{1}{M^{2\beta}}$$

for $\boldsymbol{x} \in \bigcup_{i \in [M^{2d}]} (C_{i,2})^0_{1/M^{2p+2}}$ and

$$|f_{w_{\mathcal{P}_2}}(\boldsymbol{x})| \leq 2$$

for $\boldsymbol{x} \in [-a,a)^d$, where c_{19}, c_{20}, c_{21} and c_{22} are constants not depending on M.

Proof. For $\hat{\boldsymbol{\phi}}_{1,2} = \left(\hat{\phi}_{1,2}^{(1)}, \dots, \hat{\phi}_{1,2}^{(d)}\right)$ and $\hat{\boldsymbol{\phi}}_{2,2} = \left(\hat{\phi}_{2,2}^{(1)}, \dots, \hat{\phi}_{2,2}^{(d)}\right)$ considered on (A.9), we define

$$\begin{aligned} f_{w_{\mathcal{P}_2,j}}(\boldsymbol{x}) := f_{\rho_0}\left(\frac{M^2}{a} \cdot \left(\hat{\phi}_{1,2}^{(j)} - \hat{\phi}_{2,2}^{(j)}\right)\right) &- 2f_{\rho_0}\left(\frac{M^2}{a} \cdot \left(\hat{\phi}_{1,2}^{(j)} - \hat{\phi}_{2,2}^{(j)} - \frac{a}{M^2}\right)\right) \\ &+ f_{\rho_0}\left(\frac{M^2}{a} \cdot \left(\hat{\phi}_{1,2}^{(j)} - \hat{\phi}_{2,2}^{(j)} - \frac{2a}{M^2}\right)\right) \end{aligned}$$

for $j \in [d]$, where $f_{\rho_0}(\cdot)$ is defined on Lemma A.2. Since

$$\max(0, x) - 2\max(0, x-1) + \max(0, x-2) = \max(0, 1 - |1-x|)$$

holds for every $x \in \mathbb{R}$, we have

$$f_{w_{\mathcal{P}_2,j}}(\boldsymbol{x}) = \max\left(0, 1 - \frac{M^2}{a} \cdot \left|\hat{\phi}_{2,2}^{(j)} + \frac{a}{M^2} - \hat{\phi}_{1,2}^{(j)}\right|\right).$$

Then, we define the network

$$\tilde{f}_{w_{\mathcal{P}_2}}(\boldsymbol{x}) := f_{mult,d}\left(f_{w_{\mathcal{P}_2,1}}(\boldsymbol{x}), \dots, f_{w_{\mathcal{P}_2,d}}(\boldsymbol{x})\right)$$

using Lemma A.4, where R in Lemma A.4 is chosen as $R = \lceil \log_2(M^\beta) \rceil$. By Lemma A.4, we obtain

$$\left|\tilde{f}_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - \prod_{j=1}^{d} \max\left(0, 1 - \frac{M^2}{a} \cdot \left|\hat{\phi}_{2,2}^{(j)} + \frac{a}{M^2} - \hat{\phi}_{1,2}^{(j)}\right|\right)\right| \le c_{22}\frac{1}{M^{2\beta}},$$

where c_{22} is a constant not depending on M. Using this fact, we have

$$\begin{aligned} |\tilde{f}_{w_{\mathcal{P}_2}}(\boldsymbol{x})| \le& \left|\tilde{f}_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - \prod_{j=1}^{d} f_{w_{\mathcal{P}_2,j}}(\boldsymbol{x})\right| + \left|\prod_{j=1}^{d} f_{w_{\mathcal{P}_2,j}}(\boldsymbol{x})\right| \\ \le& 1 + \prod_{j=1}^{d} \max\left(0, 1 - \frac{M^2}{a} \cdot \left|\hat{\phi}_{2,2}^{(j)} + \frac{a}{M^2} - \hat{\phi}_{1,2}^{(j)}\right|\right) \\ \le& 2 \end{aligned} \tag{A.17}$$

for $\boldsymbol{x} \in [-a, a)^d$. Also, for $\boldsymbol{x} \in \bigcup_{i \in [M^{2d}]} (C_{i,2})^0_{1/M^{2p+2}}$, since $\hat{\phi}_{1,2}^{(j)} = x^{(j)}$ and $\hat{\phi}_{2,2}^{(j)} = (C_{\mathcal{P}_2}(\boldsymbol{x}))^{(j)}_{left}$ holds, we have

$$|\tilde{f}_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x})| \le c_{22}\frac{1}{M^{2\beta}} \tag{A.18}$$

for $\boldsymbol{x} \in \bigcup_{i \in [M^{2d}]} (C_{i,2})^0_{1/M^{2p+2}}$. Note that $\tilde{f}_{w_{\mathcal{P}_2}}(\cdot)$ requires $4 + 1 + \lceil \log_2(d) \rceil \lceil \log_2(M^\beta) \rceil$ hidden layers and $\max(2d + M^d \cdot d \cdot (8d+4), 6d, 24d)$ neurons in each of the hidden layers.

Also, we define $U_l(\tilde{f}_{w_{\mathcal{P}_2}}) \in \mathbb{R}^+$ for $l \in [L(\tilde{f}_{w_{\mathcal{P}_2}})+1]$ as

$$\begin{aligned} U_l(\tilde{f}_{w_{\mathcal{P}_2}}) &:= \frac{M^2}{1-\nu^2} && l \in \{1, \dots, 6\}, \\ U_l(\tilde{f}_{w_{\mathcal{P}_2}}) &:= c_4 && l \in \{7, \dots, L(\tilde{f}_{w_{\mathcal{P}_2}})+1\}, \end{aligned}$$

where c_4 is a constant defined on Lemma A.4. Then, $U_l(\tilde{f}_{w_{\mathcal{P}_2}})$ satisfies

$$\max(1, |\operatorname{vec}(W_l(\tilde{f}_{w_{\mathcal{P}_2}}))|_\infty) \le U_l(\tilde{f}_{w_{\mathcal{P}_2}}).$$

Now we choose ζ_l as

$$\zeta_l := \frac{1}{U_l(\tilde{f}_{w_{\mathcal{P}_2}})} \left(c_4^{L(\tilde{f}_{w_{\mathcal{P}_2}})+1-6} \left(\frac{M^2}{1-\nu^2} \right)^6 \right)^{\frac{1}{L(\tilde{f}_{w_{\mathcal{P}_2}})+1}}$$

and re-scale the parameters of $\tilde{f}_{w_{\mathcal{P}_2}}$ using Lemma A.1. We denote this DNN model as $f_{w_{\mathcal{P}_2}}$. Since $\prod_{l=1}^{L(\tilde{f}_{w_{\mathcal{P}_2}})+1} \zeta_l = 1$, (A.17) and (A.18) also hold for $f_{w_{\mathcal{P}_2}}$ by Lemma A.1. Also note that

$$\begin{aligned} \left(c_4^{L(\tilde{f}_{w_{\mathcal{P}_2}})-5} \left(\frac{M^2}{1-\nu^2} \right)^6 \right)^{\frac{1}{L(\tilde{f}_{w_{\mathcal{P}_2}})+1}} &\le \frac{c_4}{1-\nu^2} (M^{12})^{\frac{1}{\beta \log_2(M)}} \\ &= \frac{c_4}{1-\nu^2} 2^{12/\beta} =: c_{23}, \end{aligned}$$

which means $|\boldsymbol{\theta}_w(f_{w_{\mathcal{P}_2}}|_\infty \le c_{23}$ holds. Finally, since $\prod_{l'=1}^{l} \zeta_{l'} \le 1$ for every $l \in [L(f_{w_{\mathcal{P}_2}})+1]$, we have $|\boldsymbol{\theta}_b(f_{w_{\mathcal{P}_2}})|_\infty \le a+1$. By choosing $c_{21} = \max(c_{23}, a+1)$, we get the last assertion. ■

The last step is to construct a network f_{net} which approximates $w_{\mathcal{P}_2} \cdot f$ on $\boldsymbol{x} \in [-a,a)^d$. In this step, we use the networks $f_{net,\mathcal{P}_2}$, $f_{check,\mathcal{P}_2}$ and $f_{w_{\mathcal{P}_2}}$, which are defined on Lemma A.5, Lemma A.6 and Lemma A.7, respectively.

Lemma A.8. *For $\beta > 0$ and $K \ge 1$, let $f \in \mathcal{H}_d^\beta(K)$ be the β-Hölder class function defined on $[-a,a]^d$. For any $\nu \in [0,1)$ and sufficiently large $M \in \mathbb{N}$, there exists a neural network $f_{net}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ such that*

$$f_{net}(\cdot) \in \mathcal{F}_{\rho_\nu}^{\mathrm{DNN}} \left(\lceil c_{24} \log_2 M \rceil, c_{25} M^d, c_{26} \right)$$

and

$$|f_{net}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \cdot f(\boldsymbol{x})| \le c_{27} \cdot \frac{1}{M^{2\beta}}$$

hold for $\boldsymbol{x} \in [-a,a)^d$, where c_{24}, c_{25}, c_{26} and c_{27} are constants not depending on M.

Proof. We construct

$$\begin{aligned} f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) :=& f_{\rho_0}\Big(f_{net,\mathcal{P}_2}(\boldsymbol{x}) - c_{13} \cdot f_{check,\mathcal{P}_2}(\boldsymbol{x})\Big) \\ & - f_{\rho_0}\Big(- f_{net,\mathcal{P}_2}(\boldsymbol{x}) - c_{13} \cdot f_{check,\mathcal{P}_2}(\boldsymbol{x})\Big), \end{aligned}$$

where $f_{\rho_0}(\cdot)$ is defined on Lemma A.2, $f_{net,\mathcal{P}_2}(\cdot)$ and c_{13} are defined on Lemma A.5 and $f_{check,\mathcal{P}_2}(\cdot)$ is defined on Lemma A.6. Note that Since $|f_{net,\mathcal{P}_2}(\boldsymbol{x})| \leq c_{13}$, we have

$$f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) = 0$$

for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} C_{i,2} \setminus (C_{i,2})^0_{1/M^{2\beta+2}}$,

$$f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) = f_{net,\mathcal{P}_2}(\boldsymbol{x})$$

for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} (C_{i,2})^0_{2/M^{2\beta+2}}$ and

$$|f_{net,\mathcal{P}_2,true}(\boldsymbol{x})| \leq |f_{net,\mathcal{P}_2}(\boldsymbol{x})| \leq c_{13}$$

for $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} (C_{i,2})^0_{1/M^{2\beta+2}} \setminus (C_{i,2})^0_{2/M^{2\beta+2}}$. For $L_{\text{diff}} := L(f_{net,\mathcal{P}_2,true}) - L(f_{w_{\mathcal{P}_2}})$, we define the network $f_{net}(\cdot)$ as

$$f_{net}(\boldsymbol{x}) := f_{mult}\left(f_{id}^{\max(0,L_{\text{diff}})}(f_{w_{\mathcal{P}_2}}(\boldsymbol{x})), f_{id}^{\max(0,-L_{\text{diff}})}(f_{net,\mathcal{P}_2,true}(\boldsymbol{x}))\right),$$

where f_{mult} is defined on Lemma A.4 with $R = \lceil \log_2(M^\beta) \rceil$, f_{id} is defined on Lemma A.2 and $f_{w_{\mathcal{P}_2}}$ is defined on Lemma A.7. Then, by Lemma A.4, we have

$$\left| f_{net}(\boldsymbol{x}) - f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) \right| \leq \frac{c_{28}}{M^{2\beta}}$$

for some constant c_{28} not depending on M. From

$$\begin{aligned} |f_{net}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \cdot f(\boldsymbol{x})| \leq & \left| f_{net}(\boldsymbol{x}) - f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) \right| \\ & + \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \cdot f(\boldsymbol{x}) \right|, \end{aligned} \tag{A.19}$$

1. For $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} (C_{i,2})^0_{2/M^{2\beta+2}}$,

$$\begin{aligned} (A.19) \leq & \frac{c_{28}}{M^{2\beta}} + \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) \right| \cdot |f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) - f(\boldsymbol{x})| + |f(\boldsymbol{x})| \cdot \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \right| \\ \leq & \frac{c_{28}}{M^{2\beta}} + 2 \left| f_{net,\mathcal{P}_2}(\boldsymbol{x}) - f(\boldsymbol{x}) \right| + F \cdot \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \right| \\ \leq & \frac{c_{28}}{M^{2\beta}} + \frac{2c_{12}}{M^{2\beta}} + \frac{Fc_{22}}{M^{2\beta}} \end{aligned}$$

holds by Lemma A.5 and Lemma A.7.

