How do noise tails impact on deep ReLU networks?

Jianqing Fan*, Yihong Gu† and Wen-Xin Zhou‡

Princeton University and University of California, San Diego

Abstract

This paper investigates the stability of deep ReLU neural networks for nonparametric regression under the assumption that the noise has only a finite $p$-th moment. We unveil how the optimal rate of convergence depends on $p$, the degree of smoothness and the intrinsic dimension in a class of nonparametric regression functions with hierarchical composition structure when both the adaptive Huber loss and deep ReLU neural networks are used. This optimal rate of convergence cannot be obtained by the ordinary least squares but can be achieved by the Huber loss with a properly chosen parameter that adapts to the sample size, smoothness, and moment parameters. A concentration inequality for the adaptive Huber ReLU neural network estimators with allowable optimization errors is also derived. To establish a matching lower bound within the class of neural network estimators using the Huber loss, we employ a different strategy from the traditional route: constructing a deep ReLU network estimator that has a better empirical loss than the true function and the difference between these two functions furnishes a low bound. This step is related to the Huberization bias, yet more critically to the approximability of deep ReLU networks. As a result, we also contribute some new results on the approximation theory of deep ReLU neural networks.

Keywords: Robustness, Truncation, Heavy Tails, Optimal Rates, Approximability of ReLU networks, Composition of functions.

1 Introduction

Consider a nonparametric regression model $Y = f_0(X) + \varepsilon$, where $Y$ is the response variable, $X$ is a $d$-dimensional covariate vector, $f_0 : \mathbb{R}^d \to \mathbb{R}$ is an unknown function of interest, and the observation noise $\varepsilon$ satisfies $\mathbb{E}[\varepsilon|X] = 0$. A fundamental statistical problem is to estimate the unknown regression function $f_0$ based on a sample of $n$ independent observations $\{(X_i, Y_i)\}_{i=1}^n$ from the above model. From early 1960s to late 1990s, there has been a rich development on nonparametric regression methods, including kernel and local polynomial regressions (Nadaraya, 1964; Watson, 1964; Haberman, 1977; Cleveland, 1979; Fan, 1992, 1993), spline-based methods (Friedman, 1991),...
tree-based methods (Breiman et al., 2017), regression in reproducing kernel Hilbert spaces (RKHS) (Wahba, 1990) and (shallow) neural network regression (Barron, 1994; McCaffrey & Gallant, 1994), to name a few. We refer to Fan & Gijbels (1996), Györfi et al. (2002) and Tsybakov (2009) for a comprehensive exposition on nonparametric regression.

With the development of computational capability that makes training deep neural network possible and efficient, deep neural network has shown superior performance over classical methods in many machine learning tasks, such as image classification and speech recognition (LeCun et al., 2015). From a statistical view, a key reason for the great success of neural networks is their approximation ability in the sense that many complex nonlinear functions (of several variables) can be effectively approximated by neural networks. The well-known universal approximation theorem (Cybenko, 1989; Hornik, 1991; Barron, 1993) shows that a neural network with one hidden layer can approximate any continuous function up to any specified precision. To analyze the statistical properties of neural network estimators, it is necessary to derive nonasymptotic approximation error bounds for neural networks. For fully-connected deep neural networks with ReLU activation function (ReLU-DNN), Telgarsky (2016) demonstrated the great benefits of using deep neural networks. As an illustrative example, a tooth function with $O(2^k)$ oscillations can be realized as a ReLU-DNN with depth $O(k)$ and width $O(1)$, and a one-hidden-layer ReLU-DNN but with $\Omega(2^k)$ nodes. Since smooth functions can be well approximated by Taylor series, based on the idea of point fitting and approximating polynomials via tooth function, Yarotsky (2017) proved a near-optimal approximation error (in terms of the number of parameters) for Lipschitz functions. This idea is widely used in deriving nonasymptotic approximation error bounds for various classes of functions (Shen et al., 2019, 2022; Suzuki, 2018; Yarotsky, 2018; Lu et al., 2020). Via more delicate point fitting techniques, Lu et al. (2020) established a nonasymptotic $L_\infty$ approximation error bound for smooth functions using ReLU-DNN with arbitrary depth and width, which is optimal (in terms of width and depth) up to logarithmic factors.

Another important reason for the success of neural networks is their ability of being adaptive to unknown low-dimensional structure. By employing the compositional nature of deep neural network and the aforementioned approximation results for smooth functions, Bauer & Kohler (2019), Schmidt-Hieber (2020) and Kohler & Langer (2021) showed that neural networks can circumvent the curse of dimensionality if the intrinsic dimension $d^*$ of the regression function $f_0$ is much smaller than the input dimension $d$. Specifically, suppose $f_0$ can be represented as a hierarchical composition of several smooth functions, with either a high degree of smoothness condition or low input dimension, then the neural network can automatically adapt to the intrinsic low-dimensional structure without knowing the composition structure explicitly. Schmidt-Hieber (2020) also showed that neural network estimates achieve minimax-optimal rate of convergence when the regression function has such structure. Furthermore, deep neural network can also be used to estimate the nonlinear component of a semi-parametric model, which circumvents the curse of dimensionality and facilitates the inference on the linear component (Farrell et al., 2021; Zhong et al., 2022).

The existing results for the least squares ReLU-DNN regression estimates rely on a sub-Gaussian moment condition on the regression error (Schmidt-Hieber, 2020; Kohler & Langer, 2021). A natural question is:

**Whether the least squares ReLU-DNN estimator achieves the same convergence rate when the error distribution is heavy-tailed, and if not, whether there exists a robust alternative that can.**
When regressing directly over the nonparametric function class to which \( f_0 \) belongs, some recent works discuss the effect of heavy-tailed errors on the convergence rate of (constrained) least squares estimator (LSE) (Han & Wellner, 2018, 2019; Kuchibhotla & Patra, 2019). They argue that the convergence rate of LSE may depend on both the complexity of the function class and the moment condition on the regression error. Specifically, for a uniformly bounded function class satisfying a standard “entropy condition” with exponent \( \alpha \in (0, 2) \), Han & Wellner (2019) showed that the corresponding LSE converges at a rate \( \mathcal{O}(n^{-1/(2+\alpha)} \lor n^{-1/(1-p)}) \) (in \( L_2 \) error) when \( \epsilon \) has bounded \((p + \epsilon)\)-th \((p \geq 1)\) moment and is independent of \( X \). Therefore, for certain function class that is not sufficiently complex, the heavy tailedness of the errors is the main cause for the LSE to converge at a slower rate (when \( p < 1 + 2/\alpha \)). Moreover, Han & Wellner (2019) also proved the sharpness of this rate by constructing some non-smooth function classes that witness the worst case rate \( \mathcal{O}(n^{-1/(1-p)}) \). Kuchibhotla & Patra (2019) proved similar results when \( \epsilon \) may also depend on \( X \). Specifically, they provided a detailed characterization of the convergence rate when \( \mathcal{F} \) is a uniform VC-type function class (indexed by \( \alpha \geq 0 \)) and \( p = 2 \). Note that an ReLU-DNN with fixed depth and width belongs to a parametric function class with finite VC-dimension, which corresponds to the case of \( \alpha = 0 \). In this case, Han & Wellner (2019) showed that the convergence rate is of order \( \mathcal{O}(n^{-1/4}) \), achieved by a highly non-smooth function class. Kuchibhotla & Patra (2019) claimed that the degree of smoothness for the function class, measured via local envelope function, determines the convergence rate which is \( \mathcal{O}(n^{-1/2}) \) for the most smooth function class and \( \mathcal{O}(n^{-1/4}) \) for the most non-smooth class. However, due to the more complex nature of neural network classes, the impact of the tails of the noise on the LSE estimator remains unclear.

To robustify least squares estimates, several robust loss functions have been widely used, including but not limited to the \( L_1 \) loss, Huber’s loss (Huber, 1973), the Cauchy loss and Tukey’s biweight loss (Beaton & Tukey, 1974). Originally these robust methods were introduced to guard against outliers in the observations, say under Huber’s contamination model. When there is no contamination but the underlying distribution itself is heavy-tailed and skewed, Fan et al. (2017) and Sun et al. (2020) revisited the Huber regression method, and proposed the use of an adaptive robustification parameter \( \tau \) for bias-robustness tradeoff. Via a deviation study, Sun et al. (2020) showed that the adaptive Huber (linear) regression estimator satisfies sub-Gaussian-type concentration bounds even when the error only has low-order moments. It should be noted that for linear models, both the LSE and its robust alternative admit the same rate of convergence as long as the errors have finite variance, while the advantage of the latter is that it achieves exponential-type deviation bounds even when the error variable does not have exponentially thin tails. The main reason for this is that linear functions of the form \( f(x) = \beta^\top x \) not only have simple structures but also are sufficiently smooth if \( \|\beta\|_2 \) is bounded. For nonparametric models, it is unclear whether a robust regression estimator can achieve a faster rate of convergence than LSE when \( p \geq 2 \).

The shortcomings of the nonparametric LSE, specifically the lack of robustness, have motivated the development of robust methods when ReLU-DNN is used (Shen et al., 2021a,b; Hernan Madrid Padilla et al., 2020; Lederer, 2020). Using deep neural networks, these papers studied nonparametric robust regression with a \( L_1 \)-Lipschitz continuous loss \( \rho \), typified by the Huber loss and the check loss (Shen et al., 2021a; Hernan Madrid Padilla et al., 2020), and established upper bounds on the excess risk \( \mathbb{E}[\rho(\hat{f}(X) - Y) - \rho(f^*(X) - Y)] \) where \( f^* \) is the population risk minimizer. Shen et al. (2021b) showed that when the observation noise has bounded \( p \)-th moment, the empirical risk
minimizer $\hat{f}$ satisfies the excess risk bound
\[
\mathbb{E}[\rho(Y - \hat{f}(X)) - \rho(Y - f^*(X))] \leq \frac{\lambda_L (NL)^2}{n^{1-1/p}} + \omega_f((NL)^{-2/\nu})
\]
up to logarithmic factors, where $\omega_f(\cdot)$ is the modulus of continuity of function $f$, i.e., $\omega_f(\delta) = \sup_{\|x-y\| \leq \delta} |f(x) - f(y)|$. These results provide a first glance on the impact of noise tails on regression with ReLU-DNN, but still leave several loopholes as follows: (i) the convergence rate cannot take advantage of the low-dimensional structure of $f_0$ since $f^*$ and $f_0$ in general are not the same; (ii) the convergence rate (under $L_2$ loss) for estimating $f_0$, i.e., $\mathbb{E}_X[\hat{f}(X) - f_0(X)]^2$, is still unclear; (iii) the theoretical benefit of using a robust ReLU-DNN estimator is ambiguous from the above result because the obtained convergence rate turns out to be slower than that of the least squares counterpart when $f_0 = f^*$ is $(\beta, C)$-smooth; see the discussions in Section 4.4.

In this paper, we attempt to address the aforementioned questions by providing a comprehensive analysis of the impact of heavy-tailed noise on the convergence rate of ReLU-DNN estimators. Inspired by Fan et al. (2017), we focus on the adaptive Huber ReLU-DNN estimator, defined as the empirical Huber loss minimizer over the ReLU-DNN function class. Collectively, the network parameters (width and depth) and the robustification parameter play a crucial role in reaching the optimal statistical rate of convergence when $\varepsilon$ has bounded $p$-th moment. By carefully balancing the Huberization bias, the ReLU-DNN approximation error of the regression function and the statistical error, we unveil a subtle dependence of the convergence rate (under $L_2$ risk) on the moment index $p$ and the intrinsic dimension of the function class of interest. Moreover, we establish lower bounds on the convergence rates of both adaptive Huber ReLU-DNN estimator and its least squares counterpart, thus justifying the tightness of the obtained upper bounds. Although both estimators adapt to the low-dimensional structure of $f_0$, the adaptive Huber estimator with a properly chosen hyper-parameter achieves a faster convergence rate than the LSE in the presence of heavy-tailed errors. These results demonstrate the advantage of adaptive Huber estimator over LSE in the context of nonparametric regression with ReLU-DNN. We summarize the main contributions of this work as follows.

(i) We establish nonasymptotic error bounds on the $L_2$ risk $\|\hat{f}_n - f_0\|_2^2 = \mathbb{E}_X[\hat{f}_n(X) - f_0(X)]^2$ for the adaptive Huber estimator using deep ReLU neural networks; see Theorem 3.4. Specifically, we show that if the error $\varepsilon$ has bounded $p$-th moment ($p \geq 2$), and with properly tuned hyper-parameters (robustification parameter, network depth and width), any approximate (within a given order of optimization error) empirical (Huber) risk minimizer satisfies, up to logarithmic factors, that
\[
\mathbb{P}(\|\hat{f}_n - f_0\|_2 \geq D\delta_n) \leq \exp(-n^\gamma D^{2p}) \quad \text{with} \quad \delta_n = n^{-\nu^*/\nu^*}, \quad (1.1)
\]
where $\gamma^*$ depends on the intrinsic dimension of $f_0$, $\nu^* = 1 - 1/(2p - 1)$, and $c$ is a positive constant depending only on $\gamma^*$. Our results reveal the following two advantages of the adaptive Huber estimator:

(a). With properly chosen hyper-parameters, the adaptive Huber ReLU-DNN estimator circumvents the curse of dimensionality in the heavy-tailed setting the same way as the LSE does with sub-Gaussian errors: the convergence rate depends only on the intrinsic dimension.
(b). Under the $p$-th ($p \geq 2$) moment condition, the adaptive Huber estimator achieves a faster convergence rate than the LSE. As we shall see, the latter converges at the rate $O_p(n^{-2/p'}(2p' + 1))$ with $p' = 1 - 1/p < p^*$. In addition, from a nonasymptotic perspective, the adaptive Huber estimator admits exponential-type deviation bounds whereas the worst case probability for LSE decays polynomially.

In the special case where the heavy-tailed observation noise is symmetric, the Huber ReLU-DNN estimator with a fixed robustification parameter achieves the optimal convergence rate $O_p(n^{-p^*/(2p^* + 1)})$ as if the noise is sub-Gaussian. This cannot be achieved by the LSE.

(ii) We provide a lower bound for the convergence rate of adaptive Huber ReLU-DNN estimator with an arbitrary configuration of hyper-parameters in Theorem 4.1. The result shows that the obtained upper bound is tight up to logarithmic factors. By letting the robustification parameter go to infinity, it also provides a lower bound for the convergence rate of the least squares ReLU-DNN estimator.

(iii) We present more characterizations about the approximation ability of deep ReLU neural networks. These results play a key role in our proof of the lower bound, and are of independent interest. The existing approximation results for deep ReLU neural networks are primarily about their representation power, i.e., whether it can efficiently approximate a group of functions. In contrast, Theorem 4.6 characterizes one of their approximation abilities from a different angle, which is related to the non-robustness nature of ReLU-DNN. Furthermore, we provide a lower bound on the approximation error under $L_2$ norm (see Theorem 4.2), while the previous lower bound results are under $L_\infty$ norm.

Organization  The rest of the paper is as follows. Section 2 describes the Huber ReLU-DNN estimator, which is main objective of this work. In Section 3, we first provide an upper bound on the Huberization bias, and then establish nonasymptotic error bounds for the adaptive Huber ReLU-DNN estimator. Both symmetric and asymmetric error cases are considered. The lower bounds on the convergence rate of a nonparametric Huber estimator with any sufficiently large hyper-parameter are presented in Section 4. All the technical proofs are collected in the supplemental material.

Notations  The following notations will be used throughout this paper. We use $c_1, c_2, \ldots$ to denote the global constants that appear in the statement of any theorem, proposition, corollary, and lemma. We use $C_1, C_2, \ldots$ to denote the local intermediate constants in the proof. Hence all the $c_1, c_2, \ldots$ have unique referred numbers, while all the $C_1, C_2, \ldots$ will have different referred numbers in respective proofs. We use $a \lesssim b$ if there exists some universal constant $C$ such that $a \leq Cb$, we use $a \gtrsim b$ if there exists some universal constant $C > 0$ such that $a \geq Cb$, we use $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$.

2 Methodology

Consider a non-parameteric regression model

$$Y = f_0(X) + \epsilon, \quad (2.1)$$
where \( X \in [0,1]^d \) is the \( d \)-dimensional covariate vector, and \( \varepsilon \) is the noise variable satisfying
\[
\mathbb{E}[\varepsilon|X = x] = 0 \quad \text{and} \quad \mathbb{E}[|\varepsilon|^p|X = x] \leq v_p < \infty \quad \text{for all} \ x \in [0,1]^d.
\] (2.2)

Let \( \{(X_i,Y_i)\}_{i=1}^n \) be i.i.d. observations from model (2.1). Our goal is to estimate the unknown regression function \( f_0 : [0,1]^d \to \mathbb{R} \). Within a suitably chosen function class \( \mathcal{F}_n \), the nonparametric least squares method aims to find some \( \hat{f}_n \) that minimizes the \( L_2 \)-loss
\[
\hat{R}(f) = \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2.
\] (2.3)

The accuracy of the estimator \( \hat{f}_n \) can be evaluated through the mean squared error \( \mathbb{E}_X[\hat{f}_n(X) - f_0(X)]^2 \), which is the excess risk
\[
\mathcal{R}(\hat{f}) = \mathcal{R}(f_0) = \mathbb{E}_X[\hat{f}_n(X) - f_0(X)]^2,
\] (2.4)

where \( \mathcal{R}(f) = \mathbb{E}_{(X,Y)}[|Y - f(X)|^2] \) is the prediction error of \( f \). It is well known that the minimax rate of convergence is of order \( n^{-\frac{2\beta}{2\beta + d^*}} \) (Györfi et al., 2002), where \( \beta \) is some smoothness parameter of the nonparametric function class. This is often referred to as the curse of dimensionality in nonparametric regression as the rate is substantially slower when \( d \) is moderately large. Under a hierarchical composition assumption on \( f_0 \), Schmidt-Hieber (2020) showed that the estimator obtained from learning a deep neural network with ReLU activation and sparse connection achieves a faster convergence rate \( n^{-\frac{2\beta^*}{2\beta^* + d^*}} \), where \( d^* \) is the intrinsic dimension that can be much smaller than the input dimension \( d \), and \( \beta^* \) is corresponding smoothness parameter. Further, by using advanced neural network approximation results, Kohler & Langer (2021) proved that the sparse connection requirement can be relaxed. More specifically, they showed that if \( f_0 \) satisfies an hierarchical composition model with smoothness and order parameters \((\beta^*,d^*)\), the convergence rate is \( n^{-\frac{2\beta^*}{2\beta^* + d^*}} \) (up to some logarithmic factor).

Notably, most of the existing results on estimation rates are established under the assumption that the noise variable \( \varepsilon \), or equivalently, the response variable \( Y \), is sub-Gaussian. Such an assumption would raise legitimate concerns when heavy-tailed data is observed. A natural question is that how well would deep neural networks work in the context of nonparametric regression with heavy-tailed errors, or how critical this sub-Gaussian condition is so as to achieve faster convergence rate via DNN. To approach this question, we start with the Huber loss (Huber, 1964), which robustifies the \( L_2 \)-loss through a truncation parameter \( \tau > 0 \).

**Definition 2.1** (Huber Loss). Given some parameter \( \tau > 0 \), the Huber loss \( \ell_\tau(\cdot) \) is defined as
\[
\ell_\tau(x) = \begin{cases} 
\frac{1}{2}x^2 & |x| \leq \tau \\
\tau|x| - \frac{1}{2}\tau^2 & |x| > \tau
\end{cases}
\] (2.5)

Note that the Huber loss is continuously differentiable with score function \( \ell_\tau'(x) = \min\{\max(-\tau,x),\tau\} \).

Given a robustification parameter \( \tau = \tau_n > 0 \), consider the empirical Huber loss
\[
\hat{R}_\tau(f) = \frac{1}{n} \sum_{i=1}^n \ell_\tau(Y_i - f(X_i)).
\] (2.6)
The corresponding nonparametric Huber estimator is defined as
\[
\hat{f}_n \in \arg\min_{f \in \mathcal{F}_n(d, L, N, M)} \hat{R}_n(f)
\]  
(2.7)
where \( \mathcal{F}_n(d, L, N, M) \) is a set of truncated deep ReLU neural networks with width \( N \), depth \( L \), input dimension \( d \) and a truncation parameter \( M \). More specifically, we begin with the architecture of a deep neural network, which can be written as
\[
f(x) = L_{L+1} \circ \sigma \circ L_L \circ \sigma \circ L_{L-1} \circ \sigma \circ \cdots \circ L_2 \circ \sigma \circ L_1(x),
\]  
(2.8)
where \( L_i(x) = W_i x + b_i \) is a linear transformation with \( W_i \in \mathbb{R}^{d_i \times d_{i-1}}, b_i \in \mathbb{R}^{d_i} \) and \((d_0, d_1, \cdots, d_L, d_{L+1}) = (d, N, \cdots, N, 1)\), and \( \sigma : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i} \) applies the ReLU function \( \sigma(x) = \max(0, x) \) to each entry of an \( \mathbb{R}^{d_i} \)-valued vector. We refer to this type of networks as deep ReLU network with width \( N \) and depth \( L \), and \(((W_i, b_i))_{i=1}^{L+1} \) are the network weights or parameters. Let \( \|f\|_\infty \) be the supremum norm of \( f(\cdot) \) over the unit cube \([0, 1]^d\). Now we are ready to define the following two classes of network functions:
\[
\mathcal{F}_n(d, L, N) = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \text{ is of the form (2.8) with width } N \text{ and depth } L \right\}
\]
and
\[
\mathcal{F}_n(d, L, N, M) = T_M \mathcal{F}_n(d, L, N) = \{ f = T_M \hat{g} : g \in \mathcal{F}_n(d, L, N) \},
\]
where \( T_M \) is the truncation (Winsorization) operator at level \( M > 0 \), defined as \( T_M u = \text{sgn}(u)(|u| \wedge M) \). For convenience, we write \( \mathcal{F}_n = \mathcal{F}_n(d, L, N, M) \) throughout the rest of the paper.

3 Statistical Analysis

Let \( \beta = r + s \) for some nonnegative integer \( r \) and \( 0 < s \leq 1 \). A \( d \)-variate function \( g \) is called \((\beta, C)\) smooth if for every sequence of nonnegative integers \( \{a_j\}_{j=1}^d \) such that \( \sum_{j=1}^d a_j = r \), the partial derivative \( \partial^a g / (\partial x_1^{a_1} \cdots \partial x_d^{a_d}) \) exists and satisfies
\[
\left| \frac{\partial^a g}{\partial x_1^{a_1} \cdots \partial x_d^{a_d}}(x) - \frac{\partial^a g}{\partial z_1^{a_1} \cdots \partial z_d^{a_d}}(z) \right| \leq C \|x - z\|^s_2.
\]

**Condition 1** (Hierarchical composition model). The regression function \( f_0 : [0, 1]^d \rightarrow \mathbb{R} \) is uniformly bounded with \( \|f_0\|_\infty \leq M \) and belongs to the function class \( \mathcal{H}(l, \mathcal{P}) \) (Kohler & Langer, 2021), where \( l \in \mathbb{N}^+ \) and \( \mathcal{P} \) is a subset of \( \mathbb{N}^+ \times \mathbb{N}^+ \). For \( l = 1 \),
\[
\mathcal{H}(1, \mathcal{P}) = \{ h : \mathbb{R}^d \rightarrow \mathbb{R} : h(x) = g(x_{\pi(1)}, \ldots, x_{\pi(t)}), \text{ where } g : \mathbb{R}^l \rightarrow \mathbb{R} \text{ is (} \beta, C \text{)-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } \pi : [t] \rightarrow [d] \}\].
It consists of all \( t \)-variate functions with \((\beta, C)\) smoothness. For \( l > 1 \), \( \mathcal{H}(l, \mathcal{P}) \) is defined recursively as
\[
\mathcal{H}(l, \mathcal{P}) = \{ h : \mathbb{R}^d \rightarrow \mathbb{R} : h(x) = g(f_1(x), \ldots, f_l(x)), \text{ where } g : \mathbb{R}^l \rightarrow \mathbb{R} \text{ is (} \beta, C \text{)-smooth for some } (\beta, t) \in \mathcal{P} \text{ and } f_i \in \mathcal{H}(l - 1, \mathcal{P}) \}\].
Here \( C \) is some positive constant.
Remark 3.1. Our definition of the hierarchical composition model is slightly different from that in Bauer & Kohler (2019) and Kohler & Langer (2021). For the smoothness parameter $\beta$, they require $\beta \in [1, \infty)$ while in this paper we assume $\beta \in \mathbb{N}^+$. This is because our analysis relies on an ReLU-DNN approximation result for $d$-variate $(\beta, C)$-smooth function with $d, \beta \in \mathbb{N}^+$ from Lu et al. (2020). The proof strategy therein can be extended to cover all real-valued $\beta \in [1, \infty)$. To avoid having lengthy and repetitive technical proofs with little new statistical insights, we will focus on the specific case $\beta \in \mathbb{N}^+$.

Condition 2 (Moment conditions). The random covariate vector $X \in \mathbb{R}^d$ follows some distribution $P_X$ over the unit cube $[0,1]^d$. The noise variable $\varepsilon$ has zero mean and uniformly bounded (conditional) $p$-th moments for some $p \geq 2$, that is,

$$
\mathbb{E}[\varepsilon|X = x] = 0 \quad \text{and} \quad \mathbb{E}[|\varepsilon|^p|X = x] \leq v_p < \infty \quad \text{for all} \quad x \in [0,1]^d.
$$

(3.1)

For any given $\tau > 0$, define the population Huber risk

$$
R_\tau(f) = \mathbb{E}_X \{ \ell_\tau(Y - f(X)) \}.
$$

(3.2)

Our goal is to derive the rate of convergence for $\hat{f}_n$ under the $\| \cdot \|_2 = \| \cdot \|_{L_2(P_X)}$-norm, that is, $\|f\|_2 = \sqrt{\mathbb{E}_X \|f(X)\|^2}$. Under Condition 2, it is easy to see that

$$
\|f - f_0\|_2^2 = R(f) - R(f_0),
$$

(3.3)

where $R(f) = R_\infty(f)$. In other words, the excess risk coincides with the $L_2$-error.

For the Huber loss, let $\psi_\tau(x) = \ell_\tau'(x) = \text{sgn}(x)(|x| \wedge \tau)$ be the corresponding score function, which is Lipschitz continuous and has a derivative almost everywhere, that is,

$$
\psi_\tau'(x) = 1[|x| \leq \tau].
$$

(3.4)

The parameter $\tau$ plays an important role in robustness-bias tradeoff (Sun et al., 2020), and depends on the scale of $\varepsilon$ (e.g., standard deviation). If the distribution of $\varepsilon$ is symmetric around zero, Huberization will not introduce bias because the underlying regression function $f_0$ is also the population Huber risk minimizer for any $\tau$. In this case, $\tau$ will only depend on the noise scale. Under a bounded second moment condition, we show that the Huber estimator based on deep ReLU networks achieves the same convergence rate as its least squares counterpart for sub-Gaussian noise. In the case of asymmetric noise, the Huberization bias can no longer be disregarded, and needs to be balanced with the statistical error. As a result, the robustification parameter $\tau$ should adapt to the sample size $n$ in a suitable way so as to achieve a bias-robustness trade-off. The amount of bias that should be traded for robustness depends on the moment parameter $p$. As we shall see, the final convergence rate is slower than but infinitely close to the minimax rate as $p \to \infty$.

In addition to the moment condition (3.1), we sometimes further impose the following symmetry assumption that is of independent interest.

Condition 3 (Symmetric noise). For each $x \in [0,1]^d$, the noise $\varepsilon|X = x$ is symmetric, i.e., the distribution of $\varepsilon|X = x$ is same with respect to $-\varepsilon|X = x$. 

8
3.1 Lower bound on excess risk and upper bound on Huberization bias

We first examine the population Huber loss and quantify the bias induced by Huberization. Denote by $\Theta = \{ f : \mathbb{R}^d \to \mathbb{R} : \|f\|_\infty \leq M, \|f\|_2 < \infty \}$ the global function space. For any $r > 0$, define $\Theta_0(r) = \{ f \in \Theta : \|f - f_0\|_2 \leq r \}$ and $\Theta_0^c(r) = \{ f \in \Theta : \|f - f_0\|_2 > r \}$. For each function $f$, write $\Delta_f(x) = f_0(x) - f(x)$, so that

$$R_r(f) = \mathbb{E}_{(X,Y)}[\ell_r(Y - f(X))] = \mathbb{E}_{(X,Y)}[\ell_r(\epsilon + \Delta_f(X))].$$

\textbf{Proposition 3.1.} Assume Condition 2 holds, $\|f_0\|_\infty \leq M$, and let $\tau \geq c_1 = 2 \max\{2M, (2\nu_p)^{1/p}\}$. Then

$$R_r(f) - R_r(f_0) \geq \frac{1}{8} \|f - f_0\|_2^2 \quad \text{for all } f \in \Theta_0^c(c_2\tau^{1-p})$$

with $c_2 = 8\nu_p$. Assume, in addition, that Condition 3 holds and $p \geq 1$ in Condition 2. Then

$$R_r(f) - R_r(f_0) \geq \frac{1}{4} \|f - f_0\|_2^2 \quad \text{for all } f \in \Theta.$$

Proposition 3.1 provides lower bounds for the population excess risk over some subset of the function space. Let $f_{0,\tau}$ be the global minimizer of the population Huber risk, i.e.,

$$f_{0,\tau} \in \arg\min_{f \in \Theta} R_r(f).$$

If the distribution of $\epsilon$ is asymmetric, $f_{0,\tau}$ generally differs from $f_0$ with $r_\tau = \|f_{0,\tau} - f_0\|_2 > 0$. By the optimality of $f_{0,\tau}$, $R_r(f_{0,\tau}) - R_r(f_0) \leq 0$. This explains why the lower bound (3.6) holds only outside a local neighborhood of $f_0$. On the other hand, if the (conditional) distribution of $\epsilon$ is symmetric, it is easy to see that $f_{0,\tau} = f_0$. The lower bound (3.6) can be viewed as a form of the restricted (outside the local neighborhood) strong convexity, provided that the robustification parameter $\tau$ is sufficiently large.

The following proposition provides an upper bound for the Huberization bias $\|f_{0,\tau} - f_0\|_2$.

\textbf{Proposition 3.2.} Assume Condition 2 holds, and let $\tau \geq c_1 = 2 \max\{2M, (2\nu_p)^{1/p}\}$. Then, the global minimizer $f_{0,\tau}$ of the population Huber loss satisfies

$$\|f_{0,\tau} - f_0\|_2 \leq 4\nu_p\tau^{1-p}. \quad (3.9)$$

In addition, assume there exists some constant $\sigma > 0$ such that

$$\mathbb{P}(|\epsilon| \geq t|X = x) \leq 2e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \in \mathbb{R} \text{ and } x \in [0,1]^d. \quad (3.10)$$

Then

$$\|f_{0,\tau} - f_0\|_2 \leq 2.75(\tau + \sigma^2/\tau)e^{-\frac{\sigma^2}{2\tau^2}}. \quad (3.11)$$

From the above result we see that the Huberization bias, at least an upper bound of it, depends on both the tuning parameter $\tau$ and the moment index $p \geq 2$. If $\epsilon$ only has bounded moments up to order $p$, the bias decays at polynomial rates; if $\epsilon$ is sub-Gaussian as assumed in Kohler & Langer (2021), the bias decays exponentially fast as a function of $\tau$. By choosing $\tau = \sigma \sqrt{2\log n}$, the bias satisfies

$$\|f_{0,\tau} - f_0\|_2 \leq c_\tau n^{-1}(\log n)^{1/2},$$

and hence is negligible compared to the statistical error. In the case of heavy-tailed noise, say $p = 2$, the bias will play a bigger role, and may result in a slower rate of convergence.
3.2 Rate of convergence

We first present the general results in the absence of symmetry—Condition 3. The following neural network approximation result provides the key to establish the convergence rate.

**Proposition 3.3** (Neural network approximation error, upper bound). Assume Condition 1 holds, and define

\[(\beta^*, d^*) = \arg\min_{(\beta, d) \in \mathcal{P}} \frac{\beta}{d}, \quad \gamma^* = \frac{\beta^*}{d^*}.\]

Then, for any \(L, N \geq 2\), there exists some \(f^\dagger \in \mathcal{F}(d, c_3L \log L, c_4N \log N, M)\) such that

\[\|f^\dagger - f_0\|_\infty \leq c_5(NL)^{-2\gamma^*},\]

(3.12)

where \(c_2\)–\(c_5\) are positive constants that are independent of \((n, L, N)\).

