Posterior Contraction for Sparse Neural Networks in Besov Spaces with Intrinsic Dimensionality

Kyeongwon Lee[1] **Lizhen Lin**[1] **Jaewoo Park**[2] **Seonghyun Jeong**[2]*
[1]Department of Mathematics, University of Maryland
[2]Department of Statistics and Data Science, Yonsei University
{kwlee,lizhen01}@umd.edu {jwpark88,sjeong}@yonsei.ac.kr

Abstract

This work establishes that sparse Bayesian neural networks achieve optimal posterior contraction rates over anisotropic Besov spaces and their hierarchical compositions. These structures reflect the intrinsic dimensionality of the underlying function, thereby mitigating the curse of dimensionality. Our analysis shows that Bayesian neural networks equipped with either sparse or continuous shrinkage priors attain the optimal rates which are dependent on the intrinsic dimension of the true structures. Moreover, we show that these priors enable *rate adaptation*, allowing the posterior to contract at the optimal rate even when the smoothness level of the true function is unknown. The proposed framework accommodates a broad class of functions, including additive and multiplicative Besov functions as special cases. These results advance the theoretical foundations of Bayesian neural networks and provide rigorous justification for their practical effectiveness in high-dimensional, structured estimation problems.

1 Introduction

Neural networks (NNs) have been widely used to extract features from complex datasets, such as visual recognition and language modeling [1]. Due to their approximation ability [2–7], NNs exhibit remarkable flexibility in representing complex multivariate functions. Given their network structure, NNs are trained by minimizing empirical risk defined through a suitable loss function, which can be viewed as maximum likelihood estimation from a statistical perspective [8]. However, such a frequentist approach may lead to miscalibration and overconfidence [9], particularly when the model is trained on out-of-distribution samples [10]. On the other hand, the Bayesian approach quantifies predictive uncertainty via the posterior distribution [11, 12] and can improve calibration and robustness in many cases [13, 14]. Another key advantage of the Bayesian perspective is that, unlike their frequentist counterparts, Bayesian neural networks (BNNs) can easily achieve the optimal rate of convergence without knowledge of the smoothness of the function. This property, known as *rate adaptation*, is inherently attained by the nature of Bayesian inference.

To capture the flexibility of NNs, it is desirable to consider function classes that are richer than the intuitively smooth ones, such as Hölder or Sobolev spaces. In particular, Besov spaces, which encompass non-smooth or even discontinuous functions, are well suited for this purpose. For example, images often exhibit local inhomogeneities and sharp edges naturally represented within Besov spaces [15]. Moreover, additional flexibility arises from accounting for intrinsic dimensionality, which helps explain the robustness of NNs to the curse of dimensionality. For instance, images frequently possess low intrinsic dimensionality [16]. Such an intrinsic dimensional structure can be implicitly modeled through anisotropy or composite function constructions. Although frequentist

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

approaches have extensively investigated theoretical properties in these complex settings [17, 18], Bayesian analyses remain largely limited to simpler cases [19–22]. This study aims to fill this gap by establishing that BNNs achieve optimal posterior contraction and exhibit rate adaptation under the fully Bayesian paradigm in such complex scenarios.

1.1 Related works and our contribution

Rates of convergence provide a fundamental means of assessing the quality of estimation procedures in both frequentist and Bayesian inference. While frequentist analyses focus on the convergence rates of estimators, Bayesian approaches examine the contraction rates of posterior distributions, which describe how quickly the posterior concentrates around the truth. In the deep neural network (DNN) literature, such notions of optimality have been investigated under a variety of settings from both frequentist and Bayesian perspectives.

Frequentist works. Under the frequentist paradigm, Schmidt-Hieber [8] showed that a carefully constructed DNN estimator achieves the near-minimax optimal rate over Hölder classes, provided that the network architecture has sufficient depth and sparsity adapted to the underlying function complexity. Their result extends to composite structures, encompassing a wide range of structured function classes, including additive models. Suzuki [17] further extended these results to Besov spaces, and Suzuki and Nitanda [18] demonstrated that DNNs mitigate the curse of dimensionality by adapting to anisotropic smoothness and composite structures in Besov spaces. In contrast to the aforementioned works that utilize sparse networks, Kohler and Langer [23] investigated dense NNs for composite Hölder spaces. Collectively, these frequentist results demonstrate that DNNs can achieve optimal convergence rates under appropriately specified conditions. However, attaining such optimality requires careful tuning of the network architecture to the underlying function complexity, and rate-adaptive results are currently not available.

Bayesian works. The Bayesian literature remains comparatively limited. A pioneering contribution by Polson and Ročková [21] established that BNNs equipped with spike-and-slab priors achieve optimal posterior contraction over Hölder classes. This work was extended to Besov spaces by Lee and Lee [20], who considered both spike-and-slab and continuous shrinkage priors, although the application of shrinkage priors remains largely limited. Kong and Kim [19] obtained similar results using dense networks with non-sparse priors, further extending the theory to composite Hölder spaces. To the best of our knowledge, the only Bayesian study that considers anisotropic Besov spaces is Egels and Castillo [22]. However, that work primarily employs the fractional posterior approach [24, 25], and the standard posterior under the fully Bayesian framework has been investigated only under strong restrictions. Moreover, they focused on Hölder classes in the composite setting, and, to our knowledge, no Bayesian study has yet considered composite structures within anisotropic Besov spaces. Despite this limited scope, these Bayesian approaches achieve rate adaptation, attaining optimal posterior contraction without prior knowledge of the smoothness level.

Our contribution. We establish that sparse BNNs achieve the near-minimax optimal rates over anisotropic and composite Besov spaces. This result demonstrates that BNNs adapt to intrinsic dimensional structures, thereby avoiding the curse of dimensionality. We show that sparse BNNs can accommodate a broader class of realistic functions that have not been fully addressed in earlier works. Compared to Egels and Castillo [22], we adopt the pure Bayesian framework, relying on the standard posterior distribution for inference. This choice aligns more closely with conventional Bayesian practice and facilitates broader acceptance within the Bayesian community. Furthermore, we show that sparse BNNs adapt to the underlying model complexity and attain the optimal rate for the target function without requiring oracle knowledge. As a result, users are not required to specify the exact network architecture; rate adaptation is achieved through an appropriately designed prior distribution. A summary of related works and our contributions is provided in Table 1.

2 Preliminaries

2.1 Setup

Notation. For $a \in \mathbb{R}$, let $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the floor and ceiling functions, respectively. For $n \in \mathbb{N}$, the notation $[n]$ stands for the set $\{1, 2, \cdots, n\}$. For $a, b \in \mathbb{R}$, we write $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For a real vector v, $\|v\|_p$ denotes the ℓ_p-norm for $p \in [1, \infty]$,

Table 1: Summary of related works and this study. The abbreviations Iso., Aniso., and Ada. denote *isotropic*, *anisotropic*, and *adaptation*, respectively.

			Single Function		Composite Function	
Study	Approach	Architecture	Function Space	Ada.	Function Space	Ada.
Schmidt-Hieber [8]	Frequentist	Sparse	Iso. Hölder	✗	Iso. Hölder	✗
Kohler and Langer [23]	Frequentist	Dense	Iso. Hölder	✗	Iso. Hölder	✗
Suzuki [17]	Frequentist	Sparse	Iso. Besov	✗	-	-
Suzuki and Nitanda [18]	Frequentist	Sparse	Aniso. Besov	✗	Aniso. Besov	✗
Polson and Ročková [21]	Bayesian	Sparse	Iso. Hölder	✓	-	-
Kong and Kim [19]	Bayesian	Dense	Iso. Hölder	✓	Iso. Hölder	✓
Lee and Lee [20]	Bayesian	Sparse	Iso. Besov	✓	-	-
Egels and Castillo [22]	Bayesian*	Dense	Aniso. Besov**	✓	Iso. Hölder	✓
This work	Bayesian	Sparse	Aniso. Besov	✓	Aniso. Besov	✓

∗ Egels and Castillo [22] primarily employ the fractional posterior to circumvent issues of model complexity.
∗∗ Egels and Castillo [22] impose stronger restrictions on the smoothness parameter than other studies on anisotropic Besov spaces (see Section 3.1 for details).

and $\|v\|_0$ denotes the number of nonzero components. For a measurable function $f : [0,1]^d \to \mathbb{R}$ and a measure μ, define $\|f\|_{L^p(\mu)} = (\int_{[0,1]^d} |f|^p d\mu)^{1/p}$ for $p \in (0,\infty)$. When μ is the Lebesgue measure, we write $\|f\|_{L^p}$ for brevity. For $p = \infty$, we define $\|f\|_{L^\infty} = \operatorname{ess\,sup}_{x\in[0,1]^d} |f(x)|$. We also define the supremum norm by $\|f\|_\infty = \sup_{x\in[0,1]^d} |f(x)|$. The Dirac delta at zero is denoted by δ_0, and the indicator function of a set A is denoted by $I(A)$. For sequences a_n and b_n, we write $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ to mean that $a_n \le C b_n$ for some universal constant $C > 0$. If $a_n \lesssim b_n \lesssim a_n$, we write $a_n \asymp b_n$. Let $\mathcal{UB} = \{f : [0,1]^d \to \mathbb{R}; \|f\|_\infty \le 1\}$ denote the family of uniformly bounded functions. For a normed space $\mathcal{F}$, $U(\mathcal{F})$ denotes the unit ball of $\mathcal{F}$.

Model. We consider a nonparametric regression problem with a d-dimensional input variable $X_i \in [0,1]^d$ and an output variable $Y_i \in \mathbb{R}$, for $i \in [n]$. The observations $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ are independent and identically distributed according to the model,

$$Y_i = f_0(X_i) + \xi_i, \quad X_i \sim P_X, \quad \xi_i \sim \mathrm{N}(0, \sigma_0^2), \quad i \in [n], \tag{1}$$

where $f_0 : [0,1]^d \to \mathbb{R}$ is the true regression function, $\sigma_0^2 > 0$ is the noise variance, and P_X is the distribution of X_i. We denote the joint distribution of $\mathcal{D}_n$ under this model by $P^{(n)}_{f_0,\sigma_0}$.

Neural network. We denote the ReLU activation function by $\zeta(\cdot)$ and use the same notation for its vectorized version. The parameter space of L-layered NNs with B-bounded and S-sparse weights is defined as

$$\Theta(L,D,S,B) = \Big\{\theta = \big(W^{(1)}, b^{(1)}, \dots, W^{(L+1)}, b^{(L+1)}\big) : W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}},\ b^{(l)} \in \mathbb{R}^{d_l},$$
$$d_l = D,\ l \in [L],\ d_0 = d,\ d_{L+1} = 1,\ \|\theta\|_0 \le S,\ \|\theta\|_\infty \le B\Big\}.$$

Let $f_\theta(\cdot) = (W^{(L+1)}(\cdot) + b^{(L+1)}) \circ \zeta \circ \cdots \circ \zeta \circ (W^{(1)}(\cdot) + b^{(1)}) : [0,1]^d \to \mathbb{R}$ be the input-output mapping of an L-layered NN. We define the corresponding feedforward NN class as $\Phi(L,D,S,B) = \{\mathrm{clip} \circ f_\theta : \theta \in \Theta(L,D,S,B)\}$, where $\mathrm{clip}(x) = \min\{1, \max\{-1, x\}\}$ clips its input to the interval $[-1,1]$. We also define the unbounded sparse and unbounded dense parameter spaces as $\Theta(L,D,S) = \lim_{B\to\infty} \Theta(L,D,S,B)$ and $\Theta(L,D) = \lim_{S\to\infty} \Theta(L,D,S)$, respectively. By further taking union over the width, we define $\Theta(L) = \bigcup_{D=1}^\infty \Theta(L,D)$. The corresponding NN classes are defined analogously as $\Phi(L,D,S) = \lim_{B\to\infty} \Phi(L,D,S,B)$, $\Phi(L,D) = \lim_{S\to\infty} \Phi(L,D,B)$, and $\Phi(L) = \bigcup_{D=1}^\infty \Phi(L,D)$.

2.2 Anisotropic Besov spaces

We first consider anisotropic Besov spaces, which accommodate varying smoothness across different directional components [26, 18]. Let $s = (s_1, \dots, s_d) \in \mathbb{R}^d_{++}$, $0 < p, q \le \infty$, and $r = \lfloor \max_j s_j \rfloor + 1$. For $h \in \mathbb{R}^d$, the r-th order difference is defined as $\Delta_h^r(f)(x) = \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} f(x + jh)$ if $x, x + rh \in [0,1]^d$, and 0 otherwise. Given $t = (t_1, \dots, t_d) \in \mathbb{R}^d_{++}$, the anisotropic modulus of smoothness is defined as $w_{r,p}(f,t) = \sup_{h\in\mathbb{R}^d : |h_j| \le t_j} \|\Delta_h^r(f)\|_{L^p}$. The anisotropic Besov

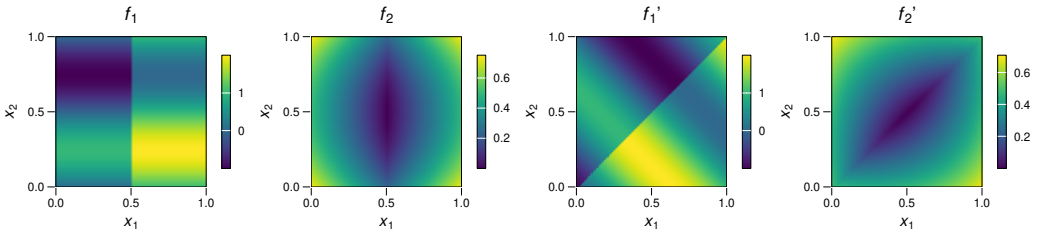


Figure 1: We illustrate two example functions, $f_1(x) = I(\{x_1 \in [1/2, 1]\}) + \sin(2\pi x_2)$ and $f_2(x) = |x_1 - 1/2| + (x_2 - 1/2)^2$, and their rotated counterparts f_1' and f_2'.

seminorm is then defined as $\|f\|_{\mathcal{B}^s_{p,q}} := \|f\|_{L^p} + \|f\|^*_{\mathcal{B}^s_{p,q}}$, where

$$\|f\|^*_{\mathcal{B}^s_{p,q}} = \begin{cases} \left(\sum_{k=0}^{\infty} \left[2^k w_{r,p}(f, (2^{-k/s_1}, \ldots, 2^{-k/s_d}))\right]^q\right)^{1/q} & \text{if } q < \infty, \\ \sup_{k \geq 0} \left[2^k w_{r,p}(f, (2^{-k/s_1}, \ldots, 2^{-k/s_d}))\right] & \text{if } q = \infty. \end{cases}$$

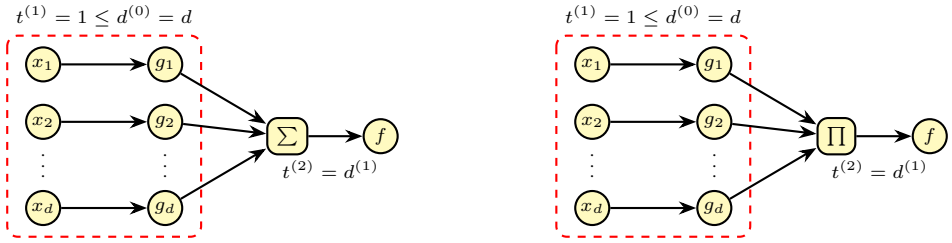
The anisotropic Besov space $\mathcal{B}^s_{p,q}$ is the collection of all functions $f \in L^p$ such that $\|f\|_{\mathcal{B}^s_{p,q}}$ is finite. For an anisotropic smoothness parameter s, we define the smallest smoothness as $\underline{s} = \min_j s_j$, the largest smoothness as $\overline{s} = \max_j s_j$, and the intrinsic smoothness (exponent of global smoothness) as $\tilde{s} = (\sum_{j=1}^d s_j^{-1})^{-1}$. In the special case where $s = (s_0, \ldots, s_0)$ for some $s_0 > 0$, the space $\mathcal{B}^s_{p,q}$ reduces to an isotropic Besov space. Besov spaces generalize classical notions of differentiability and continuity and are more flexible than Hölder spaces, which can be viewed as particular subspaces of Besov spaces. For further discussion of the continuous embedding properties, see Remark B.2.

It is well known that the minimax optimal rate of convergence over isotropic Besov classes with smoothness s_0 is $n^{-s_0/(2s_0+d)}$ [27, 28]. For anisotropic Besov spaces, the minimax rate is given by $n^{-\tilde{s}/(2\tilde{s}+1)}$ [29, 30], which can also be expressed as $n^{-\underline{s}/(2\underline{s}+d^*)}$, where $d^* := \underline{s}/\tilde{s}$. This form resembles the minimax rate over isotropic Besov spaces with smoothness $\underline{s}$ and dimension d^*, in which the anisotropic space is continuously embedded. In this sense, d^* can be interpreted as an intrinsic dimension associated with anisotropic Besov spaces [18]. For example, the two functions f_1 and f_2 in Figure 1 belong to $\mathcal{B}^{(s_1,s_2)}_{1,\infty}$ for any $0 < s_1 < 1$ and $s_2 > s_1$. The anisotropic Besov space $\mathcal{B}^{(s_1,s_2)}_{1,\infty}$ is continuously embedded in the isotropic Besov space with s_1. By accounting for anisotropy, the rate exponent can be improved from $s_1/(2s_1 + 2)$ to $\tilde{s}/(2\tilde{s} + 1)$, where $\tilde{s} = s_1 s_2/(s_1 + s_2)$. The intrinsic dimension satisfies $d^* = s_1/\tilde{s} = s_1/s_2 + 1 < 2$, indicating a reduced effective dimension compared to the ambient dimension $d = 2$.

