Nonparametric Estimation for High-Dimensional Space Models Based on a Deep Neural Network

Hongxia Wang, Xiao Jin, Jianian Wang and Hongxia Hao *

School of Statistics and Data Science, Nanjing Audit University, Nanjing 211815, China; hzwang@nau.edu.cn (H.W.); mg2108123@stu.nau.edu.cn (X.J.); mg2008118@stu.nau.edu.cn (J.W.)
* Correspondence: hshao@nau.edu.cn

Abstract: With high dimensionality and dependence in spatial data, traditional parametric methods suffer from the curse of dimensionality problem. The theoretical properties of deep neural network estimation methods for high-dimensional spatial models with dependence and heterogeneity have been investigated only in a few studies. In this paper, we propose a deep neural network with a ReLU activation function to estimate unknown trend components, considering both spatial dependence and heterogeneity. We prove the compatibility of the estimated components under spatial dependence conditions and provide an upper bound for the mean squared error (MSE). Simulations and empirical studies demonstrate that the convergence speed of neural network methods is significantly better than that of local linear methods.

Keywords: deep neural network; spatial dependence; spatial heterogeneity; ReLU activation function

MSC: 62G05; 68T07

1. Introduction

Spatial data arise in many fields, including environmental science, econometrics, epidemiology, image analysis, oceanography, geography, geology, plant ecology, archaeology, agriculture, and psychology. Spatial correlation and spatial heterogeneity are two significant features of spatial data. Various spatial modeling methods have been applied to explore the effect of spatial heterogeneity. Notably, numerous local spatial techniques have been proposed to accommodate spatial heterogeneity. For example, Hallin et al. [1] and Biau and Cadre [2] proposed a local linear method for the modeling of spatial heterogeneity. Bentsen et al. [3] used a graph neural network architecture to extract spatial dependencies with different update functions to learn temporal correlations.

Spatial data often exhibit high dimensionality, a large scale, heterogeneity, and strong complexity. These challenges often make traditional statistical methods ineffective. Statistical machine learning methods can effectively address such challenges. Du et al. [4] pointed out the issues of traditional spatial data being large-scale and generally complex and summarized the effectiveness and application potential of four advanced machine learning methods—support vector machine (SVM)-based kernel learning, semi-supervised and active learning, ensemble learning, and deep learning—in handling complex spatial data. Farrell et al. [5] highlighted the challenges that high-dimensional spatial data, large volumes of data, and multicollinearity among covariates pose to traditional statistical models in variable selection. Three machine learning algorithms—maximum entropy (MaxEnt), random forests (RFs), and support vector machines (SVMs)—were employed to mitigate the issues of multicollinearity in high-dimensional spatial data. Nikparvar et al. [6] pointed out that the properties of spatially explicit data are often ignored or inadequately addressed in machine learning applications within spatial domains. They argued that the future prospects for spatial machine learning are very promising.
Statistical machine learning methods have advanced rapidly, while theoretical ones are not well established. Schmidt-Hieber [7] investigated the following nonparametric model:

$$Y_i = f_0(X_i) + \epsilon_i, i = 1, \ldots, n,$$

where the noise variables $\epsilon_i$ are assumed to be i.i.d., $X_i \in [0,1]^d, i = 1, \ldots, n$, are independently and identically distributed, $Y_i \in \mathbb{R}, i = 1, \ldots, n$ and are independently and identically distributed. It was shown that estimators based on sparsely connected deep neural networks with the ReLU activation function and a properly chosen network architecture achieve minimax rates of convergence (up to log n-factors) under a general composition assumption on the regression function.

Considering the dependencies and heterogeneity of spatial models, we study nonparametric high-dimensional spatial models as follows:

$$Y_i = m(X_i) + R_i,$$

where $i \in \Lambda_n = \{(i_1, i_2, \ldots, i_N): i_1 = 1, 2, \ldots, n_j, j = 1, 2, \ldots N\}, Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d, m(X_i)$ represents the trend function, $R_i$ satisfies the $\alpha$-mixing condition (the definition of the $\alpha$-mixing condition can be found at the beginning of Section 2.2), $n = n_1 + \ldots + n_N$. For example, $Y$ denotes the hourly ozone concentrations; $X$ is a $5 \times 1$ vector that consists of the following explanatory variables observed at each station: wind speed, air pressure, air temperature, relative humidity, and elevation. The observation locations are recorded in longitude $i_1$ and latitude $i_2$. In this case, $d = 5$ and $N = 2$; see [8].

Under general assumptions, we prove the consistency of the estimator and provide bounds for the mean squared error $MSE$. In the simulation aspect, a comparison with the local linear regression method demonstrates that the neural network method converges much faster than does the local linear regression. In the empirical study, considering the air pollution index, air pollutants, and environmental factors, the effectiveness of the neural network is demonstrated through a comparison with the local linear regression method. In this case, $d = 5$ and $N = 2$; see [8].

Throughout the rest of the paper, bold letters are used to represent vectors; for example, $x := (x_1, \ldots, x_d)^T$. We define $|x|_\infty := \max_i |x_i|$, $|x|_0 := \sum I(x_i \neq 0)$, where $I$ represents the indicator function:

$$I_A(w) = \begin{cases} 0 & \text{if } w \notin A, \\ 1 & \text{if } w \in A, \end{cases}$$

$|x|_0$ denotes the total number of $x_i$, which is not equal to zero. $|x|_p := \left( \sum |x_i|^p \right)^{1/p}$, and we write $|f|_p := |f|_{L^p(D)}$ as the $L^p$ norm on $D$; $D$ is some domain, and different situations may be different. For two sequences $(a_n)_n$ and $(b_n)_n$, we write $a_n \lesssim_R b_n$ if there exists a constant $C$ such that for all $n, a_n \leq Cb_n$. Moreover, $a_n \asymp_R b_n$ means $a_n \lesssim_R b_n$ and $b_n \lesssim_R a_n$. $\log_2$ denotes the logarithm base 2, $\ln$ denotes the logarithm base $e$, $[x]$ represents the largest integer $\leq x$. $\lfloor x \rfloor$ represents the largest integer $\leq x$.

2. Nonparametric High-Dimensional Space Model Estimation

2.1. Mathematical Modeling of Deep Network Features

**Definition 1.** Fitting a multilayer neural network requires the choice of an activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ and the network architecture. Motivated by its importance in deep learning, we study the rectifier linear unit (ReLU) activation function; see [7].

$$\sigma(x) = \max(x, 0).$$

For $v = (v_1, \ldots, v_r) \in \mathbb{R}^r$, the displacement activation function $\sigma_v : \mathbb{R}^r \rightarrow \mathbb{R}^r$ is defined as follows:
\( \sigma_v : \mathbb{R}^r \to \mathbb{R}^r, \)
\[
\sigma_v \begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \begin{pmatrix} \sigma(y_1 - v_1) \\ \vdots \\ \sigma(y_r - v_r) \end{pmatrix}.
\]

The neural network architecture \((L, p)\) consists of a positive integer \(L\) known as the number of hidden layers or depth and a width vector \(p = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}\). A neural network with the network structure \((L, p)\) is any function of the following form:

\[
f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad x \mapsto f(x) = W_L \sigma_{v_L} W_{L-1} \sigma_{v_{L-1}} \cdots W_1 \sigma_{v_1} W_0 x,
\]

where \(W_i\) is a \(p_{i+1} \times p_i\) weight matrix and \(v_i \in \mathbb{R}^{p_i}\) is a displacement vector, where \(i = 1, 2, \ldots, L\). Therefore, the network function is constructed by alternating matrix vector multiplications and the action of nonlinear activation functions \(\sigma_v\). In Equation (3), the shift vectors can also be omitted by considering the input as \((x, 1)\) and augmenting the weight matrices with an additional row and column. To fit the network to data generated by a \(d\)-dimensional nonparametric regression model, it is required to have \(p_0 = d\) and \(p_{L+1} = 1\).

Given a network function \(f(x) = W_L \sigma_{v_L} W_{L-1} \sigma_{v_{L-1}} \cdots W_1 \sigma_{v_1} W_0 x\), the network parameters are the elements of the matrices \((W_i)_{i=0,\ldots,L}\) and the vectors \((v_i)_{i=1,\ldots,L}\). These parameters need to be estimated/learned from the data. In this context, “estimate” and “learn” can be used interchangeably, as the process of estimating the parameters from data is often referred to as learning in the context of neural networks and machine learning.

The purpose of this paper is to consider a framework that encompasses the fundamental characteristics of modern deep network architectures. In particular, in this paper, we allow for a large depth \(L\) and a significant number of potential network parameters without requiring an upper bound on the number of network parameters for the main results. Consequently, this approach deals with high-dimensional settings that have more parameters than training data. Another characteristic of trained networks is that the learned parameters are typically not very large; see [7]. In practice, the weights of trained networks often do not differ significantly from the initialized weights. As all elements in orthogonal matrices are bounded by 1, the weights of trained networks also do not become excessively large. However, existing theoretical results often demand that the size of the network parameters tends to infinity. To be more consistent with what is observed in practice, all parameters considered in this paper are bounded by 1. By projecting the network parameters at each iteration onto the interval \([-1, 1]\), this constraint can be easily incorporated into deep learning algorithms.

Let \(\|W_j\|_{\infty}\) denote the maximum element norm of \(W_j\), and let us consider the network function space with a given network structure and network parameters bounded by 1 as follows:

\[
\mathcal{F}(L, p) := \left\{ f \text{ in the form of } (2.1) : \max_{j=0,\ldots,L} \|W_j\|_{\infty} \vee \|v_j\|_{\infty} \leq 1 \right\},
\]

where \(v_0\) is a vector with all components being 0.

In this work, we model the network sparsity assuming that there are only a few nonzero/active network parameters. If \(\|W_j\|_0\) denotes the number of nonzero entries of \(W_j\) and \(\|f\|_\infty\) stands for the supnorm of the function \(x \mapsto |f(x)|_\infty\), then the \(s\)-sparse networks are given by

\[
\mathcal{F}(L, p, s) := \mathcal{F}(L, p, s, F) := \left\{ f \in \mathcal{F}(L, p) : \sum_{j=0}^{L} \|W_j\|_0 + \|v_j\|_0 \leq s, \|f\|_\infty \leq F \right\},
\]

where \(F\) is a constant; the upper bound on the uniform norm of the function \(f\) is often unnecessary and is thus omitted in the notation. Here, we consider cases where the number
of network parameters $s$ is very small compared to the total number of parameters in the network.

For any estimate $\hat{m}_n$ that returns a network in class $\mathcal{F}(L, p, s, F)$, the corresponding quantity is defined as follows:

$$\Delta_n(\hat{m}_n, m) := E_n \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - \hat{m}_n(X_i))^2 - \inf_{f \in \mathcal{F}(L, p, s, F)} \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2 \right].$$  \hspace{1cm} (5)

The sequence $\Delta_n(\hat{m}_n, m)$ measures the discrepancy between the expected empirical risk of $\hat{m}_n$ and the global minimum of this class over all networks. The subscript $m$ in $E_n$ indicates the sample expectation with respect to the nonparametric regression model generated by the regression function $m$. Notice that $\Delta_n(\hat{m}_n, m) \geq 0$, and $\Delta_n(\hat{m}_n, m) = 0$ if $\hat{m}_n$ is an empirical risk minimizer.

Therefore, $\Delta_n(\hat{m}_n, m)$ is a critical quantity that, together with the minimax estimation rates, determines the convergence rate of $\hat{m}_n$.

To evaluate the statistical performance of $\hat{m}_n$ under general assumptions, the mean squared error of the estimator is defined as

$$R(\hat{m}_n, m) := E_n \left[ (\hat{m}_n(X) - m(X))^2 \right].$$  \hspace{1cm} (6)

### 2.2. Estimation and Theoretical Properties

In order to obtain asymptotic results, we will assume throughout this paper that $X_i$, $i \in \Lambda_n$ satisfies the following $\alpha$-mixing condition: there exists a function $\varphi(t) \downarrow$ as $t \to \infty$ with $\varphi(0) = 1$, such that whenever $\Xi, (\tilde{\Xi}) \subseteq \Lambda_n$ are finite sets, it is the case that

$$\tau(\mathcal{B}(\Xi), \mathcal{B}(\tilde{\Xi})) = \sup\{|Pr(AB) - Pr(B)| A \in \mathcal{B}(\Xi), B \in \mathcal{B}(\tilde{\Xi})\} \leq \varphi(\text{card}(\Xi), (\tilde{\Xi})) \varphi(d(\Xi, \tilde{\Xi})),$$

where $\mathcal{B}(\Xi)$ denotes the Borel $\sigma$-field generated by $\{X_i, i \in \Xi\}$, card$(\Xi)$ is the cardinality of $\Xi$, and $d(\Xi, (\tilde{\Xi})) = \min\{|i - i'| : i \in \Xi, i' \in (\tilde{\Xi})\}$ is the distance between $\Xi$ and $(\tilde{\Xi})$, where $|i| = (i_2^2 + \cdots + i_d^2)^{1/2}$ stands for the Euclidean norm and $\varphi : \mathbb{N}^2 \to \mathbb{R}^+$ is a symmetric positive function that is nondecreasing in each variable; see [8].

The theoretical performance of neural networks depends on the underlying function class, and a classic approach in nonparametric statistics is to assume that the regression function is $\beta$-smooth. In this paper, we assume that the regression function $m(X_i)$ is a composition of multiple functions, i.e.,

$$m = g_{\beta} \circ g_{\beta-1} \circ \cdots \circ g_1 \circ g_0,$$

where $g_i : [a, b]^d_i \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$. We denote the components of $g_i$, as $g_{ij} = (g_{ij})^T_{j=1, \ldots, d_{i+1}}$, and we let $t_i$ be the maximum variable that each $g_i$ depends on. Thus, each $g_i$ is a function with $t_i$ variables.

If all partial derivatives up to order $|\beta|$ of a function exist and are bounded, the $|\beta|$-th order partial derivatives are $\beta^{|\beta|}$ Hölder, where $|\beta|$ represents the largest integer strictly less than $\beta$. Then, the ball of $\beta$-Hölder functions with radius $K$ is defined as follows:

$$\mathcal{G}_K^\beta(D, K) = \left\{ f : D \subset \mathbb{R}^d \to \mathbb{R} : \sum_{\alpha : |\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha : |\alpha| = |\beta|} \sup_{x \neq y \in D} \frac{\|\partial^\alpha f(x) - \partial^\alpha f(y)\|_\infty}{\|x - y\|_{|\beta|}} \leq K \right\},$$

where $\partial^\alpha f$ denotes the total $|\alpha|$-th order partial derivative of $f$.
where we use multi-index notation, i.e., $\partial^\alpha = \partial^{\alpha_1} \ldots \partial^{\alpha_r}$, where $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$; see [7].

We assume that each function $g_{ij}$ has Hölder smoothness $\beta_i$. Since $g_{ij}$ is also a function of $t_i$ variables, $g_{ij} \in \mathcal{C}^\beta_i\left([a_i, b_i]^t_i, K_i\right)$, the underlying function space is then defined as

$$
\mathcal{G}(q, d, t, \beta, K) := \left\{ m = g_q \circ \ldots \circ g_0 : g_i = (g_{ij}) : [a_i, b_i]^{d_i} \rightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}},
\right.
$$

where $d := (d_0, \ldots, d_q), t := (t_0, \ldots, t_q), \beta := (\beta_0, \ldots, \beta_q).

**Theorem 1.** We consider the nonparametric regression model with $d$ variables for the composite regression function $m = g_q \circ g_{q-1} \circ \ldots \circ g_1 \circ g_0$ in the class $\mathcal{G}(q, d, t, \beta, K)$, as described in Equation (2). Let $\hat{m}_n$ be an estimator from the function class $\mathcal{F}(L, (p_i)_{i=0,..,L+1}, s, F)$ satisfying the following conditions:

1. $F \geq \max(K, 1),$
2. $\sum_{q=0}^q \log_2 (4t_q \vee 4\beta_q) \log_2 n \leq L \lesssim n\phi_n,$
3. $n\phi_n \lesssim \min_{i=1,..,L} p_i,$
4. $s \asymp n\phi_n \ln n,$

where $\phi_n$ is a positive sequence; then, there exist constants $C$ and $C'$ depending only on $q, d, t, \beta, F,$ such that if $\Delta_n(\hat{m}_n, m) \leq C\phi_n L \ln^2 n$, then

$$
R(\hat{m}_n, m) \leq C'\phi_n L \ln^2 n,
$$

if $\Delta_n(\hat{m}_n, m) \geq C\phi_n L \ln^2 n$, then

$$
\frac{1}{C'} \Delta_n(\hat{m}_n, m) \leq R(\hat{m}_n, m) \leq C' \Delta_n(\hat{m}_n, m).
$$

To minimize $\phi_n L \ln^2 n$, let $\Delta_n(\hat{m}_n, m) \leq C\phi_n \ln^3 n$; then,

$$
R(\hat{m}_n, m) \leq C'\phi_n \ln^3 n.
$$

The convergence rate in Theorem 1 depends on $\phi_n$ and $\Delta_n(\hat{m}_n, m)$. The following reasoning shows that $\phi_n$ serves as a lower bound for the supremum infimum estimation risk over this class. For any empirical risk minimizer, where the definition of the $\Delta_n$ term becomes zero, the following corollary holds.