2. For $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]}(C_{i,2})^0_{1/M^{2\beta+2}} \setminus (C_{i,2})^0_{2/M^{2\beta+2}}$,

$$\begin{aligned}&(A.19)\\ &\leq \frac{c_{28}}{M^{2\beta}} + |f_{net,\mathcal{P}_2,true}(\boldsymbol{x})| \cdot \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x})\right| + |w_{\mathcal{P}_2}(\boldsymbol{x})| \cdot |f_{net,\mathcal{P}_2,true}(\boldsymbol{x}) - f(\boldsymbol{x})|\\ &\leq \frac{c_{28}}{M^{2\beta}} + c_{13} \cdot \left| f_{w_{\mathcal{P}_2}}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x})\right| + |w_{\mathcal{P}_2}(\boldsymbol{x})| \cdot (c_{13}+F)\\ &\leq \frac{c_{28}}{M^{2\beta}} + \frac{c_{13}c_{22}}{M^{2\beta}} + \frac{2(c_{true}+F)}{M^{2\beta}}\end{aligned}$$

holds by Lemma A.5, Lemma A.7 and (A.13).

3. For $\boldsymbol{x} \in \bigcup_{i\in[M^{2d}]} C_{i,2} \setminus (C_{i,2})^0_{1/M^{2\beta+2}}$,

$$\begin{aligned}(A.19) &= |f_{net}(\boldsymbol{x})| + |w_{\mathcal{P}_2}(\boldsymbol{x}) \cdot f(\boldsymbol{x})|\\ &\leq \frac{c_{28}}{M^{2\beta}} + F \cdot |w_{\mathcal{P}_2}(\boldsymbol{x})|\\ &\leq \frac{c_{28}}{M^{2\beta}} + \frac{2F}{M^{2\beta}},\end{aligned}$$

holds by Lemma A.4 and (A.13).

In conclusion, there exists c_{27} not depending on M such that

$$|f_{net}(\boldsymbol{x}) - w_{\mathcal{P}_2}(\boldsymbol{x}) \cdot f(\boldsymbol{x})| \leq c_{27} \cdot \frac{1}{M^{2\beta}}.$$

Also, since we have

$$f_{net,\mathcal{P}_2,true} \in \mathcal{F}^{\mathrm{DNN}}_{\rho_\nu}\left(\lceil c_9 \log_2 M\rceil + 1, (c_{10}+c_{16})M^d, \max(c_{11}, c_{17}, c_{13})\right),$$

we obtain

$$f_{net} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}\left(\lceil c_{24} \log_2 M\rceil, c_{25}M^d, c_{26}\right),$$

where c_{24}, c_{25} and c_{26} are constants not depending on M. ■

Proof of Theorem 1. We partition $[-2a, 2a)^d$ into M^{2d} equal-sizes half-open cubes (i.e., length of $4a/M^2$). We denote this partition as

$$\mathcal{P}_3 := \{C_{j,3}\}_{j\in[M^{2d}]}.$$

Furthermore, let

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{2^d} \in \left\{0, \frac{2a}{M^2}\right\}^d$$

be the d-dimensional 2^d different vectors. For $k \in [2^d]$ and $j \in [M^{2d}]$, we define

$$C_{j,3,k} := \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} - \boldsymbol{u}_k \in C_{j,3}\}$$

as the slightly shifted block, and define

$$\mathcal{P}_{3,k} := \{C_{j,3,k}\}_{j\in[M^{2d}]}$$

as the slightly shifted partition. In other words, $\mathcal{P}_{3,k}$ is the partition of

$$\mathcal{X}_k := \left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} - \boldsymbol{u}_k \in [-2a, 2a)^d \right\}$$

with equal-sizes half-open cubes $C_{1,3,k}, \ldots, C_{M^{2d},3,k}$.

We can apply Lemma A.8 to partitions $\mathcal{P}_{3,k}$ for $k \in 2^d$, instead of $\mathcal{P}_2$. In other words, there exist neural networks $f_{net,k}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ for $k \in [2^d]$ such that

$$f_{net,k}(\cdot) \in \mathcal{F}^{\mathrm{DNN}}_{\rho_\nu}\left(c_{29} \log_2 M, c_{30} M^d, c_{31}\right)$$

and

$$\left| f_{net,k}(\boldsymbol{x}) - w_{\mathcal{P}_{2,k}}(\boldsymbol{x}) \cdot f(\boldsymbol{x}) \right| \leq c_{32} \cdot \frac{1}{M^{2\beta}}$$

holds for $\boldsymbol{x} \in \mathcal{X}_k$ and hence for $\boldsymbol{x} \in [-a, a)^d$, where c_{29}, c_{30}, c_{31} and c_{32} are constants not depending on M and k and

$$w_{\mathcal{P}_{2,k}}(\boldsymbol{x}) = \prod_{j=1}^{d} \max\left(0, 1 - \frac{M^2}{2a} \cdot \left| (C_{\mathcal{P}_{2,k}}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} - x^{(j)} \right| \right).$$

For any $j \in [d]$, note that

1. If $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} \leq x^{(j)} < (C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2}$, then the half of $\{(C_{\mathcal{P}_{2,k}}(\boldsymbol{x}))^{(j)}_{left}\}_{k=1}^{2^d}$ have value $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left}$ and the other half have value $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} - \frac{2a}{M^2}$. Also,

$$\frac{M^2}{2a}\left| (C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} - x^{(j)} \right| + \frac{M^2}{2a}\left| \left((C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} - \frac{2a}{M^2} \right) + \frac{2a}{M^2} - x^{(j)} \right| = 1.$$

2. If $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} \leq x^{(j)} < (C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{4a}{M^2}$, the half of $\{(C_{\mathcal{P}_{2,k}}(\boldsymbol{x}))^{(j)}_{left}\}_{k=1}^{2^d}$ have value $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left}$ and the other half have value $(C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2}$. Also,

$$\frac{M^2}{2a}\left| (C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} - x^{(j)} \right| + \frac{M^2}{2a}\left| \left((C_{\tilde{\mathcal{P}}_2}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} \right) + \frac{2a}{M^2} - x^{(j)} \right| = 1$$

Hence, by factorization we have

$$\begin{aligned}
\sum_{k=1}^{2^d} w_{\mathcal{P}_{2,k}}(\boldsymbol{x}) &= \sum_{k=1}^{2^d} \prod_{j=1}^{d} \max\left(0, 1 - \frac{M^2}{2a} \cdot \left| (C_{\mathcal{P}_{2,k}}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} - x^{(j)} \right| \right) \\
&= \sum_{k=1}^{2^d} \prod_{j=1}^{d} \left(1 - \frac{M^2}{2a} \cdot \left| (C_{\mathcal{P}_{2,k}}(\boldsymbol{x}))^{(j)}_{left} + \frac{2a}{M^2} - x^{(j)} \right| \right) \\
&= \prod_{j=1}^{d} (1 + 1 - 1) = 1.
\end{aligned}$$

Hence, by defining

$$f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}(\boldsymbol{x}) := \sum_{k=1}^{2^d} f_{net,k}(\boldsymbol{x}),$$

we obtain

$$\left\| f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu} - f \right\|_{\infty,[-a,a]^d} \leq c_{32} \cdot 2^d \cdot \frac{1}{M^{2\beta}}$$

and

$$f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\rho_\nu}\left(\lceil c_{29} \log_2 M \rceil, c_{30} 2^d M^d, c_{31} \right).$$

■

Appendix B. Proofs for Section 4

In this section, we prove the examples and theorems presented in Section 4. In Section B.1, we demonstrate that prior distributions in Example 1, 2 and 3 satisfy Assumption 1. In Section B.2, we describe auxiliary lemmas for demonstrating the concentration results. In Section B.3, B.4 and B.5, we prove Theorem 2, 3 and 4, respectively.

B.1 Proofs for examples in Section 4.1

Proof of Example 1. For any $\boldsymbol{\theta} \in [-\kappa, \kappa]^J$, we have

$$\pi(\boldsymbol{\theta}) = \prod_{j=1}^{J} \pi^{(j)}(\theta^{(j)}) \geq \delta_\kappa^J$$

and hence Assumption 1 holds. ■

Proof of Example 2. We denote $\pi(\cdot|\boldsymbol{\psi})$ as the probability density function of $\Pi_{\boldsymbol{\psi}}$. Then, for every $\boldsymbol{\theta} \in [-\kappa, \kappa]^J$, we have

$$\begin{aligned}\pi(\boldsymbol{\theta}) &= \int_{\boldsymbol{\psi} \in \mathbb{R}^S} \pi(\boldsymbol{\theta}|\boldsymbol{\psi}) d\Xi \\ &\geq \int_{\boldsymbol{\psi} \in \Psi} \pi(\boldsymbol{\theta}|\boldsymbol{\psi}) d\Xi \geq (\delta_1 \delta_\kappa)^J.\end{aligned}$$

Hence, Assumption 1 holds by $\delta_1 \delta_\kappa$. ■

Proof of Example 3. We denote $\boldsymbol{\mu} \in [-B, B]^J$ and $\Sigma \in \mathbb{R}^{J \times J}$ as the mean vector and covariance matrix of Π, respectively. Also, we denote $\lambda_1, \ldots, \lambda_J$ as the eigenvalues of Σ. For any $\boldsymbol{\theta} \in [-\kappa, \kappa]^J$, we have

$$|\boldsymbol{\theta} - \boldsymbol{\mu}|_2 \leq (B + \kappa)\sqrt{J}$$

and hence

$$\begin{aligned}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) &= (B + \kappa)^2 J \cdot \left(\frac{\boldsymbol{\theta} - \boldsymbol{\mu}}{(B + \kappa)\sqrt{J}} \right)^\top \Sigma^{-1} \left(\frac{\boldsymbol{\theta} - \boldsymbol{\mu}}{(B + \kappa)\sqrt{J}} \right) \\ &\leq \frac{(B + \kappa)^2 J}{\lambda_{\min}}.\end{aligned}$$

Also, we have

$$\det(\Sigma) = \prod_{j=1}^{J} \lambda_j \leq (\lambda_{\max})^J.$$

To sum up, we obtain

$$\begin{aligned}\pi(\boldsymbol{\theta}) =& (2\pi)^{-\frac{J}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}) \right) \\ \geq& (2\pi\lambda_{\max})^{-\frac{J}{2}} \exp\left(-\frac{(B + \kappa)^2 J}{2\lambda_{\min}} \right).\end{aligned}$$

Hence, Assumption 1 holds by

$$\delta_\kappa = \frac{1}{\sqrt{2\pi\lambda_{\max}}} \exp\left(-\frac{(B+\kappa)^2}{2\lambda_{\min}}\right).$$

■

B.2 Auxiliary lemmas for posterior concentration results

Lemma B.1. *Consider two DNN models $f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu} : [-a,a]^d \to \mathbb{R}, f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu} : [-a,a]^d \to \mathbb{R}$ with the $(L, \boldsymbol{r})$ architecture, where $L \in \mathbb{N}$, $\boldsymbol{r} = (d, r, r, ..., r, 1)^\top \in \mathbb{N}^{L+2}$ for some $r \in \mathbb{N}$ and $\nu \in [0,1)$. If $|\boldsymbol{\theta}_1|_\infty \le B$, $|\boldsymbol{\theta}_2|_\infty \le B$ and $|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_\infty \le \delta$ holds for some $B > 0$ and $\delta > 0$, then*

$$\left\| f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu} - f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu} \right\|_{\infty,[-a,a]^d} \le a(d+1)(r+1)^L B^L (L+1)\delta$$

holds.