Compared with the approximation results in Kohler & Langer (2021), the above result applies to a wider range of neural network architectures. To be specific, in order to obtain similar approximation errors, Kohler & Langer (2021) suggests to use neural networks with a ‘special shape’, which is either thin and deep or wide and shallow. In contrast, Proposition 3.3 allows one to tune the width and depth of a neural network more flexibly, which weakens the condition and provides deeper insights.

Based on the above neural network approximation result, the following theorem establishes the statistical rate of convergence for the adaptive Huber ReLU-DNN estimator.

**Theorem 3.4** (Convergence rate under asymmetric noise). Assume Conditions 1 and 2 hold, and let \((X_1, Y_1), \cdots, (X_n, Y_n)\) be i.i.d. samples generated from \(Y = f_0(X) + \varepsilon\). Let \(\beta^*, d^*\) and \(\gamma^*\) be as in Proposition 3.3, and \(L, N \geq 2\) be such that

\[LN \approx n^{\frac{2\gamma^*}{\nu^*}}\] with \(\nu^* = 1 - \frac{1}{2p-1}\). (3.13)

Consider the function class \(\mathcal{F}_n = \mathcal{F}_n(d, \bar{L}, \bar{N}, M)\) with depth and width

\[
\bar{L} = c_3L \log(L) \quad \text{and} \quad \bar{N} = c_4N \log(N),
\]

(3.14)

where \(c_3\)–\(c_4\) are positive constants from Proposition 3.3. Then, there exist positive constants \(c_6\) and \(c_7\) independent of \((n, t)\) such that for any \(t > c_7\), any approximate empirical risk minimizer \(\hat{f}_n\) with optimization error of order \(\delta^2_n\), more specifically,

\[
P\{\hat{\mathbb{R}}(\hat{f}_n) \geq \inf_{f \in \mathcal{F}_n} \hat{\mathbb{R}}_\tau(f) + 2^{-8}\delta^2_n\} \leq \exp\left\{-c_6n^{\frac{\nu^*}{2\gamma^*}}(\log n)^6\tau^2\right\}
\]

with

\[
\tau \approx \left(n^{\frac{\gamma^*}{\nu^*}}(\log n)^{-3}\right)^{2(1-\nu^*)} \quad \text{and} \quad \delta^2_n \approx \left(n^{-\frac{\gamma^*}{\nu^*}}(\log n)^3\right)^{\nu^*},
\]

(3.15)

has statistical error

\[
P\{\|\hat{f}_n - f_0\| \geq t\delta_n\} \leq \exp\left\{-c_6n^{\frac{\nu^*}{2\gamma^*}}(\log n)^6\tau^2\right\}.
\]

(3.16)
If the noise variable $\varepsilon$ has bounded (conditional) $p$-th ($p \geq 2$) moment, Theorem 3.4 shows that the adaptive Huber estimator based on deep ReLU networks and with a suitably chosen robustification parameter admits a convergence rate

$$||\hat{f}_n - f_0||_2 = O_P\left(n^{-\frac{\beta^*}{2p-1}} (\log n)^{3\nu'}\right), \quad \text{where} \quad \nu' = 1 - \frac{1}{2p-1} \in [2/3, 1). \quad (3.17)$$

Compared with the nonparametric least squares estimator over the same function class when $\varepsilon$ is sub-Gaussian, there is a statistical price to be paid by allowing for heavy-tailed errors that only have finite low order.

**Remark 3.2.** For any $\delta \geq 0$, let $S(\delta) = \{\tilde{f} \in \mathcal{F}_n : \tilde{R}_t(\tilde{f}) \leq \inf_{f \in \mathcal{F}_n} \tilde{R}_t(f) + 2^{-8}\delta^2\}$ denote a set of approximate empirical risk minimizers. From the proof of Theorem 3.4, it can be further shown that with properly chosen hyper-parameters (network and robustification parameters), it holds

$$\sup_{f \in \mathcal{H}(\mathcal{F}_n), \mathbb{E}|\varepsilon|^p |X=x| \leq \nu_p} \mathbb{P}\{ ||\hat{f}_n \in S(t\delta_n) \text{ s.t.} ||\hat{f}_n - f_0||_2 \geq t\delta_n \} \leq \exp\{-c_6n^{\frac{\beta^*}{2p-1}} (\log n)^{6\nu'}\} \quad (3.18)$$

for all $t > c_7$, where $\delta_n$ is as in (3.15).

**Remark 3.3.** When the noise variable $\varepsilon$ satisfies the sub-Gaussian tail assumption (3.10), by Proposition (3.2) we may choose $\tau = \tau_n = \sigma \sqrt{\log n}$, which is much smaller than that in (3.15), so that the Huberization bias is negligible. Following the proof of Theorem 3.4, it can be shown that the resulting nonparametric Huber estimator satisfies

$$||\hat{f}_n - f_0||_2 = O_P\left(n^{-\frac{\beta^*}{2p-1}} (\log n)^{3}\right).$$

**Remark 3.4.** If $p = p_n$ grows with $n$ and satisfies $p \asymp \log n$, it follows that $\nu' = 1 - (2p - 1)^{-1} \asymp 1 - (\log n)^{-1}$. Then, the convergence rate in Theorem 3.4 coincides with that under the sub-Gaussian tail assumption by noting that

$$n^{-\frac{\beta^*}{2p-1}} \leq n^{-\frac{\beta^*}{2p-1}} \left(1 - \frac{1}{2p-1}\right) \leq n^{-\frac{\beta^*}{2p-1}} n^{\frac{\beta^*}{2p-1}} \asymp n^{-\frac{\beta^*}{2p-1}},$$

where the last step follows from the fact that $n^{\alpha/\log n} = e^{\alpha \log n / \log n} = e^{\alpha}$ for any constant $\alpha$.

In the robust regression literature, the case of symmetric noise is sometimes of independent interest. The following result shows that the Huber estimator will benefit from the blessing of symmetry: with a robustification parameter of constant level, it achieves the same rate of convergence as its least squares counterpart when the noise is sub-Gaussian.

**Theorem 3.5** (Convergence rate under symmetric noise). Assume Conditions 1–3 hold with $p \geq 1$, and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples generated from $Y = f_0(X) + \varepsilon$. Moreover, let the parameters $d^*, \beta^*$, $\gamma^*$ be as in Theorem 3.4, and set $\alpha^* = \frac{\gamma^*}{2p-1}$. Consider the function class $\mathcal{F}_n = \mathcal{F}_n(d, \bar{L}, \bar{N}, M)$, where $\bar{L} = c_3L \log L$ and $\bar{N} = c_4N \log N$ with $L, N \geq 2$ satisfying $LN \asymp n^{1/2-\alpha^*}$. Then, there exist some positive constants $c_8$ and $c_9$ independent of $(n, t)$ such that for any $t > c_9$, any approximate estimator $\hat{f}_n$ satisfying

$$\mathbb{P}\{ R_t(\hat{f}_n) \geq \inf_{f \in \mathcal{F}_n} \tilde{R}_t(f) + 2^{-7}\delta_n^2 t^2 \} \leq \exp\{-c_8n^{1-2\alpha^*} (\log n)^6 t^2\}$$

with $c_1 \leq \tau \leq 1$ and $\delta_n \asymp n^{-\alpha^*} (\log n)^3$ satisfies

$$\mathbb{P}\{ ||\hat{f}_n - f_0||_2 \geq t\delta_n \} \leq \exp\{-c_8n^{1-2\alpha^*} (\log n)^6 t^2\}, \quad (3.19)$$

...
Remark 3.5. It should be noted that the Huber estimator is not the only estimator that can achieve \( n^{-\frac{2\nu}{2\nu+2}} \) rate of convergence under asymmetric and heavy-tailed noise. For example, one may also use the robust loss considered in Catoni (2012) or a pseudo-Huber loss that is twice continuously differentiable everywhere. The theoretical analysis of these estimators will follow the same argument. On the other hand, a simpler robustification strategy is to apply adaptive truncation on \( Y \) (Fan et al., 2021), resulting in the empirical risk

\[
\hat{R}_{T,\tau}(f) = \frac{1}{n} \sum_{i=1}^{n} \{T_{\tau}(Y_i) - f(X_i)\}^2.
\]

This corresponds to the LSE with truncated response and facilitates the neural network training. It is possible to obtain a result that is comparable to Theorem 3.4 using a similar argument. The first key step is to show a lower bound of the population excess risk as in Proposition 3.1, i.e.,

\[
R_{T,\tau}(f) - R_{T,\tau}(f_0) \gtrsim \|f - f_0\|_2^2 \quad \text{if} \quad \|f - f_0\|_2 \gtrsim \frac{1}{\tau^{p-1}},
\]

where \( R_{T,\tau}(f) = \mathbb{E}_{(X,Y)}\{T_{\tau}(Y) - f(X)\}^2 \) is the population risk. The proof of this is similar to that for the Huber risk by using the fact that

\[
\left| \mathbb{E}[T_{\tau}(f_0(X) + \epsilon)|X] - f_0(X) \right| \lesssim \frac{1}{\tau^{p-1}}
\]

under Condition 2. The second key step is to show that the variance term grows linearly with \( \tau \) as we see from Lemma A.7 for the Huber estimator. Combining these with a modified version of Lemma A.2 yields an upper bound on the convergence rate. However, when the noise is symmetric, the LSE with truncated response cannot achieve the rate \( n^{-\frac{2\nu}{2\nu+2}} \) as in Theorem 3.5. The main reason is that the (population) excess risk lower bound (3.7) does not hold in this case because \( \mathbb{E}[T_{\tau}(f_0(X) + \epsilon)|X] \) is not necessarily equal to \( f_0(X) \).

Remark 3.6. The adaptive Huber ReLU-DNN estimator is easy to implement using the Python library TensorFlow. Specifically, we can use the function \texttt{tf.keras.losses.Huber} instead of mean-square loss \texttt{tf.keras.losses.MeanSquaredError}.

4 Lower bound under heavy-tailed error

This section provides a lower bound of the adaptive Huber ReLU-DNN regression estimator for any given neural network under a bounded (conditional) \( p \)-th moment. The relationships among the results in this section are depicted in Figure 1 and our organization of this section is as follows.

1. We present the setting and the main result (Theorem 4.1) of the lower bound on the convergence rate for adaptive Huber ReLU-DNN estimators in Section 4.1. Subsequently, we develop tools, including approximation theory of ReLU-DNN network, to prove the result.

2. In Section 4.2, we first provide some intuitive explanations on the three terms in the lower bound, namely, the bias induced by Huber loss, the approximation error of ReLU-DNN to \( d \)-variate \((\beta, 1)\)-smooth function, and the statistical error. Then we give the key ideas on the proof of the main result—Theorem 4.1, and present the lower bound on convergence rate for least squares ReLU-DNN estimator (Corollary 4.4).

3. In Section 4.3, we furnish a new ReLU-DNN approximation result that is related to ReLU-DNN’s non-robustness nature. This result plays a key role in analyzing the statistical error.
Figure 1: The connections of all the results presented in Section 4. The arrow from claim A to claim B means we will use claim A to prove claim B. The main result is presented in Theorem 4.1 and subsequently we develop tools, including approximation theory of ReLU-DNN, to prove it.

term in Section 4.2. We provide insights why the ReLU-DNN approximation result is related to a slower convergence rate under heavy-tailed errors.

(4) The above discussion focuses only on the best convergence rate for the adaptive Huber ReLU-DNN estimator under heavy-tailed errors. A natural question is that what is the best possible rate an ReLU-DNN estimator can achieve under the same scenario. We summarize our results and present some preliminary answers to this question in Section 4.4.

4.1 Main result on lower bound

Theorem 3.4 establishes an upper bound on the convergence rate of the adaptive Huber ReLU-DNN estimator under a bounded (conditional) $p$-th moment assumption. A natural question arises: is this upper bound tight under the assumed moment condition? To answer this question, in this section we provide several lower bounds for both ReLU-DNN least squares and Huber estimators under the same moment conditions.

Without loss of generality, in this section we assume $\|f_0\|_\infty \leq 1$, $X$ follows a uniform distribution on $[0, 1]^d$ and $\mathbb{E}[|e|^p | X = x] \leq 1$ (almost surely). Moreover, assume that the underlying regression function $f_0$ belongs to the class $C(d, \beta)$ given by

$$C(d, \beta) = \{ f : [0, 1]^d \rightarrow \mathbb{R} : \| f \|_\infty \leq 1, |\partial^\alpha f(x)| \leq \beta, x \in [0, 1]^d | \| \alpha \|_1 \leq \beta \}. \tag{4.1}$$

It should be noted that our analysis can also be extended to the case where $f_0 \in \mathcal{H}(l, \mathcal{P})$ with $\beta$, $d^*$ satisfying $(\beta^*, d^*) = \arg\min_{(\beta, d) \in \mathcal{P}} (\beta, d)$; see Remark 4.3 for details. To facilitate the presentation, we focus on $C(d, \beta)$ to make the proof simpler. Under the above settings, we first present a lower bound for the adaptive Huber ReLU-DNN estimator.

**Theorem 4.1** (A lower bound on the convergence rate for adaptive Huber estimator). For any $\beta, d \in \mathbb{N}^+$, let $X$ be uniformly distributed on $[0, 1]^d$ and $C(d, \beta)$ be the set of functions defined in (4.1). Let $S_n^{\text{RE}}(\delta)$ be the set of all the approximate adaptive Huber ReLU-DNN estimates with robustification parameter $\tau$, depth $L$ and width $\bar{N}$, i.e.,

$$S_n^{\text{RE}}(\delta) = \left\{ \tilde{f}_n \in \mathcal{F}_n(d, \bar{L}, \bar{N}, 1) : \tilde{R}_\tau (\tilde{f}_n) \leq \min_{f \in \mathcal{F}_n(d, L, \bar{N}, 1)} \inf_{f \in \mathcal{F}_n(d, L, \bar{N}, 1)} \tilde{R}_\tau (f) + c_{10} \delta^2 \right\}. \tag{4.2}$$
Then, for any $\epsilon > 0$, there exists some $n_\epsilon$ such that for any $N, L \geq c_{11}$, and $\tau \geq c_{12}$,

$$\sup_{f_0 \in C(d, \beta), \mathbb{E}[|X-x|] = 0, \mathbb{E}[|\epsilon| |X-x|] \leq 1} \mathbb{P} \left( \exists \hat{f}_n \in \mathcal{S}_n^{\text{HH}}(\delta_n), \|\hat{f}_n - f_0\|_2 \geq \delta_n \right) \geq 1 - \epsilon$$

holds for all $n \geq n_\epsilon$, where

$$\delta_n = \frac{1}{(NL)^{\beta/d} \log^{5\beta/d}(NL)} \sqrt{\left\{ \frac{NL}{\log(n)} \sqrt{n} \left( \tau \wedge \left( \frac{n \log^2 n}{(NL)^2} \right)^{1/p} \right) \wedge 1 \right\} \sqrt{\frac{1}{\tau^{p-1} \log^2 n}}}. \quad (4.3)$$

Here $c_{10} - c_{12}$ are positive constants independent of $n, L, N$ and $\tau$.

**Remark 4.1.** Note that our lower bound result holds for sufficiently large $\tau$, i.e., $\tau \geq c_{12}$. When $\tau < c_{12}$, by following the similar proof strategy, we can find some $f_0$ and asymmetric noise distribution such that $\|f_0 - f_0, \tau\|_2 \geq \frac{1}{\tau^{p-1}} \geq 1$, while $\|\hat{f}_n - f_0, \tau\|_2 \to 0$. But in this case, we will rely on a different convergence analysis of $\|\hat{f}_n - f_0, \tau\|_2$ because Proposition 3.2 does not necessarily hold. This will make our proof much longer. Therefore, we omit such a discussion to keep the proof simple.

Theorem 4.1 provides a lower bound on the adaptive Huber ReLU-DNN estimator that matches the upper bound we derived in Theorem 3.4 up to logarithmic factors by substituting the parameters $N, L$ and $\tau$ therein. Therefore for any $p \geq 2$, the rate of convergence in Theorem 3.4 is the fastest rate of convergence we can obtain using adaptive Huber ReLU-DNN estimator under heavy-tailed errors. To see this, because the lower bound (4.3) holds for arbitrary $N, L$ that are greater than or equal to some constant, and robustification parameter $\tau \geq 32$, then for large enough $n$, we can just plug in $N, L$ and $\tau$ in Theorem 3.4 to obtain its lower bound and easily verify that it has the same rate as $\delta_n$ in Theorem 3.4 up to logarithmic factors. This specific choice also corresponds to the fastest rate of convergence in the lower bound by letting

$$\frac{1}{\tau^{p-1} \log^2 n} \leq \frac{1}{(NL)^{\beta/d} \log^{5\beta/d}(NL)} \leq \frac{NL}{\log(n)} \sqrt{n} \sqrt{\frac{\tau}{\tau}}$$

i.e.,

$$\tilde{N}L \approx n^{\frac{d-\nu^*}{d+\beta - 1}} (\log n)^{\frac{\beta - \nu^*}{d+\beta - 1}} \quad \text{and} \quad \tau \approx n^{\frac{2(1-\nu^*)}{d+\beta - 1}} (\log n)^{\frac{\beta - \nu^*}{d+\beta - 1}},$$

and is given by $n^{-\beta/\nu^*(-d+1)} (\log n)^{\frac{\beta - \nu^*}{d+\beta - 1}}$, where $\nu^* = 1 - (2p - 1)^{-1}$.

It is worth noticing that the above lower bound is tailored to the Huber regression estimator trained on ReLU-DNN to reveal the impact of the tails of noise on ReLU-NN estimators. For the the Hölder class $C(\beta, d)$, Kuchibhotla & Patra (2019) show that the $C(\beta, d)$-constrained LSE achieves the optimal rate $n^{-\beta/\nu^*(-d+1)}$ when $p \geq 2 + d/\beta$. In Section 4.4, we will discuss whether ReLU-DNN-based estimators can achieve such optimal rate.

### 4.2 Explanations on the lower bound

From the previous upper bound analysis, we see that the overall convergence rate is determined by a trade-off among three terms: bias introduced by the Huber loss, neural network approximation error, and statistical error. Then, the estimation error can be controlled by

$$\|\hat{f}_n - f_0\|_2 \leq \|f_0 - f_0, \tau\|_2 + \|f_n - f_0, \tau\|_2 + \|\hat{f}_n - f_0\|_2,$$

where

- $B(\tau)$: Bias introduced by the Huber loss
- $A(N, L)$: Neural network approximation error
- $V_\epsilon(\tau, N, L)$: Statistical error
The result in Theorem 4.2 can be extended to an arbitrary function class as long as the corresponding minimax rate is known. For example, suppose the minimax rate of convergence over the nonparametric function class $G$ is $n^{-\alpha'}$, then it can be shown that

$$\sup_{f_0 \in G} \inf_{f \in F} \|f - f_0\|_{\infty} \geq (\bar{N}L)^2 \log(\bar{N}L)^{-\beta'/d},$$

for any $\epsilon > 0$. We refer to Schmidt-Hieber (2020) for the minimax estimation rate for a class of composite Hölder continuous functions.
Figure 2: An illustration on the construction of $f_0$, which is zero (blue), and noise $\varepsilon$ that follows a Trinomial distribution such that $y = \varepsilon$ (black circle), and $\tilde{f}$ (red function), which is capped at 1 due to the constraint $\|\tilde{f}\|_{\infty} \leq 1$. The red function has a better fit than the true function (blue) when $u$ is sufficiently small. This is related to the approximation ability of ReLU neural network with a given depth and width.

Statistical error. The next proposition establishes a lower bound for the statistical error term $V_n(\tau, \bar{N}, \bar{L})$. For this purpose, we let $X$ be uniformly distributed on $[0, 1]^d$ and $f_0 = 0$ and furnish a lower bound on $\|\hat{f}_n - f_0\|_2$. It serves as a supporting lemma to Theorem 4.1; see Figure 1.

Proposition 4.3 (Lower bound of convergence rate in the null case). Let $f_0 = 0$, $X$ be uniformly distributed on $[0, 1]^d$, and $M = 1$. Moreover, let $n \geq \sqrt{2(d+1)}$, $p \geq 2$ and $\tau \geq c_{12}$ be fixed, and suppose that $N, L$ are any positive integers satisfying $(NL)^2 \geq c_{14}$. Then, there exists some symmetric distribution of $\varepsilon$ with $\mathbb{E}[|\varepsilon|^p | X = x] \leq 1$ such that with probability at least $1 - c_{15} n \wedge (NL)^2$, there exists some approximate empirical risk minimizer $\tilde{f}_n$ (over the function class $\mathcal{F}_n$) satisfying

$$\hat{R}_{\tau}(\tilde{f}_n) \leq \min \left\{ \hat{R}_{\tau}(f_0), \inf_{\|f\|_{\infty} \leq 1} \hat{R}_{\tau}(f) + c_{16} \delta_n^2 \right\}$$

(4.4)

and the lower bound $\|\tilde{f}_n - f_0\|_2 = \|f_0\|_2 \geq \delta_n$, where $\mathcal{F}_n$ is either $\mathcal{F}_n(d, c_{17} L \log n, c_{18} N, 1)$ or $\mathcal{F}_n(d, c_{19} L, c_{20} N \log n, 1)$, and

$$\delta_n = \frac{NL}{\sqrt{n}} \left\{ \tau \wedge \left( \frac{n}{(NL)^2} \right)^{1/p} \right\} \wedge 1.$$  

Here $c_{12}$ is the constant in Theorem 4.1, $c_{14} - c_{20}$ are positive constants independent of $N, L, \tau, n$.

Remark 4.4. Note that $\inf_{\|f\|_{\infty} \leq 1} \hat{R}_{\tau}(f) \leq \inf_{f \in \mathcal{F}_n} \hat{R}_{\tau}(f)$. Thus, in Proposition 4.3 we show the existence of an approximate empirical risk minimizer $\tilde{f}_n$, instead of the exact minimizer, satisfying the desired lower bound. We believe this is an artifact of the proof technique. It still matches the upper bound stated in Theorem 3.4 because $\hat{f}_n$ therein only needs to be an approximate minimizer satisfying $\tilde{R}_{\tau}(\hat{f}_n) - \inf_{f \in \mathcal{F}_n} \tilde{R}_{\tau}(f) \leq \delta_n^2$ with high probability.

To gain the idea how such a lower bound is established, we consider the simplest case where $d = 1$ and use Huber loss over some uniformly bounded function class $\mathcal{F}$. It reveals that the approximation ability of ReLU neural network is essential in our construction of lower bound and
demystify how this is related to the statistical rate of convergence for heavy-tailed noises. To complete the proof of Proposition 4.3 for more general cases, we need to establish Theorem 4.6, which says a deep ReLU neural network with depth \( \bar{L} \) and width \( \bar{N} \) can approximate \( S = (\bar{N}\bar{L})^2 \) points arbitrarily in a highly non-smooth manner.

**Proof Sketch of Proposition 4.3.** Our target is to find a distribution of \( \varepsilon \) satisfying \( \mathbb{E}[|\varepsilon|^p | X = x] \leq 1 \) such that there exists some \( \tilde{f} \in \mathcal{F} \) satisfying (with \( f_0 = 0 \))

\[
\|\tilde{f} - f_0\|_2 \geq \delta_n \quad \text{and} \quad \mathcal{R}_\varepsilon(\tilde{f}) - \mathcal{R}_\varepsilon(f_0) \leq \inf_{\|f\|_2 \leq 1} \mathcal{R}_\varepsilon(f) + \delta_n^2. \tag{4.5}
\]

Let the random noise

\[
\varepsilon = \begin{cases} 
\left(\frac{u}{S}\right)^{1/p} & \text{with probability } \frac{S}{2n}, \\
-\left(\frac{u}{S}\right)^{1/p} & \text{with probability } \frac{S}{2n}, \\
0 & \text{with probability } 1 - \frac{S}{n},
\end{cases}
\]

be independent of \( X \sim \text{Uniform}[0, 1] \). The observed data are \( \{(X_i, Y_i)\}_{i=1}^n \) with \( Y_i = \varepsilon_i \) satisfying \( \mathbb{E}[|\varepsilon_i|^p | X_i] = 1 \).

If \( S = o(n) \), then by a concentration result, we have that there are approximately \( k = S \) samples with non-zero \( \varepsilon_i \) while the rest \( n - k = n - S \) samples all have zero \( \varepsilon_i \). Without loss of generality, let \( |\varepsilon_1| = |\varepsilon_2| = \cdots = |\varepsilon_k| = (n/S)^{1/p} \) and \( \varepsilon_{k+1} = \varepsilon_{k+2} = \cdots = \varepsilon_n = 0 \). Now that we need to find some \( \tilde{f} \) satisfying (4.5). The key idea is construct a red function as shown in Figure 2 in which \( \tau \) is fixed. Given the large outliers, fitting those by sign(\( \varepsilon_i \)) (the maximum magnitude allowed in our function class) has much smaller losses than by zero (true value), and they can even compensate the loses elsewhere by fitting a non-zero constant \( u \). As a result, the red function has a better fit to the data than the true function (blue), yet \( \|\tilde{f} - f_0\|_2 \geq u \), satisfying (4.5).

Let us now formally implement the above idea. For an arbitrary \( \Delta > 0 \), suppose we can take \( \tilde{f} \in \mathcal{F} \) such that \( \tilde{f}(X_i) = \text{sgn}(\varepsilon_i) \) for \( i \in \{1, \ldots, k\} \), and \( \tilde{f}(x) \equiv u \) in \([0, 1] \setminus \bigcup_{i=1}^k (X_i - \Delta, X_i + \Delta) \). Then, we have \( \|\tilde{f} - f_0\|_2 \geq u \) if \( \Delta \) is small. If we further require that \( \Delta < \min_{i \neq j} |X_i - X_j| \) with probability tending to 1, then \( \tilde{f}(X_i) = u \) for all \( i \in \{k + 1, \ldots, n\} \). By a second-order Taylor expansion, we have

\[
\mathcal{R}_\varepsilon(f) = \frac{1}{n} \sum_{i=1}^k \left( \psi_{\varepsilon_i} f(X_i) + \frac{1}{2} \tilde{f}(X_i)^2 \right) + \sum_{i=k+1}^n \frac{1}{2} \tilde{f}(X_i)^2 \biggr) 
\leq -k \left( \frac{u}{n} \right)^{1/p} \Delta + \frac{k}{2n} \Delta^2 + \frac{1}{2} u^2 \leq \frac{S}{n} \left( \frac{n}{S} \right)^{1/p} \Delta \biggr],
\]

as long as \( \tau \geq 1 \). Thus, \( \mathcal{R}_\varepsilon(\tilde{f}) \leq \mathcal{R}_\varepsilon(f_0) \) if \( u \leq \sqrt{\frac{S}{n}} \left( \frac{n}{S} \right)^{1/p} \Delta \). Now by choosing \( \delta_n \equiv u \equiv \sqrt{\frac{S}{n}} \left( \frac{n}{S} \right)^{1/p} \Delta \), we also have

\[
\mathcal{R}_\varepsilon(f_0) - \inf_{\|f\|_2 \leq 1} \mathcal{R}_\varepsilon(f) = \frac{2k}{n} \left( \frac{n}{S} \right)^{1/p} \Delta - \frac{k}{n} \leq \delta_n^2.
\]

Therefore, Claim (4.5) holds with \( \delta_n \equiv \sqrt{\frac{S}{n}} \left( \frac{n}{S} \right)^{1/p} \Delta \) and \( 1 \). \( \square \)

In Section 4.3, we will show further that for multiple dimension \( (d \geq 2) \), the class of deep ReLU neural network with depth \( \bar{L} \) and width \( \bar{N} \) shares a similar approximation ability as \( \mathcal{F} \) when \( S = (\bar{L}\bar{N})^2 \), so Proposition 4.3 can be proved in a similar way.
By combining Theorem 4.2 and Proposition 4.3, we also provide a lower bound for the convergence rate of the least squares estimator. This corresponds to the case that $\tau = \infty$ in Theorem 4.1.

**Corollary 4.4** (Convergence rate for least squares estimator, lower bound). For any $d, \beta \in \mathbb{N}^+$, let $X$ be uniformly distributed on $[0, 1]^d$ and $C(d, \beta)$ be the set of functions defined in (4.1). Consider the set $S_n^L(\delta_n)$ of all the approximate least squares ReLU-DNN estimates with depth $\bar{L}$ and width $\bar{N}$, i.e.,

$$S_n^L(\delta_n) = \left\{ \hat{f}_n \in \mathcal{F}(d, \bar{L}, \bar{N}, 1) : \sqrt{\mathcal{R}(\hat{f}_n)} \leq \min \left\{ \sqrt{\mathcal{R}(f_0)} \quad \min_{f \in \mathcal{F}(d, \bar{L}, \bar{N}, 1)} \sqrt{\mathcal{R}(f)} + c_21(\delta_n^2) \right\} \right\} \quad (4.6)$$

Then, for any $\epsilon > 0$, there exists some $n_\epsilon$ such that for any $\bar{N}, \bar{L} \geq c_{11}$,

$$\sup_{f_0 \in \mathcal{C}(d, \beta), \mathbb{E}[\epsilon(x) = 0, \mathbb{E}[\epsilon(x)|X = x] \leq 1} \mathbb{P}\left\{ \exists \hat{f}_n \in S_n^L(\delta_n) \text{ s.t. } ||\hat{f}_n - f_0||_2 \geq \delta_n \right\} \geq 1 - \epsilon$$

holds for all $n \geq n_\epsilon$, where

$$\delta_n = \frac{1}{(\bar{N}\bar{L})^{\alpha/d} \log^{5\beta/d}(\bar{N}\bar{L})} \vee \left\{ \left( \frac{(\bar{N}\bar{L})^2}{n(\log n)^2} \right)^{1 - \frac{1}{p^\beta}} \wedge 1 \right\}. \quad (4.7)$$

Here $c_{11}$ is the constant in Theorem 4.1, and $c_{21}$ is a positive constant independent of $n, \bar{L}, \bar{N}$.

**Remark 4.5.** Corollary 4.4 reveals the best convergence rate that a least squares ReLU-DNN estimator can achieve. According to the approximation error and stochastic error trade-off in (4.7), the best convergence can be obtained by letting $\bar{N}, \bar{L}$ satisfy

$$\bar{N}\bar{L} \asymp n^{\frac{1}{1 - \frac{1}{p\beta}} - \frac{1}{p\beta}} (\log n)^{\frac{1 - \frac{1}{p\beta} - \frac{3}{2}}{1 - \frac{1}{p\beta}}},$$

and is given by

$$n^{\frac{1 - \frac{1}{p\beta}}{1 - \frac{1}{p\beta} + \frac{2}{p\beta}} (\log n)^{-\frac{1}{1 - \frac{1}{p\beta} + \frac{2}{p\beta}}} \cdot \frac{\log n}{\bar{N}\bar{L}}^{\frac{1}{p\beta}},$$

This is slower than that of the adaptive Huber estimator in (3.17).

**Bias.** Suppose $X$ and $\epsilon$ are independent. It is easy to construct an asymmetric noise $\epsilon$ such that $\|\hat{f}_n - f_0\|_2 \gtrsim \frac{1}{\tau^{1/2}}$, so the difficulty here is to show that $\|\hat{f}_n - f_0\|_2 \gtrsim \|f_0 - f_0, r\|_2 \gtrsim \frac{1}{\tau^{1/2}}$. The key idea of the bias analysis is that if we choose $f_0 = 0$ and let $\epsilon$ be independent of $X$, then $f_0, r(x) \equiv C_\tau$ and $\|\hat{f}_n - f_0, r\|_2$ can be upper bounded by $\sqrt{\frac{\bar{N}\bar{L}}{n}} \tau$ up to logarithmic terms. This result is formally stated in Lemma C.5. We now divide the discussion in two cases.

1. $\hat{f}_n$ converges to $f_0$ faster than $f_0, r$ does, i.e., $\sqrt{\frac{\bar{N}\bar{L}}{n}} \tau \leq \frac{1}{\tau^{1/2}}$. In this case, using our construction of $\epsilon$ and the triangle inequality, the $L_2$ risk satisfies $\|\hat{f}_n - f_0\|_2 \gtrsim \|f_0 - f_0, r\|_2 - \|f_0, r - \hat{f}_n\|_2 \gtrsim \frac{1}{\tau^{1/2}}$.