**Corollary 1.** Let $\hat{m}_n \in \arg \min_{f \in \mathcal{F}(L, (p_i), F)} \sum_{i \in \Lambda_n} (Y_i - m(X_i))^2$ be an empirical risk minimizer under the same conditions as in Theorem 1. There exists a constant $C'$, depending only on $q, d, t, \beta, F,$ such that

$$
R(\hat{m}_n, m) \leq C'\phi_n L \ln^2 n.
$$

Condition (1) in Theorem 1 is very mild and states only that the network functions should have at least the same supremum norm as the regression function. From the other assumptions in Theorem 1, it becomes clear that there is a lot of flexibility in selecting a good network architecture as long as the number of active parameters $s$ is taken to be in the right order.

In a fully connected network, the number of network parameters is $\sum_{i=0}^L p_i p_{i+1}$. This implies that Theorem 1 requires a sparse network. More precisely, the network must have at least $\sum_{i=1}^L p_i - s$ completely inactive nodes, meaning that all incoming signals are zero. Condition (4) chooses $s \asymp n\phi_n \ln n$ to balance the mean squared error and variance. From the proof of this theorem (Appendix B), convergence rates for various orders of $s$ can also be derived.
Deep learning excels over other methods only in the large sample regime. This suggests that the method may be adaptable to the underlying structures in the data. This may produce rapid convergence rates, but with larger constants or remainders, which can lead to relatively poor performance in small sample scenarios.

The proof of the risk bounds in Theorem 1 is based on the following oracle-type inequality.

**Theorem 2.** Let us consider the d-dimensional nonparametric regression model given by Equation (2) with an unknown regression function \( m \), where \( F \geq 1 \) and \( |m|_\infty \leq F \), let \( \hat{m}_n \) be an arbitrary estimator taking values in the class \( \mathcal{F}(L, p, s, F) \), and let

\[
\Delta_n(\hat{m}_n, m) := E_m \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - \hat{m}_n(X_i))^2 - \inf_{f \in \mathcal{F}(L, p, s, F)} \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2 \right],
\]

and for any \( \epsilon \in (0, 1] \), there exists a constant \( C_\epsilon \), depending only on \( \epsilon \), such that

\[
\tau_{\epsilon,n} := C_\epsilon F^2 \frac{(s + 1) \ln(n(s + 1)^L p_0 p_{L+1})}{n},
\]

and we have

\[
(1 - \epsilon)^2 \Delta_n(\hat{m}_n, m) - \tau_{\epsilon,n} \leq R(\hat{m}_n, m) \\
\leq (1 + \epsilon)^2 \left( \inf_{f \in \mathcal{F}(L, p, s, F)} \|f - m\|_\infty^2 + \Delta_n(\hat{m}_n, m) \right) + \tau_{\epsilon,n}.
\]

In the context of oracle-type inequalities, an increase in the number of layers can lead to a deterioration in the upper bound on the risk. In practice, it has also been observed that having too many layers can result in a decline in performance. We refer to Section 4.4 in He et al. [9] and He and Sun [10] for more details.

The proof relies on two specific properties of the ReLU activation function rather than other activation functions. The first property is its projection property, which is expressed as

\[
\sigma \circ \sigma = \sigma,
\]

where the composite of the ReLU activation function is considered, given that the foundation of approximation theory lies in constructing smaller networks to perform simpler tasks, which may not all require the same network depth. To combine these subnetworks, it is necessary to synchronize the network depth by adding hidden layers that do not alter the output. This can be achieved by selecting weight matrices in the network (assuming an equal width for consecutive layers) and by utilizing the projection property of the ReLU activation function, given by \( \sigma \circ \sigma = \sigma \). This property is beneficial not only theoretically but also in practice, as it greatly aids in passing a result to deeper layers through skip connections.

Next, we prove that \( \phi_n \) serves as a lower bound for the supremum infimum estimation risk over class \( \mathcal{F}(q, d, t, \beta, K) \) with \( t_i \leq \min(d_0, \ldots, d_{i-1}) \). This means that in the composition of functions, no additional dimensions are added at deeper abstract layers. In particular, this approach avoids the case where \( t_i \) exceeds the input dimension \( d_0 \).

**Theorem 3.** Let us consider the nonparametric regression model (2), where \( X_i \) is drawn from a distribution with a Lebesgue density on \([0, 1]^d\), and the lower and upper bounds of this distribution are positive constants. For any nonnegative integer \( q \), arbitrary dimension vectors \( d \) and \( t \), and for all \( i \) such that \( t_i \leq \min(d_0, \ldots, d_{i-1}) \), and any smoothness vector \( \beta \), along with all sufficiently large constants \( K > 0 \), there exists a positive constant \( c \) such that

\[
\inf_{\hat{m}_n} \sup_{m \in \mathcal{F}(q, d, t, \beta, K)} R(\hat{m}_n, m) \geq c \phi_n,
\]
where \( \inf \) is taken over all estimators \( \hat{m}_n \).

By combining the supremum infimum lower bound with the oracle-type inequality, we can easily obtain the following result.

**Lemma 1.** Given \( \beta, K > 0 \), and \( d \in \mathbb{N} \), there exist constants \( c_1, c_2 \) depending only on \( \beta, K, d \), such that for \( \varepsilon \leq c_2 \), we have

\[
\frac{\varepsilon}{k \ln(1/\varepsilon)} \leq c_1 \left( 1 + \frac{c_2}{\varepsilon} \right),
\]

and then, for any width vector \( \mathbf{p} \), where \( p_0 = d \) and \( p_{L+1} = 1 \), we know that

\[
\inf_{m \in \mathcal{G}_0(0,1)^{d,L}} \sup_{f \in \mathcal{G}(L,p)} \left\| f - m \right\|_\infty \geq \varepsilon.
\]

### 2.3. Suboptimality of Wavelet Series Estimation

In this section, we show that wavelet series estimators are unable to take advantage of the underlying composition structure in the regression function and achieve, in some setups, much slower convergence rates. Wavelet estimation is susceptible to the curse of dimensionality, whereas neural networks can achieve faster convergence rates.

We consider a compressed wavelet system \((\psi, \Lambda) \), restricted to \( L^2[0,1] \) from \( L^2(\mathbb{R}) \), as referred to in Cohen et al. [11]. Here, \( \Lambda = \{ (j,k) : j = -1,0,1,\ldots; k \in I_j \} \), and \( \psi_{-1,k} := \psi(-k) \) denotes the shift-scaled function. For any function \( f \in L^2[0,1]^d \), we have

\[
f(x) = \sum_{(\lambda_1,\ldots,\lambda_d) \in \Lambda \times \cdots \times \Lambda} d_{\lambda_1,\ldots,\lambda_d}(f) \prod_{r=1}^d \psi_{\lambda_r}(x_r),
\]

and the convergence on \( L^2[0,1] \) entails wavelet coefficients.

\[
d_{\lambda_1,\ldots,\lambda_d}(f) := \int f(x) \prod_{r=1}^d \psi_{\lambda_r}(x_r) dx.
\]

To construct a counterexample, it is sufficient to consider the nonparametric regression model \( Y_i = m(X_i) + R_i, i \in \mathcal{A}_n = \{ (i_1,i_2,\ldots,i_N) : i_j = 1,2,\ldots,n, j = 1,2,\ldots,N \} \), \( X_i := (U_{i,1},\ldots,U_{i,d}) \sim U[0,1]^d \). The empirical wavelet coefficients are obtained; furthermore,

\[
\hat{d}_{\lambda_1,\ldots,\lambda_d}(f_0) = \frac{1}{n} \sum_{i \in \mathcal{A}_n} Y_i \prod_{r=1}^d \psi_{\lambda_r}(U_{i,r}).
\]

Since \( E\left[ \hat{d}_{\lambda_1,\ldots,\lambda_d}(f_0) \right] = d_{\lambda_1,\ldots,\lambda_d}(f_0) \), an unbiased estimate for the wavelet coefficients is obtained; furthermore,

\[
\text{Var}\left( \hat{d}_{\lambda_1,\ldots,\lambda_d}(f_0) \right) = \frac{1}{n} \text{Var}\left( Y_1 \prod_{r=1}^d \psi_{\lambda_r}(U_{1,r}) \right) \\
\geq \frac{1}{n} E \left[ \text{Var}\left( Y_1 \prod_{r=1}^d \psi_{\lambda_r}(U_{1,r}) \mid U_{1,1},\ldots,U_{1,d} \right) \right] \\
= \frac{1}{n}.
\]

We study the estimators of the following form

\[
f_n(x) = \sum_{(\lambda_1,\ldots,\lambda_d) \in \Lambda} \hat{d}_{\lambda_1,\ldots,\lambda_d}(f_0) \prod_{r=1}^d \psi_{\lambda_r}(x_r), \quad (11)
\]
and for any subset \( I \subset \Lambda \times \ldots \times \Lambda \), we have

\[
R\left( \hat{f}_n, f_0 \right) = \sum_{(\lambda_1, \ldots, \lambda_d) \in I} E \left[ \left( \hat{d}_{\lambda_1 \ldots \lambda_d}(f_0) - d_{\lambda_1 \ldots \lambda_d}(f_0) \right)^2 \right] + \sum_{(\lambda_1, \ldots, \lambda_d) \in \Lambda^d} d_{\lambda_1 \ldots \lambda_d}(f_0)^2 
\geq \frac{1}{n} \left( \sum_{(\lambda_1, \ldots, \lambda_d) \in \Lambda^d} d_{\lambda_1 \ldots \lambda_d}(f_0)^2 \right). 
\tag{12}
\]

Thus, for any \( j \), we can find a function \( f \in \mathcal{X} \) where

- If \( \psi \) is zero outside \( [0, 2^q] \) for some integer \( q > 0 \).

\section*{Lemma 2.} For any integer \( q > 0 \), \( v := \lceil \log_2 d \rceil + 1 \), and any \( 0 < \alpha \leq 1 \) and \( K > 0 \), there exists a nonzero constant \( c(\psi, d) \), which depends solely on \( d \) and the properties of the wavelet function \( \psi \).

Thus, for any \( j \), we can find a function \( f_j := h_j(x_1 + \ldots + x_d) \), where \( h_j \in \mathcal{C}^2([0, d], K) \), such that for all \( p_1, \ldots, p_d \in \{0, 1, \ldots, 2^{j-v-1} \} \), we have

\[
d_{(j,2^{j+v}p_1, \ldots, j,2^{j+v}p_d)}(f_j) = c(\psi, d)K2^{-\frac{j}{2}(2^v+d)}.
\]

\section*{Theorem 4.} If \( \hat{m}_n \) represents a wavelet estimator with compact support \( \psi \) and an arbitrary index set \( I \),

\[
\hat{m}_n(x) = \sum_{(\lambda_1, \ldots, \lambda_d) \in I} \hat{d}_{\lambda_1 \ldots \lambda_d}(f_0) \mathcal{P}_{\lambda}(x),
\]

therefore, for any \( 0 < \alpha \leq 1 \) and any Hölder radius \( K > 0 \), we have

\[
\sup_{m(x) = h(\sum_{r=1}^d x_r), h \in \mathcal{C}^2([0,d],K)} R(\hat{m}_n, m) \gtrsim n^{1-\frac{2\alpha}{2\alpha+2d}}.
\]

As a result, the convergence rate of the wavelet series estimation is slower than \( n^{-2\alpha/(2\alpha+d)} \). If \( d \) is large, this rate becomes significantly slower. Therefore, wavelet estimation is sensitive to the curse of dimensionality, while neural networks can achieve rapid convergence.

\section*{3. Simulation Experiments and Case Study}

\subsection*{3.1. Simulation Experiments}

In this section, we conduct a comparative study through 100 repeated experimental simulations, evaluating the mean squared error of the estimation using both the local linear regression method and deep neural networks with the ReLU activation function. We consider the following models.

\textbf{Model 1:} \[ Y_i = 2 + 2X_i \beta + e_i, \]

where \( X_i \) follows a uniform distribution on \([-1, 1]\), \( \beta \) is the coefficient, simulated from a uniform distribution on \([0, 1] \), and \( e_i \) is the noise variable following a standard normal distribution.

A neural network with a depth of 3 and a width of 32 was created, where the first two layers are fully connected layers, and the output layer uses the ReLU activation function.
We consider four scenarios with the same sample size but different dimensions:

Scenario 1: \( \text{dim} = 3, n = [200, 600, 1000, 1400, 1800, 2200] \);
Scenario 2: \( \text{dim} = 5, n = [200, 600, 1000, 1400, 1800, 2200] \);
Scenario 3: \( \text{dim} = 8, n = [200, 600, 1000, 1400, 1800, 2200] \);
Scenario 4: \( \text{dim} = 10, n = [200, 600, 1000, 1400, 1800, 2200] \);

where \( \text{dim} \) represents the dimension and \( n \) is the sample size. For each scenario, the mean squared error (MSE) is calculated to compare the performance of the local linear regression method and the deep neural network method. The MSE in this study is defined as follows:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2.
\]

Table 1 presents the estimates at different values of \( n \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.1452</td>
<td>0.0363</td>
<td>0.9120</td>
<td>0.0695</td>
</tr>
<tr>
<td>600</td>
<td>0.0438</td>
<td>0.0150</td>
<td>0.2422</td>
<td>0.0280</td>
</tr>
<tr>
<td>1000</td>
<td>0.0242</td>
<td>0.0110</td>
<td>0.1343</td>
<td>0.0207</td>
</tr>
<tr>
<td>1400</td>
<td>0.0186</td>
<td>0.0097</td>
<td>0.0964</td>
<td>0.0187</td>
</tr>
<tr>
<td>1800</td>
<td>0.0146</td>
<td>0.0080</td>
<td>0.0755</td>
<td>0.0171</td>
</tr>
<tr>
<td>2200</td>
<td>0.0112</td>
<td>0.0075</td>
<td>0.0613</td>
<td>0.0153</td>
</tr>
</tbody>
</table>

In this table, \( \text{NE} \) represents the MSE of the nonparametric estimation method, and \( \text{DNN} \) represents the MSE of the deep neural network method. It is evident that as \( \text{MSE} \) approaches 0, the estimation accuracy increases. From Table 1, we observe that for the same dimension, as the sample size increases, \( \text{DNN} \) tends towards 0. For the same sample size, \( \text{DNN} \) is significantly smaller than \( \text{NE} \), and with the increase in dimension, the superiority of the neural network method over the local linear regression method becomes more pronounced. Therefore, the neural network method achieves much higher estimation accuracy, especially for large sample sizes and high dimensions.

As shown in Figure 1, with the increase in dimension, the performance of the neural network fitting surpasses the local linear regression method, where the x-axis represents the sample size \( n \), and the y-axis represents the mean squared error (MSE). In higher dimensions, the MSE of the neural network method approaches almost zero, which is attributed to the avoidance of the curse of dimensionality by deep neural networks. This demonstrates that the convergence rate of the deep neural network is superior to that of the local linear regression method and approaches the optimal convergence rate.

Next, we consider high-dimensional spatial models with dependency structures to compare the MSEs of the two methods mentioned above.

Model 2:

\[
Y_{ij} = 2 \sin(2\pi X_{ij}) + R_{ij},
\]

where \( i = 1, \ldots, n_1, j = 1, \ldots, n_2 \), and \( X_{ij} \) follows a zero-mean second-order stationary process. \( R_{ij} \) follows a standard normal distribution. Similarly to the work of Cressie and Wikle [12], high-dimensional spatial processes \( X_{ij} \) are generated using spectral methods:

\[
X_{ij} = (2/M)^{1/2} \sum_{k=1}^{M} \cos(w(1,k) \ast i + w(2,k) \ast j + r(k)).
\]
In this case, \( w(i, k) \) for \( i = 1, 2 \) follows a standard normal distribution and is independent of \( r(k) \) for \( k = 1, \ldots, M \), where \( r(k) \) are independently and identically distributed from a uniform distribution on \([-\pi, \pi]\). As \( n \to \infty \), \( X_{i,j} \) converges to a Gaussian random process. We consider the case with dimension 5 and sample sizes \([200, 600, 1000, 1400, 1800, 2200]\). The network structure is the same as in Model 1.

