Journal of Machine Learning Research 23 (2022) 1-42 Submitted 3/21; Revised 7/22; Published 8/22

# Quantile regression with ReLU Networks: Estimators and minimax rates

**Oscar Hernan Madrid Padilla** OSCAR.MADRID@STAT.UCLA.EDU
*Department of Statistics*
*Univeristy of California, Los Angeles*
*520 Portola Plaza, Los Angeles, California, USA*

**Wesley Tansey** TANSEYW@MSKCC.ORG
*Computational Oncology Department of Epidemiology and Biostatistics*
*1275 York Avenue New York, NY 10065*

**Yanzhen Chen** IMYANZHEN@UST.HK
*Department of Information Systems, Business Statistics and Operations Management*
*Hong Kong University of Science and Technology*
*Clear Water Bay, Kowloon, Hong Kong*

**Editor:** Pradeep Ravikumar

## Abstract

Quantile regression is the task of estimating a specified percentile response, such as the median ($50^{\text{th}}$ 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 conditioning 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,

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

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 $5^{\text{th}}$ or $95^{\text{th}}$ percentile outcome. Estimating the $5^{\text{th}}$, $50^{\text{th}}$ (median), $95^{\text{th}}$, 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 $f_\tau^* : \mathbb{R}^d \to \mathbb{R}$ given as

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

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 $\tau^{\text{th}}$ 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 $\tau$ corresponds to the $(\tau \times 100)^{\text{th}}$ percentile response.

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

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

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 $x_1, \ldots, x_n$. 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 (see Schmidt-Hieber (2020)). 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 $(X, Y)$. Consequently, quantile regression with ReLU networks can be directly applied to models with heavy-tailed error distributions.

- Suppose that $f_\tau^*(\cdot)$ belongs to a Besov space $B^s_{p,q}([0,1]^d)$, 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 in balls of the space $B^s_{p,q}([0,1]^d)$. 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.

In attaining these results, we highlight the novelty of the assumptions and the proof argument discussed in Section 5. Additionally, we propose a novel method for estimating multiple quantiles simultaneously using ReLU networks while enforcing noncrossing constraints. Our experiments suggest that this proposed approach is especially useful for estimating quantiles close to 0 and 1.

### 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 space 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. These works are 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 were 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 (2020) for classes of functions constructed from Hölder functions. For such classes, the corresponding approximation results were exploited by Schmidt-Hieber (2020) 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 intrinsict 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 $W_1^1([0,1]^d)$ ($\ell_1$-integrable functions with domain $[0,1]^d$ and $\ell_1$-integrable first order partial derivatives). In a related work, Chen et al. (2020) considered quantile treatment effect estimation. Despites 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 $\phi_v : \mathbb{R}^r \to \mathbb{R}^r$ 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 $\phi : \mathbb{R} \to \mathbb{R}$ given as $\phi(x) = \max\{x, 0\}$ is the ReLU activation function. By convention, when $v = 0$ we write $\phi$ to denote $\phi_v$. With this notation, we consider neural network functions $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ of the form

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

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

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

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

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 $\tau_0 < \ldots < \tau_m$ be the elements of $\Lambda$, we solve

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

and set

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

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 $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^{d'}$, with $d' > 1$. Furthermore, consider the Euclidean unit ball in $\mathbb{R}^{d'}$, namely $B^{(d')} = \{u \in \mathbb{R}^{d'} \,:\, \|u\| \leq 1\}$, with $\|\cdot\|$ the usual Euclidean norm. Chaudhuri (1996) defines the function $\Psi(\cdot,\cdot) : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$,

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

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

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

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

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

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,

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

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

Marginals quantile 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 $f_\tau^* : \mathbb{R}^d \to \mathbb{R}^{d'}$, $f_\tau^*(x) = (f_{\tau,1}^*(x), \ldots, f_{\tau,d'}^*(x))^\top$, where

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

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

$$(\hat{f}_{\tau,1}, \cdots, \hat{f}_{\tau,d'})^\top = \underset{f=(f_1,\ldots,f_{d'})^\top \in \mathcal{F}}{\arg\min} \sum_{j=1}^{d'} \sum_{i=1}^{n} \rho_\tau(y_{i,j} - f_j(x_i)). \tag{9}$$

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

## 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 $f, g : \mathbb{R}^d \to \mathbb{R}$, we define the function $\Delta_n^2(f, g)$ as

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

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

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

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 $f, g : [0,1]^d \to \mathbb{R}$, we define $\Delta^2(f, g)$ as $\Delta^2(f, g) := \mathbb{E}\left(D^2(f(X) - g(X))\right)$, and set $\Delta(f, g) := \sqrt{\Delta^2(f, g)}$. We also write $\|f - g\|_{\ell_2} :=$

$\sqrt{\mathbb{E}\left((f(X)-g(X))^2\right)}$, and

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

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

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

We also write $\lfloor x \rfloor$ for the largest integer strictly smaller than $x$. The notation $\mathbb{N}_+$ and $\mathbb{R}_+$ indicate the set of positive natural and real numbers respectively. For sequences $a_n$ and $b_n$ we write $a_n = O(b_n)$ and $a_n \lesssim b_n$ if there exist $C > 0$ and $N > 0$ such that $n \geq N$ implies $a_n \leq Cb_n$. If $a_n = O(b_n)$ and $b_n = O(a_n)$ then we write $a_n \asymp b_n$.