2. $\hat{f}_n$ converges to $f_0$ slower than $f_0, r$ does, i.e., $\sqrt{\frac{\bar{N}\bar{L}}{n}} \tau \gtrsim \frac{1}{\tau^{1/2}}$. In this case the bias term is no longer the dominating term. Indeed, by some calculations, it is easy to show that $V_n(\bar{N}, \bar{L}, \tau) \gtrsim B(\tau)$. Using the construction of $f_0$ and $\epsilon$ in Proposition 4.3, we can show that $\|\hat{f}_n - f_0\|_2 \gtrsim V_n(\bar{N}, \bar{L}, \tau) \gtrsim B(\tau)$. 

18
4.3 Revisiting neural network approximation

The sketch of proof in Section 4.2 connects the lower bound on the convergence rate to a specific approximation ability of the function class \( \mathcal{F} \) as shown in Figure 2. To be specific, suppose the function class \( \mathcal{F} \) has the approximation ability that it can approximate any \( S = o(n) \) points \( \{(x_i,y_i)\}_{i=1}^S \), with \( y_i \) being arbitrarily chosen from \([-1, +1]\), in a way that it stays as a constant

\[
u = \sqrt{\frac{d}{n}} \left( \frac{n}{S} \right)^{1/p} \tau \approx \delta_n \quad \text{except in the region } \bigcup_{i=1}^S B_{o\nu}(x_i, \Delta)
\]

for a sufficiently small \( \Delta > 0 \), where \( B_{o\nu}(x, r) = \{y : \|y - x\|_\infty \leq r\} \) is the hypercube with the half-width \( r \). Then the convergence rate will be lower bounded by \( \delta_n \).

However, it is unclear whether aforementioned approximation ability holds. Yet, we are able to show a weak version the approximation ability by relaxing the requirements in two aspects: the \( S \) points \( \{x_i\} \) are located in \( \mathcal{S} \) distinct sub-cubes in \([0, 1]^d \) (see Figure 3(a)) and either the width or the depth can depend logarithmically on \( \Delta \). Fortunately it turns out that this weaker approximation ability is enough to prove Proposition 4.3 using a similar strategy as outlined in Section 4.2.

The next theorem claims a deep ReLU neural network with depth \( \bar{L} \) and width \( \bar{N} \) can fit any piece-wise constant function in \((\bar{N}\bar{L})^2\) sub-hypercubes, and that establishes a foundation to prove Theorem 4.6, which is the approximation ability used in Proposition 4.3.

**Theorem 4.5.** For any given \( N, L \in \mathbb{N}^+ \), let \( K = [N^{1/d} 2L^{1/d} 2^2] \), and \( \{y_a\}_{a \in \mathcal{A}} \subseteq [0, 1] \) be an arbitrary set of values indexed by \( \mathcal{A} = \{1, \ldots, K\} \). For any tolerance parameter \( \Delta \in (0, 1/(3K)] \), and precision parameter \( \epsilon \in (0, 1) \), let

\[
Q_\epsilon(\Delta) = \left\{ x = (x_1, \cdots, x_d) : (a_i - 1)/K \leq x_i \leq a_i/K - 1_{[a_i < K]}/\Delta \right\};
\]

Then, there exist a deep ReLU neural network \( f_1^s \) with depth \( (5L + 7)(\lceil \log_2(1/\epsilon) \rceil + 2) \) and width \( (4N + 3)d \lor (8N + 10) \), and a deep ReLU neural network \( f_2^s \) with depth \( 9L + 12 \) and width \( (4N + 3)d \lor (8N + 6)(\lceil \log_2(1/\epsilon) \rceil + 1) \) such that

\[
|f_1^s(x) - y_a| \leq \epsilon \quad \text{for all } x \in Q_\epsilon; \quad s = 1, 2.
\]

Moreover, if \( y_a = \sum_{i=1}^d 2^{-\theta_i} \) for \( \theta = (\theta_1, \ldots, \theta_d) \in [0, 1]^d \) with \( r \leq \lceil \log_2(1/\epsilon) \rceil \), we have \( f_1^s(x) = y_a \) instead of \( |f_1^s(x) - y_a| \leq \epsilon \) in (4.9). In this case, the term \( \lceil \log_2(1/\epsilon) \rceil + 1 \) in the width and depth can further be reduced to \( r \) if all the \( y_a \) can be written as the above form.

Theorem 4.5 indicates that for any \( N, L \in \mathbb{N}^+ \), if we divide the unit cube \([0, 1]^d \) uniformly into \( K^d \) sub-cubes with length \( \frac{1}{K} \) for \( K = (NL)^{2/d} \), then it is possible to find a deep ReLU neural network \( f^\dagger \) with depth \( \leq L \) and width \( \leq N \) such that \( f^\dagger \) is approximately equal to an arbitrary pre-specified value for each sub-cube except in a small “bad” region, i.e., \([0, 1]^d \setminus \bigcup_{a \in \mathcal{A}} Q_\epsilon(\Delta) \). Moreover, such bad region can be arbitrarily small, and in order to achieve \( \epsilon \)-accuracy, we only need to multiply a factor of \( \log(1/\epsilon) \) to either the width or the depth of the neural network.

Figure 3(a) illustrates the result of Theorem 4.5. In the case where \( d = 2, N^2 L^2 = 16 \), we divide \([0, 1]^2 \) into \( K^2 \) sub-squares of equal size with \( K = 4 \), so that each sub-square has length \( 1/K = 0.25 \). The values of \( f^\dagger \) in the “bad” region (green) can be arbitrary. However, in the blue regions with index \( \alpha \in \{1, \ldots, 4\} \), we can find some \( f^\dagger \) such that it approximates the pre-specified \( y_a \) in each sub-square.

Based on Theorem 4.5, we are ready to give a formal statement of the approximation ability of a deep ReLU neural network.
In this section, we summarize the main results obtained in this paper and leave an interesting open question. Assume that the regression function $f_0 \in C(d, \beta)$, and the noise variable satisfies $\mathbb{E}[|e|^p | X = 0] < \infty$, where $1 \leq p < \infty$.

\section{Neural network meets heavy-tailed error: an comprehensive picture.}

In this section, we summarize the main results obtained in this paper and leave an interesting open question. Assume that the regression function $f_0 \in C(d, \beta)$, and the noise variable satisfies $\mathbb{E}[|e|^p | X = 0] < \infty$, where $1 \leq p < \infty$. Theorem 4.6 shows that if the $S = K^d \approx L^2 N^2$ points $\{x_i\}_{i=1}^S$ are located on distinguished regions of $K^d$ sub-cubes with length $1/K$ in $[0, 1]^d$, then we can find some $f^* \in \mathcal{F}(d, N, L, 1)$ such that the value of $f^* (x_i)$ equals a pre-specified constant in $[-1, +1]$ respectively for $i = 1, \cdots, S$, while maintaining constant outside the cubes (with length $2\Delta_2$) of these $N^2 L^2$ points and a “bad” region $[0, 1]^d \setminus Q$. Note that $\Delta_1 \in (0, 1/(3K)]$ can be arbitrarily chosen, and the network depth or width only depends on $\Delta_2$ logarithmically. Figure 3(b) gives an example of such kind of approximation ability when $d = 1$. 

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3}
\caption{Illustrative explanations of Theorems 4.5 and 4.6. The left panel shows that there exists a deep ReLU neural network with depth $L$ and width $N$ that is approximately piecewise constant in $N^2 L^2 \approx K^d$ blue regions with pre-specified values except green regions. Right panel indicates that there exists a deep ReLU neural network that takes value 1 at $x_k$ and value −1 at $x_{k+2}$ and constant value $u$ in other regions except in small regions around the points $x_k$ and $x_{k+2}$ with radius $\Delta_2$ (red) and regions $(j/K - \Delta_1, j/K) (j = k, k + 1, k + 2)$ (green). The second part of exception region is similar to the green region in the left panel for 1-dimension.

Theorem 4.6. For any given $N, L \in \mathbb{N}^+$, letting $K = [N^{1/d}]^2 [L^{1/d}]^2$, then for any $\Delta_1 \in (0, 1/(3K)]$, $\Delta_2 > 0$, suppose $(x_{\alpha}, y_{\alpha})_{\alpha \in \tilde{A}}$ is a set of arbitrary points whose index set $\tilde{A}$ is a subset of $\{1, \cdots, K\}^d$. Each element $(x_{\alpha}, y_{\alpha})$ satisfies $x_{\alpha} \in Q_{\alpha}(\Delta_1)$ defined by (4.8) and $y_{\alpha} \in \{-1, 1\}$. Then there exist some constants $c_{22} - c_{25}$ independent of $N, L, \Delta_1, \Delta_2$ such that for any $u \in [-1, 1]$, we can find a deep ReLU neural network $f_s^\dagger$ with depth $c_{22} L \log_2(1/\Delta_2)$ and width $c_{23} N$ and a deep ReLU neural network $f_s^\ddagger$ with depth $c_{24} L$, $c_{25} N \log_2(1/\Delta_2)$ satisfying

$$f_s^\dagger (x_{\alpha}) = y_{\alpha} \text{ for all } \alpha \in \tilde{A}, \quad s = 1, 2,$$

and

$$f_s^\ddagger (x) = u \text{ if } x \in Q, \text{ and } \|x - x_{\alpha}\|_\infty \geq \Delta_2 \text{ for all } \alpha \in \tilde{A}, \quad s = 1, 2,$$

where $Q = \bigcup_{\alpha \in \{1, \cdots, K\}^d} Q_{\alpha}(\Delta_1)$.
Figure 4: Comparison of rate of convergence $\alpha(p,d,\beta)$, given in Table 1, for the least squares estimator in Hölder function class (red), the adaptive Huber ReLU-DNN estimator (blue), and the least squares ReLU-DNN estimator (green) as a function of $p$. Here, we take $\beta = 2$ and $d = 1$ (dash) and $d = 3$ (solid).

$x \leq 1$ for some $p \geq 2$. Table 1 summarizes our main results, and for comparison purposes, also includes an upper bound result on the least squares estimator (LSE) for Hölder classes (Kuchibhotla & Patra, 2019). Our paper provides proofs of lower bound and upper bound for adaptive Huber ReLU-DNN estimator and lower bound for least squares ReLU-DNN estimator. The proof of upper bound for least squares ReLU-DNN estimator is similar and thus omitted. Clearly, the empirical adaptive Huber minimizer outperforms the corresponding least-squares estimator under the same moment conditions. Figure 4 depicts how the optimal rate changes with the moment index $p$ for each estimator.

![Comparison of rate of convergence](image)

Table 1: A summary of convergence rates for nonparametric estimators when $f_0 \in C(d,\beta)$ and $E(|\varepsilon|^p X)$ is bounded.

| Function class       | Empirical minimizer | Upper bound                          | Lower bound                          |
|----------------------|---------------------|--------------------------------------|--------------------------------------|
| Hölder class         | Square loss         | $n^{-\beta/2(2+\beta)} (p \geq 2)$ | $n^{-\beta/p} (p \geq 2)$           |
| ReLU neural network  | Square loss         | $n^{-\frac{1}{p}} (p \geq 2)$       | $n^{-\frac{1}{p}} (p \geq 2)$       |
| ReLU neural network  | Huber loss          | $n^{-\frac{1}{2p} \pm \frac{1}{4p}} (p \geq 2)$ | $n^{-\frac{1}{2p} \pm \frac{1}{4p}} (p \geq 2)$ |

A natural question is whether we can construct a robust neural network regression estimator that achieves the optimal rate $n^{-\beta/(2\beta+d)}$ under heavy-tailed errors. The answer to this question is related to how well neural networks can approximate a function with additional smoothness control. For example, if we are able to approximate any $f_0 \in C(d,\beta)$ using $\mathcal{F}_n^{(C)} = \{f \in \mathcal{F}_n : |f(x) - f(y)| \leq C|x - y|\}$ with some constant $C$ independent of $(N,L)$, while maintaining similar approximation error, i.e.,

$$
\sup_{f_0 \in C(d,\beta)} \inf_{f \in \mathcal{F}_n^{(C)}} \|f_0 - f\|_2 \leq \sup_{f_0 \in C(d,\beta)} \inf_{f \in \mathcal{F}_n} \|f_0 - f\|_2,
$$

then we only need to consider the estimator

$$
\tilde{f}_n \in \arg\min_{f \in \mathcal{F}_n^{(C)}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.
$$

Now all the functions in class $\mathcal{F}_n^{(C)}$ are $C$-Lipschitz functions. Thus the corresponding LSE is expected to achieve a similar convergence rate as the LSE for Hölder class discussed in Kuchibhotla
& Patra (2019). Using a similar argument it can be shown that the LSE defined in (4.11) achieves the rate $n^{-\frac{d}{2p}}$ when $p \geq 2 + d$, if (4.10) holds. Even if (4.10) is valid, we will lose the successful implementations of neural network training algorithms when we optimize (4.11).

5 Conclusion

In this paper, we have studied how the heavy-tailed errors impact on the rate of convergence of nonparametric regression estimators fitted on deep ReLU neural networks. We consider the adaptive Huber estimator and establish non-asymptotic error bounds on the $L_2$ risk. By presenting a matching lower bound, we further show that this is the best possible convergence rate the adaptive Huber ReLU-DNN estimator can obtain, which is faster than that of the least squares counterpart. This provides a comprehensive picture of the stability of deep ReLU neural networks under polynomial-tail errors.

Several interesting questions are worth exploring further. The first one is whether it is possible to obtain a ReLU-DNN estimator that achieves the minimax-optimal rate over the class of Lipschitz-continuous functions that have a hierarchical composition structure when the noise distribution is heavy-tailed. Our paper sheds a promising direction on this question, which can be divided into two steps. The first step is to develop a neural network approximation result (4.10) showing that constraining deep ReLU neural networks with a bounded Lipschitz constant does not reduce much the approximation error. If so, then for heavy-tailed noise distribution with $p \geq 2 + d$, the estimator in (4.11) can achieve the same rate of convergence as if the noise is sub-Gaussian. The second step is to develop a computationally feasible method to compute such an estimator (4.11). Note that the above estimator only attains the minimax-optimal rate when $p \geq 2 + d$. Even the above program is successful, the optimal rate for regression functions when $2 \leq p < 2 + d$ remains unknown. In addition, regardless of validity of (4.10), it is still interesting to understand the extend to which the ‘regularized’ estimator smooth-estimator can provide stability.

References

Anthony, M. & Bartlett, P. L. (1999). Neural Network Learning: Theoretical Foundations. Cambridge University Press.

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

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. Machine learning, 14(1), 115–133.

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

Bauer, B. & Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. The Annals of Statistics, 47(4), 2261–2285.
Beaton, A. E. & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2), 147–185.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’IHP Probabilités et statistiques*, volume 48 (pp. 1148–1185).

Chernozhukov, V., Chetverikov, D., & Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4), 1564–1597.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368), 829–836.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4), 303–314.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association*, 87(420), 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The annals of Statistics*, (pp. 196–216).

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall.

Fan, J., Li, Q., & Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(1), 247.

Fan, J., Wang, W., & Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, 49(3), 1239.

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213.

Feller, W. (2008). *An introduction to probability theory and its applications*, volume I. John Wiley & Sons.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19(1), 1–67.

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The annals of statistics*, 5(5), 815–841.

Han, Q. & Wellner, J. A. (2018). Robustness of shape-restricted regression estimators: An envelope perspective. *arXiv preprint arXiv:1805.02542*. 


Han, Q. & Wellner, J. A. (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics, 47*(4), 2286–2319.

Hernan Madrid Padilla, O., Tansey, W., & Chen, Y. (2020). Quantile regression with relu networks: Estimators and minimax rates. *arXiv e-prints*, (pp. arXiv–2010).

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

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics, 35*(1), 73–101.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The annals of statistics*, (pp. 799–821).

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

Kuchibhotla, A. K. & Patra, R. K. (2019). On least squares estimation under heteroscedastic and heavy-tailed errors. *arXiv preprint arXiv:1909.02088*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436–444.

Lederer, J. (2020). Risk bounds for robust deep learning. *arXiv preprint arXiv:2009.06202*.

Lu, J., Shen, Z., Yang, H., & Zhang, S. (2020). Deep network approximation for smooth functions. *arXiv preprint arXiv:2001.03040*.

McCaffrey, D. F. & Gallant, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks, 7*(1), 147–158.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications, 9*(1), 141–142.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function (with discussion). *The Annals of Statistics, 48*(4), 1875–1921.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., & Huang, J. (2021a). Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*.

Shen, G., Jiao, Y., Lin, Y., & Huang, J. (2021b). Robust nonparametric regression with deep neural networks. *arXiv preprint arXiv:2107.10343*.

Shen, Z., Yang, H., & Zhang, S. (2019). Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*.

Shen, Z., Yang, H., & Zhang, S. (2022). Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157, 101–135.

Sun, Q., Zhou, W.-X., & Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association, 115*(529), 254–265.
Suzuki, T. (2018). Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. arXiv preprint arXiv:1810.08033.

Telgarsky, M. (2016). Benefits of depth in neural networks. In Conference on learning theory (pp. 1517–1539).: PMLR.

Tsybakov, A. B. (2009). Introduction to Nonparametric Estimation. Springer.

Wahba, G. (1990). Spline models for observational data. SIAM.

Watson, G. S. (1964). Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, (pp. 359–372).

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

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep relu networks. In Conference on Learning Theory (pp. 639–649).: PMLR.

Zhong, Q., Mueller, J., & Wang, J.-L. (2022). Deep learning for the partially linear cox model. The Annals of Statistics.
A Proof for Section 3

A.1 Technical lemmas

We first revisit Talagrand’s inequality for suprema of empirical processes, and then provide a tail probability for the nonparametric adaptive Huber estimator.

Lemma A.1 (Talagrand inequality for suprema of empirical process). Let \( Z_1, Z_2, \ldots, Z_n \) be i.i.d. random variables from some distribution \( P \). Let \( G \) be a function class such that \( \mathbb{E}g(Z) = 0 \) for any \( g \in G \). Suppose \( \sup_{g \in G} \|g\|_\infty \leq U < \infty \) and let \( \sigma^2 \) satisfy \( \sigma^2 \geq \sup_{g \in G} \mathbb{E}g^2(Z) \). Set

\[
V = \sigma^2 + U \cdot \mathbb{E}\left\{ \sup_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} g(Z_j) \right| \right\}.
\]

Then, for any \( x > 0 \),

\[
\mathbb{P}\left\{ \sup_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} g(Z_j) \right| \geq \mathbb{E}\left\{ \sup_{g \in G} \left| \frac{1}{n} \sum_{j=1}^{n} g(Z_j) \right| \right\} + \sqrt{2Vx} + \frac{Ux}{3} \right\} \leq 2e^{-nx}.
\] (A.1)

Lemma A.2 (Tail probability for adaptive Huber estimator). Assume either Condition 2 holds or \( \tau \leq 1 \). Let \( R_\tau \) and \( \hat{R}_\tau \) be the population and empirical Huber risks, respectively, parameterized by \( \tau = \tau_n \geq 1 \). For any \( f^* \in \Theta \) given, define \( \Theta_\tau(r) = \{ f \in \Theta : \|f - f^*\|_2 \leq r \} \) and \( \Theta_\tau^*(r) = \{ f \in \Theta : \|f - f^*\|_2 > r \} \). Given some positive sequence \( \{\delta_n\}_{n \geq 1} \), assume that

1. there exists a sequence of functions \( f_n \in \mathcal{F}_n \) satisfying \( \|f_n - f^*\|_2 \leq 3\delta_n \) and

\[
R_\tau(f_n) - R_\tau(f^*) \leq 9\delta_n^2;
\] (A.2)

2. there exists some constant \( c_{26} > 0 \) such that for any \( \delta \geq \delta_n \),

\[
R_\tau(f) - R_\tau(f^*) \geq c_{26}\delta^2 \quad \text{for all} \quad f \in \Theta_\tau^*(\delta/2);
\] (A.3)

3. there exists a function \( \phi_n : \mathbb{R}^+ \to \mathbb{R}^+ \) such that for any \( \delta \geq \delta_n \),

\[
\mathbb{E}\left\{ \sup_{f \in \mathcal{F}_n, \theta \in \Theta_\tau^*(\delta)} |\Delta_n(f)| \right\} \leq c_{27}\phi_n(\delta),
\] (A.4)

for some constant \( c_{27} > 0 \) independent of \( \delta, n \), and the function \( \phi_n(\delta) \) satisfies \( \phi_n(\alpha\delta) \leq \alpha\phi_n(\delta) \) for all \( \alpha \geq 1 \), where

\[
\Delta_n(f) = \left( \hat{R}_\tau(f) - \hat{R}_\tau(f^*) \right) - \left[ R_\tau(f) - R_\tau(f^*) \right];
\] (A.5)

4. the function \( \phi_n(\delta) \) satisfies \( \delta_n^2 \geq c_{28}\phi_n(\delta_n) \) for some constant \( c_{28} > 0 \);

5. \( \delta_n \) and \( \tau \) are such that \( n\delta_n^2 \geq c_{29}\tau \) for some constant \( c_{29} > 0 \).

Then, let \( \mathcal{S}(\delta) = \{ \tilde{f} \in \mathcal{I} \in \mathcal{F}_n : \hat{R}_\tau(\tilde{f}) \leq \inf_{f \in \mathcal{F}_n} \hat{R}_\tau(f) + \frac{c_{26}}{\sqrt{26}}\delta_2 \} \), for \( D \geq \max\{2, \frac{6\sqrt{2}}{\sqrt{26}}, \frac{32}{c_{26}c_{28}}\} \),

\[
\mathbb{P}\left\{ \exists \tilde{f} \in \mathcal{S}(D\delta_n) \text{ s.t. } \|\tilde{f} - f^*\|_2 \geq D\delta_n \right\} \leq c_{31} \exp\left\{ -c_{30} \frac{n\delta_n^2D^2}{M(\tau + M)} \right\}
\] (A.6)

where \( c_{30} = c_{26}/54, c_{31} = 2(1 - e^{-4c_{26}c_{30}/(M+M^2)})^{-1} \) are constants independent of \( n, \delta_n \) and \( D \).
Proof of Lemma A.2. The proof is based on the peeling argument complemented with Talagrand’s inequality. For integers \( j = 1, 2, \ldots \), define the donut-shaped sets

\[
S_{n,j} = \{ f \in \mathcal{F}_n : 2^{j-1}D\delta_n < \| f - f^* \|_2 \leq 2^jD\delta_n \},
\]

so that \( \mathbb{P}\{ \exists f_n \in S(D\delta_n) \text{ s.t. } \| f_n - f^* \|_2 > D\delta_n \} = \sum_{j=1}^{\infty} \mathbb{P}\{ \exists f_n \in S(D\delta_n) \text{ s.t. } f_n \in S_{n,j} \} \). It suffices to bound each probability \( \mathbb{P}\{ \exists f_n \in S(D\delta_n) \text{ s.t. } f_n \in S_{n,j} \} \) separately. By definition, any \( f_n \in S_{n,j} \) satisfies \( \| f_n - f^* \|_2 > 2^{j-1}D\delta_n \geq D\delta_n \geq \delta_n \). It follows from condition (2) that

\[
c_{26}2^{2j-2}D^2\delta_n^2 \leq c_{26}\| f_n - f^* \|_2^2 \leq R_\tau(f_n) - R_\tau(f^*). \tag{A.7}
\]

Conditioning on \( \{ \| f_n - f^* \|_2 \leq 2^jD\delta_n \} \) and \( f_n \in S(D\delta_n) \), we have

\[
R_\tau(f_n) - R_\tau(f^*) = R_\tau(f_n) - R_\tau(f_n) + R_\tau(f_n) - R_\tau(f_n) + R_\tau(f_n) - R_\tau(f^*) \\
\leq R_\tau(f_n) - R_\tau(f_n) + \frac{c_{26}}{8}D^2\delta_n^2 + R_\tau(f_n) - R_\tau(f_n) + R_\tau(f_n) - R_\tau(f^*)
\]

Note that condition (1) ensures \( \| f_n - f^* \|_2 \leq 3\delta_n \leq 2D\delta_n \leq 2^jD\delta_n \) for all \( j \geq 1 \), which further implies

\[
R_\tau(f_n) - R_\tau(f_n) + R_\tau(f_n) - R_\tau(f_n) \leq 2 \sup_{f \in \mathcal{F}_n \cap \Theta_\tau(2^jD\delta_n)} |\Delta_n(f)|.
\]

Combining the above calculations with the fact \( R_\tau(f_n) - R_\tau(f^*) \leq 9\delta_n^2 \), we have

\[
R_\tau(f_n) - R_\tau(f^*) \leq (9 + c_{26}D^2/8)\delta_n^2 + 2 \sup_{f \in \mathcal{F}_n \cap \Theta_\tau(2^jD\delta_n)} |\Delta_n(f)|.
\]

This together with the lower bound (A.7) yields that conditioned on the event \( \{ \exists f_n \in S(D\delta_n) \text{ s.t. } f_n \in S_{n,j} \} \),

\[
(c_{26}2^{2j-2}D^2 - c_{26}D^2/8 - 9)\delta_n^2 \leq 2 \sup_{f \in \mathcal{F}_n \cap \Theta_\tau(2^jD\delta_n)} |\Delta_n(f)|.
\]

Provided \( D^2 \geq 72/c_{26} \), it holds for each \( j \geq 1 \) that

\[
\frac{c_{26}}{8}D^2 + 9 \leq \frac{1}{2}(c_{26}2^{-2} + c_{26}2^{-2})D^2 \leq c_{26}2^{2j-4}D^2.
\]

Substituting this into the earlier inequality gives

\[
c_{26}2^{2j-4}D^2\delta_n^2 \leq \sup_{f \in \mathcal{F}_n \cap \Theta_\tau(2^jD\delta_n)} |\Delta_n(f)|.
\]

Setting \( \delta = D\delta_n \geq 2\delta_n \), we conclude that

\[
\mathbb{P}\{ \exists f_n \in S(D\delta_n) \text{ s.t. } f_n \in S_{n,j} \} \leq \mathbb{P}\left\{ \sup_{f \in \mathcal{F}_n \cap \Theta_\tau(2\delta)} |\Delta_n(f)| \geq c_{26}2^{2j-4}\delta^2 \right\}.
\]
Next we apply Talagrand’s inequality to bound the supremum sup \( f |\Delta_n(f)| \). Note that
\[
\Delta_n(f) = (\mathcal{R}_\varepsilon(f) - \mathcal{R}_\varepsilon(f^*)) - (\mathcal{R}_\varepsilon(f) - \mathcal{R}_\varepsilon(f^*))
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \ell_\varepsilon(e_i + (f_0 - f)(X_i)) - \ell_\varepsilon(e_i + (f_0 - f^*)(X_i) - \mathbb{E}[\ell_\varepsilon(e + (f_0 - f)(X)) - \ell_\varepsilon(e + (f_0 - f^*)(X))]
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} h(X_i, e_i) - \mathbb{E}[h(X, e)],
\]
where \( h(X, e) : = \ell_\varepsilon(e + (f_0 - f)(X)) - \ell_\varepsilon(e + (f_0 - f^*)(X)) \). By the fundamental theorem of calculus,
\[
h(X, e) = \psi_\varepsilon(e + (f_0 - f^*)(X))(f - f^*)(X)
\]
\[
+ \int_{0}^{(f^* - f)(X)} 1[|e + (f_0 - f)(X) + t| \leq \tau][(f^* - f)(X) - t]dt,
\]
thus implying \( |h(X, e)| \leq 2M \cdot \tau + \frac{1}{2}(2M)^2 \leq 2M(M + \tau) \). Let \( Z_i = (X_i, e_i) \) and \( \tilde{h}(Z_i) = h(X_i, e_i) - \mathbb{E}[h(X, e)] \), and note that
\[
\sup_{f \in \mathcal{F} \cap \Theta, (2/\delta)} \|\tilde{h}\|_{\infty} \leq 4M(\tau + M) =: U.
\]
Moreover,
\[
\mathbb{E}[\tilde{h}^2(X, e)] \leq 2\mathbb{E}\left[|\psi_\varepsilon^2(e + (f_0 - f^*)(X))(f^* - f)(X)| + ||(f^* - f)(X)||^2/2\right]
\]
\[
\leq 2\mathbb{E}\left[|\psi_\varepsilon^2(e + (f_0 - f^*)(X))|X|(f^* - f)(X)^2 + M^2(f^* - f)^2(X)^2\right].
\]
We claim that
\[
\mathbb{E}[\psi_\varepsilon^2(e + (f_0 - f^*)(X))|X = x) \leq C_1, \quad (A.8)
\]
for some constant \( C_1 > 0 \). If Condition 2 holds, since \( |\psi_\varepsilon(t)| = \min(|t|, \tau) \leq |t| \), from Condition 2 with \( p \geq 2 \) it follows that
\[
\mathbb{E}[\psi_\varepsilon^2(e + (f_0 - f^*)(X))|X = x) \leq \mathbb{E}(e^2)X = x) + (f_0 - f^*)(X)^2 \leq v_2 + 4M^2.
\]
If \( \tau \leq C_1' \) for some constant \( C_1' > 0 \), because \( |\psi_\varepsilon(t)| \leq \tau \), then \( \mathbb{E}[\psi_\varepsilon^2(e + (f_0 - f^*)(X))|X = x) \leq (C_1')^2 \). Therefore, the claim \( (A.8) \) holds with \( C_1 = (v_2 + 4M^2) \vee (C_1')^2 \). Together the previous two displays imply
\[
\mathbb{E}[\tilde{h}^2(X, e)] \leq (2C_1 + 2M^2)\mathbb{E}(f^* - f)(X)^2 = C'||f - f^*||_2^2
\]
and hence
\[
\sup_{f \in \mathcal{F} \cap \Theta, (2/\delta)} \mathbb{E}[\tilde{h}^2(X, e)] \leq \sup_{f \in \mathcal{F} \cap \Theta, (2/\delta)} \mathbb{E}[\tilde{h}^2(X, e)] \leq \sup_{f \in \mathcal{F} \cap \Theta, (2/\delta)} C'||f - f_0||_2^2 \leq C'2^2\delta^2 =: \sigma^2.
\]
For notational convenience, we denote \( \mathbb{E}[\sup_{f \in \mathcal{F} \cap \Theta, (2/\delta)} |\Delta_n(f)|] \) as \( E_{\delta, j} \). On one hand, condition (3) of the lemma ensures
\[
E_{\delta, j} \leq \phi_n(2^j/\delta) \leq 2^j\phi_n(\delta).
\]

On the other hand, it follows from Talagrand’s inequality that
\[
\mathbb{P}\left\{ \sup_{f \in \mathcal{F}_n \cap \Theta_{\delta,j}(2/\delta)} |\Delta_n(f)| \geq E_{\delta,j} + \frac{\sqrt{2(\sigma^2 + U E_{\delta,j})}}{3} x \right\} \leq 2 e^{-n x}, \quad \text{valid for any } x \geq 0.
\]
By solving the equation
\[
E_{\delta,j} + \frac{\sqrt{2(\sigma^2 + U E_{\delta,j})}}{3} x = c_{26} 2^{2j-4} \delta^2,
\]
we choose \( x \geq 0 \) as
\[
\sqrt{x} = \frac{\sqrt{\frac{\sqrt{2(\sigma^2 + U E_{\delta,j})}}{3} 2^{2j} c_{26}^2 D^2}}{\frac{2}{3} U}
\]
provided \( c_{26} 2^{2j-4} \delta^2 - E_{\delta,j} \geq 0 \). To see this, using conditions (3) and (4) of the lemma and the fact that \( \delta = D \delta_n \) with \( D \geq \frac{c_2}{c_{26}^2} \), we have
\[
c_{26} 2^{2j-4} \delta^2 - E_{\delta,j} \geq c_{26} 2^{2j-4} D^2 \delta_n^2 - 2^j D \phi(\delta_n)
\geq 2^j \delta_n^2 (c_{26} 2^{2j-4} D^2 - D/c_{28})
\geq (c_{26} 2^{j-5}) 2^j D^2 \delta_n^2 \geq 0.
\]
From (A.9) and the inequality \( \sqrt{a} - \sqrt{b} \geq \sqrt{a - b} \) for \( a \geq b \geq 0 \), we see that
\[
\sqrt{x} \geq \frac{\sqrt{\frac{\sqrt{2(\sigma^2 + U E_{\delta,j})}}{3} 2^{2j} c_{26}^2 D^2}}{\frac{2}{3} U} =: \sqrt{x'},
\]
which further implies
\[
\mathbb{P}\left\{ \sup_{f \in \mathcal{F}_n \cap \Theta_{\delta,j}(2/\delta)} |\Delta_n(f)| \geq c_{26} 2^{2j-4} \delta^2 \right\} 
\leq 2 e^{-n x'} \leq 2 \exp\left(-2^{j-1} \frac{c_{26} \delta_n^2 D^2}{54} \right) = 2 \exp\left(-2^{j-1} \frac{c_{30} \delta_n^2 D^2}{54} \right).
\]
Finally, putting together the pieces we conclude that
\[
\mathbb{P}\left\{ \exists \hat{f}_n \in S(D\delta_n) \text{ s.t. } \|\hat{f}_n - f^*\|_2 > D \delta_n \right\} = \sum_{j=1}^{\infty} \mathbb{P}\left\{ \exists \hat{f}_n \in S(D\delta_n) \text{ s.t. } \hat{f}_n \in S_{n,j} \right\}
\leq 2 \sum_{j=1}^{\infty} \exp\left(-c_{30} 2^{j-1} \frac{n \delta_n^2 D^2}{2M(M + \tau)} \right)
\leq 2 \sum_{j=1}^{\infty} \exp\left(-c_{30} \frac{n \delta_n^2 D^2}{M(M + \tau)} \right)
\leq 2 \left(1 - \exp\left(-c_{30} \frac{n \delta_n^2 D^2}{M(M + \tau)} \right)\right)^{-1} \exp\left(-c_{30} \frac{n \delta_n^2 D^2}{M(M + \tau)} \right).
\]
Recall that \( \tau \geq 1 \) and \( D \geq 2 \), and by condition (5), \( n\delta_n^2/\tau \geq c_{29} \). Consequently, we have

\[
\Delta \leq 2 \left\{ 1 - \exp \left( -c_{29}c_{30} \frac{4}{M + M^2} \right) \right\}^{-1} = c_{31},
\]

which completes the proof. \( \Box \)

Lemma A.2 provides a general, high-level result for establishing the convergence rate of the nonparametric Huber estimator \( \hat{f}_n \) to some function \( f^* \). Now we focus on the case where \( f^* = f_0 \). Recall from Proposition 3.1 that a lower bound for the excess risk holds outside some local region, from which (A.3) follows. The main difficulty is then to validate condition (3), that is, inequality (A.4). To this end, we define

\[
g(x, \epsilon) = \ell_a(\epsilon + f_0(x) - f(x)) - \ell_a(\epsilon), \quad \text{(A.10)}
\]

and let \( G_n = \{ g = g_f : [0, 1]^d \times \mathbb{R} \to \mathbb{R} | f \in F_n \} \). Moreover, let \( G_n(\delta) = \{ g = g_f : [0, 1]^d \times \mathbb{R} \to \mathbb{R} | f \in F_n \cap \Theta_0(\delta) \} \). The next two lemmas characterize the properties of function \( g \) defined in (A.10) and the envelop \( G \) of \( G_n \), satisfying

\[
\sup_{g \in G_n} |g(x, \epsilon)| \leq G(x, \epsilon) \quad \text{for all } x \in [0, 1]^d, \epsilon \in \mathbb{R}. \quad \text{(A.11)}
\]

Lemma A.3. Let \( g : [0, 1]^d \times \mathbb{R} \to \mathbb{R} \) be as in (A.10).