As shown in Table 2, it can be observed that the convergence rate of the high-dimensional spatial model with dependence is worse than the convergence rate of the high-dimensional spatial model without dependence. However, in comparison to the local linear regression method, the MSE values of the neural network are much smaller, indicating that the neural network achieves better convergence performance.

Table 2. The MSE values of both methods for Model 2.

<table>
<thead>
<tr>
<th>( n )</th>
<th>200</th>
<th>600</th>
<th>1000</th>
<th>1400</th>
<th>1800</th>
<th>2200</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>36.8606</td>
<td>32.4013</td>
<td>17.6990</td>
<td>14.0570</td>
<td>10.7740</td>
<td>5.1166</td>
</tr>
<tr>
<td>DNN</td>
<td>2.0305</td>
<td>2.0078</td>
<td>2.0074</td>
<td>2.0047</td>
<td>2.0041</td>
<td>2.0020</td>
</tr>
</tbody>
</table>

Figure 1. MSEs of local linear regression method (dashed blue line) and the neural network method (solid orange line) of dimension 3, 5, 8 and 10. (a) Three dimensions. (b) Five dimensions. (c) Eight dimensions. (d) Ten dimensions.

As shown in Figure 2, we see that in the case of large sample sizes and high-dimensional spatial models, the neural network achieves superior convergence compared to that of the local linear regression method, where the x-axis represents the sample size \( n \), and the y-axis represents the mean squared error (MSE).
3.2. Case Study

To compare the consistency of the local linear regression method and deep neural network for high-dimensional spatial models, we consider the relationship between air pollution and respiratory diseases in the New Territories East of Hong Kong from 1 January 2000 to 15 January 2001, as studied by Wang et al. [13]. There is a dataset consisting of 821 observations, where we mainly consider the air pollution index $X_1$ and five pollutants, sulfur dioxide (g/m$^3$) $X_2$, inhalable particulate matter (g/m$^3$) $X_3$, nitrogen compounds (g/m$^3$) $X_4$, nitrogen dioxide (g/m$^3$) $X_5$, and ozone (g/m$^3$) $X_6$, as well as two environmental factors: temperature (°C) $X_7$ and relative humidity (%) $X_8$. In this section, we examine the relationship between the levels of chemical pollutants in the New Territories East of Hong Kong and the daily hospital admissions for respiratory diseases ($Y$). The specific parameter settings are the same as those in the numerical simulation part.

After dimensionless and standardization processing, we use 397 data points for the case study. Among these, 80% of the data are used to train the model, and the remaining 20% are used to evaluate the quality of the trained model. The $MSE$ values are shown in Table 3. It can be observed that the $MSE_2$ values are closer to 0, indicating that in real-world cases, the mean squared error of the deep neural network method is much smaller than that of the local linear regression method. Therefore, the deep neural network shows better convergence performance.

<table>
<thead>
<tr>
<th>$n$</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>$NE$</td>
<td>3.1912</td>
<td>2.2667</td>
<td>2.0146</td>
<td>1.8464</td>
<td>1.7926</td>
<td>1.7485</td>
</tr>
<tr>
<td>$DNN$</td>
<td>1.7958</td>
<td>1.5066</td>
<td>1.3923</td>
<td>1.3338</td>
<td>1.2983</td>
<td>1.2901</td>
</tr>
</tbody>
</table>

Figure 3 presents a visual representation of the $MSE$ values for both methods, where the x-axis represents the sample size $n$, and the y-axis represents the mean squared error ($MSE$). From the graph, it is evident that the deep neural network method exhibits a faster convergence rate compared to that of the local linear regression method.
4. Conclusions

In this study, we employ neural networks with ReLU activation functions for nonparametric estimation in high-dimensional spaces. By constructing suitable network architectures, we estimate unknown trend functions and prove the consistency of the estimators while also comparing and analyzing the deep neural network approach with traditional nonparametric methods.

The focus is on high-dimensional space models with unknown error distributions. Considering the spatial dependencies and heterogeneity in the space models, a deep neural network with ReLU activation functions is used to estimate the unknown trend functions. Under general assumptions, the consistency of the estimators is established, and bounds for the mean squared error (MSE) are provided. The estimators exhibit a convergence rate that is related to the sample size but independent of the dimensionality $d$, thereby avoiding the curse of dimensionality. Moreover, the proposed estimators achieve convergence speeds close to optimality.

Considering the spatial dependencies in high-dimensional settings with large sample sizes, the deep neural network method outperforms traditional nonparametric estimation methods.

Author Contributions: Methodology, H.W.; software & writing, X.J.; editing, H.H.; review J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Social Science Fund of China, grant number 22BTJ021.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Embedding Property of Network Function Classes

To approximate functions by using neural networks, we first construct smaller networks to compute simpler objects. Let $p = (p_0, \ldots, p_{L+1})$ and $p' = (p'_0, \ldots, p'_{L+1})$.

To merge networks, the following rules are commonly used in this paper.

Enlargement: $\mathcal{F}(L, p, s) \subseteq \mathcal{F}(L, q, s')$, $p \leq q$ and $s \leq s'$.

Composition: Let $f \in \mathcal{F}(L, p)$ and $g \in \mathcal{F}(L', p')$, where $p_{L+1} = p'_0$. For a vector $v \in \mathbb{R}^{p_{L+1}}$, in the space $\mathcal{F}(L + L' + 1, (p_0, \ldots, p_{L+1}, p'_0, \ldots, p'_{L+1}))$, i.e., $\mathcal{F}(L + L' + 1, (p, p'_0, \ldots, p'_{L+1}))$, the composition network $g \circ v(f)$ is defined.
Additional Layer/Depth Synchronization: To synchronize the number of hidden layers in two networks, an additional layer can be added with a unit weight matrix, such that
\[
\mathcal{F}(L, p, s) \subset \mathcal{F}\left(L + q, \left(p_0, \ldots, p_0, p\right), s + qp_0\right).
\] (A1)

Parallelization: Let \( f \) and \( g \) be two networks with the same number of hidden layers and the same input dimension, i.e., \( f \in \mathcal{F}(L, p) \) and \( g \in \mathcal{F}(L, p') \), where \( p_0 = p_0' \). The parallel network \((f, g)\) computes both \( f \) and \( g \) simultaneously in the joint network class \( \mathcal{F}(L, (p_0, p_1 + p_1', \ldots, p_{L+1} + p_{L+1}')) \).

Removing Inactive Nodes: We have
\[
\mathcal{F}(L, p, s) = \mathcal{F}(L, (p_0, p_1 \wedge s, p_2 \wedge s, \ldots, p_L \wedge s, p_{L+1}), s).
\] (A2)

In this context, we have \( p_i \wedge s = \min\{p_i, s\}, i = 1, \ldots, L \). Let \( f(x) = W_{L}v_{L}W_{L-1} \cdots v_{1}W_{0}x \in \mathcal{F}(L, p, s) \). If all entries in the \( j \)-th column of \( W_{i} \) are zero, we can remove this column along with the \( j \)-th row of \( W_{i-1} \) and the \( j \)-th element of \( v_{i} \) without changing the function. This implies that \( f \in \mathcal{F}(L, (p_0, \ldots, p_{i-1}, p_{i} - 1, p_{i+1}, \ldots, p_{L+1}), s) \). Since there are \( s \) active parameters, for any \( i = 1, \ldots, L \), we need to iterate at least \( p_i - s \) times. This proves that \( f \in \mathcal{F}(L, (p_0, p_1 \wedge s, p_2 \wedge s, \ldots, p_L \wedge s, p_{L+1}), s) \).

In this paper, we often utilize the following fact. For a fully connected network in \( \mathcal{F}(L, p) \), there are \( \sum_{\ell=0}^{L} p_{\ell}p_{\ell+1} \) weight matrix parameters and \( \sum_{\ell=1}^{L} p_{\ell} \) network parameters from bias vectors. Therefore, the total number of parameters is
\[
\sum_{\ell=0}^{L} (p_{\ell} + 1)p_{\ell+1} - p_{L+1}.
\] (A3)

**Appendix B. Approximation by Polynomial Neural Networks**

We construct a network with all parameters bounded by 1 to approximate the calculation of \( xy \) for given inputs \( x \) and \( y \). Let \( T^k : \left[0, 2^{2-2k}\right] \to \left[0, 2^{-2k}\right] \), where \( k \) is a positive integer.

\[
T^k(x) := (x/2) \wedge \left(2^{1-2k} - x/2\right) = (x/2) - \left(x - 2^{1-2k}\right)_+,
\]

and \( R^k : \left[0, 1\right] \to \left[0, 2^{-2k}\right], \) where
\[
R^k := T^k \circ T^{k-1} \circ \ldots \circ T^1.
\]

Next, we prove that \( \sum_{k=1}^{m} R^k(x) \) converges exponentially to \( x(1-x) \) as \( m \) increases, especially in \( L^{\infty}[0,1]:x(1-x) = \sum_{k=1}^{\infty} R^k(x) \). This lemma can be seen as a variation of Lemma 2.4 in Telgarsky’s work [14] and Proposition 2 in Yarotsky’s work [15]. Compared to existing results, this result allows us to construct networks with parameters equal to 1 and provides an explicit bound on the approximation error.

**Lemma A1.** For any positive integer \( m \),
\[
\left|x(1-x) - \sum_{k=1}^{m} R^k(x)\right| \leq 2^{-m}.
\]

**Proof.** Step 1: We prove by induction that \( R^k(x) \) is a triangular wave. More precisely, \( R^k(x) \) is piecewise linear on the intervals \( \left[\ell/2^k, (\ell + 1)/2^k\right] \), where \( \ell \) is an integer. If \( \ell \) is odd, the endpoints are \( R^k(\ell/2^k) = 2^{-2k} \), and if \( \ell \) is even, the endpoints are \( R^k(\ell/2^k) = 0.\)
When \( k = 1 \), the equality \( R^1 = T^1 \) holds obviously.

We assume that the statement holds for \( k \), and we let \( \ell - r \) be divisible by 4, denoted as \( \ell \equiv r \mod 4 \). Consider \( x \) in the interval \( \left[ \ell - 2^{k+1}, (\ell + 1)/2^{k+1} \right] \). When \( \ell \equiv 0 \mod 4 \), then \( R^k(x) = 2^{-k} \left( x - \ell - 2^{k+1} \right) \). When \( \ell \equiv 2 \mod 4 \), then \( R^k(x) = 2^{-2k} - 2^{-k} \left( x - \ell - 2^{k+1} \right) \).

When \( \ell \equiv 1 \mod 4 \), where \( \ell + 1 \equiv 2 \mod 4 \), then \( R^k(x) = 2^{-k} \left( x - \ell - 2^{k} \right) - 2^{-2k} \). When \( \ell \equiv 3 \mod 4 \), then \( R^k(x) = 2^{-2k} - 2^{-k} \left( x - \ell - 2^{k} \right) \). Then, for \( k + 1 \), the statement also holds.

\[
R^{k+1}(x) = T^{k+1} \circ R^k(x)
= \frac{R^k(x)}{2} \, \mathbf{1}(R^k(x) \leq 2^{-2k-1}) + \left(2^{-2k-1} - \frac{R^k(x)}{2}\right) \, \mathbf{1}(R^k(x) > 2^{-2k-1}).
\]

The statement holds for \( k + 1 \), and the induction is complete.

Step 2: For convenience, let us denote \( g(x) = x(1-x) \). We now prove that for any \( m \geq 1 \) and \( \ell \in \{0, 1, \ldots, 2^m\} \), the following holds:

\[
g(\ell 2^{-m}) = \sum_{k=1}^{m} R^k(\ell 2^{-m}).
\]

To prove this, we use mathematical induction on \( m \). For \( m = 1 \), when \( \ell = 0 \), we have \( g(0) = 0 \) and \( R^1(0) = 0 \). When \( \ell = 1 \), we have \( g(1/2) = 1/4 \) and \( R^1(1/2) = 1/2 \). When \( \ell = 2 \), we have \( g(1) = 0 \) and \( R^1(1) = 0 \). \( g(\ell 2^{-1}) = R^k(\ell 2^{-1}) \) holds. Therefore, for the inductive step, assuming that it holds for \( m + 1 \), when \( m + 1 \) is considered, if \( \ell \) is even, then \( R^{m+1}(\ell 2^{-m-1}) = 0 \), which implies \( g(\ell 2^{-m-1}) = \sum_{k=1}^{m} R^k(\ell 2^{-m-1}) = \sum_{k=1}^{m+1} R^k(\ell 2^{-m-1}) \). If \( \ell \) is odd, then the function \( x \mapsto \sum_{k=1}^{m} R^k(x) \) is linear over the interval \( [(\ell - 1)2^{-m-1}, (\ell + 1)2^{-m-1}] \). Furthermore, for any \( t \), we have

\[
g(x) - \frac{g(x + t) + g(x - t)}{2} = t^2.
\]

Since \( x = \ell 2^{-m-1} \) and \( t = 2^{-m-1} \), and considering \( \ell \) such that \( R^{m+1}(\ell 2^{-m-1}) = 2^{-2m-2} \), we can deduce

\[
g(\ell 2^{-m-1}) = 2^{-2m-2} + \sum_{k=1}^{m} R^k(\ell 2^{-m-1}) = \sum_{k=1}^{m+1} R^k(\ell 2^{-m-1}),
\]

and the result also holds for \( m + 1 \), completing the induction.

Thus, the interpolation of \( \sum_{k=1}^{m} R^k(x) \) at the point \( \ell 2^{-m} \) with function \( g \) has been proven, and it is linear over the interval \( [(\ell - 1)2^{-m}, (\ell + 1)2^{-m}] \). Let \( \sum_{k=1}^{m} R^k(x) = y \); \( g \) is a Lipschitz function with Lipschitz constant 1. Therefore, for any \( x \), there exists an \( \ell \) determined by

\[
g(2^{-m}(\ell + 1)) - y = \frac{y - g(2^{-m} \ell)}{x - 2^{-m} \ell},
\]

and we have

\[
y = 2^{-m}g(2^{-m}(\ell + 1))x - g(2^{-m}(\ell + 1))\ell - (\ell + 1)g(2^{-m} \ell) + 2^{-m}xg(2^{-m} \ell),
\]

which implies

\[
\left| g(x) - \sum_{k=1}^{m} R^k(x) \right| = \left| g(x) - (2^{-m}x - \ell)g((\ell + 1)2^{-m}) - (\ell + 1 - 2^{-m}x)g(2^{-m}) \right|
\leq 2^{-m},
\]
thus proving the lemma.

Let \( g(x) = x(1-x) \). As proven above, to construct a network that takes inputs \( x \) and \( y \) and approximates the product \( xy \), we use polar-type identities.

\[
g \left( \frac{x - y + 1}{2} \right) - g \left( \frac{x + y}{2} \right) + \frac{x + y}{2} - \frac{1}{4} = xy. \tag{A4}
\]

\( \square \)

**Lemma A2.** For any positive integer \( m \), there exists a network \( \text{Mult}_m \in \mathcal{F}(m + 4, (2, 6, 6, \ldots, 6, 1)) \) such that \( \text{Mult}_m(x, y) \in [0, 1] \) for all \( x, y \in [0, 1] \), and

\[
|\text{Mult}_m(x, y) - xy| \leq 2^{-m},
\]

and \( \text{Mult}_m(0, y) = \text{Mult}_m(x, 0) = 0. \)

**Proof.** Let \( T_k(x) = \left( \frac{x}{2} \right)_+ - \left( x - 2^{1-2k} \right)_+ = T_+(x) - T_k^-(x) \), where \( T_+(x) = \left( \frac{x}{2} \right)_+ \) and \( T_k^-(x) = \left( x - 2^{1-2k} \right)_+ \). Consider a nonnegative function \( h : [0, 1] \to [0, \infty) \).

Step 1: We prove the existence of a network \( N_m \) with \( m \) hidden layers and width vector \((3, 3, \ldots, 3, 1)\) to compute this function:

\[
\left( T_+(u), T_-^1(u), h(u) \right) \mapsto \sum_{k=1}^{m+1} R_k(u) + h(u),
\]

for all \( u \in [0, 1] \), as shown in Figure A1; it is worth noting that all parameters in this network are bounded by 1.

![Figure A1](image-url)

Figure A1. Network \( \left( T_+(u), T_-^1(u), h(u) \right) \mapsto \sum_{k=1}^{m+1} R_k(u) + h(u) \).