Proof. We denote

$$\begin{aligned} f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}(\cdot) &= A_{L+1,1} \circ \boldsymbol{\rho}_\nu \circ A_{L,1} \cdots \circ \boldsymbol{\rho}_\nu \circ A_{1,1}(\cdot), \\ f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}(\cdot) &= A_{L+1,2} \circ \boldsymbol{\rho}_\nu \circ A_{L,2} \cdots \circ \boldsymbol{\rho}_\nu \circ A_{1,2}(\cdot). \end{aligned}$$

Also, for $l \in [L]$, we define $\boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} : [-a,a]^d \to \mathbb{R}^r$ and $\boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu} : [-a,a]^d \to \mathbb{R}^r$ as the DNN models whose outputs are l-th hidden layer of $f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}$ and $f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}$, respectively. That is,

$$\begin{aligned} \boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu}(\cdot) &= A_{l,1} \circ \boldsymbol{\rho}_\nu \circ A_{l-1,1} \cdots \circ \boldsymbol{\rho}_\nu \circ A_{1,1}(\cdot), \\ \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}(\cdot) &= A_{l,2} \circ \boldsymbol{\rho}_\nu \circ A_{l-1,2} \cdots \circ \boldsymbol{\rho}_\nu \circ A_{1,2}(\cdot). \end{aligned}$$

Also, we denote $\boldsymbol{h}_{\boldsymbol{\theta}_1,L+1,\boldsymbol{\rho}_\nu}(\cdot) = f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}(\cdot)$ and $\boldsymbol{h}_{\boldsymbol{\theta}_2,L+1,\boldsymbol{\rho}_\nu}(\cdot) = f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}(\cdot)$.

For $l \in [L]$, we have

$$\left\| |\boldsymbol{h}_{\boldsymbol{\theta_1},l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \le a(d+1)(r+1)^{l-1} B^l.$$

and hence

$$\begin{aligned} &\left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,l+1,\boldsymbol{\rho}_\nu} - \boldsymbol{h}_{\boldsymbol{\theta}_2,l+1,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &= \left\| |A_{l+1,1} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - A_{l+1,2} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &\le \left\| |A_{l+1,1} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - A_{l+1,2} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &\quad + \left\| |A_{l+1,2} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - A_{l+1,2} \circ \boldsymbol{\rho}_\nu \circ \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &\le (r+1)|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|_\infty \left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &\quad + r|\boldsymbol{\theta}_2|_\infty \left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \\ &\le \delta a(d+1)(r+1)^l B^l \\ &\quad + rB \left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d}. \end{aligned}$$

Since

$$\left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,1,\boldsymbol{\rho}_\nu} - \boldsymbol{h}_{\boldsymbol{\theta}_2,1,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \le a(d+1)\delta$$

holds, we get

$$\left\| |\boldsymbol{h}_{\boldsymbol{\theta}_1,l,\boldsymbol{\rho}_\nu} - \boldsymbol{h}_{\boldsymbol{\theta}_2,l,\boldsymbol{\rho}_\nu}|_\infty \right\|_{\infty,[-a,a]^d} \le a(d+1)(r+1)^{l-1}B^{l-1}l\delta$$

for every $l \in [L+1]$ by induction. ■

Lemma B.2 (Theorem 19.3 of Györfi et al. (2002))**.** *Let $\boldsymbol{X}, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be independent and identically distributed random vectors with values in $\mathbb{R}^d$. Let $K_1, K_2 \ge 1$ be constants and let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to \mathbb{R}$ with the properties*

$$|g(\boldsymbol{x})| \le K_1 \quad (\boldsymbol{x} \in \mathbb{R}^d) \qquad \text{and} \qquad \mathbb{E}(g(\boldsymbol{X})^2) \le K_2 \mathbb{E}(g(\boldsymbol{X})).$$

Let $0 < \tau < 1$ and $\alpha > 0$. Assume that

$$\sqrt{n}\tau\sqrt{1-\tau}\sqrt{\alpha} \ge 288 \max\left\{2K_1, \sqrt{2K_2}\right\}$$

and that, for all $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ and for all $t \ge \frac{\alpha}{8}$,

$$\frac{\sqrt{n}\tau(1-\tau)t}{96\sqrt{2}\max\{K_1, 2K_2\}} \ge \int_{\frac{\tau(1-\tau)t}{16\max\{K_1,2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \left\{g \in \mathcal{G} : \frac{1}{n}\sum_{i=1}^n g(\boldsymbol{x}_i)^2 \le 16t\right\}, \|\cdot\|_{1,n}\right)} du.$$

Then,

$$\mathbf{P}\left\{\sup_{g\in\mathcal{G}} \frac{\left|\mathbb{E}\{g(\boldsymbol{X})\} - \frac{1}{n}\sum_{i=1}^n g(\boldsymbol{X}_i)\right|}{\alpha + \mathbb{E}\{g(\boldsymbol{X})\}} > \tau\right\} \le 60\exp\left(-\frac{n\alpha\tau^2(1-\tau)}{128\cdot 2304\max\left\{K_1^2, K_2\right\}}\right).$$

B.3 Proof of Theorem 2

Without loss of generality, we consider γ in $(2, \frac{5}{2})$. We define $\mathcal{F}_n$ as the set of pairs of truncated DNN with the $(L_n, \boldsymbol{r}_n)$ architecture and variance of the Gaussian noise,

$$\mathcal{F}_n := \left\{ \left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}, \sigma^2\right)^\top \ : \ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n), \ 0 < \sigma^2 \le e^{4n\varepsilon_n^2} \right\},$$

where L_n and r_n are defined on (4). We denote J_n as the number of parameters in the DNN model with the $(L_n, \boldsymbol{r}_n)$ architecture. In other words,

$$J_n := \sum_{l=1}^{L_n+1} (r_n^{(l-1)} + 1) r_n^{(l)}.$$

For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, we denote $P_{(f,\sigma^2),i}$ and $p_{(f,\sigma^2),i}$ as the probability measure and density corresponding to the Gaussian distribution $N(f(\boldsymbol{x}_i), \sigma^2)$, respectively. For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, we define the semimetric h_n on $\mathcal{F}_n$ as the average of the squares of the Hellinger distances between $P_{(f,\sigma^2),i}$. That is,

$$h_n^2\left((f_1, \sigma_1^2), (f_2, \sigma_2^2)\right) := \frac{1}{n}\sum_{i=1}^n h^2\left(P_{(f_1,\sigma_1^2),i}, P_{(f_2,\sigma_2^2),i}\right).$$

Lemma B.3. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, we have*

$$\sup_{\varepsilon > \varepsilon_n} \log \mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{(f,\sigma^2) \in \mathcal{F}_n : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) < \varepsilon\right\}, h_n\right) \lesssim n\varepsilon_n^2$$

under the conditions of Theorem 2.

Proof. We define semimetric d_n on $\mathcal{F}_n$ as

$$d_n^2\left((f_1,\sigma_1^2),(f_2,\sigma_2^2)\right) := \|f_1 - f_2\|_{1,n} + |\sigma_1^2 - \sigma_2^2|^2.$$

Then, $h_n^2(\cdot) \lesssim d_n^2(\cdot)$ holds by by Lemma B.1 of Xie and Xu (2020) and hence $\mathcal{N}\left(\varepsilon, \mathcal{F}_n, h_n\right) \lesssim \mathcal{N}\left(\varepsilon^2, \mathcal{F}_n, d_n^2\right)$. Also, by the fact that $\|f_1 - f_2\|_{1,n} \le \frac{\varepsilon^2}{2}$ and $|\sigma_1^2 - \sigma_2^2|^2 \le \frac{\varepsilon^2}{2}$ implies $\|f_1 - f_2\|_{1,n} + |\sigma_1^2 - \sigma_2^2|^2 \le \varepsilon^2$, we get

$$\begin{aligned}
\mathcal{N}\left(\varepsilon, \mathcal{F}_n, h_n\right) \lesssim & \mathcal{N}\left(\varepsilon^2, \mathcal{F}_n, d_n^2\right) \\
\le & \frac{\sqrt{2}}{\varepsilon} \cdot \exp(4n\varepsilon_n^2) \cdot \mathcal{N}\left(\frac{\varepsilon^2}{2}, T_F \circ \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n, r_n), \|\cdot\|_{1,n}\right) \\
\le & \frac{\sqrt{2}}{\varepsilon} \cdot \exp(4n\varepsilon_n^2) \cdot \mathcal{M}\left(\frac{\varepsilon^2}{2}, T_F \circ \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n, r_n), \|\cdot\|_{1,n}\right) \\
\le & \frac{\sqrt{2}}{\varepsilon} \cdot \exp(4n\varepsilon_n^2) \cdot 3\left(\frac{8eF}{\epsilon^2}\log\frac{12eF}{\epsilon^2}\right)^{V^+_{T_F \circ \mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n, r_n)}} \\
\le & \frac{\sqrt{2}}{\varepsilon} \cdot \exp(4n\varepsilon_n^2) \cdot 3\left(\frac{8eF}{\epsilon^2}\log\frac{12eF}{\epsilon^2}\right)^{V^+_{\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n, r_n)}} \\
\le & \frac{\sqrt{2}}{\varepsilon} \cdot \exp(4n\varepsilon_n^2) \cdot 3\left(\frac{8eF}{\epsilon^2}\log\frac{12eF}{\epsilon^2}\right)^{c_{33}L_n^2 r_n^2 \log(L_n r_n^2)} \qquad \text{(B.1)}
\end{aligned}$$

holds for every $\varepsilon > 0$, where the fourth and last inequalities hold by Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019), respectively. Here, $c_{33} > 0$ is a constant not depending on n. Hence, we obtain

$$\begin{aligned}
& \sup_{\varepsilon > \varepsilon_n} \log \mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{(f,\sigma^2) \in \mathcal{F}_n : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) < \varepsilon\right\}, h_n\right) \\
& \le \log \mathcal{N}\left(\frac{1}{36}\varepsilon_n, \mathcal{F}_n, h_n\right) \\
& \lesssim n\varepsilon_n^2 + L_n^2 r_n^2 \log L_n r_n^2 \log n \\
& \lesssim n\varepsilon_n^2.
\end{aligned}$$