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

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

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 $\kappa > 0$ such that for $\delta \in \mathbb{R}^n$ satisfying $\|\delta\|_\infty \leq \kappa$ we have that, a.s.,*

$$|F_{y_i|x_i}(f_\tau^*(x_i)+\delta_i) - F_{y_i|x_i}(f_\tau^*(x_i))| \geq \underline{p}|\delta_i|$$

*for $i = 1,\dots,n$ and for some constant $\underline{p} > 0$. We also require that*

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

*for some constant $c > 0$, where $p_{y_i|x_i}$ is the probability density function of $y_i$ conditioning on $x_i$*

Assumption 2 requires that there exists a neighborhood around $f_\tau^*(x_i)$ in which the cumulative distribution function of $y_i$ conditioning on $x_i$ is well behaved. Related conditions appeared as D.1 Belloni and Chernozhukov (2011), Condition 2 in He and Shi (1994), and Assumption A in Padilla and Chatterjee (2020).

Next we define $f_n$, the projection of the quantile function $f_\tau^*$ 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*

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

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

$$err_1 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \rho_\tau(z_i - f_n(x_i)) - \frac{1}{n}\sum_{i=1}^n \rho_\tau(z_i - f_\tau^*(x_i))\right].$$

Notice that when $f_\tau^* \in \mathcal{F}(W, U, L)$ and $\|f\|_\infty^* \leq F$ then 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 2** *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*

$$\mathbb{E}\left[\Delta_n^2(f_\tau^*, \hat{f}) \,\middle|\, x_1, \ldots, x_n\right] \leq c_1 F\left[\frac{\{LW\log U \cdot \log n\}}{n}\right]^{1/2} + c_1 err_1,$$

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

$$\mathbb{E}\left[\|\hat{f} - f_\tau^*\|_n^2 \,\middle|\, x_1, \ldots, x_n\right] \leq c_1 \max\{1, F\}F\left[\frac{\{LW\log U \cdot \log n\}}{n}\right]^{1/2} + c_1 \max\{1, F\} err_1.$$

Theorem 2 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 $L$, $W$ and $U$ are constants in $n$, then the rate becomes $n^{-1/2} + \text{err}_1$. We do not claim that this rate is optimal. The novelty in Theorem 2 is that it holds without any constraints on the parameters of the neural network class used for estimation and without any smoothness assumption of the quantile function. In the next two subsections, we will include some constraints in the ReLU network class, which will lead to rates that match minimax rates in regression for very broad function classes.

Finally, we highlight that Theorem 2 gives a general risk bound on the performance of quantile regression with ReLU networks. In the case of homescedastic errors with finite variance, it is posssible to translate the upper bound in Theorem 2 to be an upper bound on the genralization error. While our work here is theoretical, other authors (Jiang et al., 2019) have studied empirical aspects of the generalization error in the classification setting. It would be interesting to apply the ideas from Jiang et al. (2019) to the quantile setting,

but that is out of the scope of this paper. Theorem 2 is also limited by the fact that it is concerned with statistical guarantees for an optimal solution to the problem in (13), whereas in practice one has to use stochastic gradient descent often times accompained by early stopping or some other form of implicit regularization. We refer the reader to Soudry et al. (2018) where the authors studied the optimization behavior of gradient descent in general frameworks including clasification with neural networks.

### 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 (2020). There, the authors showed that for such class, neural networks with ReLU activation function attain minimax rates. However, the results in Schmidt-Hieber (2020) 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 3** *We define the class of ReLU neural networks $\mathcal{G}(L, p, s, F)$ as*

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

With the notation in Definition 3, we consider the estimator

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

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

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

A few comments are in order. First, notice that we assume that all the parameters are bounded by one. As discussed in Schmidt-Hieber (2020), 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 (2020) and Suzuki (2018). See Hassibi and Stork (1993); Han et al. (2015); Frankle and Carbin (2018) and 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 (2020).

**Definition 4** *For $\beta > 0$ and $r \in \mathbb{N}_+$ we define the class of Hölder functions of exponent $\beta$ as*

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

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

We are now ready to construct the function class based on composition of Hölder functions.

**Definition 5** *For $q \in \mathbb{N}_+$, $d = (d_0,\ldots,d_{q+1}) \in \mathbb{N}_+^{q+2}$, $t = (t_0,\ldots,t_q) \in \mathbb{N}_+^{q+1}$, $\beta = (\beta_0,\ldots,\beta_q) \in \mathbb{R}_+^{q+1}$ and $K \in \mathbb{R}_+$ we define the class of functions*

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

With Definitions 4–5, we now state an assumption on the true quantile function regarding its smoothness.

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

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

*for $i = 1,\ldots,q-1$ and $\beta_q^* = \beta_q$.*

Importantly, as Section 5 of Schmidt-Hieber (2020) showed, the class $\mathcal{H}(q,d,t,\beta,K)$ is challenging enough so that wavelet estimators are suboptimal for estimating $f_\tau^* \in \mathcal{H}(q,d,t,\beta,K)$. Our main result in this section shows that, in contrast, quantile regression with ReLU networks attains optimal rates.

