

Quantile regression with ReLU Networks: Estimators and minimax rates

Oscar Hernan Madrid Padilla¹, Wesley Tansey², and Yanzhen Chen³

¹Department of Statistics, University of California, Los Angeles

²Department of Epidemiology and Biostatistics, Memorial Sloan Kettering

³Department of ISOM, Hong Kong University of Science and Technology

October 28, 2020

Abstract

Quantile regression is the task of estimating a specified percentile response, such as the median (50^{th} percentile), from a collection of known covariates. We study quantile regression with rectified linear unit (ReLU) neural networks as the chosen model class. We derive an upper bound on the expected mean squared error of a ReLU network used to estimate any quantile conditional on a set of covariates. This upper bound only depends on the best possible approximation error, the number of layers in the network, and the number of nodes per layer. We further show upper bounds that are tight for two large classes of functions: compositions of Hölder functions and members of a Besov space. These tight bounds imply ReLU networks with quantile regression achieve minimax rates for broad collections of function types. Unlike existing work, the theoretical results hold under minimal assumptions and apply to general error distributions, including heavy-tailed distributions. Empirical simulations on a suite of synthetic response functions demonstrate the theoretical results translate to practical implementations of ReLU networks. Overall, the theoretical and empirical results provide insight into the strong performance of ReLU neural networks for quantile regression across a broad range of function classes and error distributions. All code for this paper is publicly available at <https://github.com/tansey/quantile-regression>.

Keywords: Deep networks, robust regression, minimax, sparse networks.

1 Introduction

The standard task in regression is to predict the mean response of some variable Y , conditioned on a set of known covariates X . Typically, this is done by minimizing the mean squared error,

$$\hat{f}^{(\text{mse})} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2, \quad (1)$$

where \mathcal{F} is a function class.

In many scenarios, this may not be the desired estimand. For instance, if the data contain outliers or the noise distribution of Y is heavy-tailed, eq. (1) will be an unstable. The median is then often a more prudent quantity to estimate, even under a squared error risk metric. Alternatively, fields such as quantitative finance and precision medicine are often concerned with extremal risk as well as expected risk. In these domains,

one may wish to estimate tail events such as the 5th or 95th percentile outcome. Estimating the 5th, 50th (median), 95th, or any other response percentile conditional on covariates is the task of quantile regression.

The goal of quantile regression is to estimate a *quantile function*. Formally, given independent measurements $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, from the random vector $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, the goal is to estimate $f_\tau^* : \mathbb{R}^d \rightarrow \mathbb{R}$ given as

$$f_\tau^*(x) = F_{Y|X=x}^{-1}(\tau), \quad x \in \mathbb{R}^d,$$

where $\tau \in (0, 1)$ is a quantile level, $F_{Y|X=x}$ is the distribution of Y conditioned on $X = x$, and $f_\tau^*(\cdot)$ is the quantile function for the τ th quantile. For example, when $\tau = 0.5$ the function $f_\tau^*(x)$ becomes the conditional median of Y given $X = x$. More generally, the quantile level τ corresponds to the $(\tau \times 100)^{\text{th}}$ percentile response.

As an estimator for f_τ^* , we consider \hat{f} of the form

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \rho_\tau(y_i - f(x_i)), \quad (2)$$

where \mathcal{F} is a class of neural network models, and $\rho_\tau(x) = \max\{\tau x, (\tau - 1)x\}$ is the quantile loss function as in [Koenker and Bassett Jr \(1978\)](#). Neural network models optimizing eq. (2) have been proposed in previous contexts (see Section 1.2). These models have shown strong empirical performance, but a theoretical understanding of neural quantile regression remains absent. In this paper, we lay the groundwork for the theoretical foundations of quantile regression with neural networks. Below we briefly summarize our contributions.

1.1 Summary of results

We establish statistical guarantees for quantile regression with multilayer neural networks built with the rectified linear unit (ReLU) activation function ($\phi(x) = \max\{x, 0\}$). Specifically, we make the following contributions:

- For the class of ReLU neural networks \mathcal{F} with W parameters, U nodes, and L layers, we provide an upper bound on the expected value of the mean squared error for estimating the quantile function $f_\tau^*(\cdot)$ at the design points x_1, \dots, x_n . The upper bound requires no assumptions about the function $f_\tau^*(\cdot)$ though it depends on the best performance possible, under the quantile loss, for functions in the class \mathcal{F} .
- Suppose that $f_\tau^*(\cdot)$ can be written as the composition of functions whose coordinates are Hölder functions with potentially different smoothness levels (see [Schmidt-Hieber \(2017\)](#)). We show that there exists a sparse ReLU neural network class \mathcal{F} such that the corresponding quantile regression estimator attains minimax rates, under squared error loss, for estimating $f_\tau^*(\cdot)$. This result holds under minimal assumptions on the distribution of (X, Y) . As a result, quantile regression with ReLU networks can be directly applied to various classes of heavy-tailed error distributions.
- Suppose that $f_\tau^*(\cdot)$ belongs to a Besov space $B_{p,q}^s([0, 1]^d)$, where $0 < p, q \leq \infty$, and $s > d/p$. We show that under mild conditions, there exists a ReLU neural network structure \mathcal{F} such that quantile regression constrained to \mathcal{F} attains the rate $n^{-\frac{2s}{2s+d}}$ under the squared error loss. The resulting rate is minimax for the class $B_{p,q}^s([0, 1]^d)$. Thus, our work advances the nonparametric regression results from [Suzuki \(2018\)](#) to the quantile regression setting where the distribution of the errors can be arbitrary distributions.

1.2 Previous work

This paper lies at the intersection of nonparametric function estimation theory and quantile regression. The theory we develop draws on two well-established, though mostly-independent, lines of research in estimation theory: (i) universality and convergence rates for neural networks, and (ii) minimax rates for estimating functions in a Besov and a space based on compositions of Hölder functions. We merge and extend results from these fields to analyze neural quantile regression. This provides a theoretical foundation for a number of proposed neural quantile methods with strong empirical performance but no prior theoretical motivation. We briefly outline relevant work in each of these areas and situate this paper within these different lines of work.

Neural networks have been shown to have attractive theoretical properties in many scenarios. Hornik et al. (1989) showed that regardless of the activation function, single-layer feedforward networks can approximate any measurable function; more thorough descriptions of approximation theory for neural networks are given in White (1989); Barron (1993, 1994); Hornik et al. (1994); Anthony and Bartlett (2009). In the statistical theory literature, McCaffrey and Gallant (1994) proved convergence rates for single-layer feedforward networks. Kohler and Krzyżak (2005) proved convergence rates for estimating a regression function with a shallow network using sigmoid activation functions; Hamers and Kohler (2006) developed risk bounds in a similar framework. Klusowski and Barron (2016a) developed risk bounds for high-dimensional ridge function combinations that include neural networks. Klusowski and Barron (2016b) studied uniform approximations by neural network models.

More recently, an emerging line of research explores approximation theory for ReLU neural networks. This work is motivated by the empirical successes of ReLU networks, which often outperform neural networks with other activation functions (e.g. Nair and Hinton, 2010; Glorot et al., 2011) and have achieved state-of-the-art performance in a number of domains (e.g. Krizhevsky et al., 2012; Devlin et al., 2019). Yarotsky (2017) provided approximation results for Sobolev spaces, which was exploited by Farrell et al. (2018) for semiparametric inference in causality related problems. Liang and Srikant (2016) and Petersen and Voigtlaender (2018) provided approximation results for piecewise smooth functions. Additional approximation results were also established in Schmidt-Hieber (2017) for classes of functions constructed from Hölder functions. For such classes, the corresponding approximation results were exploited by Schmidt-Hieber (2017) to obtain minimax rates for nonparametric regression with ReLU networks. More recently, Nakada and Imaizumi (2020) proved minimax rates for nonparametric estimation with ReLU networks in settings with intrinsic low dimension of the data. Bauer and Kohler (2019) studied theoretical properties of nonparametric regression with neural networks with sigmoid activation function.

Separate from neural network theory, related work has investigated regression when the true function is in a Besov space. Such classes of functions are widely used in statistical modeling due to their ability to capture spatial inhomogeneity in smoothness. In addition, Besov spaces include more traditional smoothness spaces such as Hölder and Sobolev spaces. Some statistical works involving Besov spaces include Donoho et al. (1998) and Suzuki (2018) which provided minimax results on Besov spaces in the context of regression based on wavelet and neural network estimators respectively. Brown et al. (2008) studied the one dimensional median regression setting when the median function belongs to a Besov space. Uppal et al. (2019) considered the context of density estimation and convergence of generative adversarial networks. A more mathematically generic treatment of Besov spaces can be found in DeVore and Popov (1988) and Lindenstrauss and Tzafriri (2013).

In a line of empirical work, quantile regression with neural networks has been shown to be a powerful nonparametric tool for modeling complex data sets. Successful applications of neural quantile regression include precipitation downscaling and wind power (Cannon, 2011; Hatalis et al., 2017), credit portfolio analysis (Feng et al., 2010), value at risk (Xu et al., 2016), financial returns (Taylor, 2000; Zhang et al., 2019), electrical industry forecasts (Zhang et al., 2018), and transportation problems (Rodrigues and Pereira,

2020).

On the theoretical side of quantile regression with neural networks, White (1992) proved convergence in probability results for shallow networks. Chen and White (1999) developed theory for estimation with a general loss and with single hidden layer neural network architectures based on a smooth activation function. For target functions in the Barron class (Barron, 1993, 1994; Hornik et al., 1994), Chen and White (1999) proved convergence rates better than $n^{-1/4}$ rate in root-mean-square error metric for time series nonparametric quantile regression. Similarly, Example 3.2.2 in Chen (2007) also established a faster than $n^{-1/4}$ rate in root-mean-square error metric for nonparametric quantile regression in the Sobolev space $W_1^1([0, 1]^d)$ (ℓ_1 -integrable functions with domain $[0, 1]^d$ and ℓ_1 -integrable first order partial derivatives). In a related work, Chen et al. (2020) considered quantile treatment effect estimation. Despite all these notable efforts, the results for quantile regression with neural networks are not known to be minimax optimal. We fill this gap by considering quantile regression with deep ReLU neural network architectures, showing minimax rates for general classes of functions.

2 Neural quantile regression with ReLU networks

2.1 Univariate response quantile regression

For a vector $v \in \mathbb{R}^r$ we define the function $\phi_v : \mathbb{R}^r \rightarrow \mathbb{R}^r$ as

$$\phi_v \begin{pmatrix} a_1 \\ \vdots \\ a_r \end{pmatrix} = \begin{pmatrix} \phi(a_1 - v_1) \\ \vdots \\ \phi(a_r - v_r) \end{pmatrix},$$

where $\phi : \mathbb{R} \rightarrow \mathbb{R}$ given as $\phi(x) = \max\{x, 0\}$ is the ReLU activation function. By convention, when $v = 0$ we write ϕ to denote ϕ_v . With this notation, we consider neural network functions $f : \mathbb{R}^{p_0} \rightarrow \mathbb{R}^{p_{L+1}}$ of the form

$$f(x) = A^{(L)} \phi_{V_L} \circ A^{(L-1)} \phi_{V_{L-1}} \circ \cdots \circ A^{(1)} \phi_{V_1} \circ A^{(0)} x, \quad (3)$$

where \circ denotes the composition of functions, and $A^{(i)} \in \mathbb{R}^{p_{i+1} \times p_i}$, $V_i \in \mathbb{R}^{p_i}$, $p_0, \dots, p_{L+1} \in \mathbb{N}$ for $i \in \{0, 1, \dots, L+1\}$. Here the matrices $\{A^{(i)}\}$ are the weights in the network, L is the number of layers, and $(p_0, \dots, p_{L+1})^\top \in \mathbb{R}^{L+2}$ the width vector.

Since we focus on quantile regression restricted to neural networks with ReLU activation functions, we briefly review how joint estimation of quantiles can be achieved. Specifically, if multiple quantile levels are given in a set $\Lambda \subset (0, 1)$, then it is natural to estimate the quantile functions $\{f_\tau^*(\cdot)\}_{\tau \in \Lambda}$ by solving the problem

$$\begin{aligned} \{\hat{f}_\tau\}_{\tau \in \Lambda} &= \arg \min_{\{f_\tau\}_{\tau \in \Lambda} \subset \mathcal{F}} \sum_{\tau \in \Lambda} \sum_{i=1}^n \rho_\tau(y_i - f_\tau(x_i)) \\ \text{subject to } &f_\tau(x_i) \leq f_{\tau'}(x_i) \quad \forall \tau < \tau', \quad \tau, \tau' \in \Lambda, \quad i = 1, \dots, n. \end{aligned} \quad (4)$$

The constraints in (4) are noncrossing restrictions that are meant to ensure the monotonicity of quantiles. However, due to the nature of stochastic subgradient descent, the monotonicity constraints in (4) can make finding a solution to this problem challenging. To address this, letting $\tau_0 < \dots < \tau_m$ be the elements of Λ , we solve

$$\begin{aligned} \{\hat{h}_\tau\}_{\tau \in \Lambda} &= \arg \min_{\{h_\tau\}_{\tau \in \Lambda} \subset \mathcal{F}} \sum_{i=1}^n \rho_\tau(y_i - h_{\tau_0}(x_i)) + \sum_{j=1}^m \sum_{i=1}^n \rho_\tau \left\{ y_i - h_{\tau_0}(x_i) - \sum_{l=1}^j \log \left(1 + e^{h_{\tau_l}(x_i)} \right) \right\} \end{aligned} \quad (5)$$

and set

$$\hat{f}_{\tau_0}(x) = \hat{h}_{\tau_0}(x), \quad \text{and} \quad \hat{f}_{\tau_j}(x) = \hat{h}_{\tau_0}(x) + \sum_{l=1}^j \log \left(1 + e^{\hat{h}_{\tau_l}(x)} \right) \quad \text{for } j = 1, \dots, m.$$

By construction, (5) implies that the quantile functions $\{\hat{f}_\tau\}_{\tau \in \Lambda}$ satisfy the monotonicity constraint in (4). We find this approach to be numerically stable as compared to other choices such as replacing the terms $\log(1 + e^{\hat{h}_{\tau_l}(x_i)})$ with $e^{\hat{h}_{\tau_l}(x_i)}$. A different alternative is to estimate the quantile functions separately and then to order their output as in [Chernozhukov et al. \(2010\)](#) and [Zhang et al. \(2019\)](#). In this paper, we will focus on solving (5) which we find to be better in practice.