■

Lemma B.4. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, we define*

$$\begin{aligned}
K_i((f_0,\sigma_0^2),(f,\sigma^2)) &= \int \log(p_{(f_0,\sigma_0^2),i}/p_{(f,\sigma^2),i}) dP_{(f_0,\sigma_0^2),i}, \\
V_i((f_0,\sigma_0^2),(f,\sigma^2)) &= \int \left(\log(p_{(f_0,\sigma_0^2),i}/p_{(f,\sigma^2),i}) - K_i((f_0,\sigma_0^2),(f,\sigma^2))\right)^2 dP_{(f_0,\sigma_0^2),i}
\end{aligned}$$

and

$$B_n^*\left((f_0,\sigma_0^2),\varepsilon_n\right) = \Bigg\{(f,\sigma^2)\in\mathcal{F}_n : \frac{1}{n}\sum_{i=1}^n K_i((f_0,\sigma_0^2),(f,\sigma^2)) \le \varepsilon_n^2, \\ \frac{1}{n}\sum_{i=1}^n V_i((f_0,\sigma_0^2),(f,\sigma^2)) \le \varepsilon_n^2\Bigg\}.$$

Then, we have

$$(\Pi\otimes\Xi)\left(B_n^*\left((f_0,\sigma_0^2),\varepsilon_n\right)\right) \gtrsim e^{-n\varepsilon_n^2}$$

under the conditions of Theorem 2.

Proof. For $\varepsilon>0$, define

$$A_n^*\left((f_0,\sigma_0^2),\varepsilon\right) := \Big\{(f,\sigma^2)\in\mathcal{F}_n : \max_i |f(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)| \le \frac{\sigma_0\varepsilon}{2}, \sigma^2\in[\sigma_0^2,(1+\varepsilon^2)\sigma_0^2]\Big\}.$$

Then for every $f\in A_n^*\left((f_0,\sigma_0^2),\varepsilon\right)$ and $i\in[n]$,

$$K_i((f_0,\sigma_0^2),(f,\sigma^2)) = \frac{1}{2}\log\frac{\sigma^2}{\sigma_0^2} + \frac{\sigma_0^2+(f_0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i))^2}{2\sigma^2} - \frac{1}{2} \le \varepsilon^2$$

and

$$\begin{aligned} V_i((f_0,\sigma_0^2),(f,\sigma^2)) =& \mathrm{Var}_{f_0,\sigma_0^2}\left(-\frac{(Y_i-f_0(\boldsymbol{x}_i))^2}{2\sigma_0^2} + \frac{(Y_i-f(\boldsymbol{x}_i))^2}{2\sigma^2}\right) \\ =& \mathrm{Var}_{f_0,\sigma_0^2}\left(-\frac{(Y_i-f_0(\boldsymbol{x}_i))^2}{2\sigma_0^2} + \frac{(Y_i-f_0(\boldsymbol{x}_i)+f_0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i))^2}{2\sigma^2}\right) \\ =& \mathrm{Var}_{f_0,\sigma_0^2}\left(-\frac{1}{2}(1-\frac{\sigma_0^2}{\sigma^2})Z_i^2 + \frac{\sigma_0(f_0(\boldsymbol{x}_i)-f(\boldsymbol{x}_i))Z_i}{\sigma^2}\right) \le \varepsilon^2 \end{aligned}$$

where $Z_i := \frac{Y_i-f_0(\boldsymbol{x}_i)}{\sigma_0}\sim N(0,1)$. Hence, we obtain

$$A_n^*\left((f_0,\sigma_0^2),\varepsilon_n\right) \subset B_n^*\left((f_0,\sigma_0^2),\varepsilon_n\right). \tag{B.2}$$

Also, by Theorem 1 with $M=n^{\frac{1}{2(2\beta+d)}}$, there exists $f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}\in\mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n,r_n,C_3)$ such that

$$\begin{aligned} \left\|f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0\right\|_{\infty,[-a,a]^d} \le& c_1 n^{-\frac{\beta}{(2\beta+d)}} \\ <& \frac{\sigma_0\varepsilon_n}{4} \end{aligned} \tag{B.3}$$

satisfies for sufficiently large n. Note that $\hat{\boldsymbol{\theta}} \in [-C_3, C_3]^{J_n}$ holds. With (B.2), (B.3) and Lemma B.1, we obtain

$$\begin{aligned}
&(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \\
&\geq (\Pi \otimes \Xi)\left(A_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \\
&= \Pi\left(\left\{\boldsymbol{\theta} : \max_i \ \left|T_F \circ f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right| \leq \frac{\sigma_0 \varepsilon_n}{2}\right\}\right) \Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right) \\
&\geq \Pi\left(\left\{\boldsymbol{\theta} : \max_i \ \left|f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right| \leq \frac{\sigma_0 \varepsilon_n}{2}\right\}\right) \Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right) \\
&\geq \Pi\left(\left\{\boldsymbol{\theta} : \max_i \ \left|f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i)\right| \leq \frac{\sigma_0 \varepsilon_n}{4}\right\}\right) \Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right) \\
&\geq \Pi\left(\boldsymbol{\theta} : \boldsymbol{\theta} \in C_n^\star(\hat{\boldsymbol{\theta}})\right) \Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right),
\end{aligned}$$

where $C_n^*(\hat{\boldsymbol{\theta}})$ is defined by

$$C_n^*(\hat{\boldsymbol{\theta}}) := \left\{\boldsymbol{\theta} \in \mathbb{R}^{J_n} : |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|_\infty \leq \frac{\sigma_0 \varepsilon_n}{4a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)}\right\}.$$

By Assumption 1, there exists a constant $\delta_1 > 0$ such that

$$\Pi\left(C_n^*(\hat{\boldsymbol{\theta}})\right) \geq \delta_1^{J_n}\left(\frac{\sigma_0 \varepsilon_n}{2a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)}\right)^{J_n}.$$

Also, since the density function of $\Xi(\sigma^2)$ is continuous and positive, there exists a constant $\delta_2 > 0$ such that $\Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right) \geq \delta_2 \varepsilon_n^2$ by the Extreme Value Theorem. Hence, we obtain

$$\begin{aligned}
(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) &\geq \Pi\left(C_n^*(\hat{\boldsymbol{\theta}})\right) \Xi\left(\left\{\sigma^2 : \sigma^2 \in [\sigma_0^2, (1+\varepsilon_n^2)\sigma_0^2]\right\}\right) \\
&\geq \delta_1^{T_n}\left(\frac{\sigma_0 \varepsilon_n}{2a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)}\right)^{J_n} \delta_2 \varepsilon_n^2 \qquad \text{(B.4)} \\
&\gtrsim \exp\left(-C_2^2 C_1 (\log n) n^{\frac{d}{2\beta+d}} (\log n)^2\right) n^{-1} \\
&\gtrsim e^{-n\varepsilon_n^2} \qquad \text{(B.5)}
\end{aligned}$$

for all but finite many n. ■

Lemma B.5. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we have*

$$\frac{(\Pi \otimes \Xi)\left(\mathcal{F}_n^c\right)}{(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right)} = o(e^{-2n\varepsilon_n^2})$$

under the conditions of Theorem 2.

Proof. From (B.5) and

$$\begin{aligned}(\Pi \otimes \Xi)\left(\mathcal{F}_n^c\right) &= \Xi\left(\sigma^2 > e^{4n\varepsilon_n^2}\right) \\ &\lesssim e^{-4n\varepsilon_n^2},\end{aligned}$$

we obtain the assertion. ∎

Proof of Theorem 2. From Lemma B.3, Lemma B.4, Lemma B.5 and Theorem 4 of Ghosal and van der Vaart (2007), we have

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right] \to 0$$

for every sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^{\infty}$, where the expectation is with respect to $\{Y_i\}_{i=1}^n$. Since

$$\begin{aligned}\left(\|f_1 - f_2\|_{2,n} + |\sigma_1^2 - \sigma_2^2|\right)^2 &\leq 2\left(\|f_1 - f_2\|_{2,n}^2 + |\sigma_1^2 - \sigma_2^2|^2\right) \\ &\lesssim h_n^2\left((f_1,\sigma_1^2),(f_2,\sigma_2^2)\right)\end{aligned}$$

holds by Lemma B.1 of Xie and Xu (2020), we obtain

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) : \|f - f_0\|_{2,n} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right] \to 0$$

for every sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^{\infty}$, where the expectation is with respect to $\{Y_i\}_{i=1}^n$. Note that we can consider

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) : \|f - f_0\|_{2,n} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)}\right] \tag{B.6}$$

as the sequence of bounded random variable. Since (B.6) is uniformly integrable, we have

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) : \|f - f_0\|_{2,n} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right)\right] \to 0, \tag{B.7}$$

where the expectation is with respect to $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$.

Next, we will check the conditions in Lemma B.2 for

$$\begin{aligned}&\mathcal{G} := \left\{g \;:\; g = (T_F \circ f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2, f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n)\right\}, \\ &\tau := \frac{1}{2},\ \alpha := \varepsilon_n^2,\ K_1 = K_2 = 4F^2.\end{aligned}$$

First, it is easy to check $\|g(\boldsymbol{x})\|_\infty \leq 4F^2$ and $\mathbb{E}(g(\boldsymbol{X})^2) \leq 4F^2\mathbb{E}(g(\boldsymbol{X}))$ for $g \in \mathcal{G}$. Also, since

$$\begin{aligned}&\left\|(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2 - (T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2\right\|_{1,n} \\ &= \left\|(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0 + T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}})\right\|_{1,n} \\ &\leq 4F\left\|T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}\right\|_{1,n}\end{aligned}$$

holds for any $f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}, f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n)$, there exists $c_{34} > 0$ such that

$$\begin{aligned}\mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right) \leq& \mathcal{N}\left(\frac{u}{4F}, T_F \circ \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n), \|\cdot\|_{1,n}\right)\\ \leq& \mathcal{M}\left(\frac{u}{4F}, T_F \circ \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n), \|\cdot\|_{1,n}\right)\\ \leq& 3\left(\frac{16eF^2}{u}\log\frac{24eF^2}{u}\right)^{\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n)^+}\\ \lesssim& n^{c_{34}L_n^2 r_n^2 \log(L_n r_n^2)}\end{aligned}$$

for $u \geq n^{-1}$ by Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019). Hence for all $t \geq \frac{\varepsilon_n^2}{8}$,

$$\begin{aligned}\int_{\frac{\tau(1-\tau)t}{16\max\{K_1, 2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right)}du \lesssim& \sqrt{t}\left(n^{\frac{d}{2\beta+d}}(\log n)^4\right)^{\frac{1}{2}}\\ =& o\left(\frac{\sqrt{n}\tau(1-\tau)t}{96\sqrt{2}\max\left\{K_1, 2K_2\right\}}\right)\end{aligned}$$

holds. To sum up, we conclude that

$$\mathbf{P}\left\{\sup_{f \in \mathcal{F}^{\mathrm{DNN}}(L_n, r_n)} \frac{\left|\|f - f_0\|^2_{2,\mathrm{P}_X} - \|f - f_0\|^2_{2,n}\right|}{\varepsilon_n^2 + \|f - f_0\|^2_{2,\mathrm{P}_X}} > \frac{1}{2}\right\} \leq 60\exp\left(-\frac{n\varepsilon_n^2/8}{128 \cdot 2304 \cdot 16F^4}\right) \tag{B.8}$$

holds for all but finite many n by Lemma B.2. By (B.7) and (B.8), we obtain the assertion. ■