1. \( \mathbb{E}[g(X, \epsilon)] \leq \|f - f_0\|_2^2 \) for all \( f \in \Theta_0(2v_\tau \tau^{1-p}) \). If, in addition, that Condition 3 holds, then \( \mathbb{E}[g(X, \epsilon)] \leq \|f - f_0\|_2^2 \) for all \( f \in \Theta \).

2. We have \( \mathbb{E}[g^2(X, \epsilon)] \leq 2(M^2 + \tau^2)\|f - f_0\|_2^2 \) for all \( f \in \Theta \). If we further assume Condition 2 holds for some \( p \geq 2 \). Then \( \mathbb{E}[g^2(X, \epsilon)] \leq c_{32}\|f - f_0\|_2^2 \) for all \( f \in \Theta \), where \( c_{32} = 2(M^2 + v_2) \).

Proof of Lemma A.3. Part (1). By the second-order Taylor series expansion, we have

\[
\ell_a(v + w) - \ell_a(v) = \psi_a(v)w + \int_0^w \psi_a(v + t)(w - t)dt,
\]

from which it follows that

\[
\mathbb{E}[g(X, \epsilon)] = \mathbb{E}\left[ \psi_a(\epsilon)\Delta_f(X) + \int_0^{\Delta_f(X)} 1(|\epsilon + t| \leq \tau)\Delta_f(X) - t)dt \right]
\]

\[
\leq \mathbb{E}\left[ |\psi_a(\epsilon)\Delta_f(X)| + \frac{1}{2}\Delta_f^2(X) \right]
\]

\[
\leq v_\tau \tau^{1-p}\|f - f_0\|_2 + \frac{1}{2}\|f - f_0\|_2^2.
\]

The constraint \( f \in \Theta_0(2v_\tau \tau^{1-p}) \) implies \( v_\tau \tau^{1-p}\|f - f_0\|_2 \leq \frac{1}{2}\|f - f_0\|_2^2 \), and hence \( \mathbb{E}[g(X, \epsilon)] \leq \|f - f_0\|_2^2 \), as claimed.

Under Condition 3, it has been shown in Proposition 3.1 that \( \mathbb{E}[\psi_a(\epsilon)|X = x] = 0 \), \( \mathcal{P}_X \)-a.e. By the tower rule, now we have

\[
\mathbb{E}[g(X, \epsilon)] = \mathbb{E}\left[ \int_0^{\Delta_f(X)} 1(|\epsilon + t| \leq \tau)\Delta_f(X) - t)dt \right] \leq \frac{1}{2}\|f - f_0\|_2^2 \quad \text{for any } f.
\]

30
Part (2). Similarly,
\[
    \mathbb{E}(g^2(X, \epsilon)) = \mathbb{E}\left[\psi_r(\epsilon)\psi(X) + \int_0^{\Delta_f(X)} 1(|\epsilon + t| \leq \tau)|\Delta_f(X) - t|\,dt\right]^2
\]
\[
    \leq \mathbb{E}\left(2\psi_r^2(\epsilon)\psi(X) + \frac{1}{2}\Delta^2_f(X)\right)
\]
\[
    \leq 2\mathbb{E}\left(M^2 + \mathbb{E}(\psi_r^2(\epsilon)|X)\right)\cdot \Delta^2_f(X),
\]
where the last inequality uses the bound \(|f - f_0|_\infty \leq 2M\). The first claim follow from \(\mathbb{E}(\psi^2_r(\epsilon)|X = x) \leq \tau^2\). For the second claim, note that \(|\psi_r(\epsilon)| = \min(|\epsilon|, \tau) \leq |\epsilon|\). Under Condition 2 with \(p \geq 2\), it holds for any \(x\) that \(\mathbb{E}(\psi_r^2(\epsilon)|X = x) \leq \mathbb{E}[|\epsilon^2|X = x] \leq v_2\). Combining the pieces yields \(\mathbb{E}[g^2(X, \epsilon)] \leq 2(M^2 + v_2)||f - f_0||_2^2\), as expected. □

Lemma A.4. Let \(G(x, \epsilon) = 2M^2 + 2M|\psi_r(\epsilon)|\) be an envelop function for the class \(\mathcal{G}_n\). Then
\[
    \mathbb{E}(G^2(X, \epsilon)) \leq c_{14} := 8M^2(M^2 + v_2) \quad \text{and} \quad ||G||_\infty \leq 2M(M + \tau).
\]
Proof of Lemma A.4. By a second-order Taylor series expansion of \(\ell_r(\cdot)\) and the triangle inequality,
\[
    |g(x, \epsilon)| \leq |\psi_r(\epsilon) \cdot \Delta_f(x)| + \frac{1}{2}\Delta^2_f(x) \leq 2M|\psi_r(\epsilon)| \leq \tau
\]
holds for any \(f \in \mathcal{F}_n\). Therefore \(G : [0, 1]^d \times \mathbb{R} \to [0, \infty)\) is an envelop of \(\mathcal{G}_n\).

Recall from the proof of Lemma A.3 that \(\mathbb{E}(\psi_r^2(\epsilon)|X| \leq v_2\). This together with the Cauchy-Schwarz inequality implies \(\mathbb{E}(G^2(X, \epsilon)) \leq 8M^2(M^2 + v_2)\). The second bound follows from the fact that \(|\psi_r(\epsilon)| \leq \tau\) almost surely. □

To apply Lemma A.2, the key is an upper bound for
\[
    \mathbb{E}\left(\sup_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^n g(X_i, \epsilon_i) - \mathbb{E}g(X, \epsilon)\right)
\]
for a suitably chosen neural network \(\mathcal{F}_n\). Specifically, let \(\mathcal{F}_n = \mathcal{F}_n(d, L, \bar{N}, M)\) be the function class realized by ReLU neural network with depth \(L = c_3 L \log(L)\) and width \(\bar{N} = c_4 N \log(N)\), respectively. The next result characterizes the complexity of the function class \(\mathcal{F}_n\) via an upper bound on the uniform covering number, defined as follows.

Definition A.1 (Uniform covering number). Let \(m \in \mathbb{N}^+\), and \(\mathcal{F} = \{f : X \to \mathbb{R}\}\) be a function class. We define the uniform covering number under \(L_\infty\)-norm for the function class \(\mathcal{F}\) as
\[
    \mathcal{N}_\infty(\epsilon, \mathcal{F}, m) = \sup_{X = (x_1, \ldots, x_m) \in X^m} \mathcal{N}(\epsilon, \mathcal{F}|X, \|\cdot\|_\infty),
\]
where \(\mathcal{F}|X := \{(f(x_1), \ldots, f(x_m)) : f \in \mathcal{F}\}\).

Lemma A.5 (Uniform covering number bound for \(\mathcal{F}_n\)). For any \(\epsilon > 0\),
\[
    \log \mathcal{N}_\infty(\epsilon, \mathcal{F}_n, m) \leq \log(emM/\epsilon)\log(L \cdot \bar{N})(L \cdot \bar{N})^2.
\]

31
Proof of Lemma A.5. By Theorem 12.2 of Anthony & Bartlett (1999), we have that for any \( \epsilon > 0 \),

\[
\log \mathcal{N}_\infty(\epsilon, \mathcal{T}_n, m) \leq \text{Pdim}(\mathcal{T}_n) \cdot \log(\epsilon M/\epsilon),
\]

where \( \text{Pdim}(\mathcal{T}_n) \) denotes the pseudo-dimension of \( \mathcal{T}_n \). Applying further Theorem 7 of Bartlett et al. (2019) yields the bound \( \text{Pdim}(\mathcal{T}_n) \leq W \hat{L} \log(W) \), where \( W \) is the number of parameters of the network \( \mathcal{T}_n \), satisfying \( W = O(\hat{L} \cdot \hat{N}^2) \).

\[
\square
\]

Lemma A.6 (A maximal inequality (Chernozhukov et al., 2014)). Consider the function class \( \mathcal{G} \) of measurable functions \( [0, 1]^d \times \mathbb{R} \to \mathbb{R} \), to which a measure envelop \( G \) is attached. Suppose that \( \|G\|_2 < \infty \), and let \( \sigma \) be any positive constant such that \( \sup_{g \in \mathcal{G}} \mathbb{E} g^2(X, \epsilon) \leq \sigma^2 \leq \|G\|_2^2 \). Moreover, define \( r = \sigma/\|G\|_2 \) and \( \mathcal{G} = \max_{1 \leq j \leq n} G(X_i, \epsilon_j) \). Then

\[
\mathbb{E} \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i, \epsilon_i) - \mathbb{E} g(X, \epsilon) \right| \right\} \leq \|G\|_2 \cdot J(r, \mathcal{G}, G) + \frac{\|G\|_2 \cdot J^2(r, \mathcal{G}, G)}{r^2 \sqrt{n}},
\]

where

\[
J(r, \mathcal{G}, G) = \int_{0}^{r} \sup_{Q_c} \sqrt{1 + \log \mathcal{N}(r, \mathcal{G}, \| \cdot \|_{2, Q_c}^2)} \, dr
\]

is the uniform entropy integral, and the supremum is taken over all \( n \)-discrete probability measures on \( [0, 1]^d \times \mathbb{R} \).

With the above preparations, we are now ready to bound the expected value of the supremum (A.12).

Lemma A.7. Suppose either Condition 2 holds or \( \tau \leq 1 \). Let \( \mathcal{T}_n = \mathcal{T}_n(d, \hat{L}, \hat{N}, M) \) be the deep ReLU neural network class whose depth \( \hat{L} \leq L \log L \) and width \( \hat{N} \leq N \log N \) satisfying \( LN \leq n^{1/2-\alpha} \) for some positive constant \( \alpha \) independent of \( n \). Then

\[
\mathbb{E} \left\{ \sup_{g \in \mathcal{G}_{n, \delta}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, \epsilon_i) - \mathbb{E} g(X, \epsilon) \right| \right\} \leq c_{33} \left\{ \frac{\delta}{n^\alpha} (\log n)^{5/2} \sqrt{\log(n\tau)} + \frac{\tau}{n^{2\alpha}} (\log n)^3 \log(n\tau) \right\}
\]

for all \( \delta \geq 1/n \) and \( \tau \geq 1 \), where \( c_{33} > 0 \) is constant independent of \( \delta \) and \( n \).

Proof of Lemma A.7. Recall our setting of neural network architecture, that is,

\[
\hat{L} \leq L \log L \quad \text{and} \quad \hat{N} \leq N \log N
\]

subject to \( NL \leq n^{1/2-\alpha} \). This implies

\[
(\hat{N} \cdot \hat{L})^2 \leq (NL)^2 (\log N \cdot \log L)^2 \leq n^{1-2\alpha} (\log n)^4 \quad \text{and} \quad \log(\hat{N} \cdot \hat{L}) \leq \log n.
\]

By Lemma A.5, we have for any \( \epsilon \in (0, M) \) that

\[
\log \mathcal{N}_\infty(\epsilon, \mathcal{T}_n, n) \leq \log(Mn/\epsilon) \cdot (\hat{N} \cdot \hat{L})^2 \log(\hat{N} \cdot \hat{L}) \leq \log(Mn/\epsilon) \cdot n^{1-2\alpha} (\log n)^5.
\]

Since \( \ell(\cdot) \) is a \( \tau \)-Lipschitz function, any \( (\epsilon/\tau) \)-net of \( \mathcal{T}_{n, L} \) under the \( \| \cdot \|_{\infty} \)-norm is also an \( \epsilon \)-net of \( \mathcal{G}_{n, L} \) under the \( \| \cdot \|_{\infty} \)-norm. Therefore, for any \( \epsilon \in (0, nM\tau) \)

\[
\log \mathcal{N}_{\infty}(\epsilon, \mathcal{G}_n, n) \leq \log(2Mn\tau/\epsilon) \cdot n^{1-2\alpha} (\log n)^5.
\]
Consequently,

\[
\sup_{Q_n} \sqrt{1 + \log N(\epsilon, \mathcal{G}_n, L_2(Q_n))} \leq \sqrt{1 + \log N_m(\epsilon, \mathcal{G}_n, n)} \leq [\log(2Mn\tau/\epsilon) + 1] \cdot n^{1/2-\alpha}(\log n)^{5/2}.
\]

Because either Condition 2 or \( \tau \leq 1 \) holds, by Lemma A.3 and Lemma A.4, we have \( \sigma^2 = C_1\delta^2 \), \( \|G\|_2 = C_2 \). \( \|\overline{G}\|_2 \leq C_3 + C_4\tau \) for constants \( C_1, \ldots, C_4 \) that are independent of \( n, \tau, \overline{L} \) and \( \overline{N} \). Taking \( r = C_5\delta \) for \( C_5 = \sqrt{C_1/C_2} \) yields

\[
J(r, \mathcal{G}_n, G) \leq \int_0^{C_5\delta} \left\{ \sqrt{\log(C_2nM\tau/\epsilon)} + 1 \right\} n^{1/2-\alpha}(\log n)^{5/2} \, d\epsilon
\]

where the last inequality follows from the fact that \( \int_0^{x} \sqrt{\log(1/y)} \, dy \leq x(\sqrt{\log(1/x)} \lor 1) \). We thus conclude that for any \( \delta \geq 1/n \) and \( \tau \geq 1 \),

\[
J(r, \mathcal{G}_n, G) \leq C_6 n^{1/2-\alpha}(\log n)^{5/2} \sqrt{\log(n\tau)} \cdot n^{1/2-\alpha}(\log n)^{5/2} \sqrt{\log(n\tau)}.
\]

for all \( \delta \geq 1/n \) and \( \tau \geq 1 \). This establishes the claim (A.16).

\[\square\]

### A.2 Proof of Proposition 3.1

To begin with, we derive from the fundamental theorem of calculus that for every \( v, w \in \mathbb{R} \),

\[
\ell_r(v + w) - \ell_r(v) = \psi_r(v)w + \int_0^w \psi_r'(v + t)(w - t) \, dt,
\]

where \( \psi_r(\cdot) \) and \( \psi'_r(\cdot) \) are given in (3.4). For any \( \tau > 0 \) and function \( f : [0, 1]^d \rightarrow \mathbb{R} \), it follows that

\[
\mathcal{R}_r(f) - \mathcal{R}_r(f_0) = \mathbb{E}[\ell_r(\epsilon + \Delta_f(X))] - \mathbb{E}[\ell_r(\epsilon)]
\]

\[
= \mathbb{E}[\psi_r(\epsilon)\Delta_f(X)] + \mathbb{E} \left[ \int_0^{\Delta_f(X)} 1(|\epsilon + t| \leq \tau)|\Delta_f(X) - t| \, dt \right].
\]

(A.18)
In the following, we bound the two terms on the right-hand side of (A.18) separately.

For the former, note that \( \psi_r(e) = 1 \{ |e| \leq \tau r \} e + 1 \{ e > \tau r \} e + 1 \{ e < -\tau r \} e \). Recall the assumption that \( \mathbb{E}(\varepsilon | X = x) = 0 \), we have \( \mathbb{E}(1 \{ |e| \leq \tau |e| X | X = x \} = -\mathbb{E}(1 \{ |e| > \tau |e| X | X = x \} \). Taking the conditional expectation of \( \psi_r(e) \) given \( X = x \) yields

\[
\mathbb{E}(\psi_r(e) | X = x) = \mathbb{E}(-1 \{ |e| > \tau |e| X | X = x \} - 1 \{ e > \tau |e| X | X = x \} + 1 \{ e < -\tau |e| X | X = x \})
\leq \mathbb{E}((|e| - \tau |e|) | X = x)
\leq \mathbb{E}(|e|) \leq v_p \tau^{1-p}.
\]  
\( \text{(A.19)} \)

Combining this with the Cauchy–Schwarz inequality, we conclude that

\[
\mathbb{E}(\psi_r(e) \Delta_f(X)) \geq -v_p \tau^{1-p} \mathbb{E}(\Delta_f(X)) \geq -v_p \tau^{1-p} \| f - f_0 \|_2.
\]  
\( \text{(A.20)} \)

Turning to the second term on the right-hand side of (A.18), for every \( f \in \Theta \) we have

\[
\mathbb{E} \left[ \int_0^{\Delta_f(X)} 1\{ |e + t| \leq \tau \} |\Delta_f(x) - t| d|X = x \right]
= \mathbb{E} \left[ \int_0^{\Delta_f(X)} [1 - 1\{ |e + t| > \tau \}] |\Delta_f(x) - t| d|X = x \right]
\geq \frac{1}{2} |\Delta_f(x)|^2 - \mathbb{E} \left[ \int_0^{\Delta_f(X)} \left[ 1\{ |e| > \tau/2 \} + 1\{ |\Delta_f(x)| > \tau/2 \} \right] |\Delta_f(x) - t| d|X = x \right]
= \frac{1}{2} |\Delta_f(x)|^2 \mathbb{P}(|e| > \tau/2 | X = x),
\]
where the last step follows from the fact that \( |\Delta_f(x)| = |f(x) - f_0(x)| \leq 2M \leq \tau/2 \) provided \( \tau \geq 4M \).

By Markov’s inequality and Condition 2,

\[
\mathbb{P}(|e| > \tau/2 | X = x) \leq \frac{\mathbb{E}(|e|^p | X = x)}{(\tau/2)^p} \leq \frac{1}{2} \text{ for all } x
\]
as long as \( \tau \geq 2(2v_p)^{1/p} \). Taking the expectation with respect to \( X \sim \mathcal{P}_X \) gives

\[
\mathbb{E} \left[ \int_0^{\Delta_f(X)} 1\{ |e + t| \leq \tau \} |\Delta_f(x) - t| d|X = x \right] \geq \frac{1}{4} \mathbb{E} |\Delta_f(X)| = \frac{1}{4} \| f - f_0 \|_2^2.
\]  
\( \text{(A.21)} \)

Together, (A.18), (A.20) and (A.21) imply that as long as \( \tau \geq \max(2M, 2v_p)^{1/p} \),

\[
\mathcal{R}_\tau(f) - \mathcal{R}_\tau(f_0) \geq -v_p \tau^{1-p} \| f - f_0 \|_2 + \frac{1}{4} \| f - f_0 \|_2^2
\]
holds for all \( f \in \Theta \). In particular, for \( f \in \Theta \setminus \Theta_0(8v_p \tau^{1-p}) \) we have \( v_p \tau^{1-p} \| f - f_0 \|_2 \leq \| f - f_0 \|_2^2/8 \), and hence

\[
\mathcal{R}_\tau(f) - \mathcal{R}_\tau(f_0) \geq \frac{1}{8} \| f - f_0 \|_2^2.
\]

If the distribution of \( \varepsilon | X = x \) is symmetric, then \( \psi_r(\cdot) \) is also symmetric (around zero), we have

\[
\mathbb{E}(\psi_r(e) | X = x) = \left\{ \int_{-\infty}^{0} + \int_{0}^{\infty} \right\} \psi_r(t) dF_{\varepsilon|X=x}(t)
= \int_{0}^{\infty} \psi_r(t) dF_{\varepsilon|X=x}(-t) + \int_{0}^{\infty} \psi_r(t) dF_{\varepsilon|X=x}(t)
= \int_{0}^{\infty} \psi_r(t) d\{ 1 - F_{\varepsilon|X=x}(t) \} + \int_{0}^{\infty} \psi_r(t) dF_{\varepsilon|X=x}(t) = 0,
\]

34
implying that $\mathbb{E}[\psi_r(\varepsilon)\Delta_f(X)] = 0$. Consequently,
\[
\mathcal{R}_r(f) - \mathcal{R}_r(f_0) = \mathbb{E}\left[ \int_0^{\Delta_f(X)} 1(|\varepsilon + t| \leq \tau)\Delta_f(X) - t|dt \right] \geq \frac{1}{4}\|f - f_0\|_2^2
\]
for all $f \in \Theta$ provided that $\tau \geq 2 \max\{2M, (2\nu_p)^{1/p}\}$. This completes the proof of Proposition 3.1. \hfill \Box

### A.3 Proof of Proposition 3.2

Recall from the proof of Proposition 3.1 that as long as $\tau \geq 2 \max\{2M, (2\nu_p)^{1/p}\}$,
\[
\mathcal{R}_r(f) - \mathcal{R}_r(f_0) \geq -v_p r^{1-p}\|f - f_0\|_2 + \frac{1}{4}\|f - f_0\|_2^2
\]
for all $f \in \Theta$. Taking $f = f_{0, \tau}$, the claimed result follows immediately from the fact that $\mathcal{R}_r(f_{0, \tau}) - \mathcal{R}_r(f_0) \leq 0$.

Next, assume that the noise variable $\varepsilon$ is sub-Gaussian satisfying (3.10). Similarly to (A.19), now we have for any $x \in [0, 1]^d$ that $\mathbb{P}(|\varepsilon| > \tau/2|X = x) \leq 2e^{-\tau^2/(8\sigma^2)}$ and
\[
\left| \mathbb{E}[\psi_r(\varepsilon)X = x] \right| \leq \mathbb{E}[|\varepsilon|1(|\varepsilon| > \tau)|X = x] \\
= \sigma \mathbb{E}\int_0^{\infty} 1(|\varepsilon|/\sigma > \tau/\sigma)1(|\varepsilon|/\sigma > \tau)dt \\
= \tau \mathbb{P}(|\varepsilon|/\sigma > \tau/\sigma) + \sigma \int_{\tau/\sigma}^{\infty} \mathbb{P}(|\varepsilon|/\sigma > t)dt \\
\leq 2\tau e^{-\tau^2/8\sigma^2} + 2\sigma \int_{\tau/\sigma}^{\infty} e^{-t^2/2\sigma^2}dt \leq 2(\tau + \alpha^2/\tau)e^{-\alpha^2/2\sigma^2}.
\]

Keeping the remaining arguments the same, we obtain that as long as $\tau \geq 4 \max\{M, \sigma\}$,
\[
\mathcal{R}_r(f) - \mathcal{R}_r(f_0) \geq -2(\tau + \alpha^2/\tau)e^{-\alpha^2/2\sigma^2}\|f - f_0\|_2 + (1 - 2e^{-2})\|f - f_0\|_2^2
\]
for all $f \in \Theta$. This proves (3.11) by taking $f = f_{0, \tau}$. \hfill \Box

### A.4 Proof of Theorem 3.4

The proof is based on the high-level result stated in Lemma A.2. To this end, we first choose appropriate $\tau$, $\delta_n$ and $f_n$ such that conditions (1)–(5) in Lemma A.2 are satisfied. Recall that $\gamma^\ast = \beta^\ast/d^\ast$ and $\nu^\ast = 1 - \frac{1}{2^{p-1}}$. Moreover, set
\[
\alpha = \frac{\gamma^\ast}{2\gamma^\ast + \nu^\ast} = \frac{\beta^\ast}{2\beta^\ast + d^\ast(1 - \frac{1}{2^{p-1}})} = \frac{1}{2} - \frac{\nu^\ast}{2(2\gamma^\ast + \nu^\ast)},
\]
and let $L, N \geq 1$ be such that $LN \asymp n^{1/2-\alpha}$. Next, we specify the robustification parameter $\tau$ and convergence rate $\delta_n$ as
\[
\tau = (1 \lor v_p^{1/p} \lor 4M)|n^\nu(\log n)^{-3}2^{(1-\nu^\ast)} \text{ and } \delta_n = 2C_1\tau^{1-p}c_5(nL)^{-2}\nu^\ast,
\]
where $C_1 = 8v_p$ and $c_5$ is as in Proposition 3.3. Noting that $n^{-(1-2\alpha)\nu^\ast} = n^{-\frac{\nu^\ast}{\gamma^\ast + \nu^\ast}} = n^{-\nu^\ast\alpha}$ and $2(1 - \nu^\ast)(p - 1) = \nu^\ast$, we have $\delta_n \asymp (n^{-\alpha}(\log n)^{3})^{\nu^\ast}$.
We start with verifying condition (1), i.e., there exists some \( f_n \in \mathcal{F}_n \) such that (A.2) holds. Since \( f_0 \in \mathcal{H}(l, \mathcal{P}) \), we have \( f_0 + c \in \mathcal{H}(l, \mathcal{P}) \) for any constant \( c \in \mathbb{R} \). By Proposition 3.3, we can find a neural network \( \mathcal{F}_n \) with depth \( \tilde{L} = c_3 L \log(L) \) and width \( \tilde{N} = c_4 N \log(N) \) such that there exists some \( f_n \in \mathcal{F}_n \) satisfying

\[
\|f_n - (f_0 - 2\delta_n)\|_\infty \leq c_5 (NL)^{-2}\gamma' \leq \delta_n.
\]

This further implies \( |f_n(x) - f_0(x)| \leq 3\delta_n \) and \( |f_n(x) - f_0(x)| \geq \delta_n \) for all \( x \in [0, 1]^d \). Consequently,

\[
2\nu_p \tau^{1-p} \leq 2C_1 \tau^{1-p} \leq \delta_n \leq \|f_n - f_0\|_2 \leq 3\delta_n.
\]

Next, recall our definition of \( \delta \) in (A.10). By Lemma A.3 and since \( \|f_n - f_0\|_2 \geq 2\nu_p \tau^{1-p} \), we have

\[
\mathcal{R}_\tau(f_n) - \mathcal{R}_\tau(f_0) = \mathbb{E}[g(X, \epsilon)] \leq \|f_n - f_0\|_2^2 \leq 9\delta_n^2
\]
as desired. For condition (2), applying Proposition 3.1 we see that

\[
\mathcal{R}_\tau(f) - \mathcal{R}_\tau(f_0) \geq \frac{1}{8}\|f - f_0\|_2^2 \text{ for any } f \in \Theta_0(\delta_n/2).
\]

More generally, for \( f \in \Theta_0(\delta/2) \) with \( \delta \geq \delta_n \), we have

\[
\mathcal{R}_\tau(f) - \mathcal{R}_\tau(f_0) \geq \frac{1}{32} \delta^2,
\]

which certifies condition (2) of Lemma A.2 with \( c_{26} = 1/32 \).

Turning to condition (3), it follows from Lemma A.7 and the fact log \( \tau \leq \log n \) that for all \( \delta \geq \delta_n \geq 1/n \),

\[
\mathbb{E}\left\{ \sup_{f \in \mathcal{F}_n} |\Delta_n(f)| \right\} \leq C'\left\{ \delta n^{-a}(\log n)^3 + \tau n^{-2a}(\log n)^6 \right\} = C_2 \phi_n(\delta).
\]

(A.22)

It is easy to see that \( \phi_n(C\delta) \leq C\phi_n(\delta) \) for all \( C \geq 1 \) because \( \phi(\cdot) \) is linear function and satisfies \( \phi(0) > 0 \).

For condition (4), recall that \( \delta_n \leq (n^{-a}(\log n)^3)^{1/\gamma'} \). Thus,

\[
\phi_n(\delta_n) = \delta_n n^{-a}(\log n)^3 + \tau n^{-2a}(\log n)^6
\]

\[
\leq \{n^{-a}(\log n)^3\}^{1+\gamma'} + \{n^{-a}(\log n)^3\}^{2-2(1-\gamma')}
\]

\[
\leq \{n^{-a}(\log n)^3\}^{2\gamma'}
\]

\[
\leq \delta_n^2,
\]

implying \( \delta_n^2 \geq c_{28} \phi_n(\delta_n) \) for some constant \( c_{28} > 0 \).

Finally, by our choice of \( \delta_n \) and \( \tau \), it holds

\[
\frac{n\delta_n^2}{\tau} = n\{n^{-a}(\log n)^3\}^{2\gamma'} = n^{1-2a}(\log n)^6 = n^{\gamma' \nu}(\log n)^6 \geq 1.
\]

This verifies condition (5). Noting that

\[
\frac{n\delta_n^2}{M(\tau + M)} \geq n^{\gamma' \nu}(\log n)^6 \text{ and } c_{26}/8 = 1/256,
\]

36
Lemma A.2 implies the bound
\[ \mathbb{P}(\|f_n - f_0\| \geq D\delta_n) \leq \mathbb{P}(\tilde{R}_r(f_n) > \tilde{R}_r(f_0) + 2^{-9}D^2\delta_n^2) + c_{31} \exp\left(-c_{30} \frac{n\delta_n^2 D^2}{M(\tau + M)}\right) \]
\[ \leq \exp\left(-c_8 n^{-1/2 - \alpha^*} (\log n)^6 D^2\right) \]
for all \( D \geq \max\{2, \frac{6\sqrt{2}}{\sqrt{c_26}}, \frac{32}{c_{26}^2}\} = c_7 \), which completes the proof. \( \square \)

### A.5 Proof of Theorem 3.5

Similarly to the proof of Theorem 3.5, again we rely on Lemma A.2 but with a different choice of \((\tau, \delta_n)\). The main difference is that we no longer need a diverging \(\tau\) to ensure a negligible bias as \(n \to \infty\).