2.2 Extension to multivariate response

The framework that we have considered so far restricts the outcome variable to be univariate. However, in many machine learning problems where neural networks are used the outcome is multivariate. In this section we discuss two simple extensions of the quantile loss to the multivariate response setting. Our experiments section will contain empirical evaluations of the proposals here.

2.2.1 Geometric quantiles

We start by considering geometric quantiles. These were introduced by [Chaudhuri \(1996\)](#) to generalize quantiles to multivariate settings. Specifically, suppose that we are given data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^p$, with $p > 1$. Furthermore, consider the Euclidean unit ball \mathbb{R}^p , namely $B^{(p)} = \{u \in \mathbb{R}^p : \|u\| \leq 1\}$. [Chaudhuri \(1996\)](#) defines the function $\Psi(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$,

$$\Psi(u, x) = \|x\| + x^\top u,$$

and proposes to minimize the empirical risk associated with this loss. Motivated by the geometric quantile framework, we define the geometric quantile based on $u \in B^{(p)}$ and a ReLU network class $\mathcal{F} \subset \{f : f : \mathbb{R}^d \rightarrow \mathbb{R}^p\}$, as

$$\hat{f}_u = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \Psi(u, y_i - f(x_i)).$$

Notice that when $u = 0$, \hat{f}_u becomes

$$\hat{f}_u = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \|y_i - f(x_i)\|. \tag{6}$$

The latter can be thought as an estimator of the mean of y_i conditioning on x_i . In fact, (6) is commonly known as the L_1 -median, see [Vardi and Zhang \(2000\)](#). The L_1 -median can be interpreted as a robust version of the usual least squares,

$$\arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \|y_i - f(x_i)\|^2. \tag{7}$$

This is due to the fact that replacing $\|\cdot\|^2$ with $\|\cdot\|$, as in (6), has the advantage that large residuals are not heavily penalized as in (7).

2.2.2 Marginal quantiles

Marginal quantiles have perhaps the advantage over geometric quantiles in that they can produce actual prediction intervals, and have probabilistic meaning. However, as their name suggests, marginal quantiles only produce predication intervals for each variable in the output marginally, and thus do not produce a prediction region for the output jointly.

Let $\tau \in (0, 1)$, and $f_\tau^* : \mathbb{R}^d \rightarrow \mathbb{R}^p$, $f_\tau^*(x) = (f_{\tau,1}^*(x), \dots, f_{\tau,p}^*(x))^\top$, where

$$f_{\tau,j}^*(x), = F_{Y_j|X=x}^{-1}(\tau), \quad (8)$$

where $Y = (Y_1, \dots, Y_p)^\top \in \mathbb{R}^p$. The functions $f_{\tau,1}^*(x), \dots, f_{\tau,p}^*(x)$ are the marginal quantiles of Y_1, \dots, Y_p respectively, conditioning on X . Marginal quantiles have been studied in the literature (c.f. [Babu and Rao, 1989](#); [Abdous and Theodorescu, 1992](#)). Given $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^p$ independent copies of (X, Y) , the multivariate function f_τ^* can be estimated with a multivariate output ReLU neural network architecture \mathcal{F} as

$$\hat{f}_\tau = \arg \min_{f \in \mathcal{F}} \sum_{j=1}^p \sum_{i=1}^n \rho_\tau(y_{i,j} - f(x_i)). \quad (9)$$

To be specific, here the class \mathcal{F} consists of functions of the form (3) with $p_0 = d$ and $p_{L+1} = p$.

3 Theory

We now proceed to provide statistical guarantees for quantile regression with ReLU networks. Our theory is organized in three parts. First, we provide a general upper bound on the mean squared error for estimating the quantile function. Second, we study a setting where the quantile function is a member of a space of compositions of functions whose coordinates are Hölder functions. Finally, we assume that the quantile function belongs to a Besov space.

3.1 Notation

Throughout this section, for functions $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$, we define the function $\Delta_n^2(f, g)$ as

$$\Delta_n^2(f, g) := \frac{1}{n} \sum_{i=1}^n D^2(f(x_i) - g(x_i)), \quad (10)$$

with $\{x_i\}_{i=1}^n$ the features and where

$$D^2(t) := \min \{|t|, t^2\}. \quad (11)$$

This function was used as performance metric in a different quantile regression context in [Padilla and Chatterjee \(2020\)](#).

Furthermore, for bounded functions f and g with $f, g : [0, 1]^d \rightarrow \mathbb{R}$, we define $\Delta^2(f, g)$ as

$$\Delta^2(f, g) := \mathbb{E}(D^2(f(X) - g(X))),$$

where $X \sim U[0, 1]^d$ and set $\Delta(f, g) = \sqrt{\Delta^2(f, g)}$. We also write

$$\|f - g\|_{\ell_2} := \sqrt{\mathbb{E}((f(X) - g(X))^2)},$$

and

$$\|f - g\|_n^2 := \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2. \quad (12)$$

For a matrix $A \in \mathbb{R}^{s \times t}$ we define

$$\|A\|_0 = |\{(i, j) : A_{i,j} \neq 0, i \in \{1, \dots, s\}, j \in \{1, \dots, t\}\}|, \quad \|A\|_\infty = \max_{i=1, \dots, s, j=1, \dots, t} |A_{i,j}|.$$

Finally, we refer to the quantities $\varepsilon_i = y_i - f_\tau^*(x_i)$ for $i = 1, \dots, n$ as the errors.

3.2 General upper bound

In this subsection we focus on quantile regression ReLU estimators of the form

$$\hat{f} = \arg \min_{f \in \mathcal{F}(W, U, L), \|f\|_\infty \leq F} \sum_{i=1}^n \rho_\tau(y_i - f(x_i)), \quad (13)$$

where $\mathcal{F}(W, U, L)$ is the class of networks of the form (3) such that the number of parameters in the network is W , the number of nodes is U , and the number of layers is L . Here, F is a fixed positive constant.

Before arriving at our first result, we start by stating some assumptions regarding the generative model. Throughout, we consider $\tau \in (0, 1)$ as fixed.

Assumption 1. We write $f_\tau^*(x_i) = F_{y_i|x_i}^{-1}(\tau)$ for $i = 1, \dots, n$. Here $F_{y_i|x_i}$ is cumulative distribution function of y_i conditioning on x_i for $i = 1, \dots, n$. Also, $y_1, \dots, y_n \in \mathbb{R}$ are assumed to be independent.

Notice that Assumption 1 simply requires that the different outcome measurements are independent conditioning on the design, which for this subsection is assumed to be fixed.

Assumption 2. There exists a constant $L > 0$ such that for $\delta \in \mathbb{R}^n$ satisfying $\|\delta\|_\infty \leq L$ we have that

$$\min_{i=1, \dots, n} p_{y_i|x_i}(f_\tau^*(x_i) + \delta_i) \geq \underline{p},$$

for some $\underline{p} > 0$, and where $p_{y_i|x_i}$ is the probability density function of y_i conditioning on x_i . We also require that

$$\sup_{t \in \mathbb{R}} p_{y_i|x_i}(t) \leq c, \text{ a.s.},$$

for some constant $c > 0$.

Assumption 2 requires that there exists a neighborhood around $f_\tau^*(x_i)$ in which the probability density function of y_i conditioning on x_i is bounded below. Related conditions appeared as D.1 [Belloni and Chernozhukov \(2011\)](#), Condition 2 in [He and Shi \(1994\)](#), and Assumption 2 in [Padilla and Chatterjee \(2020\)](#).

Next we define f_n , the projection of the quantile function f_τ^* onto the network class $\mathcal{F}(W, U, L)$ in the sense of the quantile risk.

Definition 1. We define the function f_n as

$$f_n \in \arg \min_{f \in \mathcal{F}(W, U, L), \|f\|_\infty \leq F} \mathbb{E} \left[\sum_{i=1}^n \rho_\tau(z_i - f(x_i)) - \sum_{i=1}^n \rho_\tau(z_i - f_\tau^*(x_i)) \right],$$

where $z \in \mathbb{R}^n$ is an independent copy of y . We also define the approximation error as

$$\text{err}_1 = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \rho_\tau(z_i - f_n(x_i)) - \frac{1}{n} \sum_{i=1}^n \rho_\tau(z_i - f_\tau^*(x_i)) \right].$$

Notice that when $f_\tau^* \in \mathcal{F}(W, U, L)$ and $\|f\|_\infty^* \leq F$ then $f_n = f_\tau^*$ and the approximation error is zero. However, in general $f_\tau^* \notin \mathcal{F}(W, U, L)$ and $\text{err}_1 \geq 0$.

We are now ready to present our first theorem which exploits the VC dimension results from [Bartlett et al. \(2019\)](#).

Theorem 1. *Suppose that Assumptions 1–2 hold and $n \geq CLW \log(U)$ for a large enough $C > 0$. Then \hat{f} defined in (13) satisfies*

$$\mathbb{E} \left[\Delta_n^2(f_\tau^*, \hat{f}) \mid x_1, \dots, x_n \right] \leq c_1 F \left[\frac{\{LW \log U \cdot \log n\}}{n} \right]^{1/2} + c_1 \text{err}_1,$$

with $c_1 > 0$ a constant. Furthermore, it also holds that

$$\mathbb{E} \left[\|\hat{f} - f_\tau^*\|_n^2 \mid x_1, \dots, x_n \right] \leq c_1 \max\{1, F\} F \left[\frac{\{LW \log U \cdot \log n\}}{n} \right]^{1/2} + c_1 \max\{1, F\} \text{err}_1.$$

Theorem 1 provides a general bound on the mean squared error that depends on the sample size n , the parameters of the network, and the approximation error. For instance, if we regard L , W and U as constant, then the rate becomes $n^{-\frac{1}{2}} + \text{err}_1$. In the next two subsections we will consider classes of ReLU network with more structure which will lead to rates that match minimax rates in nonparametric regression.

3.3 Space of compositions based on Hölder functions

Next we provide convergence rates for quantile regression with ReLU networks under the assumption that the quantile function belongs to a class of functions based on Hölder spaces. Such class of functions, defined below, was studied in [Schmidt-Hieber \(2017\)](#). There, the authors showed that for such class, neural networks with ReLU activation function attain minimax rates. However, the results in [Schmidt-Hieber \(2017\)](#) hold under the assumption of Gaussian errors. We now show that it is possible to attain the same rates under general error assumptions by employing the quantile loss. Before arriving at such result we start by providing some definitions.

Definition 2. We define the class of ReLU neural networks $\mathcal{G}(L, p, s, F)$ as

$$\begin{aligned} \mathcal{G}(L, p, S, F) = & \left\{ f : f \text{ is of form (3), and } \sum_{j=0}^L (\|A^{(l)}\|_0 + \|V_l\|_0) \leq S, \|f\|_\infty \leq F, \right. \\ & \left. \max_{j=0,1,\dots,L} \|A^{(j)}\|_\infty \leq 1, \max_{j=1,\dots,L} \|V_j\|_\infty \leq 1 \right\}. \end{aligned}$$

With the notation in Definition 2, we consider the estimator

$$\hat{f} = \arg \min_{f \in \mathcal{G}(L, p, S, F)} \sum_{i=1}^n \rho_\tau(y_i - f(x_i)), \quad (14)$$

and define a $\|\cdot\|_\infty$ -projection of f_τ^* , the true quantile function, onto $\mathcal{G}(L, p, S, M)$ as

$$f_n \in \arg \min_{f \in \mathcal{G}(L, p, S, F)} \|f - f_\tau^*\|_\infty.$$

A few comments are in order. First, notice that we assume that all the parameters are bounded by one. As discussed in [Schmidt-Hieber \(2017\)](#), this is standard and in practice can be achieved by projecting the parameters in $[-1, 1]$ after every iteration of stochastic subgradient descent. Second, we assume that the networks are sparse as was the case in [Schmidt-Hieber \(2017\)](#) and [Suzuki \(2018\)](#). See [Hassibi and Stork](#)

(1993); Han et al. (2015); Frankle and Carbin (2018); Gale et al. (2019) for different approaches to produce sparse networks.