B.4 Proof of Theorem 3

Without loss of generality, we consider γ in $(2, \frac{5}{2})$. We define $\mathcal{F}_n$ as the set of truncated DNN with the $(L_n, \boldsymbol{r}_n)$ architecture,

$$\mathcal{F}_n := \left\{T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \ : \ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n)\right\},$$

where L_n and r_n are defined on (4). We denote J_n as the number of parameters in the DNN model with the $(L_n, \boldsymbol{r}_n)$ architecture. In other words,

$$J_n := \sum_{l=1}^{L_n+1} (r_n^{(l-1)} + 1) r_n^{(l)}.$$

For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we denote $P_{f,i}$ and $p_{f,i}$ as the probability measure and density corresponding to the Bernoulli distribution Bernoulli $(\phi \circ f(\boldsymbol{x}_i))$, respectively. Also, for given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we define semimetric h_n on $\mathcal{F}_n$ as the average of the squares of the Hellinger distances between $P_{f,i}$. That is,

$$\begin{aligned}
&h_n^2\left(f_1, f_2\right)\\
&:=\frac{1}{n}\sum_{i=1}^{n} h^2\left(P_{f_1,i}, P_{f_2,i}\right)\\
&=\frac{1}{2n}\sum_{i=1}^{n}\left[\left(\sqrt{\phi\circ f_1(\boldsymbol{x}_i)}-\sqrt{\phi\circ f_1(\boldsymbol{x}_i)}\right)^2+\left(\sqrt{1-\phi\circ f_1(\boldsymbol{x}_i)}-\sqrt{1-\phi\circ f_1(\boldsymbol{x}_i)}\right)^2\right]
\end{aligned}$$

Lemma B.6. *For given $\boldsymbol{x}^{(n)}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$, we have*

$$\sup_{\varepsilon>\varepsilon_n}\log\mathcal{N}\left(\frac{1}{36}\varepsilon,\left\{f\in\mathcal{F}_n : h_n\left(f,f_0\right)<\varepsilon\right\},h_n\right)\lesssim n\varepsilon_n^2.$$

Proof. We define semimetric d_n on $\mathcal{F}_n$ as

$$d_n^2\left(f_1,f_2\right):=\|\phi\circ f_1-\phi\circ f_2\|_{2,n}.$$

Since the infinite norms of $f_1, f_2\in\mathcal{F}_n$ are bounded by F,

$$\begin{aligned}
&d_n^2\left(f_1,f_2\right)\\
&=\frac{1}{2n}\sum_{i=1}^{n}\left\{(\phi\circ f_1(\boldsymbol{x}_i)-\phi\circ f_2(\boldsymbol{x}_i))^2+((1-\phi\circ f_1(\boldsymbol{x}_i))-(1-\phi\circ f_2(\boldsymbol{x}_i)))^2\right\}\\
&=\frac{1}{2n}\sum_{i=1}^{n}\left\{\left(\sqrt{\phi\circ f_1(\boldsymbol{x}_i)}-\sqrt{\phi\circ f_2(\boldsymbol{x}_i)}\right)^2\left(\sqrt{\phi\circ f_1(\boldsymbol{x}_i)}+\sqrt{\phi\circ f_2(\boldsymbol{x}_i)}\right)^2\right.\\
&\qquad\left.+\left(\sqrt{1-\phi\circ f_1(\boldsymbol{x}_i)}-\sqrt{1-\phi\circ f_2(\boldsymbol{x}_i)}\right)^2\left(\sqrt{1-\phi\circ f_1(\boldsymbol{x}_i)}+\sqrt{1-\phi\circ f_2(\boldsymbol{x}_i)}\right)^2\right\}\\
&\gtrsim h_n^2\left(f_1,f_2\right)
\end{aligned}$$

holds. Hence, we obtain

$$\begin{aligned}
\mathcal{N}\left(\varepsilon,\mathcal{F}_n,h_n\right)\lesssim&\mathcal{N}\left(\varepsilon,\mathcal{F}_n,d_n\right)\\
\leq&\mathcal{N}\left(\varepsilon,\mathcal{F}_n,\|\cdot\|_{2,n}\right)\\
\leq&\mathcal{M}\left(\varepsilon,T_F\circ\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n,r_n),\|\cdot\|_{2,n}\right)\\
\leq&3\left(\frac{8eF^2}{\epsilon^2}\log\frac{12eF^2}{\epsilon^2}\right)^{V^+_{T_F\circ\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n,r_n)}}\\
\leq&3\left(\frac{8eF^2}{\epsilon^2}\log\frac{12eF^2}{\epsilon^2}\right)^{V^+_{\mathcal{F}_{\boldsymbol{\rho}}^{\mathrm{DNN}}(L_n,r_n)}}\\
\leq&3\left(\frac{8eF^2}{\epsilon^2}\log\frac{12eF^2}{\epsilon^2}\right)^{c_{35}L_n^2r_n^2\log(L_nr_n^2)}
\end{aligned}$$

holds for every $\varepsilon > 0$, where the second, fourth and last inequalities hold by 1-Lipschitz continuity of ϕ, Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019), respectively. Here, $c_{35} > 0$ is a constant not depending on n. Hence, we obtain

$$\begin{aligned}\sup_{\varepsilon>\varepsilon_n} \log\mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{f \in \mathcal{F}_n : h_n\left(f, f_0\right) < \varepsilon\right\}, h_n\right) \leq& \log\mathcal{N}\left(\frac{1}{36}\varepsilon_n, \mathcal{F}_n, h_n\right)\\ \lesssim& L_n^2 r_n^2 \log L_n r_n^2 \log n\\ \lesssim& n\varepsilon_n^2.\end{aligned}$$

■

Lemma B.7. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we define*

$$\begin{aligned}K_i(f_0, f) =& \int \log(p_{f_0,i}/p_{f,i}) dP_{f_0,i},\\ V_i(f_0, f) =& \int \left(\log(p_{f_0,i}/p_{f,i}) - K_i(f_0, f)\right)^2 dP_{f_0,i},\\ B_n^*\left(f_0, \varepsilon_n\right) =& \left\{f \in \mathcal{F}_n : \frac{1}{n}\sum_{i=1}^n K_i(f_0, f) \leq \varepsilon_n^2,\ \frac{1}{n}\sum_{i=1}^n V_i(f_0, f) \leq \varepsilon_n^2\right\}.\end{aligned}$$

Then, we have

$$\Pi\left(B_n^*\left(f_0, \varepsilon_n\right)\right) \gtrsim e^{-n\varepsilon_n^2}.$$

Proof. For $\varepsilon > 0$, define

$$A_n^*\left(f_0, \varepsilon\right) := \left\{f \in \mathcal{F}_n : \max_i |f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)| \leq \varepsilon\right\}.$$

Then by Lemma 3.2 of van der Vaart and van Zanten (2008), we have

$$A_n^*\left(f_0, \varepsilon_n\right) \subset B_n^*\left(f_0, \varepsilon_n\right). \tag{B.9}$$

Also, by Theorem 1 with $M = n^{\frac{1}{2(2\beta+d)}}$, there exists $f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n, C_3)$ such that

$$\begin{aligned}\left\|f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0\right\|_\infty \leq& c_1 n^{-\frac{\beta}{(2\beta+d)}}\\ <& \frac{\varepsilon_n}{2}\end{aligned} \tag{B.10}$$

satisfies for sufficiently large n. Note that $\hat{\boldsymbol{\theta}} \in [-C_3, C_3]^{J_n}$ holds. With (B.9), (B.10) and Lemma B.1, we obtain

$$\begin{aligned}\Pi\left(B_n^*\left(f_0, \varepsilon_n\right)\right) \geq& \Pi\left(A_n^*\left(f_0, \varepsilon_n\right)\right)\\ =& \Pi\left(\left\{\boldsymbol{\theta} : \max_i \left|T_F \circ f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right| \leq \varepsilon_n\right\}\right)\\ \geq& \Pi\left(\left\{\boldsymbol{\theta} : \max_i \left|f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right| \leq \varepsilon_n\right\}\right)\\ \geq& \Pi\left(\left\{\boldsymbol{\theta} : \max_i \left|f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i) - f_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(\boldsymbol{x}_i)\right| \leq \frac{\varepsilon_n}{2}\right\}\right)\\ \geq& \Pi\left(\boldsymbol{\theta} : \boldsymbol{\theta} \in C_n^*(\hat{\boldsymbol{\theta}})\right),\end{aligned}$$

where $C_n^*(\hat{\boldsymbol{\theta}})$ is defined by

$$C_n^*(\hat{\boldsymbol{\theta}}) := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{J_n} : |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|_\infty \leq \frac{\varepsilon_n}{2a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)} \right\}.$$

By Assumption 1, there exists a constant $\delta_3 > 0$ such that

$$\Pi\left(C_n^*(\hat{\boldsymbol{\theta}})\right) \geq \delta_3^{J_n} \left(\frac{\varepsilon_n}{a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)} \right)^{J_n}.$$

Hence, we obtain

$$\begin{aligned}\Pi\left(B_n^*\left(f_0, \varepsilon_n\right)\right) \geq & \Pi\left(C_n^*(\hat{\boldsymbol{\theta}})\right) \\ \geq & \delta_3^{J_n} \left(\frac{\varepsilon_n}{a(d+1)(r_n+1)^{L_n} C_3^{L_n}(L_n+1)} \right)^{J_n} \\ \gtrsim & \exp\left(-C_2^2 C_1 (\log n) n^{\frac{d}{2\beta+d}} (\log n)^2\right) \\ \gtrsim & e^{-n\varepsilon_n^2}\end{aligned}$$

for all but finite many n. ■

Proof of Theorem 3. From Lemma B.6, Lemma B.7, $\Pi_n(\mathcal{F}_n^c) = 0$ and Theorem 4 of Ghosal and van der Vaart (2007), we have

$$\mathbb{E}_0\left[\Pi_n\left(f : h_n\left(f, f_0\right) > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right] \to 0$$

for every sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^\infty$, where the expectation is with respect to $\{Y_i\}_{i=1}^n$. Since

$$\begin{aligned}\|\phi \circ f_1 - \phi \circ f_2\|_{2,n} = & \frac{1}{n}\sum_{i=1}^n (\phi \circ f_1(\boldsymbol{x}_i) - \phi \circ f_2(\boldsymbol{x}_i))^2 \\ = & \frac{1}{n}\sum_{i=1}^n \left(\sqrt{\phi \circ f_1(\boldsymbol{x}_i)} - \sqrt{\phi \circ f_2(\boldsymbol{x}_i)}\right)^2 \left(\sqrt{\phi \circ f_1(\boldsymbol{x}_i)} + \sqrt{\phi \circ f_2(\boldsymbol{x}_i)}\right)^2 \\ \lesssim & h_n^2\left(f_1, f_2\right),\end{aligned}$$

we obtain

$$\mathbb{E}_0\left[\Pi_n\left(f : \|\phi \circ f - \phi \circ f_0\|_{2,n} > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right] \to 0$$

for every sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^\infty$, where the expectation is with respect to $\{Y_i\}_{i=1}^n$. Since

$$\mathbb{E}_0\left[\Pi_n\left(f : \|\phi \circ f - \phi \circ f_0\|_{2,n} > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right]$$

is uniformly integrable, we have

$$\mathbb{E}_0\left[\Pi_n\left(f : \|\phi \circ f - \phi \circ f_0\|_{2,n} > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right)\right] \to 0, \tag{B.11}$$

where the expectation is with respect to $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$.