Step 2: We prove the existence of a network with \( m \) hidden layers to compute the following functions:

\[
(x, y) \mapsto \left( \sum_{k=1}^{m+1} R_k \left( \frac{x-y+1}{2} \right) - \sum_{k=1}^{m+1} R_k \left( \frac{x+y}{2} \right) + \frac{x+y}{2} - \frac{1}{4} \right)_+ \land 1.
\]

Given the input \( (x, y) \), the computation of this network in the first layer is as follows:

\[
\left( T_+ \left( \frac{x-y+1}{2} \right), T_-^1 \left( \frac{x-y+1}{2} \right), \left( \frac{x+y}{2} \right)_+, T_+ \left( \frac{x+y}{2} \right), T_-^1 \left( \frac{x+y}{2} \right), \frac{1}{2} \right).
\]
Applying the network \( N_m \) on the first three elements and the last three elements, we obtain a network with \( m + 1 \) hidden layers and a width vector of \((2,6,\ldots,6,2)\), and we compute

\[
(x, y) \mapsto \left( \sum_{k=1}^{m+1} R^k \left( \frac{x - y + 1}{2} \right) + \frac{x + y}{2} \right) \sum_{k=1}^{m+1} R^k \left( \frac{x + y}{2} \right) + \frac{1}{4},
\]

applying the two-hidden-layer network \((u,v) \mapsto (1-(1-(u-v)+)+)_+ = (u-v)_+ \land 1\) to the output. Therefore, the composite network Mult\(_m\)(x, y) has \(m + 4\) hidden layers and computes

\[
(x, y) \mapsto \left( \sum_{k=1}^{m+1} R^k \left( \frac{x - y + 1}{2} \right) - \sum_{k=1}^{m+1} R^k \left( \frac{x + y}{2} \right) + \frac{x + y - 1}{4} \right)_+ \land 1, \tag{A5}
\]

and this implies that the output is always in the interval [0, 1]. According to Equation (A4) and Lemma A1, we can obtain the following:

\[
|M_{mult}(x, y) - xy| = \left| \sum_{k=1}^{m+1} R^k \left( \frac{x - y + 1}{2} \right) - \sum_{k=1}^{m+1} R^k \left( \frac{x + y}{2} \right) + \frac{x + y - 1}{4} \right|
\]

\[
\leq \left| \sum_{k=1}^{m+1} R^k \left( \frac{x - y + 1}{2} \right) - \sum_{k=1}^{m+1} R^k \left( \frac{x + y}{2} \right) - \frac{x + y - 1}{4} \right|
\]

\[
\leq 2^{-m-1} + 2^{-m-1} \leq 2^{-m}.
\]

For all \(0 \leq u \leq 1\), we have \(R^1((1 - u)/2) = R^1((1 + u)/2)\) and \(R^2((1 + u)/2) = R^2(u/2)\). Therefore, when \(k\) is odd, \(R^k((1 - u)/2) = R^k((1 + u)/2)\); when \(k\) is even, \(R^k((1 + u)/2) = R^k(u/2)\). For all input pairs \((0, y)\) and \((0, x)\), the output in Equation (A5) becomes zero. \(\square\)

**Lemma A3.** For any positive integer \(m\), there exists a network

\[ \text{Mult}_m^r \in \mathcal{F}((m + 5)\lceil \log_2 r \rceil, (r, 6r, 6r, \ldots, 6r, 1)), \]

such that \(\text{Mult}_m^r \in [0, 1]\), for all \(x = (x_1, \ldots, x_r) \in [0, 1]^r\), and we have

\[ \left| \text{Mult}_m^r(x) - \prod_{i=1}^{r} x_i \right| \leq r^2 2^{-m}. \]

Furthermore, if one of the components of \(x\) is 0, then \(\text{Mult}_m^r(x) = 0\).

**Proof.** Let \(q := \lceil \log_2 (r) \rceil\). We now construct the network \(\text{Mult}_m^r\) and perform calculations in the first hidden layer.

\[
(x_1, \ldots, x_r) \mapsto (x_1, \ldots, x_r, 1_{2^r-r}) \tag{A6}
\]

We apply the network \(\text{Mult}^m\) from Lemma A2 to each pair \((x_1, x_2), (x_3, x_4), \ldots, (1, 1)\) to compute \(\text{Mult}_m^r(x_1, x_2), \text{Mult}_m^r(x_3, x_4), \ldots, \text{Mult}_m^r(1, 1)\) \(\in \mathbb{R}^{2^r-1}\). Now, we pair adjacent terms and apply \(\text{Mult}_m^r\) again. We continue this process until only one term remains. The resulting network is denoted as \(\text{Mult}_m^r\), which has \(q(m + 5)\) hidden layers, and all parameters are bounded by 1.
If \( a, b, c, d \in [0, 1] \), then by Lemma A2 and the triangle inequality, we have

\[
| \text{Mult}_m(a, b) - cd | \leq | \text{Mult}_m(a, b) - ab | + | ab - cd | \leq 2^{-m} + |a - c| + |b - d|;
\]

therefore, by iteration and induction, we obtain

\[
| \text{Mult}_m^r(x) - \prod_{i=1}^r x_i | \leq 3^{q-1} 2^{-m} \leq 3^{|\log_2(r)| - 1} 2^{-m} \\
\leq 3^{|\log_2(r)|} 2^{-m} \leq 4^{|\log_2(r)|} 2^{-m} \\
\leq r^2 2^{-m}.
\]

By using Lemma A2 and the construction described above, it is evident that if one of the components of \( x \) is 0, then \( \text{Mult}_m^r(x) = 0 \).

We construct a sufficiently large network to approximate all monomials \( x_1^{a_1} \cdots x_r^{a_r} \) for nonnegative integers \( a_i \) up to a certain specified degree. Typically, we use multi-index notation: \( x^\alpha := x_1^{a_1} \cdots x_r^{a_r} \), where \( \alpha = (a_1, \ldots, a_r) \) and \( |\alpha| := \sum_{\ell=1}^r |a_\ell| \) represents the degree of the monomial.

The number of monomials with degrees satisfying \( |\alpha| < \gamma \) is denoted by \( \text{Cr}_{r,\gamma} \), and since each \( a_i \) takes values in \( \{0, 1, \ldots, \lfloor \gamma \rfloor \} \), we have \( \text{Cr}_{r,\gamma} \leq (\gamma + 1)^r \). \( \square \)

**Lemma A4.** For \( \gamma > 0 \) and positive integers \( m \), there exists a network

\[
\text{Mon}_{m,\gamma}^r \in \mathcal{F}(1 + (m + 5)|\log_2(\gamma \lor 1)|, (r, 6|\gamma| \text{Cr}_{r,\gamma}, \ldots, 6|\gamma| \text{Cr}_{r,\gamma}, \text{Cr}_{r,\gamma})),
\]

such that \( \text{Mon}_{m,\gamma}^r \in [0, 1]^\text{Cr}_{r,\gamma} \), and for all \( x \in [0, 1]^r \), we have

\[
|\text{Mon}_{m,\gamma}^r(x) - (x^\alpha)_{|\alpha|<\gamma}|_{\infty} \leq \gamma 2^{-m}.
\]

**Proof.** For \( |\alpha| \leq 1 \), the monomials are either linear or constant functions. There exists a shallow network in the class \( \mathcal{F}(1, (1, 1, 1)) \) that precisely represents the monomial \( x^\alpha \).

Considering the multiplicities in Equation (A6), Lemma A3 can be directly extended to monomials. For \( |\alpha| \geq 2 \), this implies that in the class

\[
\mathcal{F}((m + 5)|\log_2|\alpha||, (r, 6|\alpha|, \ldots, 6|\alpha|, 1)),
\]

there exists a network in the class that takes values within the interval \([0, 1]\) and approximates \( x^\alpha \) to a supnorm error of \( |\alpha|^2 2^{-m} \). By utilizing the parallelization and depth synchronization properties discussed in Appendix B, the proof of Lemma A4 can be established.

Following the classical local Taylor approximation, previously used for network approximation by Yarotsky [15], for a vector \( \alpha \in [0, 1]^r \), we define

\[
P_\alpha^p f(x) = \sum_{0 \leq |\alpha| < \beta} (\partial^\alpha f)(\alpha) \frac{(x - \alpha)^\alpha}{\alpha!}.
\]

(A7)

According to Taylor’s theorem for multivariable functions, for an appropriate \( \xi \in [0, 1] \), we have

\[
f(x) = \sum_{\alpha:|\alpha| < \beta - 1} (\partial^\alpha f)(\alpha) \frac{(x - \alpha)^\alpha}{\alpha!} + \sum_{\beta - 1 \leq |\alpha| < \beta} (\partial^\alpha f)(\alpha + \xi (x - \alpha)) \frac{(x - \alpha)^\alpha}{\alpha!}.
\]
We have \(|(x - a)^a| = \prod_i |x_i - a_i|^a| \leq |x - a|^a|\). Therefore, for \(f \in \mathbb{C}^r([0,1]^r,K)\), we have
\[
|f(x) - P_a^f(x)| = \sum_{\beta-1 \leq |a| \leq \beta} \left| \frac{(x-a)^a}{a!} \right| |(\partial^a f)(a + \xi(x-a)) - (\partial^a f)(a)|
\] (A8)
\[
\leq K|x - a|^\beta.
\]
We can express Equation (A7) as a linear combination of monomials.
\[
P_a^f(x) = \sum_{0 \leq |\gamma| < \beta} x^\gamma c_\gamma,
\] (A9)
for suitable coefficients \(c_\gamma\), and, for convenience, the dependence on \(a\) in \(c_\gamma\) is omitted here. Since \(\partial^\gamma P_a^f(x)|x = 0 = \gamma!c_\gamma\), it follows that
\[
c_\gamma = \sum_{\gamma \leq a \& |a| < \beta} (\partial^a f)(a) \frac{(-a)^{a-\gamma}}{\gamma!(a-\gamma)!},
\]
and since \(a \in [0,1]^r\) and \(f \in \mathbb{C}^r([0,1]^r,K)\), we have
\[
|c_\gamma| \leq K/\gamma! \quad \text{and} \quad \sum_{\gamma \geq 0} |c_\gamma| \leq K \prod_{j=1}^r \sum_{\gamma_j \geq 0} \frac{1}{\gamma_j!} = Ke^r.
\] (A10)
We consider the grid points \(D(M) := \{x_\ell = (\ell_i/M)_{i=1,...,r} : \ell = (\ell_1,..,\ell_r) \in \{0,1,..,M\}^r\}\). The number of elements in this set is \((M+1)^r\). Let \(x_\ell = (x_\ell^j)\) represent the elements of \(X_\ell\). We define
\[
P_a f(x) := \sum_{x_\ell \in D(M)} P_a^f(x) \prod_{j=1}^r \left(1 - M|x_j - x_\ell^j|\right)_+.
\]
\[\square\]
Lemma A5. If \(f \in \mathbb{C}^r([0,1]^r,K)\), then \(|P_a f - f|_{L^\infty([0,1]^r)} \leq KM^{-\beta}\).

Proof. For all \(x = (x_1,..,x_r) \in [0,1]^r\), we have
\[
\sum_{x_\ell \in D(M)} \prod_{j=1}^r \left(1 - M|x_j - x_\ell^j|\right)_+ = \prod_{j=1}^r \sum_{\ell=0}^M \left(1 - M|x_j - \ell/M|\right)_+ = 1,
\] (A11)
and we use mathematical induction, assuming \(M = 1\). The left-hand side of (A11) is
\[
(1 - |x_1 - 0|)_+ = 1 - x_1, (1 - |x_1 - 1|)_+ = x_1 = 1,
\]
and after summing, we obtain
\[
(1 - x_1) + x_1 = 1
\]
while the middle-hand side, we have
\[
(1 - |x_j - 0|)_+ + (1 - |x_j - 1|)_+ = (1 - x_j) + x_j = 1,
\]
and therefore the equation holds when \(M = 1\).
Assuming \( M = 2 \), we have \( x_1 \in (0, \frac{1}{2}) \), \( x_2 \in (\frac{1}{2}, 1) \), with \( x_{1}^f = \frac{\ell_1}{M}, \ell_1 = 0, 1, 2; \)
\( x_{2}^f = \frac{\ell_2}{M}, \ell_2 = 0, 1, 2 \). Then, the left-hand side is
\[
\left( 1 - 2 \left| x_1 - \frac{0}{2} \right| \right)_+ \left( 1 - 2 \left| x_2 - \frac{0}{2} \right| \right)_+ = 0,
\]
\[
(1 - 2|x_1 - 0|)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = (1 - 2x_1)(1 - 2x_2 + 1) = (1 - 2x_1)(2 - 2x_2),
\]
\[
(1 - 2|x_1 - 0|)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = (1 - 2x_1)(1 - 2 + 2x_2) = (1 - 2x_1)(2x_2 - 1),
\]
\[
\left( 1 - 2 \left| x_1 - \frac{1}{2} \right| \right)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = 2x_1(2 - 2x_2),
\]
\[
\left( 1 - 2 \left| x_1 - \frac{1}{2} \right| \right)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = 2x_1(2x_2 - 1),
\]
\[
(1 - 2|x_1 - 1|)_+ \left( 1 - 2 \left| x_2 - 0 \right| \right)_+ = 0,
\]
\[
(1 - 2|x_1 - 1|)_+ \left( 1 - 2 \left| x_2 - 0 \right| \right)_+ = 0,
\]
\[
(1 - 2|x_1 - 1|)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = 0,
\]
\[
(1 - 2|x_1 - 1|)_+ \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ = 0,
\]
and after summation, we obtain
\[
(1 - 2x_1)(2 - 2x_2) + (1 - 2x_1)(2x_2 - 1) + 2x_1(2 - 2x_2) + 2x_1(2x_2 - 1) = 1.
\]

In the middle, we have
\[
\prod_{j=1}^{2} \left( \left( 1 - 2 \left| x_j - 0 \right| \right)_+ + \left( 1 - 2 \left| x_j - \frac{1}{2} \right| \right)_+ + \left( 1 - 2 \left| x_j - 1 \right| \right)_+ \right)
= \left( 1 - 2|x_1 - 0| \right)_+ + \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ + \left( 1 - 2|x_1 - 1| \right)_+ \cdot \left( 1 - 2|x_2 - 0| \right)_+ + \left( 1 - 2 \left| x_2 - \frac{1}{2} \right| \right)_+ + \left( 1 - 2|x_2 - 1| \right)_+ \cdot (1 - 2x_1) + 2x_1)((2 - 2x_2) + (2x_2 - 1)) = 1,
\]
and, therefore, when \( M = 2 \), the equation holds true.

Next, we calculate the second equation in (A11); when \( x_j \in \left( 0, \frac{1}{M} \right) \), we have
\[
\left( 1 - M \left| x_j - \frac{0}{M} \right| \right)_+ = 1 - Mx_j,
\]
\[
\left( 1 - M \left| x_j - \frac{1}{M} \right| \right)_+ = (1 - 1 + Mx_j)_+ = Mx_j,
\]
\[
\left( 1 - M \left| x_j - \frac{2}{M} \right| \right)_+ = (1 - 2 + Mx_j)_+ = (Mx_j - 1)_+ = 0,
\]
and when \( \ell \) takes values 3, 4, ..., \( M \), the sum above is zero, resulting in 1. By analogy, for \( x_j \in \left( \frac{k}{M}, \frac{k+1}{M} \right) \), where \( k = 0, \ldots, M - 1 \), the same holds true. Therefore, we can deduce that \( \prod_{j=1}^{M} \sum_{\ell=0}^{M} \left( 1 - M \left| x_j - \ell/M \right| \right)_+ = 1.\)

By using \( f(x) = \sum_{x_i \in D(M)} \mathbb{1}_{x_i \leq 1/M} f(x) \prod_{j=1}^{M} \left( 1 - M \left| x_j - x_j^f \right| \right)_+ \) and Equation (A8), we obtain
\[ |p^\beta f(x) - f(x)| = \left| \sum_{x_i \in \mathcal{D}(M)} p_{x_i}^\beta f(x) \prod_{j=1}^{\ell} \left( 1 - M |x_j - x'_j| \right)_+ - \sum_{x_i \in \mathcal{D}(M) : \|x-x_i\|_{\infty} < 1/M} f(x) \prod_{j=1}^{\ell} \left( 1 - M |x_j - x'_j| \right)_+ \right| \]
\[ \leq \max_{x_i \in \mathcal{D}(M) : \|x-x_i\|_{\infty} \leq 1/M} \left| p_{x_i}^\beta f(x) - f(x) \right| \]
\[ \leq K\|x - x_\ell\|_{\infty}^\beta \]
\[ \leq KM^\beta. \]

Then, we describe how to construct a network that approximates \( p^\beta f \).