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 $X$ 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 (2020).

**Theorem 6** *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*

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

*where*

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

*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 6 shows that the ReLU network based estimator defined in (14) attains the rate $\epsilon_n$ under the mean squared error and the $\ell_2$ metrics, ignoring $L$ and the log factors, for estimating quantile functions in the class $\mathcal{H}(q, d, t, \beta, K)$. Importantly, the rate $\epsilon_n$ is minimax for estimating functions in the class $\mathcal{H}(q, d, t, \beta, K)$. Specifically, Theorem Schmidt-Hieber (2020) showed that if $t_j \leq \min\{d_0, \ldots, d_{j-1}\}$ for all $j$ then for a constant $c > 0$ we have that

$$\inf_{\hat{f}} \sup_{f_{0.5}^* \in \mathcal{H}(q,d,t,\beta,K)} \|\hat{f} - f_{0.5}^*\|_{\ell_2}^2 \;\geq\; c\epsilon_n,$$

where the infimum is taken over all possible estimators, and with the assumption that the errors are Gaussian and the covariates are uniformly distributed in $[0, 1]^d$. Thus, Theorem 6 provides an upper bound that nearly matches the lower bound and that it allows for 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 6 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 7** *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*

$$\begin{aligned}\mathcal{I}(L, W, S, B) \quad := \quad & \Big\{(A^{(L)}\phi(\cdot) + b^{(L)}) \circ \ldots \circ (A^{(1)}x + b^{(1)}) \;:\; A^{(l)} \in \mathbb{R}^{W\times W},\; b^{(l)} \in \mathbb{R}^W \\ & \sum_{l=1}^{L}\Big(\|A^{(l)}\|_0 + \|b^{(l)}\|_0\Big) \leq S,\; \max_l \max\{\|A^{(l)}\|_\infty, \|b^{(l)}\|_\infty\} \leq B\Big\}.\end{aligned}$$

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 7, we focus on the estimator

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

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 satifies $f_\tau^* \in B_{p,q}^s([0,1]^d)$, $\|f_\tau^*\|_\infty \leq F$, 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 in Definition 14 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 8** *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*

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

*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 8 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 8 holds under general assumptions of the errors allowing for heavy-tailed distributions. Furthermore, the rate $n^{-\frac{2s}{2s+d}}$ is minimax for estimation, with Gaussian errors, of the conditional mean when such function belongs to a fixed ball of the space $B_{p,q}^s([0,1]^d)$, 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 consider two ReLU models: (i) the näive strategy of fitting separate ReLU networks for each quantile then projecting the answers to the monotone surface to enforce coherency of the quantiles, and (ii) the single multitask ReLU network we propose in eq. (5) which enforces monotonicity directly and shares statistical strength across quantiles. We compare the 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. The multitask ReLU model is shown to perform better than the separate ReLU at extreme quantiles while performing similarly or slightly worse at predicting the median.

### 4.1 Univariate response

We assess the performance of quantile regression with ReLU networks (Q-Network) 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 Network), 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 $\tau$, $\tau \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$. Since SqErr Network only estimates the mean, we only

1. `https://www.statsmodels.org`
2. `https://scikit-garden.github.io/`

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,

$$\begin{array}{rcl} y_i & = & f_0(x_i) + \epsilon_i, \;\; i = 1, \ldots, n, \\ x_i & \overset{\text{ind}}{\sim} & [0,1]^d, \end{array}$$

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

**Scenario 1.** We set

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

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

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

with $v_i \overset{\text{ind}}{\sim} t(2)$, for $i = 1, \ldots, n$, where $t(2)$ is the t-distribution with 2 degrees of freedom.

**Scenario 2.** In this scenario we specify

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

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

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

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

and setting $\epsilon_i = \sqrt{x_i^\top \beta} \nu_i$, where $\beta = (1, 1/2)^\top$ and $\nu_i \overset{\text{ind}}{\sim} t(2)$ for $i = 1, \ldots, n$.