Next, we will check the conditions in Lemma B.2 for

$$\begin{aligned}
\mathcal{G} :=& \left\{g \ : \ g = (\phi \circ T_F \circ f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ f_0)^2, f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n)\right\}, \\
\tau :=& \frac{1}{2}, \ \alpha := \varepsilon_n^2, \ K_1 = K_2 = 1.
\end{aligned}$$

First, it is easy to check $\|g(\boldsymbol{x})\|_\infty \le 1$ and $\mathbb{E}(g(\boldsymbol{X})^2) \le \mathbb{E}(g(\boldsymbol{X}))$ for $g \in \mathcal{G}$. Also, since

$$\begin{aligned}
&\left\|(\phi \circ T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ f_0)^2 - (\phi \circ T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ f_0)^2\right\|_{1,n} \\
&= \left\|(\phi \circ T_F \circ f_{\boldsymbol{\theta_1},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ f_0 + \phi \circ T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ f_0)(\phi \circ T_F \circ f_{\boldsymbol{\theta_2},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ T_F \circ f_{\boldsymbol{\theta_2},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}})\right\|_{1,n} \\
&\le 4 \left\|\phi \circ T_F \circ f_{\boldsymbol{\theta_1},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - \phi \circ T_F \circ f_{\boldsymbol{\theta_2},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}\right\|_{1,n} \\
&\le 4 \left\|T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}\right\|_{1,n}
\end{aligned}$$

holds for any $f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}, f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n)$, there exists $c_{36} > 0$ such that

$$\begin{aligned}
\mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right) \le& \mathcal{N}\left(\frac{u}{4}, T_F \circ \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n), \|\cdot\|_{1,n}\right) \\
\le& \mathcal{M}\left(\frac{u}{4}, T_F \circ \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n), \|\cdot\|_{1,n}\right) \\
\le& 3\left(\frac{16eF}{u}\log\frac{24eF}{u}\right)^{\mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r_n)^+} \\
\lesssim& n^{c_{36} L_n^2 r_n^2 \log(L_n r_n^2)}
\end{aligned}$$

for $u \ge n^{-1}$ by Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019). Hence for all $t \ge \frac{\varepsilon_n^2}{8}$,

$$\begin{aligned}
\int_{\frac{\tau(1-\tau)t}{16\max\{K_1, 2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right)} du \lesssim& \sqrt{t}\left(n^{\frac{d}{2\beta+d}}(\log n)^4\right)^{\frac{1}{2}} \\
=& o\left(\frac{\sqrt{n}\tau(1-\tau)t}{96\sqrt{2}\max\{K_1, 2K_2\}}\right)
\end{aligned}$$

holds. To sum up, we conclude that

$$\mathbf{P}\left\{\sup_{f \in \mathcal{F}^{\mathrm{DNN}}(L_n, r_n)} \frac{\left|\|\phi \circ f - \phi \circ f_0\|_{2,\mathrm{P}_X}^2 - \|\phi \circ f - \phi \circ f_0\|_{2,n}^2\right|}{\varepsilon_n^2 + \|\phi \circ f - \phi \circ f_0\|_{2,\mathrm{P}_X}^2} > \frac{1}{2}\right\} \le 60 \exp\left(-\frac{n\varepsilon_n^2/8}{128 \cdot 2304}\right) \tag{B.12}$$

holds for all but finite many n by Lemma B.2. Hence by (B.11) and (B.12), we obtain the assertion.

■

B.5 Proof for hierarchical compositional structure Theorem 4

The primary advantage of assuming such a hierarchical composition structure is that the dimensions of each g are significantly smaller compared to the overall dimensions. This fact allows the function to be approximated with a smaller DNN model, as demonstrated in the following lemma.

Lemma B.8. *For $\nu \in [0,1)$, $\boldsymbol{N} \in \mathbb{N}^q$ and $\mathcal{P} \subset [\beta_{min}, \beta_{max}] \times \{1, \ldots, d_{max}\}$, there exist positive constants $\tilde{C}_1$, $\tilde{C}_2$, $\tilde{C}_3$ and c_2 such that for every f_0 that follows the hierarchical composition structure $\mathcal{H}(\boldsymbol{N}, \mathcal{P})$, there exists $f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n, \tilde{C}_3)$, where L_n and r_n are given by*

$$L_n := \left\lceil \tilde{C}_1 \log_2 n \right\rceil,$$
$$r_n := \left\lceil \tilde{C}_2 \max_{(\beta',d') \in \mathcal{P}} n^{\frac{d'}{2(2\beta'+d')}} \right\rceil,$$

such that

$$\left\| f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}},\boldsymbol{\rho}_\nu} - f_0 \right\|_{\infty,[-a,a]^d} \leq c_2 \max_{(\beta',d') \in \mathcal{P}} n^{-\frac{\beta'}{2\beta'+d'}}$$

holds.

Proof. For each $l \in [q]$ and $i \in [N_l]$, we have $\beta_{min} \leq \beta_{l,i} \leq \beta_{max}$ and $1 \leq d_{l,i} \leq d_{max}$. Hence, there exist constants $c_{37} > 0$, $c_{38} > 0$, $c_{39} \geq 1$ and $c_{40} > 0$ such that for any sufficiently large $M_{l,i}$ there exists $\hat{g}^{\mathrm{DNN}}_{l,i} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(\lceil c_{37} \log_2 n \rceil, \lceil c_{38} n^{\frac{d_{l,i}}{2(2\beta_{l,i}+d_{l,i})}} \rceil, c_{39})$ with

$$\left\| \hat{g}^{\mathrm{DNN}}_{l,i} - g_{l,i} \right\|_{\infty,[-a',a']^d} \leq c_{40} n^{-\frac{\beta_{l,i}}{2\beta_{l,i}+d_{l,i}}} \tag{B.13}$$

by putting $M = n^{\frac{1}{2(2\beta_{l,i}+d_{l,i})}}$ in Theorem 1, where $a' := \max(a, 2F)$. From the bottom layer, we sequentially construct a network by

$$\hat{f}^{\mathrm{DNN}}_{1,i}(\boldsymbol{x}) = \hat{g}^{\mathrm{DNN}}_{1,i}\left(f_{0,\sum_{i'=1}^{i-1} d_{1,i'}+1}(\boldsymbol{x}), \ldots, f_{0,\sum_{i'=1}^{i-1} d_{1,i'}+d_{1,i}}(\boldsymbol{x}) \right)$$

for $i \in [N_1]$ and

$$\hat{f}^{\mathrm{DNN}}_{l,i}(\boldsymbol{x}) = \hat{g}^{\mathrm{DNN}}_{l,i}\left(\hat{f}^{\mathrm{DNN}}_{l-1,\sum_{i'=1}^{i-1} d_{l,i'}+1}(\boldsymbol{x}), \ldots, \hat{f}^{\mathrm{DNN}}_{l-1,\sum_{i'=1}^{i-1} d_{l,i'}+d_{l,i}}(\boldsymbol{x}) \right)$$

for $l \in \{2, \ldots, q\}$ and $i \in [N_l]$. Note that each

$$\left(f_{0,\sum_{i'=1}^{i-1} d_{1,i'}+1}(\boldsymbol{x}), \ldots, f_{0,\sum_{i'=1}^{i-1} d_{1,i'}+d_{1,i}}(\boldsymbol{x}) \right)$$

for $i \in [N_1]$ is a permutation of $\boldsymbol{x}$ (with length $d_{l,i}$). Hence, by defining $\tilde{C}_1 := 2q \cdot c_{37}$, $\tilde{C}_2 := \max_{l \in [q]} N_l \cdot c_{38}$ and $\tilde{C}_3 := c_{39}^2$, we obtain

$$\hat{f}^{\mathrm{DNN}}_{q,1} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}\left(\lceil \tilde{C}_1 \log_2 n \rceil, \lceil \tilde{C}_2 \max_{(\beta,d) \in \mathcal{P}} n^{\frac{d}{2(2\beta+d)}} \rceil, \tilde{C}_3 \right). \tag{B.14}$$

Now, for $l \in [q]$ and $i \in [N_l]$, we will show

$$\left\| \hat{f}^{\mathrm{DNN}}_{l,i} - f_{l,i} \right\|_{\infty,[-a,a]^d} \leq c_{40} l \left(C_{Lip} \sqrt{d_{\max}} \right)^{l-1} \max_{(\beta',d') \in \mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}} \tag{B.15}$$

holds by induction. First, for $l=1$, we obtain (B.15) directly from (B.13). Assume that (B.15) holds for some $l \in [q-1]$ and every $i \in [N_l]$. Then, for any $j \in [N_{l+1}]$, we have

$$\begin{aligned}
\left| \hat{f}^{\mathrm{DNN}}_{l+1,j}(\boldsymbol{x}) - f_{l+1,j}(\boldsymbol{x}) \right| =& \Bigg| \hat{g}^{\mathrm{DNN}}_{l+1,j} \left(\hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, \hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \\
&\quad - g_{l+1,j} \left(f_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, f_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \Bigg| \\
\leq& \Bigg| \hat{g}^{\mathrm{DNN}}_{l+1,j} \left(\hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, \hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \\
&\quad - g_{l+1,j} \left(\hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, \hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \Bigg| \\
&+ \Bigg| g_{l+1,j} \left(\hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, \hat{f}^{\mathrm{DNN}}_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \\
&\quad - g_{l+1,j} \left(f_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+1}(\boldsymbol{x}), \dots, f_{l,\sum_{i'=1}^{j-1} d_{l+1,i'}+d_{l+1,j}}(\boldsymbol{x}) \right) \Bigg| \\
\leq& c_{40} n^{-\frac{\beta_{l+1,j}}{2\beta_{l+1,j}+d_{l+1,j}}} \\
&+ C_{Lip} \sqrt{d_{l+1,j}} c_{40} l \left(C_{Lip} \sqrt{d_{\max}} \right)^{l-1} \max_{(\beta',d') \in \mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}} \\
\leq& c_{40}(l+1) \left(C_{Lip} \sqrt{d_{\max}} \right)^{l} \max_{(\beta',d') \in \mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}}
\end{aligned}$$

for any $\boldsymbol{x} \in [-a,a]^d$, where the second inequality holds by $\|\hat{f}^{\mathrm{DNN}}_{l,i}\|_{\infty,[-a',a']^d} \leq 2F \leq a'$ for $i \in [N_l]$ and the Lipschitz condition of $g_{l+1,j}$. By defining $c_2 := c_{40} q \left(C_{Lip} \sqrt{d_{\max}} \right)^{q-1}$, we obtain

$$\left\| \hat{f}^{\mathrm{DNN}}_{q,1} - f_0 \right\|_{\infty,[-a,a]^d} \leq c_2 \max_{(\beta',d') \in \mathcal{P}} n^{-\frac{\beta'}{(2\beta'+d')}}. \tag{B.16}$$

By (B.14) and (B.16), we obtain the assertion. ■

Proof of Theorem 4. We only present results for nonparametric Gaussian regression. Extending to nonparametric logistic regression can be done similarly to those in Appendix B.4. Without loss of generality, we consider γ in $(2, \frac{5}{2})$. We define $\mathcal{F}_n$ as the set of pairs of truncated DNN with the $(L_n, \boldsymbol{r}_n)$ architecture and variance of the Gaussian noise,

$$\mathcal{F}_n := \left\{ \left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}, \sigma^2 \right)^\top \ : \ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n), \ 0 < \sigma^2 \leq e^{4n\varepsilon_n^2} \right\}, \tag{B.17}$$

where L_n and r_n are defined on (8). We denote T_n as the number of parameters in the DNN model with the $(L_n, \boldsymbol{r}_n)$ architecture.