Let \(c_1 \leq \tau \times 1\) for \(c_1\) given in Proposition 3.2, and let \(L, N \geq 1\) be such that \(NL \asymp n^{1/2 - \alpha^*}\) with \(\alpha^* = \gamma^*/(2\gamma^* + 1)\). Set
\[ \delta_n = \tau n^{-\alpha^*} (\log n)^3 \sqrt{c_5(NL)^{-2\gamma^*}} \asymp n^{-\alpha^*} (\log n)^3. \]
Recall that \(f_0 \in \mathcal{H}(L, \mathcal{P})\). To verify condition (1) of Lemma A.2, from Proposition 3.3 we see that there exist a neural network \(\mathcal{F}_n\) with depth \(\bar{L} = c_3L\log(L)\) and width \(\bar{N} = c_4N\log(N)\) and some \(f_n \in \mathcal{F}_n\) satisfying
\[ \|f_n - f_0\| \leq c_5(NL)^{-2\gamma^*} \leq \delta_n. \]
Together, Lemma A.3 and Condition 3 imply
\[ \mathcal{R}_n(f_n) - \mathcal{R}_n(f_0) \leq \delta_n^2 \leq 9\delta_n^2. \]

For condition (2), Proposition 3.1 ensures that \(\mathcal{R}_n(f) - \mathcal{R}_n(f_0) \geq \|f - f_0\|^2/4\) for all \(f \in \Theta\). In particular, if \(f \in \Theta^\epsilon(\delta/2)\), this further implies
\[ \mathcal{R}_n(f) - \mathcal{R}_n(f_0) \geq \frac{1}{4}\|f - f_0\|^2 \geq \frac{1}{16}\delta^2, \]
thus verifying condition (2) with \(c_{26} = 2^{-4}\). Turning to condition (3), it follows from Lemma A.7 that, for all \(\delta \geq \delta_n \geq 1/n\) and \(n \geq \tau\),
\[ \mathbb{E}\left(\sup_{f \in \mathcal{F}_n(\delta_0)} |\Delta_n(f)|\right) \leq C_1 \phi_n(\delta), \quad (A.23) \]
where \(\phi_n(\delta) = \delta n^{-\alpha^*} (\log n)^3 + n^{-2\alpha^*} (\log n)^6\) satisfies \(\phi_n(C\delta) \leq C\delta\) for all \(C \geq 1\). For conditions (4) and (5), by the above choice of \(\delta_n\) and \(\phi_n(\delta)\), we have
\[ \phi_n(\delta_n) \leq \delta_n n^{-\alpha^*} (\log n)^3 + n^{-2\alpha^*} (\log n)^6 \leq \delta_n^2 \]
and \(n\delta_n^2/\tau = n^{1-2\alpha^*} (\log n)^6 \geq 1\).

Finally, applying Lemma A.2 with \(c_{26} = 2^{-4}\) yields
\[ \mathbb{P}(\|f_n - f_0\| \geq D\delta_n) \leq \mathbb{P}(\tilde{R}_r(f_n) > \tilde{R}_r(f_0) + 2^{-7}D^2\delta_n^2) + c_{31} \exp\left(-\frac{c_{30}}{c_8} \frac{n\delta_n^2 D^2}{M(\tau + M)}\right) \]
\[ \leq \exp\left(-c_{8} n^{1-2\alpha^*} (\log n)^6 D^2\right) \]
for all \(D \geq c_9 := \max\{2, \frac{6\sqrt{2}}{\sqrt{c_26}}, \frac{32}{c_{26}^2}\}\), which completes the proof. \( \square \)
B Proof of Neural Network Approximation

We include all the proofs of the neural network approximation result in this section. We first introduce some additional notations in Section B.1, which simplify the proofs of neural network approximation. Then, we prove Proposition 3.3 in Section 3.2. Sections 4.5 and 4.6 provide proofs of Theorem 4.5 and Theorem 4.6, respectively.

B.1 Notations about the construction of neural network

In this subsection we introduce several notations and simple facts on the construction of neural network. 

Representation of neural network. In the section, we consider a more general representation of neural network, which has output dimension $o \in \mathbb{N}^+$ rather than 1. In this case, the neural network can be viewed as a function $f : \mathbb{R}^d \rightarrow \mathbb{R}^o$. We write $f \in \mathcal{F}(d, L, N, o)$ if $f$ is a neural network with input dimension $d$, output dimension $o$, depth $\leq L$, and at most $N$ hidden nodes at each layer. Formally, suppose

$$f(x) = \mathcal{L}_{L+1} \circ \sigma \circ \mathcal{L}_L \circ \sigma \circ \mathcal{L}_{L-1} \circ \sigma \circ \cdots \circ \mathcal{L}_2 \circ \sigma \circ \mathcal{L}_1(x)$$

where $\mathcal{L}_i(x) = W_{i}x+b_{i}$ is a linear transformation with $W_i \in \mathbb{R}^{d \times d_{i-1}}$, $b_i \in \mathbb{R}^d$ and $(d_0, d_1, \cdots, d_L, d_{L+1})$. Then $\mathcal{F}(d, L, N, o)$ can be written as

$$\mathcal{F}(d, L, N, o) = \{f \text{ is of the form (B.1) with } d_0 = d, d_{L+1} = o \text{ and } d_i \leq N \text{ for any } i \in \{1, \cdots, L\}\}.$$  

Neural network padding. If $f$ is a neural network with depth between 1 and $L$, and at most $N$ hidden nodes at each layer, then there exists some neural network $g$ with depth $L$ and $N$ hidden nodes at each layer such that $f(x) = g(x)$ for all the input $x$. We refer to this construction as neural network padding. The padding with respect to width is trivial. For the padding with respect to depth, assume that the neural network has $L' \geq 1$ hidden layers. We can apply the identity map together with the activation function $L - L'$ times between the first hidden layer and the layer next to it. This will not change $f(x)$, but will increase the number of layers by $L - L'$. Hence $\mathcal{F}(d, L, N, o)$ can also be seen as the set of all neural networks with input dimension $d$, output dimension $o$, depth $L$ and width $N$. From the above discussion, we also have that $\mathcal{F}(d, L, N, o) \subset \mathcal{F}(d, L', N', o)$ if $L' \geq L$ and $N' \geq N$.

Network composition. Suppose $f \in \mathcal{F}(d_1, L_1, N_1, d_2)$ and $g \in \mathcal{F}(d_2, L_2, N_2, d_3)$, we use $h = g \circ f$ to denote the neural network which uses the input of $g$ as the output of $f$. It should be noted that $h$ is a neural network with width $N_1 \lor N_2 \lor d_2$ and depth $L_1 + L_2$. This is because we can combine the weight connecting the final hidden layer and the output layer of $g$ and the weight connecting the input layer and the first hidden layer of $f$ as a single weight, i.e. $W_2(W_1x + b_1) + b_2 = W_1'x + b_1'$.

Network parallelization. Suppose $f_i \in \mathcal{F}(d_i, N_i, L_i, o_i)$ for $i \in \{1, \cdots, k\}$. We use $h = (f_1, \cdots, f_k)$ to denote the neural network that takes $x \in \mathbb{R}^{\sum_{i=1}^{k} d_i}$ as the input, feeds the entries $x^{(i)} = (x_{\sum_{j=1}^{i-1} d_j+1}, \cdots, x_{\sum_{j=1}^{i} d_j})$ to the $i$-th sub-network $f_i$ that returns $y^{(i)}$, and combines these $y^{(i)}$ as the output. Such an $h$ is a neural network with input dimension $\sum_{i=1}^{k} d_i$, output dimension $\sum_{i=1}^{k} o_i$, depth at most $\max_{1 \leq i \leq d} L_i$.

38
Proof of Lemma B.1. Similarly, we conclude that \( h \) is a neural network with input dimension \( d \), output dimension \( \sum_{i=1}^{k} o_i \), depth at most \( \max_{1 \leq i \leq d} L_i \) and width at most \( \sum_{i=1}^{d} N_i \).

**Simple functions.** At last, we introduce some simple functions that can be parameterized using ReLU neural networks:

**Lemma B.1** (Identity, Absolute value, Min, Max). For any \( x, y \in \mathbb{R} \), the following properties hold:

1. \( x \in \mathcal{F}(1, 1, 2, 1) \);
2. \( |x| \in \mathcal{F}(1, 1, 2, 1) \);
3. \( \min(x, y) \in \mathcal{F}(2, 1, 4, 1) \);
4. \( \max(x, y) \in \mathcal{F}(2, 1, 4, 1) \).

**Proof of Lemma B.1.** For claims (1) and (2), recall that \( \sigma(x) = (x)_+ \), we thus have \( x = \sigma(x) - \sigma(-x) \), \( |x| = \sigma(x) + \sigma(-x) \). For claims (3) and (4), note that \( \min(x, y) = \frac{1}{2}(x + y - |x - y|) \) and \( \max(x, y) = \frac{1}{2}(x + y + |x - y|) \). It follows that

\[
\min(x, y) = \frac{1}{2}(x + y - |x - y|)
\]

\[
= \frac{1}{2} \left( \sigma(x + y) - \sigma(-x - y) - \sigma(x - y) - \sigma(y - x) \right),
\]

hence proving claim (3). Claim (4) can be similarly proved.

**Lemma B.2** (\( \| \cdot \|_{\infty} \) norm in \( \mathbb{R}^d \)). For any \( d \geq 1 \), the map \( \|x\|_{\infty} : \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}(d, \lfloor \log_2 d \rfloor + 1, 2d, 1) \).

**Proof of Lemma B.2.** The univariate case \( d = 1 \) is trivial because \( \|x\|_{\infty} = |x| \) so that the result follows directly from Lemma B.1.

Assume \( d \geq 2 \) and let \( x = (x_1, \ldots, x_d) \). For each \( x_i \), by Lemma B.1 we can construct a neural network \( \phi_i(x) \in \mathcal{F}(1, 1, 2, 1) \) such that \( \phi_i(x_i) = |x_i| \). Via a parallelization argument, we can then construct \( f = (\phi_1, \ldots, \phi_d) \) such that \( f(x) = (|x_1|, \ldots, |x_d|) \) and \( f \in \mathcal{F}(d, 1, 2d, d) \). Since \( \|x\|_{\infty} = \max_{1 \leq i \leq d} |x_i| \), it suffices to build some \( h^{(d)}(x) \) such that \( h^{(d)}(x_1, \ldots, x_d) = \max_{1 \leq i \leq d} x_i \). We construct such an \( h^{(d)} \) by induction. We claim that \( h^{(d)} \in \mathcal{F}(d, \lfloor \log_2 d \rfloor, 2d, 1) \).

Starting from \( d = 2 \), Lemma B.1 ensures that there exists some \( h^{(2)}(x_1, x_2) \in \mathcal{F}(2, 1, 4, 1) \) such that \( h^{(2)}(x_1, x_2) = \max\{x_1, x_2\} \). Next we consider the general case of \( d > 2 \). If \( d = 2k \) for some \( k \geq 1 \), by Lemma B.1 we can construct \( \psi_i(x_{2i-1}, x_{2i}) \in \mathcal{F}(2, 1, 4, 1) \) such that \( \psi_i(x_{2i-1}, x_{2i}) = x_{2i-1} \lor x_{2i} \) for each \( 1 \leq i \leq k \). Again, using the parallelization argument we can construct \( g = (\psi_1, \ldots, \psi_k) \in \mathcal{F}(d, 1, 2d, d) \) such that \( g(x_1, \ldots, x_d) = (x_1 \lor \ldots \lor x_{2d-1} \lor x_{2d}) \). This means that \( h^{(d)} \) can be constructed as a composition of \( g \) and \( h^{(d/2)} \). Since \( h^{(d/2)} \in \mathcal{F}(d/2, \lfloor \log_2 (d/2) \rfloor, 1, d) \) and \( g \in \mathcal{F}(d, 1, 2d, d/2) \), it follows from a composition argument that \( h^{(d)} = g \circ h^{(d/2)} \in \mathcal{F}(d, \lfloor \log_2 d \rfloor, 2d, 1) \).

The case \( d = 2k + 1 \) for some \( k \geq 1 \) can be dealt with similarly. By Lemma B.1, we can build \( \psi_i(x_{2i-1}, x_{2i}) \in \mathcal{F}(2, 1, 4, 1) \) such that \( \psi_i(x_{2i-1}, x_{2i}) \) for each \( 1 \leq i \leq k \), and \( \psi_{k+1}(x_{2k+1}) = x_{2k+1} \in \mathcal{F}(1, 1, 2, 1) \). Parallelization argument ensures that we can construct \( g = (\psi_1, \ldots, \psi_{k+1}) \in \mathcal{F}(d, \lfloor \log_2 d \rfloor, 2d, 1) \).
\( F(d, 1, 2d, (d + 1)/2) \) such that \( g(x_1, \ldots, x_d) = (x_1 \lor x_2, \ldots, x_{d-2} \lor x_{d-1}, x_d) \). By the composition argument, we further have \( h^{(d)} = g \circ h^{(d+1)/2} \in F(d, [\log_2 d], 2d, 1) \).

Now we have \( f(x) = (|x_1|, \ldots, |x_d|) \in F(d, 1, 2d, d) \) and \( h^{(d)}(x) = \max_{1 \leq i \leq d} x_i \in F(d, [\log_2 d], 2d, 1) \). Finally, taking \( f^\dagger = h^{(d)} \circ f \), we have \( f^\dagger(x) = ||x||_\infty \) and \( f^\dagger \in F(d, 1 + [\log_2 d], 2d, 1) \), as claimed. \( \square \)

### B.2 Proof of Proposition 3.3

The proof is based on the following neural network approximation result for \( d \)-variate \((\beta, C)\) smooth functions.

**Lemma B.3** (Theorem 1.1 in Lu et al. (2020)). Let \( f_0 \) be a \( d \)-variate, \((s, C)\)-smooth function with \( s \in \mathbb{N} \). For any \( N, L \in \mathbb{N} \), there exists a function \( \hat{f} \) from a deep ReLU network with depth \( c_{34}(L + 2) \log_2(4L) + 2d \) and width \( c_{35}(N + 2) \log_2(8N) \) such that

\[
||\hat{f} - f||_\infty \leq c_{36}(NL)^{-\frac{2\beta}{d}},
\]

where \( c_{34} = 17s^{d+1}3^d, \ c_{35} = 18s^2, \) and \( c_{36} = 85(s + 1)^d8^sC \).

Now we are ready to prove Proposition 3.3. The proof is based on a similar argument to that in Section 4 of Kohler & Langer (2021). The key idea is that neural network approximation is preserved under compositions. To be specific, if \( f \) and \( g \) can be approximated by neural networks \( \hat{f} \) and \( \hat{g} \), each with an \( || \cdot ||_\infty \)-error of \( \epsilon \), and \( g \) is an \( L \)-Lipschitz function, then \( \hat{g} \circ \hat{f} \) approximates \( g \circ f \) with an \( || \cdot ||_\infty \)-error of \( (L + 1)\epsilon \). The former ‘\( \circ \)’ refers to the network composition introduced in Section B.1, and the latter ‘\( \circ \)’ refers to function composition. Therefore, suppose the target \( f_0 \) is a composition of several low-dimensional smooth functions \( g_1, \ldots, g_l \), then in order to approximate \( f_0 \), we only need to approximate each \( g_i \) sufficiently well.

The current argument differs from that in Kohler & Langer (2021) in two aspects. First, we rely on Lemma B.3 in terms of approximating smooth functions. Compared with the proof in Kohler & Langer (2021), it has three benefits. (1) It leads to more flexible choices of the depth and width parameters, whereas their proof requires \( N = O(1) \) or \( L = O(\log N) \). (2) The sub-networks \( \hat{g}_i \) that approximate the components \( g_i \) now have the same order of depth and width, i.e., for each \( \hat{g}_i \), it has depth \( L \log L \) and width \( N \log N \). (3) The connection between the approximation error of the target function \( f_0 \) and the approximation errors of its components \( g_i \), and the connection between the approximation error of \( f_0 \) and the network architecture configuration (i.e., depth and width) are described more clearly in our proof. To be specific, assume each \( g_i \) is a \( d_i \)-variate, \((\beta_i, C)\)-smooth function. From Lemma B.3 we see that \( ||\hat{g}_i - g_i||_\infty \leq (NL)^{-2\beta_i/d_i} \). Then the approximation error for \( f_0 \) can be controlled by \( \max_i(NL)^{-2\beta_i/d_i} = (NL)^{-2\min_i\beta_i/d_i} \). Secondly, in order to better evaluate the approximation error based on compositions, it is necessary to truncate each smooth function it approximates because the smooth function approximation result only holds in the region \([0, 1]^d\).

**Proof of Proposition 3.3.** By Condition 1, \( f_0(x) = h_1^{(d)}(x) \) belongs to the function class \( \mathcal{H}(l, \mathcal{P}) \) for some integer \( l > 1 \). To obtain \( h_1^{(d)}(x) \in \mathcal{H}(l, \mathcal{P}) \), one needs to compute various hierarchical composition models at level \( i \in \{1, \ldots, l-1\} \), the number of which is denoted by \( M_i \). At level \( i \in \{1, \ldots, l\} \), let \( h_j^{(d)} : \mathbb{R}^d \to \mathbb{R} \) be the \( j \)-th \( (j \in \{1, \ldots, M_i\}) \) hierarchical composition model. The dependence of \( h_j^{(d)} \) on \( h_{j-1}^{(d)} \) depends on a \((\beta_j^{(d)}, C)\)-smooth function \( g_j^{(d)} : \mathbb{R}^{d_j} \to \mathbb{R} \) for some \((\beta_j^{(d)}, l_j^{(d)}) \in \mathcal{P} \). Recursively,
\( h_1^{(i)}(\cdot) \) is defined as
\[
    h_1^{(i)}(x) = g_j^{(i)}\left( \frac{h^{(i-1)}_{\xi+1}}{\sum_{\xi+1} h^{(i-1)}_{\xi+1}}(x), \ldots, \frac{h^{(i-1)}_{\xi_j+1}}{\sum_{\xi_j+1} h^{(i-1)}_{\xi_j+1}}(x) \right)
\]
for \( j \in \{1, \ldots, M_j\} \) and \( i \in \{2, \ldots, l\} \), and
\[
    h_1^{(1)}(x) = g_j^{(1)}\left( x_{\pi(\xi+1)}, \ldots, x_{\pi(\xi_j+1)} \right)
\]
for some \( \pi : \{1, \ldots, M_1\} \to \{1, \ldots, d\} \). The quantities \( M_1, \ldots, M_l \) can be defined recursively as
\[
    M_i = \begin{cases} 
        1 & i = l, \\
        \sum_{M_{i+1}} t^{(i+1)}_j & i \in \{1, \ldots, l-1\}.
    \end{cases}
\]
Moreover, define
\[
    K = \max_{i \in \{1, \ldots, l\}, j \in \{1, \ldots, M_i\}} \|g_j^{(i)}\|_{\infty} + 1
\]
and let \( \mathcal{D}_j^{(i)} \) be the domain of function \( g_j^{(i)} \) under the hierarchical composition model, i.e.,
\[
    \mathcal{D}_j^{(i)} = \left\{ \left( \frac{h^{(i-1)}_{\xi+1}}{\sum_{\xi+1} h^{(i-1)}_{\xi+1}}(x), \ldots, \frac{h^{(i-1)}_{\xi_j+1}}{\sum_{\xi_j+1} h^{(i-1)}_{\xi_j+1}}(x) : x \in [0, 1]^d \right) \right\} \quad i \in \{2, \ldots, l\}
\]
\[
    \mathcal{D}_j^{(1)} = \left\{ \left( \frac{h^{(i-1)}_{\xi+1}}{\sum_{\xi+1} h^{(i-1)}_{\xi+1}}(x), \ldots, \frac{h^{(i-1)}_{\xi_j+1}}{\sum_{\xi_j+1} h^{(i-1)}_{\xi_j+1}}(x) : x \in [0, 1]^d \right) \right\} \quad i = 1.
\]
We thus have \( \mathcal{D}_j^{(i)} \subseteq [-K, K]^{(i)} \). Without loss of generality we may assume \( \mathcal{D}_j^{(i)} = [-K, K]^{(i)} \); otherwise we can simply extend \( g_j^{(i)} \) to the cube \([-K, K]^{(i)} \) and the following analysis remains valid.

\textbf{Step 1. Construction of neural network.} In the rest of the proof, for notational convenience we use \( \mathcal{F}(N, L) \) to denote a deep ReLU neural network with width \( N \), depth \( L \).

Fix \( i \in \{1, \ldots, l\} \) and \( j \in \{1, \ldots, M_i\} \). Note that \( g_j^{(i)} \) is a \( \tau_j^{(i)} \)-variate, \((\beta_j^{(i)}, C)\)-smooth function defined on \([-K, K]^{(i)} \). Define the function
\[
    \tilde{g}_j^{(i)}(z) = g_j^{(i)}(2Kz - K) \quad \text{for } z \in [0, 1]^d,
\]
so that \( \tilde{g}_j^{(i)} \) is a \( \tau_j^{(i)} \)-variate, \((\beta_j^{(i)}, KC)\)-smooth function defined on \([0, 1]^{(i)} \), and satisfies
\[
    g_j^{(i)}(z) = \tilde{g}_j^{(i)}\left( \frac{z + K}{2K} \right) \quad \text{for } z \in \mathcal{D}_j^{(i)}.
\]
For any given \( N, L \in \mathbb{N} \), Lemma B.3 ensures that there exists a function \( \tilde{g}_j^{(i)} \) from some deep ReLU neural network \( \tilde{g}_j^{(i)} \) with depth \( L' = C_1 L \log_2 L + 2\tau_j^{(i)} \) and width \( N' = C_2 N \log_2 N \) such that
\[
    \left\| \tilde{g}_j^{(i)}\left( \frac{z + K}{2K} \right) - \tilde{g}_j^{(i)}\left( \frac{z + K}{2K} \right) \right\|_{\infty} \leq C_3 (NL)^{-2\tau_j^{(i)} \gamma} \quad \text{for all } z \in \mathcal{D}_j^{(i)}.
\]
It should be noted that the constants \( C_1, C_2 \) and \( C_3 \) may depend on the parameters \((\beta_j^{(i)}, \tau_j^{(i)})\). Since there are only finitely many \( g_j^{(i)} \), we can simply choose \((C_1, C_2)\) to be the largest among all \((C_1, C_2)\)
depending on \((g^{i(j)}, t^{i(j)})\) because \(\mathcal{F}(N, L) \subseteq \mathcal{F}(N', L')\) if \(N \leq N'\) and \(L \leq L'\). Similarly, we also choose \(C_3\) to be the largest among all \(C_3\)’s.

Next, consider a ‘truncated’ version of \(g_j^{i(j)}\), defined as

\[
\tilde{g}_j^{i(j)}(z) = \max\{\min[g_j^{i(j)}(z), K], -K\} = \sigma(2K - \sigma(K - \tilde{g}_j^{i(j)}(z))) - K,
\]

where \(\sigma(x) = \max(x, 0)\) is the ReLU activation function. Note that \(\|T_K f - g\|_\infty \leq \epsilon\) if \(\|g\|_\infty \leq K\) and \(\|f - g\|_\infty \leq \epsilon\). Therefore, we have \(g_j^{i(j)} \in \mathcal{F}(N', L' + 2)\) and

\[
\left\|\tilde{g}_j^{i(j)}\left(\frac{z + K}{2K}\right) - \frac{\tilde{g}_j^{i(j)}(z)}{2K}\right\|_\infty \leq C_3(NL)^{-2\gamma_i^{i(j)}(1)} \leq C_3(NL)^{-2\gamma} \quad \text{for all} \quad z \in D_j^{i(j)}.
\]

Now we are ready to construct a neural network \(f^\dagger\) to approximate \(f_0 = h_1^{i(1)}\). To be specific, our construction proceeds recursively as

\[
\tilde{h}_1^{i(1)}(x) = \tilde{g}_j^{i(1)}\left(\frac{x_1^{(1)} + K}{2K}, \ldots, \frac{x_K^{(1)} + K}{2K}\right)
\]

and

\[
\tilde{h}_j^{i(j)}(x) = \tilde{g}_j^{i(j)}\left(\frac{\tilde{h}_1^{(i-1)}(x) + K}{2K}, \ldots, \frac{\tilde{h}_K^{(i-1)}(x) + K}{2K}\right).
\]

The corresponding composited network, denoted by \(\tilde{f} = \tilde{g}(\tilde{h}(x), \beta_1, \ldots, \alpha_k \tilde{h}_k + \beta_k)\), is realized by first applying network composition \(L_1 \circ \tilde{h}_1\) for each \(i \in \{1, \ldots, k\}\), where \(L_i(x) = \alpha_i x + \beta_i\), followed by network parallelization \((L_1 \circ \tilde{h}_1(x), \ldots, L_k \circ \tilde{h}_k(x))\), and then followed by network composition \(\tilde{g} \circ (L_1 \circ \tilde{h}_1(x), \ldots, L_k \circ \tilde{h}_k(x))\). For \(i \in \{1, \ldots, k\}\), assume the deep ReLU neural network \(\tilde{h}_i : \mathbb{R}^d \rightarrow \mathbb{R}\) has depth \(L_{h_i}\) and width \(N_{h_i}\), and the deep ReLU neural network \(\tilde{g}\) has depth \(L_g\) and width \(N_g\). Following the discussions on the depth and width of network composition and network parallelization in Section B.1, we conclude that the composited network \(\tilde{f}\) has depth \((\max L_{h_i}) + L_g\) and width \((\sum_{i=1}^k N_{h_i}) \lor N_g\).

Based on the recursive construction of neural networks, we set \(f^\dagger\) to be \(h_1^{i(1)}\). Now it suffices to calculate the width, depth and approximation error of \(h_1^{i(1)}\). These quantities will also be calculated recursively.

**Step 2. Specifying Width and Depth.** The goal is to calculate the width and depth of each \(h_j^{i(j)}\) from \(i = 1\) to \(i = l\). Let \(L_j^{i(j)}\) and \(N_j^{i(j)}\) be the width and depth of the network \(h_j^{i(j)}\). First, by Lemma B.3 and the discussion before, for each \(j \in \{1, \ldots, M_j\}\), the depth and width satisfy

\[
L_j^{(1)} = C_1 L \log_2 L + 2(t_j^{(1)} + 1), \quad N_j^{(1)} = C_2 N \log_2 N.
\]

Now suppose we have already calculated the depth and width for all \(h_j^{(i-1)}\). Then, based on our discussion of the composited network before, for any given \(j \in \{1, \ldots, M_j\}\), the depth and width of \(h_j^{i(j)}\) satisfy

\[
L_j^{i(j)} = \max_{j' \in P_{i(j)}} L_j^{(i-1)} + C_1 L \log_2 L + 2(t_j^{(i-1)} + 1), \quad N_j^{i(j)} = \sum_{j' \in P_{i(j)}} N_j^{(i-1)}.
\]
where $P(i, j) = (\sum_{\ell=1}^{i-1} t^{(i)}_{\ell} + 1, \ldots, \sum_{\ell=1}^{j} t^{(i)}_{\ell})$. Using the above recursive calculation, the depth of $f^t = \tilde{h}^{(i)}_1$ can be written as

$$L = (lC_1)L \log L + \sum_{j=1}^{m} \max_{i=1}^{l} t^{(i)}_{j} \leq C_4L \log_2 L + C_5 \leq c_3L \log L,$$

while the depth of $f^t = \tilde{h}^{(i)}_1$ can be written as

$$\tilde{N} = N^{(i)}_1 \leq \left( \frac{M_1C_2}{\log 2} \right) N \log N.$$

**Step 3. Approximation Error.** Let $t^* = \max_i t^{(i)}_j$. We claim that

$$\|\tilde{h}^{(i)}_j - h^{(i)}_j\|_{\infty} \leq C_3(C \sqrt{t^*} + 1)^{i-1}(NL)^{-2\gamma}.$$  

We prove inequality (B.3) by mathematical induction, starting with the case of $i = 1$. By our discussion in Step 1, let $z = (x_{\pi(\sum_{t=1}^{i-1} t^{(i)}_{t} + 1), \ldots, x_{\pi(\sum_{t=1}^{i} t^{(i)}_{t})})$, we have for all $x \in [0, 1]^d$ that

$$\tilde{h}^{(i)}_j(x) - h^{(i)}_j(x) = \left| g^{(i)}_j \left( \frac{z + K}{2K} \right) - g^{(i)}_j(z) \right|
= \left| g^{(i)}_j \left( \frac{z + K}{2K} \right) - g^{(i)}_j(z) \right|
\leq C_3(NL)^{-2\gamma},$$

where the last step follows from (B.2).

Suppose (B.3) holds for $i - 1$ and $j \in [1, \ldots, M_{i-1}]$. Write $z = (h^{(i-1)}_{\sum_{t=1}^{i-1} t^{(i)}_{t} + 1}(x), \ldots, h^{(i-1)}_{\sum_{t=1}^{i} t^{(i)}_{t}}(x))$ and $\tilde{z} = (\tilde{h}^{(i-1)}_{\sum_{t=1}^{i-1} t^{(i)}_{t} + 1}(x), \ldots, \tilde{h}^{(i-1)}_{\sum_{t=1}^{i} t^{(i)}_{t}}(x))$ for $x \in [0, 1]^d$, we have

$$\tilde{h}^{(i)}_j(x) - h^{(i)}_j(x) = \left| g^{(i)}_j \left( \frac{\tilde{z} + K}{2K} \right) - g^{(i)}_j(z) \right|
\leq \left| \frac{g^{(i)}_j \left( \frac{\tilde{z} + K}{2K} \right) - g^{(i)}_j(\tilde{z})}{g^{(i)}_j \left( \frac{\tilde{z} + K}{2K} \right) - g^{(i)}_j(\tilde{z})} \right| + |g^{(i)}_j(\tilde{z}) - g^{(i)}_j(z)|.$$

Together, (B.2) and the fact that $\tilde{z} \in [-K, K)^{i}$ imply

$$\left| g^{(i)}_j \left( \frac{\tilde{z} + K}{2K} \right) - g^{(i)}_j(\tilde{z}) \right| \leq C_3(NL)^{-2\gamma}.$$  

Since $g^{(i)}_j$ is at least $C$-Lipschitz, we further have

$$|g^{(i)}_j(\tilde{z}) - g^{(i)}_j(z)| \leq C\|\tilde{z} - z\|_2
\leq C \sqrt{t^*}\|\tilde{z} - z\|_\infty
\leq C \sqrt{t^*}(1 + C \sqrt{t^*})^{i-2}C_3(NL)^{-2\gamma},$$

43
where the last inequality follows from the induction. Putting together the pieces, we obtain
\[
\hat{h}_j^{(i)}(x) - h_j^{(i)}(x) \leq \left| \frac{s_j^{(i)}(\tilde{z} + K)}{2K} - s_j^{(i)}(\tilde{z}) \right| + |s_j^{(i)}(\tilde{z}) - s_j^{(i)}(z)| \\
\leq C_3(NL)^{-2}\gamma + C_3C \sqrt{r}(1 + C \sqrt{r})^{-2}(NL)^{-2}\gamma \\
\leq C_3(1 + C \sqrt{r})^{-1}(NL)^{-2}\gamma.
\]
Finally, we conclude that
\[
\|f^\dagger - f_0\|_\infty = \|\hat{h}_1^{(i)} - h_1^{(i)}\|_\infty \leq \frac{C_3(C \sqrt{r} + 1)^{d-1}(NL)^{-2}\gamma}{c_5},
\]
as claimed. \qed

B.3 Proof of Theorem 4.5

To prove theorem 4.5, we need several technical lemmas in Lu et al. (2020) to build some basic modules via deep ReLU neural networks. Having these basic modules in hand, we can apply parallelization or composition to construct more complicated functions that we are interested in.

**Lemma B.4** (Step function). For any $N, L, d \in \mathbb{N}^+$, and $\Delta \in (0, 1/3K]$ with $K = [L^{2/d}][N^{1/d}]^2$, there exists an ReLU neural network $\phi$ with depth $4L + 5$ and width $4[N^{1/d} + 3]$ such that
\[
\phi(x) = k \quad \text{if} \quad x \in [k/K, (k + 1)/K - 1_{[k+1<K]\Delta}]
\]
for $k = 0, 1, \ldots, K - 1$.