**Scenario 4.** The function $f_0$ is chosen as

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

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

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

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

| | | **Scenario 1** | | | | | **Scenario 2** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Method | $\tau = 0.05$ | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ | $\tau = 0.95$ | $\tau = 0.05$ | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ | $\tau = 0.95$ |
| 100 | SqErr Network | * | * | 0.99 | * | * | * | * | 0.33 | * | * |
| | Q-Network (Separate) | 4.66 | 0.60 | 0.20 | 0.36 | 5.01 | 6.08 | 1.96 | 1.15 | 2.00 | 6.40 |
| | Q-Network (Multitask) | 0.92 | 0.68 | 0.30 | 0.48 | 2.83 | 3.02 | 1.58 | 1.27 | 1.96 | 3.40 |
| | Quantile Spline | 2.87 | 0.25 | 0.14 | 0.20 | 5.40 | 5.41 | 1.15 | 0.63 | 1.09 | 6.13 |
| | Quantile Forest | 2.22 | 0.72 | 0.41 | 0.55 | 3.49 | 4.08 | 1.59 | 1.02 | 1.60 | 4.81 |
| 1000 | SqErr Network | * | * | 0.96 | * | * | * | * | 0.24 | * | * |
| | Q-Network (Separate) | 0.46 | 0.07 | 0.05 | 0.05 | 0.43 | 1.59 | 0.25 | 0.13 | 0.35 | 1.64 |
| | Q-Network (Multitask) | 0.26 | 0.10 | 0.04 | 0.06 | 0.19 | 0.51 | 0.20 | 0.18 | 0.22 | 0.43 |
| | Quantile Spline | 0.17 | 0.07 | 0.06 | 0.07 | 0.18 | 0.58 | 0.12 | 0.05 | 0.10 | 0.73 |
| | Quantile Forest | 1.72 | 0.20 | 0.11 | 0.68 | 3.10 | 3.92 | 1.24 | 0.68 | 1.23 | 3.97 |
| 10000 | SqErr Network | * | * | 0.95 | * | * | * | * | 0.18 | * | * |
| | Q-Network (Separate) | 0.05 | 0.01 | 0.01 | 0.01 | 0.07 | 0.28 | 0.03 | 0.01 | 0.03 | 0.22 |
| | Q-Network (Multitask) | 0.06 | 0.01 | 0.01 | 0.01 | 0.03 | 0.07 | 0.04 | 0.03 | 0.04 | 0.06 |
| | Quantile Spline | 0.08 | 0.06 | 0.06 | 0.06 | 0.07 | 0.09 | 0.01 | 0.00 | 0.01 | 0.06 |
| | Quantile Forest | 6.07 | 1.66 | 0.04 | 0.14 | 2.20 | 3.58 | 1.04 | 0.55 | 1.03 | 3.48 |
| | | **Scenario 3** | | | | | **Scenario 4** | | | | |
| $n$ | Method | $\tau = 0.05$ | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ | $\tau = 0.95$ | $\tau = 0.05$ | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ | $\tau = 0.95$ |
| 100 | SqErr Network | * | * | * | * | * | * | * | 0.22 | * | * |
| | Q-Network (Separate) | 14.66 | 0.86 | 0.33 | 0.91 | 5.83 | 6.64 | 2.82 | 1.62 | 3.13 | 9.99 |
| | Q-Network (Multitask) | 1.63 | 1.32 | 0.97 | 1.60 | 3.95 | 5.02 | 3.34 | 2.95 | 3.83 | 5.29 |
| | Quantile Spline | 4.24 | 0.41 | 0.25 | 0.42 | 7.05 | * | * | * | * | * |
| | Quantile Forest | 6.80 | 0.95 | 0.33 | 1.69 | 13.50 | 3.77 | 1.81 | 1.01 | 1.61 | 5.32 |
| 1000 | SqErr Network | * | * | 0.21 | * | * | * | * | 0.12 | * | * |
| | Q-Network (Separate) | 1.30 | 0.08 | 0.06 | 0.11 | 1.02 | 2.28 | 0.37 | 0.18 | 0.30 | 1.64 |
| | Q-Network (Multitask) | 0.21 | 0.10 | 0.08 | 0.10 | 0.22 | 0.77 | 0.38 | 0.29 | 0.38 | 0.87 |
| | Quantile Spline | 0.41 | 0.09 | 0.08 | 0.09 | 0.37 | * | * | * | * | * |
| | Quantile Forest | 11.08 | 0.74 | 0.20 | 1.35 | 9.98 | 2.82 | 0.82 | 0.41 | 0.76 | 2.55 |
| 10000 | SqErr Network | * | * | 0.17 | * | * | * | * | 0.06 | * | * |
| | Q-Network (Separate) | 0.26 | 0.02 | 0.02 | 0.02 | 0.19 | 0.39 | 0.07 | 0.03 | 0.06 | 0.32 |
| | Q-Network (Multitask) | 0.05 | 0.02 | 0.02 | 0.02 | 0.05 | 0.15 | 0.06 | 0.05 | 0.07 | 0.18 |
| | Quantile Spline | 0.10 | 0.07 | 0.07 | 0.07 | 0.11 | * | * | * | * | * |
| | Quantile Forest | 36.17 | 11.78 | 2.26 | 0.57 | 9.43 | 1.75 | 0.43 | 0.19 | 0.42 | 1.72 |

Table 1: Univariate Responses Tasks. Performances of different methods in Scenarios 1–4, in terms of squared error from the true quantile averaged over 25 independent trials.

| Scenario 5 | | | | | | |
|---|---|---|---|---|---|---|
| $n$ | Model | $\tau = 0.05$ | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ | $\tau = 0.95$ |
| 100 | SqErr Network | * | * | 31.26 | * | * |
| | Q-Network (Separate) | 18.51 | 2.81 | 1.41 | 2.85 | 17.31 |
| | Q-Network (Multitask) | 4.16 | 3.99 | 2.37 | 3.27 | 10.08 |
| | Quantile Spline | * | * | * | * | * |
| | Quantile Forest | 43.32 | 21.75 | 15.13 | 20.21 | 49.29 |
| 1000 | SqErr Network | * | * | 30.72 | * | * |
| | Q-Network (Separate) | 1.82 | 0.52 | 0.34 | 0.52 | 1.76 |
| | Q-Network (Multitask) | 1.53 | 0.96 | 0.45 | 0.88 | 2.49 |
| | Quantile Spline | * | * | * | * | * |
| | Quantile Forest | 29.46 | 12.55 | 7.33 | 10.85 | 32.85 |
| 10000 | SqErr Network | * | * | 30.63 | * | * |
| | Q-Network (Separate) | 0.30 | 0.09 | 0.07 | 0.08 | 0.28 |
| | Q-Network (Multitask) | 0.24 | 0.16 | 0.09 | 0.18 | 0.37 |
| | Quantile Spline | * | * | * | * | * |
| | Quantile Forest | 17.36 | 7.01 | 3.67 | 5.76 | 18.47 |

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

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

A visualization of the perfomance of the different approaches and true quantile functions can be found in Section E of the Appendix.