Before stating our main result of this subsection, we provide the definition of the function class that we consider. Such class requires that we introduce some notation that comes from Schmidt-Hieber (2017).

Definition 3. For $\beta > 0$ and $r \in \mathbb{N}_+$ we define the class of Hölder functions of exponent β as

$$\mathcal{C}_r^\beta(I, K) = \left\{ f : I \subset \mathbb{R}^r \rightarrow \mathbb{R} : \sum_{\alpha: \|\alpha\|_1 < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha: \|\alpha\|_1 = \lfloor \beta \rfloor} \sup_{x \neq y, x, y \in I} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leq K \right\},$$

where $\partial^\alpha = \partial^{\alpha_1} \cdots \partial^{\alpha_r}$ with $(\alpha_1, \dots, \alpha_r) \in \mathbb{N}^r$.

Definition 4. For $q \in \mathbb{N}_+$, $d = (d_0, \dots, d_{q+1}) \in \mathbb{N}_+^{q+2}$, $t = (t_0, \dots, t_q) \in \mathbb{N}_+^{q+1}$, $\beta = (\beta_0, \dots, \beta_q) \in \mathbb{R}_+^{q+1}$ and $K \in \mathbb{R}_+$ we define the class of functions

$$\begin{aligned} \mathcal{H}(q, d, t, \beta, K) = & \left\{ f = g_q \circ \cdots \circ g_0 : g_i = (g_{i,j})_j : [a_i, b_i]^{d_i} \rightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}} \right. \\ & \left. g_{i,j} \in \mathcal{C}_{t_i}^{\beta_i}([a_i, b_i]^{t_i}, K), \quad \text{and} \quad |a_i|, |b_i| \leq K \right\}. \end{aligned}$$

With Definitions 3–4 in hand, we now state an assumption on the true quantile function regarding its smoothness.

Assumption 3. The quantile function f_τ^* satisfies $f_\tau^* \in \mathcal{H}(q, d, t, \beta, K)$ for some $q \in \mathbb{N}_+$, $d = (d_0, \dots, d_{q+1}) \in \mathbb{N}_+^{q+2}$, $t = (t_1, \dots, t_q) \in \mathbb{N}_+^q$, $\beta = (\beta_0, \dots, \beta_q) \in \mathbb{R}_+^{q+1}$ and $K \in \mathbb{R}_+$. We also require that $\|f_\tau^*\| \leq F$ where F is the same appearing in (14). Moreover, we define the smoothness indices

$$\beta_i^* = \beta_i \prod_{l=i+1}^q \min\{\beta_l, 1\}.$$

As for the distribution of the data, our next assumption requires that the covariates have a probability density function that is bounded by above and below.

Assumption 4. We assume that $\{(x_i, y_i)\}_{i=1}^n$ are independent copies of (X, Y) , with x_i having a probability density function g_X with support in $[0, 1]^d$ and such that

$$c_1 \leq \inf_{x \in [0, 1]^d} g_X(x) \leq \sup_{x \in [0, 1]^d} g_X(x) \leq c_2,$$

for some constants $c_1, c_2 > 0$.

We are now ready to state our main result of this subsection that exploits the approximation results from Schmidt-Hieber (2017).

Theorem 2. Suppose that Assumptions 1–4 hold. In addition, suppose that for the class $\mathcal{G}(L, p, S, F)$ the parameters are chosen to satisfy

$$\begin{aligned} \sum_{i=0}^q \log_2(4 \max\{t_i, \beta_i\}) \log_2(n) &\leq L \lesssim n\epsilon_n, \quad \max\{1, K\} \leq F, \\ n\epsilon_n &\lesssim \min_{i=1, \dots, L} p_i, \quad S \asymp n\epsilon_n \log n, \quad \max_{i=1, \dots, L} p_i \lesssim n, \end{aligned}$$

where

$$\epsilon_n = \max_{i=0, 1, \dots, q} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}. \tag{15}$$

Then there exists a constant $C > 0$ such that with probability approaching one, we have that

$$\max \left\{ \|\hat{f} - f_\tau^*\|_{\ell_2}^2, \|\hat{f} - f_\tau^*\|_n^2 \right\} \leq C\epsilon_n L \log^2 n,$$

where \hat{f} is the estimator defined in (14). Hence, if in addition $L \asymp \log n$, then

$$\max \left\{ \|\hat{f} - f_\tau^*\|_{\ell_2}^2, \|\hat{f} - f_\tau^*\|_n^2 \right\} \leq C\epsilon_n \log^3 n,$$

with probability approaching one.

Notice that Theorem 2 shows that the ReLU network based estimator defined in (14) attains the rate ϵ_n under the mean squared error and the ℓ_2 metrics, ignoring L and the log factors, for estimating quantile functions in the class $\mathcal{G}(L, p, S, F)$. Importantly, the rate ϵ_n is minimax for estimating functions in the class $\mathcal{G}(L, p, S, F)$, see Theorem 3 in Schmidt-Hieber (2017). However, it is important to emphasize here that the result in Theorem 2 holds under more general assumptions than those in Corollary 1 from Schmidt-Hieber (2017). In particular, Theorem 2 allows for the possibility of heavy-tailed error distributions.

3.4 Besov spaces

Next we study quantile regression with ReLU networks in the context of Besov spaces. Our main result from this subsection will be similar in spirit to Theorem 2 but under the assumption that the quantile function belongs to a Besov space. To arrive at our main result, we first introduce some notation regarding the ReLU class of networks that we consider.

Definition 5. For $W, L \in \mathbb{N}_+$, $S, B \in \mathbb{R}$ we define the class of sparse networks $\mathcal{I}(L, W, S, B)$ as

$$\mathcal{I}(L, W, S, B) := \left\{ (A^{(L)}\phi(\cdot) + b^{(L)}) \circ \dots \circ (A^{(1)}x + b^{(1)}) : A^{(l)} \in \mathbb{R}^{W \times W}, b^{(l)} \in \mathbb{R}^W \right. \\ \left. \sum_{l=1}^L (\|A^{(l)}\|_0 + \|b^{(l)}\|_0) \leq S, \max_l \max\{\|A^{(l)}\|_\infty, \|b^{(l)}\|_\infty\} \leq B \right\}.$$

Notice that the space of networks $\mathcal{I}(L, W, S, B)$ is actually similar to $\mathcal{G}(L, p, S, F)$. The main difference is that the networks in the former class have weight matrices of the same size across the different layers. This minor differences are only necessary in order to achieve the theoretical guarantees under the different classes to which the quantile function belongs.

We the notation from Definition 5, we focus on the estimator

$$\hat{f} = \arg \min_{f \in \mathcal{I}(L, W, S, B), \|f\|_\infty \leq F} \sum_{i=1}^n \rho_\tau(y_i - f(x_i)), \quad (16)$$

where $F > 0$ is fixed.

Before providing a statistical guarantee for \hat{f} in (16), we first state the required assumptions imposed on the generative model.

Assumption 5. The quantile function satisfies $f_\tau^* \in B_{p,q}^s([0, 1]^d)$, $\|f_\tau^*\|_\infty \leq F$, where for $0 < p, q \leq \infty$, and $0 < s < \infty$ we have $s \geq d/p$. Furthermore, the exists $m \in \mathbb{N}$ such that $0 < s < \min\{m, m-1+1/p\}$. Here, $B_{p,q}^s([0, 1]^d)$ is a Besov space in $[0, 1]^d$ as Definition 9 in the Appendix.

We are now ready to state the main result concerning estimation of a quantile function that belongs to a Besov space.

Theorem 3. Suppose that Assumptions 1–2 and 4–5 hold. In addition, suppose that for the class $\mathcal{I}(L, W, S, B)$ the parameters are chosen as

$$L = 3 + 2 \lceil \log_2 \left(\frac{3^{\max\{d,m\}}}{\epsilon c_{d,m}} \right) + 5 \rceil \lceil \log_2 \max\{d, m\} \rceil, \quad W = W_0 N, \\ S = (L - 1)W_0^2 N + N, \quad B = O \left(N^{(v^{-1} + d^{-1})(\max\{1, (d/p-s)_+\})} \right),$$

for a constant $c_{d,m} > 0$ that depends on d and m , a constant $W_0 > 0$, and where $v = (s - \delta)/\delta$,

$$\delta = d/p, \quad \epsilon = N^{-s/d - (v^{-1} + d^{-1})(d/p-s)_+} + \{\log N\}^{-1}, \quad N \asymp n^{\frac{d}{2s+d}}.$$

Then there exists a constant $C > 0$ such that with probability approaching one, we have that

$$\max \left\{ \|\hat{f} - f_\tau^*\|_{\ell_2}^2, \|\hat{f} - f_\tau^*\|_n^2 \right\} \leq C \frac{(\log n)^2}{n^{\frac{2s}{2s+d}}},$$

with \hat{f} defined as (16).

Notably, Theorem 3 shows that the neural network based quantile estimator \hat{f} attains the rate $n^{-\frac{2s}{2s+d}}$, ignoring logarithmic factors, for estimating the quantile function. This results generalizes Theorem 2 from Suzuki (2018) to the quantile regression setting. In particular, Theorem 3 holds under general assumptions of the errors allowing for heavy-tailed distributions. Furthermore, the rate $n^{-\frac{2s}{2s+d}}$ is minimax for the space $B_{p,q}^s([0, 1]^d)$ satisfying the conditions in Assumption 5, see Donoho et al. (1998) and Suzuki (2018).

4 Experiments

We study the performance of ReLU networks for quantile regression across a suite of heavy-tailed synthetic and real-data benchmarks. The benchmarks include both univariate and multivariate responses. For univariate responses, we compare ReLU methods against quantile regression versions of random Forests (Meinshausen, 2006) and splines (Koenker et al., 1994; He and Shi, 1994). In the univariate synthetic benchmarks, ReLU networks are shown to outperform random Forests in all of the tested settings; splines outperform ReLU networks only when the true response function is smooth. For multivariate responses, we consider the two different loss functions for multivariate quantiles proposed in section 2.2. In both univariate and multivariate responses, the ReLU networks with quantile-based losses perform better when estimating the mean than using a squared error loss.

4.1 Univariate response

We assess the performance of quantile regression with ReLU networks (Quantile Networks) on five different generative models. Each model involves a set of covariates and a univariate response target. The covariates determine the location of the response and a zero-mean, symmetric function with heavy tails is used as the noise distribution; we focus here on Student's t and Laplace distributions.

We compare Quantile Networks with three other nonparametric methods: (i) mean squared error regression with ReLU networks (SqErr Networks), as in Problem 1; (ii) quantile regression with natural splines (Quantile Splines, Koenker et al., 1994; He and Ng, 1999); and quantile regression with random Forests (Quantile Forests, Meinshausen, 2006). For the two neural network methods, we train the models using stochastic gradient descent (SGD) as implemented in PyTorch (Paszke et al., 2019) with Nesterov momentum of 0.9, starting learning rate of 0.1, and stepwise decay 0.5. The neural network models also use the same architecture: two hidden layers of 200 units each, with dropout rate of 0.1 and batch normalization

in each layer. For the other two nonparametric methods, we choose parameters to be flexible enough to capture a large number of nonlinearities while still computationally feasible on a laptop for moderate-sized problems. For Quantile Splines, we use a natural spline basis with 3 degrees of freedom; we use the implementation available in the `statsmodels` package.¹ For Regression Forests, we use 100 tree estimators and a minimum sample count for splits of 10; these are defaults in the `scikit-garden` package.²

We assess the performance of all methods using the mean squared error (MSE) between the estimated and true quantile functions. In each experiment, the methods are estimated at different training sample sizes n , $n \in \{100, 1000, 10000\}$, and different quantile levels τ , $\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$. Since the SqErr Network only estimates the mean, we only evaluate it at $\tau = 0.50$, which is equivalent to the mean in all benchmarks. For each benchmark, we generate 25 datasets independently from the same generative model and evaluate performance using 10000 sampled covariates with the corresponding true quantile. In each scenario the data are generated following the same location-plus-noise template,

$$\begin{aligned} y_i &= f_0(x_i) + \epsilon_i, \quad i = 1, \dots, n, \\ x_i &\stackrel{\text{ind}}{\sim} [0, 1]^d, \end{aligned}$$

where $\epsilon_i \sim G_i$ for a distribution G_i in \mathbb{R} , and with $f_0 : [0, 1]^d \rightarrow \mathbb{R}$ for a choice of d that is scenario dependent. We consider 5 different scenarios following this template:

Scenario 1. We set

$$\begin{aligned} f_0(q) &= g_2 \circ g_1(q), \quad \forall q \in \mathbb{R}^2 \\ g_1(q) &= (\sqrt{q_1} + q_1 \cdot q_2, \cos(2\pi q_2))^\top, \quad \forall q \in \mathbb{R}^2, \\ g_2(q) &= \sqrt{q_1 + q_2^2} + q_1^2 \cdot q_2, \quad \forall q \in \mathbb{R}^2, \end{aligned}$$

and $\epsilon_i = v_i g_3(x_i)$ where

$$g_3(q) = \|q - (1/2, 1/2)^\top\|, \quad \forall q \in \mathbb{R}^2,$$

with $v_i \stackrel{\text{ind}}{\sim} U[0, 1]^2$, for $i = 1, \dots, n$, where $t(2)$ is the t-distribution with 2 degrees of freedom.