**Lemma A6.** For any positive integers \( M \) and \( m \), there exists a network \( \text{Hat}^\ell \in \mathcal{F}(2 + (m + 5) \lfloor \log_2 r \rfloor, (r, 6r(M + 1)^\ell)^\ell, \ldots, 6r(M + 1)^\ell, (M + 1)^\ell), s, 1) \), where \( s \leq 49 \)
\[ r^2(M + 1)^\ell \left( 1 + (m + 5) \lfloor \log_2 r \rfloor \right) \], such that \( \text{Hat}^\ell \in [0, 1](M+1)^\ell \), and for any \( x = (x_1, \ldots, x_r) \in [0, 1]^r \), we have
\[ \left| \text{Hat}^\ell(x) - \left( \prod_{j=1}^{\ell} \left( 1/M - |x_j - x'_j| \right)_+ \right)_{x_i \in \mathcal{D}(M)} \right|_{\infty} \leq r^2 2^{-m}. \]

For any \( x_i \in \mathcal{D}(M) \), the support of the function \( x \mapsto (\text{Hat}^\ell(x))_{x_i} \) is contained within the support of the function \( x \mapsto \prod_{j=1}^{\ell} \left( 1/M - |x_j - x'_j| \right)_+ \).

**Proof.** The first hidden layer uses \( 2r(M + 1) \) units and \( 4r(M + 1) \) nonzero parameters to compute the functions \( (x_j - \ell/M)_+ \) and \( (\ell/M - x_j)_+ \). The second hidden layer uses \( r(M + 1) \) units and \( 3r(M + 1) \) nonzero parameters to compute the function \( (1/M - |x_j - \ell/M|)_+ = (1/M - (x_j - \ell/M)_+ - (\ell/M - x_j)_+) \). These functions take values in the interval \([0, 1]\), and the result holds when \( r = 1 \).

For \( r > 1 \), we combine the obtained network with the network approximating the product \( \prod_{j=1}^{\ell} \left( 1/M - |x_j - \ell/M| \right)_+ \). According to Lemma A3, there exists a network \( \text{Mult}^\ell_m \) in the following class:
\[ \mathcal{F}(m + 5) \lfloor \log_2 r \rfloor, (r, 6r, 6r, \ldots, 6r, 1) \).

We compute \( \prod_{j=1}^{\ell} \left( 1/M - |x_j - x'_j| \right)_+ \) with an error bounded by \( r^2 2^{-m} \). From Equation (A3), it follows that a \( \text{Mult}^\ell_m \) network has nonzero parameters
\[ (36r^2 + 6r)(1 + ((m + 5) \lfloor \log_2 r \rfloor)) \leq 42r^2(M + 1)^\ell \left( 1 + (m + 5) \lfloor \log_2 r \rfloor \right), \]
as a bound, and since these networks have \((M + 1)^\ell \) parallel instances, each hidden layer requires \( 6(M + 1)^\ell \) units and \( 42r^2(M + 1)^\ell \left( 1 + (m + 5) \lfloor \log_2 r \rfloor \right) \) nonzero parameters for multiplication operations. Adding the \( 7(M + 1)^\ell \) nonzero parameters from the first two layers, the total bound on the number of nonzero parameters is
\[ 49r^2(M + 1)^\ell \left( 1 + (m + 5) \lfloor \log_2 r \rfloor \right). \]
According to Lemma A3, if one of the components of $x$ is zero, then $\text{Mult}_m^r(x) = 0$. This implies that for any $x_i \in D(M)$, the support of the function $x \mapsto (\text{Hat}'(x))_{x_i}$ is contained within the support of the function $x \mapsto \prod_{j=1}^r (1/M - |x_j - x'_j|)_+$. $\square$

**Theorem A1.** For any function $f \in \mathcal{F}^r_1([0,1]^r, K)$ and any integers $m \geq 1$ and $N \geq (\beta + 1)r \vee (K + 1)e'$, there exists a network

$$\hat{f} \in \mathcal{F}(L, (r, 6(r + \lceil \beta \rceil))N, \ldots, 6(r + \lceil \beta \rceil)N, 1, s, \infty),$$

with a depth of

$$L = 8 + (m + 5)(1 + \lceil \log_2 (\beta \vee r) \rceil),$$

and the number of parameters

$$s \leq 141(r + \beta + 1)^{3+r}N(m + 6),$$

such that

$$\|\hat{f} - f\|_{L^\infty([0,1]^r)} \leq (2K + 1)(1 + r^2 + \beta^2)6^rN2^{-m} + K3^\beta N^{-\frac{1}{6}}.$$

**Proof.** In this proof, all constructed networks take the form $\mathcal{F}(L, p, s) = \mathcal{F}(L, p, s, \infty)$, where $F = \infty$. Let $M$ be the largest integer such that $(M + 1)' \leq N$, and we define $L^* := (m + 5)\lceil \log_2 (\beta \vee r) \rceil$. With the help of Equations (A9) and (A10), and Lemma A4, we can add a hidden layer to the network $\text{Mon}_{m, \beta}^r$, resulting in a new network

$$Q_1 \in \mathcal{F}(2 + L^*, (r, 6\beta[C_{r, \beta}], \ldots, 6\beta[C_{r, \beta}, C_{r, \beta}, (M + 1)']),$$

such that $Q_1(x) \in [0,1]^{|M + 1}'$ and for any $x \in [0,1]^r$, we have

$$\left| Q_1(x) - \left( \frac{p_{n,f}(x)}{B} + \frac{1}{2} \right) \right|_{x_i \in D(M)} \leq \beta^2 2^{-m},$$

(A12)

where $B := [2Ke']$ and $e$ is the natural logarithm. According to Equation (A3), the number of nonzero parameters in network $Q_1$ is bounded by

$$6r(\beta + 1)C_{r, \beta} + 42(\beta + 1)^2C_{r, \beta}^2(L^* + 1) + C_{r, \beta}(M + 1)'$$

According to Lemma A6, the network $\text{Hat}'$ calculates the product $\prod_{j=1}^r (1/M - |x_j - x'_j|)_+$ with an error bounded by $r^2 2^{-m}$. It requires at most $49r^2 N(1 + L')$ active parameters. Now, consider the parallel network $(Q_1, \text{Hat}')$. Based on the definition of $C_{r, \beta}$ and the assumption on $N$, we observe that $C_{r, \beta} \leq (\beta + 1)' \leq N$. According to Lemma A6, networks $Q_1$ and $\text{Hat}'$ can be embedded into a joint network $(Q_1, \text{Hat}')$ with $2 + L^*$ hidden layers. The weight vector $(r, 6(r + \lceil \beta \rceil)N, \ldots, 6(r + \lceil \beta \rceil)N, 2(M + 1)')$ and all parameters are bounded by $1$. By using $C_{r, \beta} \vee (M + 1)' \leq N$, the bound on the number of nonzero parameters in the combined network $(Q_1, \text{Hat}')$ is

$$6r(\beta + 1)C_{r, \beta} + 42(\beta + 1)^2C_{r, \beta}^2(L^* + 1) + C_{r, \beta}(M + 1)' + 49r^2 N(1 + L^*)$$

$$\leq 49(r + \beta + 1)^2C_{r, \beta} N(1 + L^*)$$

$$\leq 98(r + \beta + 1)^{3+r}N(m + 5),$$

(A13)

where, for the last inequality, we use $C_{r, \beta} \leq (\beta + 1)'$, the definition of $L^*$, and the property that for any $x \geq 1$, we have $1 + \lceil \log_2(x) \rceil \leq 2 + \log_2(x) \leq 2(1 + \ln(x)) \leq 2x$.

Next, we pair the outputs of $Q_1$ and $\text{Hat}'$ corresponding to the $x_i$ term and apply the Mult$_m$ network described in Lemma A2 to each of the $(M + 1)'$ pairs. In the final layer, we sum all the terms together. According to Lemma A2, this requires at most $42(m + 5)(M + 1)' + (M + 1)' \leq 43(m + 5)N$ active parameters for the total $(M + 1)'$
multiplications. By using Lemmas A2 and A6, Equation (A12), and the triangle inequality, we can construct a network

\[ Q_2 \in \mathcal{F}(3 + (m + 5)(1 + \lceil \log_2(\beta \vee r) \rceil), (r, 6(r + \lceil \beta \rceil)N, \ldots, 6(r + \lceil \beta \rceil)N, 1)), \]

such that, for any \( \mathbf{x} \in [0,1]^r \), we have

\[
\left| Q_2(\mathbf{x}) - \sum_{\mathbf{x}_i \in \mathcal{D}(M)} \left( \frac{p_i^\beta f(\mathbf{x})}{B} + \frac{1}{2} \prod_{j=1}^{r} \left( \frac{1}{M} - |x_j - x_j'| \right) \right) \right| + \sum_{\mathbf{x}_i \in \mathcal{D}(M) : \|\mathbf{x} - \mathbf{x}_i\|_\infty \leq 1/M} (1 + r^2 + \beta^2)2^{-m} \leq (1 + r^2 + \beta^2)2^{-m},
\]

where the first inequality follows from the fact that the support of \((\mathbf{H}^i(\mathbf{x}))\mathbf{x}_i\) is contained within the support of \(\prod_{j=1}^{r} (1/M - |x_j - x_j'|)\), as stated in Lemma A6. Due to Equation (A3), the network \(Q_2\) has at most

\[
98(r + \beta + 1)^{3+r}N(m + 5) + 43(m + 5)N \leq 141(r + \beta + 1)^{3+r}N(m + 5).
\]

To obtain the network reconstruction of the function \(f\), it is necessary to apply scaling and shifting to the output terms. This is primarily due to the finite parameter weights in the network. We recall that \(B = [2Ke^r]\). The network \(\mathbf{x} \mapsto BM'\mathbf{x}\) belongs to the class \(\mathcal{F}(3, (1, M', 1, [2Ke^r], 1))\), where the shift vectors \(\mathbf{v}_j\) are all zero and all entries of the weight matrices \(W_j\) are equal to 1. Since \(N \geq (K + 1)e^r\), the number of parameters in this network is bounded by \(2M' + 2[2Ke^r] \leq 6N\). This implies the existence of a network in the class \(\mathcal{F}(4, (1, 2, 2M', 2, [2Ke^r], 1))\), which computes \(a \mapsto BM'(a - c)\), where \(c := 1/(2M')\). This network computes \((a - c)_+\) and \((c - a)_+\) in the first hidden layer, and then applies the network \(\mathbf{x} \mapsto BM'\mathbf{x}\) to these two units. In the output layer, the first value is subtracted from the second value. This requires at most \(6 + 12N\) active parameters.

Due to Equations (A11) and (A14), there exists a network \(Q_3\) in the following class

\[ \mathcal{F}(8 + (m + 5)(1 + \lceil \log_2(r \vee \beta) \rceil), (r, 6(r + \lceil \beta \rceil)N, \ldots, 6(r + \lceil \beta \rceil)N, 1)), \]

and for all \( \mathbf{x} \in [0,1]^r \), we have

\[
\left| Q_3(\mathbf{x}) - \sum_{\mathbf{x}_i \in \mathcal{D}(M)} p_i^\beta f(\mathbf{x}) \prod_{j=1}^{r} \left( 1 - M|x_j - x_j'| \right) \right| + \sum_{\mathbf{x}_i \in \mathcal{D}(M) : \|\mathbf{x} - \mathbf{x}_i\|_\infty \leq 1/M} (2K + 1)M^r(1 + r^2 + \beta^2)(2e)^r2^{-m}.
\]

Under the condition of Equation (A15), the bound for the nonzero parameters of \(Q_3\) is

\[ 141(r + \beta + 1)^{3+r}N(m + 6). \]

By constructing \((M + 1)^r \leq N \leq (M + 2)^r \leq (3M)^r\), it follows that \(M^{-\beta} \leq N^{-\beta/3\beta}\). Combined with Lemma A5, we have
\[ \| \hat{f} - f \|_{L^\infty([-1,1]^d)} = |Q_3 - f(x)| \]

\[ = |Q_3(x) - \sum_{x_i \in \mathcal{D}(M)} P_{x_i}^\beta f(x) \prod_{j=1}^{r} (1 - M|x_j - x_j^f|) + \]

\[ \sum_{x_i \in \mathcal{D}(M)} P_{x_i}^\beta f(x) \prod_{j=1}^{r} (1 - M|x_j - x_j^f|) - f(x) | \]

\[ \leq |Q_3(x) - \sum_{x_i \in \mathcal{D}(M)} P_{x_i}^\beta f(x) \prod_{j=1}^{r} (1 - M|x_j - x_j^f|) + \]

\[ \sum_{x_i \in \mathcal{D}(M)} P_{x_i}^\beta f(x) \prod_{j=1}^{r} (1 - M|x_j - x_j^f|) - f(x) | \]

\[ \leq (2K + 1) (1 + r^2 + \beta^2) (2e^r M'2^{-m} + K M^{-\beta} \]

\[ \leq (2K + 1) (1 + r^2 + \beta^2) 6^r N 2^{-m} + K 3^3 N^{-\beta}. \]

Thus, the result is proven.

Based on Theorem A1, we can now construct a network that approximates \( f = g_q \circ \ldots \circ g_0 \). In the first step, we show that \( f \) can always be represented as a composition of functions defined on hypercubes \([-1,1]^q\). As in the previous theorem, let \( g_{ij} \in \mathcal{C}_j^\beta([a_i, b_i]) \), and we assume \( K_i \geq 1 \) for \( i = 1, \ldots, q - 1 \). Define

\[ h_0 := \frac{g_0}{2 K_0} + \frac{1}{2}, \quad h_i := \frac{g_i(2K_i - 1 - K_i - 1)}{2 K_i^\beta} + \frac{1}{2} , \quad h_q = g_q \big((2K_q - 1 - K_q - 1)\big), \]

where \( 2K_i x - K_i - 1 \) means applying the transformation \( 2K_i x - K_i - 1 \) to all \( j \). It is evident that

\[ f = g_q \circ \ldots \circ g_0 = h_q \circ \ldots \circ h_0. \quad \text{(A16)} \]

From the definition of the Hölder ball \( \mathcal{C}_i^\beta(D, K) \), we can see that \( h_{0j} \) takes values in the interval \([0, 1]\).

\[ h_{0j} \in \mathcal{C}_0^\beta([0,1]^{h_0}, 1), h_{ij} \in \mathcal{C}_j^\beta([0,1]^{h_i}, (2K_j - 1)^{\beta_j}), \]

where, for \( i = 1, \ldots, q - 1 \), we have \( h_{ij} \in \mathcal{C}_j^\beta([0,1]^{h_i}, K_j (2K_j - 1)^{\beta_j}) \). Without loss of generality, we can always assume that the radius of the Hölder ball is at least 1, i.e., \( K_i \geq 1 \).

**Lemma A7.** Let \( h_{ij} \) be as defined above, with \( K_i \geq 1 \). Then, for any function \( \tilde{h}_i \in (h_{ij})_j^\beta \), where \( \tilde{h}_{ij} : [0, 1]^{h_i} \to [0, 1] \), we have

\[ \|h_q \circ \ldots \circ h_0 - \tilde{h}_0\|_{L^\infty[0,1]^d} \leq K_q \prod_{i=0}^{q-1} (2K_i)^\beta_{i+1} \sum_{i=0}^{q-1} \|h_i - \tilde{h}_i\|_{L^\infty[0,1]}^\beta_{i+1}. \]

**Proof.** Let \( H_i = h_i \circ \ldots \circ h_0 \) and \( \tilde{H}_i = \tilde{h}_i \circ \ldots \circ \tilde{h}_0 \). If \( Q_i \) is an upper bound of the Hölder seminorm of \( h_{ij} \) for \( j = 1, \ldots, d_i + 1 \), then, by the triangle inequality, we have

\[ |H_i(x) - \tilde{H}_i(x)|_\infty \leq |h_i \circ H_{i-1}(x) - h_i \circ \tilde{H}_{i-1}(x)|_\infty + |h_i \circ \tilde{H}_{i-1}(x) - \tilde{h}_i \circ \tilde{H}_{i-1}(x)|_\infty \]

\[ \leq Q_i |H_{i-1}(x) - \tilde{H}_{i-1}(x)|_\infty + \|h_i - \tilde{h}_i\|_{L^\infty[0,1]}^\beta_{i+1}. \]
Combining this with the inequality \((y + z)^a \leq y^a + z^a\), which holds for all \(y, z \geq 0\) and all \(a \in [0, 1]\), the lemma is proven. □

**Proof of Theorem 1.** Here, all \(n\) are assumed to be sufficiently large. Throughout the entire proof, \(C'\) is a constant that depends only on the variation of \((q, d, t, \beta, F)\). Combining Theorem 2 with the bounds on the depth \(L\) and network sparsity \(s\) assumed, for \(n \geq 3\), we have

\[
\frac{1}{4} \Delta_n(\hat{m}_n, m) - C' \phi_n L \ln^2 n \leq R(\hat{m}, m)
\]

\[
\leq 4 \inf_{m' \in \mathcal{F}(L, p, s, F)} \|m^* - m\|_{\infty}^2 + 4 \Delta_n(\hat{m}_n, m) + C' \phi_n L \ln^2 n,
\]

where, for the lower bound, we set \(\varepsilon = 1/2\), and for the upper bound, we set \(\varepsilon = 1\). We take \(C = 8C'\); then, when \(\Delta_n(\hat{m}_n, m) \geq C \phi n L \ln^2 n\), we have \(\frac{1}{8} \Delta_n(\hat{m}_n, m) \geq C' \phi n L \ln^2 n\). Substituting this into the left-hand side of Equation (A17), we obtain

\[
\frac{1}{4} \Delta_n(\hat{m}_n, m) - \frac{1}{8} \Delta_n(\hat{m}_n, m) \leq R(\hat{m}, m),
\]

that is,

\[
\frac{1}{8} \Delta_n(\hat{m}_n, m) \leq R(\hat{m}, m).
\]

Thus, the lower bound for Equation (8) is established.