2.3 Composite Besov spaces

Beyond anisotropic spaces, we consider composite function spaces in which functions are represented as hierarchical compositions of simpler components. For a composite function $f = f_H \circ \cdots \circ f_1 : [0,1]^d \to \mathbb{R}$ with composition depth H, each intermediate function f_h maps between successive spaces, potentially involving dimension reduction through sparse connectivity. This compositional structure naturally aligns with the layered architecture of NNs, where each layer progressively transforms its inputs into higher-level representations. Figure 1 presents examples in which rotated functions can be interpreted as compositions of Besov functions with affine transformations. Compositional function spaces offer a useful framework for understanding the ability of NNs to approximate high-dimensional functions without suffering from the curse of dimensionality.

Let $d_\circ = (d^{(0)}, d^{(1)}, \ldots, d^{(H)})$ be a sequence of dimensions with $d^{(0)} = d$ and $d^{(H)} = 1$. Let $t_\circ = (t^{(1)}, \ldots, t^{(H)})$ denote the effective dimensions, where each $t^{(h)}$ satisfies $1 \leq t^{(h)} \leq d^{(h-1)}$. Let $s_\circ = (s^{(1)}, \ldots, s^{(H)})$ be a sequence of anisotropic smoothness vectors, where each $s^{(h)} \in \mathbb{R}^{t^{(h)}}_{++}$. We define the composite anisotropic Besov space $\mathcal{B}^{d_\circ, t_\circ, s_\circ}_{p,q}$ as the collection of composite functions $f = f_H \circ \cdots \circ f_1 : [0,1]^d \to \mathbb{R}$ such that the intermediate functions $f_h = (f_{h,1}, \ldots, f_{h,d^{(h)}}) : [0,1]^{d^{(h-1)}} \to [0,1]^{d^{(h)}}$ for $h = 1, \ldots, H-1$, and $f_H : [0,1]^{d^{(H-1)}} \to \mathbb{R}$ satisfy the following condition: for each h and each $j \in [d^{(h)}]$, there exists a subset $I_{h,j} \subset [d^{(h-1)}]$ with $|I_{h,j}| = t^{(h)}$



(a) Additive structure

(b) Multiplicative structure

Figure 2: Illustration of additive ($f(x) = \sum_{i=1}^{d} g_i(x_i)$) and multiplicative ($f(x) = \prod_{i=1}^{d} g_i(x_i)$) composite Besov functions. Each component function g_i depends on a single input dimension ($t^{(1)} = 1$), although the ambient dimension $d^{(0)} = d$ may be much larger.

and a function $\tilde{f}_{h,j} \in U(\mathcal{B}_{p,q}^{s^{(h)}})$ such that $f_{h,j}(x) = \tilde{f}_{h,j}(x_{I_{h,j}})$, where x_I denotes the subvector of x restricted to the coordinates in I. Additive and multiplicative Besov spaces serve as simple examples of composite Besov spaces, as illustrated in Figure 2.

Note that Suzuki and Nitanda [18] did not impose effective low-dimensionality in their definition of composite Besov spaces. In contrast, our definition aligns more with the composite Hölder spaces introduced by Schmidt-Hieber [8], and it is well suited for analyzing functions approximated by NNs, where dimension reduction may occur within layers. By incorporating anisotropy within each f_h, the framework captures low-dimensional structures even when the ambient or latent dimensions are high. This interplay between compositional structure and anisotropy enables efficient function representation.

3 Main results

In this section, we present our main results on posterior contraction. Table 2 summarizes the corresponding assumptions and theorems for the anisotropic and composite anisotropic Besov spaces.

3.1 Posterior contraction in anisotropic Besov spaces

We first establish theoretical results for anisotropic Besov spaces. The following assumptions are imposed.

(A1) The true regression function f_0 satisfies $f_0 \in U(\mathcal{B}_{p,q}^{s}) \cap \mathcal{UB}$ for some $0 < p, q \le \infty$ and $s \in \mathbb{R}_{++}^{d}$, such that $(1/p - 1/2)_+ < \tilde{s}$.

(A2) The distribution P_X has a bounded density p_X such that $\|p_X\|_\infty \le R$ for some constant $R > 0$.

(A3) The true standard deviation σ_0 satisfies $\underline{\sigma} < \sigma_0 < \overline{\sigma}$ for some constants $\underline{\sigma}$ and $\overline{\sigma}$, and the prior for σ is supported on $[\underline{\sigma}, \overline{\sigma}]$ with a positive density throughout.

Assumption (A1) states that f_0 lies in a bounded Besov class. The condition $(1/p - 1/2)_+ < \tilde{s}$ aligns with those in other theoretical studies on Besov spaces [20, 17, 18], whereas Egels and Castillo [22] impose a stronger requirement $1/p < \tilde{s}$ to achieve a sharper approximation result. Assumption (A2) is minor given that the dimension d is fixed. Assumption (A3) ensures that the prior for σ is supported on a compact interval containing σ_0. We first show that the optimal rate is achieved by suitably specified spike-and-slab and shrinkage priors, and then demonstrate that rate adaptation can be attained for both priors with slight modifications.

3.1.1 Spike-and-slab prior

Our results rely on the fact that for every $f_0 \in U(\mathcal{B}_{p,q}^{s})$, there exists a NN approximator $\hat{f} \in \Phi(L_{1n}, D_{1n}, S_{1n}, B_1)$ achieving the optimal approximation error, where the network parameters satisfy

$$L_{1n} \asymp \log n, \quad D_{1n} \asymp N_n, \quad S_{1n} \asymp N_n \log n, \quad B_1 \propto 1, \tag{2}$$

Table 2: Summary of the main results. Along with the listed assumptions, all results also require Assumptions (A2) and (A3), which are taken as common assumptions.

	Anisotropic Besov		Composite Anisotropic Besov	
Prior	Assumptions	Result	Assumptions	Result
Spike-and-Slab	(A1), (B1), (B2)	Theorem 3.3	(A4)/(A5), (B1), (B2)	Theorem 3.14(i)
Shrinkage	(A1), (C1)–(C3)	Theorem 3.8	(A4)/(A5), (C1)–(C3)	Theorem 3.14(ii)
Adaptive Spike-and-Slab	(A1), (B1), (B2)	Theorem 3.9(i)	(A4)/(A5), (B1), (B2)	Theorem 3.14(i)
Adaptive Shrinkage	(A1), (C1)–(C3)	Theorem 3.9(ii)	(A4)/(A5), (C1)–(C3)	Theorem 3.14(ii)

with $N_n = \lceil n^{1/(2\tilde{s}+1)} \rceil$. These network parameters depend on the unknown Besov parameters $\tilde{s}$ and p. See Lemma B.4 and Remark B.6 for details.

We place a spike-and-slab prior over $\Theta(L_{1n}, D_{1n}, S_{1n})$ using the network parameters in (2). This implies that the network structure is determined based on the smoothness parameter and thus the procedure does not attain rate adaptation. A spike-and-slab prior for θ is given by

$$\begin{aligned} \pi(\theta \mid \gamma, L, D, S) &= \prod_{j=1}^{T} \left[\gamma_j \tilde{\pi}_{SL}(\theta_j) + (1-\gamma_j)\delta_0(\theta_j)\right], \\ \pi(\gamma \mid L, D, S) &= \frac{1}{\binom{T}{S}} I(\gamma \in \{0,1\}^T,\ \|\gamma\|_0 = S), \end{aligned} \tag{3}$$

where $\tilde{\pi}_{SL}$ is the slab density for the nonzero components and $T = |\Theta(L, D)|$. The slab distribution is required to satisfy the following assumptions.

(B1) $\log \int_{|u|>K_n} \tilde{\pi}_{SL}(u)du \lesssim -K_n$ for any $K_n \to \infty$.

(B2) $\log \inf_{|u|\le B_1} \tilde{\pi}_{SL}(u) \gtrsim -(\log n)^2$.

Assumption (B1) implies that $\tilde{\pi}_{SL}$ has exponential tails on both sides. From a nonparametric Bayesian perspective, this condition is necessary to control the prior mass outside a chosen sieve. Assumption (B2) requires that $\tilde{\pi}_{SL}$ to place sufficient mass around the NN approximator, which is essential for ensuring adequate prior concentration in a Kullback-Leibler neighborhood. If $\tilde{\pi}_{SL}$ is independent of n and bounded away from zero on $[-B_1, B_1]$, then the assumption holds trivially.

Example 3.1 (Uniform slab prior [20, 21])**.** If $\tilde{\pi}_{SL}$ is the density of the uniform distribution $U(-C_u, C_u)$ for $C_u > B_1$, then $\tilde{\pi}_{SL}$ satisfies Assumptions (B1)–(B2).

Example 3.2 (Gaussian slab prior)**.** If $\tilde{\pi}_{SL}$ is the density of a zero-mean Gaussian distribution with fixed variance, then $\tilde{\pi}_{SL}$ satisfies Assumptions (B1)–(B2).

Example 3.1 is simple but has the drawback that the prior depends on B_1. Example 3.2 mitigates this issue. We now present the posterior contraction result.

Theorem 3.3 (Spike-and-slab prior)**.** *Suppose that Assumptions* (A1)–(A3) *hold, and that the prior distribution in* (3) *is placed over* $\Theta(L_{1n}, D_{1n}, S_{1n})$*. Assume further that the slab density* $\tilde{\pi}_{SL}$ *satisfies Assumptions* (B1)–(B2)*. Then, the posterior distribution concentrates at the rate* $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$*, in the sense that*

$$\Pi\Big((f,\sigma) \in \Phi(L_{1n}, D_{1n}, S_{1n}) \times [\underline{\sigma}, \overline{\sigma}] : \|f - f_0\|_{L^2(P_X)} + |\sigma^2 - \sigma_0^2| > M_n\epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P^{(n)}_{f_0,\sigma_0}$*-probability as* $n \to \infty$ *for any* $M_n \to \infty$*.*

Proof. See Appendix C.8. □

The contraction rate ϵ_n is near-minimax rate and only depends on intrinsic smoothness $\tilde{s}$ of the function. Suzuki and Nitanda [18] showed that the empirical risk minimizer under the squared loss achieves the same rate ϵ_n. Therefore, BNNs attain the same theoretical optimality as their frequentist counterpart for anisotropic Besov spaces.

Remark 3.4. The exponential tail condition in (B1) is imposed to effectively control model complexity, as measured by the metric entropy of a suitably chosen sieve [31, 32]. If one instead uses a polynomial-tailed density, the entropy calculation can be bypassed using the Vapnik-Chervonenkis (VC)-dimension technique [33]. In that case, the posterior contraction rate is slightly degraded to $n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^2$. See Appendix C.9 for further details.

3.1.2 Shrinkage prior

The main bottleneck in using spike-and-slab priors lies in the computational burden introduced by the point-mass component of the prior. To address this issue, we extend the results to continuous shrinkage priors, which offer a more computationally efficient alternative. These priors retain the sparsity-promoting nature of spike-and-slab priors while avoiding the complexity of variable-dimensional posterior inference. This leads to scalable and practical Bayesian procedures suitable for high-dimensional settings, where computational tractability is essential.

Using the network parameters in (2), we place a prior over $\Theta(L_{1n}, D_{1n})$. That is, sparsity is not imposed explicitly but is instead induced implicitly through the use of shrinkage priors. We express a shrinkage prior as

$$\pi(\theta \mid L, D) = \prod_{j=1}^{T} \tilde{\pi}_{SH}(\theta_j), \tag{4}$$

where $\tilde{\pi}_{SH}$ is a density on $\mathbb{R}$ and $T = |\Theta(L, D)|$. Specifically, $\tilde{\pi}_{SH}$ is assumed to satisfy the following assumptions.

(C1) $\log \int_{|u|>K_n} \tilde{\pi}_{SH}(u)du \lesssim -K_n$ for any $K_n \to \infty$.

(C2) $\log \inf_{|u|\le B_1} \tilde{\pi}_{SH}(u) \gtrsim -(\log n)^2$.

(C3) $\log \int_{|u|>a_n} \tilde{\pi}_{SH}(u)du \le -C_A(\log n)^2$ for a sufficiently large constant $C_A > 0$, where $a_n = e^{-2L_{1n}\log n}$.

Assumptions (C1)–(C2) serve similar roles to Assumptions (B1)–(B2). Assumption (C3) imposes an additional constraint that serves as a continuous analogue of the spike component in spike-and-slab priors. This assumption is indeed restrictive: for rapidly decreasing a_n, it requires the prior to concentrate nearly all its mass on $[-a_n, a_n]$. As a result, widely used shrinkage priors such as the horseshoe [34] do not satisfy Assumptions (C1)–(C3) directly. Nevertheless, the assumptions can be satisfied by designing priors that place most of their mass near zero, as required in Assumption (C3), while also satisfying the exponential tail decay in Assumption (C1). The following examples illustrate such constructions, with verification given in Appendices C.10 and C.11.

Example 3.5 (Relaxed spike-and-slab [20])**.** Let φ_k denote a sub-Weibull density with tail index $0 < k \le 1$ [35] satisfying $\int_{|u|>K} \varphi_k(u)\,du \le C_1 \exp(-C_2 K^{1/k})$ for any $K > 0$ and some constants $C_1, C_2 > 0$. For $C_u > B_1$, define $\tilde{\pi}_{SH}$ as

$$\tilde{\pi}_{SH}(u) = \pi_{1n}\frac{1}{\sigma_{1n}}\varphi_k\left(\frac{u}{\sigma_{1n}}\right) + \pi_{2n}U(u; -C_u, C_u), \tag{5}$$

where $\pi_{1n} = 1 - e^{-2C_A(\log n)^2}$, $\pi_{2n} = e^{-2C_A(\log n)^2}$, $\sigma_{1n} = a_n(2C_A(\log n)^2)^{-k}$ and a sufficiently large $C_A > 0$. Then, $\tilde{\pi}_{SH}$ satisfies Assumptions (C1)–(C3).

Example 3.6 (Relaxed spike-and-slab; Gaussian slab)**.** If the uniform component in (5) is replaced by the density of a zero-mean Gaussian distribution with fixed variance, then $\tilde{\pi}_{SH}$ satisfies Assumptions (C1)–(C3).

Remark 3.7. For $k = 1$, a sub-Weibull density φ_k corresponds to a sub-exponential density. For $k = 1/2$, it corresponds to a sub-Gaussian density. In particular, if φ_k is chosen as a Gaussian density, the prior in Example 3.6 becomes a Gaussian mixture prior.

Note that the mixture weight π_{1n} in (5) approaches 1. Although this may be practically undesirable, it ensures that Assumption (C3) is satisfied. Similar constraints have also been adopted in the literature on sparse BNNs with shrinkage priors [36, 20]. We now formalize the contraction result for shrinkage priors. The following theorem shows that BNNs equipped with shrinkage priors achieve the near-minimax rate.

Theorem 3.8 (Shrinkage prior)**.** *Suppose that Assumptions* (A1)–(A3) *hold, and that the prior distribution in* (4) *is placed over* $\Theta(L_{1n}, D_{1n})$*. Assume further that the continuous density* $\tilde{\pi}_{SH}$ *satisfies Assumptions* (C1)–(C3)*. Then, the posterior distribution concentrates at the rate* $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$*, in the sense that*

$$\Pi\Big((f,\sigma) \in \Phi(L_{1n}, D_{1n}) \times [\underline{\sigma}, \overline{\sigma}] : \|f - f_0\|_{L^2(P_X)} + |\sigma^2 - \sigma_0^2| > M_n\epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P_{f_0,\sigma_0}^{(n)}$*-probability as* $n \to \infty$ *for any* $M_n \to \infty$*.*

Proof. See Appendix C.12. □

3.1.3 Rate adaptation

In Section 3.1.1 and Section 3.1.2, the priors depend on the smoothness parameter s, indicating that the procedures are not rate-adaptive. By placing suitable priors on the network parameters, the results can be extended to achieve rate adaptation. Therefore, the optimal contraction rate is achieved without knowing the characteristics of the true function.

Specifically, instead of the network parameters in (2), we consider the depth $\tilde{L}_n = \lceil C_L \log n \rceil$ for a sufficiently large $C_L > 0$ and the priors π_D and π_S on the width D and sparsity S, respectively, given by

$$\pi_D(D) \propto e^{-\lambda_D D(\log D)^3}, \quad \pi_S(S) \propto e^{-\lambda_S S(\log S)^2}, \tag{6}$$

for constants $\lambda_D > 0$ and $\lambda_S > 0$.