Scenario 2. In this scenario we specify

$$f_0(q) = q_1^2 + q_2^2, \quad q \in [0, 1]^2,$$

and generate $\epsilon_i \stackrel{\text{ind}}{\sim} \text{Laplace}(0, 2)$ for $i = 1, \dots, n$.

Scenario 3. This is constructed by defining $f_0 : [0, 1]^2 \rightarrow \mathbb{R}$ as

$$f_0(q) = \begin{cases} \sqrt{q_1 + q_2} + 1 & \text{if } q_1 < 0.5, \\ \sqrt{q_1 + q_2} & \text{otherwise,} \end{cases}$$

and setting

$$\epsilon_i = \sqrt{x_i^\top \beta} \nu_i,$$

where $\beta = (1, 1/2)^\top$ and $\nu_i \stackrel{\text{ind}}{\sim} t(2)$ for $i = 1, \dots, n$.

¹<https://www.statsmodels.org>

²<https://scikit-garden.github.io/>

Scenario 4. The function f_0 is chosen as

$$f_0(q) = \sqrt{q_1 + q_2 + q_3 + q_4 + q_5}, \quad q \in [0, 1]^d,$$

with $d = 5$, and the errors as $\epsilon_i \stackrel{\text{ind}}{\sim} \text{Laplace}(0, 2)$ for $i = 1, \dots, n$. Here, $\text{Laplace}(0, 2)$ is the Laplace distribution with mean zero and scale parameter 2.

Scenario 5. The function $f_0 : [0, 1]^{10} \rightarrow \mathbb{R}$ is defined as $f_0(q) = g_3 \circ g_2 \circ g_1(q)$ where

$$\begin{aligned} g_1(q) &= (\sqrt{q_1^2 + \sum_{j=2}^{10} q_j}, (\sum_{j=1}^{10} q_j)^3)^\top, \quad q \in [0, 1]^{10}, \\ g_2(q) &= (|q_1|, q_2 \cdot q_1)^\top, \quad q \in [0, 1]^2, \\ g_3(q) &= q_1 + \sqrt{q_1 + q_2}, \quad q \in [0, 1]^2, \end{aligned}$$

where $\epsilon_i \stackrel{\text{ind}}{\sim} t(3)$ for $i = 1, \dots, n$, and $t(3)$ denotes the t -distribution with 3 degrees of freedom.

We visualize the performances different approaches and true quantile functions in Figures 2–3. There, we can see that Quantile Network is a better estimate for the quantile functions in general, compared to the other methods.

We report the results for Scenarios 1–3 in Table 1 and Scenarios 4–5 in Table 2. From Table 1 we can see that in Scenario 1, the Quantile Network method outperforms the competitors for most quantiles and sample sizes. The advantage becomes more evident as the sample size grows. The closest competitor is Quantile Spline which is the best method in some small sample problems. Furthermore, in Scenario 2 the best method is Quantile Spline, with Quantile Network as second best. This is not surprising since Scenario 2 consists of very smooth quantile functions defined in a low dimensional domain ($d = 2$). In contrast, Scenario 3 consists of a quantile function with discontinuities and heteroscedastic errors. In this more challenging setting, Quantile Network outperforms others for larger values of n .

We do not compare against Quantile Splines in Scenarios 4–5 as such method does not scale up to 5 dimensional problems or above. In Table 2, we show that for Scenarios 4–5 the clear best method is Quantile Network (for $n > 100$), with Quantile Forests as the second best.

Overall, the results in Tables 1–2 demonstrate a clear advantage of the Quantile Network method. This method generally outperforms SqErr Network in all examples, presumably due to the heavy-tailed or heteroscedastic error distributions. At the same time, Quantile Network also outperforms the other competitors with larger sample size or more complicated quantile functions.

4.2 Multivariate response

We explore the performance of different quantile ReLU network approaches for multivariate responses as discussed in Section 2.2. We refer to the estimator in equation (6) as Geometric Quantile, and the estimator in equation (9) as Quantile Network. As a benchmark, once again, we consider the estimator based on the squared error loss and as defined in equation (7). For all these estimators, the corresponding network class is chosen as in section 4.1.

We conduct simulations in two different scenarios. In each scenario, we evaluate performance based on mean squared error defined as

$$\frac{1}{np} \sum_{j=1}^p \sum_{i=1}^n \left(f_{\tau,j}^*(x_i) - \hat{f}(x_i) \right)^2,$$

where the quantile functions $f_{\tau,j}^*(\cdot)$, $j = 1, \dots, p$, are defined in (8), and $\tau = 0.5$.

We consider two multivariate response generative models as the follows.

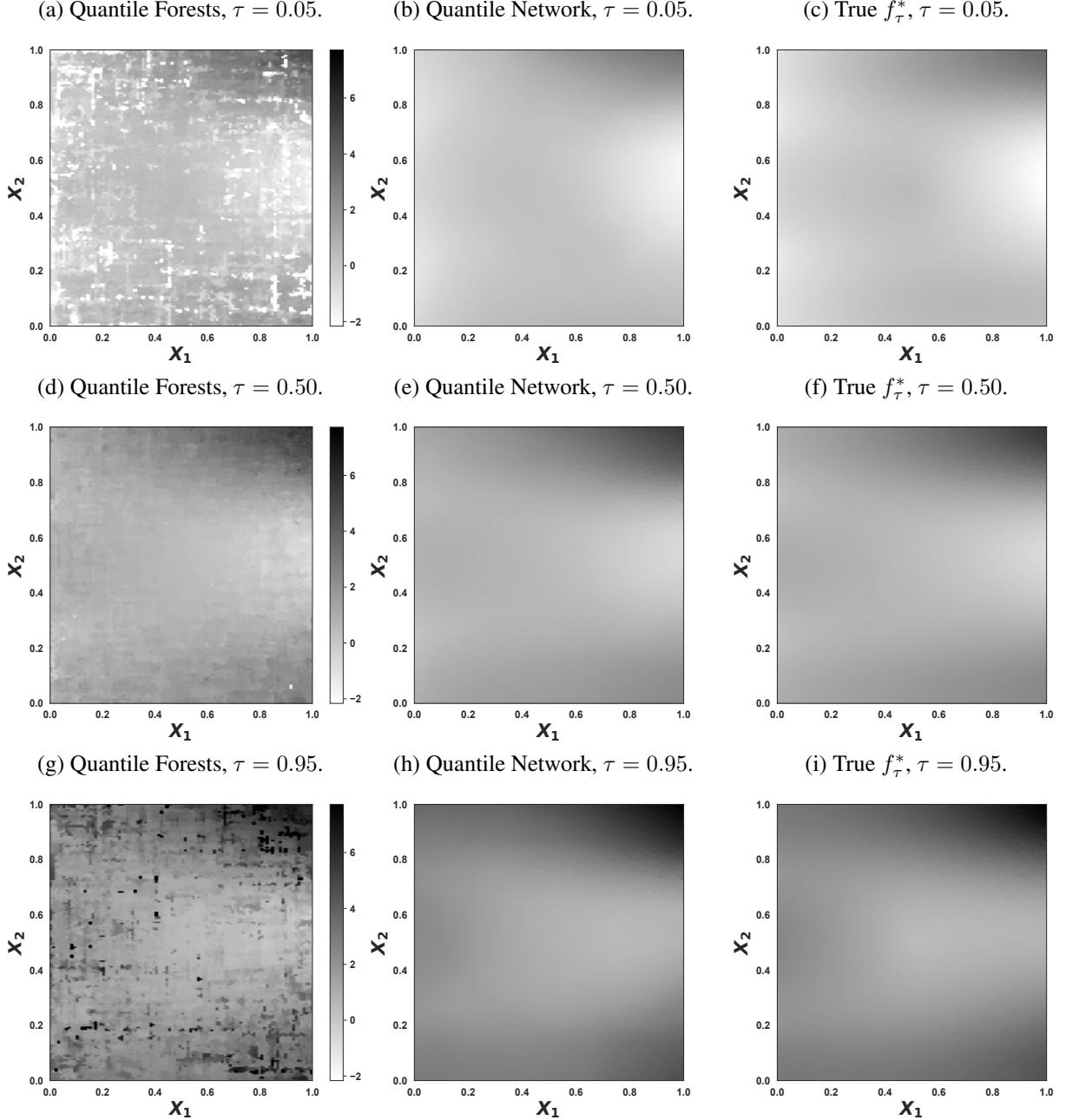


Figure 1: One instance of the true quantile function with $\tau \in \{0.05, 0.50, 0.95\}$ and its corresponding estimates based on Quantile Network and Quantile Forests. Here $n = 10000$ and the data are generated under Scenario 1.

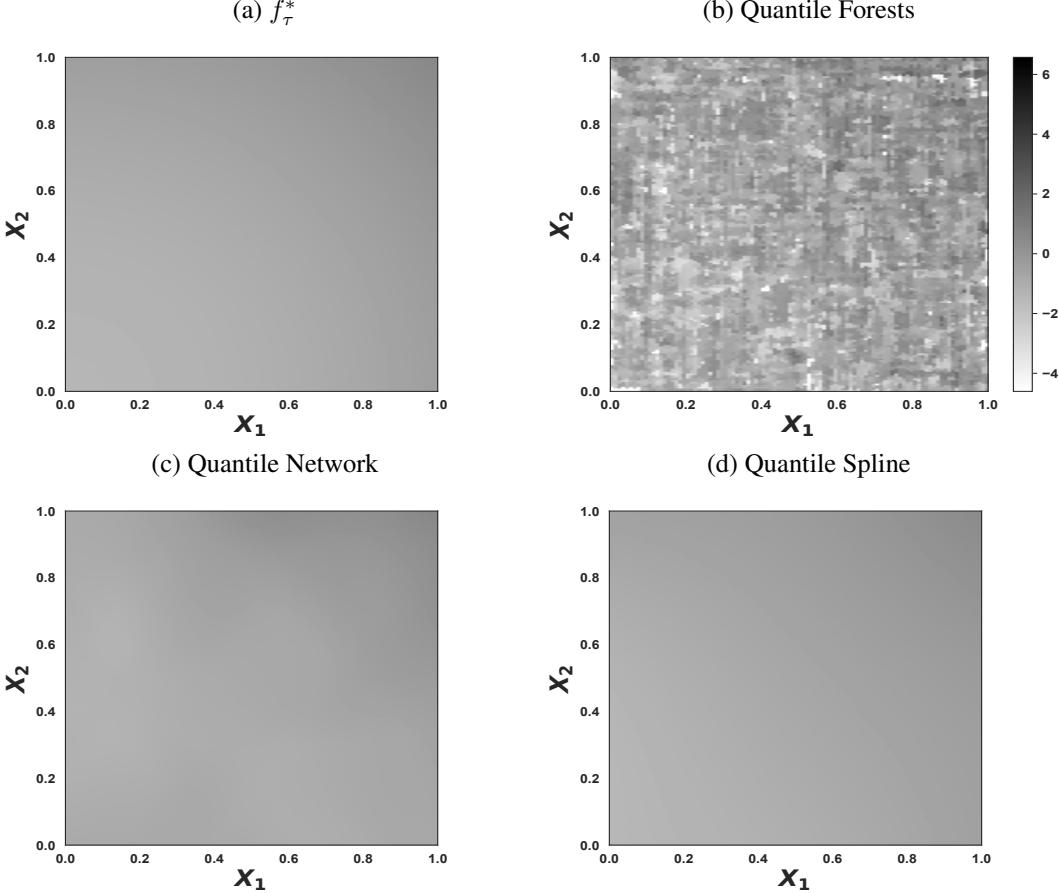


Figure 2: One instance of the true quantile function with $\tau = 0.25$ and its corresponding estimates obtained from different methods. Here $n = 10000$ and the data are generated under Scenario 2.

Scenario 6.

$$\begin{aligned}
 y_i &= g_2 \circ g_1(x_i) + \epsilon_i \\
 g_1(q) &= (|q_1|, q_2 \cdot q_1)^\top \\
 g_2(q) &= (\sqrt{q_2^2 + q_1}, (q_1 + q_2)^3)^\top \\
 x_i &\stackrel{\text{ind}}{\sim} U[0, 1]^2, \quad i = 1, \dots, n, \\
 \epsilon_i &\stackrel{\text{ind}}{\sim} Mt_3(0, I_2), \quad i = 1, \dots, n,
 \end{aligned}$$

where $Mt_3(0, I_2)$ is the multivariate t -distribution with 3 degrees of freedom and scale matrix identity $I_2 \in \mathbb{R}^{2 \times 2}$.