**Lemma B.5** (Point fitting). For any $N, L \in \mathbb{N}^+$, and $\theta_i \in \{0, 1\}$ for $i \in \{0, \ldots, N^2L^2 - 1\}$, there exists a function $\phi : \mathbb{R} \to \mathbb{R}$ determined by an ReLU neural network with depth $5L + 7$ and width $8N + 6$ such that
\[
\phi(i) = \theta_i \quad \text{for} \quad i = 0, 1, \ldots, (NL)^2 - 1.
\]

With the above step function and point fitting modules, we are ready to prove Theorem 4.5.

**Proof of Theorem 4.5.** Our target point fitting network $f^\dagger$ consists of two modules: the encoder module $f_e(\cdot)$ and the decoder module $f_d(\cdot)$. The encoder module $f_e(\cdot)$ takes $x \in [0, 1]^d$ as input and outputs an integer index $I(\alpha) = \sum_{i=1}^{d} (\alpha_i - 1)K^{d-1}$ with $\alpha$ satisfying $x \in Q_\alpha(\Delta)$. The decoder module takes the index $I(\alpha) \in \{0, \ldots, K^d - 1\}$ as input and outputs a value that approximates $y_a$.

**Step 1. Construct Encoder $f_e(\cdot)$.** Given any $N, L \in \mathbb{N}^+$, let $\tilde{L} = [L^{1/d}]^d$ so that $\tilde{L} \leq L$ and $[\tilde{L}^{2/d}] = [L^{1/d}]^2$. For any $i \in \{1, \ldots, d\}$, applying Lemma B.4 with $L = \tilde{L}$ and $K = [\tilde{L}^{2/d}][N^{1/d}]^2 = [L^{1/d}]^2[N^{1/d}]^2$, there exists an ReLU neural network $\phi_i$ with depth at most $4\tilde{L} + 5 \leq 4L + 5$ and width $4N + 3$ such that
\[
\phi_i(x_i) = k \quad \text{if} \quad x_i \in [k/K, (k + 1)/K - 1_{[k+1<K]\Delta}]
\]
for $k = 0, \ldots, K - 1$. Via parallelization, the function $g_1 = (\phi_1, \ldots, \phi_d) \in \mathcal{F}(d, 4L+5, (4N+3)d, d) : \mathbb{R}^d \to \mathbb{R}^d$ satisfies
\[
g_1(x) = (\alpha_1 - 1, \ldots, \alpha_d - 1) \quad \text{if} \quad x \in Q_\alpha(\delta).
\]

44
Moreover, let \( g_2(x) = \sum_{i=1}^{d} x_i K_i^{-1} \) so that \( g_2 \in \mathcal{F}(d, 0, 0, 1) \). By the composition argument, we have \( f = g_2 \circ g_1 \in \mathcal{F}(d, 4L + 5, (4N + 3)\delta, 1) \), satisfying
\[
f_\gamma(x) = I(\alpha) \quad \text{if} \quad x \in Q_\alpha(\delta)
\]
for all \( \alpha \in \{1, \ldots, K\}^d \).

**Step 2. Construct Decoder \( f_d(\cdot) \) in Two Ways.** First, note that every \( u \in [0, 1] \) can be written as \( u = \sum_{i=0}^{\infty} 2^{-i} \theta_i \). Set \( u^{(r)} = \sum_{i=0}^{r} 2^{-i} \theta_i \) with \( r = \lceil \log(1/\epsilon) \rceil \) such that
\[
|u - u^{(r)}| \leq \sum_{i=r+1}^{\infty} 2^{-i} \leq \sum_{i=r+1}^{\infty} 2^{-r} = 2^{-r}.
\]
Therefore, we have \( |u^{(r)} - u| \leq \epsilon \). Moreover, for any \( u = \sum_{i=0}^{s} 2^{-i} \theta_i \) with \( \theta_i \in \{0, 1\} \) and some \( s \leq r \), we have \( u - u^{(r)} = 0 \).

By the above discussions, we only need to build a neural network to fit \( y^{(r)}_\alpha \) for each \( \alpha \). To be specific, in this part, we target to build a neural network \( f_d(\cdot) \) such that
\[
f_d(I(\alpha)) = y^{(r)}_\alpha \quad \text{for all} \quad \alpha \in \{1, \ldots, K\}^d.
\]

Let \( S = ([L^{1/d}]^d[N^{1/d}]^d)^2 \) and note that \( f(\cdot) \) is a bijective map from \( [1, \ldots, K]^d \) to \( [0, \ldots, S - 1] \). Thus we let \( y^{(r)}_\alpha = \sum_{i=0}^{r} 2^{-i} \theta_i f(\alpha) \). Then for each \( i \in \{0, \ldots, r\} \), it follows from Lemma B.5 that there exists some \( \psi_i(\cdot) \in \mathcal{F}(1, 5[L^{1/d}]^d + 7, 8[N^{1/d}]^d + 6, 1) \subset \mathcal{F}(1, 5L + 7, 8N + 6, 1) \) such that
\[
\psi_i(I(\alpha)) = \theta_i f(\alpha) \quad \text{for all} \quad \alpha \in \{1, \ldots, K\}^d.
\]
Finally, we use \( \psi_i(\cdot) \) to construct \( f_d(\cdot) \) in two ways.

**Case 1 (Parallel).** We claim that there exists some \( f_d \in \mathcal{F}(1, 5L + 7, (8N + 6)(r + 1), 1) \) such that \( f_d(I(\alpha)) = y^{(r)}_\alpha \). Via the parallelization argument, we have \( g = (\psi_0, \cdots, \psi_r) \in \mathcal{F}(1, 5L + 7, (8N + 6)(r + 1), r + 1) \). Moreover, we have \( h(x) = \sum_{i=0}^{r} 2^{-i} x_i \in F(r + 1, 0, 0, 1) \). Then using the composition argument we conclude that \( f_d^{(1)} = h \circ g \in \mathcal{F}(1, 5L + 7, (8N + 6)(r + 1), 1) \) and
\[
f_d^{(1)}(x) = \sum_{i=0}^{r} 2^{-i} \psi_i(x) \quad \text{for any} \quad x.
\]

Combining this with (B.6), it is easy to see that \( f_d^{(1)} \) meets the requirements in (B.5).

**Case 2 (Series).** Here we consider a different construction. Since \( \text{Id}(x) = x \in \mathcal{F}(1, 1, 2, 1) \), and by the parallelization argument, we have \( g_0(x) = (\psi_0(x), \text{Id}(x)) \in \mathcal{F}(1, 5L + 7, 8N + 8, 2) \). Moreover, define
\[
g_i(x, y) = \begin{cases} 2^{-i} \psi_i(y) + x, & \text{Id}(y) \\ 2^{-i} \psi_i(y) + x, & i = r. \end{cases}
\]
Note that \( g_i \) can be constructed by first applying parallelization to \( (\psi_i, \text{Id}) \), followed by a composition with a linear function. We thus have \( g_i(x, y) \in \mathcal{F}(2, 5L + 7, 8N + 10, 2) \). Finally, by the composition argument and induction, we conclude that \( f_d^{(2)} = g_d \circ g_{d-1} \circ \cdots \circ g_0 \in \mathcal{F}(1, (5L + 7)(r + 1), 8N + 10, 1) \) and
\[
f_d^{(2)}(x) = \sum_{i=0}^{r} 2^{-i} \psi_i(x).
\]
Combined with the property of $\psi(\cdot)$ in (B.6), our constructed $f^{(2)}_d$ satisfies the requirements in (B.5).

**Step 3. Composition of $f_\epsilon(\cdot)$ and $f_d(\cdot)$.** To conclude, let $f^\dagger_1 = f_\epsilon^{(1)} \circ f_\epsilon$ and $f^\dagger_2 = f_\epsilon^{(2)} \circ f_\epsilon$. By the composition argument, we have

$$f^\dagger_1 \in \mathcal{F}(d, 9L + 12, (4N + 3)d \lor (8N + 6)(r + 1), 1)$$

and

$$f^\dagger_2 \in \mathcal{F}(d, 4L + 5 + (5L + 7)(r + 1), (4N + 3)d \lor (8N + 10), 1).$$

Moreover, for each $s \in \{1, 2\}$,

$$f^\dagger_s(x) = f_d(f_\epsilon(x)) = f_d^{(s)}(I(\alpha)) = y^{(s)}_\alpha \quad \text{if} \quad x \in Q_\alpha(\Delta), \quad \alpha = 1, \ldots, K.$$

From the discussions in Step 2, the claimed approximation error of $f^\dagger_1$ holds.  \hfill \square

**B.4 Proof of Theorem 4.6**

We first prove a weaker version of Theorem 4.6.

**Proposition B.1.** For any given $N, L \in \mathbb{N}^+$, let $K = [N^{1/d}]^2[1^{L/d}]^2$. Then for any $\Delta_1 \in (0, 1/3K]$, $\Delta_2 > 0$, suppose $(x_\alpha)_{\alpha \in \mathcal{A}}$ is an arbitrary set of points indexed by $\mathcal{A} = \{1, \ldots, K\}$ satisfying $x_\alpha \in Q_\alpha(\Delta_1)$, where $Q_\alpha(\Delta)$ is defined in (4.8). Then there exist an ReLU neural network $f^\dagger_1$ with depth $c_{37}L[\log_2(1/\Delta_2)]$ and width $c_{38}N$ and an ReLU neural network $f^\dagger_2$ with depth $c_{39}L$, $c_{40}N[\log_2(1/\Delta_2)]$ satisfying

$$f^\dagger_1(x_\alpha) = 1 \quad \text{for all} \quad \alpha \in \mathcal{A},$$

and

$$f^\dagger_2(x) = 0 \quad \text{if} \quad x \in Q_\alpha(\Delta) \quad \text{for some} \quad \alpha \in \mathcal{A} \quad \text{subject to} \quad \|x - x_\alpha\|_\infty \geq \delta_2,$$

where $s = 1, 2$.

**Proof of Proposition B.1.** Suppose we have already constructed an ReLU neural network $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ with depth $\bar{L}_g$ and width $\bar{N}_g$, satisfying

$$g(x) = \tilde{x}_\alpha \quad \text{if} \quad x \in Q_\alpha,$$

where $\|\tilde{x}_\alpha - x_\alpha\|_\infty \leq \Delta_2/3$ holds for all $\alpha \in \mathcal{A}$. Consider the function

$$h(x, y) = \sigma\left(2 - (3/\Delta_2)\|x - y\|_\infty\right) \land 1, \quad x, y \in \mathbb{R}^d.$$

By the composition argument and Lemmas B.1 and B.2, we have

$$h \in \mathcal{F}(2d, [\log_2 d] + 3, 4d, 1).$$

Next we claim that $f^\dagger(x) = h \circ (g(x), \text{Id}(x))$ is the function of interest, where $\text{Id}(x) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the identify function. Combining the parallelization argument with Lemma B.1 yields $\text{Id}(x) \in \mathcal{F}(d, 1, 2d, d)$, which further implies

$$f^\dagger \in \mathcal{F}(d, \bar{L}_g + [\log_2 d] + 3, (\bar{N}_g + 2d) \lor 4d, 1).$$

46
To prove the claim, note that

\[
2 - (3/\Delta_2)\|x - \bar{x}_\alpha\|_{\infty} \in \begin{cases} [1, 2] & \text{if } \|x - \bar{x}_\alpha\|_{\infty} \leq \frac{\Delta_1}{3} \\
[0, 1) & \text{if } \frac{\Delta_1}{3} < \|x - \bar{x}_\alpha\|_{\infty} \leq \frac{2\Delta_1}{3} \\
(-\infty, 0) & \text{if } \frac{2\Delta_1}{3} < \|x - \bar{x}_\alpha\|_{\infty}
\end{cases}
\]

This means \(f(x_\alpha) = \sigma(2 - (3/\Delta_2)\|x_\alpha - \bar{x}_\alpha\|_{\infty}) 1 = 1\). Moreover, if \(x \in Q_\alpha(\Delta_1)\) but \(\|x - x_\alpha\|_{\infty} \geq \Delta_2\), it follows that \(\|x - \bar{x}_\alpha\|_{\infty} \geq \|x - x_\alpha\|_{\infty} - \|x_\alpha - \bar{x}_\alpha\|_{\infty} \geq \frac{2\Delta_1}{3}\). This implies \(f(x) = \sigma(2 - (3/\Delta_2)\|x_\alpha - \bar{x}_\alpha\|_{\infty}) = 0\).

It remains to show that we can implement \(g\) using an ReLU neural network. To be specific, we will use the two different configurations in Theorem 4.5 to construct \(g\). Consider first the one that multiplies the \(\log(1/\epsilon)\) factor to the depth. For any \(N, L \in \mathbb{N}^+\) and \(i \in \{1, \ldots, d\}\), applying Theorem 4.5 with \(\Delta = \Delta_1\) and \(\epsilon = \Delta_2/3\) to the set \(\{(x_\alpha)_{\alpha \in \mathcal{A}}\}\), we see that there exists a neural network \(\phi_i \in \mathcal{F}(d, C_1L \log(3/\Delta_2), C_2N, 1)\) such that

\[
|\phi_i(x) - (x_\alpha)| \leq \Delta_2/3 \quad \text{if } x \in Q_\alpha(\Delta_1)
\]

(B.7)

for all \(\alpha \in \mathcal{A}\). By the parallelization argument, \(g_1 = (\phi_1, \cdots, \phi_d) \in \mathcal{F}(d, C_1L \log(3/\Delta_2), C_2dN, d)\) and \(\|g_1(x) - x_\alpha\|_{\infty} \leq \max_{1 \leq i \leq d} |\phi_i(x) - (x_\alpha)| \leq \Delta_2/3\) if \(x \in Q_\alpha(\Delta_1)\). In this case, the final \(f^1_1 = h \circ (g_1(x), Id(x))\) has depth at most \(C_1L \log_2(3/\Delta_2) + \lfloor \log d \rfloor + 3 \leq c_{37}L \log_2(3/\Delta_2)\) and width at most \((C_2dN + 2d) \vee 4d \leq c_{38}N\).

For the construction of \(f^1_2\), we follow the same arguments except use the neural network \(\phi_i \in \mathcal{F}(d, C_1^2L, C_2^2N \log(3/\Delta_2), 1)\) such that (B.7) holds for each \(i \in \{1, \ldots, d\}\). Therefore, \(f^1_2 = h \circ (g_2(x), Id(x))\) satisfies all the requirements and has depth at most \(C_1^2L + \log(3/\Delta_2) + 3 \leq c_{39}L\) and width at most \((C_2^2dN \log_2(3/\Delta_2) + 2d) \vee 4d \leq c_{40}N \log_2(3/\Delta_2)\). This completes the proof. \(\Box\)

Now having the constructed module in Proposition B.1, we are ready to prove Theorem 4.6.

Proof of Theorem 4.6. Let \(N, L \in \mathbb{N}^+, \Delta_1 \in (0, 1/3K]\), \(\Delta_2 > 0\) be arbitrary, and write \(\mathcal{A} = \{1, \ldots, K\}^d\) and

\[
\mathcal{A}^+ = \{\alpha \in \mathcal{A} : y_\alpha = 1\} \quad \text{and} \quad \mathcal{A}^- = \{\alpha \in \mathcal{A} : y_\alpha = -1\}.
\]

We first construct \(f^1_1\). Let \(X^+ = \{x^+_{\alpha}\}_{\alpha \in \mathcal{A}^+}\) satisfy \(x^+_{\alpha} = x_\alpha\) if \(\alpha \in \mathcal{A}^+\) and we choose any \(x^+_{\alpha} \in Q_\alpha(\Delta_1)\) if \(\alpha \notin \mathcal{A}^+\). By Proposition B.1 with given \(N, L, \Delta_1, \Delta_2\), and our sample set \(X^+\), there exists a neural network \(f^+\) with depth at most \(C_1L \log(1/\Delta_2)\) and width at most \(C_2N\) such that

\[
f^+(x_\alpha) = f^+(x^+_{\alpha}) = 1 \quad \text{for } \alpha \in \mathcal{A}^+
\]

and

\[
f^+(x) = 0 \quad \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \in \mathcal{A}^+ \text{ but } \|x - x_\alpha\|_{\infty} \geq \Delta_2.
\]

Now we try to apply Theorem 4.5 with given \(N, L, \Delta = \Delta_1, \epsilon = 1/2\), and sample set \(Y^+ = \{y^+_{\alpha}\}_{\alpha}\) defined as \(y^+_{\alpha} = 1_{\alpha \in \mathcal{A}^+}\). Because \(y^+_{\alpha}\)'s are all in \([0, 1]\), there exists an ReLU neural network \(m^+(x)\) with depth at most \(C_3L\) and width at most \(C_4N\) such that

\[
m^+(x) = 1_{\alpha \in \mathcal{A}^+} \quad \text{if } x \in Q_\alpha(\Delta_1)
\]

and
for all $\alpha \in \mathcal{A}$.

Here we consider the function $g^+(x) = f^+(x) \land m^+(x)$, then we have the following

$$f^+(x) \land m^+(x) = \begin{cases} 0 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \notin \mathcal{A}^+ \\ 1 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \in \mathcal{A}^+ \text{ and } x = x_\alpha \\ 0 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \in \mathcal{A}^+ \text{ but } ||x - x_\alpha||_\infty \geq \Delta_2 \end{cases}.$$ 

Combining the parallelization and composition arguments with Lemma B.1, we have $g^+(x) \in \mathcal{F}(d,(C_1 L \log(1/\Delta_2) \lor C_3 L) + 1, (C_2 + C_4)N, 1)$. Similarly, applying Proposition B.1 to the sample set $X^- = \{x_\alpha\}_{\alpha \in \mathcal{A}}$ satisfying $x_\alpha = x_\alpha$ if $\alpha \in \mathcal{A}^-$, and applying Theorem 4.5 to the sample set $Y^- = \{y_\alpha\}_{\alpha \in \mathcal{A}}$ with $y_\alpha = 1_{[0,1]}$, we can construct $f^-$ and $m^-$ satisfying

$$g^-(x) = f^-(x) \land m^-(x) = \begin{cases} 0 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \notin \mathcal{A}^- \\ 1 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \in \mathcal{A}^- \text{ and } x = x_\alpha \\ 0 & \text{if } x \in Q_\alpha(\Delta_1) \text{ with some } \alpha \in \mathcal{A}^- \text{ but } ||x - x_\alpha||_\infty \geq \Delta_2 \end{cases}$$

and $g^- \in \mathcal{F}(d,(C'_1 L \log(1/\Delta_2) \lor C'_3 L) + 1, (C'_2 + C'_4)N, 1)$ for some constants $C'_1 - C'_4$. Now we are able to conclude that the function

$$f^+_1(x) = (2g^+(x) - 2g^-(x) + u) \lor (-1) \lor 1$$

satisfies the conditions stated in Theorem 4.6. If $x = x_\alpha$ with $\alpha \in \mathcal{A}^+$, we have $(2g^+(x) - 2g^-(x) + u) = 2 + u - 0 \geq 1$, which implies $f^+_1(x) = 1$. Meanwhile, if $x = x_\alpha$ with $\alpha \in \mathcal{A}^-$, we have $(2g^+(x) - 2g^-(x) + i) = u - 2 \leq -1$, implying $f^+_1(x) = -1$.

At last, if $x \in Q_\alpha(\Delta_1)$ for some $\alpha \in \mathcal{A}$, and $||x - x_\alpha||_\infty \geq \Delta_2$ for all the $\tilde{\alpha} \in \tilde{\mathcal{A}}$, then we can divide it into two cases: for the first case, that $\alpha \notin \tilde{\mathcal{A}}$, by the condition that $g^+$ and $g^-$ satisfying, we have $g^+(x) = g^-(x) = 0$, this means $(2g^+(x) - 2g^-(x) + u) = u$ and $f^+_1(x) = u$ because $u \in [-1, 1]$; for the second case, that $\alpha \in \tilde{\mathcal{A}}$, then we have $||x - x_\alpha||_\infty \geq \Delta_2$ because it is hold for all the $\tilde{\alpha} \in \tilde{\mathcal{A}}$ by assumption, this gives $g^+(x) = 0$ if $\alpha \in \mathcal{A}^+$ or $g^-(x) = 0$ if $\alpha \in \mathcal{A}^-$, by using the same reason in the first case for another $g$ function, we can further conclude that $g^+(x) = g^-(x) = 0$. So we can conclude $f^+_1(x) = u$.

Finally, let us conclude by specifying the depth and width for $f^+_1$, by using parallelization of $g^+$ and $g^-$ together with the composition of a given ReLU neural network with sum, min, max function implemented by ReLU neural network, we have $f^+_1$ has depth at most

$$\left( (C_1 L \log(1/\Delta_2) \lor C_3 L) + 1 \right) \lor \left( (C'_1 L \log(1/\Delta_2) \lor C'_3 L) + 1 \right) + 2 \leq c_{22} L \log_2(1/\Delta_2),$$

and width at most

$$(C'_2 + C'_4)N + (C_2 + C_4)N \leq c_{23} N.$$ 

The construction of $f^+_2$ is almost the same except that we choose different configurations to implement $g^+$ and $g^-$. \qed
C Proof of Lower Bound Result

C.1 Proof of Theorem 4.2

In this part, we consider the following nonparameteric regression task

\[ Y = f_0(X) + \varepsilon, \]

where \( \varepsilon \) is a standard normal noise variable, \( X \sim \text{Uniform}[0, 1]^d \), and the regression function \( f_0 \in C(d, \beta) \). The key idea to prove Theorem 4.2 is that if we can achieve a faster approximation rate, i.e.,

\[ \omega(L, N) \equiv \sup_{f_0 \in C(d, \beta)} \inf_{f \in F} \| f - f_0 \|_2 = o\left((N^2 L^2 \log^4(NL))^{-\beta/d}\right), \]

then we can obtain a better convergence rate than the minimax optimal rate, which is impossible.

Before proving our main theorem, we need some results from the nonparameteric regression literature.

The following Lemma characterizes the minimax lower bound for this nonparameteric regression task.

**Lemma C.1** (Theorem 3.2 in Györfi et al. (2002)). There exists a positive constant \( c_{41} > 0 \) such that

\[ \liminf_{n \to \infty} \inf_{f_0 \in C(d, \beta)} \sup_{f \in F} \frac{E[\| \hat{f}_n - f_0 \|_2^2]}{n^{-2\beta/d}} \geq c_{41}, \]

where the infimum is over any estimator \( \hat{f}_n \) of \( f_0 \) based on i.i.d. observations \( \{(X_i, Y_i)\}_{i=1}^n \) from model (C.1).

**Lemma C.2** below provides an upper bound on \( E[\| \hat{f}_n - f_0 \|_2^2] \) using a combination of \( \omega(L, N) \) and a statistical error term.

**Lemma C.2** (Lemma 18 in Kohler & Langer (2021)). Assume that the response variable \( Y \) satisfies \( E[\exp(c_{42} Y^2)] < \infty \) for some constant \( c_{42} > 0 \) and the regression function \( f_0 \) is uniformly bounded. Let \( \tilde{f}_n \) be the least squares estimator

\[ \tilde{f}_n = \arg\min_{f \in F_n} \frac{1}{n} \sum_{i=1}^n |Y_i - f(X_i)|^2 \]

constrained on some model class \( F_n \), and let \( \hat{f}_n = T_{c_41 \log n} \tilde{f}_n \) for some constant \( c_{41} > 0 \), where \( T_c \) is a truncation operator defined as \( (T_c f)(x) = \text{sgn}(f(x))(|f(x)| \wedge c) \). Then \( \hat{f}_n \) satisfies

\[ E[\| \hat{f}_n - f_0 \|_2^2] \leq \frac{c_{44}(\log n)\left(\log N_{\text{approx}}(\frac{1}{c_{43} \log(n)}; T_{c_41 \log(n)} F_n, n) + 1\right)}{n} + 2 \inf_{f \in F_n} ||f - f_0||_2^2, \]

where \( T_c F = \{T_c f : f \in F\} \).

Now we are ready to present the following Proposition, which establishes the lower bound of the approximation error either as \( NL \to \infty \) or when \( NL \leq M \) for some fixed \( M > 0 \).
Proposition C.1. Given $d, \beta \in \mathbb{N}^+$, let $C(d, \beta)$ be the set of functions defined in (4.1), and $\omega(L, N) = \sup_{f \in C(d, \beta)} \inf_{f \in \mathcal{F}(d, L, N, 1)} \|f - f_0\|_2$, then we have the following holds for $\omega(L, N)$:

(i) For any $\epsilon > 0$, it holds

$$\liminf_{NL \to \infty} \frac{\omega(L, N)}{(N^2 L^2 \log^{4+\epsilon}(NL))^{-\beta/d}} = c_{45} > 0$$

for some positive constant $c_{45}$ depending on $\epsilon$.

(ii) For any given $M > 0$, we have

$$\liminf_{Nk \leq M} \omega(L, N) = c_{46} > 0,$$

where $c_{46}$ is a constant that depends on $M$.

Proof. Part (i). We prove (C.5) by contradiction. Suppose the LHS of (C.3) is 0, then there exists a sequence of $(L_k, N_k)_{k=1}^\infty$ satisfying $N_k L_k \to \infty$ and

$$q_k = \frac{\omega(L_k, N_k)}{((N_k L_k)^2 \log^{4+\epsilon}(N_k L_k))^{-\beta/d}} \to 0.$$  (C.5)

Now we choose a sequence of $n_k$ such that

$$(N_k L_k)^2 \geq \frac{n_k^{\frac{4}{1+\epsilon}}}{(\log n_k)^{4+\epsilon}}.$$  (C.6)

Since $N_k L_k \to \infty$, we have $n_k \to \infty$.

Now we try to conduct nonparametric regression. In particular, consider the data generating process defined (C.1) with sample size $n_k$, and construct a truncated least squares estimator over the model class $\mathcal{F}(d, L_k, N_k, 1)$ as in Lemma C.2. Since $\epsilon$ is Gaussian and $f_0 \in C(d, \beta)$ is (uniformly) bounded, $Y = f_0(X) + \epsilon$ is a sub-Gaussian random variable, i.e. $\mathbb{E} \exp(C_1 Y^2) < \infty$ for some constant $C_1 > 0$. Moreover, Lemma A.5 implies

$$\log N_k \left(\frac{1}{n_k c_{43}} \log(n_k), T_{c_{43}} \log(n_k), \mathcal{F}_L, n_k\right) + 1 \leq \log \left(n_k^2 c_{43} \log(n_k)\right) (N_k L_k)^2 \log(N_k L_k).$$

Then it follows from Lemma C.2 that there exists an estimator $\hat{f}_{n_k}$ such that

$$\mathbb{E}[\|\hat{f}_{n_k} - f_0\|^2] \leq \frac{(n_k L_k)^2}{n_k} \log N_k L_k + \omega^2(L_k, N_k).$$  (C.7)

Plugging (C.5) and (C.6) into (C.7) yields

$$\mathbb{E}[\|\hat{f}_{n_k} - f_0\|^2] \leq \frac{(n_k L_k)^2}{n_k} \log N_k L_k + \omega_0^2((N_k L_k)^2 \log^{4+\epsilon}(N_k L_k))^{-2\beta/d}$$

$$\leq \frac{(n_k L_k)^2}{n_k} \frac{n_k^{\frac{4}{1+\epsilon}}}{(\log n_k)^{4+\epsilon}} + \omega_0^2(n_k^{\frac{4}{1+\epsilon}})^{-2\beta/d}$$

$$\leq n_k^{\frac{4}{1+\epsilon}} \left(\frac{q_k^2}{(\log n_k)^{1+\epsilon}} + \frac{1}{(\log n_k)^{1+\epsilon}}\right).$$
Hence, there exist some $n_k \to \infty$ and estimator $\hat{f}_{n_k}$ based on i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^{n_k}$ such that

$$\lim_{k \to \infty} \frac{\mathbb{E}[\|f_{n_k} - f_0\|^2]}{n_k^{2d/2d+1}} = \lim_{k \to \infty} \left(\frac{c_k^2}{\log n_k} + \frac{1}{(\log n_k)^c}\right) = 0.$$ 

This contradicts the minimax lower bound (C.2) stated in Lemma C.1, implying that the assumption

$$\lim_{NL \to \infty} \frac{\omega(L, N)}{(N^2L^2 \log^{4+\epsilon}(NL))^{-\beta/d}} = 0$$

cannot be true. This concludes the proof.

Part (ii). The proof of (ii) proceeds in a similar way via a contradiction argument. Suppose the LHS of (C.4) is 0, then there exists a sequence of $\{(L_n, N_n)\}_{n=1}^{\infty}$ such that $L_n N_n \leq M$ and $\omega(L_n, N_n) \to 0$ as $n \to \infty$. Without loss of generality, we assume that $\omega(L_n, N_n) \leq \frac{1}{\sqrt{n}}$; otherwise we can choose a sub-sequence of $\{(L_n, N_n)\}_{n=1}^{\infty}$. Again, we consider the data generating process specified in (C.1) and the same truncated least squares estimator over model class $\mathcal{T}_n(d, L_n, N_n, 1)$ as in Lemma C.2. The boundedness assumption $L_n N_n \leq M$ implies

$$\log N_{\infty}\left(\frac{1}{nc_{43} \log(n)}, T_{c_{43} \log(n)} \mathcal{T}_n, n\right) + 1 \leq \log \left(n^2c_{43} \log(n)\right) N_n L_n \log(N_n L_n) \leq \log n.$$ 

By Lemma C.2, there exists an estimator $\hat{f}_n$ such that

$$\mathbb{E}[\|\hat{f}_n - f_0\|^2] \leq \frac{(\log n)^3}{n} + \omega^2(L_n, N_n) = \frac{(\log n)^3}{n} + \frac{1}{n} \leq \frac{(\log n)^3}{n}.$$ 

Therefore, we have

$$\lim_{n \to \infty} \frac{\mathbb{E}[\|\hat{f}_n - f_0\|^2]}{n^{2d/2d+1}} = \lim_{n \to \infty} (\log n)^3 n^{-\frac{d}{2d+1}} = 0,$$

which contradicts the result in Lemma C.1. This completes the proof. 

With Proposition C.1 in hand, we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. The goal is to show that there exists some constant $c_{13} > 0$ such that

$$\omega(L, N) \geq c_{13}(N^2L^2 \log^5(NL))^{-\beta/d}.$$ 

Applying part (i) of Proposition C.1 with $\epsilon = 1$, we have

$$\lim_{NL \to \infty} \frac{\omega(L, N)}{(N^2L^2 \log^5 NL)^{-\beta/d}} = C_1 > 0.$$ 

If $C_1 = \infty$, we can choose any $C_1 > 0$ so that $\lim_{NL \to \infty} \frac{\omega(L, N)}{(N^2L^2 \log^5 NL)^{-\beta/d}} \geq C_1$. Then there exists some $M > 0$ such that

$$\frac{\omega(L, N)}{(N^2L^2 \log^5 NL)^{-\beta/d}} \geq \frac{C_1}{2}.$$
as long as $NL \geq M$. At the same time, part (ii) of Proposition C.1 shows that there exists a positive constant $C_2$ such that

$$\omega(N, L) \geq C_2 \geq C_2 \frac{(N^2 L^2 \log^5 NL)^{-\beta/d}}{\sup_{NL \leq M}(N^2 L^2 \log^5 NL)^{-\beta/d}} = \frac{C_2}{e^{2\beta/d}} (N^2 L^2 \log^5 NL)^{-\beta/d}$$

provided $e \leq NL \leq M$. Putting together the pieces, we conclude that

$$\omega(N, L) \geq \left(\frac{C_1}{2}\right) \wedge \left(\frac{C_2}{e^{2\beta/d}}\right) (N^2 L^2 \log^5 (NL))^{-\beta/d}$$

for all $N, L$ satisfying $NL \geq e$. □

C.2 Proof of Proposition 4.3

We need the following facts from probability theory to prove Proposition 4.3.