Theorem 3.9 (Adaptation)**.** *Suppose that Assumptions* (A1)–(A3) *hold, and that the prior distributions satisfy either one of the following conditions:*

(i) The priors π_D *and* π_S *in* (6) *are assigned to* (D, S)*, respectively, and the prior in* (3) *is placed over* $\Theta(\tilde{L}_n, D, S)$ *conditional on* (D, S)*. The slab part* $\tilde{\pi}_{SL}$ *satisfies Assumptions* (B1)–(B2)*.*

(ii) The prior π_D *in* (6) *is assigned to* D*, and the prior in* (4) *is placed over* $\Theta(\tilde{L}_n, D)$ *conditional on* D*. The continuous density* $\tilde{\pi}_{SH}$ *satisfies Assumptions* (C1)–(C3)*.*

If C_L *is sufficiently large so that* $\tilde{L}_n \geq L_{1n}$*, then the posterior distribution concentrates at the rate* $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$*, in the sense that*

$$\Pi\Big((f,\sigma) \in \Phi(\tilde{L}_n) \times [\underline{\sigma}, \overline{\sigma}] : \|f - f_0\|_{L^2(P_X)} + |\sigma^2 - \sigma_0^2| > M_n\epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P_{f_0,\sigma_0}^{(n)}$*-probability as* $n \to \infty$ *for any* $M_n \to \infty$*.*

Proof. See Appendix C.13. □

In both cases, the priors are placed over $\Theta(\tilde{L}_n)$, and the procedures achieve rate adaptation once C_L is chosen sufficiently large. Although the priors do not depend on the Besov parameters for the true function, the conditions on $\tilde{\pi}_{SL}$ and $\tilde{\pi}_{SH}$ still involve these unknown quantities. Nevertheless, the required conditions can be satisfied without explicit knowledge of them. Specifically, if Example 3.1 and Example 3.5 are defined with a sufficiently large constant C_v such that $C_v > B_1$, the conditions hold regardless of the value of B_1, which depends on the Besov parameters. In contrast, Example 3.2 and Example 3.6 are already independent of these parameters. Therefore, rate adaptation can be readily achieved in practice.

3.2 Posterior contraction in composite Besov spaces

In this section, we show that BNNs attain the minimax optimal rate over composite Besov classes. In other words, we demonstrate that BNNs outperform traditional statistical approaches by achieving the minimax optimal convergence rate. Following Suzuki and Nitanda [18], we assume that the true regression function satisfies one of the following two structural conditions.

(A4) The true regression function f_0 satisfies $f_0 \in \mathcal{B}_{p,q}^{d_\circ, t_\circ, s_\circ} \cap \mathcal{UB}$ for some $0 < p, q \leq \infty$, $d_\circ$, $t_\circ$, and $s_\circ$, such that $\tilde{s}^{(1)} > (1/p - 1/2)_+$ and $\tilde{s}^{(h)} > 1/p$ for $h = 2, \ldots, H$.

(A5) The true regression function f_0 is defined as $f_0 = f_2 \circ f_1$, where $f_1 = A \cdot + b$ with $A \in \mathbb{R}^{d^{(1)} \times d^{(0)}}$ and $b \in \mathbb{R}^{d^{(1)}}$ such that $d^{(1)} \leq d^{(0)}$ and $f_1(x) \in [0,1]^{d^{(1)}}$ for all $x \in [0,1]^{d^{(0)}}$, and $f_2 \in U(\mathcal{B}_{p,q}^{s^{(2)}}) \cap \mathcal{UB}$ with $0 < p, q \leq \infty$, and $s^{(2)} \in \mathbb{R}_{++}^{d^{(1)}}$ such that $\tilde{s}^{(2)} > (1/p - 1/2)_+$.

Assumption (A4) corresponds to the composite Besov space defined in Section 2.3. We note that our assumptions are comparable to those in Suzuki and Nitanda [18]. Assumption (A5) defines a nontrivial Besov class involving an affine transformation, which is a specific case of the general class $\mathcal{B}_{p,q}^{d_\circ, t_\circ, s_\circ}$. However, it is not contained within Assumption (A4) owing to the relaxed smoothness requirement on $\tilde{s}^{(2)}$. Under Assumption (A4), define $t^{*(h)} = \underline{s}^{(h)}/\tilde{s}^{(h)}$, $\tilde{s}^{*(h)} = \tilde{s}^{(h)} \prod_{k=h+1}^{H} \{(\underline{s}^{(k)} - t^{*(k)}/p) \wedge 1\}$, $h \in [H]$, and $h^* = \arg\min_{h \in [H]} \tilde{s}^{*(h)}$. We define the intrinsic dimension as $t^* = t^{(h^*)}$ and the intrinsic smoothness as $\tilde{s}^* = \tilde{s}^{*(h^*)}$. For Assumption (A5), we adopt the same definitions with $h^* = 2$. Suzuki and Nitanda [18] obtained a lower bound of the minimax risk that roughly matches $n^{-\tilde{s}^*/(2\tilde{s}^*+1)}$. The following examples illustrate practically relevant functions that are covered by either Assumption (A4) or Assumption (A5).

Example 3.10 (Additive function; Figure 2a)**.** Suppose $f(x) = \sum_{i=1}^{d} g_i(x_i)$, where $g_i \in U(\mathcal{B}_{p,q}^{s_0})$ for $s_0 > (1/p - 1/2)_+$. Then, f is covered by Assumption (A4) in the form $f \in \mathcal{B}_{p,q}^{(d,d,1),(1,d),(s_0,s^{(2)})}$ for $s^{(2)}$ such that $\tilde{s}^{(2)}$ can be taken arbitrarily large, reflecting the additive structure.

Example 3.11 (Multiplicative function; Figure 2b)**.** Suppose $f(x) = \prod_{i=1}^{d} g_i(x_i)$, where $g_i \in U(\mathcal{B}_{p,q}^{s_0})$ for $s_0 > (1/p - 1/2)_+$. This f is also covered by Assumption (A4), similar to the additive case in Example 3.10.

Example 3.12 (Rotation; Figure 1)**.** Let $R_\tau \in \mathbb{R}^{d \times d}$ be the rotation matrix by angle τ. Suppose $f = g \circ (A \cdot + b)$, where $g \in U(\mathcal{B}_{p,q}^{s^{(2)}}) \cap \mathcal{UB}$, $A = R_\tau/\sqrt{d}$ and $b = (I - R_\tau)(1/2, \cdots, 1/2)^T$. This f is covered by Assumption (A5).

Example 3.13 (Piecewise function)**.** The indicator function of a hyper-rectangle lies in $\mathcal{B}_{p,\infty}^{s}$ if $\overline{s} \leq 1/p$ for $1 \leq p < \infty$ [37]. For hyper-rectangles $A_i \subset [0,1]^d$, suppose $f(x) = \sum_{i=1}^{H} I(x \in A_i)\, g_i(x)$, where $g_i \in U(\mathcal{B}_{p,\infty}^{s_0})$ for $(1/p - 1/2)_+ < \tilde{s}_0$ and $\overline{s}_0 \leq 1/p$. This f is covered by Assumption (A4) using Examples 3.10 and 3.11.

We now establish posterior contraction properties for the composite Besov spaces. Similar to the anisotropic Besov case, our results rely on the existence of an optimal NN approximator $\hat{f} \in \Phi(L_{2n}, D_{2n}, S_{2n}, B_2)$ under Assumption (A4), and $\hat{f} \in \Phi(L_{3n}, D_{3n}, S_{3n}, B_3)$ under Assumption (A5), where

$$L_{kn} \asymp \log n, \quad D_{kn} \asymp N_n^*, \quad S_{kn} \asymp N_n^* \log n, \quad B_k \propto 1, \quad k = 2, 3,$$

with $N_n^* = \lceil n^{1/(2\tilde{s}^{(*)}+1)} \rceil$. Therefore, the network parameters share the same asymptotic orders for $k = 2, 3$, although their specific values differ. For further details, see Lemma B.4 and Remark B.6. The posterior contraction results are formalized below.

Theorem 3.14 (Composite anisotropic Besov)**.** *Suppose that Assumptions* (A2)–(A3) *hold, and that either Assumption* (A4) *or Assumption* (A5) *is satisfied as follows.*

(i) If Assumption (A4) *holds, assume that the priors and conditions in one of Theorems 3.3, 3.8, and 3.9 are satisfied with* (L_{2n}, D_{2n}, S_{2n}) *in place of* (L_{1n}, D_{1n}, S_{1n}).

(ii) If Assumption (A5) *holds, assume that the priors and conditions in one of Theorems 3.3, 3.8, and 3.9 are satisfied with* (L_{3n}, D_{3n}, S_{3n}) *in place of* (L_{1n}, D_{1n}, S_{1n}).

Then, the posterior distribution concentrates at the rate $\epsilon_n = n^{-\tilde{s}^*/(2\tilde{s}^*+1)} (\log n)^{3/2}$.

Proof. See Appendix C.14. □

The contraction rate in Theorem 3.14 matches that of Suzuki and Nitanda [18], indicating that BNNs attain the same level of optimality. However, achieving the optimal rate in Suzuki and Nitanda [18] requires a correctly specified network structure. In contrast, our Bayesian procedure offers a clear advantage by learning the unknown compositional depth, connectivity, and anisotropic smoothness directly from data, without requiring prior structural knowledge.

4 Discussion

In this paper, we have established that sparse BNNs achieve optimal posterior contraction rates over anisotropic Besov spaces and their hierarchical compositions. We show that BNNs equipped with either spike-and-slab or continuous shrinkage priors attain a near-minimax rate that depends only on the intrinsic dimension, thereby overcoming the curse of dimensionality. Furthermore, BNNs are shown to achieve rate adaptation over both anisotropic and composite Besov classes.

Practice meets theory. Our work provides a rigorous theoretical foundation for the empirical success of NNs on complex real-world data. By establishing optimal rates over anisotropic Besov spaces, we shed light on why NNs perform well in high-dimensional settings where classical smoothness-based theory falls short. Although our analysis focuses on BNNs, which additionally offer rate adaptation and uncertainty quantification, the insights extend more broadly to understanding NN performance. The results provide actionable guidance: practitioners dealing with such data may benefit from employing sparse architectures with sparsity-inducing priors, potentially inspiring new algorithmic advances for efficient inference.

Extension. Our theoretical results are established under the Gaussian nonparametric regression model in (1). These results are readily extended to other statistical models in which the Hellinger distance can be translated into an L^2-type distance for the underlying function. As one such extension, we present posterior contraction results for nonparametric binary classification. We also show that the same contraction rates derived for the L^2-norm with respect to P_X hold for the empirical L^2-norm by applying empirical process theory. See Appendix A for details.

Future work. This study has several limitations that suggest promising directions for future research. While the spike-and-slab prior offers desirable theoretical properties, its practical implementation is hampered by the point mass at zero, which introduces substantial computational challenges. Variational approximations may partially alleviate this burden [38, 39]. Shrinkage priors address the issue differently, but our theoretical framework does not accommodate widely used choices such as the horseshoe prior [34]. Extending our results to incorporate such popular shrinkage priors is an important direction. Furthermore, a natural direction for future work is to extend our theoretical framework to modern deep learning architectures widely used in practice. Recent developments in the statistical theory of convolutional NNs [40, 41] and transformers [42, 43] offer valuable foundations for establishing posterior contraction results beyond fully connected architectures.

Acknowledgment

This research was supported by grants from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (2022R1C1C1006735, RS-2023-00217705, RS-2025-00513129) and the National Science Foundation (NSF) (DMS-2503119).

References

[1] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[2] M. Bianchini and F. Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, 2014.

[3] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[4] O. Delalleau and Y. Bengio. Shallow vs. deep sum-product networks. In *Advances in Neural Information Processing Systems*, volume 24, pages 666–674, 2011.

[5] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[6] M. Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

[7] M. Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pages 1517–1539, 2016.

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

[9] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694, 2021.

[10] M. Hein, M. Andriushchenko, and J. Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019.

[11] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1613–1622, 2015.

[12] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.

[13] J. Arbel, K. Pitas, M. Vladimirova, and V. Fortuin. A primer on Bayesian neural networks: Review and debates. *Statistical Science*, 2024.

[14] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antoran, and J. M. Hernandez-Lobato. Bayesian deep learning via subnetwork inference. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2510–2521, 2021.

[15] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[16] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[17] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[18] T. Suzuki and A. Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621, 2021.

[19] I. Kong and Y. Kim. Posterior concentrations of fully-connected Bayesian neural networks with general priors on the weights. *Journal of Machine Learning Research*, 26(94):1–60, 2025.

[20] K. Lee and J. Lee. Asymptotic properties for Bayesian neural network in Besov space. In *Advances in Neural Information Processing Systems*, volume 35, pages 5641–5653, 2022.

[21] N. G. Polson and V. Ročková. Posterior concentration for sparse deep learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[22] P. Egels and I. Castillo. Posterior and variational inference for deep neural networks with heavy-tailed weights. *Journal of Machine Learning Research*, 26(122):1–58, 2025.

[23] 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, 2021.

[24] A. Bhattacharya, D. Pati, and Y. Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

[25] Y. Yang, D. Pati, and A. Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

[26] S. M. Nikol'skii. *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, 1975.

[27] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

[28] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

[29] M. Hoffman and O. Lepski. Random rates in anisotropic regression. *The Annals of Statistics*, 30(2):325–396, 2002.

[30] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer, 2013.

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

[32] S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

[33] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

[34] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[35] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

[36] 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.

[37] W. Sickel. On the regularity of characteristic functions. In *Anomalies in Partial Differential Equations*, pages 395–441. Springer, 2021.

[38] B.-E. Chérief-Abdellatif. Convergence rates of variational inference in sparse deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1831–1842, 2020.

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

[40] M. Kohler, A. Krzyżak, and B. Walter. On the rate of convergence of image classifiers based on convolutional neural networks. *Annals of the Institute of Statistical Mathematics*, 74(6): 1085–1108, 2022.

[41] Z. Fang and G. Cheng. Optimal convergence rates of deep convolutional neural networks: Additive ridge functions. *Transactions on Machine Learning Research*, 2023.

[42] J. Kim, T. Nakamaki, and T. Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[43] Y. Jiao, Y. Lai, D. Sun, Y. Wang, and B. Yan. Approximation bounds for transformer networks with application to regression. *arXiv preprint arXiv:2504.12175*, 2025.

[44] Y. Yang. Minimax nonparametric classification. I. rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999.

[45] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[46] Y. Kim, I. Ohn, and D. Kim. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

[47] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2006.

[48] 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*, pages 17462–17491, 2023.

[49] S. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

[50] S. Jeong and V. Ročková. The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research*, 24(337):1–65, 2023.

[51] 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.

[52] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

[53] A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our main theoretical contributions on posterior contraction rates for Bayesian neural networks over anisotropic and composite Besov spaces.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations of the work in Section 4, including computational challenges, theoretical constraints, and the necessity to extend our framework to modern DNN architectures. Furthermore, we provide future avenues for research based on these limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems in the paper explicitly state assumptions, which are clearly labeled throughout the manuscript. Each theoretical result is presented with comprehensive statements that reference these assumptions. Complete proofs of all theoretical results are provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper focuses on the theoretical analysis of posterior contraction rates for Bayesian neural networks. As no numerical experiments were conducted, this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The main contribution of our work is to establish the theoretical properties of BNNs, making it purely mathematical; thus, this question is not applicable.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Since the paper does not include numerical experiments, there is no need to specify implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This is a theoretical study centered on the mathematical analysis of posterior contraction rates. As no experiments were conducted, error bars and statistical significance analyses are not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Since no computational experiments were conducted, information regarding computer resources, such as processor types, memory requirements, or execution times, is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research is fully in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper presents a purely theoretical analysis of the mathematical properties of Bayesian neural networks. It has no direct societal applications or impacts that would require discussion of positive or negative outcomes.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any models, data, or assets that could potentially be misused.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: Our paper is purely theoretical and does not utilize existing code, data, or models that require licensing. All relevant literature and theoretical works are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper is purely theoretical and does not release any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our research is purely theoretical, focusing on mathematical analysis of Bayesian neural networks, and does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This research is purely theoretical and does not involve human subjects; therefore, IRB approval is not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLMs to derive any core research findings or mathematical results.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Supplement on posterior contraction

A.1 Nonparametric regression

In this section, we present theorems that apply to general settings with independent and identically distributed (IID) data. For a semi-metric space (A, ρ), we let $\mathcal{N}(\epsilon, A, \rho)$ denote the ϵ-covering number. The following result builds on the foundational work of Ghosal et al. [32] and Ghosal and van der Vaart [31].