Scenario 7.

$$\begin{aligned}
 y_i &= f_0(x_i) + \epsilon_i \\
 f_0(q) &= (\sqrt{q_1^2 + q_2^2}, \sqrt{q_3^2 + q_4^2})^\top \\
 x_i &\stackrel{\text{ind}}{\sim} U[0, 1]^4, \\
 \epsilon_{i,j} &\stackrel{\text{ind}}{\sim} \text{Laplace}(0, 2).
 \end{aligned}$$

Table 3 illustrates the performance of different methods in Scenarios 6 and 7 with sample sizes $n = \{100, 1000, 10000\}$. We can see that both Quantile Network and Geometric Quantile outperform the ℓ_2 -based approach SqErr Network. This corroborates the results earlier in Section 4.1. Particularly, it demon-

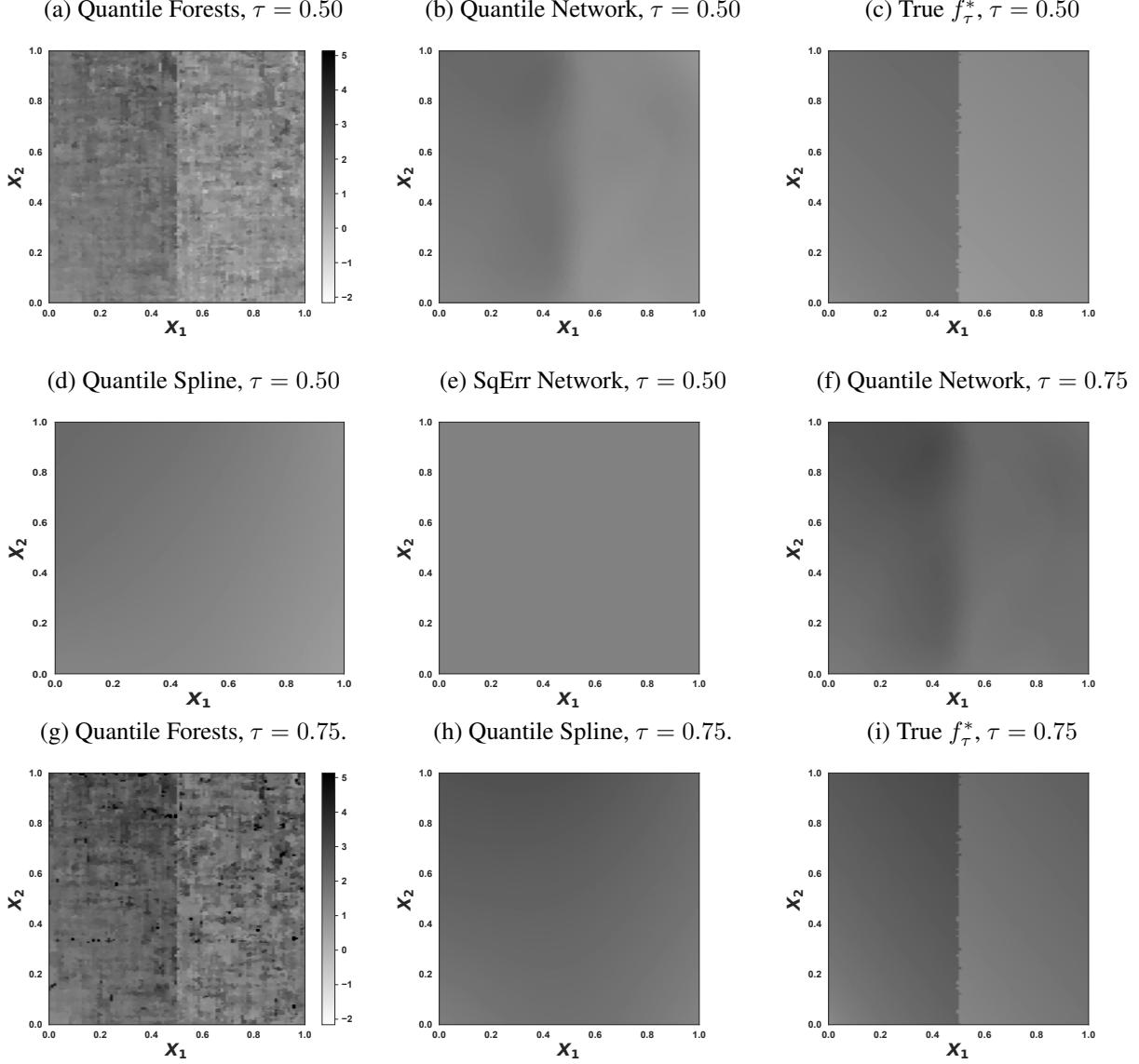


Figure 3: One instance of the true quantile function with $\tau \in \{0.50, 0.75\}$ and of the corresponding estimates obtained with the different methods. Here $n = 10000$ and the data are generated under Scenario 3.

strates that both Quantile Network and Geometric Quantile are robust estimators when dealing with heavy-tailed distributions.

5 Conclusion and Future Work

In this paper we have studied, both theoretically and empirically, the statistical performance of ReLU networks for quantile regression. Our main theorems establish minimax estimation rates under general classes of functions and distributions of the errors. These results rely on the approximation theory from [Schmidt-Hieber \(2017\)](#) and [Suzuki \(2018\)](#). Future work can extend these results to other function classes provided that the corresponding approximation theory is used or developed. Empirically, experiments in both uni-

Scenario 1						
n	Method	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
100	SqErr Network	*	*	0.98	*	*
	Quantile Network	0.60	0.50	0.46	0.68	1.85
	Quantile Spline	1.79	0.19	0.14	0.17	3.36
	Quantile Forests	1.69	0.54	0.34	0.69	2.37
1000	SqErr Network	*	*	0.95	*	*
	Quantile Network	0.36	0.06	0.04	0.06	0.18
	Quantile Spline	0.17	0.07	0.06	0.07	0.17
	Quantile Forests	1.76	0.20	0.07	0.22	1.77
10000	SqErr Network	*0	*	0.94	*	*
	Quantile Network	0.04	0.01	0.01	0.01	0.03
	Quantile Spline	0.07	0.06	0.06	0.06	0.07
	Quantile Forests	2.90	0.16	0.08	0.67	13.89
Scenario 2						
n	Method	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
100	SqErr Network	*	*	0.28	*	*
	Quantile Network	3.82	2.37	2.21	2.64	4.43
	Quantile Spline	5.51	1.28	0.75	1.28	7.27
	Quantile Forests	4.13	1.68	1.10	1.66	4.42
1000	SqErr Network	*	*	0.26	*	*
	Quantile Network	0.41	0.23	0.21	0.24	0.48
	Quantile Spline	0.72	0.11	0.04	0.11	0.53
	Quantile Forests	4.08	1.24	0.68	1.18	3.62
10000	SqErr Network	*	*	0.19	*	*
	Quantile Network	0.07	0.03	0.03	0.04	0.08
	Quantile Spline	0.06	0.01	0.00	0.01	0.06
	Quantile Forests	3.47	1.03	0.56	1.04	3.55
Scenario 3						
n	Method	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
100	SqErr Network	*	*	0.24	*	*
	Quantile Network	1.33	0.80	0.93	1.57	2.83
	Quantile Spline	6.87	0.39	0.30	0.43	4.75
	Quantile Forests	5.49	0.81	0.37	0.81	4.42
1000	SqErr Network	*	*	0.25	*	*
	Quantile Network	0.21	0.09	0.08	0.09	0.32
	Quantile Spline	0.50	0.10	0.08	0.09	0.53
	Quantile Forests	11.26	0.93	0.20	0.77	9.96
10000	SqErr Network	*	*	0.17	*	*
	Quantile Network	0.05	0.03	0.03	0.03	0.07
	Quantile Spline	0.09	0.07	0.07	0.07	0.11
	Quantile Forests	17.75	0.67	0.15	0.49	12.48

Table 1: Univariate Responses Tasks. Performances of different methods in Scenarios 1 – 3, in terms of squared error from the true quantile averaged over 25 independent trials.

Scenario 4						
n	Method	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
100	SqErr Network	*	*	0.30	*	*
	Quantile Network	5.96	3.06	3.44	4.02	5.13
	Quantile Spline	*	*	*	*	*
	Quantile Forests	3.65	1.85	1.11	1.60	3.55
1000	SqErr Network	*	*	0.13	*	*
	Quantile Network	1.09	0.54	0.44	0.58	1.33
	Quantile Spline	*	*	*	*	*
	Quantile Forest	2.53	0.76	0.37	0.82	2.69
10000	SqErr Network	*	*	0.05	*	*
	Quantile Network	0.20	0.09	0.07	0.10	0.25
	Quantile Spline	*	*	*	*	*
	Quantile Forest	1.71	0.42	0.18	0.42	1.71
Scenario 5						
n	Model	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.95$
100	SqErr Network	*	*	30.59	*	*
	Quantile Network	5.05	4.37	3.61	4.26	6.96
	Quantile Spline	*	*	*	*	*
	Quantile Forest	41.86	21.94	15.36	20.69	53.68
1000	SqErr Network	37.42	31.85	30.97	31.26	35.60
	Quantile Network	1.61	1.15	0.65	0.95	2.05
	Quantile Spline	*	*	*	*	*
	Quantile Forest	29.01	12.53	7.33	11.05	33.29
10000	SqErr Network	35.97	31.13	30.60	31.24	36.31
	Quantile Network	0.28	0.22	0.17	0.24	0.39
	Quantile Spline	*	*	*	*	*
	Quantile Forest	17.28	7.03	3.70	5.77	18.50

Table 2: Univariate Responses Tasks: Performance of different methods in Scenarios 4 and 5, in terms of squared error from the true quantile averaged over 25 independent trials.

variate and multivariate response quantile regression with ReLU networks show an advantage over other quantile regression methods. Quantile regression networks were also shown to outperform ℓ_2 -based regression with the same neural network architecture when the error distribution is heavy-tailed. In the case of multivariate responses, quantile networks were also shown to perform well in benchmarks. A theoretical study of statistical rates of convergence for multivariate response neural quantile regression is left for future work.

A Notation

For an $\epsilon > 0$ and a metric $\text{dist}(\cdot, \cdot)$ on the class of functions \mathcal{F} , we define the covering number $N(\epsilon, \mathcal{F}, \text{dist}(\cdot, \cdot))$ as the minimum number of balls of the form $\{g : \text{dist}(g, f) \leq \epsilon\}$, with $f \in \mathcal{F}$, needed to cover \mathcal{F} .

We also write

$$B(f, \|\cdot\|_{\ell_2}, r) = \{g : \|f - g\|_{\ell_2} \leq r\}.$$

Furrthermore, if a_n and b_n are positive sequences, we say that $a_n \lesssim b_n$ if there exists m such that $n \geq m$

Scenario 6		
n	Method	$\tau = 0.50$
100	SqErr Network	0.92
	Quantile Network	0.49
	Geometric Quantile	0.54
1000	SqErr Network	0.91
	Quantile Network	0.06
	Geometric Quantile	0.07
10000	SqErr Network	0.89
	Quantile Network	0.01
	Geometric Quantile	0.01
Scenario 7		
n	Method	$\tau = 0.50$
100	SqErr Network	0.24
	Quantile Network	1.67
	Geometric Quantile	2.36
1000	SqErr Network	0.11
	Quantile Network	0.23
	Geometric Quantile	0.14
10000	SqErr Network	0.09
	Quantile Network	0.03
	Geometric Quantile	0.03

Table 3: Multivariate Responses Tasks: Performance of different methods in Scenarios 6 and 7.

Note: We measure performance using the averaged mean squared error based on 25 Monte Carlo simulations for the two synthetic multivariate benchmarks.

implies $a_n \leq cb_n$ for a constant $c > 0$.

B Theorem 1

Througouth this section we write \mathcal{F} to refer to $\mathcal{F}(W, U, L)$.

B.1 Auxiliary results

Before stating our first result we first state some definitions and an auxiliary lemma.

Definition 6. We define the empirical loss function

$$\hat{M}_n(\theta) = \sum_{i=1}^n \hat{M}_{n,i}(\theta),$$

where

$$\hat{M}_{n,i}(f) = \frac{1}{n} (\rho_\tau(y_i - f(x_i)) - \rho_\tau(y_i - f_n(x_i))),$$

with

$$f_n \in \arg \min_{f \in \mathcal{F}} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - f(x_i)) - \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - f_\tau^*(x_i)) \right].$$

We also set

$$M_{n,i}(f) = \frac{1}{n} \mathbb{E}[\rho_\tau(z_i - f(x_i)) - \rho_\tau(z_i - f_n(x_i))],$$

where $z \in \mathbb{R}^n$ is an independent copy of y .

Lemma 4. Suppose that Assumption 1–2 hold. Then there exists a constant c_τ such that for all $\delta \in \mathbb{R}^n$, we have

$$\sum_{i=1}^n \mathbb{E}[\rho_\tau(z_i - f_\tau^*(x_i) - \delta_i)] - \sum_{i=1}^n \mathbb{E}[\rho_\tau(z_i - f_\tau^*(x_i))] \geq c_\tau n \Delta_n^2(\delta),$$

where $z \in \mathbb{R}^n$ is an independent copy of y .

Proof. See Lemma 8 in Padilla and Chatterjee (2020). \square

Definition 7. Let \mathcal{H} be a class of functions from \mathcal{X} to \mathbb{R} . We define the pseudodimension of \mathcal{H} , denoted as $\text{Pdim}(\mathcal{H})$, as the largest integer m for which there exist $(a_1, b_1), \dots, (a_m, b_m) \in \mathcal{X} \times \mathbb{R}$ such that for all $\eta \in \{0, 1\}^m$ there exists $f \in \mathcal{H}$ such that

$$f(a_i) > b_i \iff \eta_i,$$

for $i = 1, \dots, m$.