To obtain upper bounds for Equations (7) and (8), it is necessary to constrain the approximation error. For this purpose, the regression function \(m\) is rewritten as Equation (A16), i.e., \(m = h_q \circ \ldots \circ h_0\), where \(h_i = (h_{ij})_j\) and \(h_{ij}\) is defined on \([0, 1]^{\beta_i}\) and maps to \([0, 1]\) for any \(i < q\).

Here, we apply Theorem A1 to each function \(h_{ij}\) separately. Let \(m = [\log_2 n]\) and consider the following.

\[
L' := 8 + ([\log_2 n] + 5)(1 + [\log_2 (t_i \lor \beta_i)]);
\]

this means that there exists a network

\[
\tilde{h}_{ij} \in \mathcal{F}(L', (t_i, 6(t_i + [\beta_i])N, \ldots, 6(t_i + [\beta_i])N, 1, s_i),
\]

where \(s_i \leq 141(t_i + \beta_i + 1)^{3+\varepsilon} N([\log_2 n] + 6)\), such that

\[
\|\tilde{h}_{ij} - h_{ij}\|_{L^\infty([0, 1])} \leq (2 Q_1 + 1) \left(1 + t_i^2 + \beta_i^2\right) 6^i N^{-1} + Q_2 3^{\beta_i} N^{-\beta_i/\nu},
\]

where \(Q_2\) is the Hölder norm upper bound of \(h_{ij}\). If \(i < q\), two additional layers \(1 - (1 - x)_+\) are applied to the output, requiring four additional parameters. The resulting network is denoted as

\[
h_{ij} := \mathcal{F}(L' + 2, (t_i, 6(t_i + [\beta_i])N, \ldots, 6(t_i + [\beta_i])N, 1, s_i + 4),
\]

and it is observed that \(\sigma(h_{ij}) = (h_{ij}(x) \lor 0) \land 1\). Since \(h_{ij}(x) \in [0, 1]\), we have

\[
\|\sigma(h_{ij}) - h_{ij}\|_{L^\infty([0, 1])} \leq \|\tilde{h}_{ij} - h_{ij}\|_{L^\infty([0, 1])}.
\]

If the network \(h_{ij}\) is parallelized, \(h_i = (h_{ij}^*)_{j = 1, \ldots, d_i + 1}\) belongs to the class

\[
\|\sigma(h_{ij}^*) - h_{ij}\|_{L^\infty([0, 1])} \leq \|\tilde{h}_{ij} - h_{ij}\|_{L^\infty([0, 1])}.
\]
where \( r_i := \max_j d_{i+1}(t_i + [\beta_i]) \). Finally, constructing the composite network \( m^* = \tilde{h}_{q1} \circ \sigma \left( h_{q-1}^* \circ \ldots \circ \sigma(h_{q}^*) \right) \), according to the construction rules in Appendix A, we can realize it in the following class:

\[
\mathcal{F} \left( E, (d, 6r_1N, \ldots, 6r_lN, 1), \sum_{i=0}^{q} d_{i+1}(s_i + 4) \right),
\]

where \( E := 3(q - 1) + \sum_{i=0}^{q} L_i \). By observation, there exists an \( A_n \) bounded by \( n \) such that

\[
E = A_n + \log_2 n \left( \sum_{i=0}^{q} \left( \log_2 (t_i \lor \beta_i) \right) + 1 \right).
\]

For all sufficiently large \( n \), utilizing the inequality \( [x] < x + 1 \), we have \( E \leq \sum_{i=0}^{q} \left( \log_2 4 + \log_2 (t_i \lor \beta_i) \right) \log_2 n \leq L \), according to Equation (A1), and for sufficiently large \( n \), the space defined in Equation (A20) can be embedded into \( \mathcal{F}(L, p, s) \), where \( L, p, s \) satisfies the assumptions of the theorem. We choose \( N = \left\lceil c \max_{i=0, \ldots, q} n \right\rceil \) with a sufficiently small constant \( c > 0 \), depending only on \( q, d, t, \beta \). Combining Theorem A1 with Equations (A18) and (A19), we have

\[
\begin{align*}
\inf_{f_\ast \in \mathcal{F}(L, p, s)} \| m_\ast - m \|^2_{\sigma_0} \\
= \inf_{f_\ast \in \mathcal{F}(L, p, s)} \| \tilde{h}_{q1} \circ \sigma(h_{q-1}^* \circ \ldots \circ \sigma(h_{q}^*) - h_{q} \circ \ldots \circ h_0) \|^2_{\sigma_0} \leq \inf_{f_\ast \in \mathcal{F}(L, p, s)} \| \tilde{h}_{q1} \circ h_{q-1} \circ \ldots \circ h_0 - h_{q} \circ \ldots \circ h_0 \|^2_{\sigma_0} \\
\leq K_\eta \prod_{\ell=0}^{q-1} (2K_l)^{\bar{\beta}_{\ell+1}} \sum_{i=0}^{q} \| h_i - \tilde{h}_i \|^{2}_{\sigma_0} \\
\leq \left( (2Q_i + 1) \left( 1 + \bar{t}_{\ell}^2 + \beta_{\ell}^2 \right) 6^11N^{-1} + Q_i3^\delta N^{-\beta_{\ell}^1} \right)^2 \leq C' \max_{i=0, \ldots, d} N^{-\frac{2\delta_{\ell}^1}{1 - \bar{t}_{\ell}^2}}. \tag{A21}
\end{align*}
\]

For the approximation error in Equation (A17), we need to find a network function \( m \) bounded by the supnorm of \( F \). According to the previous inequalities, there exists a sequence of functions \( (\hat{m}_n) \) such that, for all sufficiently large \( n \), \( \hat{m}_n \in \mathcal{F}(L, p, s) \), and \( |\hat{m}_n - m|^2_{\sigma_0} \leq 2C \max_{i=0, \ldots, q} \beta_{\ell}^2 n^{-2} \). Let us define \( m^*_n = \hat{m}_n(|m|_{\sigma_0} / |m_\ast|_{\sigma_0} \land 1) \). Then, \( |m^*_n|_{\sigma_0} \leq |m|_{\sigma_0} = |g_q|_{\sigma_0} \leq K \leq F \), where the last inequality is based on Assumption (1). Additionally, \( m^*_n \in \mathcal{F}(L, p, s, F) \). We can denote \( m_n^* - m = (m^*_n - \hat{m}_n) + (\hat{m}_n - m) \), and we have \( |m^*_n - m|_{\sigma_0} \leq |m^*_n - \hat{m}_n|_{\sigma_0} + |\hat{m}_n - m|_{\sigma_0} \), which implies \( |m^*_n - m|_{\sigma_0} \leq 2|\hat{m}_n - m|_{\sigma_0} \). This shows that if we take the lower bound on a smaller space \( \mathcal{F}(L, p, s, F) \), then Equation (A21) also holds. Combining this with the upper bound of Equation (A17), we obtain when \( \Delta_n(\hat{m}_n, m) \leq C\phi_F L \ln^2 n \),

\[
R(\hat{m}_n, m) \leq C' \phi_F L \ln^2 n,
\]

and when \( \Delta_n(\hat{m}_n, m) \geq C\phi_F L \ln^2 n \),

\[
R(\hat{m}_n, m) \leq C' \Delta_n(\hat{m}_n, m);
\]
therefore, the upper bounds in Equations (7) and (8) hold for any constant $C > 0$. This completes the proof.

We begin by utilizing several oracle inequalities for the least squares estimators, as presented in Gyo et al. [16–20]. However, these inequalities assume bounded response variables, which are violated in the nonparametric regression model with Gaussian measurement noise. Additionally, we provide a lower bound for the risk and offer proof that can be easily generalized to any noise distribution. Let $\mathcal{N}(\delta, \mathcal{F}, \| \cdot \|_\infty)$ be the covering number, which represents the minimum number of $\| \cdot \|_\infty$ balls with radius $\delta$ needed to cover $\mathcal{F}$ (where the center does not necessarily have to be within $\mathcal{F}$). □

Lemma A8. We consider the nonparametric regression model in $d$-dimensional variables given by Equation (2) with an unknown regression function $m$. Let $\hat{m}$ be an arbitrary estimator taking values in $\mathcal{F}$. Let us define

$$\Delta_n := \Delta_n(\hat{m}_n, m, \mathcal{F}) := E_m \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - \hat{m}(X_i))^2 - \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2 \right].$$

For $F \geq 1$ and assuming $\{ m \} \cup \mathcal{F} \subset \{ f : [0, 1]^d \to [-F, F] \}$. If $\mathcal{N}_n := \mathcal{N}(\delta, \mathcal{F}, \| \cdot \|_\infty) \geq 3$, for any $\delta, \epsilon \in (0, 1]$, then

$$(1 - \epsilon)^2 \Delta_n - F^2 \frac{18 \ln \mathcal{N} + 76}{n} - 38\delta F \leq R(\hat{m}, m) \leq (1 + \epsilon)^2 \left[ \inf_{f \in \mathcal{F}} E \left[ (f(X) - m(X))^2 \right] + F^2 \frac{18 \ln \mathcal{N} + 72}{n} + 32\delta F + \Delta_n \right].$$

Proof. Throughout the proof, let $E = E_m$. Define $\| g \|^2 := \frac{1}{n} \sum_{i=1}^{n} g(X_i)^2$. For any estimator $\hat{m}$, we introduce the empirical risk $\hat{R}_n(\hat{m}, m) := E \left[ \| \hat{m} - m \|^2 \right]$. Step 1: We show that the upper bound holds under the restriction $\ln \mathcal{N}_n \leq n$. Since $R(\hat{m}, m) \leq 4F^2$, the upper bound naturally holds when $\ln \mathcal{N}_n \geq n$. In this case, let $\hat{m} \in \arg \min_{f \in \mathcal{F}} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2$ be a global risk minimizer. We observe that

\begin{align*}
\hat{R}_n(\hat{m}, m) - \hat{R}_n(\hat{m}_n, m) &= E \left[ \| \hat{m} - m \|^2 \right] - E \left[ \| \hat{m}_n - m \|^2 \right] \\
&= E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (\hat{m} - m)^2 \right] - E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (\hat{m}_n - m)^2 \right] \\
&= E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (\hat{m} - Y_i + R_i)^2 \right] - E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (\hat{m}_n - Y_i + R_i)^2 \right] \\
&= \Delta_n + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m} - Y_i) \right] + \left( \frac{1}{n} \sum_{i \in \Lambda_n} R_i^2 \right) \\
&\quad - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m}_n - Y_i) \right] + \left( \frac{1}{n} \sum_{i \in \Lambda_n} R_i^2 \right) \\
&= \Delta_n + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m} - m + R_i) \right] - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m}_n - m + R_i) \right] \\
&= \Delta_n + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m} \right] - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i m \right] + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}_n \right] - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i m \right] \\
&= \Delta_n - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}(X_i) \right] - E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}_n(X_i) \right].
\end{align*}

From this equation, we see that $\Delta_n \leq 8F^2$, which implies a lower bound on the logarithm $\ln \mathcal{N}_n \geq n$ in the argument.
Therefore, we assume that \( \ln \mathcal{N}_n \leq n \). The proof is divided into four parts, denoted as (I)–(IV).

(I) Establishing a connection between risk \( R(\hat{m}, m) = E \left[ (\hat{m}(X) - m(X))^2 \right] \) and its empirical counterpart \( \hat{R}_n(\hat{m}, m) \) through inequalities

\[
(1 - \varepsilon) \hat{R}_n(\hat{m}, m) = \frac{F^2}{n \varepsilon} (15 \ln \mathcal{N}_n + 75) - 26\delta F
\]

\[
\leq R(\hat{m}, m) \leq (1 + \varepsilon) \left( \hat{R}_n(\hat{m}, m) + (1 + \varepsilon) \frac{F^2}{n \varepsilon} (12 \ln \mathcal{N}_n + 70) + 26\delta F \right).
\]

(II) For any estimate \( \hat{m} \) taking values in \( \mathcal{F} \), we know that

\[
\left| E \left[ \frac{2}{n} \sum_{i \in \mathcal{N}_n} R_i \hat{m}(X_i) \right] \right| \leq 2 \sqrt{\frac{\hat{R}_n(\hat{m}, m)(3 \ln \mathcal{N}_n + 1)}{n}} + 6\delta.
\]

(III) We have

\[
\hat{R}_n(\hat{m}, m) \leq (1 + \varepsilon) \left[ \inf_{f \in \mathcal{F}} E \left[ (f(X) - m(X))^2 \right] + 6\delta + F^2 \frac{6 \ln \mathcal{N}_n + 2}{n \varepsilon} + \Delta_n \right].
\]

(IV) We have

\[
\hat{R}_n(\hat{m}, m) \geq (1 - \varepsilon) \left( \Delta_n - \frac{3 \ln \mathcal{N}_n + 1}{n \varepsilon} - 12\delta \right).
\]

Since \( F \geq 1 \), the lower bound of the lemma can be obtained by combining (I) and (IV), while the upper bound can be obtained from (I) and (III).

(I) Given a minimum \( \delta \)-covering of \( \mathcal{F} \), let \( f_j \) represent the centers of the balls. According to the construction, there exists a random \( j^* \) such that \( |\hat{m} - f_{j^*}|_\infty \leq \delta \). Without loss of generality, we can assume that \( |f_j|_\infty \leq F \). The random variables \( X'_i, i \in 1, 2, \ldots, n \) have the same distribution as \( X \) and are independent of \( (X_i), i \in 1, 2, \ldots, n \). We can use \( |f_j|_\infty, |m|_\infty, \delta \leq F \)

\[
|R(\hat{m}, m) - \hat{R}_n(\hat{m}, m)|
\]

\[
= \left| E \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( \hat{m}(X'_i) - m(X'_i) \right)^2 - \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( \hat{m}(X_i) - m(X_i) \right)^2 \right|
\]

\[
= \left| E \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( \hat{m}(X'_i) - f_{j^*}(X'_i) + f_{j^*}(X'_i) - m(X'_i) \right)^2 \right|
\]

\[
- \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( \hat{m}(X_i) - f_{j^*}(X_i) + f_{j^*}(X_i) - m(X_i) \right)^2 \right|
\]

\[
= \left| E \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( 2(\hat{m}(X'_i) - f_{j^*}(X'_i))(f_{j^*}(X'_i) - m(X'_i)) + (f_{j^*}(X'_i) - m(X'_i))^2 \right) \right|
\]

\[
- \frac{1}{n} \sum_{i \in \mathcal{N}_n} \left( 2(\hat{m}(X_i) - f_{j^*}(X_i))(f_{j^*}(X_i) - m(X_i)) + (f_{j^*}(X_i) - m(X_i))^2 \right) \right|
\]

\[
\leq E \left[ \frac{1}{n} \sum_{i \in \mathcal{N}_n} g_{j^*}(X'_i, X'_i) \right] + 9\delta F,
\]

where \( g_j(X_i, X'_i) := (f_{j^*}(X'_i) - m(X'_i))^2 - (f_{j^*}(X_i) - m(X_i))^2 \). We replace \( f_j \) with \( f_{j^*} \), and we define \( g_j \) by using the same method. Similarly, we set \( r_j := \sqrt{n^{-1} \ln \mathcal{N}n} \vee E^{1/2} \left[ (f_j(X) - m(X))^2 \right] \), and define \( r^* \) as \( r_j \) when \( j = j^* \).
\[ r^* := \sqrt{n^{-1} \ln \mathcal{A}_n} \vee E^{1/2} \left[ (f_j(X) - m(X))^2 \mid (X_i, Y_i) \right] \]
\[ \leq \sqrt{n^{-1} \ln \mathcal{A}_n} + E^{1/2} \left[ (\tilde{m}(X) - m(X))^2 \mid (X_i, Y_i) \right] + \delta. \]

In the last part, we use the triangle inequality and \( |f_j - \tilde{m}| \leq \delta \).