First, for given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we obtain

$$\sup_{\varepsilon > \varepsilon_n} \log \mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{(f, \sigma^2) \in \mathcal{F}_n : h_n\left((f, \sigma^2), (f_0, \sigma_0^2)\right) < \varepsilon\right\}, h_n\right) \lesssim n\varepsilon_n^2$$

under the conditions of Theorem 4, by following the proof of Lemma B.3. Also, we define $K_i((f_0, \sigma_0^2), (f, \sigma^2))$, $V_i((f_0, \sigma_0^2), (f, \sigma^2))$ and $B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)$ in the same way as in Lemma B.4, with the only change in the definition of $\mathcal{F}_n$ by (B.17). By Lemma B.8, there exists $f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r_n, C_3)$ such that

$$\begin{aligned}\left\|f^{\mathrm{DNN}}_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu} - f_0\right\|_{\infty, [-a,a]^d} \leq & c_2 \max_{(\beta', d') \in \mathcal{P}} n^{-\frac{\beta'}{2\beta' + d'}} \\ < & \frac{\sigma_0 \varepsilon_n}{4}\end{aligned}$$

satisfies for sufficiently large n. Hence, we obtain

$$(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \gtrsim e^{-n\varepsilon_n^2}$$

and

$$\frac{(\Pi \otimes \Xi)\left(\mathcal{F}_n^c\right)}{(\Pi \otimes \Xi)\left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right)} = o(e^{-2n\varepsilon_n^2})$$

under the conditions of Theorem 4, by following the proofs of Lemma B.4 and B.5. The rest of the proof can be completed along the lines of that of Theorem 2. ■

Appendix C. Proof for Theorem 5

Extending to nonparametric logistic regression is straightforward by following Appendix B.4, therefore, we only present results for nonparametric Gaussian regression. Without loss of generality, we consider γ in $(\frac{5}{2}, 3)$. We define

$$\xi_n := \left\lceil \tilde{C}_2 \max_{(\beta',d')\in\mathcal{P}} n^{\frac{d'}{2(2\beta'+d')}} \log^{\frac{1}{2}} n \right\rceil$$

and

$$\boldsymbol{\xi}_n := (d, \xi_n, \ldots, \xi_n, 1)^\top \in \mathbb{N}^{L_n+2},$$

where $\tilde{C}_2$ is a constant (depending on β_{min}, β_{max} and d_{max}) defined in Lemma B.8. We denote S_n as the number of parameters in the DNN model with the $(L_n, \boldsymbol{\xi}_n)$ architecture. In other words,

$$S_n := \sum_{l=1}^{L_n+1} (\xi_n^{(l-1)} + 1)\xi_n^{(l)}.$$

We define the sieve $\mathcal{F}_n$ as

$$\mathcal{F}_n := \Big\{ \left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}, \sigma^2\right)^\top \ : \ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \bigcup_{r=1}^{\xi_n} \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r), \ 0 < \sigma^2 \leq e^{4n\varepsilon_n^2} \Big\}, \tag{C.1}$$

where L_n is defined on (10).

Lemma C.1. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we have*

$$\sup_{\varepsilon > \varepsilon_n} \log \mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{(f, \sigma^2) \in \mathcal{F}_n : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) < \varepsilon\right\}, h_n\right) \lesssim n\varepsilon_n^2$$

under the conditions of Theorem 5, where $\mathcal{F}_n$ is defined on (C.1).

Proof. We have

$$\begin{aligned}
&\sup_{\varepsilon > \varepsilon_n} \log \mathcal{N}\left(\frac{1}{36}\varepsilon, \left\{(f, \sigma^2) \in \mathcal{F}_n : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) < \varepsilon\right\}, h_n\right) \\
&\leq \log \mathcal{N}\left(\frac{1}{36}\varepsilon_n, \mathcal{F}_n, h_n\right) \\
&\leq \log\left(\sum_{r=1}^{\xi_n} \mathcal{N}\left(\frac{1}{36}\varepsilon_n, \Big\{ \left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}, \sigma^2\right)^\top : f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r), \ 0 < \sigma^2 \leq e^{4n\varepsilon_n^2}\Big\}, h_n\right)\right) \\
&\leq \log\left(\xi_n \mathcal{N}\left(\frac{1}{36}\varepsilon_n, \Big\{ \left(T_F \circ f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}, \sigma^2\right)^\top : f^{\mathrm{DNN}}_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu} \in \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, \xi_n), \ 0 < \sigma^2 \leq e^{4n\varepsilon_n^2}\Big\}, h_n\right)\right) \\
&\lesssim \log \xi_n + n\varepsilon_n^2 + L_n^2 \xi_n^2 \log L_n \xi_n^2 \log n \\
&\lesssim n\varepsilon_n^2,
\end{aligned}$$

where the fourth inequality holds by (B.1). ■

Lemma C.2. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we define $K_i((f_0, \sigma_0^2), (f, \sigma^2))$, $V_i((f_0, \sigma_0^2), (f, \sigma^2))$ and $B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)$ in the same way as in Lemma B.4, with the only change in the definition of $\mathcal{F}_n$ by (C.1). Then, we have*

$$\sum_{r=1}^{\infty} (\Pi_r \otimes \Xi) \left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \Gamma(r) \gtrsim e^{-n\varepsilon_n^2}$$

under the conditions of Theorem 5.

Proof. We define

$$\xi_n' := \left\lceil \tilde{C}_2 \max_{(\beta', d') \in \mathcal{P}} n^{\frac{d'}{2(2\beta' + d')}} \right\rceil$$

and

$$\begin{aligned} \boldsymbol{\xi}_n' &:= (d, \xi_n', \ldots, \xi_n', 1)^\top \in \mathbb{N}^{L_n + 2}, \\ S_n' &:= \sum_{l=1}^{L_n + 1} ({\xi_n'}^{(l-1)} + 1){\xi_n'}^{(l)}. \end{aligned}$$

Then, by Lemma B.8, there exists $f_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, \xi_n', \tilde{C}_3)$ such that

$$\begin{aligned} \left\| f_{\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0 \right\|_{\infty, [-a, a]^d} \leq & c_2 \max_{(\beta', d') \in \mathcal{P}} n^{-\frac{\beta'}{2\beta' + d'}} \\ < & \frac{\sigma_0 \varepsilon_n}{4} \end{aligned}$$

holds for sufficiently large n. Hence, we obtain

$$\begin{aligned} & \sum_{r=1}^{\infty} (\Pi_r \otimes \Xi) \left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \Gamma(r) \\ & \geq \Gamma(\xi_n') \cdot (\Pi_{\xi_n'} \otimes \Xi) \left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \\ & \gtrsim \frac{1}{(\log n)^5} e^{-(\log n)^5 (\xi_n')^2} \delta_1^{S_n'} \left(\frac{\sigma_0 \varepsilon_n}{2a(d+1)(\xi_n' + 1)^{L_n} C_3^{L_n} (L_n + 1)} \right)^{S_n'} \delta_2 \varepsilon_n^2 \\ & \gtrsim \exp\left(-(\log n)^5 C_2^2 \max_{(\beta', d') \in \mathcal{P}} n^{\frac{d'}{2\beta' + d'}} \right) \exp\left(-C_2^2 C_1 (\log n) \max_{(\beta', d') \in \mathcal{P}} n^{\frac{d'}{2\beta' + d'}} (\log n)^2 \right) n^{-1} \\ & \gtrsim e^{-n\varepsilon_n^2} \end{aligned} \tag{C.2}$$

for all but finite many n, where the third inequality holds by the proof of (B.4). ■

Lemma C.3. *For given $\boldsymbol{x}^{(n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, we have*

$$\frac{\sum_{r=1}^{\infty} (\Pi_r \otimes \Xi) \left(\mathcal{F}_n^c\right) \Gamma(r)}{\sum_{r=1}^{\infty} (\Pi_r \otimes \Xi) \left(B_n^*\left((f_0, \sigma_0^2), \varepsilon_n\right)\right) \Gamma(r)} = o(e^{-2n\varepsilon_n^2})$$

under the conditions of Theorem 5, where $\mathcal{F}_n$ is defined on (C.1).

Proof. Since

$$\left(\frac{1}{2kr} - \frac{1}{4k^2r^3}\right) e^{-kr^2} \leq \int_r^{\infty} e^{-kt^2} dt \leq \frac{1}{2kr} e^{-kr^2}$$

for any $k > 0$ and $s > 0$,

$$\begin{aligned}\Gamma(r > \xi_n) \leq & \frac{\sum_{r=\xi_n+1}^{\infty} e^{-(\log n)^5 r^2}}{\sum_{r=1}^{\infty} e^{-(\log n)^5 r^2}} \\ \lesssim & \frac{e^{-(\log n)^5 \xi_n^2}}{\xi_n e^{-(\log n)^5}} \\ \lesssim & e^{-(\log n)^5 {\xi_n}^2} e^{(\log n)^5}\end{aligned}$$

holds. From (C.2) and

$$\begin{aligned}\sum_{r=1}^{\infty} (\Pi_r \otimes \Xi) \left(\mathcal{F}_n^c\right) \Gamma(r) \leq & \Gamma\left(r > \xi_n\right) + \Xi\left(\sigma^2 > e^{4n\varepsilon_n^2}\right) \\ \lesssim & e^{-(\lambda \log n)^5 {\xi_n}^2} e^{(\lambda \log n)^5} + e^{-4n\varepsilon_n^2} \\ \lesssim & e^{-4n\varepsilon_n^2},\end{aligned}$$

we obtain the assertion. ■

Proof of Theorem 5. We only present results for nonparametric Gaussian regression. From Theorem 4 of Ghosal and van der Vaart (2007), Lemma C.1 and Lemma C.2, we have

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) \in \mathcal{F}_n : h_n\left((f,\sigma^2),(f_0,\sigma_0^2)\right) > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right) \middle| \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\right] \to 0$$

for every sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^{\infty}$, where the expectation is with respect to $\{Y_i\}_{i=1}^{n}$. Similar with the proof of (B.7), we have

$$\mathbb{E}_0\left[\Pi_n\left((f,\sigma^2) \in \mathcal{F}_n : \|f - f_0\|_{2,n} + |\sigma^2 - \sigma_0^2| > M_n\varepsilon_n \middle| \mathcal{D}^{(n)}\right)\right] \to 0, \tag{C.3}$$

where the expectation is with respect to $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^{n}$.