We report the results for Scenarios 1-4 in Table 1 and Scenario 5 in Table 2. From Table 1 we can see that in Scenario 1, the Q-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 Q-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, Q-Network outperforms other methods 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 Tables 1-2, we show that for Scenarios 4–5, and most quantiles, the clear best method is Q-Network (for $n > 100$), with Quantile Forests as the second best. Across all scenarios, the multitask Q-Network is less sample efficient than the separate Q-Network when predicting the median, though still typically outperforms other methods. On the extremal quantiles (5% and 95%) the multitask Q-Network outperforms all methods in all scenarios at most sample sizes. This suggests that sharing statistical strength between quantiles through the multitask architecture has the benefit of reducing the variance of the estimates for the extremes at minor cost to the median prediction accuracy; exploring these tradeoffs is an interesting direction for future work.

Overall, the results in Tables 1-2 demonstrate a clear advantage of the Q-Network method. This method generally outperforms SqErr Network in most examples, presumably due to the heavy-tailed or heteroscedastic error distributions. At the same time, Q-Network

also outperforms the other competitors with larger sample size or more complicated quantile functions.

One disconnect between our theoretical and empirical results is the structure of the model class. Our theory results assume the ReLU networks are sparse, whereas our benchmarks initialize the neural network models to be dense, randomly-weighted representations. The success of the Q-Network in these conditions may therefore seem surpising. An examination of the distribution of the learned weights of the Q-Network (fig. 1) reveals that the vast majority of weights in the Q-Network lie a small epsilon ($\epsilon = 10^{-6}$) of zero. The nonzero weights tend to be thousands of times larger than the numerically zero weights and therefore likely dominate the computation. Thus, the learned Q-Network is functionally sparse with most weights having no meaningful effect on the predictions. In contrast, the SqErr network typically has a uniform distribution of weights, suggesting the internal representation is dense.

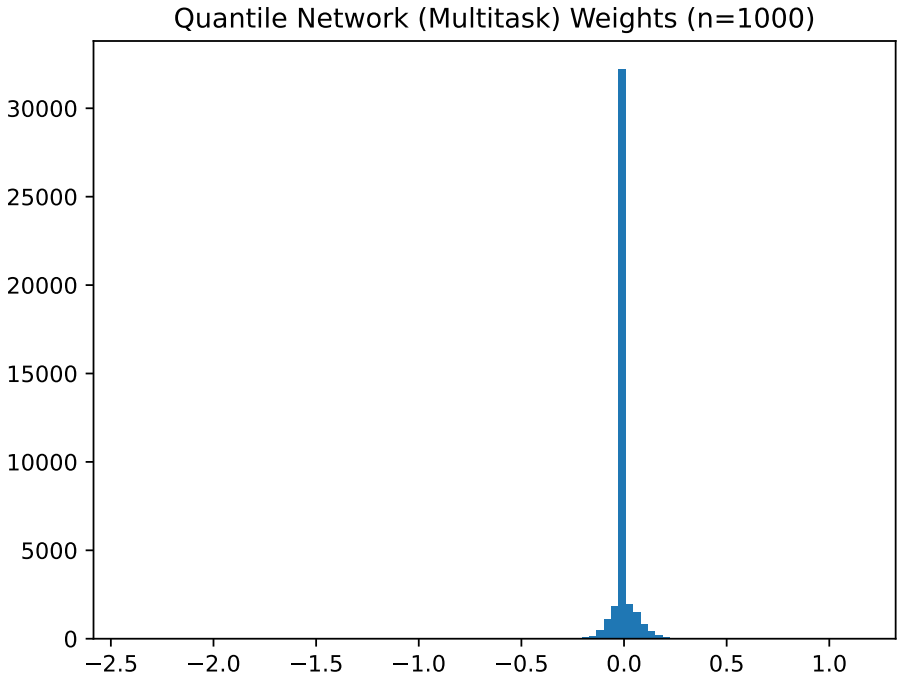


Figure 1: Quantile network weights are functionally sparse. Above: distribution of weights for one instance of the Q-Network in Scenario 1 with 1000 training samples; other scenarios and sample sizes produce qualitatively similar results. Approximately 75% of the weights are less than $10^{-6}$, suggesting they are dominated by the larger weights and therefore the model is functionally sparse.

### 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

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

Table 3: Multivariate Responses Tasks: Performance of different methods in Scenarios 6 and 7. We measure performance using the averaged mean squared error based on 25 Monte Carlo simulations for the two synthetic multivariate benchmarks.