For random variables \( U \) and \( T \), the Cauchy–Schwarz inequality states that \( E[UT] \leq E^{1/2}[U^2] E^{1/2}[T^2] \). Let

\[ U = E^{1/2} \left[ (\tilde{m}(X) - m(X))^2 \mid (X_i, Y_i) \right], \]

and

\[ T := \max_j \left| \sum_{j \in \Lambda_n} g_j(X_i, X'_i) / (r_jF) \right|. \]

By using \( E[U^2] = R(\tilde{m}, m) \), we have

\[ \left| R(\tilde{m}, m) - \tilde{R}_n(\tilde{m}, m) \right| \]
\[ \leq E \left( \frac{F}{n} T_{r_j} \right) + 9\delta F \]
\[ = E \left( \frac{F}{n} T(U - U + r_j) \right) + 9\delta F \]
\[ = E \left( \frac{F}{n} T(U + T(r_j - U)) \right) + 9\delta F \]
\[ \leq \frac{F}{n} R(\tilde{m}, m)^{1/2} E^{1/2} \left[ T^2 \right] + \frac{F}{n} \left( \sqrt{\frac{\ln \mathcal{A}_n}{n}} + \delta \right) E[T] + 9\delta F. \]

(A23)

Observing that \( E[g_j(X_i, X'_i)] = 0 \),

\[ |g_j(X_i, X'_i)| = \left| (f_j(X'_i) - m(X'_i))^2 - (f_j(X_i) - m(X_i))^2 \right| \leq 4f_j(X_i)m(X_i) \leq 4t^2 \]

and

\[ \text{Var}(g_j(X_i, X'_i)) = 2 \text{Var} \left( (f_j(X_i) - m(X_i))^2 \right) \]
\[ \leq 2E \left[ (f_j(X_i) - m(X_i))^4 \right] \leq 8F^2 r_j^2. \]

Bernstein’s inequality states that for independent and centered random variables \( U_i, \ldots, U_n \), if \( |U_i| \leq M \), then it holds true that [21]

\[ P \left( \left| \sum_{i \in \Lambda_n} U_i \right| \geq t \right) \leq 2 \exp \left( -t^2 / [2Mt/3 + 2 \sum_{i \in \Lambda_n} \text{Var}(U_i)] \right). \]

Combining Bernstein’s inequality and the bound argument, we obtain

\[ P(T \geq t) \leq 1 \land 2A_n \max_j \left( \frac{t^2}{8t/(3r_j) + 16n} \right), \]

and since \( r_j \geq \sqrt{n^{-1} \ln \mathcal{A}_n} \), for all \( t \geq 6\sqrt{n \ln \mathcal{A}_n} \), we have

\[ P(T \geq t) \leq 2A_n \exp \left( -\frac{6t \sqrt{n \ln \mathcal{A}_n}}{48 \sqrt{n \ln \mathcal{A}_n} / \left( 3 \sqrt{n^{-1} \ln \mathcal{A}_n} \right) + 16n} \right) \]
\[ \leq 2A_n \exp \left( -3t \sqrt{\ln \mathcal{A}_n} / (16\sqrt{n}) \right). \]
Thus, for large values of \( t \), the denominator in the exponential is dominated by the first term. We have

\[
E[T] = \int_0^\infty P(T \geq t) \, dt = \int_0^{6\sqrt{n \ln N_n}} P(T \geq t) \, dt + \int_0^\infty P(T \geq t) \, dt \\
\leq 6\sqrt{n \ln N_n} + \int_0^\infty 2M_n \exp \left( -\frac{3t \ln N_n}{16\sqrt{n}} \right) \, dt \\
\leq 6\sqrt{n \ln N_n} + \frac{32}{3} \sqrt{\frac{n}{\ln N_n}}.
\]

According to the assumption, \( N_n \geq 3 \); hence, \( \ln N_n \geq 1 \). By using a similar approach to the upper bound for \( E[T] \), we can obtain the quadratic case.

\[
E[T^2] = \int_0^\infty P(T \geq u) \, du = \int_0^\infty P(T \geq \sqrt{n}) \, du \\
\leq 36n \ln N_n + \int_0^\infty 2M_n \exp \left( -\frac{3\sqrt{n \ln N_n}}{16\sqrt{n}} \right) \, du \\
\leq 36n \ln N_n + 2^8 n.
\]

Step 2: The identity \( \int_0^\infty e^{-\sqrt{n}u} \, du = 2 \int_0^\infty e^{-\sqrt{n}s} \, ds = 2(ba + 1)e^{-ba} / a^2 \). In this case, we set \( a = \frac{3\sqrt{n \ln N_n}}{16\sqrt{n}} \) and \( b = 6\sqrt{n \ln N_n} \) to obtain the aforementioned inequality.

\[
\left| R(\hat{\mu}, \mu) - \hat{R}_n(\hat{\mu}, \mu) \right| \\
\leq F \left( \frac{F}{n} R(\hat{\mu}, \mu) \right)^{1/2} \left( 36n \ln N_n + 2^8 n \right)^{1/2} + F \left( 6\sqrt{n \ln N_n} + \frac{32}{3} \sqrt{\frac{n}{\ln N_n}} \right) + 9\delta F \quad (A24)
\]

Let \( a, b, c, d \) be positive real numbers such that \( |a - b| \leq 2\sqrt{ac} + d \). We have

\[
b - 2\sqrt{a} - d \leq a \leq 2\sqrt{a} + b + d,
\]

\[
b - d - 2\left( \sqrt{\frac{\epsilon}{1 - \epsilon}} \sqrt{\alpha} \right) \left( \sqrt{\frac{1 - \epsilon}{\epsilon}} \right) \leq a \leq b + d + 2\left( \sqrt{\frac{\epsilon}{1 + \epsilon}} \sqrt{\alpha} \right) \left( \sqrt{\frac{1 + \epsilon}{\epsilon}} \right),
\]

Consequently, for any \( \epsilon \in (0, 1] \), we have the following.

\[
(1 - \epsilon)b - d - \frac{\epsilon^2}{\epsilon} \leq a \leq (1 + \epsilon)(b + d) + \frac{(1 + \epsilon)^2}{\epsilon} c^2. \quad (A25)
\]

According to Equation (A24), we take \( a = R(\hat{\mu}, \mu) \), \( b = \hat{R}_n(\hat{\mu}, \mu) \), and \( c = F(9n \ln N_n + 64n)^{1/2} / n, d = F(6 \ln N_n + 11) / n + 26\delta F \). Substituting \( a, b, c, d \) into Equation (A25) (denoted as (I)), we complete the proof of (I).
(II) Given an estimation \( \hat{f} \) that takes values in \( \mathcal{F} \), let \( f' \) be such that \( |\hat{m} - f'| \leq \delta \). Then, \[ E \left[ \sum_{i \in \Lambda_n} R_i \left( \hat{m}(X_i) - f'_i(X_i) \right) \right] \leq \delta E \left[ \sum_{i \in \Lambda_n} |R_i| \right] \leq n\delta. \] Since \( E[R_i \ m(X_i)] = \text{cov}(X_i, R_i) \neq 0 \), we have

\[
\left| E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}(X_i) \right] \right| \\
\leq E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \left( \hat{m}(X_i) - m(X_i) + m(X_i) \right) \right] \\
\leq E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \left( \hat{m}(X_i) - m(X_i) \right) \right] + \left| E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i m(X_i) \right] \right| \\
= E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \left( \hat{m}(X_i) - f'_i(X_i) + f'_i(X_i) - m(X_i) \right) \right] + \left| E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i m(X_i) \right] \right| \\
\leq 2\delta + \frac{2}{n} \text{cov}(X_i, R_i) + \left| E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \left( f'_i(X_i) - m(X_i) \right) \right] \right| \\
\leq 2\delta + \frac{2}{\sqrt{n}} E \left[ \left( \left\| \hat{m} - m \right\|_n + \delta \right) \left| \xi_f \right| \right],
\]

where
\[
\xi_f := \frac{\text{cov}(X_i, R_i) + \sum_{i \in \Lambda_n} R_i \left( f'_i(X_i) - m(X_i) \right)}{\sqrt{n} \left\| f'_i - m \right\|_n}.
\]

Under the condition \( (X_i)_{i \in \mathbb{N}} \sim \mathcal{N}(0, 1) \). According to Lemma A9, we obtain \( E \left[ \xi_f^2 \right] \leq E \left[ \max_j \xi_j^2 \right] \leq 3 \ln N_n + 1 \). By using Cauchy–Schwarz, we have

\[
E \left[ \left( \left\| \hat{m} - m \right\|_n + \delta \right) \left| \xi_f \right| \right] \leq \left( \hat{R}_n(\hat{m}, m)^{1/2} + \delta \right) \sqrt{3 \ln N_n + 1} \tag{A27}
\]

Since \( \ln N_n \leq n \), we have \( 2n^{-1/2} \delta \sqrt{3 \ln N_n + 1} \leq 4\delta \). Combining Equations (A26) and (A27), we have

\[
\left| E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}(X_i) \right] \right| \leq 2\delta + \frac{2}{\sqrt{n}} E \left[ \left( \left\| \hat{m} - m \right\|_n + \delta \right) \left| \xi_f \right| \right] \\
\leq 2\delta + \frac{2}{\sqrt{n}} \left( \hat{R}_n(\hat{m}, m)^{1/2} + \delta \right) \sqrt{3 \ln N_n + 1} \\
\leq 2\sqrt{\frac{\hat{R}_n(\hat{m}, m) (3 \ln N_n + 1)}{n}} + 6\delta.
\]

(II) is proven.

(III) For any fixed \( f \in \mathcal{F} \), we have \( E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - \hat{m}(X_i))^2 \right] \leq E \left[ \frac{1}{n} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2 \right] + \Delta_n \). Since \( X_i \overset{d}{=} X \) and \( f \) are deterministic, we have
\[ E[\|f - m\|^2_n] = E \left[ (f(X) - m(X))^2 \right]. \] Since \( E[R_i m(X_i)] = \text{cov}(X_i, R_i) \neq 0 \), we have

\[ \hat{R}_n(\hat{m}, m) = E \left[ \|\hat{m} - m\|^2_n \right] = E \left[ \|\hat{m} - Y_i + R_i\|^2_n \right] \]

\[ \leq E \left[ \|\hat{m} - Y_i\|^2_n \right] + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m} - Y_i) \right] \]

\[ \leq E \left[ \|f - m\|^2_n \right] + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i(\hat{m} - m) \right] + \Delta_n \]

\[ \leq E \left[ (f(X) - m(X))^2 \right] + \Delta_n + E \left[ \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m} \right] + \frac{2}{n} \text{cov}(X_i, R_i) \]

\[ \leq E \left[ (f(X) - m(X))^2 \right] + \frac{2}{n} \text{cov}(X_i, R_i) + 2 \sqrt{\frac{\hat{R}_n(\hat{m}, m)(3 \ln \mathcal{N}_n + 1)}{n}} + 6 \delta + \Delta_n. \]

By setting \( a := \hat{R}_n(\hat{m}, m) \), \( b := 0 \), \( c := \sqrt{(3 \ln \mathcal{N}_n + 1)} / n \), \( d := E \left[ (f(X) - m(X))^2 \right] + 6 \delta + \Delta_n + \frac{2}{n} \text{cov}(X_i, R_i) \) in Equation (A25), we obtain the result for (III).

(IV) Let \( \hat{m} \in \arg \min_{f \in \mathcal{F}} \sum_{i \in \Lambda_n} (Y_i - f(X_i))^2 \) be the empirical risk minimizer. By using Equation (A22), (II), and \((1 - \epsilon)/\epsilon + 1 = 1/\epsilon\), we have

\[ \hat{R}_n(\hat{m}, m) - \hat{R}_n(\hat{m}, m) \]

\[ \geq \Delta_n - \frac{2}{n} \sum_{i \in \Lambda_n} R_i \hat{m}(X_i) \]

\[ \geq \Delta_n - 2 \sqrt{\frac{\hat{R}_n(\hat{m}, m)(3 \ln \mathcal{N}_n + 1)}{n}} - 2 \sqrt{\frac{\hat{R}_n(\hat{m}, m)(3 \ln \mathcal{N}_n + 1)}{n}} - 12 \delta \]

\[ \geq \Delta_n - \frac{\epsilon}{1 - \epsilon} \hat{R}_n(\hat{m}, m) - \frac{n \epsilon}{3 \ln \mathcal{N}_n + 1} - 12 \delta. \]

After rearranging, we have \( \hat{R}_n(\hat{m}, m) \geq (1 - \epsilon) \left( \Delta_n - \frac{3 \ln \mathcal{N}_n + 1}{n \epsilon} - 12 \delta \right) \), which completes the proof of (IV). \( \square \)

**Lemma A9.** Let \( \eta_j \sim \mathcal{N}(0, 1) \), and then \( E \left[ \max_{j=1, \ldots, M} \eta_j^2 \right] \leq 3 \ln M + 1. \)

**Proof.** Let \( Z = \max_{j=1, \ldots, M} \eta_j^2 \). Since \( Z \leq \sum_j \eta_j^2 \), and we have \( E[Z] \leq M \). For \( M \in \{1, 2, 3\} \), it is evident that \( M \leq 3 \ln (M) + 1 \) holds. Therefore, we consider the case when \( M \geq 4 \).

By using Mill’s ratio, we obtain \( P \left( |\eta_1| \geq \sqrt{T} \right) = 2P \left( \eta_1 \geq \sqrt{T} \right) \leq 2e^{-T/2}/(\sqrt{2\pi T}) \). For any \( T \), we have

\[
E[Z] = \int_0^\infty P(Z \geq t)dt \leq T + \int_T^\infty P(Z \geq t)dt \leq T + M \int_T^\infty P(\eta_1^2 \geq t)dt \]

\[
\leq T + M \int_T^\infty \frac{2}{\sqrt{2\pi t}} e^{-t/2}dt \leq T + \frac{2M}{\sqrt{2\pi T}} \int_T^\infty e^{-t/2}dt \]

\[
= T + \frac{4M}{\sqrt{2\pi T}} e^{-T/2}. \]
For $T = 2 \ln M$ and $M \geq 4$, we have

$$E[Z] \leq 2 \ln M + 2/\sqrt{\pi \ln M} \leq 3 \ln M + 1.$$ 

Since $2/\sqrt{\pi \ln M} \leq \ln M + 1$, we can deduce that $2/\sqrt{\pi} \leq (\ln M)^2 + \ln M$. Considering that the function is monotonically increasing with respect to $M$, it holds for all $M \geq 4$. \hfill \Box

Lemma A10. If $V := \prod_{\ell=0}^{L+1} (p_\ell + 1)$, then for any $\delta > 0$, we have

$$\ln \mathcal{M}(\delta, \mathcal{F}(L, p, s, \infty), \| \cdot \|_\infty) \leq (s + 1) \ln \left( 2\delta^{-1}(L + 1)V^2 \right).$$

Proof. Given a network

$$f(x) = W_1 \sigma_{v_1} W_{L-1} \sigma_{v_{L-1}} \cdots W_0 \sigma_{v_0},$$

define $k \in \{1, \ldots, L\}$, $A_k^+ f : [0, 1]^d \to \mathbb{R}^{p_k}$,

$$A_k^+ f(x) = \sigma_{v_k} W_{k-1} \sigma_{v_{k-1}} \cdots W_1 \sigma_{v_1},$$

and $A_k^- f : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_{k+1}}$,

$$A_k^- f(y) = W_1 \sigma_{v_1} W_{L-1} \sigma_{v_{L-1}} \cdots W_{k-1} \sigma_{v_{k-1}} W_{k-1} y.$$

Let $A_0^+ f(x) = A_{L+2}^- f(x) = x$, and we note that for $f \in \mathcal{F}(L, p)$, we have $|A_k^+ f(x)|_\infty \leq \prod_{\ell=0}^{k-1} (p_\ell + 1)$. For a multivariate function $h(\cdot)$, $h(\cdot)$ is said to be a Lipschitz function if, for all $x, y$ in its domain, $|h(x) - h(y)|_\infty \leq L|x - y|_\infty$, where the smallest $L$ is the Lipschitz constant. The combination of two Lipschitz functions with Lipschitz constants $L_1$ and $L_2$ results in a new Lipschitz function with a Lipschitz constant of $L_1 L_2$. Therefore, the Lipschitz constant of $A_k^+ f$ is bounded by $\prod_{\ell=k}^{L+1} p_\ell$. Given $\epsilon > 0$, let $f, f \in \mathcal{F}(L, p, s)$ be two network functions with parameters differing from each other by at most $\epsilon$. Let $f$ have parameters $(v_k, W_k)$ and $f^* \in \mathcal{F}(L, p, s)$ have parameters $(v_k^*, W_k^*)$. Then, we have

$$|f(x) - f^*(x)| = |W_1 \sigma_{v_1} W_{L-1} \sigma_{v_{L-1}} \cdots W_0 \sigma_{v_0} - W_1 \sigma_{v_1} W_{L-1} \sigma_{v_{L-1}} \cdots W_0 \sigma_{v_0}|$$

$$= |A_k^+ f(\sigma_{v_k} W_{k-1} \sigma_{v_{k-1}} A_{k-1}^+ f(x)) - A_k^+ f^*(\sigma_{v_k} W_{k-1} \sigma_{v_{k-1}} A_{k-1}^+ f^*(x))|$$

$$\leq \sum_{k=1}^{L+1} |A_k^+ f(\sigma_{v_k} W_{k-1} A_{k-1}^+ f(x)) - A_k^+ f^*(\sigma_{v_k} W_{k-1} A_{k-1}^+ f^*(x))||$$

$$\leq \sum_{k=1}^{L+1} \left( \prod_{\ell=k}^{L+1} p_\ell \right) |\sigma_{v_k} W_{k-1} A_{k-1}^+ f(x) - \sigma_{v_k} W_{k-1} A_{k-1}^+ f^*(x)|_\infty$$

$$\leq \sum_{k=1}^{L+1} \left( \prod_{\ell=k}^{L+1} p_\ell \right) \left( |W_{k-1} A_{k-1}^+ f(x) - \psi_k| + |W_{k-1} A_{k-1}^+ f^*(x) - \psi_k^*| \right)_\infty$$

$$\leq \epsilon \sum_{k=1}^{L+1} \left( \prod_{\ell=k}^{L+1} p_\ell \right) (p_{k-1} |A_{k-1}^+ f(x)|_\infty + 1)$$

$$\leq \epsilon \sum_{k=1}^{L+1} \left( \prod_{\ell=k}^{L+1} p_\ell + 1 \right)$$

$$\leq \epsilon V(L + 1).$$
The final step uses $V := \prod_{t=0}^{L+1} (p_t + 1)$. Therefore, according to Equation (A3), the total number of parameters is bounded by $T := \sum_{t=0}^{L} (p_t + 1) p_{t+1} \leq (L + 1) 2^{-L} \prod_{t=0}^{L+1} (p_t + 1) \leq V$, and there are $\left( \frac{T}{s} \right) \leq V^s$ combinations to select $s$ nonzero parameters.