Theorem 5 (Theorem 7 from Bartlett et al. (2019)). *With the notation from before, we have that*

$$\text{Pdim}(\mathcal{F}(W, U, L)) = O(LW \log(U)).$$

B.2 Proof of Theorem 1

Proof. Throughout this proof the covariates x_1, \dots, x_n are fixed. Let $\hat{\delta}_i = \hat{f}(x_i) - f_\tau^*(x_i)$ for $i = 1, \dots, n$. Notice that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n D^2\left\{f_\tau^*(x_i) - \hat{f}(x_i)\right\}\right] &\leq \frac{1}{c_\tau n} \mathbb{E}\left(\sum_{i=1}^n \mathbb{E}\left[\rho_\tau\{z_i - f_\tau^*(x_i) - \hat{\delta}_i\}\right] - \sum_{i=1}^n \mathbb{E}[\rho_\tau\{z_i - f_\tau^*(x_i)\}]\right) \\ &= \frac{1}{c_\tau} E\left\{M_n(\hat{f})\right\} + \frac{1}{c_\tau} \text{err}_1, \end{aligned} \tag{17}$$

where the inequality follows from Lemma 4.

Next we proceed to bound $E\{M_n(\hat{f})\}$. To that end, notice that for a constant $C > 0$,

$$\begin{aligned} \mathbb{E}\left\{M_n(\hat{f})\right\} &\leq 4\mathbb{E}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \xi_i f(x_i)\right\} \\ &\leq F \mathbb{E}\left\{\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \xi_i \frac{f(x_i)}{F}\right\} \\ &\leq \frac{CF}{\sqrt{n}} \int_0^2 \sqrt{\log N(\mu, \mathcal{F}/F, \|\cdot\|_n)} d\mu \\ &\leq \frac{CF}{\sqrt{n}} \int_0^2 \sqrt{\log\left(\left(\frac{2 \cdot e \cdot n}{\mu \cdot \text{Pdim}(\mathcal{F})}\right)^{\text{Pdim}(\mathcal{F})}\right)} d\mu \\ &\leq \tilde{C}F \sqrt{\frac{LW \log U \cdot \log n}{n}} \end{aligned}$$

for some constant $\tilde{C} > 0$, where the first inequality follows by simmetrization and Talgrand's inequality ([Ledoux and Talagrand \(2013\)](#)) similarly to Theorem 12 in [Padilla and Chatterjee \(2020\)](#), the third inequality follows from Dudley's theorem, the fourth holds because of Lemma 4 in [Farrell et al. \(2018\)](#), and the last from Theorem 5.

□

C Theorem 3

The proof is in the spirit of the proof of Theorem 1 in [Farrell et al. \(2018\)](#) combined with results and ideas from [Padilla and Chatterjee \(2020\)](#) and [Suzuki \(2018\)](#).

C.1 Notation

Througout we let $\mathcal{X} = [0, 1]^d$. For $p > 0$ and $f : \mathcal{X} \rightarrow \mathbb{R}$ we let

$$\|f\|_p := \left(\int_{\mathcal{X}} (f(x))^p dx \right)^{1/p}$$

and

$$L^p(\mathcal{X}) = \{f : f : \mathcal{X} \rightarrow \mathbb{R}, \text{ and } \|f\|_p < \infty\}.$$

Definition 8. For a function $f \in L^p(\mathcal{X})$ and $p \in (0, \infty]$ we define the r -modulus of continuity as

$$w_{r,p}(f, t) = \sup_{\|h\|_2 \leq t} \|\Delta_h^r(f)\|_p,$$

with

$$\Delta_h^r(f) = \begin{cases} \sum_{j=0}^r \frac{r!}{j!(r-j)!} (-1)^{r-j} f(x + hj) & \text{if } x \in \mathcal{X}, x + rh \in \mathcal{X}, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 9. For $0 < p, q \leq \infty$, $\alpha > 0$, $r = \lfloor \alpha \rfloor + 1$, we define the Besov space $B_{p,q}^\alpha(\mathcal{X})$ as

$$B_{p,q}^\alpha(\mathcal{X}) = \left\{ f \in L^p(\mathcal{X}) : \|f\|_{B_{p,q}^\alpha(\mathcal{X})} < \infty \right\},$$

where

$$\|f\|_{B_{p,q}^\alpha(\mathcal{X})} = \|f\|_p + |f|_{B_{p,q}^\alpha(\mathcal{X})},$$

with

$$|f|_{B_{p,q}^\alpha(\mathcal{X})} = \begin{cases} \left(\int_0^\infty (t^{-\alpha} w_{r,p}(f, t))^q t^{-1} dt \right)^{\frac{1}{q}} & \text{if } q < \infty, \\ \sup_{t>0} t^{-\alpha} w_{r,p}(f, t) & \text{if } q = \infty. \end{cases}$$

Throughout we denote by $f_n \in \mathcal{I}(L, W, S, B)$ a function satisfying

$$f_n \in \arg \min_{f \in \mathcal{I}(L, W, S, B), \|f\|_\infty \leq F} \|f - f_\tau^*\|_\infty.$$

We also write

$$\tilde{\mathcal{I}}(L, W, S, B) = \{f \in \mathcal{I}(L, W, S, B) : \|f\|_\infty \leq F\}.$$

C.2 Auxiliary lemmas

Lemma 6. Suppose that $\|f_n - f_\tau^*\|_\infty \leq c$ for a small enough constant c . With the notation in (11) we have that

$$\Delta^2(f, f_n) \leq \frac{1}{c_\tau} \left[\mathbb{E}(\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X))) + \|f_n - f_\tau^*\|_\infty \Delta(f, f_n) \sqrt{F} \right],$$

and

$$\|f - f_n\|_{\ell_2}^2 \leq \frac{2F}{c_\tau} \left[\mathbb{E}(\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X))) + \|f_n - f_\tau^*\|_\infty \|f - f_n\|_{\ell_2} \sqrt{F} \right],$$

for any $f \in \tilde{\mathcal{I}}(L, W, S, B)$ and for some constant c_τ .

Proof. Notice that by Equation B.3 in [Belloni and Chernozhukov \(2011\)](#),

$$\begin{aligned} \rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X)) &= -(f(X) - f_n(X))(\tau - 1\{Y \leq f_n(X)\}) + \\ &\quad \int_0^{f(X)-f_n(X)} [1\{Y \leq f_n(X) + z\} - 1\{Y \leq f_n(X)\}] dz \\ &= -(f(X) - f_n(X))(\tau - 1\{Y \leq f_\tau^*(X)\}) - \\ &\quad (f(X) - f_n(X))(1\{Y \leq f_\tau^*(X)\} - 1\{Y \leq f_n(X)\}) + \\ &\quad \int_0^{f(X)-f_n(X)} [1\{Y \leq f_n(X) + z\} - 1\{Y \leq f_n(X)\}] dz. \end{aligned}$$

Hence, taking expectations and using Fubini's theorem,

$$\begin{aligned} \mathbb{E}(\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X))) &= \mathbb{E}\left(-(f(X) - f_n(X))\mathbb{E}\left((\tau - 1\{Y \leq f_\tau^*(X)\}) \middle| X\right)\right) - \\ &\quad \mathbb{E}\left((f(X) - f_n(X))\mathbb{E}\left((1\{Y \leq f_\tau^*(X)\} - 1\{Y \leq f_n(X)\}) \middle| X\right)\right) + \\ &\quad \mathbb{E}\left(\int_0^{f(X)-f_n(X)} \left[\mathbb{E}\left(1\{Y \leq f_n(X) + z\} \middle| X\right) - \right.\right. \\ &\quad \left.\left.\mathbb{E}\left(1\{Y \leq f_n(X)\} \middle| X\right)\right] dz\right) \\ &\geq -c_1 \mathbb{E}[|f(X) - f_n(X)| \cdot |f_\tau^*(X) - f_n(X)|] + \\ &\quad c_\tau \mathbb{E}(D^2(f(X) - f_n(X))) \\ &\geq -c_1 \sqrt{\mathbb{E}[|f(X) - f_n(X)|^2]} \sqrt{\mathbb{E}[|f_\tau^*(X) - f_n(X)|^2]} + \\ &\quad c_\tau \mathbb{E}(D^2(f(X) - f_n(X))) \\ &\geq -c_1 \|f_n - f_\tau^*\|_\infty \sqrt{F \Delta^2(f, f_n)} \\ &\quad c_\tau \mathbb{E}(D^2(f(X) - f_n(X))) \end{aligned}$$

for a constant $c_1 > 0$, where the first inequality holds since the cumulative distribution function of Y conditioning on X is Lipschitz around $f_\tau^*(X)$ by Assumption 2, and by the same argument from the proof of Lemma 8 in [Padilla and Chatterjee \(2020\)](#). □

Definition 10. We define the empirical loss function

$$\hat{M}_n(\theta) = \sum_{i=1}^n \hat{M}_{n,i}(\theta),$$

where

$$\hat{M}_{n,i}(f) = \frac{1}{n} [\rho_\tau\{y_i - f(x_i)\} - \rho_\tau\{y_i - f_n(x_i)\}],$$

and we set

$$M_n(f) = \mathbb{E}[\rho_\tau\{Y - f(X)\} - \rho_\tau\{Y - f_n(X)\}].$$

Lemma 7. Suppose that $\|f_n - f_\tau^*\|_\infty \leq c$ for a small enough constant c . The estimator \hat{f} defined in (16) satisfies

$$\Delta^2(\hat{f}, f_n) \leq \frac{1}{c_\tau} \left[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \Delta(\hat{f}, f_n) \sqrt{F} \right].$$

Furthermore,

$$\|\hat{f} - f_n\|_{\ell_2}^2 \leq \frac{2F}{c_\tau} \left[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} \sqrt{F} \right].$$

Proof. By Lemma 6 we have that

$$\begin{aligned} \Delta^2(\hat{f}, f_n) &\leq \frac{1}{c_\tau} \left[\mathbb{E} \left(\rho_\tau(Y - \hat{f}(X)) - \rho_\tau(Y - f_n(X)) \right) + \|f_n - f_\tau^*\|_\infty \Delta(\hat{f}, f_n) \sqrt{F} \right] \\ &\leq \frac{1}{c_\tau} \left[\mathbb{E} \left(\rho_\tau(Y - \hat{f}(X)) - \rho_\tau(Y - f_n(X)) \right) + \|f_n - f_\tau^*\|_\infty \Delta(\hat{f}, f_n) \sqrt{F} - \right. \\ &\quad \left. \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \hat{f}(x_i)) + \sum_{i=1}^n \rho_\tau(y_i - f_n(x_i)) \right], \end{aligned}$$

where the last inequality follows by the optimality of \hat{f} . \square

Lemma 8. Suppose that

$$3\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_{\ell_2}^2 \leq r^2} \frac{1}{n} \sum_{i=1}^n \xi_i(f(x_i) - f_n(x_i))^2 \right) \leq r^2, \quad (18)$$

for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$, and

$$\max \left\{ 4F \sqrt{\frac{\gamma}{n}}, 4F \sqrt{\frac{\gamma}{3n}} \right\} \leq r. \quad (19)$$

Then with probability at least $1 - e^{-\gamma}$, $\|f - f_n\|_{\ell_2}^2 \leq r^2$ with $f \in \tilde{\mathcal{I}}(L, W, S, B)$ implies

$$\|f - f_n\|_n^2 \leq (2r)^2.$$

Proof. First notice that

$$|(f(x) - f_n(x))^2| \leq 2 [F^2 + \|f_n\|_\infty^2],$$

for all x . Hence,

$$\begin{aligned} \mathbb{E}((f(X) - f_n(X))^4) &\leq 2(F^2 + \|f_n\|_\infty^2) \mathbb{E}((f(X) - f_n(X))^2) \\ &\leq 4F^2 \|f - f_n\|_{\ell_2}^2 \end{aligned}$$

Then by Theorem 2.1 in [Bartlett et al. \(2005\)](#) with probability at least $1 - \exp(-\gamma)$,

$$\begin{aligned} &\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_{\ell_2}^2 \leq r^2} \left\{ \|f - f_n\|_n^2 - \|f - f_n\|_{\ell_2}^2 \right\} \\ &\leq 3\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_{\ell_2}^2 \leq r^2} \frac{1}{n} \sum_{i=1}^n \xi_i(f(x_i) - f_n(x_i))^2 \right) \\ &\quad + 4rF \sqrt{\frac{\gamma}{n}} + \frac{16F^2\gamma}{3n} \end{aligned} \quad (20)$$

and the claim follows. \square

Lemma 9. Suppose that $\|\hat{f} - f_n\|_{\ell_2} \leq r_0$, with r_0 satisfying (18)-(19) and Assumption 5 holds. Also, with the notation of Assumption 4, suppose that for the class $\tilde{\mathcal{I}}(L, W, S, B)$ the parameters are chosen as

$$L = 3 + 2 \lceil \log_2 \left(\frac{3^{\max\{d,m\}}}{\epsilon c_{d,m}} \right) + 5 \rceil \lceil \log_2 \max\{d, m\} \rceil, \quad W = W_0 N,$$

$$S = (L - 1)W_0^2 N + N, \quad B = O \left(N^{(v^{-1} + d^{-1})(\max\{1, (d/p-s)_+\})} \right),$$

for a constant $c_{d,m}$ that depends on d and m , a constant W_0 , and where $v = (s - \delta)/\delta$,

$$\delta = \frac{d}{p}, \quad N \asymp n^{\frac{d}{2s+d}}.$$

Then for some positive constant C_0 it holds that

$$\begin{aligned} \|\hat{f} - f_n\|_{\ell_2}^2 &\leq C_0 \left[r_0 F^{2.5} \sqrt{\frac{\gamma}{n}} + \frac{F^{2.5}\gamma}{n} + \right. \\ &\quad \left. r_0 F \sqrt{\frac{N(\log N)^2}{n}} + r_0 F \sqrt{\frac{N[(\log N)^2 + \log r_0^{-1} + \log n]}{n}} + N^{-s/d} r_0 F^{1.5} \right] \end{aligned}$$

with probability at least $1 - \exp(-\gamma)$, where $N \asymp n^{\frac{d}{2s+d}}$.