Lemma A.1. *Consider model* (1) *with $f_0 \in \mathcal{UB}$ and $\sigma_0 \in [\underline{\sigma}, \overline{\sigma}]$ for constants $0 < \underline{\sigma} \le \overline{\sigma}$. The prior for σ is supported on $[\underline{\sigma}, \overline{\sigma}]$. For $\mathcal{F} \subset \mathcal{UB}$, define*

$$A_\epsilon = \left\{ (f, \sigma) \in \mathcal{F} \times [\underline{\sigma}, \overline{\sigma}] : \|f - f_0\|_{L^2(P_X)} \le \frac{\epsilon}{2}, |\sigma - \sigma_0| \le \frac{\epsilon}{2} \right\}.$$

Suppose there exist a subset $\mathcal{F}_n \subset \mathcal{F}$ and a sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$ such that

$$\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2, \tag{S1}$$

$$-\log \Pi(A_{\epsilon_n}) \lesssim n\epsilon_n^2, \tag{S2}$$

$$\sup_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = o\Big(e^{-Cn\epsilon_n^2}\Big), \tag{S3}$$

for a sufficiently large $C > 0$. Then, the posterior satisfies

$$\Pi\Big((f, \sigma) \in \mathcal{F} \times [\underline{\sigma}, \overline{\sigma}] : \|f - f_0\|_{L^2(P_X)} + |\sigma^2 - \sigma_0^2| > M_n \epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P^{(n)}_{f_0, \sigma_0}$-probability as $n \to \infty$ for any $M_n \to \infty$.

Proof. See Appendix C.1. □

We focus on the case where $\Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma)$ is invariant with respect to σ, and we write it simply as $\Pi(\mathcal{F} \setminus \mathcal{F}_n)$.

A.2 Nonparametric binary classification

Suppose we have a set of n input-output observations, where each pair is an independent random sample from a binary classification model with a d-dimensional input variable $X_i \in [0,1]^d$ and an output variable $Y_i \in \{0, 1\}$:

$$P(Y_i = 1 \mid X_i) = 1 - P(Y_i = 0 \mid X_i) = (\psi_M \circ f_0)(X_i), \quad X_i \sim P_X, \quad i \in [n], \tag{S4}$$

where ψ_M is the sigmoid function $\psi_M(x) = (1 + e^{-Mx})^{-1}$ for a large $M > 0$ and $f_0 : [0,1]^d \to \mathbb{R}$ denotes the true regression function. The scaling factor M is introduced because f_0 and the prior support are restricted to $\mathcal{UB}$, whose elements have unit sup-norm. Let $P_f^{(n)}$ denote the joint distribution of $\mathcal{D}_n$ with a regression function f.

Lemma A.2. *Consider model* (S4) *with $f_0 \in \mathcal{UB}$. For $\mathcal{F} \subset \mathcal{UB}$, define*

$$A'_\epsilon = \left\{ f \in \mathcal{F} : \|f - f_0\|_{L^2(P_X)} \le \epsilon \right\}.$$

Suppose there exist a subset $\mathcal{F}_n \subset \mathcal{F}$ and a sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$ such that

$$\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2,$$

$$-\log \Pi(A'_{\epsilon_n}) \lesssim n\epsilon_n^2,$$

$$\Pi(\mathcal{F} \setminus \mathcal{F}_n) = o\Big(e^{-Cn\epsilon_n^2}\Big),$$

for a sufficiently large $C > 0$. Then, the posterior satisfies

$$\Pi\Big(f \in \mathcal{F} : \|(\psi_M \circ f) - (\psi_M \circ f_0)\|_{L^2(P_X)} > M_n \epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P^{(n)}_{f_0}$-probability as $n \to \infty$ for any $M_n \to \infty$.

Proof. See Appendix C.2. □

The following theorem shows that a plug-in classifier can achieve results analogous to those in Theorems 3.3, 3.8, 3.9 and 3.14 for classification problems, under the same assumptions and notation as in Section 3. The proof closely parallels that of the Gaussian case and is therefore omitted.

Theorem A.3 (Nonparametric classification)**.** *Consider model* (S4) *with the same prior specifications and assumptions as in Theorem 3.3 (or in Theorems 3.8, 3.9 and 3.14), except for Assumption* (A3)*. Then, the posterior distribution concentrates at the rate* $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$*, in the sense that*

$$\Pi\Big(f \in \Phi : \|(\psi_M \circ f) - (\psi_M \circ f_0)\|_{L^2(P_X)} > M_n\epsilon_n \mid \mathcal{D}_n\Big) \to 0$$

in $P_{f_0}^{(n)}$*-probability as* $n \to \infty$ *for any sequence* $M_n \to \infty$*, where* Φ *denotes the corresponding NN model space defined in the referenced theorem.*

It can be shown that the minimax rate is also attained for the excess risk for the misclassification error, defined as

$$\mathcal{E}(f_0, f) := \mathbb{E}_{f_0}[I(Y \neq I((\psi_M \circ f)(X) \geq 1/2))] - \mathbb{E}_{f_0}[I(Y \neq I((\psi_M \circ f_0)(X) \geq 1/2))],$$

implying that optimal performance is achieved in general nonparametric classification settings. It is well known that [44] the excess risk satisfies

$$\mathcal{E}(f_0, f) \leq 2\|(\psi_M \circ f) - (\psi_M \circ f_0)\|_{L^2(P_X)}.$$

This provides the posterior contraction rate with respect to the misclassification error as follows.

Corollary A.4 (Nonparametric classification with respect to the misclassification error)**.** *Consider model* (S4) *with the same prior specifications and assumptions as in Theorem 3.3 (or in Theorems 3.8, 3.9 and 3.14), except for Assumption* (A3)*. Then, the posterior distribution concentrates at the rate* $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$*, such that*

$$\Pi(f \in \Phi : \mathcal{E}(f_0, f) > M_n\epsilon_n \mid \mathcal{D}_n) \to 0$$

in $P_{f_0}^{(n)}$*-probability as* $n \to \infty$ *for any sequence* $M_n \to \infty$*, where* Φ *denotes the corresponding NN model space defined in the referenced theorem.*

Remark A.5. This result can be extended to achieve optimal convergence rates under boundary conditions, particularly under *Tsybakov's margin condition* [45], as demonstrated in frequentist approaches [46]. However, incorporating the likelihood corresponding to this loss function poses substantial challenges, and we leave this extension for future work.

A.3 Empirical norm

The following lemma describes the relationship between expectations of empirical processes.

Lemma A.6 (Theorem 19.3 of Györfi et al. [47])**.** *Let* $X,\ X_1, \cdots, X_n \in \mathbb{R}^d$ *be IID random vectors drawn from the distribution* P_X*. Let* $K_1,\ K_2 \geq 1$ *be constants and let* $\mathcal{G}$ *be a class of functions* $g : \mathbb{R}^d \to \mathbb{R}$ *such that*

$$|g(x)| \leq K_1, \quad \mathbb{E}[g(X)^2] \leq K_2\mathbb{E}[g(X)].$$

For $0 < \alpha < 1$ *and* $\epsilon > 0$*, assume that*

$$\sqrt{n}\alpha\sqrt{1-\alpha}\sqrt{\epsilon} \geq 288 \max\Big\{2K_1, \sqrt{2K_2}\Big\},$$

and that, for all $x_1, \cdots, x_n \in \mathbb{R}^d$ *and for all* $t \geq \epsilon/8$*,*

$$\frac{\sqrt{n}\alpha(1-\alpha)t}{96\sqrt{2}\max\{K_1, 2K_2\}} \geq \int_{\frac{\alpha(1-\alpha)t}{16\max\{K_1,2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \left\{g \in \mathcal{G} : \frac{1}{n}\sum_{i=1}^n g(x_i)^2 \leq 16t\right\}, \|\cdot\|_{1,n}\right)}du,$$

where $\|g\|_{1,n} = \frac{1}{n}\sum_{i=1}^n |g(X_i)|$*. Then,*

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\mathbb{E}[g(X)] - \frac{1}{n}\sum_{i=1}^n g(X_i)\right|}{\epsilon + \mathbb{E}[g(X)]} > \alpha\right) \leq 60\exp\left(-\frac{n\epsilon\alpha^2(1-\alpha)}{128 \cdot 2304\max\{K_1^2, K_2\}}\right).$$

For a function $f : [0,1]^d \to \mathbb{R}$, we define the empirical L^2-norm as $\|f\|_n = (n^{-1}\sum_{i=1}^n |f(X_i)|^2)^{1/2}$. As a corollary of Lemma A.6, we present the following result, adapted from Kong et al. [48], which establishes a relationship between the empirical norm $\|\cdot\|_n$ and the population norm $\|\cdot\|_{L^2(P_X)}$.

Lemma A.7 (Change of norm)**.** *Let $N_n = \lceil n^{1/(2\tilde{s}+1)} \rceil$, $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$, and Φ be an NN model space in Theorem 3.3 (or in Theorems 3.8, 3.9 and 3.14). Then,*

$$P_{f_0,\sigma_0}^{(n)}\left(\sup_{f\in\Phi}\frac{\left|\|f-f_0\|_{L^2(P_X)}^2 - \|f-f_0\|_n^2\right|}{M_0\epsilon_n^2 + \|f-f_0\|_{L^2(P_X)}^2} > \frac{1}{2}\right) \le 60\exp\left(-\frac{nM_0\epsilon_n^2}{8\cdot 128\cdot 2304\cdot 16}\right)$$

holds for a sufficiently large constant $M_0 > 0$ and sufficiently large n.

Proof. See Appendix C.3. □

Using Lemma A.7, we deduce that for any $f \in \Phi$ and all sufficiently large n, the following inequalities hold with probability $1 - \exp(-M_0^* n\epsilon_n^2)$, for some $M_0^* > 0$:

$$\|f-f_0\|_n^2 \le \frac{3}{2}\|f-f_0\|_{L^2(P_X)}^2 + \frac{M_0}{2}\epsilon_n^2,$$
$$\|f-f_0\|_{L^2(P_X)}^2 \le 2\|f-f_0\|_n^2 + M_0\epsilon_n^2.$$

This yields a contraction rate with respect to the empirical norm.

Theorem A.8 (Contraction rate with respect to the empirical norm)**.** *Consider model* (1)*, prior and the assumptions as in Theorem 3.3 (or in Theorems 3.8, 3.9 and 3.14). Then, the posterior distribution concentrates at the rate $\epsilon_n = n^{-\tilde{s}/(2\tilde{s}+1)}(\log n)^{3/2}$, in the sense that*

$$\Pi\big((f,\sigma)\in\Phi\times[\underline{\sigma},\overline{\sigma}] : \|f-f_0\|_n + |\sigma^2-\sigma_0^2| > M_n\epsilon_n \mid \mathcal{D}_n\big) \to 0$$

in $P_{f_0,\sigma_0}^{(n)}$-probability as $n\to\infty$ for any sequence $M_n\to\infty$, where Φ denotes the corresponding NN model space defined in the referenced theorem.

Proof. See Appendix C.4. □

B Properties of ReLU networks for Besov spaces

This section describes key properties of ReLU networks, highlighting their capacity to approximate a broad class of functions. In particular, ReLU networks exhibit the universal approximation property: with sufficient depth or width, they can approximate any anisotropic Besov function on $[0,1]^d$ to arbitrary accuracy. We develop the theoretical foundations of this property and discuss its implications for both the design and theoretical understanding of deep learning models. Let $\tilde{\Phi}(\cdot) = \{f_\theta : \theta \in \Theta(\cdot)\}$ denote the NN model space without the clip function.

For the first step, note that a ReLU NN can approximate a cardinal B-spline with arbitrary precision [18, 17]. Define $\psi(x) = 1$ for $x \in [0,1]$ and $\psi(x) = 0$ otherwise. The cardinal B-spline of order m is obtained recursively by convolution:

$$\psi_m(x) = \begin{cases}(\psi * \psi_{m-1})(x) & \text{for } m > 0,\\ \psi(x) & \text{for } m = 0,\end{cases}$$

where $(f*g)(x) := \int f(x-t)g(t)dt$. For $k \in \mathbb{Z}_+$ and $j = (j_1,\dots,j_d) \in \mathbb{Z}_+^d$, define

$$M_{k,j}^{d,m}(x) = \prod_{i=1}^d \psi_m(2^{\lfloor k s_i^{-1}\rfloor}x_i - j_i),$$

for $x \in \mathbb{R}^d$. Here, $s = (s_1,\dots,s_d) \in \mathbb{R}_{++}^d$ denotes the smoothness parameter, k controls the spatial resolution, and j determines the location at which the basis function is centered. A function f in an anisotropic Besov space can thus be approximated by a superposition of $M_{k,j}^{d,m}(x)$, which is closely related to wavelet basis functions [49]. The following lemma establishes the approximation of the cardinal B-spline basis by ReLU activations.

Lemma B.1 (Approximation of B-spline basis by NNs; Lemma 1 of Suzuki and Nitanda [18])**.** *There exists a constant $c(d,m)$, depending only on d and m, and an NN $\hat{M} \in \tilde{\Phi}(L_0, D_0, S_0, B_0)$ with*

$$\begin{aligned} L_0 &:= 3 + 2\left\lceil \log_2\left(\frac{3^{d\vee m}}{\epsilon c(d,m)}\right) + 5\right\rceil \lceil \log_2(d\vee m)\rceil, \\ D_0 &:= 6dm(m+2) + 2d, \\ S_0 &:= L_0 D_0^2, \\ B_0 &:= 2(m+1)m \end{aligned}$$

that satisfies $\hat{M}(x) = 0$ for all $x \notin [0, m+1]^d$ and $\|M_{0,0}^{d,m} - \hat{M}\|_\infty \le \epsilon$ for all $\epsilon > 0$.

For order $m \in \mathbb{N}$ of the cardinal B-spline bases, define

$$J(k) := \prod_{j=1}^{d} J_j(k), \quad J_i(k) = \{-m, -m+1, \cdots, 2^{\lfloor k s_i^{-1}\rfloor}\},$$

and define the quasi-norm (*Besov sequence norm*) of the sequence $(\alpha_{k,j})_{k\in\mathbb{Z}_+, j\in J(k)}$ as

$$\left\|(\alpha_{k,j})_{k,j}\right\|_{b^s_{p,q}} = \left\{\sum_{k=0}^{\infty}\left[2^{k\left[\underline{s}-\left(\sum_{i=1}^d \lfloor k s_i^{-1}\rfloor/k\right)/p\right]}\left(\sum_{j\in J(k)}|\alpha_{k,j}|^p\right)^{1/p}\right]^q\right\}^{1/q}. \tag{S5}$$

For $p = \infty$ or $q = \infty$, the norm is modified in the standard way.

As noted in Remark B.2, anisotropic Besov spaces admit useful continuous embeddings.

Remark B.2 (Continuous embeddings in anisotropic Besov spaces; Proposition 1 of Suzuki and Nitanda [18])**.** Let $s = (s_1, \cdots, s_d) \in \mathbb{R}^d_{++}$, Then, the following continuous embeddings hold on a bounded domain $\Omega \subset \mathbb{R}^d$.

- If $s_1 = \cdots = s_d = s_0 \notin \mathbb{N}$, then $C^{s_0}(\Omega) = \mathcal{B}^{(s_0,\ldots,s_0)}_{\infty,\infty}(\Omega)$.
- For $0 < p_1 \le p_2 \le \infty$, $0 < q \le \infty$, if $\tilde{s} > \left(1/p_1 - 1/p_2\right)_+$, then $\mathcal{B}^s_{p_1,q}(\Omega) \hookrightarrow \mathcal{B}^{\gamma\, s}_{p_2,q}(\Omega)$, $\gamma = 1 - (1/p_1 - 1/p_2)_+/\tilde{s}$.
- For $0 < p \le \infty$ and $0 < q_1 < q_2 \le \infty$, $\mathcal{B}^s_{p,q_1}(\Omega) \hookrightarrow \mathcal{B}^s_{p,q_2}(\Omega)$.
- If $0 < p, q \le \infty$ and $\tilde{s} > 1/p$, $\mathcal{B}^s_{p,q}(\Omega) \hookrightarrow C^0(\Omega)$.

The notation $\hookrightarrow$ denotes a continuous embedding from the space on the left into the space on the right. In particular, if $\tilde{s} > 1/p$, then $\mathcal{B}^s_{p,q}(\Omega) \hookrightarrow C^{\gamma\underline{s}}(\Omega)$ with $\gamma = 1 - 1/(\tilde{s}p)$. In this case, every function in $\mathcal{B}^s_{p,q}(\Omega)$ is continuous. Otherwise, the functions may display discontinuities or non-smooth behavior such as jumps or spikes.

Building on these embedding properties, the following lemma establishes the approximation of anisotropic Besov functions by a cardinal B-spline basis, thereby extending the classical spline approximation results for isotropic Besov spaces [17] to the anisotropic setting.