**Lemma C.3** (Concentration for Binomial distribution). Suppose $X$ is a binomial random variable with parameter $n \in \mathbb{N}^+$ and $p \in (0, 1)$. For any $\epsilon \in (0, 1)$,

$$\mathbb{P}\{(1 - \epsilon)np \leq X \leq (1 + \epsilon)np\} \geq 1 - \frac{3}{npe^2}.$$

**Proof of Lemma C.3.** By (3.5) of Feller (2008) (page 151), for any $r \geq np$, we have

$$\mathbb{P}(X \geq r) \leq \frac{r(1 - p)}{(r - np)^2}. \tag{C.8}$$

Choosing $r = (1 + \epsilon)np$, this implies

$$\mathbb{P}(X \geq (1 + \epsilon)np) \leq \frac{(1 + \epsilon)np(1 - p)}{(\epsilon np)^2} = \frac{2(1 - p)}{np} \leq \frac{1}{np} \left(1 + \frac{\epsilon}{e^2}\right).$$

Moreover, note that

$$\mathbb{P}(X \leq (1 - \epsilon)np) = \mathbb{P}(n - X \geq n - (1 - \epsilon)np).$$

Here $n - X$ is a Bernoulli random variable with parameter $n$ and $(1 - p)$. Therefore, using (C.8) with $r = n(1 - (1 - \epsilon)p) \geq n(1 - p)$ gives

$$\mathbb{P}(n - X \geq n - (1 - \epsilon)np) \leq \frac{n(1 - (1 - \epsilon)p)p}{[n(1 - (1 - \epsilon)p) - n(1 - p)]^2} = \frac{np(1 - (1 - \epsilon)p)}{(\epsilon np)^2} \leq \frac{1}{np} \frac{1}{e^2}.$$

Putting these pieces together, we have

$$\mathbb{P}((1 - \epsilon)np \leq X \leq (1 + \epsilon)np) = 1 - \mathbb{P}(X > (1 + \epsilon)np) - \mathbb{P}(X < (1 - \epsilon)np) \geq 1 - \frac{1}{npe^2} \geq 1 - \frac{3}{npe^2},$$

as claimed. □
**Lemma C.4** (Concentration for number of boxes containing balls). Suppose we throw the \(m\) balls into the \(n\) boxes one by one independently with equal probability, and let \(Z\) be the number of boxes that have at least one ball. Then we have \(\mathbb{E}[Z] = n(1 - (1 - 1/n)^m)\) and

\[
\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq \exp(-2t^2/m)
\]

**Proof of Lemma C.4.** We first calculate \(\mathbb{E}[Z]\). Let \(Z_k\) be the number of boxes that contains at least one ball after \(k\) balls have been thrown. By convention, \(Z_0 = 0\). For any \(k \geq 1\), because the balls are thrown into boxes independently with equal probability, given \(Z_k\), the next ball has a probability of \(n Z_k/n\) being thrown into an empty box, leading to \(Z_{k+1} = Z_k + 1\), and has a probability of \(Z_k/n\) being thrown into a box that already contains at least one ball, leading to \(Z_{k+1} = Z_k\). We thus have

\[
\mathbb{E}[Z_{k+1}] = \mathbb{E}[Z_k + n - Z_k] = (1 - 1/n)\mathbb{E}[Z_k] + 1. \quad \text{(C.9)}
\]

We prove by induction that \(\mathbb{E}[Z_k] = n[1 - (1 - 1/n)^k]\). For \(k = 0\), \(\mathbb{E}[Z_0] = 0 = n[1 - (1/n)^0]\). If this holds for a general \(k \geq 1\), the recursive equation (C.9) yields

\[
\mathbb{E}[Z_{k+1}] = (1 - 1/n)n[1 - (1 - 1/n)^k] + 1 = n[1 - 1/n - (1 - 1/n)^{k+1}] + 1/n = n[1 - (1 - 1/n)^{k+1}].
\]

This proves the first claim that \(\mathbb{E}[Z] = n[1 - (1 - 1/n)^m]\).

Let \(X_i \in \{1, \ldots, n\}\) denote the index of the box where the \(i\)-th ball is thrown, and write \(Z = f(X_1, \ldots, X_n)\) for some function \(f\). Note that for any \(i \in \{1, \ldots, m\}\), \(|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \leq 1\). By McDiarmid’s inequality, we have

\[
\mathbb{P}(|f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)]| \geq t) \leq \exp(-2t^2/m).
\]

This completes the proof. \(\square\)

Now we are ready to prove Proposition 4.3. The key is the neural network approximation ability stated in Theorem 4.6 that neural network is able to fit arbitrary values at “uniformly located” \(Θ(N^2L^2)\) points while staying as a constant in most areas.

**Proof of Proposition 4.3.** Step 1. **Construct Noise Distribution.** We first construct the distribution of the noise \(\epsilon\). Let \(S = (\widetilde{N}L)^{2d}\) for some \(\widetilde{N}, \tilde{L} \in \mathbb{N}^+\) to be determined, and assume \(S \leq n\). Then let \(\epsilon\) be a discrete random variable independent of \(X\), satisfying

\[
\epsilon = \begin{cases} 
(\frac{S}{n})^{1/p} & \text{with probability } \frac{S}{2n}, \\
-(\frac{S}{n})^{1/p} & \text{with probability } \frac{S}{2n}, \\
0 & \text{with probability } 1 - \frac{S}{n}.
\end{cases}
\]

It is easy to see that

\[
\mathbb{E}[\epsilon|X = x] = \mathbb{E}[\epsilon] = 0 \quad \text{and} \quad \mathbb{E}[|\epsilon|^p|X = x] = 2 \cdot \frac{S}{2n} \cdot \frac{n}{S} = 1.
\]

Moreover, the above \(\epsilon\) is symmetric, thus implying \(f_0 = f_{0,\tau}\).
**Step 2. Construct “Good” Event \( \mathcal{E} \).** Let \( X \sim \text{Uniform}([0,1]^d) \) and \( M_1 \) be the number of non-zero \( \varepsilon_i \)'s. Without loss of generality, assume \( \varepsilon_1, \ldots, \varepsilon_{M_1} \) are non-zero and \( \varepsilon_{M_1+1} = \varepsilon_{M_1+2} = \cdots = \varepsilon_n = 0 \). Then \( X_1, \ldots, X_{M_1} \) and \( X_{M_1+1}, \ldots, X_n \) are the corresponding covariate vectors. Let \( K = (\tilde{N}L)^2 \), and

\[
Q_\alpha(\Delta) = \{ x = (x_1, \ldots, x_d) : (\alpha_i - 1)/K \leq x_i \leq \alpha_i/K - \Delta \}.
\]

We also define the set of the indexes \( \alpha \) such that each hypercube \( Q_\alpha(\Delta) \) contains at least one \( X_i \) with \( i \in \{1, \ldots, M_1\} \), that is,

\[
\mathcal{A} = \left\{ \alpha \in [1, \ldots, K]^d : \exists i \in \{1, \ldots, M_1\} \text{ s.t. } X_i \in Q_\alpha(1/(n^2 K)) \right\}.
\]

Set \( M_2 = |\mathcal{A}| \).

Define the event \( \mathcal{E} = \cap_{i=1}^4 \mathcal{E}_i \), where

\[
\mathcal{E}_1 = \{0.5S \leq M_1 \leq 1.5S\},
\]

\[
\mathcal{E}_2 = \{ \forall i \in \{1, \ldots, n\}, X_i \in Q_\alpha(1/(n^2 K)) \text{ for some } \alpha \in [1, \ldots, K]^d \},
\]

\[
\mathcal{E}_3 = \{ M_2 \geq S/8 \},
\]

\[
\mathcal{E}_4 = \{ \|X_i - X_j\|_\infty \geq 1/(2n^3) \text{ for all } i \neq j \}.
\]

In the following we aim to show that \( \mathcal{E} \) occurs with high probability if \( S \) and \( n \) are sufficiently large.

We first consider event \( \mathcal{E}_1 \). Note that \( M_1 \) is a Binomial random variable with parameter \( n \) and \( S/n \). Applying Lemma C.3 with \( \epsilon = 1/2 \), we have

\[
\mathbb{P}(\mathcal{E}_1) = \mathbb{P}(0.5S \leq M_1 \leq 1.5S) \geq 1 - \frac{12}{S}.
\]

For \( \mathcal{E}_2 \), since \( X_1, \ldots, X_n \) are i.i.d. from Uniform([0,1]^d), it follows from the union bound that

\[
\mathbb{P}(\mathcal{E}_2) \geq 1 - \sum_{i=1}^n \mathbb{P}(X_i \notin Q_\alpha(1/(n^2 K)) \text{ for all } \alpha \in [1, \ldots, K]^d)
\]

\[
\geq 1 - n \mathbb{P}(X_1 \notin Q_{(1, \ldots, 1)}(1/(n^2 K)) | X_1 \in [0,1]^d)
\]

\[
\geq 1 - n \cdot \left( 1 - \left( \frac{1}{n} \right)^d \right)
\]

\[
= 1 - n \left( 1 - \left( \frac{1}{n^2} \right)^d \right)
\]

\[
\geq 1 - \frac{d}{n},
\]

where the last inequality follows from the fact that \( (1 - x)^d \geq 1 - dx \) for \( x \in [0,1] \) with \( x = 1/n^2 \).

For \( \mathcal{E}_3 \), we need to bound the probability \( \mathbb{P}(\mathcal{E}_3 | \mathcal{E}_1 \cap \mathcal{E}_2) \) from below. Conditioned on \( \mathcal{E}_1 \cap \mathcal{E}_2 \), \( X_1, X_2, \ldots, X_{M_1} \) are independent and uniformly distributed on \( \bigcup_{\alpha \in [1, \ldots, K]^d} Q_\alpha(1/(n^2 K)) \), and the probability that \( X_1 \) lies in \( Q_\alpha(1/(n^2 K)) \) is the same for each \( \alpha \). Then we can apply Lemma C.4 to provide an lower bound for \( M_2 \). To this end, note that \( X_1, \ldots, X_{M_1} \) can be viewed as \( M_1 \) balls, and \( Q_\alpha(1/(n^2 K)) \) with \( \alpha \in [1, \ldots, K]^d \) can be treated as \( K^d = S \) boxes. Hence, Lemma C.4 directly implies a concentration result for \( M_2 \). Specifically, we have

\[
\mathbb{E}[M_2] = S \{ 1 - (1 - 1/S)^{M_1} \} \geq S \{ 1 - (1 - 1/S)^{S/2} \}.
\]
Because \( \lim_{x \to \infty} (1 - 1/x)^x = 1/e \), there exists some constant \( C_1 \) such that for any \( x \geq C_1 \), \((1 - 1/x)^x - 1/e| \leq (9/16 - 1/e) \), which implies \((1 - 1/x)^x \leq 9/16 \) for \( x \geq C_1 \). Consequently,

\[
\mathbb{E}[M_2] = S \{1 - (1 - 1/S)^{\frac{t}{2}}\} \geq S \{1 - \sqrt{9/16}\} = S/4
\]
as long as \( S \geq C_1 \). This together with the tail probability in Lemma C.4 with \( t = S/8 \) yields

\[
P(M_2 \leq S/8|E_1 \cap E_2) \leq P(M_2 \leq \mathbb{E}[M_2] - S/8|E_1 \cap E_2) \leq \exp\left( -\frac{S^2}{32M_1}\right) \leq e^{-S/48},
\]
where the last inequality follows from the fact that \( M_1 \leq 1.5S \) conditioned on \( E_1 \). We thus conclude that \( P(E_3|E_1 \cap E_2) \geq 1 - e^{-S/48} \).

Turning to \( E_4 \), applying the union bound yields

\[
P(E_4) \leq \sum_{i \neq j} P[\|X_i - X_j\|_\infty \leq 1/(2n^3)]
\leq \frac{n(n-1)}{2} P[\|X_1 - X_2\|_\infty \leq 1/(2n^3)]
\leq \frac{n(n-1)}{2} \left(\frac{2}{2n^3}\right)^d \leq \frac{1}{n}.
\]

Putting together the pieces we obtain

\[
P(E) \geq P(E_1 \cap E_2 \cap E_3) + P(E_4) - 1
= P(E_3|E_1 \cap E_2)P(E_1 \cap E_2) + P(E_4) - 1
\geq (1 - e^{-S/48})(1 - d/n - 6/S) - \frac{1}{n}
\geq 1 - e^{-S/48} - \frac{d+1}{n} - \frac{6}{S}.
\]

Moreover, the following properties hold conditioned on \( E \):

1. \( S/8 \leq M_2 \leq M_1 \leq 1.5S \);
2. \( \inf_{i \neq j} \|X_i - X_j\|_\infty \geq 1/(2n^3) \).

**Step 3. Construct Neural Network \( \tilde{f}_n \).** Here we use our neural network approximation result Theorem 4.6 to construct \( \tilde{f}_n \). For each \( \alpha \in \mathcal{A} \), we can choose arbitrary \( X_i \in Q_\alpha(1/(n^2K)) \) with corresponding \( |e| = (n/S)^{1/p} \) and construct our point \((x_\alpha, y_\alpha)\) to be \( x_\alpha = X_i \) and \( y_\alpha = \text{sgn}(e_i) \), where \( \text{sgn}(x) = 1 \) if \( x > 0 \) and \( \text{sgn}(x) = -1 \) if \( x < 0 \). Let \( I_{\mathcal{A}} \) be the index set that contains all the index \( i \) we selected for all the \( \alpha \in \mathcal{A} \). Now we apply Theorem 4.6 with \( \Delta_1 = 1/(n^2K) \), \( \Delta_2 = 1/(2n^3) \) and \( u \) that we will specify later, then there exist some \( \tilde{f}_n \in \mathcal{T}_n(d, C_2 \tilde{L}^d \log_2 n, C_3 \tilde{N}^d, 1) \), or \( \tilde{f}_n \in \mathcal{T}_n(d, C_2 \tilde{L}^d, C_3 \tilde{N}^d \log_2 n) \), such that

\[
\tilde{f}_n(X_i) = \text{sgn}(e_i) \quad \text{for all } i \in I_{\mathcal{A}},
\]
and for any \( x \in Q = \bigcup_{\alpha \in \{1, \ldots, K^d\}} Q_\alpha(1/(n^2K)) \), we have

\[
\tilde{f}_n(x) = u \quad \text{if } \|x - x_\alpha\|_\infty \geq 1/(2n^3) \quad \text{for all } \alpha \in \mathcal{A}.
\]
This means that if $\mathcal{E}$ occurs,
\[
\tilde{f}_n(X_i) = u \quad \text{for all} \quad i \notin I_{\mathcal{R}}.
\]

The remaining proof of Step 3 proceeds conditioned on $\mathcal{E}$. Using a second-order Taylor expansion of $\ell_r(\cdot)$, for any $f$ we have
\[
\tilde{\mathcal{R}}_r(f) - \tilde{\mathcal{R}}_r(f_0) = \frac{1}{n} \sum_{i=1}^n (\ell_r(e_i - f(X_i)) - \ell_r(e_i)) \leq \frac{1}{n} \sum_{i=1}^n \left\{ -\psi_r(e_i) f(X_i) + \frac{1}{2} f^2(X_i) \right\}. \tag{C.11}
\]

By the definition of $M_1$ and $I_{\mathcal{R}}$, we have
\[
- \sum_{i=1}^n \psi_r(e_i) \tilde{f}_n(X_i) = \sum_{i \in I_{\mathcal{R}}} \psi_r(e_i) \tilde{f}_n(X_i) + \sum_{i \notin I_{\mathcal{R}}} \psi_r(e_i) \tilde{f}_n(X_i)
\leq -M_2 [\tau \wedge \left( n/S \right)^{1/p}] + (M_1 - M_2) u [\tau \wedge \left( n/S \right)^{1/p}] + (n - M_1) \cdot 0
\leq \left\{ - \frac{S}{8} + (2 - 1/8) S u \right\} [\tau \wedge \left( n/S \right)^{1/p}]
=(2 - 1/8) u - 1/8 S [\tau \wedge \left( n/S \right)^{1/p}].
\]

On the other hand,
\[
\sum_{i=1}^n \tilde{f}_n(X_i)^2 = \sum_{i \in I_{\mathcal{R}}} \tilde{f}_n(X_i)^2 + \sum_{i \notin I_{\mathcal{R}}} \tilde{f}_n(X_i)^2 = M_2 + (n - M_2) u \leq S + nu^2.
\]

Taking $f = \tilde{f}_n$ in (C.11), it follows from the above inequalities that
\[
\tilde{\mathcal{R}}_r(\tilde{f}_n) - \tilde{\mathcal{R}}_r(f_0) \leq \frac{S}{n} [\tau \wedge \left( n/S \right)^{1/p}] (2 - 1/8) u - 1/8 + \frac{S}{n} + u^2.
\]

Suppose $S/n \leq (4/225)^2$, because $p \geq 2$, we choose
\[
u = \sqrt{\frac{1}{32} \frac{S}{n} [\tau \wedge \left( n/S \right)^{1/p}]} \leq \sqrt{\frac{1}{32} (S/n)^{1/4}} \leq \frac{1}{30}.
\]

Under the assumption $\tau \geq c_{12} = 32$, and if $S/n \leq 32^{-p}$, it follows that
\[
\tilde{\mathcal{R}}_r(\tilde{f}_n) - \tilde{\mathcal{R}}_r(f_0) \leq \frac{S}{n} [\tau \wedge \left( n/S \right)^{1/p}] \left( \frac{15}{8} u - 1/8 \right) + \frac{S}{n} + u^2
\leq \left( \frac{1}{16} - \frac{1}{8} \right) \frac{S}{n} \left\{ [\tau \wedge \left( n/S \right)^{1/p}] - 16 \right\} + u^2
\leq \frac{1}{32} \frac{S}{n} [\tau \wedge \left( n/S \right)^{1/p}] + u^2
\leq 0.
\]

On the other hand, by the convexity of $\ell_r(\cdot)$ we have for any $f$ that
\[
\tilde{\mathcal{R}}_r(f) - \tilde{\mathcal{R}}_r(f_0) = \frac{1}{n} \sum_{i=1}^n \ell_r(e_i + f(X_i)) - \ell_r(e_i) \geq \frac{1}{n} \sum_{i=1}^n \psi_r(e_i) f(X_i).
\]

This implies
\[
\inf_{\|f\| \leq 1} \tilde{\mathcal{R}}_r(f) - \tilde{\mathcal{R}}_r(f_0) \geq -M_1/n [\tau \wedge \left( n/S \right)^{1/p}] \geq -1.5 [\tau \wedge \left( n/S \right)^{1/p}] \frac{S}{n} = -48 u^2,
\]

56
and hence

$$\hat{R}_c(\tilde{f}_n) \leq \hat{R}_c(f_0) \leq \inf_{\|f\|_{\infty} \leq 1} \hat{R}_c(f) + 48u^2.$$  

Now we are ready to provide a lower bound on $$\|\tilde{f}_n - f_0\|_2 = \|\tilde{f}_n\|_2.$$ Let $$\mu(\cdot)$$ denote the uniform measure on $$[0, 1]^d$$. By our construction of $$\tilde{f}_n$$ satisfying (C.10), it follows that

$$\mu(\{\tilde{f}_n = u\}) = \sum_{a \in [1, \ldots, K]^d} \mu(\{x : \|x - x_a\|_{\infty} \geq 1/(2n^3)\} \cap Q_a(1/(n^2K)))$$

$$= K^d \left( \frac{1}{K} - \frac{1}{n^2K} \right)^d \left( \frac{1}{n^3} \right)$$

$$\geq \left( 1 - \frac{1}{n^2} \right)^d - \frac{1}{n^{3d-1}}$$

$$\geq 1 - \frac{d}{n^2} - \frac{1}{n^{3d-1}}.$$  

Under the condition $$n \geq \sqrt{2}(d + 1)$$, it holds

$$\|\tilde{f}_n\|_2 \geq u \cdot \mu(\{\tilde{f}_n = u\}) \geq 0.5u.$$  

**Step 4. Conclude by Choosing Different S.** From the previous analysis, we have $$S = (NL)^{2d}$$ for $$N, L \in \mathbb{N}^+.$$ If $$C_1 \leq S \leq (32^{n} \wedge (4/255)^2)n = C_4n$$ for some positive constant $$C_1$$ and $$n \geq \sqrt{2}(d + 1)$$, then we can find some $$\tilde{f}_n^{(1)} \in \mathcal{F}_n(d, C_2 L \log n, C_3 N, 1)$$ and $$\tilde{f}_n^{(2)} \in \mathcal{F}_n(d, C_2 L \log n, C_3 N, 1)$$ such that conditioned on event $$E$$ with $$u = \sqrt{(\tau \wedge (n/S)^{1/p}) \cdot S/(32n)}$$,

$$\|\tilde{f}_n^{(s)}\|_2 \geq 0.5u \quad \text{and} \quad \hat{R}_c(\tilde{f}_n^{(s)}) \leq \hat{R}_c(f_0) \leq \inf_{\|f\|_{\infty} \leq 1} \hat{R}_c(f) + 48u^2$$

for $$s \in \{1, 2\}$$. Moreover, event $$E$$ occurs with probability at least $$1 - e^{-S/128} - \frac{d + 1}{n} - 6/S.$$  

Note that $$(|\mathcal{L}|^{1/d}|N|^{1/d})^{2d} \geq C_5 N^{2d} L^2$$ for some constant $$C_5 > 0$$. We prove the final conclusion by considering the following two cases, in which $$\delta_n$$ is chosen as 0.5u but with different $$S$$.

**Case I.** $$C_{14} = C_1/C_5 \leq (NL)^2 \leq C_4n.$$ In this case, let $$\tilde{N} = [N^{1/d}]$$ and $$\tilde{L} = [L^{1/d}]$$ so that

$$C_1 \leq C_5 N^{2d} L^2 \leq S = (|\mathcal{L}|^{1/d}|N|^{1/d})^{2d} \leq N^{2d} L^2 \leq C_4n.$$  

Then, there exist some $$\tilde{f}_n^{(1)} \in \mathcal{F}_n(d, C_2 L^{1/d} \log n, C_3 N^{1/d}, 1) \subseteq \mathcal{F}_n(d, C_2 L \log n, C_3 N, 1)$$, and some $$\tilde{f}_n^{(2)} \in \mathcal{F}_n(d, C_2 L^{1/d} \log n, C_3 N^{1/d} \log n, 1) \subseteq \mathcal{F}_n(d, C_2 L_3 L, C_3 N \log n, 1),$$ such that

$$\|\tilde{f}_n^{(s)}\|_2 \geq \delta_n \quad \text{and} \quad \hat{R}_c(\tilde{f}_n^{(s)}) \leq \hat{R}_c(f_0) \leq \inf_{\|f\|_{\infty} \leq 1} \hat{R}_c(f) + 192\delta_n^2$$

hold with probability at least

$$1 - \exp\left(\frac{-N^{2d} L^2/(48C_5)}{(NL)^2} \right) - \frac{6C_5}{(NL)^2} = \frac{d + 1}{n}.$$  

The prescribed $$\delta_n$$ satisfies

$$C_5 \cdot \frac{0.5 NL}{\sqrt{32} \sqrt{n}} \sqrt{\tau \wedge \left( \frac{n}{N^{2d} L^2} \right)^{1/p}} \leq \delta_n \leq \frac{0.5 NL}{\sqrt{32} \sqrt{n}} \sqrt{\tau \wedge \left( \frac{n}{N^{2d} L^2} \right)^{1/p}}.$$
Case 2. $(NL)^2 ≥ C_4n$. In this case, let $\tilde{N}$ and $\tilde{L}$ be the maximum integers satisfying $(\tilde{N}\tilde{L})^2 ≤ C_4n$, $\tilde{N}^d ≤ N$ and $\tilde{L}^d ≤ L$. Similarly, we have $(\tilde{N}\tilde{L})^2 ≥ C_6n$ for some constant $C_6 > 0$. Then with $S = (\tilde{N}\tilde{L})^2$, we have

$$\frac{1}{8\sqrt{2}} \sqrt{C_6τ ∧ C_6^{1−1/p}} ≤ δ_n = 0.5 \sqrt{\frac{1}{32} \left( \frac{τS}{n} ∧ \left( \frac{S}{n} \right)^{1−1/p} \right)} ≤ \frac{1}{8\sqrt{2}} \sqrt{C_4τ ∧ C_4^{1−1/p}}.$$ 

Because the conditions on $S$ automatically hold by our choice of $\tilde{N}$ and $\tilde{L}$, there exist $\tilde{f}_n^{(1)} ∈ F_n(d, C_4\tilde{L}^d \log_2 n, C_3\tilde{N}^d, 1) ⊆ F_n(d, C_2L^d, C_3N, 1)$ and $\tilde{f}_n^{(2)} ∈ F_n(d, C_4'\tilde{L}^d, C_3'\tilde{N}^d \log_2 n, 1) ⊆ F_n(d, C_4'\tilde{L}, C_3'\tilde{N}^d \log_2 n, 1)$ such that

$$\|\tilde{f}_n^{(s)}\|_2 ≥ δ_n \quad \text{and} \quad \tilde{R}_s(\tilde{f}_n^{(s)}) ≤ \tilde{R}_s(f_0) ≤ \inf_{\|f\|_∞ ≤ 1} \tilde{R}_s(f) + 192δ_n^2 \quad (s ∈ \{1, 2\})$$

hold with probability at least

$$1 − \exp(-C_6n/48) − \frac{1}{C_6n} − \frac{d + 1}{n}.$$ 

Finally, combining the above two cases we choose $C_7 = \min\{C_5 \cdot \frac{0.5}{\sqrt{8}}, \frac{1}{8\sqrt{2}} \sqrt{C_6τ ∧ C_6^{1−1/p}}\}$, $C_8 = \max\{\frac{1}{8\sqrt{2}}, \frac{1}{8\sqrt{2}} \sqrt{C_4τ ∧ C_4^{1−1/p}}\}$ so that $δ_n = 0.5u$ satisfies

$$C_7 \left( \frac{NL}{\sqrt{n}} \sqrt{τ ∧ \left( \frac{n}{N^2L^2} \right)^{1/p} ∧ 1} \right) ≤ δ_n ≤ C_8 \left( \frac{NL}{\sqrt{n}} \sqrt{τ ∧ \left( \frac{n}{N^2L^2} \right)^{1/p} ∧ 1} \right).$$

Moreover, setting $c_{15} = 2\max\{128/C_6, 1/C_6 + d + 1, 6/C_5 + d + 1, 128/C_5\}$, we see that

$$\mathbb{P}(E) ≥ 1 − \exp(-2((NL)^2 ∧ n)/c_{15}) − \frac{c_{15}}{2((NL)^2 ∧ n)} ≥ 1 − \frac{c_{15}}{(NL)^2 ∧ n}$$

due to the fact $e^{−x} ≤ \frac{1}{x}$. The constants $c_{16}−c_{20}$ are set to be $c_{16} = 192, c_{17} = C_2/\log 2, c_{18} = C_3, c_{19} = C_2', c_{20} = C_3'/\log 2$. □

C.3 Proof of Theorem 4.1

Before proving Theorem 4.1, we need a result stating the rate of convergence for $\tilde{f}_n$ to $f_{0,τ}$ under special cases.

Lemma C.5 (Convergence rate of $\tilde{f}_n$ to $f_{0,τ}$). Let $n, \tilde{N}, \tilde{L} ∈ \mathbb{N}^+ \setminus \{1, 2\}$, $p ≥ 2$ be arbitrary, $X$ be uniformly distributed on $[0, 1]^d$, $f_0 = 0$. Suppose the noise $ε$ is independent of $X$, and satisfies $\mathbb{E}[|ε|^p]|X = x| ≤ 1$ for all $x ∈ [0, 1]^d$. Then for any $τ$ satisfying $\log τ ≤ \log n$ and $τ ≥ 8$, we have

$$\mathbb{P}\left( \left\| \tilde{f}_n − f_{0,τ} \right\|_2 ≥ D \sqrt{\frac{τN^2L^2 \log(\tilde{N}\tilde{L}) \log n}{n}} \right) ≤ c_{48} \exp\left( - c_{49}D^2N^2L^2 \log(\tilde{N}\tilde{L}) \log n \right)$$

for any $D ≥ c_{47} ≥ 1$, where $c_{47}−c_{49}$ are constants independent of $τ, n, \tilde{N}$ and $\tilde{L}$, and $\tilde{f}_n ∈ \arg\min_{f ∈ \mathcal{F}_n(d, L, N, 1)} \tilde{R}_s(f)$. 

58
\textbf{Proof of Lemma C.5.} The proof relies on the tail probability given in Lemma A.2 with $f^* = f_{0, \tau}$. Since $X$ is independent of $\varepsilon$, we claim that there exists some constant $C_1 \in [-1/2, 1/2]$ depending only on $\tau$ such that $f_{0, \tau} = f_0 + C_1$. By the first-order condition and with a sufficiently large $\tau$, $f_{0, \tau} = \arg\min_{f \in \Theta} R(f)$ satisfies

$$E[\psi_{\tau}(\varepsilon + f_{0, \tau}(X))|X = x] = 0.$$ 

Because $\varepsilon$ and $X$ are independent, we must have $f_0 - f_{0, \tau} \equiv C_1$, where $C_1$ is such that $E[\psi_{\tau}(\varepsilon + C_1)] = 0$. By Proposition 3.2, we have $\|f_0 - f_{0, \tau}\|_2 \leq \frac{4}{\tau^{p-1}}$ so that $|C_1| = \|f_0 - f_{0, \tau}\|_2 \leq \frac{1}{2}$ if $\tau \geq 8$. We thus choose $\delta_n = \sqrt{\frac{\tau N^2 L^2 \log(N)}{n}}$, $f_n \equiv -C_1 \in \mathcal{F}_n(d, \bar{L}, \bar{N}, 1)$, and it remains to verify conditions (1)-(5) of Lemma A.2.

For condition (1), note that $f_n = f_{0, \tau}$, we have $\|f_n - f_{0, \tau}\|_2 = 0$ and $R_n(f_n) - R_n(f_{0, \tau}) = 0$.