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

$$\frac{1}{nd'}\sum_{j=1}^{d'}\sum_{i=1}^{n}\left(f^*_{\tau,j}(x_i)-\hat{f}_j(x_i)\right)^2,$$

where the quantile functions $f^*_{\tau,j}(\cdot)$, $j=1,\ldots,d'$, are defined in (8), and $\tau=0.5$.

We consider two multivariate response generative models as follows.

**Scenario 6.**

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

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.**

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

Table 3 illustrates the performance of different methods in Scenarios 6 and 7 with sample sizes $n=\{100,1000,10000\}$. We can see that for large $n$ both Quantile Network and Geometric Quantile outperform the $\ell_2$-based approach SqErr Network. This corroborates the results earlier in Section 4.1. Particularly, it demonstrates that both Quantile Network and Geometric Quantile are robust estimators when dealing with heavy-tailed distributions.

## 5. Proof ideas

In this section we give an overview behind the ideas of the proof of Theorems 6 and 8. For both of these results we start by defining the empirical loss function

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

for $f \in \mathcal{F}$, where

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

and we set

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

for the population loss. With this notation, we now proceed to break down the proof of Theorems 6 and 8 into different steps, the details of which can be found in the Appendices. Througout, for sequences $a_n$ and $b_n$ we write $a_n \lesssim b_n$ if $a_n \leq cb_n$ for a constant $c > 0$.

**Step 1.** In Lemma 15 we show that

$$\Delta^2(f, f_n) \lesssim M_n(f) + \|f_n - f_\tau^*\|_\infty \Delta(f, f_n)\sqrt{F}.$$

Thus, we relate the population loss function $M_n$ with the function $\Delta^2$.

**Step 2.** Next, we use the optimality of $\hat{f}$ to show in Lemma 17 that

$$\Delta^2(\hat{f}, f_n) \lesssim M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \Delta(\hat{f}, f_n)\sqrt{F}.$$

**Step 3.** We then show a localization property in Lemma 18. Specifically, we show that if

$$3\mathbb{E}\left(\sup_{f\in\mathcal{F},\ \|f-f_n\|_{\ell_2}^2 \leq r^2} \frac{1}{n}\sum_{i=1}^{n} \xi_i (f(x_i) - f_n(x_i))^2\right) \leq r^2, \tag{17}$$

for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$, then, with high probability, it holds that if $f \in \mathcal{F}$, and $\|f - f_n\|_{\ell_2}^2 \leq r^2$ then

$$\|f - f_n\|_n^2 \leq 4r^2.$$

**Step 4.** Using Step 3, we show that if $\|\hat{f} - f_n\|_{\ell_2} \leq r_0$ for some constant $r_0$ satisfying (17), then, with high probability,

$$\|\hat{f} - f_n\|_{\ell_2}^2 \ \leq \ r_0 \Psi(n)$$

for a sequence $\Psi(n)$ that depends on $n$ and the complexity of the class $\mathcal{F}$.

**Step 5.** In Lemma 20 we give an upper bound on $r^* > 0$, where $r^*$ satisfies that if $r_0 > r^*$ then (17) holds with $r_0$.

**Step 6.** The final step starts by assuming that $\|\hat{f} - f_\tau^*\|_{\ell_2}$ is upper bounded by some large quantity. Then we repeatedly apply Steps 2 to 5 to arrive at the desired upper bound in Theorem 6 (and similarly in Theorem 8).

As for the novelty of our argument, we highlight that Steps 1–2 have not appeared in the literature before, whereas Steps 3–6 are also novel in the context of quantile regression.

## 6. Discussion

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 (2020) 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 univariate and multivariate response quantile regression with ReLU networks show an advantage over other quantile regression methods. Quantile regression networks were also shown to outperform $\ell_2$-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.

However, we acknowledge a certain mismatch between our theory and empirical findings. Specifically, some open questions that we leave for future work are the following.

- In our analysis, we study the statistical properties of a minimizer of the quantile empirical loss restricted to the class of neural networks. It is unknown if the same properties hold for the solutions used in practice based on stochastic gradient descent, say with early stopping (Prechelt, 1998).

- In Theorems 6 and 8, the parameters of the neural networks classes considered are restricted to be sparse. However, in our simulations, we do not enforce this constraint. Despite that, Figure 1 shows that the estimated networks are sparse. This leads to the fundamental question of whether or not it is possible to drop the sparsity assumption from Theorems 6 and 8 and perhaps replace it with some more practical condition. For instance, Soudry et al. (2018) argues that the model complexity can be controlled by the dynamics of stochastic gradient descent instead of imposing sparsity penalties. Unfortunately, we are not aware of how to do this while maintaining minimax rates, and even in the regression setting, existing works enforcing the sparsity constraint are very common in the literature (Suzuki, 2018; Schmidt-Hieber, 2020). Though our upper bound in Theorem 12 holds without sparsity constraints.

- A theoretical study of statistical rates of convergence for multivariate response quantile regression with ReLU networks is left for future work.

Finally, we emphasize the sensitivity of Theorems 2, 6 and 8 to $\tau$. Specifically, while the rate in those theorems does not depend on $\tau$, the constants involved in the rates become larger as $\tau$ becomes closer 0 or 1. This intuitively explains the higher errors in the performance of Q-Network when $\tau \in \{0.05, 0.95\}$ in Section 4.