Lemma B.3 (Approximation of anisotropic Besov function by cardinal B-spline basis; Lemma 2 of Suzuki and Nitanda [18])**.** *Suppose that $0 < p, q, r \le \infty$, $\omega := (1/p - 1/r)_+ < \tilde{s}$, and $\nu = (\tilde{s} - \omega)/(2\omega)$. Assume that $0 < \overline{s} < \min\{m, m - 1 + 1/p\}$, where $m \in \mathbb{N}$ is the order of the cardinal B-spline bases. For any $f \in \mathcal{B}^s_{p,q}$ and $N > 0$, define $K(N)$ such that $2^{\sum_{i=1}^d \lfloor K(N)\underline{s}/s_j\rfloor} = N$ and $K^*(N) = \lceil K(N)(1 + 1/\nu)\rceil$. Then, there exist $E_N \subset \{(k,j) : 1 \le k \le K^*(N), j \in J(k)\}$ with $|E_N| \le N$ and f_N such that*

$$\begin{aligned} f_N(x) &= \sum_{(k,j)\in E_N} \alpha_{k,j} M^{d,m}_{k,j}(x), \\ \|f - f_N\|_{L^r} &\lesssim N^{-\tilde{s}} \|f\|_{\mathcal{B}^s_{p,q}}, \end{aligned}$$

where the coefficient $(\alpha_{k,j})$ yields the following norm equivalence

$$\|f\|_{\mathcal{B}^s_{p,q}} \simeq \|(\alpha_{k,j})_{k,j}\|_{b^s_{p,q}}. \tag{S6}$$

The following lemma extends Proposition 2 of Suzuki and Nitanda [18], establishing that for any function f_0 in an anisotropic Besov space, there exist NNs that approximate f_0 with arbitrary accuracy. In this version, the dependence of the hyperparameter B on N in the original proposition is removed.

Lemma B.4 (Approximation of anisotropic Besov functions by NNs)**.** *Suppose that $0 < p, q, r \leq \infty$, $\omega := (1/p - 1/r)_+ < \tilde{s}$ and $\nu = (\tilde{s} - \omega)/(2\omega)$. Assume that $N \in \mathbb{N}$ is sufficiently large and $m \in \mathbb{N}$ satisfies $0 < \bar{s} < \min\{m, m - 1 + 1/p\}$. Then,*

$$\sup_{f_0 \in U(\mathcal{B}^s_{p,q}([0,1]^d)) \cap \mathcal{UB}} \inf_{f \in \Phi(L_1, D_1, S_1, B_1)} \|f_0 - f\|_{L^r} \lesssim N^{-\tilde{s}}$$

with

$$\begin{aligned} L_1 = L_1(d, m, p, r, s, N) &:= 3 + 3\left\lceil \log_2\left(\frac{3^{d \vee m}}{\epsilon c(d, m)}\right) + 5\right\rceil \lceil \log_2(d \vee m) \rceil, \\ D_1 = D_1(d, m, N) &:= N D_0, \\ S_1 = S_1(d, m, p, r, s, N) &:= L_1 D_0^2 N, \\ B_1 = B_1(d, m, p, r, s) &:= B_0 \exp\left(\frac{(1 + \nu^{-1})[(1/\underline{s}) \vee (1/p - \tilde{s})_+] \log 2}{\lceil \tilde{s} \log_2(d \vee m) \rceil}\right) \end{aligned} \tag{S7}$$

where $D_0 = D_0(d, m) := 6dm(m + 2) + 2d$, $B_0 = B_0(m) := 2(m + 1)m$, $\epsilon = N^{-\tilde{s} - (1+\nu^{-1})(1/p - \tilde{s})_+} \log(N)^{-1}$, and constant $c(d, m)$ does not depend on N.

Proof. See Appendix C.5 □

Remark B.5. The key strength of Lemma B.4 lies in that the approximation accuracy depends not on the ambient dimension d but on the intrinsic dimension. When NN models are employed, if the target function exhibits anisotropic smoothness, the approximation rate is governed by the intrinsic rather than the ambient dimension. This leads to minimax-optimal performance in estimation theory and illustrates how NNs can mitigate the curse of dimensionality. In the isotropic case where $s = (s_0, s_0, \ldots, s_0)$, we have $\tilde{s} = s_0/d$, recovering the classical results for ordinary Besov functions in Suzuki [17].

Remark B.6. Note that the NN space $\Phi(L, D, S, B)$ is monotonically increasing with respect to L, D, S, and B. Thus, the parameters in (S7) can be replaced by

$$L_1 = C_L \log N, \quad D_1 = C_D N, \quad S_1 = C_S N \log N, \quad B_1 = C_B$$

for sufficiently large constants C_L, C_D, C_S, and C_B.

NN models with ReLU activation permit redistribution of weights across layers without altering the output function. The following lemma formalizes this property, which will be used to redistribute the magnitudes of the parameters.

Lemma B.7 (Re-scaling Lemma for ReLU NNs; Lemma A.1 of Kong and Kim [19])**.** *Let $L \geq 3$ and $D \geq 3$. For positive constants $c_1, \cdots, c_{L+1}$, we define*

$$\tilde{W}^{(l)} := c_l W^{(l)}, \ \tilde{b}^{(l)} = \left(\prod_{l'=1}^{l} c_{l'}\right) b^{(l)}$$

for $l \in [L + 1]$ and define $\tilde{\theta} = (\tilde{W}^{(1)}, \tilde{b}^{(1)}, \ldots, \tilde{W}^{(L+1)}, \tilde{b}^{(L+1)})$. If $\prod_{l=1}^{L+1} c_l = 1$, $f_\theta(x) = f_{\tilde{\theta}}(x)$ holds for every x.

Remark B.8. The clipping function can be implemented using a simple ReLU network. Specifically,

$$\mathrm{clip}(x) = \max\{\min\{x, 1\}, -1\} = \zeta(x + 1) - \zeta(x - 1) - 1,$$

where $\zeta(\cdot)$ denotes the ReLU activation. Consequently, the clipped network space $\Phi(\cdot)$ is contained in the unrestricted network space $\tilde{\Phi}(\cdot)$. Furthermore, to ensure that each partial block of a composite Besov function produces outputs within $[0, 1]$, we apply

$$\mathrm{clip}_{[0,1]}(x) := \zeta(x) - \zeta(x - 1) = \max\{x, 0\} - \max\{x - 1, 0\}.$$

With this construction, ReLU networks are capable of approximating composite Besov functions.

The following lemma extends Theorem 1 of Suzuki and Nitanda [18], which establishes an approximation bound for any f_0 in a composite anisotropic Besov space by NNs. Unlike their setting, which imposes no restriction on the input dimension of each component, we adopt the composite Hölder framework of Schmidt-Hieber [8] and assume that each layer map depends only on a small subset of its inputs. By combining this effective low-dimensionality constraint with anisotropic smoothness, we exploit both compositional structure and reduced intrinsic dimension to efficiently approximate functions in high ambient dimensions.

Lemma B.9. *Suppose that $\tilde{s}^{(1)} > (1/p - 1/r)_+$ for some $r > 0$ and $\tilde{s}^{(h)} > 1/p$ for all $h \geq 2$. Then, the approximation error over the composite anisotropic Besov space is bounded as*

$$\sup_{f_0 \in \mathcal{B}_{p,q}^{d_\circ, t_\circ, s_\circ}} \inf_{f \in \Phi(L_2, D_2, S_2, B_2)} \|f_0 - f\|_{L^r} \lesssim N^{-\tilde{s}^*},$$

where $\tilde{s}^{*(h)} = \tilde{s}^{(h)} \prod_{k=h+1}^{H} \{(\underline{s}^{(k)} - t^{*(k)}/p) \wedge 1\}$, $\tilde{s}^* = \min_{h \in [h]} \tilde{s}^{*(h)}$ and

$$\begin{aligned}
L_2 &= \sum_{h=1}^{H} \Big\{ L_1(t^{(h)}, m^{(h)}, p, r^{(h)}, s^{(h)}, N) + 1 \Big\}, \\
D_2 &= \max_{h \geq 1} \{ D_1(t^{(h)}, m^{(h)}, N) \vee t^{(h+1)} \}, \\
S_2 &= \sum_{h=1}^{H} d^{(h)} \{ S_1(t^{(h)}, m^{(h)}, p, r^{(h)}, s^{(h)}, N) + 4 \}, \\
B_2 &= \max_{h \geq 1} B_1(t^{(h)}, m^{(h)}, p, r^{(h)}, s^{(h)}).
\end{aligned}$$

Here, $r^{(1)} = r$, $r^{(h)} = \infty$ for $h \geq 2$, and $m^{(h)} \in \mathbb{N}$ satisfies $0 < \tilde{s}^{(h)} < \min\{m^{(h)}, m^{(h)} - 1 + 1/p\}$.

Proof. See Appendix C.6. □

Lemma B.10 (Theorem 6 of Suzuki and Nitanda [18])**.** *Suppose that $d^{(1)} \leq d^{(0)}$ and $s^{(2)} \in \mathbb{R}_{++}^{d^{(1)}}$, $\tilde{s}^{(2)} > (1/p - 1/r)_+$ for some $r > 0$. Consider the anisotropic Besov space that involves an affine transformation*

$$\mathcal{F} := \Big\{ f_2 \circ f_1 : f_1 = A \cdot + b,\ A \in \mathbb{R}^{d^{(1)} \times d^{(0)}},\ b \in \mathbb{R}^{d^{(0)}},\ \|A\|_\infty \vee \|b\|_\infty \leq C_a,\ f_2 \in \mathcal{B}_{p,q}^{s^{(2)}} \Big\}.$$

Then, the approximation error on the function space is bounded as

$$\sup_{f_0 \in \mathcal{F}} \inf_{f \in \Phi(L_3, D_3, S_3, B_3)} \|f_0 - f\|_{L^r} \lesssim N^{-\tilde{s}^{(2)}},$$

where

$$\begin{aligned}
L_3 &= L_1(d^{(1)}, m^{(2)}, p, r, s^{(2)}, N), \\
D_3 &= D_1(d^{(1)}, m^{(2)}, N), \\
S_3 &= S_1(d^{(1)}, m^{(2)}, p, r, s^{(2)}, N), \\
B_3 &= (C_a d^{(1)} + 1) B_1(d^{(1)}, m^{(2)}, p, r, s^{(2)})
\end{aligned}$$

for $m^{(2)} \in \mathbb{N}$ satisfying $0 < \tilde{s}^{(2)} < \min\{m^{(2)}, m^{(2)} - 1 + 1/p\}$.

To apply Lemma A.1, we require results concerning the complexity of the model space and the asymptotic properties of neural network models. By the Lipschitz continuity of clip,

$$|(\mathrm{clip} \circ f_1)(x) - (\mathrm{clip} \circ f_2)(x)| \leq |f_1(x) - f_2(x)|.$$

It follows that the covering number of the clipped network class can be bounded by that of the unrestricted class.

Lemma B.11 (Covering number; Lemma 3 of Suzuki [17])**.** *For any $\epsilon > 0$, $D \geq 3$, and $L \geq 3$.*

$$\log \mathcal{N}\Big(\epsilon, \tilde{\Phi}(L, D, S, B), \|\cdot\|_\infty\Big) \leq (S+1) \log\big(2\epsilon^{-1} L (B \vee 1)^L (D+1)^{2L}\big).$$

To allow for relaxed sparsity with a margin $a > 0$, we define

$$\Theta(L, D, S, B, a) = \left\{\theta : (\theta_i I(|\theta_i| > a))_{i=1}^T \in \Theta(L, D, S, B)\right\}, \tag{S8}$$

where $T = |\Theta(L, D)|$, and denote the corresponding NN class by $\tilde{\Phi}(L, D, S, B, a)$ and $\Phi(L, D, S, B, a)$.

Lemma B.12 (Lemma 5 of Lee and Lee [20])**.** *For all $\epsilon \geq 2aL(B \vee 1)^{L-1}(D+1)^L$, $D \geq 3$, and $L \geq 3$.*

$$\log \mathcal{N}\Big(\epsilon, \tilde{\Phi}(L, D, S, B, a), \|\cdot\|_\infty\Big) \leq (S+1)\log\big(2\epsilon^{-1}L(B \vee 1)^L(D+1)^{2L}\big).$$

The following lemma, introduced by Schmidt-Hieber [8], plays a key role in the proofs of Lemma B.11 and Lemma B.12. It characterizes the distance between NN models in terms of their parameters.

Lemma B.13. *For any $\epsilon > 0$ and $\theta,\ \theta^* \in \Theta(L, D, S, B)$ satisfying $\|\theta - \theta^*\|_\infty < \epsilon$,*

$$\|f_\theta - f_{\theta^*}\|_\infty \leq \epsilon L(B \vee 1)^{L-1}(D+1)^L.$$

Proof. See Appendix C.7. □

C Proofs of the technical results

C.1 Proof of Lemma A.1

Proof. We follow the proof of Theorem 2 in Jeong and Ročková [50]. Since there exist upper and lower bounds for σ_0^2 and σ^2, the rate for $|\sigma^2 - \sigma_0^2|$ is equivalent to that for $|\sigma - \sigma_0|$; hence we use the latter. For every $(f_1, \sigma_1), (f_2, \sigma_2) \in \mathcal{F} \times [\underline{\sigma}, \overline{\sigma}]$, define

$$\rho_L^2((f_1, \sigma_1), (f_2, \sigma_2)) := \|f_1 - f_2\|_{L^2(P_X)}^2 + |\sigma_1 - \sigma_2|^2.$$

By direct calculation, it is straightforward to verify that for some $C_H > 1$,

$$C_H^{-1}\rho_L((f_1, \sigma_1), (f_2, \sigma_2)) \leq \rho_H(p_{f_1,\sigma_1}, p_{f_2,\sigma_2}) \leq C_H\rho_L((f_1, \sigma_1), (f_2, \sigma_2), \tag{S9}$$

where ρ_H denotes the Hellinger distance and $p_{f,\sigma}$ is the density with f and σ. Therefore, the Hellinger distance is equivalent to ρ_L up to a constant factor and it suffices to establish the theorem with respect to the Hellinger distance, which admits an exponentially powerful test function.

Using (S9), the Hellinger entropy can instead be estimated by replacing it with ρ_L. Note that $\rho_L((f_1, \sigma_1), (f_2, \sigma_2)) \leq \epsilon$ holds if $\|f_1 - f_2\|_{L^2(P_X)} \leq \epsilon/2$ and $|\sigma_1 - \sigma_2| \leq \epsilon/2$. Hence, by (S1),

$$\log \mathcal{N}(\epsilon_n, \mathcal{F}_n \times [\underline{\sigma}, \overline{\sigma}], \rho_L) \lesssim \log \mathcal{N}\Big(\epsilon_n/2, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) - \log(\epsilon_n/2) + \log \overline{\sigma} \lesssim n\epsilon_n^2$$

for all sufficiently large n. Using (S2), there exists $C' > 0$ such that $\Pi(A_{\epsilon_n}) \geq \exp\big(-C'n\epsilon_n^2\big)$. Moreover, by Lemma B.2 of Xie and Xu [51],

$$\max\left\{-\mathbb{E}_{f_0,\sigma_0}\left[\log \frac{p_{f,\sigma}^{(n)}}{p_{f_0,\sigma_0}^{(n)}}\right], \mathbb{E}_{f_0,\sigma_0}\left[\left(\log \frac{p_{f,\sigma}^{(n)}}{p_{f_0,\sigma_0}^{(n)}}\right)^2\right]\right\} \leq C''\rho_L^2((f, \sigma), (f_0, \sigma_0))$$

for some constant $C'' > C'/(C-4)$ and all sufficiently large n. We obtain

$$\begin{aligned}
&\Pi\left(\max\left\{-\mathbb{E}_{f_0,\sigma_0}\left[\log \frac{p_{f,\sigma}^{(n)}}{p_{f_0,\sigma_0}^{(n)}}\right], \mathbb{E}_{f_0,\sigma_0}\left[\left(\log \frac{p_{f,\sigma}^{(n)}}{p_{f_0,\sigma_0}^{(n)}}\right)^2\right]\right\} \leq \epsilon_n^2\right) \\
&\geq \Pi\big(\rho_L^2((f, \sigma), (f_0, \sigma_0)\} \leq \epsilon_n^2/C''\big) \\
&\geq \Pi(A_{\epsilon_n/\sqrt{C''}}) \\
&\geq \exp\big(-n\epsilon_n^2 C'/C''\big) \\
&\geq \exp\big(-(C-4)n\epsilon_n^2\big)
\end{aligned}$$

for all sufficiently large n. In addition, (S3) implies

$$\Pi(\mathcal{F} \setminus \mathcal{F}_n) = \int \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma^2) d\Pi(\sigma^2) = o(e^{-Cn\epsilon_n^2}).$$

We get the desired result using Theorem 2.1 of Ghosal et al. [32]. □

C.2 Proof of Lemma A.2

Proof. Let $\rho_L(f_1, f_2) = \|(\psi_M \circ f_1) - (\psi_M \circ f_2)\|_{L^2(P_X)}$ for $f_1, f_2 \in \mathcal{F}$. The proof proceeds by first establishing the equivalence between the Hellinger distance ρ_H and ρ_L, and then deriving the contraction rate with respect to ρ_H. For any $f_1, f_2 \in \mathcal{F}$, let $\eta_1 = \psi_M \circ f_1$ and $\eta_2 = \psi_M \circ f_2$. Then,

$$\rho_L^2(f_1, f_2) = \frac{1}{2}\int \left(\sqrt{\eta_1(x)} - \sqrt{\eta_2(x)}\right)^2 \left(\sqrt{\eta_1(x)} + \sqrt{\eta_2(x)}\right)^2 P_X(dx) + \frac{1}{2}\int \left(\sqrt{1-\eta_1(x)} - \sqrt{1-\eta_2(x)}\right)^2 \left(\sqrt{1-\eta_1(x)} + \sqrt{1-\eta_2(x)}\right)^2 P_X(dx).$$