Proof. Let

$$\mathcal{G} = \left\{ g : g(x, y) = \rho_\tau(y - f(x)) - \rho_\tau(y - f_n(x)), \quad f \in \tilde{\mathcal{I}}(L, W, S, B), \quad \|f - f_n\|_{\ell_2} \leq r_0 \right\}.$$

Then for ξ_1, \dots, ξ_n independent Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$, we have that

$$\begin{aligned} c_\tau \|\hat{f} - f_n\|_{\ell_2}^2 &\leq 2F[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} \sqrt{F}] \\ &\leq 2F \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}(g(X, Y)) - \frac{1}{n} \sum_{i=1}^n g(x_i, y_i) \right\} + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} F^{3/2} \\ &\leq 12F \mathbb{E} \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \middle| (x_1, y_1), \dots, (x_n, y_n) \right) \\ &\quad + 4r_0 F^{2.5} \sqrt{\frac{\gamma}{n}} + \frac{100F^{2.5}\gamma}{3n} + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} F^{1.5}, \end{aligned} \tag{21}$$

where the first inequality follows from Lemmas 7, and the third happens with probability at least $1 - e^{-\gamma}$ and holds by Theorem 2.1 in Bartlett et al. (2005).

Next, notice that for a constant $C > 0$,

$$\begin{aligned} \mathbb{E}_\xi \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \right) &\leq \mathbb{E}_\xi \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f\|_\infty \leq F, \|f - f_n\|_{\ell_2} \leq r_0} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i)) \right) \\ &\leq \mathbb{E}_\xi \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f\|_\infty \leq F, \|f - f_n\|_n \leq 2r_0} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i)) \right) \\ &\leq \inf_{0 < \alpha < 2r_0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{2r_0} \sqrt{\log \mathcal{N}(\delta, \tilde{\mathcal{I}}(L, W, S, B), \|\cdot\|_n)} d\delta \right\} \\ &\leq \inf_{0 < \alpha < 2r_0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{2r_0} \sqrt{\log \mathcal{N}(\delta, \tilde{\mathcal{I}}(L, W, S, B), \|\cdot\|_\infty)} d\delta \right\} \\ &\leq \inf_{0 < \alpha < r_0} \left\{ 4\alpha + \frac{24r_0}{\sqrt{n}} \sqrt{\log \mathcal{N}(\alpha, \tilde{\mathcal{I}}(L, W, S, B), \|\cdot\|_\infty)} \right\}, \\ &\leq C \inf_{0 < \alpha < r_0} \left\{ \alpha + r_0 \sqrt{\frac{N[(\log N)^2 + \log \alpha^{-1}]}{n}} \right\}, \end{aligned} \tag{22}$$

where the first inequality follows by Talagrand's inequality (Ledoux and Talagrand, 2013), and the second holds with probability at least $1 - \exp(-\gamma)$ by Lemma 8, the third by Dudley's chaining inequality, and the last by the proof Theorem 2 in Suzuki (2018). The latter theorem also gives $N \asymp n^{\frac{d}{2s+d}}$, and

$$\|f_n - f_\tau^*\|_\infty \leq C_1 N^{-s/d},$$

for some constant $C_1 > 0$. Hence, taking

$$\alpha = r_0 \sqrt{\frac{N(\log N)^2}{n}},$$

(21) and (22) imply

$$\begin{aligned} \|\hat{f} - f_\tau^*\|_{\ell_2}^2 &\leq \frac{1}{c_\tau} \left[4r_0 F^{2.5} \sqrt{\frac{\gamma}{n}} + \frac{100F^{2.5}\gamma}{3n} + \right. \\ &\quad \left. Cr_0 F \sqrt{\frac{N(\log N)^2}{n}} + Cr_0 F \sqrt{\frac{N[(\log N)^2 + \log r_0^{-1} + \log n]}{n}} + C_1 r_0 N^{-s/d} F^{1.5} \right] \end{aligned}$$

with probability at least $1 - 2e^{-\gamma}$. \square

Lemma 10. Let r^* be defined as

$$r^* = \inf \left\{ r > 0 : 3\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_{\ell_2} \leq s} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \right) < s^2, \forall s \geq r \right\},$$

for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$. Then under the conditions of Lemma 9,

$$r^* \leq \tilde{C} \left[\sqrt{\frac{N(\log N)^2}{n}} + \sqrt{\frac{N[(\log N)^2 + \log n]}{n}} \right],$$

for a constant $\tilde{C} > 0$ and with N satisfying $N \asymp n^{\frac{d}{2s+d}}$.

Proof. Consider the set

$$\mathcal{G}_{r^*} = \left\{ f \in \tilde{\mathcal{I}}(L, W, S, B) : \|f - f_n\|_{\ell_2}^2 \leq (r^*)^2 \right\},$$

and define the event

$$E = \left\{ \sup_{f \in \mathcal{G}_{r^*}} \|f - f_n\|_n^2 \leq (2r^*)^2 \right\}.$$

If r^* satisfies (19) with $\gamma = \log n$ then we have that $\mathbb{P}(E) \geq 1 - 1/n$ by Lemma 8.

Also,

$$\begin{aligned} (r^*)^2 &\leq 3\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_{\ell_2}^2 \leq (r^*)^2} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \right) \\ &\leq 3\mathbb{E} \left(\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_n^2 \leq (2r^*)^2} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \middle| x_1, \dots, x_n \right) 1_E \right) + \\ &\quad \frac{12F^2}{n} \\ &\leq 3\mathbb{E} \left(\mathbb{E} \left(\sup_{f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_n^2 \leq (2r^*)^2} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \middle| x_1, \dots, x_n \right) 1_E \right) + \\ &\quad \frac{12F^2}{n}. \end{aligned} \tag{23}$$

Next, let

$$\mathcal{G} = \left\{ g : g(x) = (f(x) - f_n(x))^2, \text{ for some } f \in \tilde{\mathcal{I}}(L, W, S, B), \|f - f_n\|_n \leq 2r^* \right\}.$$

Notice that if $g_1, g_2 \in \mathcal{G}$ with $g_j = f_j - f_n$, $j = 1, 2$, then

$$|g_1(x) - g_2(x)| = |f_1(x) - f_2(x)| \cdot |f_1(x) + f_2(x) - 2f_n(x)| \leq 4F\|f_1 - f_2\|_\infty.$$

Hence, combining this with (23), using Dudley's chaining we obtain that

$$\begin{aligned} (r^*)^2 &\leq 3\mathbb{E} \left(\inf_{0 < \alpha < 2r^*} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r^*} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_n)} d\alpha \right\} \right) + \frac{12F}{n} \\ &\leq 3\mathbb{E} \left(\inf_{0 < \alpha < 2r^*} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r^*} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty)} d\alpha \right\} \right) + \frac{12F}{n} \\ &\leq C \inf_{0 < \alpha < 2r^*} \left\{ \alpha + r^* \sqrt{\frac{N[(\log N)^2 + \log \alpha^{-1}]}{n}} \right\} + \frac{12F}{n}, \end{aligned} \quad (24)$$

where the last inequality follows from Theorem 2 in [Suzuki \(2018\)](#), and with N satisfying $N \asymp n^{\frac{d}{2s+d}}$. Hence, if r^* satisfies (19) then for a constant $\tilde{C} > 0$

$$r^* \leq \tilde{C} \left[\sqrt{\frac{N(\log N)^2}{n}} + \sqrt{\frac{N[(\log N)^2 + \log n]}{n}} \right],$$

which follows from (24) by taking

$$\alpha = r^* \sqrt{\frac{N(\log N)^2}{n}}.$$

The claim follows. \square

C.3 Proof of Theorem 3

Proof. Throughout we use the notation from the proof of Lemma 9. Then we proceed as in [Farrell et al. \(2018\)](#). Specifically, we divide the space $\tilde{\mathcal{I}}(L, W, S, B)$ into sets of increasing radius

$$\mathbf{B}(f_n, \|\cdot\|_{\ell_2}, \bar{r}), \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, 2\bar{r}) \setminus \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, \bar{r}), \dots, \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, 2^l\bar{r}) \setminus \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, 2^{l-1}\bar{r}),$$

where

$$l = \left\lfloor \log_2 \left(\frac{2F}{\sqrt{(\log n)/n}} \right) \right\rfloor.$$

Next, if $\bar{r} > r^*$, the by Lemma 8, with probability at least $1 - le^{-\gamma}$, we have that

$$\|f - f_n\|_{\ell_2}^2 \leq 2^j\bar{r} \text{ implies } \|f - f_n\|_n^2 \leq 2^{j+1}\bar{r}.$$

Then if for some $j \leq l$ it holds that

$$\hat{f} \in \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, 2^j\bar{r}) \setminus \mathbf{B}(f_n, \|\cdot\|_{\ell_2}, 2^{j-1}\bar{r}),$$

then by Lemma 9, with probability at least $1 - 4e^{-\gamma}$, we have that

$$\begin{aligned} \|\hat{f} - f_n\|_{\ell_2}^2 &\leq \tilde{C} \left[2^j \bar{r} F^{2.5} \sqrt{\frac{\gamma}{n}} + \frac{F^{2.5} \gamma}{n} + \right. \\ &\quad \cdot 2^j \bar{r} F \sqrt{\frac{N(\log N)^2}{n}} + \cdot 2^j \bar{r} F \sqrt{\frac{N[(\log N)^2 + 2 \log n]}{n}} + 2^j \bar{r} N^{-s/d} F^{1.5} \left. \right] \\ &\leq 2^{2j-2} \bar{r}^2, \end{aligned}$$

provided that

$$\tilde{C} \left[F^{2.5} \sqrt{\frac{\gamma}{n}} + F \sqrt{\frac{N(\log N)^2}{n}} + F \sqrt{\frac{N[(\log N)^2 + 2 \log n]}{n}} + N^{-s/d} F^{1.5} \right] \leq \frac{1}{8} 2^j \bar{r},$$

and

$$\tilde{C} \frac{F^{2.5} \gamma}{n} \leq \frac{1}{4} 2^{2j} \bar{r}^2,$$

both of which for all j hold if

$$\bar{r} = 8\tilde{C} \left[F^{2.5} \sqrt{\frac{\gamma}{n}} + D \sqrt{\frac{N(\log N)^2}{n}} + D \sqrt{\frac{N[(\log N)^2 + 2 \log n]}{n}} + N^{-s/d} F^{1.5} \right] + 2\sqrt{\frac{\tilde{C} F^{2.5} \gamma}{n}} + r^*. \quad (25)$$

Therefore, by Lemmas 8–9, with probability at least $1 - e^{-\gamma}$, we have that

$$\|\hat{f} - f_n\|_{\ell_2} \leq 2^l \bar{r}, \quad \text{and} \quad \|\hat{f} - f_n\|_n \leq 2^{l+1} \bar{r},$$

which by the previous argument implies that

$$\|\hat{f} - f_n\|_{\ell_2} \leq 2^{l-1} \bar{r}, \quad \text{and} \quad \|\hat{f} - f_n\|_n \leq 2^l \bar{r},$$

and continuing recursively we arrive at

$$\|\hat{f} - f_n\|_{\ell_2} \leq \bar{r}, \quad \text{and} \quad \|\hat{f} - f_n\|_n \leq 2\bar{r}.$$

The claim follows by noticing that Lemma 10 and (25) imply that

$$\begin{aligned} \bar{r} &\leq 8\tilde{C} \left[F^{2.5} \sqrt{\frac{\gamma}{n}} + F \sqrt{\frac{N(\log N)^2}{n}} + F \sqrt{\frac{N[(\log N)^2 + 2 \log n]}{n}} + N^{-s/d} F^{1.5} \right] + 2\sqrt{\frac{\tilde{C} F^{2.5} \gamma}{n}} + \\ &\quad \tilde{C} \left[\sqrt{\frac{N(\log N)^2}{n}} + \sqrt{\frac{N[(\log N)^2 + 2 \log n]}{n}} \right], \end{aligned}$$

and the claim follows since $\|f_n - f^*\|_\infty \leq N^{-s/d}$. \square

D Proof of Theorem 2

The proof is similar to that of Theorem 3 relying in the lemmas given next.

Just as in Lemma 6, we have the following result.