For condition (2), similar to the proof of Proposition 3.1, by Taylor’s expansion, if $f \in \Theta$ with $M = 1$, we have

$$R_n(f) - R_n(f_{0, \tau}) = E[\psi_{\tau}(\varepsilon + C_1)\Delta_{f, \tau}(X) + \int_0^{\Delta_{f, \tau}(X)} 1[|\varepsilon + C_1 + t| \leq \tau](\Delta_{f, \tau}(X) - t)dt],$$

where $\Delta_{f, \tau}(X) = (f_{0, \tau} - f)(X)$. By the definition of $C_1$ and tower rule,

$$E[\psi_{\tau}(\varepsilon + C_1)(f - f_{0, \tau})(X)] = E[E[\psi_{\tau}(\varepsilon + C_1)|X](f - f_{0, \tau})(X)] = 0. \quad \text{(C.13)}$$

Moreover,

$$E\left[\int_0^{\Delta_{f, \tau}(X)} 1[|\varepsilon + C_1 + t| \leq \tau](\Delta_{f, \tau}(X) - t)dt\big|X = x\right]$$

$$= E\left[\int_0^{\Delta_{f, \tau}(X)} (1 - 1[|\varepsilon + C_1 + t| > \tau])(\Delta_{f, \tau}(X) - t)dt\big|X = x\right]$$

$$\geq \frac{1}{2}(\Delta_{f, \tau}(x))^2 - E\left[\int_0^{\Delta_{f, \tau}(X)} (1[|\varepsilon| > \tau/2] + 1[|C_1 + t| > \tau/2])(\Delta_{f, \tau}(X) - t)dt\big|X = x\right]$$

$$\geq \frac{1}{2}(\Delta_{f, \tau}(x))^2(1 - \mathbb{P}(|\varepsilon| \geq \tau/2|X = x)),$$

where the last inequality follows from $|C_1 + t| \leq 2.5 < 4 \leq \tau/2$. Together with the fact that

$$\mathbb{P}(|\varepsilon| > \tau/2|X = x) \leq \frac{E(|\varepsilon|^p|X = x)}{\tau^p} \leq \frac{1}{4^p} \leq \frac{1}{2},$$

this yields

$$E\left[\int_0^{\Delta_{f, \tau}(X)} 1[|\varepsilon + C_1 + t| \leq \tau](\Delta_{f, \tau}(X) - t)dt\right] \geq \frac{1}{4}\|f - f_0\|_2^2. \quad \text{(C.14)}$$

Substituting (C.13) and (C.14) into (C.12), we conclude that

$$R_n(f) - R_n(f_{0, \tau}) \geq \frac{1}{4}\|f - f_{0, \tau}\|_2^2$$

holds for all the $f \in \Theta$ with $M = 1$. This validates condition (2).
For condition (3), by letting \( \tilde{e} = e + C_1 \) it suffices to bound

\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}_n(\delta)} \left| \frac{1}{n} \sum_{i=1}^{n} h(X_i, \tilde{e}_i) - \mathbb{E} h(X, \tilde{e}) \right| \right] \tag{C.15}
\]

where \( h(X, \tilde{e}) = \ell_r(\tilde{e} + f_0, r(X) - f(X)) - \ell_r(\tilde{e}) \) satisfies

\[
h_f(X, \tilde{e}) = \ell_r(\tilde{e} + \Delta_{f,r}(X)) - \ell_r(\tilde{e}) = \psi_r(\tilde{e}) \Delta_{f,r}(X) + \int_0^{\Delta_{f,r}(X)} 1([\tilde{e} + t \leq r(\Delta_{f,r}(X) - t) dt,
\]

and \( \mathcal{H}_n(\delta) = \{ h : f \in \mathcal{T}_n \cap \Theta_n(\delta) \} \) whose envelop function \( H \) can be chosen as

\[
\sup_{f \in \mathcal{F}_n \cap \Theta_n(\delta)} h(X, \tilde{e}) \leq H(X, \epsilon) = 2|\psi_r(\tilde{e})| + 2.
\]

Similar to the proof of Lemma A.2, we have for any \( f \in \Theta \) that

\[
\mathbb{E}|h(X, \tilde{e})|^2 \leq 2\mathbb{E}|\psi_r(\tilde{e})\Delta_{f,r}(X)|^2 + 2\mathbb{E}|\frac{1}{2}|\Delta_{f,r}|^2|^2
\]

\[
\leq 2\mathbb{E}|\mathbb{E}[\psi_r(\tilde{e})^2|X]|\Delta_{f,r}(X)^2| + 2\mathbb{E}\Delta_{f,r}(X)^2
\]

\[
\leq 2(\mathbb{E}|\tilde{e}|^2 + 1)||f - f_0, r||_2^2
\]

\[
\leq 2(2v_2 + 2C_2^2 + 1)||f - f_0, r||_2^2 = C_2 ||f - f_0, r||_2^2,
\]

where \( C_2 > 0 \) is a constant. Next we use Lemma A.6 to bound (C.15). Similar to the proof of Lemma A.7, because \( \ell_r(\cdot) \) is \( \tau \)-Lipschitz, it follows from Lemma A.5 that

\[
\log \mathcal{N}_\infty(\epsilon, \mathcal{H}_n, n) \leq \log \mathcal{N}_\infty(\epsilon/\tau, \mathcal{T}_n, n) \leq C_3 \log(\epsilon/\epsilon)(\sqrt{\mathcal{N}}\log(\mathcal{N})) \log(\mathcal{N})
\]

for all \( \epsilon \in (0, n\tau) \). From previous discussions, we have \( \|H\|_2 \leq 2\tau + 2 \leq 4\tau, \|H\|_2 \leq \sqrt{8\mathbb{E}|\tilde{e}|^2 + 8} = C_4 \), and

\[
\alpha^2 = \sup_{h \in \mathcal{H}_n(\delta)} \mathbb{E}|h(X, e)|^2 \leq C_2 \sup_{f \in \mathcal{F}_n \cap \Theta_n(\delta)} ||f - f_0, r||_2 \leq C_2 \delta^2.
\]

Therefore, letting \( r = \alpha/\|H\|_2 = C_5 \delta \), we have

\[
J(r, \mathcal{H}_n, H) = \int_0^r \sup_{Q \in \mathcal{P}(n)} \sqrt{1 + \log \mathcal{N}(\epsilon/\|H\|_2, \mathcal{H}_n, \|Q\|_2, \|\cdot\|_2, \mathcal{Q}))} d\epsilon
\]

\[
\leq \int_0^{C_5 \delta} \sqrt{1 + \log \mathcal{N}_\infty(\epsilon/\|H\|_2, \mathcal{H}_n, n)} d\epsilon
\]

\[
\leq \sqrt{\mathcal{N}} \log(\sqrt{\mathcal{N}}) \int_0^{C_5 \delta} \sqrt{1 + \log \left( \frac{\epsilon \tau}{C_4} \right)} d\epsilon = \sqrt{\mathcal{N}} \log(\sqrt{\mathcal{N}}) \left( C_5 \delta + \sqrt{\log(\epsilon/\tau)} C_5 \delta + C_5 \delta \left( \sqrt{\log(1/C_5 \delta)} + 1 \right) \right).
\]

When \( \delta \geq \frac{1}{n} \) and \( \log \tau \leq \log n \), we conclude that

\[
J(r, \mathcal{H}_n, H) \leq C_7 \delta \sqrt{\mathcal{N}} \log(\sqrt{\mathcal{N}}) \log n \tag{C.16}
\]
for some constant $C_7 > 0$. By the maximal inequality in Lemma A.6,
\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}_n(\delta)} \frac{1}{n} \sum_{i=1}^{n} h(X_i, \tilde{e}_i) - \mathbb{E} h(X, \tilde{e}) \right] \leq \|H\|_2 J(r, \mathcal{H}_n, H) + \|H\|_\infty J(r, \mathcal{H}_n, H)^2 \frac{\rho^2}{\sqrt{n}} \leq C_8 \left( \delta \tilde{N} \tilde{L} \sqrt{\log(\tilde{N}\tilde{L}) \log n} + \frac{\tilde{N}^2 \tilde{L}^2 \log(\tilde{N}\tilde{L}) \log n}{\sqrt{n}} \right).
\]
This further implies that for any $\delta \geq \delta_n \geq 1/n$,
\[
\mathbb{E} \left[ \sup_{h \in \mathcal{H}_n(\delta)} \frac{1}{n} \sum_{i=1}^{n} h(X_i, \tilde{e}_i) - \mathbb{E} h(X, \tilde{e}) \right] \leq \phi_n(\delta) = C_8 \left( \delta \tilde{N} \tilde{L} \sqrt{\log(\tilde{N}\tilde{L}) \log n} + \frac{\tilde{N}^2 \tilde{L}^2 \log(\tilde{N}\tilde{L}) \log n}{n} \right).
\]
It is easy to see that $\phi_n(\alpha \delta) \leq \alpha \phi_n(\delta)$ for all $\alpha \geq 1$ because $\phi_n(\delta)$ is linear in $\delta$ and $\phi_n(0) > 0$.

For condition (4), substituting $\delta_n$ into $\phi_n$ gives $\phi_n(\delta_n) = C_8 (\delta_n^2 / \sqrt{\tau} + \delta_n^2) \leq 2C_8 \delta_n^2$.

For condition (5), it is easy to show that $\delta_n^2 \tau = (\tilde{N}\tilde{L})^2 \log(n) \log(NL) \leq n$.

Since $\tilde{f}_n$ is an empirical risk minimizer, by Lemma A.2 with $M = 1$, there exist some constants $C_9$–$C_{11}$ such that for any $D \geq C_9 = c_{47}$, it holds
\[
\mathbb{P}(\|\tilde{f}_n - f_0\|_2 \geq D \delta_n) \leq C_{10} \exp \left( - C_{11} \frac{\tilde{N} \tilde{L}^2 D^2}{1 + \tau} \right) \leq \frac{C_{10}}{c_{48}} \exp \left( - \frac{C_{11}}{2} \frac{(\tilde{N}\tilde{L})^2 \log n \log NL}{c_{49}} \right),
\]
as claimed. \hfill \Box

Now we are ready to prove Theorem 4.1.

**Proof of Theorem 4.1.** Let $\mathcal{P}(\beta, d, p)$ be the set of all regression functions $f_0 \in C(d, \beta)$ and noise distributions that have zero mean and variance bounded by 1, i.e., $\mathcal{P}(\beta, d, p) = \{(f, \varepsilon) : f \in C(d, \beta), \mathbb{E}[\varepsilon | X = x] = 0, \mathbb{E}[\varepsilon^p | X = x] \leq 1, \forall x \}$. Let $\tilde{N}, \tilde{L}, \tilde{R} \in \mathbb{N}^+$ be such that $\tilde{N}, \tilde{L} \geq \max\{c_{18}, c_{19}\} = c_{11}$ and $\tau \geq c_{12}$, where the constants $c_{12}$, $c_{18}$, $c_{19}$ are from Theorem 4.6 and Proposition 4.3. Moreover, let $S_n(\delta) = \{ \tilde{f} \in \mathcal{T}_n(d, \tilde{N}, \tilde{L}, 1) : \tilde{R}_c(f) \leq \tilde{R}_c(f_0, \tau, \tilde{N}, \tilde{L}, \bar{C}', \tilde{C}') \}$, where $\bar{C}'$ is a constant to be specified.

**Step 1. Approximation Error Lower Bound.** We claim that
\[
\sup_{(f_0, \varepsilon) \in \mathcal{P}(\beta, d, p)} \mathbb{P}(\tilde{f}_n \in S_n(\delta_1), \|\tilde{f}_n - f_0\|_2 \geq \delta_1) = 1,
\]
where $\delta_1 = C_1 (N^2 L^2 \log^4(NL))^{-\beta/d}$ for some constant $C_1 > 0$. This is a direct consequence of Theorem 4.2 by noting that
\[
\|\tilde{f}_n - f_0\|_2 \geq \inf_{f \in \mathcal{T}_n(d, \tilde{N}, \tilde{L}, 1)} \|f - f_0\|_2.
\]

**Step 2. Stochastic Error Lower Bound.** We claim that if $\tilde{N}\tilde{L} \geq (C_2 \log n)^2$ with $C_2 = \max\{c_{17}, c_{18}, c_{19}, c_{20}\}$, $c_{14}$, it holds
\[
\sup_{(f_0, \varepsilon) \in \mathcal{P}(\beta, d, p)} \mathbb{P}(\exists \tilde{f}_n \in S_n(\delta_2), \|\tilde{f}_n - f_0\|_2 \geq \delta_2) \geq 1 - \frac{C_3}{(\log n)^2} \quad \text{for all } n \geq \sqrt{2(d + 1)} \quad \text{(C.17)}
\]
and some constant $C_3 > 0$, where $\delta_2$ satisfies

$$\delta_2 \asymp \frac{\tilde{N}L}{\sqrt{n} \log n} \sqrt{\tau \wedge \left( \frac{n(\log n)^2}{N^2 L^2} \right)^{1/p}} \wedge 1.$$ 

Because $\tilde{N}L \geq (C_2 \log n)^2$ implies $\max(\tilde{L}, \tilde{N}) \geq C_2 \log n$, we can divide the discussion into two cases. We first consider the case where $\tilde{L} \geq C_2 \log n$. Let $L = [\tilde{L}/(c_{17} \log n)]$ and $N = [\tilde{N}/c_{18}]$. Combined with the fact that $\tilde{N} \geq c_{18}$, this implies $L, N \in \mathbb{N}^+$, $L \geq c_{14}$ and

$$c_{14} \leq NL = \left[ \frac{\tilde{L}}{c_{17} \log_2 n} \right][\tilde{N}/c_{18}] \leq \frac{\tilde{N}L}{\log n}.$$ 

Proposition 4.3 implies that there exists some pair $(f_0, \varepsilon) \in \mathcal{P}(d, \beta, p)$ with symmetric noise $\varepsilon$ such that there exists some $\tilde{f}_n \in \mathcal{F}_n(d, c_{17} L \log n, c_{18} N, 1) \subseteq \mathcal{F}_n(d, \tilde{L}, \tilde{N}, 1)$ satisfying

$$\|\tilde{f}_n - f_0\|_2 \geq \delta_2 \quad \text{and} \quad \mathcal{R}_r(\tilde{f}_n) \leq \min \left\{ \mathcal{R}_r(f_0, \tau), \inf_{\|f\|_1 \leq 1} \mathcal{R}_r(f) + c_{16} \delta_2^2 \right\}$$ 

with probability at least

$$1 - \frac{c_{15}}{(NL)^{2 \wedge n}} \geq 1 - \frac{C_5}{(\log n)^2}$$

for some constant $C_5 > 0$, and $\delta_2$ is defined as

$$\delta_2 \asymp \frac{NL}{\sqrt{n}} \sqrt{\tau \wedge \left( \frac{n}{N^2 L^2} \right)^{1/p}} \wedge \frac{\tilde{N}L}{\sqrt{n} \log n} \sqrt{\tau \wedge \left( \frac{n(\log n)^2}{N^2 L^2} \right)^{1/p}} \wedge 1 = \left( \frac{C_2 \log n}{L} \right)^2.$$ 

For the case $\tilde{N} \geq C_2 \log n$, we follow a similar argument by letting $L = [\tilde{L}/c_{19}]$ and $N = [\tilde{N}/(c_{20} \log n)]$. This time, we also have $NL = \tilde{N}L/\log n$. By Proposition 4.3, there exists some pair $(f_0, \varepsilon) \in \mathcal{P}(d, \beta, p)$ such that with probability at least $1 - \frac{C_4}{(\log n)^2}$, there exists some $\tilde{f}_n \in \mathcal{F}_n(d, c_{19} L, c_{20} N \log n, 1) \subseteq \mathcal{F}_n(d, \tilde{L}, \tilde{N}, 1)$ satisfying (C.18) with $\delta_2$ satisfying (C.19). Therefore, the claim (C.17) follows immediately by taking $C_3 = C_5 \vee C_6$.

**Step 3. Bias Lower Bound.** In this part, we assert that if $(\tilde{N}L)^2 \leq n$ and $\tau$ satisfies

$$\frac{C_7}{\tau^{2p-1}} \geq \frac{\tilde{N}^2 L^2}{n} (\log n)^2$$

for some constant $C_7 \in (0, 1)$, then it follows

$$\sup_{(f_0, \varepsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P} \left( \forall \tilde{f}_n \in S_n(0), \|\tilde{f}_n - f_0\|_2 \geq \delta_3 \right) \geq 1 - C_8 \exp(-C_9 \log n),$$

where $\delta_3 = \sqrt{C_7 \tau^{1-p}}$, and $C_8, C_9$ are some positive constants.

Consider the null case where $f_0 = 0$, and let the noise $\varepsilon$ be a discrete random variable independent of $X$, satisfying

$$\varepsilon = \begin{cases} -1 & \text{with probability } \theta \\ 2\tau & \text{with probability } \frac{\theta}{2\tau} \\ 0 & \text{with probability } 1 - \theta(1 + \frac{1}{2\tau}) \end{cases}.$$
where \( \theta = 2^{-p} r^{1-p} \). It is easy to show that for any \( \tau \geq 1 \) given,
\[
\mathbb{E}[\epsilon] = 0 \quad \text{and} \quad \mathbb{E}[|\epsilon|^2] = \theta(1 + (2\tau)^{p-1}) \leq \theta 2^p r^{p-1} \leq 1.
\]
Moreover, because \( X \) and \( \epsilon \) are independent, we have \( f_0(x) - f_0,\tau(x) = \Delta, \) where \( \Delta \) only depends on \( \tau \) and satisfies \( \mathbb{E}[\psi, (\epsilon + \Delta)] = 0 \). This implies
\[
(-1 + \Delta)\theta + (0 + \Delta)(1 - \theta(1 + 1/(2\tau))) = 0.
\]
Since \( \tau \geq 1 > \theta \), we further have
\[
\Delta = \frac{\theta}{1 - \theta/(2\tau)} \geq \theta = \frac{1}{2^p r^{p-1}}.
\]
It follows from Lemma C.5 with \( D = c_{47} \geq 1 \) that any empirical risk minimizer \( \tilde{f}_n \) satisfies
\[
\mathbb{P}\left( \|f_n - f_0,\tau\|_2 \leq c_{47} \sqrt{\frac{(N\mathcal{L} \log n)^2}{n}} \right) \geq 1 - c_{48} \exp\left(-c_{49} c_{47}^2 N^2 \mathcal{L}^2 \log n \log(N\mathcal{L})\right)
\]
\[
= 1 - C_8 \exp(-C_9 \log n),
\]
where \( C_8 = c_{48} \) and \( C_9 = c_{49} c_{47}^2 \). By the triangle inequality,
\[
\|f_0,\tau - f_0\|_2 \leq \|f_0,\tau - \tilde{f}_n\|_2 + \|	ilde{f}_n - f_0\|_2,
\]
so that conditioned on the event \( \|\tilde{f}_n - f_0,\tau\|_2 \leq c_{47} \sqrt{\frac{(N\mathcal{L} \log n)^2}{n}} \), the following inequality holds
\[
\|	ilde{f}_n - f_0\|_2 \geq \|f_0,\tau - f_0\|_2 - \|	ilde{f}_n - f_0,\tau\|_2
\]
\[
\geq \frac{1}{2^p r^{p-1}} - c_{47} \sqrt{\frac{(N\mathcal{L} \log n)^2}{n}}.
\]
Setting \( c_7 = 2^{-2p-2} c_{47}^{-2} \), we have
\[
\|	ilde{f}_n - f_0\|_2 \geq \frac{1}{2^p r^{p-1}} - \sqrt{\frac{\tau}{2^2 p^2 r^{2p-1}} = \frac{1}{2^{p+1} r^{p-1}} = \frac{\sqrt{c_7}}{\tau^{p-1}} \geq \frac{\sqrt{c_7}}{\tau^{p-1}} = \delta_3,}
\]
with probability at least \( 1 - C_8 \exp(-C_9 \log n) \). This proves the claim.

**Step 4. Combining the Separate Lower Bounds.** For any \( \epsilon > 0 \), from Step 1 and Step 2 we have
\[
\sup_{(f_0,\epsilon) \in \mathcal{P}(d,\beta,p)} \mathbb{P}\left( \forall \tilde{f}_n \in \mathcal{S}_n(0), \|\tilde{f}_n - f_0\|_2 \geq \delta_1 \right) = 1 \quad (C.20)
\]
and
\[
\sup_{(f_0,\epsilon) \in \mathcal{P}(d,\beta,p)} \mathbb{P}\left( \exists \tilde{f}_n \in \mathcal{S}_n(\delta_2), \|\tilde{f}_n - f_0\|_2 \geq \delta_2 \right) \geq 1 - \epsilon
\]
for all \( \bar{\mathcal{L}} \mathcal{N} \geq (C_2 \log n)^2 \) and \( n \geq n_{(1)}^\epsilon := \lceil \sqrt{2(d+1)} \vee \exp(\sqrt{c_5}/\epsilon) \rceil \).

We first combine the results in Step 2 and Step 3. We argue that if \( \bar{\mathcal{L}} \mathcal{N} \geq (C_2 \log n)^2 \), then
\[
\sup_{(f_0,\epsilon) \in \mathcal{P}(d,\beta,p)} \mathbb{P}\left( \exists \tilde{f}_n \in \mathcal{S}_n(\delta_{2,3}), \|\tilde{f}_n - f_0\|_2 \geq \delta_{2,3} \right) \geq 1 - \epsilon \quad \text{for} \quad \delta_{2,3} = \delta_2 \vee \frac{C_{10}}{\tau^{p-1} \log^2 n} \quad (C.21)
\]
holds for all \( n \geq n_e^{(1)} \lor n_e^{(2)} \), where \( C_{10} \) and \( n_e^{(2)} \) are some constants. Moreover, if \( \bar{N}L \leq (C_2 \log n)^2 \), then (C.21) also holds for all \( n \geq n_e^{(1)} \lor n_e^{(2)} \) with \( \delta_{2,3} > \delta_2 \).

By our construction of \( \delta_2 \), there exists a constant \( C_{11} > 1 \) such that

\[
\frac{\delta_2}{C_{11}} \leq \frac{\bar{N}L}{\sqrt{n} \log n} \sqrt{\tau \wedge \left( \frac{n(\log n)^2}{N^2 L^2} \right)^{1/p}} \wedge 1 \leq C_{11} \delta_2.
\]

The remaining of the proof will be divided into three cases regarding the relationship between \( \bar{N}L, n \) and \( \tau \).

**Case 1.** \( \bar{N}L \leq \sqrt{n} \) and \( \tau \) satisfies \( \frac{C_2}{\tau^{p-1}} \geq \frac{N_{12}}{n} (\log n)^2 \). The latter further yields

\[
\frac{\sqrt{C_7}}{\tau^{p-1} (\log n)^2} \geq \frac{\sqrt{C_7 N_{11}}}{\tau^{p-1} (\log n)^2} \geq \frac{\delta_2}{C_{11}} C_{11} = \delta_2
\]

with probability at least \( 1 - \epsilon \) if \( n \geq n_e^{(2)} \), where \( n_e^{(2)} \) is such that \( C_8 \exp(-C_9 \log n_e^{(2)}) \leq \epsilon \) and \( \log(n_e^{(2)}) \geq \sqrt{C_7} \). We thus conclude that if \( n \geq n_e^{(2)} \),

\[
P\left( \forall \hat{f}_n \in S_n(0), \| \hat{f}_n - f_0 \|_2 \geq \frac{\sqrt{C_7 C_{11}}}{\tau^{p-1} (\log n)^2} \wedge \delta_2 \right) \geq 1 - \epsilon.
\]

**Case 2.** \( \bar{N}L \leq \sqrt{n} \) and \( \tau \) satisfies \( \frac{C_2}{\tau^{p-1}} \leq \frac{N_{12}}{n} (\log n)^2 \). We claim that \( \delta_2 \vee \frac{\sqrt{C_7}}{\tau^{p-1} (\log n)^2 C_{11}} = \delta_2 \). To see this, following the same argument as in Case 1 we obtain

\[
\frac{\sqrt{C_7}}{\tau^{p-1} (\log n)^2} \leq \frac{\bar{N}L}{\sqrt{n} \log n} \sqrt{\tau} \wedge 1
\]

which means \( \frac{\sqrt{C_7}}{\tau^{p-1} (\log n)^2} \leq \delta_2 \) if \( \tau \leq \frac{\delta_{2,3}^1}{n(\log n)^2} \). It suffices to show that the same inequality also holds if \( \tau \geq \left( \frac{n(\log n)^2}{(N^2 L)^2} \right)^{1/p} \). If so, it is easy to see that

\[
\frac{1}{\tau^{p-1}} \leq \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1-1/p}.
\]

Multiplying both sides by \( \sqrt{C_7} / (\log n)^2 \) gives

\[
\frac{\sqrt{C_7}}{\tau^{p-1} (\log n)^2} \leq \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1-1/p} \frac{\sqrt{C_7}}{(\log n)^2}
\]

\[
\leq \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1/2(1-1/p)} \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1/2(1-1/p)} \frac{\sqrt{C_7}}{(\log n)^2}
\]

\[
\leq \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1/2(1-1/p)} \frac{1}{\log n} \frac{1}{(\log n)^2} \frac{\sqrt{C_7}}{(\log n)^2}
\]

\[
\leq \left( \frac{N^2 L^2}{n(\log n)^2} \right)^{1/2(1-1/p)}.
\]
by noting that $C_7 < 1$, and $\log n \geq 1$ if $n \geq \epsilon$. We thus conclude that
\[
\frac{\sqrt{C_7}}{\tau^{p-1}(\log n)^2} \leq \sqrt{\frac{\bar{N}^2 \bar{L}^2}{n(\log n)^2} \left( \tau \wedge \left( \frac{n(\log n)^2}{\bar{N}^2 \bar{L}^2} \right)^{1/p} \right)} \leq C_{11} \delta_2,
\]
which validates our claim. Therefore, by using the result in Step 2, we obtain that
\[
\sup_{(f_0, \epsilon) \in P(\beta, d, p)} P \left( \exists \hat{f}_n \in S_n(\delta_2), \|\hat{f}_n - f_0\|_2 \geq \delta_2 \vee \frac{\sqrt{C_7}}{\tau^{p-1}(\log n)^2 C_{11}} \right) \geq 1 - \epsilon
\]
for all $n \geq n^{(1)}_\epsilon$.

Case 3. $\bar{N} \bar{L} \geq \sqrt{n}$. In this case, by assumption we have $\delta_2 \geq 1/(C_{12} \log n)$ for some constant $C_{12} > 0$, then
\[
\delta_2 \geq \frac{1}{C_{12} \log n} \geq \frac{1}{C_{12} \tau^{p-1}(\log n)^2}
\]
if $\tau \geq 1$ and $n \geq \epsilon$. From the result in Step 2 we see that
\[
\sup_{(f_0, \epsilon) \in P(\beta, d, p)} P \left( \exists \hat{f}_n \in S_n(\delta_2), \|\hat{f}_n - f_0\|_2 \geq \delta_2 \vee \frac{1}{\tau^{p-1}(\log n)^2 C_{12}} \right) \geq 1 - \epsilon
\]
for all $n \geq n^{(1)}_\epsilon$.

Combining the above three cases, we choose $C_{10} = \sqrt{C_7} C_{11} \wedge \sqrt{C_{12}} \wedge \frac{1}{C_{12}}$. Because $\delta_2 \geq \frac{1}{\tau^{p-1}(\log n)^2}$ in Cases 2 and 3, (C.21) holds for all $n \geq n^{(1)}_\epsilon \vee n^{(2)}_\epsilon$ and $\bar{N} \bar{L} \geq (C_2 \log n)^2$. On the other hand, if $\bar{N} \bar{L} \leq (C_2 \log n)^2$ or $\delta_{2,3} > \delta_2$, then we must have Case 1 in which the result of Step 2 is not involved. Thus (C.21) also holds for all $n \geq n^{(1)}_\epsilon \vee n^{(2)}_\epsilon$.

Finally, we combine (C.20) and (C.21) to prove the main statement. Specifically, we claim that for all $n \geq n^{(1)}_\epsilon \vee n^{(2)}_\epsilon \vee n^{(3)}_\epsilon$, where $n^{(3)}_\epsilon$ is some constant, we have
\[
\sup_{(f_0, \epsilon) \in P(\beta, d, p)} P \left( \exists \hat{f}_n \in S_n(\delta^*), \|\hat{f}_n - f_0\|_2 \geq \delta^* \right) \geq 1 - \epsilon \quad \text{with } \delta^* = \delta_1 \vee \delta_{2,3}. \tag{C.22}
\]
To see this, we first consider the case where $\bar{N} \bar{L} \geq (C_2 \log n)^2$ or $\delta_{2,3} > \delta_2$. In this case, if $\delta_1 \geq \delta_{2,3}$, then for all $\hat{f}_n \in S_n(0) \subseteq S_n(\delta^*)$, $\|\hat{f}_n - f_0\|_2 \geq \delta_1 = \delta^*$; if $\delta_1 \leq \delta_{2,3}$, (C.21) ensures that with probability at least $1 - \epsilon$, there exists some $\hat{f}_n \in S_n(\delta_{2,3}) = S_n(\delta^*)$ such that $\|\hat{f}_n - f_0\|_2 \geq \delta_{2,3} = \delta^*$. Hence (C.22) holds in this case.

It remains to prove (C.22) if $\bar{N} \bar{L} \leq (C_2 \log n)^2$ and $\delta_{2,3} = \delta_2$. In this case, we have
\[
\frac{\delta_1}{\delta_{2,3}} = \frac{\delta_1}{\delta_2} \geq \frac{(\bar{N}^2 \bar{L}^2 \log(\bar{N} \bar{L}))^{\beta/d}}{\sqrt{(\bar{N} \bar{L})^{(1-1/p)}}} = n^{\frac{1}{2}(1-1/p)}(\bar{N} \bar{L})^{\frac{\beta}{2}-\frac{1}{2}(1-\frac{1}{p})} \log(\bar{N} \bar{L})^{-\beta/d} \geq n^{\frac{1}{2}(1-1/p)(\log n)^{C_{13}}}.
\]
Then we can choose a sufficiently large $n^{(3)}_\epsilon$ such that $\delta_1 > \delta_{2,3}$ for all $n \geq n^{(3)}_\epsilon$. Consequently, for all $\hat{f}_n \in S_n(0) \subseteq S_n(\delta^*)$, the inequality $\|\hat{f}_n - f_0\|_2 \geq \delta_1 = \delta^*$ holds. This proves the claim (C.22), and hence completes the proof of Theorem 4.1. □
C.4 Proof for Corollary 4.4

Proof of Corollary 4.4. The proof is almost identical to that of Theorem 4.1 except we choose \( \tau = \infty \). Let \( \mathcal{P}(\beta, d, p) \) and \( S_n(\delta) \) be the same as in the proof of Theorem 4.1 but with \( \tau = \infty \). Moreover, let \( \bar{N}, \bar{L} \in \mathbb{N}^+ \) be such that \( \bar{N}, \bar{L} \geq \max(c_{18}, c_{19}) = c_{11} \).

**Step 1. Approximation Error Lower Bound.** The part is identical to Step 1 in the proof of Theorem 4.1. We argue that

\[
\sup_{(f_0, \epsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P}\left( \forall f_n \in S_n(0), \|f_n - f_0\|_2 \geq \delta_1 \right) = 1,
\]

where \( \delta_1 = C_1(N^2L^2 \log^5(\bar{N}L))^{-\beta/d} \) for some constant \( C_1 > 0 \).

**Step 2. Stochastic Error Lower Bound.** We claim that if \( \bar{N}L \geq (C_2 \log n)^2 \) with \( C_2 = \max\{c_{17}, c_{18}, c_{19}, c_{20}\} \), it holds

\[
\sup_{(f_0, \epsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P}\left( \exists \tilde{f}_n \in S_n(\delta), \|\tilde{f}_n - f_0\|_2 \geq \delta_2 \right) \geq 1 - \frac{C_3}{(\log n)^2} \text{ for all } n \geq \sqrt{2(d + 1)} \quad (C.23)
\]

for some constant \( C_3 > 0 \), where \( \delta_2 \) satisfies

\[
\delta_2 = \left( \frac{N^2L^2}{n(\log n)^2} \right)^{\frac{1}{\beta} + \frac{1}{2}} \wedge 1.
\]

The proof of this is the same as before (when \( \tau < \infty \)), and thus is omitted.

**Step 3. Combining Two Parts Together.** Let \( \epsilon \in (0, 1) \). From Step 1 and Step 2, we have

\[
\sup_{(f_0, \epsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P}\left( \forall f_n \in S_n(0), \|f_n - f_0\|_2 \geq \delta_1 \right) = 1
\]

and

\[
\sup_{(f_0, \epsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P}\left( \exists \tilde{f}_n \in S_n(\delta), \|\tilde{f}_n - f_0\|_2 \geq \delta_2 \right) \geq 1 - \epsilon
\]

for all \( \bar{N}L \geq (C_2 \log n)^2 \) and \( n \geq n(1) := \lceil \sqrt{2(d + 1)} \vee \exp(C_3/\epsilon) \rceil \). We aim to show that for all \( n \geq n(1) \wedge n(2), \) where \( n(2) \) is a constant, it holds

\[
\sup_{(f_0, \epsilon) \in \mathcal{P}(d, \beta, p)} \mathbb{P}\left( \exists \tilde{f}_n \in S_n(\delta^*), \|\tilde{f}_n - f_0\|_2 \geq \delta^* \right) \geq 1 - \epsilon \quad \text{with} \quad \delta^* = \delta_1 \vee \delta_2. \quad (C.24)
\]

Similarly, we first consider the case where \( (\bar{N}L) \geq (C_2 \log n)^2 \). In this case, if \( \delta_1 \geq \delta_2 \), then for all \( \tilde{f}_n \in S_n(0) \subseteq S_n(\delta^*) \), Step 1 implies \( \|\tilde{f}_n - f_0\|_2 \geq \delta_1 = \delta^* \); if \( \delta_1 \leq \delta_2 \), the result of Step 2 implies that with probability at least \( 1 - \epsilon \), there exists some \( \tilde{f}_n \in S_n(\delta) = S_n(\delta^*) \) such that \( \|\tilde{f}_n - f_0\|_2 \geq \delta_2 = \delta^* \). So (C.24) holds if \( (\bar{N}L) \geq (C_2 \log n)^2 \).

It suffices to prove (C.24) if \( (\bar{N}L) \leq (C_2 \log n)^2 \) and \( n \geq n(2) \). Note that

\[
\frac{\delta_1}{\delta_2} \geq \frac{(n(\log n)^2)^{\frac{1}{\beta} + \frac{1}{2}}}{\bar{N}L^{1 - \frac{1}{\beta} + 2\beta/d} \log(\bar{N}L)^{3\beta/d}} \geq n^{\frac{1}{\beta} + \frac{1}{2}} (\log n)^{-C_4}
\]

for some constant \( C_4 > 0 \). Then \( \delta_1 \geq \delta_2 \) if \( n \geq n(2) \) for some \( n(2) \) large enough. In this case, for all \( \tilde{f}_n \in S_n(0) \subseteq S_n(\delta^*) \), the lower bound \( \|\tilde{f}_n - f_0\|_2 \geq \delta_1 = \delta^* \) holds. This certifies the claim (C.24) also in this case. The proof of the corollary is completed. \( \square \)