Since both η_1 and η_2 lie in $[\delta, 1-\delta]$ for some $\delta > 0$, there exists a constant $C_H > 0$ such that

$$C_H^{-1}\rho_L(f_1, f_2) \le \rho_H(p_{f_1}, p_{f_2}) \le C_H \rho_L(f_1, f_2).$$

Hence, it suffices to work with the Hellinger distance. Since the sigmoid function ψ_M is $M/4$-Lipschitz, we obtain

$$\rho_L(f_1, f_2) = \|\eta_1 - \eta_2\|_{L^2(P_X)} \le \frac{M}{4}\|f_1 - f_2\|_{L^2(P_X)}.$$

Consequently, the Hellinger entropy is bounded by $\log \mathcal{N}(\epsilon_n, \mathcal{F}_n, \rho_L) \lesssim \log(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}) \lesssim n\epsilon_n^2$. Let $p_f^{(n)}$ denote the joint density under model (S4) with f. By Lemma 2.8 of Ghosal and van der Vaart [52], for any measurable functions f_1 and f_2,

$$\max\left\{-\mathbb{E}_{f_1}\left[\log \frac{p_{f_1}^{(n)}}{p_{f_2}^{(n)}}\right], \mathbb{E}_{f_1}\left[\left(\log \frac{p_{f_1}^{(n)}}{p_{f_2}^{(n)}}\right)^2\right]\right\} \lesssim \|f_1 - f_2\|_{L^2(P_X)}.$$

The remainder of the proof follows the same argument as in Lemma A.1. □

C.3 Proof of Lemma A.7

Proof. We prove the case corresponding to Theorem 3.3; the remaining cases follow by analogous arguments. For Lemma A.6, set $\alpha = 1/2$, $\epsilon = M_0\epsilon_n^2$, $K_1 = K_2 = 4$, and

$$\mathcal{G} = \big\{g(x) = (f(x) - f_0(x))^2 : f \in \Phi\big\},$$

where $\Phi = \Phi(L_{1n}, D_{1n}, S_{1n})$. The choice $K_1 = K_2 = 4$ is justified because

$$|g(x)| \le 4, \quad \mathbb{E}[g(X)^2] \le 4\mathbb{E}[g(X)]$$

for every $g \in \mathcal{G}$. Since $n\epsilon_n^2 \to \infty$, we have

$$\sqrt{n}\alpha\sqrt{1-\alpha}\sqrt{\epsilon_n^2} \ge 288 \max\Big\{2K_1, \sqrt{2K_2}\Big\}$$

for all sufficiently large n. Moreover, for any $f_1, f_2 \in \Phi$,

$$\big\|(f_1 - f_0)^2 - (f_2 - f_0)^2\big\|_{1,n} \le 4\|f_1 - f_2\|_{1,n},$$

which implies

$$\log \mathcal{N}\left(u, \left\{g \in \mathcal{G} : \frac{1}{n}\sum_{i=1}^n g(x_i)^2 \le 16t\right\}, \|\cdot\|_{1,n}\right) \le \log \mathcal{N}\Big(u/4, \Phi, \|\cdot\|_{1,n}\Big).$$

Using Lemma B.11, we obtain for every $u \gtrsim \epsilon_n^2$,

$$\begin{aligned}
\log \mathcal{N}\Big(u/4, \Phi, \|\cdot\|_{1,n}\Big) &\le \log \mathcal{N}(u/4, \Phi, \|\cdot\|_\infty) \\
&\le (S_n + 1)\left(\log L_n + L_n \log\big((B_n \vee 1)(D_n + 1)^2\big) - \log \frac{u}{8}\right) \\
&\lesssim N_n (\log n)^3 \\
&\lesssim n\epsilon_n^2.
\end{aligned}$$

Therefore, for all $t \ge M_0^2\epsilon_n^2/8$,

$$\int_{\frac{\alpha(1-\alpha)t}{16\max\{K_1, 2K_2\}}}^{\sqrt{t}} \sqrt{\log \mathcal{N}\left(u, \left\{g \in \mathcal{G} : \frac{1}{n}\sum_{i=1}^n g(x_i)^2 \le 16t\right\}, \|\cdot\|_{1,n}\right)}\, du \lesssim \sqrt{tn\epsilon_n^2}.$$

Since $\sqrt{\epsilon_n^2/t} \le \sqrt{8/M_0^2}$ is sufficiently small, we conclude the desired result. □

C.4 Proof of Theorem A.8

Proof. Let $\mathcal{E}_n$ denote the event

$$\mathcal{E}_n = \left\{ \sup_{f\in\Phi} \frac{\left| \|f-f_0\|_{L^2(P_X)}^2 - \|f-f_0\|_n^2 \right|}{M_0\epsilon_n^2 + \|f-f_0\|_{L^2(P_X)}^2} \le \frac{1}{2} \right\}.$$

Then $P_{f_0,\sigma_0}^{(n)}(\mathcal{E}_n^c) \lesssim \exp(-Cn\epsilon_n^2)$ for some constant $C > 0$ by Lemma A.7. On the event $\mathcal{E}_n$, for any $f \in \Phi$,

$$\|f-f_0\|_n \le \sqrt{\frac{3}{2}\|f-f_0\|_{L^2(P_X)}^2 + \frac{M_0}{2}\epsilon_n^2} \le 2\|f-f_0\|_{L^2(P_X)} + \sqrt{\frac{M_0}{2}}\epsilon_n.$$

We now express

$$\begin{aligned}\Pi\big(\|f-f_0\|_n + \left|\sigma^2-\sigma_0^2\right| > M_n\epsilon_n \mid \mathcal{D}_n\big) &= \Pi\big(\|f-f_0\|_n + \left|\sigma^2-\sigma_0^2\right| > M_n\epsilon_n \mid \mathcal{D}_n\big)\, I(\mathcal{E}_n) \\ &\quad + \Pi\big(\|f-f_0\|_n + \left|\sigma^2-\sigma_0^2\right| > M_n\epsilon_n \mid \mathcal{D}_n\big)\, I(\mathcal{E}_n^c).\end{aligned}$$

For the first term, observe that

$$\begin{aligned}&\Pi\big(\|f-f_0\|_n + \left|\sigma^2-\sigma_0^2\right| > M_n\epsilon_n \mid \mathcal{D}_n\big)\, I(\mathcal{E}_n) \\ &\le \Pi\Big(\|f-f_0\|_{L^2(P_X)} + \left|\sigma^2-\sigma_0^2\right| > (M_n - C')\epsilon_n/2 \mid \mathcal{D}_n\Big)\, I(\mathcal{E}_n)\end{aligned}$$

for some constant $C' > 0$. By Theorem 3.3 (or in Theorems 3.8, 3.9 and 3.14), this converges to 0 in $P_{f_0,\sigma_0}^{(n)}$-probability as $n \to \infty$ for any $M_n \to \infty$. For the second term, since $P_{f_0,\sigma_0}^{(n)}(\mathcal{E}_n^c) \to 0$, it also converges to 0 in $P_{f_0,\sigma_0}^{(n)}$-probability as $n \to \infty$. This proves the assertion. □

C.5 Proof of Lemma B.4

Proof. Using Lemma B.3, for $K = K(N)$, $K^* = \lceil K(1+\nu^{-1})\rceil$, and E_N with $|E_N| \le N$, there exists f_N such that

$$\begin{aligned}f_N(x) &= \sum_{(k,j)\in E_N} \alpha_{k,j} M_{k,j}^{d,m}(x), \\ \|f - f_N\|_{L^r} &\lesssim N^{-\tilde{s}}.\end{aligned}$$

Since for every k and j,

$$M_{k,j}^{d,m}(x) = M_{0,0}^{d,m}(2^{\lfloor ks_1^{-1}\rfloor}x_1 - j_1, \ldots, 2^{\lfloor ks_d^{-1}\rfloor}x_d - j_d), \tag{S10}$$

Lemma B.1 implies that there exists an NN approximator $\hat{M}_{k,j}^{d,m}$ satisfying $\|M_{k,j}^{d,m} - \hat{M}_{k,j}^{d,m}\|_\infty \le \epsilon$ for any $\epsilon > 0$. By the definition of B-splines, it follows that

$$\sum_{(k,j)\in E_N} I\Big(M_{k,j}^{d,m}(x) \ne 0\Big) \le (m+1)^d(K^*+1).$$

Moreover, using (S5), (S6), and the inequality $N = 2^{\sum_{j=1}^d \lfloor K\underline{s}/s_j\rfloor} \gtrsim 2^{Kd^*}$, we obtain

$$|\alpha_{k,j}| \lesssim 2^{K^*d^*(1/p-\tilde{s})_+} \lesssim N^{(1+\nu^{-1})(1/p-\tilde{s})_+}. \tag{S11}$$

We now define the approximator

$$\hat{f} = \sum_{(k,j)\in E_N} \alpha_{k,j}\hat{M}_{k,j}^{d,m}.$$

Combining the above bounds yields

$$\begin{aligned}|f_N(x) - \hat{f}(x)| &\le \sum_{(k,j)\in E_N} |\alpha_{k,j}|\Big|M_{k,j}^{d,m}(x) - \hat{M}_{k,j}^{d,m}(x)\Big| \\ &\le \epsilon \sum_{(k,j)\in E_N} |\alpha_{k,j}| I\Big(M_{k,j}^{d,m}(x) \ne 0\Big) \\ &\lesssim \epsilon(m+1)^d(K^*+1)2^{K^*d^*(1/p-\tilde{s})_+} \\ &\lesssim \epsilon\log(N)N^{(1+\nu^{-1})(1/p-\tilde{s})_+} \\ &= N^{-\tilde{s}}.\end{aligned}$$

Using the Lipschitz continuity of clip, we get the desired approximation bound:

$$\|f - (\mathrm{clip} \circ \hat{f})\|_{L^r} = \|(\mathrm{clip} \circ f) - (\mathrm{clip} \circ \hat{f})\|_{L^r} \leq \|f - f_N\|_{L^r} + \|f_N - \hat{f}\|_{L^r} \lesssim N^{-\tilde{s}}.$$

Next, we verify that the network parameters of $\hat{f}$ are given as in (S7). Since the identity function $\mathrm{id}(x) = x$ can be represented using the ReLU function as $\mathrm{id}(x) = \zeta(x) - \zeta(-x)$, it follows that for any $L \in \mathbb{N}$,

$$\mathrm{id}^{\circ L} := \underbrace{\mathrm{id} \circ \cdots \circ \mathrm{id}}_{L \text{ times}} = \mathrm{id} \in \tilde{\Phi}(L, 2, 4L, 1).$$

Then, using $\bar{L}_0 = \bar{L}_0(N) = \lceil \log_2(d \vee m) \rceil \lceil \tilde{s} \log_2 N \rceil$, we can rewrite $\hat{f}$ as

$$\hat{f}(x) = \sum_{(k,j) \in E_N} \mathrm{id}^{\circ \bar{L}_0}(\alpha_{k,j} \hat{M}_{k,j}^{d,m}(x)), \tag{S12}$$

which is also an NN approximator. Since $\hat{M}_{k,j}^{d,m}$ requires the same depth, width, and sparsity as $\hat{M}$ in Lemma B.1, each summand $\mathrm{id}^{\circ \bar{L}_0}(\alpha_{k,j} \hat{M}_{k,j}^{d,m}(x))$ has depth $L_0 + \bar{L}_0$, width D_0, and sparsity $L_0 D_0^2 + 4\bar{L}_0 + 1$. Therefore, $\hat{f}$ expressed as in (S12) requires depth $L_0 + \bar{L}_0 \leq L_1$, width ND_0, and sparsity $(L_0 D_0^2 + 4\bar{L}_0 + 1)N \leq S_1$. To evaluate the magnitude of the network parameters, note from (S10) that the magnitudes of the parameters in the first layer of $\hat{M}_{k,j}^{d,m}$ are bounded by $2^{K^* \tilde{s}} \leq C_{11} N^{(1+\nu^{-1})/\underline{s}}$ for some constant $C_{11} > 0$. All remaining layers have the same magnitude as those of $\hat{M}$, namely B_0. In addition, by (S11), the magnitudes of the parameters in the first layer of $t \mapsto \mathrm{id}^{\circ \bar{L}_0}(\alpha_{k,j} t)$ are bounded by $C_{12} N^{(1+\nu^{-1})(1/p - \tilde{s})_+}$ for some constant $C_{12} > 0$. Let $B_{11}(N) = C_{11} N^{(1+\nu^{-1})/\underline{s}} B_0^{L_0}$ and $B_{12}(N) = C_{12} N^{(1+\nu^{-1})(1/p-\tilde{s})_+} B_0^{\bar{L}_0}$, and define

$$c_{1l} = \begin{cases} B_{11}(N)^{(L_0+1)^{-1}} / (C_{11} N^{(1+\nu^{-1})/\underline{s}}), & l = 1, \\ B_{11}(N)^{(L_0+1)^{-1}} / B_0 & l = 2, \ldots, L_0 + 1, \end{cases}$$

$$c_{2l} = \begin{cases} B_{12}(N)^{(\bar{L}_0+1)^{-1}} / (C_{12} N^{(1+\nu^{-1})(1/p-\tilde{s})_+}), & l = 1, \\ B_{12}(N)^{(\bar{L}_0+1)^{-1}} / B_0 & l = 2, \ldots, \bar{L}_0 + 1. \end{cases}$$

It is easy to verify that $\prod_{l'=1}^{l} c_{1l'} \leq 1$ and $\prod_{l'=1}^{l} c_{2l'} \leq 1$ for any l and sufficiently large N. Therefore, by Lemma B.7, the rescaled $\hat{h}$ has parameter magnitudes bounded by $B_{11}(N)^{(L_0+1)^{-1}} \vee B_{12}(N)^{(L_0+1)^{-1}}$. Since $L_0(d, m, N) \geq \bar{L}_0 = \lceil \log_2(d \vee m) \rceil \lceil \tilde{s} \log_2 N \rceil$ for sufficiently large N, we obtain

$$\begin{aligned} B_{11}(N)^{(L_0+1)^{-1}} \vee B_{12}(N)^{(\bar{L}_0+1)^{-1}} &\leq B_0 \left((C_{11} \vee C_{12}) N^{(1+\nu^{-1})[(1/\underline{s}) \vee (1/p-\tilde{s})_+]} \right)^{(\bar{L}_0+1)^{-1}} \\ &\leq B_0 N^{(1+\nu^{-1})[(1/\underline{s}) \vee (1/p-\tilde{s})_+] \bar{L}_0^{-1}} \\ &\leq B_0 \exp\left(\frac{(1+\nu^{-1})[(1/\underline{s}) \vee (1/p - \tilde{s})_+] \log 2}{\lceil \tilde{s} \log_2(d \vee m) \rceil} \right) \\ &= B_1, \end{aligned}$$

which concludes $\hat{f} \in \Phi(L_1, D_1, S_1, B_1)$. □

C.6 Proof of Lemma B.9

Proof. Our proof is similar to the proof of Theorem 1 in Suzuki and Nitanda [18]. Let $f_0 = f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,1}$. Since $\tilde{s}^{(h)} > 1/p$ for $h \geq 2$, by Lemma B.4, for each $f_{0,h,k}$ there exists

$$\begin{aligned} f_{h,k} \in \Phi(&L_1(t^{(h)}, m^{(h)}, p, \infty, s^{(h)}, N), D_1(t^{(h)}, m^{(h)}, N), \\ &S_1(t^{(h)}, m^{(h)}, p, \infty, s^{(h)}, N), B_1(t^{(h)}, m^{(h)}, p, \infty, s^{(h)})) \end{aligned}$$

such that $\|f_{0,h,k} - f_{h,k}\|_\infty \lesssim N^{-\tilde{s}^{(h)}}$. For $h = 1$, each $f_{0,1,k}$ there exists

$$\begin{aligned} f_{1,k} \in \Phi(&L_1(d^{(1)}, m^{(1)}, p, r, s^{(1)}, N), D_1(d^{(1)}, m^{(1)}, N), \\ &S_1(d^{(1)}, m^{(1)}, p, r, s^{(1)}, N), B_1(d^{(1)}, m^{(1)}, p, r, s^{(1)})) \end{aligned}$$

such that $\|f_{0,1,k} - f_{1,k}\|_{L^r} \lesssim N^{-\tilde{s}^{(1)}}$. To ensure that each block produces outputs in $[0,1]$, we apply the clipping function $\mathrm{clip}_{[0,1]}$, as described in Remark B.8, which adds one layer to each block. Let $f = f_H \circ f_{H-1} \cdots \circ f_1 \in \Phi(L_2, D_2, S_2, B_2)$ as in Schmidt-Hieber [8]. Then,

$$\begin{aligned}
&\|f_0 - f\|_{L^r} \\
&= \|f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,1} - f_H \circ f_{H-1} \cdots \circ f_1\|_{L^r} \\
&\le \sum_{h=1}^{H} \|f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h} \circ f_{h-1} \cdots f_1 - f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h+1} \circ f_h \cdots f_1\|_{L^r} \\
&\le \sum_{h=1}^{H} \|f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h} - f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h+1} \circ f_h\|_{L^{r_h}},
\end{aligned}$$

where $r_1 = r$ and $r_h = \infty$ for $h \ge 2$. By Remark B.2,

$$f_{0,h,k} \in C^{\underline{s}^{(h)}[1-1/(p\tilde{s}^{(h)})]\wedge 1} = C^{(\underline{s}^{(h)} - t^{*(h)}/p)\wedge 1} = C^{\gamma'_h}, \quad h \ge 2,$$

where $\gamma'_h := (\underline{s}^{(h)} - t^{*(h)}/p) \wedge 1$. Therefore,

$$\begin{aligned}
\|f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h} - f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,h+1} \circ f_h\|_\infty &\lesssim \|f_{0,h} - f_h\|_\infty^{\prod_{h'=h+1}^{H} \gamma'_{h'}} \\
&\lesssim N^{-\tilde{s}^{(h)} \prod_{h'=h+1}^{H} \gamma'_{h'}} \\
&= N^{-\tilde{s}^{*(h)}}
\end{aligned}$$

for $h \ge 2$. For $h = 1$,

$$\begin{aligned}
\|f_{0,H} \circ f_{0,H-1} \cdots \circ f_{0,1} - f_{0,H} \circ f_{0,H-1} \cdots \circ f_1\|_{L^r} &\lesssim \|f_{0,1} - f_1\|_{L^r}^{\prod_{h'=2}^{H} \gamma'_{h'}} \\
&\lesssim N^{-\tilde{s}^{*(1)}}.
\end{aligned}$$