Lemma 11. Suppose that $\|f_n - f_\tau^*\|_\infty \leq c$ for a small enough constant c . Then

$$\Delta^2(f, f_n) \leq \frac{1}{c_\tau} \left[\mathbb{E}(\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X))) + \|f_n - f_\tau^*\|_\infty \Delta(f, f_n) \sqrt{F} \right],$$

and

$$\|f - f_n\|_{\ell_2}^2 \leq \frac{2F}{c_\tau} \left[\mathbb{E}(\rho_\tau(Y - f(X)) - \rho_\tau(Y - f_n(X))) + \|f_n - f_\tau^*\|_\infty \|f - f_n\|_{\ell_2} \sqrt{F} \right],$$

for any $f \in \mathcal{G}(L, p, S, F)$, and for some constant c_τ .

Similarly to Lemmas 8–10, we have the following three lemmas.

Lemma 12. Suppose that $\|f_n - f_\tau^*\|_\infty \leq c$ for a small enough constant c . The estimator \hat{f} satisfies

$$\Delta^2(\hat{f}, f_n) \leq \frac{1}{c_\tau} \left[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \Delta(\hat{f}, f_n) \sqrt{F} \right].$$

Furthermore,

$$\|\hat{f} - f_n\|_{\ell_2}^2 \leq \frac{2F}{c_\tau} \left[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} \sqrt{F} \right].$$

Lemma 13. If

$$3\mathbb{E} \left(\sup_{f \in \mathcal{G}(L, p, S, F), \|f - f_n\|_{\ell_2} \leq r} \frac{1}{n} \sum_{i=1}^n \xi_i (f(X_i) - f_n(X_i))^2 \right) \leq r^2, \quad (26)$$

for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$, and

$$\max \left\{ 4F \sqrt{\frac{\gamma}{n}}, 4F \sqrt{\frac{\gamma}{3n}} \right\} \leq r. \quad (27)$$

Then with probability at least $1 - e^{-\gamma}$, $\|f - f_n\|_{\ell_2}^2 \leq r^2$ with $f \in \mathcal{G}(L, p, S, F)$ implies

$$\|f - f_n\|_n^2 \leq (2r)^2.$$

Lemma 14. Suppose that $\|\hat{f} - f_n\| \leq r_0$, with r_0 satisfying (26)–(27) and Assumption 3 holds. Also, with the notation of Definition 4, suppose that for the class $\mathcal{G}(L, p, S, F)$ the parameters are chosen to satisfy

$$\begin{aligned} \sum_{i=0}^q \log_2 (4 \max\{t_i, \beta_i\}) \log_2(n) &\leq L \lesssim n\epsilon_n, \quad \max\{1, K\} \leq F, \\ n\epsilon_n &\lesssim \min_{i=1, \dots, L} p_i, \quad S \asymp n\epsilon_n \log n, \quad \max_{i=1, \dots, L} p_i \lesssim n, \end{aligned}$$

where

$$\epsilon_n = \max_{i=0,1,\dots,q} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}}.$$

Then

$$\begin{aligned} \|\hat{f} - f_n\|_{\ell_2}^2 &\lesssim r_0 \sqrt{F\epsilon_n \log n [\log L + L \log n]} + r_0 \sqrt{F\epsilon_n \log n [\log L + L \log n + \log(n)]} \\ &\quad + r_0 F^2 \sqrt{\frac{\gamma}{n}} + \frac{F^2 \gamma}{n} + r_0 F \max_{i=0,1,\dots,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \end{aligned}$$

with probability at least $1 - 2 \exp(-\gamma)$.

Proof. Let

$$\mathcal{G} = \{g : g(x, y) = \rho_\tau(y - f(x)) - \rho_\tau(y - f_n(x)), \quad f \in \mathcal{G}(L, p, S, F), \quad \|f - f_n\| \leq r_0\}.$$

Prooceeding as in the proof of Lemma 9, we obtain that

$$\mathbb{E}_\xi \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \right) \leq C \inf_{0 < \alpha < r_0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r_0} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}(L, p, S, F), \|\cdot\|_n)} d\delta \right\}, \quad (28)$$

for some positive constant C .

However, defining $V = \prod_{l=0}^{L+1} (p_l + 1)$, then Lemma 5 in Schmidt-Hieber (2017) and (28) imply that

$$\begin{aligned} \mathbb{E}_\xi \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \right) &\leq C \inf_{0 < \alpha < r_0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r_0} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}(L, p, S, F), \|\cdot\|_\infty)} d\delta \right\} \\ &\leq 4C\alpha + \frac{24Cr_0}{\sqrt{n}} \sqrt{(S+1) \log(2\alpha^{-1}(L+1)V^2)}, \end{aligned} \quad (29)$$

where $V = \prod_{l=0}^{L+1} (p_l + 1)$. Therefore, setting

$$\alpha = r_0 \sqrt{\frac{(S+1) \log((L+1)V^2)}{n}},$$

(29) implies

$$\mathbb{E}_\xi \left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \right) \leq 4Cr_0 \sqrt{\frac{(S+1) \log((L+1)V^2)}{n}} + \frac{24r_0C}{\sqrt{n}} \sqrt{(S+1) \log(2(L+1)V^2n)}.$$

Hence, as in (21), we obtain that

$$\begin{aligned} \|\hat{f} - f_n\|_{\ell_2}^2 &\leq \frac{2\sqrt{F}}{c_\tau} \left[24r_0C \sqrt{\frac{F(S+1) \log((L+1)V^2)}{n}} + \frac{144\sqrt{F}r_0C}{\sqrt{n}} \sqrt{(S+1) \log(2(L+1)V^2n)} + \right. \\ &\quad \left. + 4r_0F^2 \sqrt{\frac{\gamma}{n}} + \frac{100F^2\gamma}{3n} + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} F \right], \end{aligned} \quad (30)$$

with probability at least $1 - e^{-\gamma}$.

Furthermore, by Equation (26) in the proof of Theorem 1 in Schmidt-Hieber (2017) and the argument therein, we have that

$$\|f_n - f_\tau^*\|_\infty \leq C' \max_{i=0,1,\dots,q} c^{-\frac{\beta_i^*}{t_i}} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}}, \quad (31)$$

for positive constants c and C' . Hence, combining (30) with (31) we arrive at

$$\begin{aligned} \|\hat{f} - f_n\|_{\ell_2}^2 &\leq \frac{s\sqrt{F}}{c_\tau} \left[24r_0C \sqrt{\frac{F(S+1) \log((L+1)V^2)}{n}} + \frac{144\sqrt{F}r_0C}{\sqrt{n}} \sqrt{(S+1) \log(2(L+1)V^2n)} + \right. \\ &\quad \left. + 4r_0F^2 \sqrt{\frac{\gamma}{n}} + \frac{100F^2\gamma}{3n} + r_0F C' \max_{i=0,1,\dots,q} c^{-\frac{2\beta_i^*}{t_i}} n^{-\frac{2\beta_i^*}{2\beta_i^* + t_i}} \right], \\ &\lesssim r_0 \sqrt{F\epsilon_n \log n [\log L + \log V]} + r_0 \sqrt{F\epsilon_n \log n [\log L + \log V + \log(n)]} \\ &\quad + r_0F^2 \sqrt{\frac{\gamma}{n}} + \frac{F^2\gamma}{n} + r_0F \max_{i=0,1,\dots,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \\ &\lesssim r_0 \sqrt{F\epsilon_n \log n [\log L + L \log n]} + r_0 \sqrt{F\epsilon_n \log n [\log L + L \log n + \log(n)]} \\ &\quad + r_0F^2 \sqrt{\frac{\gamma}{n}} + \frac{F^2\gamma}{n} + r_0F \max_{i=0,1,\dots,q} n^{-\frac{\beta_i^*}{2\beta_i^* + t_i}} \end{aligned}$$

where in the last four inequalities we have used the choice of the network parameters. \square

Lemma 15. Let r^* be defined as

$$r^* = \inf \left\{ r > 0 : 3\mathbb{E} \left(\sup_{f \in \mathcal{G}(L, p, S, F), \|f - f_n\|_{\ell_2} \leq s} \frac{1}{n} \sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \right) < s^2, \forall s \geq r \right\},$$

for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$. Then under the conditions of Lemma 14,

$$r^* \leq \tilde{C} \epsilon_n L \log^2 n,$$

for a constant $\tilde{C} > 0$.

Proof. Proceeding as in the proof of Lemma 10, we obtain that for a constant $C > 0$,

$$\begin{aligned} (r^*)^2 &\leq C \inf_{0 < \alpha < 2r^*} \left\{ \alpha + \frac{r^*}{\sqrt{n}} \sqrt{\log N(\alpha, \mathcal{G}(L, p, S, F), \|\cdot\|_\infty)} \right\} + \frac{12F^2}{n} \\ &\leq C \inf_{0 < \alpha < 2r^*} \left\{ \alpha + \frac{r^*}{\sqrt{n}} \sqrt{(S+1) \log(2\alpha^{-1}(L+1)V^2)} \right\} + \frac{12F^2}{n} \\ &\lesssim \inf_{0 < \alpha < 2r^*} \left\{ \alpha + r^* \sqrt{\epsilon_n \log n [L \log n + \log \alpha^{-1}]} \right\} + \frac{12F^2}{n} \end{aligned}$$

where the second inequality follows from Lemma 5 in Schmidt-Hieber (2017), and the second by the choice of the parameters in the network. Hence, setting

$$\alpha = r^* \sqrt{\epsilon_n L \log^2 n},$$

we obtain that

$$r^* \lesssim \sqrt{\epsilon_n L \log^2 n}.$$

\square

References

- B Abdous and R Theodorescu. Note on the spatial quantile of a random vector. *Statistics & Probability Letters*, 13(4):333–336, 1992.
- Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.
- G Jogesh Babu and C Radhakrishna Rao. Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population. In *Multivariate Statistics and Probability*, pages 15–23. Elsevier, 1989.
- Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.
- Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.
- Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

- Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20:63–1, 2019.
- Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.
- Alexandre Belloni and Victor Chernozhukov. ℓ_1 -penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.
- Lawrence D Brown, T Tony Cai, Harrison H Zhou, et al. Robust nonparametric estimation via wavelet median regression. *Annals of Statistics*, 36(5):2055–2084, 2008.
- Alex J Cannon. Quantile regression neural networks: Implementation in r and application to precipitation downscaling. *Computers & Geosciences*, 37(9):1277–1284, 2011.
- Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996.
- Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.
- Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.
- Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.
- Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.
- Ronald A DeVore and Vasil A Popov. Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, 305(1):397–414, 1988.
- David L Donoho, Iain M Johnstone, et al. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.
- Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.
- Yijia Feng, Runze Li, Agus Sudjianto, and Yiyun Zhang. Robust neural network with applications to credit portfolio data analysis. *Statistics and its Interface*, 3(4):437, 2010.
- Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.
- Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Artificial Intelligence and Statistics*, pages 315–323, 2011.

Michael Hamers and Michael Kohler. Nonasymptotic bounds on the l_2 error of neural network regression estimates. *Annals of the Institute of Statistical Mathematics*, 58(1):131–151, 2006.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pages 164–171, 1993.

Kostas Hatalis, Alberto J Lamadrid, Katya Scheinberg, and Shalinee Kishore. Smooth pinball neural network for probabilistic forecasting of wind power. *arXiv preprint arXiv:1710.01720*, 2017.

Xuming He and Pin Ng. Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, 75(2):343–352, 1999.

Xuming He and Peide Shi. Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3(3-4):299–308, 1994.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Kurt Hornik, Maxwell Stinchcombe, Halbert White, and Peter Auer. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6(6):1262–1275, 1994.

Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016a.

Jason M Klusowski and Andrew R Barron. Uniform approximation by neural networks activated by first and second order ridge splines. *arXiv preprint arXiv:1607.07819*, 2016b.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

Michael Kohler and Adam Krzyżak. Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics*, 17(8):891–913, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer Science & Business Media, 2013.

Shiyu Liang and Rayadurgam Srikant. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

Joram Lindenstrauss and Lior Tzafriri. *Classical Banach spaces II: Function spaces*, volume 97. Springer Science & Business Media, 2013.

- Daniel F McCaffrey and A Ronald Gallant. Convergence rates for single hidden layer feedforward networks. *Neural Networks*, 7(1):147–158, 1994.
- Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.
- Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning*, 2010.
- Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.
- Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. Adaptive quantile trend filtering. [arXiv preprint arXiv:2007.07472](#), 2020.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.
- Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.
- Filipe Rodrigues and Francisco C Pereira. Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. [arXiv preprint arXiv:1708.06633](#), 2017.
- Taiji Suzuki. 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](#), 2018.
- James W Taylor. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4):299–311, 2000.
- Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. In *Advances in Neural Information Processing Systems*, pages 9089–9100, 2019.
- Yehuda Vardi and Cun-Hui Zhang. The multivariate l_1 -median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.
- Halbert White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989.
- Halbert White. Nonparametric estimation of conditional quantiles using neural networks. In *Computing Science and Statistics*, pages 190–199. Springer, 1992.
- Qifa Xu, Xi Liu, Cuixia Jiang, and Keming Yu. Quantile autoregression neural network model with applications to evaluating value at risk. *Applied Soft Computing*, 49:1–12, 2016.
- Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Wenjie Zhang, Hao Quan, and Dipti Srinivasan. An improved quantile regression neural network for probabilistic load forecasting. *IEEE Transactions on Smart Grid*, 10(4):4425–4434, 2018.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Extending deep learning models for limit order books to quantile regression. *arXiv preprint arXiv:1906.04404*, 2019.