## Appendix 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 $\text{N}(\epsilon, \mathcal{F}, \text{dist}(\cdot, \cdot))$ 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

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

Furrthermore, if $a_n$ and $b_n$ are positive sequences, we say that $a_n \lesssim b_n$ if there exists $m$ such that $n \geq m$ implies $a_n \leq cb_n$ for a constant $c > 0$.

## Appendix B. Theorem 2

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 9** *We define the empirical loss function*

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

*where*

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

*with*

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

*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 10** *Suppose that Assumption 1–2 hold. Then there exists a constant $c_\tau$ such that for any $\delta \in \mathbb{R}^n$, we have*

$$\sum_{i=1}^{n} \mathbb{E}\left[\rho_\tau(z_i - f^*_\tau(x_i) - \delta_i) - \rho_\tau(z_i - f^*_\tau(x_i))\right] \geq c_\tau \sum_{i=1}^{n} D^2(\delta_i),$$

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

**Proof** See Lemma 13 in Padilla and Chatterjee (2020). ■

**Definition 11** *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. We define the pseudodimension of $\mathcal{H}$, denoted as $\operatorname{Pdim}(\mathcal{H})$, as the largest integer $m$ for which there exist $(a_1, b_1) \ldots, (a_m, b_m) \in \mathcal{X} \times \mathbb{R}$ 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 $i = 1, \ldots, m$.*

**Theorem 12 (Theorem 7 from Bartlett et al. (2019))** *With the notation from before, we have that*

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

### B.2 Proof of Theorem 2

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

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} D^2\left\{f_\tau^*(x_i) - \hat{f}(x_i)\right\}\right] &\leq \frac{1}{c_\tau n}\mathbb{E}\left(\sum_{i=1}^{n}\mathbb{E}\left[\rho_\tau\{z_i - f_\tau^*(x_i) - \hat{\delta}_i\}\right] - \sum_{i=1}^{n}\mathbb{E}\left[\rho_\tau\{z_i - f_\tau^*(x_i)\}\right]\right) \\ &= \frac{1}{c_\tau}E\left\{M_n(\hat{f})\right\} + \frac{1}{c_\tau}\mathrm{err}_1, \end{aligned} \tag{18}$$

where the inequality follows from Lemma 10.

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

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

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 12.

■

## Appendix C. Theorem 8

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 $f : \mathcal{X} \to \mathbb{R}$ we let

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

and

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

**Definition 13** *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*

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

**Definition 14** *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

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

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

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

*and*

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

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

**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,

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

for a constant $c_1>0$, where the first inequality holds since the cumulative distribution function of $Y$ conditioning on $X$ is Lipchitz around $f_\tau^*(X)$ by Assumption 2, and by the same argument from the proof of Lemma 13 in Padilla and Chatterjee (2020).

■

**Definition 16** *We define the empirical loss function*

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

*where*

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

*and we set*

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

**Lemma 17** *Suppose that $\|f_n-f_\tau^*\|_\infty\le c$ for a small enough constant $c$. The estimator $\hat{f}$ defined in (16) satisfies*

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

*Furthermore,*

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

**Proof** By Lemma 15 we have that

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

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

**Lemma 18** *Suppose that*

$$3\mathbb{E}\left(\sup_{f\in\tilde{\mathcal{I}}(L,W,S,B),\ \|f-f_n\|_{\ell_2}^2\leq r^2}\frac{1}{n}\sum_{i=1}^n\xi_i(f(x_i)-f_n(x_i))^2\right)\leq r^2, \tag{19}$$

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

$$\max\left\{4F\sqrt{\frac{\gamma}{n}}, 4F\sqrt{\frac{\gamma}{3n}}\right\}\leq r. \tag{20}$$

*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

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

for all $x$. Hence,

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

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

$$\begin{aligned}&\sup_{f\in\tilde{\mathcal{I}}(L,W,S,B),\ \|f-f_n\|_{\ell_2}^2\leq r^2}\left\{\|f-f_n\|_n^2-\|f-f_n\|_{\ell_2}^2\right\}\\ &\leq 3\mathbb{E}\left(\sup_{f\in\tilde{\mathcal{I}}(L,W,S,B),\ \|f-f_n\|_{\ell_2}^2\leq r^2}\frac{1}{n}\sum_{i=1}^n\xi_i(f(X_i)-f_n(X_i))^2\right)\\ &+4rF\sqrt{\frac{\gamma}{n}}+\frac{16F^2\gamma}{3n}\end{aligned} \tag{21}$$

and the claim follows. ■

**Lemma 19** *Suppose that $\|\hat{f} - f_n\|_{\ell_2} \leq r_0$, with $r_0$ satisfying (19)-(20) and Assumption 5 holds. Also, with the notation of Assumption 4, suppose that for the class $\mathcal{I}(L, W, S, B)$ the parameters are chosen as*