Therefore, $\|f_0 - f\|_{L^r} \lesssim \max_{h\in[H]} N^{-\tilde{s}^{*(h)}} = N^{-\tilde{s}^*}$. □

C.7 Proof of Lemma B.13

Proof. For $f \in \Phi(L, D, S, B)$ expressed as $f(x) = (W^{(L)}\zeta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(2)}\zeta(\cdot) + b^{(2)}) \circ (W^{(1)}x + b^{(1)})$, define

$$\begin{aligned}
\mathcal{A}_k^+(f)(x) &= \zeta \circ (W^{(k-1)}\zeta(\cdot) + b^{(k-1)}) \circ \cdots \circ (W^{(2)}\zeta(\cdot) + b^{(2)}) \circ (W^{(1)}x + b^{(1)}), \\
\mathcal{A}_k^-(f)(x) &= (W^{(L)}\zeta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(k+1)}\zeta(\cdot) + b^{(k+1)}) \circ (W^{(k)}x + b^{(k)}),
\end{aligned}$$

for $k = 2, \cdots, L$, and let $\mathcal{A}_{L+1}^-(f)(x) = \mathcal{A}_1^+(f)(x) = x$. Then,

$$f(x) = \mathcal{A}_{k+1}^-(f) \circ (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f)(x).$$

Using the definition of $\Phi(L, D, S, B)$,

$$\begin{aligned}
\big\|\mathcal{A}_k^+(f)(x)\big\|_\infty &\le \max_j \Big\|W_{j,:}^{(k-1)}\Big\|_1 \big\|\mathcal{A}_{k-1}^+(f)(x)\big\|_\infty + \|b^{(k-1)}\|_\infty \\
&\le DB\big\|\mathcal{A}_{k-1}^+(f)(x)\big\|_\infty + B \\
&\le (D+1)(B \vee 1)\big\|\mathcal{A}_{k-1}^+(f)(x)\big\|_\infty \\
&\le (D+1)^{k-1}(B \vee 1)^{k-1},
\end{aligned}$$

where $A_{j,:}$ denotes the j-th row of matrix A. Similarly,

$$\big|\mathcal{A}_k^-(f)(x_1) - \mathcal{A}_k^-(f)(x_2)\big| \le (BD)^{L-k+1}\|x_1 - x_2\|_\infty.$$

Fix $\epsilon > 0$ and $\theta \in \Theta(L, D, S, B)$. For any $\theta^* \in \Theta(L, D, S, B)$ satisfying $\|\theta - \theta^*\|_\infty < \epsilon$, we obtain

$$
\begin{aligned}
&|f_\theta(x) - f_{\theta^*}(x)| \\
&= \left| \sum_{k=1}^{L} \mathcal{A}_{k+1}^{-}(f_{\theta^*}) \circ (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^{+}(f_\theta)(x) - \mathcal{A}_{k+1}^{-}(f_{\theta^*}) \circ (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^{+}(f_\theta)(x) \right| \\
&\leq \sum_{k=1}^{L} (BD)^{L-k} \left\| (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^{+}(f_\theta)(x) - (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^{+}(f_\theta)(x) \right\|_\infty \\
&\leq \sum_{k=1}^{L} (BD)^{L-k} \epsilon \left[D(B \vee 1)^{k-1} (D+1)^{k-1} + 1 \right] \\
&\leq \sum_{k=1}^{L} (BD)^{L-k} \epsilon (B \vee 1)^{k-1} (D+1)^{k} \\
&\leq \epsilon L (B \vee 1)^{L-1} (D+1)^{L}.
\end{aligned}
$$

This proves the assertion. □

C.8 Proof of Theorem 3.3

Proof. Let $\mathcal{F} = \Phi(L_{1n}, D_{1n}, S_{1n})$. By Lemma A.1, it suffices to show that there exists a subset $\mathcal{F}_n \subset \mathcal{F}$ such that

(a) $\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n}) \lesssim n\epsilon_n^2$

(c) $\sup_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = o\Big(e^{-Cn\epsilon_n^2}\Big)$

for a sufficiently large C. Let $\mathcal{F}_n = \Phi(L_{1n}, D_{1n}, S_{1n}, B_n)$, where $B_n = n$. To verify (c), observe that

$$\Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = \Pi(\mathcal{F} \setminus \mathcal{F}_n) = \Pi(\exists j : |\theta_j| > B_n \mid L_{1n}, D_{1n}, S_{1n}) = 1 - (1 - v_n)^{S_{1n}},$$

where $v_n = \int_{|u| > B_n} \tilde{\pi}_{SL}(u) du$. The direct calculation of the number of network parameters yields

$$T = |\Theta(L, D)| = D^2(L-1) + D(d + L + 1) + 1,$$

which implies

$$T_{1n} := |\Theta(L_{1n}, D_{1n})| \asymp L_{1n} D_{1n}^2. \tag{S13}$$

Moreover, $v_n \leq e^{-k_1 B_n}$ for some $k_1 > 0$ by Assumption (B1). Therefore, by Bernoulli's inequality, we obtain

$$1 - (1 - v_n)^{S_{1n}} \leq S_{1n} v_n \leq T_{1n} v_n \leq e^{-k_1 B_n + \log T_{1n}} = o\Big(e^{-Cn\epsilon_n^2}\Big),$$

for a sufficiently large C. By Lemma B.11,

$$
\begin{aligned}
\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) &\leq \log \mathcal{N}(\epsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \\
&\leq (S_{1n} + 1) \left[\log L_{1n} + L_{1n} \log\big((B_n \vee 1)(D_{1n} + 1)^2\big) - \log \frac{\epsilon_n}{2} \right] \\
&\lesssim N_n (\log n)^3 \\
&\lesssim n\epsilon_n^2
\end{aligned}
$$

for sufficiently large n. Thus, (a) holds. Given Lemma B.4, Remark B.6, and Assumption (A1), there exists $\hat{f}_n := f_{\hat{\theta}} \in \Phi(L_{1n}, D_{1n}, S_{1n}, B_1) \subset \mathcal{F}_n$ for $\hat{\theta} \in \Theta(L_{1n}, D_{1n}, S_{1n}, B_1)$ such that

$$\|\hat{f}_n - f_0\|_{L^2(P_X)} \leq C_1 N^{-\tilde{s}} \leq \epsilon_n / 4, \tag{S14}$$

for some $C_1 > 0$ and all sufficiently large n. Let $\hat{\gamma}$ denote the index set of nonzero components in $\hat{\theta}$, and let $\hat{\theta}_{\hat{\gamma}}$ denote the corresponding nonzero values. We define $\Theta(\hat{\gamma}; L_{1n}, D_{1n}, S_{1n}, B_1) \subset \Theta(L_{1n}, D_{1n}, S_{1n}, B_1)$ as the subset of parameter space in which only the components indexed by $\hat{\gamma}$ are nonzero. The corresponding NN space is denoted by

$$\tilde{\mathcal{F}}_n(\hat{\gamma}) = \Phi(\hat{\gamma}; L_{1n}, D_{1n}, S_{1n}, B_1).$$

Using Assumption (A3), we have

$$\log \Pi(|\sigma - \sigma_0| \leq \epsilon_n/2) \gtrsim \log \epsilon_n \gtrsim -\log n$$

and hence

$$\Pi(A_{\epsilon_n}) \gtrsim \Pi(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \leq \epsilon_n/2) - \log n.$$

Given (S14), there exists a constant $C_1 > 0$ such that

$$\begin{aligned}
\Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \leq \epsilon_n/2\Big) &\geq \Pi\Big(f \in \mathcal{F}_n : \|f - \hat{f}_n\|_{L^2(P_X)} \leq \epsilon_n/4\Big) \\
&\geq \Pi\Big(f \in \mathcal{F}_n : \|f - \hat{f}_n\|_{\infty} \leq \epsilon_n/4\Big) \\
&\geq \Pi\Big(f \in \tilde{\mathcal{F}}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_{\infty} \leq \epsilon_n/4\Big).
\end{aligned}$$

Using Lemma B.13,

$$\begin{aligned}
&\Pi\Big(f \in \tilde{\mathcal{F}}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_{\infty} \leq \epsilon_n/4\Big) \\
&\geq \Pi\left(\theta \in \mathbb{R}^{T_{1n}} : \theta_{\hat{\gamma}^c} = 0, \|\theta_{\hat{\gamma}}\|_{\infty} \leq B_1, \|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\|_{\infty} \leq \frac{\epsilon_n}{4(D_{1n}+1)^{L_{1n}} L_{1n} (B_1 \vee 1)^{L_{1n}-1}}\right) \\
&\geq \left(t_n \inf_{u \in [-B_1, B_1]} \tilde{\pi}_{SL}(u)\right)^{S_{1n}} \binom{T_{1n}}{S_{1n}}^{-1} \\
&ge\left(t_n (D_{1n}+1)^{-L_{1n}} \inf_{u \in [-B_1, B_1]} \tilde{\pi}_{SL}(u)\right)^{S_{1n}}
\end{aligned}$$

where $t_n = \epsilon_n / [2(D_{1n}+1)^{L_{1n}} L_{1n} (B_1 \vee 1)^{L_{1n}-1}]$. The last inequality follows from

$$\binom{T_{1n}}{S_{1n}} = \frac{T_{1n}(T_{1n}-1)\cdots(T_{1n}-S_{1n}+1)}{S_{1n}!} \leq T_{1n}^{S_{1n}} \leq (D_{1n}+1)^{L_{1n} S_{1n}}.$$

Using

$$-\log t_n \leq \log\left(\frac{(D_{1n}+1)^{L_{1n}} L_{1n} (B_1 \vee 1)^{L_{1n}-1}}{2\epsilon_n}\right) \lesssim (\log n)^2$$

and Assumption (B2), we obtain

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n}) &\lesssim -S_{1n} \log\left(t_n (D_{1n}+1)^{-L_{1n}} \inf_{u \in [-B_1, B_1]} \tilde{\pi}_{SL}(u)\right) + \log n \\
&\lesssim S_{1n} (\log n)^2 \\
&\lesssim N_n (\log n)^3 \\
&\lesssim n\epsilon_n^2,
\end{aligned}$$

which verifies (b). □

C.9 Verification of Remark 3.4

Note that Kong and Kim [19] established posterior consistency over unbounded parameter spaces. Their analysis allows for greater flexibility in prior specification by leveraging a complexity bound, and the result builds upon the earlier work of Kohler and Langer [23]. Motivated by their approach, we show that posterior concentration at the rate $n\epsilon_n^2 = N_n(\log n)^4$ remains valid (in contrast to the original rate $n\epsilon_n^2 = N_n(\log n)^3$) by employing the covering number bound stated below.

Lemma C.1 (Covering number for sparse unbounded NNs)**.** *For all $\epsilon > 0$, the following holds:*

$$\log \mathcal{N}(\epsilon, \Phi(L, D, S), \|\cdot\|_{L^p}) \leq C_{V_1} LS \log S \log\left(\frac{K}{\epsilon^p}\right) + \log\left(C_{V_2} LS \log S\right),$$

for some positive constants $C_{V_1} > 0,\ C_{V_2} > 0$ and $K > 0$.

Proof. By Theorem 7 of Bartlett et al. [33], the VC-dimension $V^+_{\Phi(L,D,S)}$ of the class $\Phi(L, D, S)$ satisfies $V^+_{\Phi(L,D,S)} \leq C_V LS \log S$ for some constant $C_V > 0$. Then, by Theorem 2.6.4 of van der Vaart and Wellner [53], it follows that

$$\log \mathcal{N}(\epsilon, \Phi(L, D, S), \|\cdot\|_{L^p}) \leq \log\Big(C' V^+_{\Phi(L,D,S)} (K/\epsilon^p)^{V^+_{\Phi(L,D,S)}}\Big),$$

for some positive constants $C' > 0$ and $K > 0$. Therefore, the stated result follows. □

Now we are ready to prove the assertion in Remark 3.4.

Proof. Let $\mathcal{F} = \Phi(L_{1n}, D_{1n}, S_{1n})$. By Lemma A.1, it suffices to construct a set $\mathcal{F}_n \subset \mathcal{F}$ such that the following conditions hold:

(a) $\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2$,

(b) $-\log \Pi(A_{\epsilon_n}) \lesssim n\epsilon_n^2$,

(c) $\sup_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = o\left(e^{-Cn\epsilon_n^2}\right)$,

for some constant $C > 0$. Let $\mathcal{F}_n = \Phi(L_{1n}, D_{1n}, S_{1n})$. Note that condition (c) is trivially satisfied. Using Lemma C.1 and Assumption (A2), we obtain

$$\begin{aligned}
\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) &\leq \log \mathcal{N}(\epsilon_n/R, \mathcal{F}_n, \|\cdot\|_{L^2}) \\
&\lesssim L_{1n} S_{1n} \log S_{1n} \log \epsilon_n^{-1} \\
&\lesssim N_n (\log n)^4 \\
&\lesssim n\epsilon_n^2,
\end{aligned}$$

where the last line follows since $L_{1n} \asymp \log n$ and $S_{1n} \asymp N_n \log n$. The remaining parts follow identically from the proof of Theorem 3.3. □

C.10 Verification of Example 3.5

Assumption (C2) holds trivially. To verify Assumption (C1), observe that

$$\int_{[-K_n, K_n]^c} \tilde{\pi}_{SH}(u)\, du \leq \pi_{1n} \Pr(|Z| > K_n/\sigma_{1n}) \leq C_1 \exp\left(-C_2 K_n^{1/k}/\sigma_{1n}^{1/k}\right),$$

for all sufficiently large n, where Z is a random variable having density φ_k. Since σ_{1n} is decreasing,

$$\log\left(\int_{[-K_n, K_n]^c} \tilde{\pi}_{SH}(u)\, du\right) \lesssim -K_n^{1/k}/\sigma_{1n}^{1/k} \lesssim -K_n^{1/k} \lesssim -K_n,$$

for $k \leq 1$. This verifies Assumption (C1). Next, observe that

$$\begin{aligned}
\int_{[-a_n, a_n]^c} \tilde{\pi}_{SH}(u)\, du &\leq \pi_{2n}\left(1 - \frac{a_n}{C_u}\right) + \pi_{1n} \Pr(|Z| > (2C_A(\log n)^2)^k) \\
&\leq 2e^{-2C_A(\log n)^2} \\
&\leq e^{-C_A(\log n)^2}.
\end{aligned}$$

Hence, Assumption (C3) is satisfied.