Next, we will check the conditions in Lemma B.2 for

$$\begin{aligned}\mathcal{G} :=& \left\{g \;:\; g = (T_F \circ f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2, f_{\boldsymbol{\theta},\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} \in \bigcup_{r=1}^{\xi_n} \mathcal{F}_{\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}(L_n, r)\right\}, \\ \tau :=& \frac{1}{2},\ \alpha := \varepsilon_n^2,\ K_1 = K_2 = 4F^2.\end{aligned}$$

First, it is easy to check $\|g(\boldsymbol{x})\|_\infty \leq 4F^2$ and $\mathbb{E}(g(\boldsymbol{X})^2) \leq 4F^2\mathbb{E}(g(\boldsymbol{X}))$ for $g \in \mathcal{G}$. Also, since

$$\begin{aligned}&\left\|(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2 - (T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)^2\right\|_{1,n} \\ &= \left\|(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0 + T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - f_0)(T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}})\right\|_{1,n} \\ &\leq 4F \left\|T_F \circ f_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}} - T_F \circ f_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu}^{\mathrm{DNN}}\right\|_{1,n}\end{aligned}$$

holds for any $f^{\mathrm{DNN}}_{\boldsymbol{\theta}_1,\boldsymbol{\rho}_\nu}, f^{\mathrm{DNN}}_{\boldsymbol{\theta}_2,\boldsymbol{\rho}_\nu} \in \bigcup_{r=1}^{\xi_n} \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r)$, there exists $c_{34} > 0$ such that

$$\begin{aligned}
\mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right) \leq & \mathcal{N}\left(\frac{u}{4F}, \bigcup_{r=1}^{\xi_n} T_F \circ \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, r), \|\cdot\|_{1,n}\right) \\
\leq & \xi_n \mathcal{N}\left(\frac{u}{4F}, T_F \circ \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, \xi_n), \|\cdot\|_{1,n}\right) \\
\leq & \xi_n \mathcal{M}\left(\frac{u}{4F}, T_F \circ \mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n, \xi_n), \|\cdot\|_{1,n}\right) \\
\leq & 3\xi_n \left(\frac{16eF^2}{u} \log \frac{24eF^2}{u}\right)^{\mathcal{F}^{\mathrm{DNN}}_{\boldsymbol{\rho}_\nu}(L_n,\xi_n)^+} \\
\lesssim & \xi_n n^{c_{34} L_n^2 \xi_n^2 \log(L_n \xi_n^2)}
\end{aligned}$$

for $u \geq n^{-1}$ by Theorem 9.4 of Györfi et al. (2002) and Theorem 7 of Bartlett et al. (2019). Hence for all $t \geq \frac{\varepsilon_n^2}{8}$,

$$\begin{aligned}
\int_{\frac{\tau(1-\tau)t}{16\max\{K_1,2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \mathcal{G}, \|\cdot\|_{1,n}\right)} du \lesssim & \sqrt{t}\left(n^{\frac{d}{2\beta+d}} (\log n)^6\right)^{\frac{1}{2}} \\
= & o\left(\frac{\sqrt{n}\tau(1-\tau)t}{96\sqrt{2}\max\{K_1, 2K_2\}}\right)
\end{aligned}$$

holds. Hence, we have

$$\mathbf{P}\left\{\sup_{f \in \mathcal{F}^{\mathrm{DNN}}(L_n, r_n)} \frac{\left|\|f - f_0\|_{2,\mathrm{P}_X}^2 - \|f - f_0\|_{2,n}^2\right|}{\varepsilon_n^2 + \|f - f_0\|_{2,\mathrm{P}_X}^2} > \frac{1}{2}\right\} \leq 60 \exp\left(-\frac{n\varepsilon_n^2/8}{128 \cdot 2304 \cdot 16F^4}\right) \tag{C.4}$$

holds for all but finite many n by Lemma B.2. By (C.3) and (C.4), we obtain

$$\mathbb{E}_0\left[\Pi_n\left((f, \sigma^2) \in \mathcal{F}_n : \|f - f_0\|_{2,\mathrm{P}_X} + |\sigma^2 - \sigma_0^2| > M_n \varepsilon_n \middle| \mathcal{D}^{(n)}\right)\right] \to 0,$$

where the expectation is with respect to $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$.

Finally, Lemma C.3 and Lemma 1 of Ghosal and van der Vaart (2007) imply that

$$\mathbb{E}_0\left[\Pi_n\left((f, \sigma^2)^\top \in \mathcal{F}_n^c \middle| \mathcal{D}^{(n)}\right)\right] \to 0.$$

Hence, we obtain the assertion. ■

References

J. Bai, Q. Song, and G. Cheng. Efficient variational inference for sparse deep learning with theoretical guarantee. *Advances in Neural Information Processing Systems*, 33:466–476, 2020.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261, 2019.

W. Beker, A. Wołos, S. Szymkuć, and B. A. Grzybowski. Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks. *Nature Machine Intelligence*, 2(8):457–465, 2020.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

M. Chen, H. Jiang, W. Liao, and T. Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

B.-E. Chérief-Abdellatif. Convergence rates of variational inference in sparse deep learning. In *International Conference on Machine Learning*, pages 1831–1842. PMLR, 2020.

M. Cranmer, D. Tamayo, H. Rein, P. Battaglia, S. Hadden, P. J. Armitage, S. Ho, and D. N. Spergel. A Bayesian neural network predicts the dissolution of compact planetary systems. *Proceedings of the National Academy of Sciences*, 118(40), 2021.

S. C. Douglas and J. Yu. Why ReLU units sometimes die: analysis of single-unit error backpropagation in neural networks. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 864–868. IEEE, 2018.

M. Dusenberry, G. Jerfel, Y. Wen, Y. Ma, J. Snoek, K. Heller, B. Lakshminarayanan, and D. Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020.

R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

S. Farquhar, M. A. Osborne, and Y. Gal. Radial Bayesian neural networks: Beyond discrete support in large-scale bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1352–1362. PMLR, 2020.

V. Fortuin. Priors in Bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

V. Fortuin, A. Garriga-Alonso, S. W. Ober, F. Wenzel, G. Ratsch, R. E. Turner, M. van der Wilk, and L. Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

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

S. Ghosal and A. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

S. Ghosh, J. Yao, and F. Doshi-Velez. Model selection in Bayesian neural networks via horseshoe priors. *J. Mach. Learn. Res.*, 20(182):1–46, 2019.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

A. Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

J. Heek and N. Kalchbrenner. Bayesian inference for large scale image classification. *arXiv preprint arXiv:1908.03491*, 2019.

J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

M. Imaizumi and K. Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd international conference on artificial intelligence and statistics*, pages 869–878. PMLR, 2019.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

S. Jantre, S. Bhattacharya, and T. Maiti. Layer adaptive node selection in Bayesian neural networks: Statistical guarantees and implementation details. *Neural Networks*, 167:309–330, 2023.

Y. Jiao, G. Shen, Y. Lin, and J. Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691–716, 2023.

L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

M. Kohler and S. Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021a.

M. Kohler and S. Langer. Supplement A to "On the rate of convergence of fully connected deep neural network regression estimates.", 2021b.

M. Kohler and S. Langer. Supplement B to "On the rate of convergence of fully connected deep neural network regression estimates.", 2021c.

M. Kohler, A. Krzyżak, and S. Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE transactions on information theory*, 68(6):4032–4042, 2022.

I. Kong, D. Yang, J. Lee, I. Ohn, G. Baek, and Y. Kim. Masked Bayesian neural networks : Theoretical guarantee and its posterior inference. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 17462–17491, 2023.

K. Lee and J. Lee. Asymptotic properties for Bayesian neural network in Besov space. *Advances in Neural Information Processing Systems*, 35:5641–5653, 2022.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6 (6):861–867, 1993.

C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

J. Liu. Variable selection with rigorous uncertainty quantification using deep bayesian neural networks: Posterior concentration and bernstein-von mises phenomenon. In *International Conference on Artificial Intelligence and Statistics*, pages 3124–3132. PMLR, 2021.

C. Louizos and M. Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR, 2017.

C. Louizos, K. Ullrich, and M. Welling. Bayesian compression for deep learning. *Advances in neural information processing systems*, 30, 2017.

J. Lu, Z. Shen, H. Yang, and S. Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis. Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020.

A. L. Maas, A. Y. Hannun, A. Y. Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

D. J. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 27:2924–2932, 2014.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning*, pages 807–814, 2010.

R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174): 1–38, 2020.

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

C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

S. W. Ober and L. Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep gaussian processes. In *International Conference on Machine Learning*, pages 8248–8259. PMLR, 2021.

C. Oh, K. Adamczewski, and M. Park. Radial and directional posteriors for Bayesian neural networks. *AAAI Conference on Artificial Intelligence*, 2020.

I. Ohn and Y. Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627, 2019.

I. Ohn and Y. Kim. Nonconvex sparse regularization for deep neural networks and its optimality. *Neural computation*, 34(2):476–517, 2022.

I. Ohn and L. Lin. Adaptive variational bayes: Optimality, computation and applications. *The Annals of Statistics*, 52(1):335–363, 2024.

P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

N. G. Polson and V. Ročková. Posterior concentration for sparse deep learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 938–949, 2018.

J. Rousseau and B. Szabo. Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *The Annals of Statistics*, 48(4):2155–2179, 2020.

J. Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

S. Seto, M. T. Wells, and W. Zhang. Halo: Learning to prune neural networks with shrinkage. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 558–566. SIAM, 2021.

M. Špendl and K. Pirc. Easy Bayesian transfer learning with informative priors. In *Neural Information Processing Systems*, 2022.

Y. Sun, Q. Song, and F. Liang. Consistent sparse deep learning: Theory and computation. *Journal of the American Statistical Association*, 117(540):1981–1995, 2022.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.

J. Swiatkowski, K. Roth, B. Veeling, L. Tran, J. Dillon, J. Snoek, S. Mandt, T. Salimans, R. Jenatton, and S. Nowozin. The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in Bayesian neural networks. In *International Conference on Machine Learning*, pages 9289–9299. PMLR, 2020.

B. Szabó, A. van der Vaart, and J. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015.

T. Tran, T.-T. Do, I. Reid, and G. Carneiro. Bayesian generative active deep learning. In *International Conference on Machine Learning*, pages 6295–6304. PMLR, 2019.

A. B. Tsybakov. Introduction to nonparametric estimation, 2009.

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.

H. Wang and D.-Y. Yeung. A survey on Bayesian deep learning. *ACM computing surveys (csur)*, 53(5):1–37, 2020.

H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1235–1244, 2015.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning*, pages 681–688. Citeseer, 2011.

P. M. Williams. Bayesian regularization and pruning using a laplace prior. *Neural computation*, 7(1):117–143, 1995.

A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernandez-Lobato, and A. L. Gaunt. Deterministic Variational Inference for Robust Bayesian Neural Networks. In *International Conference on Learning Representations*, 2019.

F. Xie and Y. Xu. Adaptive Bayesian nonparametric regression using a kernel mixture of polynomials with application to partial linear models. *Bayesian Analysis*, 15(1):159–186, 2020.

B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

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

R. Zhang, C. Li, J. Zhang, C. Chen, and A. G. Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2020.