Since all parameters are bounded by 1 in absolute value, we can discretize the nonzero parameters by using a grid size of $\delta/(2(L + 1)V)$, and the covering number

$$\mathcal{N}(\delta, \mathcal{F}(L, p, s, \infty), \| \cdot \|_{\infty}) \leq \sum_{s=1}^{\infty} \left( 2\delta^{-1}(L + 1) V^2 \right)^s \leq \left( 2\delta^{-1}(L + 1) V^2 \right)^{s+1}.$$  

Taking the logarithm yields the proof.

Note 1: Similarly, applying Equation (A4) to Lemma A10 gives

$$\ln \mathcal{N}(\delta, \mathcal{F}(L, p, s, \infty), \| \cdot \|_{\infty}) \leq (s + 1) \ln \left( 2^{2L+5} \delta^{-1}(L + 1) p_0^2 p_{L+1} s^{2L} \right).$$  

Proof of Theorem 2. Let $\delta = 1/n$. The proof follows directly from Lemmas A8 and A10, and Note 1.

Proof of Theorem 3. In this proof, we define $\| \cdot \| = \| \cdot \|_{L^2([0,1]^d)}$. Let us assume that there exist positive constants $\gamma \leq \Gamma$ such that the Lebesgue density of $X$ over $[0,1]^d$ is bounded below by $\gamma$ and above by $\Gamma$. For this particular design, we have $R(\hat{m}_n, m) \geq \gamma |\hat{m}_n - m|^2$. Let $P_f$ represent the data mechanism in the nonparametric regression model given by Equation (13). For the Kullback–Leibler divergence, we have $KL(P_f, P_\gamma) = nE[(f(X_1) - \gamma(X_1))^2] = \Gamma|f - \gamma|^2$. In Alexandre’s work [22], Theorem 2.7 states that if, for $M \geq 1$ and $\kappa > 0$, we have $f_0, \ldots, f_M \in \mathcal{G}(\xi, d, t, \beta, K)$, then

(i) $\sup_{j \leq M} \| f_j \|_{\infty} \leq \kappa \sqrt{M}$, where $0 \leq j \leq M$;

(ii) $\| \sum_{j=1}^M (f_j - f_n) \|_{2} \leq M \ln(M)/(9\Gamma)$.

Then, there exists a positive constant $c = c(\kappa, \gamma)$ such that

$$\inf_{\hat{m}_n} \sup_{m \in \mathcal{F}(\xi, d, t, \beta, K)} R(\hat{m}_n, m) \geq c \phi_n.$$  

In the next step, we construct functions $f_0, \ldots, f_M \in \mathcal{G}(\xi, d, t, \beta, K)$ satisfying (i) and (ii). We define

$$t^* \in \arg \min_{t \in [0,\ldots,d]} \beta_t^*/(2\beta_t^* + t).$$

The exponent $t^*$ determines the rate of estimation, i.e., $\phi_n = n^{-2\beta^*/(2\beta^* + t^*)}$. For convenience, we denote $\beta^* := \beta_1^*, \beta^{**} := \beta_2^*$, and $t^* := t_1$. We note the distinction between $\beta^*$ and $\beta^{**}$. Let $K \in L^2([\mathbb{R}]) \cap C_1^1(\mathbb{R})$ be supported on $[0,1]$. It is easy to see that such a function $K$ exists. Furthermore, we define $m_n := \left\lfloor \rho n^{1/(2\beta^* + t^*)} \right\rfloor$ and $h_n := 1/m_n$, where $\rho$ is a constant chosen so that $m_n^{2\beta^* + t^*} \leq 1/(72\Gamma |K|^B 2^{2m})$, with $B := \prod_{\ell=t^*}^q (\beta_\ell \wedge 1)$. For any $u = (u_1, \ldots, u_t) \in \mathcal{U}_n := \{ (u_1, \ldots, u_t) : u_i \in \{0, h_n, 2h_n, \ldots, (m_n - 1)h_n \} \}$, we define

$$\psi_u(x_1, \ldots, x_t) := h_n^{t^*} \prod_{j=1}^t \left( \frac{x_j - u_j}{h_n} \right),$$

and for any $\alpha$ and $|\alpha| < \beta^*$, we have $|\partial^\alpha \psi_u|_{\infty} \leq 1$, by using the fact that $K \in C_1^{\beta^*}(\mathbb{R})$.

For $\alpha = (\alpha_1, \ldots, \alpha_t)$ with $|\alpha| = |\beta^*|$, the triangle inequality and the property $K \in C_1^{\beta^*}(\mathbb{R})$ give
We frequently use \( \beta^* \) to ensure that \((\psi, (\beta^*)^T + t^*)\). For a vector \( w = (w_u)_{u \in \mathcal{U}_n} \in 0, 1^{\lvert \mathcal{U}_n \rvert} \), we define
\[
\phi_w = \sum_{u \in \mathcal{U}_n} w_u \psi_u.
\]

By constructing \( \psi_u \) and \( \psi_{u'} \) for \( u, u' \in \mathcal{U}_n \) such that \( u \neq u' \) are mutually disjoint, we ensure that \( (\phi_w \in \mathcal{E}_t^\beta \left( [0,1]^t, (\beta^*)^T + t^* \right) \).

For \( i < i^* \), let \( g_i(x) := (x_1, \ldots, x_d_i) \). For \( i = i^* \), define \( g_{i^*,w}(x) = (\psi_w(x_1, \ldots, x_{t_{i^*}}, 0, \ldots, 0)^T \). For \( i > i^* \), let \( g_i(x) := (x_1^i \land 1, 0, \ldots, 0)^T \). Here, \( B = \prod_{i=1}^{\beta} (\beta_i \land 1) \). We frequently use \( \beta^* = \beta^* B \). Since \( t_i \leq \min(d_0, \ldots, d_{i-1}) \) and the mutually disjoint \( \psi_u \) ensures
\[
f_w(x) = g_{\phi_w} \circ \cdots \circ g_{i^*+1} \circ g_{i^*,w} \circ g_{i^*-1} \circ \cdots \circ g_0(x)
= \phi_w(x_1, \ldots, x_{t_{i^*}})^B
= \sum_{u \in \mathcal{U}_n} w_u \psi_u (x_1, \ldots, x_{t_{i^*}})^B,
\]
by choosing a sufficiently large \( K \), we ensure that \( f_w \in \mathcal{F}(q, d, t, \beta, K) \).

For all \( u, \mid \psi_u \mid_2^2 = h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*} \). Let \( \text{Ham}(w, w') = \sum_{u \in \mathcal{U}_n} 1(w_u \neq w'_u) \) denote the Hamming distance; then,
\[
\left\| f_w - f_{w'} \right\|_2^2 = \text{Ham}(w, w') h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*}.
\]

According to the Varshamov–Gilbert bound (see [22], Lemma 2.9) and \( m_n^t = \lvert \mathcal{U}_n \rvert \), there exists a subset \( \mathcal{W} \subset \{0,1\}^{m_n^t} \) with a cardinality of \( \lvert \mathcal{W} \rvert \geq 2^{m_n^t}/8 \). For all \( w, w' \in \mathcal{W} \) such that \( w \neq w' \), it holds that \( \text{Ham}(w, w') \geq m_n^t/8 \). This implies that for all \( \kappa = \left\| K^B \right\|_2^{t^*} / (\sqrt{3} \rho^{\beta^*}) \), \( w, w' \in \mathcal{W}, \ w \neq w' \), we have
\[
\left\| f_w - f_{w'} \right\|_2^2 = \text{Ham}(w, w') h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*} \geq \frac{1}{8} \frac{m_n^t}{h_n} h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*} \geq \frac{1}{8} \frac{2^{\beta^*}}{h_n} \left\| K^B \right\|_2^{2t^*} \geq \kappa^2 \psi_u.
\]

According to the definitions of \( h_n \) and \( \rho \), we have
\[
n\left\| f_w - f_{w'} \right\|_2^2 = n \text{Ham}(w, w') h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*} \leq nm_n^t h_n^{2\beta^* + t^*} \left\| K^B \right\|_2^{2t^*} \leq m_n^t \left( \frac{1}{72} \left\| K^B \right\|_2^{2t^*} \right) \left\| K^B \right\|_2^{2t^*} \leq \frac{m_n^t}{72 \Gamma} \leq \frac{\ln |\mathcal{W}|}{9\Gamma}.
\]
This indicates that the functions $f_w$ with $w \in W$ satisfy (i) and (ii), and, thus, the lemma is proven. $\square$

**Proof of Lemma 1.** Let $c_2 \leq 1$. Since $|m|_\infty \leq K$, we need to consider only the lower bound on $F(L, p, s, F)$, where $F = K + 1$. Let $\hat{f}_n$ be the empirical risk minimizer. Recall that $\Delta_n(\hat{m}_n, m) = 0$. Due to the minimax lower bound in Theorem 3, there exists a constant $c_3$ such that for all sufficiently large $n$, we have $c_3n^{-2\beta/(2\beta + d)} \leq \sup_{m \in \mathcal{D}^{\beta}([0,1],K)} R(\hat{m}_n, m)$. Since $p_0 = d$ and $p_{L+1} = 1$, by Theorem 2, we can conclude that

$$c_3 n^{-2\beta/(2\beta + d)} \leq \sup_{m \in \mathcal{D}^{\beta}([0,1],K)} R(\hat{m}_n, m) \leq 4 \sup_{m \in \mathcal{D}^{\beta}([0,1],K)} \inf_{f \in \mathcal{F}(L,p,s,K+1)} \| f - m \|^2_\infty + C(K + 1)^2 \frac{(s + 1) \ln(n) (s + 1)^L d}{n},$$

where $C$ is a constant. Given $\varepsilon$, let $n_\varepsilon := \left[ \frac{\sqrt{8\varepsilon} / \sqrt{c_3}}{(2\beta + d)/\beta} \right]$. If $\varepsilon \leq \sqrt{c_3} / 8$, and $n_\varepsilon^{-1} \leq 2 \left( \frac{\sqrt{8\varepsilon} / \sqrt{c_3}}{(2\beta + d)/\beta} \right)^{(2\beta + d)/\beta}$, and $8\varepsilon^2 / c_3 \leq n_\varepsilon^{-2\beta/(2\beta + d)}$, then, for sufficiently small $c_2 > 0$ and all $\varepsilon \leq c_2$, we can insert $n_\varepsilon$ into the previous inequalities, and we have

$$8\varepsilon^2 \leq 4 \sup_{m \in \mathcal{D}^{\beta}([0,1],K)} \inf_{f \in \mathcal{F}(L,p,s,K+1)} \| f - m \|^2_\infty + C_1 \varepsilon^{2\beta/d} \frac{1}{s} \left( \ln(1/s) \right) + C_2.$$
Therefore, $|h_{j,a}|_{\infty} \leq K/2$ and $h_{j,a} \in C^{\ell}(0,d,K)$. Let $f_{j,a}(x) = h_{j,a}(x_1 + \ldots + x_d)$. The support of $\psi$ is contained in $[0,2^{\ell}]$ and $2^\nu \geq 2d$. Based on the definition of wavelet coefficients, Equation (A28), the definition of $h_{j,a}$, and using $u_r = \int x^r \psi(x)dx$, for $p_1, \ldots, p_d \in \{0,1, \ldots, 2^{\nu-2} - 1\}$, we have

$$d_{(j,2^{\nu+p_1}) \ldots (j,2^{\nu+p_d})}(f_{j,a}) = 2^{-\frac{d}{2}} \int_{[0,2^{\ell}]^d} h_{j,a} \left( 2^{-\frac{d}{2}} \left( \sum_{\ell=1}^{d} x_\ell + 2^{\nu+p_\ell} \right) \right) \prod_{\ell=1}^{d} \psi(x_\ell)dx = K2^{-\frac{d}{2} - j}\int_{[0,2^{\ell}]^d} \left( \sum_{\ell=1}^{d} x_\ell + p_\ell \right) \prod_{\ell=1}^{d} \psi(x_\ell)dx = r^{-1}2^{-qr - v - 1} K2^{-\frac{d}{2} + \nu} \int_{[0,2^{\ell}]^d} \left( (x_1 + \ldots + x_d)^r + p_\ell \right) \prod_{\ell=1}^{d} \psi(x_\ell)dx = dr^{-1}2^{-qr - v - 1} K\left( p_0d^{-1} \mu_r + p_1d^0 \right) 2^{-\frac{d}{2} + \nu}.$$

In the last equation, according to the definition of $\mu_1 = \ldots = \mu_r = 0$.

In the case of $p_0 = 0$, we take $q(u) = (dr)^{-1}u^{d}I_{[0,1/2]}(u) + (dr)^{-1} (1 - u)^{dr}I_{[1/2,1]}(u)$. Following the same reasoning as above and by using the binomial theorem, we obtain

$$d_{(j,2^{\nu+p_1}) \ldots (j,2^{\nu+p_d})}(f_{j,a}) = \left( \frac{dr}{r} \right) 2^{-qdr - dv - 1} K\mu_r 2^{-\frac{d}{2} + \nu},$$

therefore, the lemma is proven. □

**Proof of Theorem 4.** We define $c(\psi, d)$ as in Lemma 2. We choose an integer $j^*$ such that

$$\frac{1}{n} \leq c(\psi, d) 2^2 2^{-j^*} \leq \frac{2^\nu + d}{n}.$$

This implies that $2^{j^*} \geq \frac{1}{2} c(\psi, d) 2^2 n^{1/(2^\nu + d)}$. According to Lemma 2, there exists a function $f_{j^*,a}$ of the form $h(x_1 + \ldots + x_d)$, where $h \in C^{\ell}(0,d,K)$, such that

$$R(h_{j^*,a}, f_{j^*,a}) \geq \sum_{p_1, \ldots, p_d \in \{0,1, \ldots, 2^{\nu-2} - 1\}} \frac{1}{n} \leq 2^{-qdr - dv - 1} K\mu_r 2^{-\frac{d}{2} + \nu} \geq \frac{1}{n} c(\psi, d)^2 2^2 n^{d/(2^\nu + d)} n^{2d} 2^{-d} 2^{-qdr - dv} \geq \left( c(\psi, d)^2 2^2 n^{d/(2^\nu + d)} n^{2d} \right) 2^{-qdr - dv} \geq n^{-\frac{2}{2^{\nu-1}}};$$

therefore, the lemma is proven. □

**References**


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.