C.11 Verification of Example 3.6

Assumptions (C1) and (C3) can be verified in the same manner as in Example 3.5. It therefore suffices to verify Assumption (C2). Note that $\tilde{\pi}_{SH}(u) \geq \pi_{2n}\sigma_2^{-1}\phi(u/\sigma_2)$, where ϕ denotes the density of the standard normal distribution and σ_2 is a positive standardized deviation. Therefore,

$$-\inf_{u\in[-B_1,B_1]} \log \tilde{\pi}_{SH}(u) \leq -\log \pi_{2n} + \frac{1}{2}\log(2\pi\sigma_2^2) + \frac{B_1^2}{2\sigma_2^2} \lesssim (\log n)^2.$$

C.12 Proof of Theorem 3.8

Proof. Let $\mathcal{F} = \Phi(L_{1n}, D_{1n})$. By Lemma A.1, it suffices to show that there exists a subset $\mathcal{F}_n \subset \mathcal{F}$ such that

(a) $\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n}) \lesssim n\epsilon_n^2$

(c) $\sup_{\sigma\in[\underline{\sigma},\overline{\sigma}]} \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = o\Big(e^{-C'n\epsilon_n^2}\Big)$

for a sufficiently large C'. Let $\mathcal{F}_n = \Phi(L_{1n}, D_{1n}, S_{1n}, B_n, a_n)$ as defined in (S8), where $B_n = n$. Since $B_n^{-(L_{1n}-1)} \geq \exp(-L_{1n}\log n)$ and $\epsilon_n/[2L_{1n}(D_{1n}+1)^{L_{1n}}] \geq \exp(-L_{1n}\log n)$, we obtain

$$\epsilon_n/[2L_{1n}(B_n \vee 1)^{L_{1n}-1}(D_{1n}+1)^{L_{1n}}] \geq \exp(-2L_{1n}\log n) = a_n \tag{S15}$$

for all sufficiently large n. Hence, it is easy to show that (a) holds using Lemma B.12 as in the proof of Theorem 3.3. Let $v_n = \int_{|u|>B_n} \tilde{\pi}_{SH}(u)\,du$ and $u_n = \int_{|u|>a_n} \tilde{\pi}_{SH}(u)du$. Then,

$$\begin{aligned}
\Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) &= \Pi(\mathcal{F} \setminus \mathcal{F}_n) \\
&\leq \Pi(\exists j : |\theta_j| > B_n \mid L_{1n}, D_{1n}) + \Pi\left(\sum_{j=1}^{T_{1n}} I(|\theta_j| > a_n) > S_{1n} \mid L_{1n}, D_{1n}\right) \\
&= (1-(1-v_n)^{T_{1n}}) + \Pr(S > S_{1n}),
\end{aligned}$$

where $S \sim B(T_{1n}, u_n)$ is a binomially distributed random variable. Since $v_n \leq e^{-k_1 B_n}$ for some $k_1 > 0$ by Assumption (C1), we have

$$1-(1-v_n)^{T_{1n}} \leq T_{1n}v_n \leq e^{-k_1 B_n + \log T_{1n}} = o\Big(e^{-K_2 n\epsilon_n^2}\Big)$$

for a sufficiently large $K_2 > 0$. Let $r_n = S_{1n}/(T_{1n}u_n)$. Using the multiplicative Chernoff bound of binomial distributions,

$$\Pr(S > S_{1n}) \leq \exp(-T_{1n}u_n[r_n \log r_n - (r_n - 1)]) \leq \exp(-C_1 S_{1n}\log r_n)$$

for some $C_1 > 0$. By (S13) together with Assumption (C3), we have $\log r_n \geq K_3(\log n)^2$ for a sufficiently large constant $K_3 > 0$. Therefore,

$$\Pr(S > S_{1n}) \leq \exp(-K_3 n\epsilon_n^2).$$

Next, as in the proof of Theorem 3.3, there is a constant $C_1 > 0$ and $f_{\hat{\theta}} \in \Phi(L_{1n}, D_{1n}, S_{1n}, B_1)$ for $\hat{\theta} \in \Theta(L_{1n}, D_{1n}, S_{1n}, B_1)$ such that

$$\big\|f_{\hat{\theta}} - f_0\big\|_{L^2(P_X)} \leq C_1 N_n^{-\tilde{s}} \leq \epsilon_n/8$$

for all sufficiently large n. Let $\hat{\gamma}$ denote the index set of nonzero components in $\hat{\theta}$. Define the subset of the parameter space $\Theta(\hat{\gamma}; L_{1n}, D_{1n}, S_{1n}, B_1, a_n) \subset \Theta(L_{1n}, D_{1n}, S_{1n}, B_1, a_n)$ whose magnitudes exceed a_n only at $\hat{\gamma}$, and let $\tilde{\mathcal{F}}_n(\hat{\gamma}) = \Phi(\hat{\gamma}; L_{1n}, D_{1n}, S_{1n}, B_1, a_n)$ be the corresponding NN space. Then, by (S15) and Lemma B.13, there exists $\hat{\theta}(\hat{\gamma}) \in \Theta(\hat{\gamma}, L_{1n}, D_{1n}, S_{1n}, B_1, a_n)$ such that $\|f_{\hat{\theta}} - f_{\hat{\theta}(\hat{\gamma})}\|_{L^2(P_X)} \leq \epsilon_n/8$. Let $\hat{f}_n = f_{\hat{\theta}(\hat{\gamma})}$. For all sufficiently large n, we obtain

$$\begin{aligned}
\Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \leq \epsilon_n/2\Big) &\geq \Pi\Big(f \in \mathcal{F}_n : \|f - \hat{f}_n\|_{L^2(P_X)} \leq \epsilon_n/4\Big) \\
&\geq \Pi\Big(f \in \tilde{\mathcal{F}}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_{\infty} \leq \epsilon_n/4\Big).
\end{aligned}$$

By Lemma B.13,

$$\begin{aligned}
&\Pi\Big(f \in \tilde{\mathcal{F}}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_\infty \le \epsilon_n/4\Big) \\
&\ge \Pi\left(\theta_{\hat{\gamma}^c} \in [-a_n, a_n]^{T_{1n}-S_{1n}}, \|\theta_{\hat{\gamma}}\|_\infty \le B_1, \|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\|_\infty \le \frac{\epsilon_n/4}{(D_{1n}+1)^{L_{1n}} L_{1n}(B_1 \vee 1)^{L_{1n}-1}}\right) \\
&\ge (1-u_n)^{T_{1n}-S_{1n}} \left(t_n \inf_{u \in [-B_1, B_1]} \tilde{\pi}_{SH}(u)\right)^{S_{1n}},
\end{aligned}$$

where $t_n = \epsilon_n/[2(D_{1n}+1)^{L_{1n}} L_{1n}(B_1 \vee 1)^{L_{1n}-1}]$. Therefore,

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n}) &\lesssim -S_{1n} \log\left(t_n \inf_{u \in [-B_1, B_1]} \tilde{\pi}_{SH}(u)\right) \\
&\quad - (T_{1n} - S_{1n}) \log(1-u_n) + \log n \\
&\lesssim S_{1n}(\log n)^2 - T_{1n} \log(1 - S_{1n}/T_{1n}) + \log n \\
&= S_{1n}(\log n)^2 + T_{1n}\left(S_{1n}/T_{1n} + o\left(S_{1n}/T_{1n}\right)\right) + \log n \\
&\lesssim S_{1n}(\log n)^2 + S_{1n} + o\left(S_{1n}\right) + \log n \\
&\lesssim n\epsilon_n^2.
\end{aligned} \tag{S16}$$

For the second inequality, we used the fact that

$$-\log(1-u_n) \le -\log\left(1 - e^{-C_A(\log n)^2}\right) \le -\log\left(1 - S_{1n}/T_{1n}\right)$$

for sufficiently large n, as implied by Assumption (C3). □

C.13 Proof of Theorem 3.9

Proof. Let $\mathcal{F} = \Phi(\tilde{L}_n)$. By Lemma A.1, it suffices to show that there exists a subset $\mathcal{F}_n \subset \mathcal{F}$ such that

(a) $\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) \lesssim n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n}) \lesssim n\epsilon_n^2$

(c) $\sup_{\sigma \in [\underline{\sigma}, \overline{\sigma}]} \Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) = o\Big(e^{-C'' n\epsilon_n^2}\Big)$

for a sufficiently large C''. We first prove (i), the case corresponding to the spike-and-slab prior. Define $\tilde{D}_n = C_D N_n$ and $\tilde{S}_n = C_S N_n \log n$ for sufficiently large $C_D, C_S > 0$. Define the sieve

$$\mathcal{F}_n = \bigcup_{D \le \tilde{D}_n} \bigcup_{S \le \tilde{S}_n} \Phi(\tilde{L}_n, D, S, B_n),$$

where $B_n = n$. To verify the entropy bound, apply Lemma B.11 to obtain

$$\begin{aligned}
\mathcal{N}(\epsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) &\le \sum_{D \le \tilde{D}_n} \sum_{S \le \tilde{S}_n} \left(\frac{2}{\epsilon_n} \tilde{L}_n (B_n \vee 1)^{\tilde{L}_n} (D+1)^{2\tilde{L}_n}\right)^{S+1} \\
&\le \tilde{D}_n \tilde{S}_n \left(\frac{2}{\epsilon_n} \tilde{L}_n (B_n \vee 1)^{\tilde{L}_n} (\tilde{D}_n+1)^{2\tilde{L}_n}\right)^{\tilde{S}_n+1}.
\end{aligned} \tag{S17}$$

Hence, (a) holds as in the proof of Theorem 3.3. To verify (b), observe that there exists a constant $C''' > 0$ such that

$$\min\{\pi_D(\tilde{D}_n), \pi_S(\tilde{S}_n)\} \gtrsim \exp\big(-C''' n\epsilon_n^2\big).$$

Since $D_{1n} \le \tilde{D}_n$ and $S_{1n} \le \tilde{S}_n$, by Lemma B.4 and Remark B.6, there exists $\hat{f}_n = f_{\hat{\theta}} \in \Phi(\tilde{L}_n, \tilde{D}_n, \tilde{S}_n, B_n)$ satisfying (S14). Following the proof of Theorem 3.3, it can be shown that

$$-\log \Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2 \mid \tilde{D}_n, \tilde{S}_n\Big) \lesssim n\epsilon_n^2.$$

Therefore,

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n}) &\lesssim -\log \Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2\Big) - \log n \\
&\le -\log \pi_D(\tilde{D}_n) - \log \pi_S(\tilde{S}_n) \\
&\quad - \log \Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2 \mid \tilde{D}_n, \tilde{S}_n\Big) - \log n \\
&\lesssim n\epsilon_n^2,
\end{aligned} \tag{S18}$$

which verifies (b). Now, observe that

$$\begin{aligned}
\Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) &= \Pi(\mathcal{F} \setminus \mathcal{F}_n) \\
&\le \Pi(D > \tilde{D}_n) + \Pi(S > \tilde{S}_n) \\
&\quad + \sum_{D \le \tilde{D}_n} \sum_{S \le \tilde{S}_n} \pi_D(D)\pi_S(S)\Pi\Big(\exists j : |\theta_j| > B_n \mid \tilde{L}_n, D, S\Big).
\end{aligned}$$

First, we bound the tail probability of D as

$$\Pi(D > \tilde{D}_n) \lesssim \sum_{D > \tilde{D}_n} e^{-\lambda_D D(\log D)^3} \le \sum_{D > \tilde{D}_n} e^{-\lambda_D D(\log \tilde{D}_n)^3} \lesssim e^{-\lambda_D \tilde{D}_n (\log \tilde{D}_n)^3} \le e^{-C_1 n\epsilon_n^2},$$

for a sufficiently large $C_1 > 0$, provided that C_D is large enough. Similarly, we obtain $\Pi(S > \tilde{S}_n) \lesssim e^{-C_2 n\epsilon_n^2}$ for a sufficiently large $C_2 > 0$, provided that C_S is sufficiently large. Lastly, following the argument used in the proof of Theorem 3.3,

$$\begin{aligned}
&\sum_{D \le \tilde{D}_n} \sum_{S \le \tilde{S}_n} \pi_D(D)\pi_S(S)\Pi\Big(\exists j : |\theta_j| > B_n \mid \tilde{L}_n, D, S\Big) \\
&= \sum_{D \le \tilde{D}_n} \sum_{S \le \tilde{S}_n} \pi_D(D)\pi_S(S)[1 - (1 - v_n)^S] \\
&\le 1 - (1 - v_n)^{\tilde{S}_n} \\
&= o\Big(e^{-C_3 n\epsilon_n^2}\Big),
\end{aligned}$$

for a sufficiently large $C_3 > 0$. Therefore, (c) holds.

To prove (ii), define the sieve

$$\mathcal{F}_n = \bigcup_{D \le \tilde{D}_n} \Phi(\tilde{L}_{1n}, D, S_{1n}, B_n, a_n).$$

Using Lemma B.11, (a) is verified as in (S17). Moreover, following (S16) and (S18), we obtain

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n}) &\le -\log \pi_D(\tilde{D}_n) - \log \Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2 \mid \tilde{L}_n, \tilde{D}_n\Big) - \log n \\
&\lesssim n\epsilon_n^2,
\end{aligned}$$

which verifies (b). Lastly, we obtain

$$\begin{aligned}
&\Pi(\mathcal{F} \setminus \mathcal{F}_n \mid \sigma) \\
&= \Pi(\mathcal{F} \setminus \mathcal{F}_n) \\
&\le \Pi(D > \tilde{D}_n) \\
&\quad + \sum_{D \le \tilde{D}_n} \pi_D(D) \left[\Pi\Big(\exists j : |\theta_j| > B_n \mid \tilde{L}_{1n}, D\Big) + \Pi\left(\sum_{j=1}^{T_{1n}} I(|\theta_j| > a_n) > S_{1n} \mid \tilde{L}_{1n}, D\right)\right].
\end{aligned}$$

Following the calculation in the proof of Theorem 3.8, we verify (c). □

C.14 Proof of Theorem 3.14

Proof. We only prove the part corresponding to Theorem 3.3 under Assumption (A4). Other arguments proceed similarly. Let $\mathcal{F} = \Phi(L_{2n}, D_{2n}, S_{2n})$ and $\mathcal{F}_n = \Phi(L_{2n}, D_{2n}, S_{2n}, B_2)$. We verify conditions (a) and (b) of Lemma A.1. Using Lemma B.11,

$$\begin{aligned}
\log \mathcal{N}\Big(\epsilon_n, \mathcal{F}_n, \|\cdot\|_{L^2(P_X)}\Big) &\le \log \mathcal{N}(\epsilon_n, \mathcal{F}_n, \|\cdot\|_\infty) \\
&\le (S_{2n}+1)\left[\log L_{2n} + L_{2n}\log\big((B_2 \vee 1)(D_{2n}+1)^2\big) - \log\frac{\epsilon_n}{2}\right] \\
&\lesssim N_n (\log n)^3 \\
&\lesssim n\epsilon_n^2
\end{aligned}$$

for sufficiently large n. The last inequalities hold because

$$L_{2n} \asymp \log n, \quad D_{2n} \asymp N_n, \quad S_{2n} \asymp N_n \log n$$

for fixed $d^{(h)}$. Thus, (a) holds. Using Lemma B.9 and Assumption (A1), there exists $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}_n$ such that

$$\|\hat{f}_n - f_0\|_{L^2(P_X)} \le C_1 N_n^{-\tilde{s}^*} \le \epsilon_n/4$$

for some $C_1 > 0$ and all sufficiently large n. Let $\hat{\gamma}$ denote the index set of nonzero components in $\hat{\theta}$, and let $\hat{\theta}_{\hat{\gamma}}$ denote the corresponding nonzero values. We define $\Theta(\hat{\gamma}; L_{2n}, D_{2n}, S_{2n}, B_2) \subset \Theta(L_{2n}, D_{2n}, S_{2n}, B_2)$ as the subset of parameter space in which only the components indexed by $\hat{\gamma}$ are nonzero. The corresponding NN space is denoted by

$$\mathcal{F}_n(\hat{\gamma}) = \Phi(\hat{\gamma}; L_{2n}, D_{2n}, S_{2n}, B_2).$$

Using Assumption (A3), we have

$$\log \Pi(|\sigma - \sigma_0| \le \epsilon_n/2) \gtrsim \log \epsilon_n \gtrsim -\log n$$

and hence

$$\Pi(A_{\epsilon_n}) \gtrsim \Pi(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2) - \log n.$$

Using the earlier result,

$$\Pi\Big(f \in \mathcal{F}_n : \|f - f_0\|_{L^2(P_X)} \le \epsilon_n/2\Big) \ge \Pi\Big(f \in \mathcal{F}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_\infty \le \epsilon_n/4\Big).$$

Using Lemma B.13,

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_\infty \le \epsilon_n/4\right) \\
&\ge \Pi\left(\theta \in \mathbb{R}^{T_{2n}} : \theta_{\hat{\gamma}^c} = 0, \|\theta_{\hat{\gamma}}\|_\infty \le B_2, \|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\|_\infty \le \frac{\epsilon_n}{4(D_{2n}+1)^{L_{2n}} L_{2n}(B_2 \vee 1)^{L_{2n}-1}}\right) \\
&\ge \left(t_{2n} \inf_{u \in [-B_2, B_2]} \tilde{\pi}_{SL}(u)\right)^{S_{2n}} \binom{T_{2n}}{S_{2n}}^{-1} \\
&\ge \left(t_{2n}(D_{2n}+1)^{-L_{2n}} \inf_{u \in [-B_2, B_2]} \tilde{\pi}_{SL}(u)\right)^{S_{2n}},
\end{aligned}$$

where $t_{2n} = \epsilon_n / [2(D_{2n}+1)^{L_{2n}} L_{2n}(B_2 \vee 1)^{L_{2n}-1}]$. Therefore,

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n}) &\lesssim -S_{2n} \log\left(t_{2n}(D_{2n}+1)^{-L_{2n}} \inf_{u \in [-B_2, B_2]} \tilde{\pi}_{SL}(u)\right) + \log n \\
&\le S_{2n}(\log n)^2 \\
&\lesssim n\epsilon_n^2,
\end{aligned}$$

which verifies (b). □