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

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

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

*Then for some positive constant $C_0$ it holds that*

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

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

**Proof** Let

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

Then for $\xi_1, \ldots, \xi_n$ independent Rademacher variables independent of $\{(x_i, y_i)\}_{i=1}^n$, we have that

$$\begin{aligned} c_\tau \|\hat{f} - f_n\|_{\ell_2}^2 \quad \leq \quad & 2F[M_n(\hat{f}) - \hat{M}_n(\hat{f}) + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} \sqrt{F}] \\ \leq \quad & 2F \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}(g(X,Y)) - \frac{1}{n} \sum_{i=1}^n g(x_i, y_i) \right\} + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} F^{3/2} \\ \leq \quad & 12F\mathbb{E}\left( \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(x_i, y_i) \middle| (x_1, y_1), \ldots, (x_n, y_n) \right) \\ & + 4 r_0 F^{2.5} \sqrt{\frac{\gamma}{n}} + \frac{100 F^{2.5} \gamma}{3n} + \|f_n - f_\tau^*\|_\infty \|\hat{f} - f_n\|_{\ell_2} F^{1.5}, \end{aligned} \tag{22}$$

where the first inequality follows from Lemmas 17, 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 $C > 0$,

$$\begin{aligned}
\mathbb{E}_\xi\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \xi_i g(x_i,y_i)\right) &\leq \mathbb{E}_\xi\left(\sup_{f\in\mathcal{I}(L,W,S,B),\ \|f\|_\infty\leq F\ \|f-f_n\|_{\ell_2}\leq r_0}\frac{1}{n}\sum_{i=1}^n \xi_i(f(x_i)-f_n(x_i))\right)\\
&\leq \mathbb{E}_\xi\left(\sup_{f\in\mathcal{I}(L,W,S,B),\ \|f\|_\infty\leq F,\ \|f_n-f\|_n\leq 2r_0}\frac{1}{n}\sum_{i=1}^n \xi_i(f(x_i)-f_n(x_i))\right)\\
&\leq \inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\tilde{\mathcal{I}}(L,W,S,B),\|\cdot\|_n)}d\delta\right\}\\
&\leq \inf_{0<\alpha<2r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\tilde{\mathcal{I}}(L,W,S,B),\|\cdot\|_\infty)}d\delta\right\}\\
&\leq \inf_{0<\alpha<r_0}\left\{4\alpha+\frac{24r_0}{\sqrt{n}}\sqrt{\log\mathcal{N}(\alpha,\tilde{\mathcal{I}}(L,W,S,B),\|\cdot\|_\infty)}\right\},\\
&\leq C\inf_{0<\alpha<r_0}\left\{\alpha+r_0\sqrt{\frac{N\left[(\log N)^2+\log\alpha^{-1}\right]}{n}}\right\},
\end{aligned} \tag{23}$$

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 18, 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

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

(22) and (23) imply

$$\begin{aligned}
\|\hat{f} - f_\tau^*\|_{\ell_2}^2 \leq\ & \tfrac{1}{c_\tau}\Bigg[4r_0F^{2.5}\sqrt{\tfrac{\gamma}{n}}+\tfrac{100F^{2.5}\gamma}{3n}+\\
& Cr_0F\sqrt{\tfrac{N(\log N)^2}{n}}+Cr_0F\sqrt{\tfrac{N\left[(\log N)^2+\log r_0^{-1}+\log n\right]}{n}}+C_1r_0N^{-s/d}F^{1.5}\Bigg]
\end{aligned}$$

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

**Lemma 20** *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 19,*

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

*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 the define the event

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

If $r^*$ satisfies (20) with $\gamma = \log n$ then we have that $\mathbb{P}(E) \geq 1 - 1/n$ by Lemma 18.

Also,

$$\begin{aligned}
(r^*)^2 &\leq 3\mathbb{E}\left( \sup_{f \in \tilde{\mathcal{I}}(L,W,S,B),\, \|f-f_n\|_{\ell_2}^2 \leq (r^*)^2} \frac{1}{n}\sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \right) \\
&\leq 3\mathbb{E}\left( \mathbb{E}\left( \sup_{f \in \tilde{\mathcal{I}}(L,W,S,B),\, \|f-f_n\|_n^2 \leq (2r^*)^2} \frac{1}{n}\sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \middle| x_1, \ldots, x_n \right) 1_E \right) + \\
&\quad \frac{12F^2}{n} \\
&\leq 3\mathbb{E}\left( \mathbb{E}\left( \sup_{f \in \tilde{\mathcal{I}}(L,W,S,B),\, \|f-f_n\|_n^2 \leq (2r^*)^2} \frac{1}{n}\sum_{i=1}^n \xi_i (f(x_i) - f_n(x_i))^2 \middle| x_1, \ldots, x_n \right) 1_E \right) + \\
&\quad \frac{12F^2}{n}.
\end{aligned} \tag{24}$$

Next, let

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

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 (24), using Dudley's chaining we obtain that

$$\begin{aligned}
(r^*)^2 &\leq 3\mathbb{E}\left( \inf_{0<\alpha<2r^*} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r^*} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_n)} \right\} \right) + \frac{12F}{n} \\
&\leq 3\mathbb{E}\left( \inf_{0<\alpha<2r^*} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2r^*} \sqrt{\log \mathcal{N}(\delta, \mathcal{G}, \|\cdot\|_\infty)} \right\} \right) + \frac{12F}{n} \qquad (25) \\
&\leq C \inf_{0<\alpha<2r^*} \left\{ \alpha + r^* \sqrt{\frac{N\left[(\log N)^2 + \log \alpha^{-1}\right]}{n}} \right\} + \frac{12F}{n},
\end{aligned}$$

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 (20) then for a constant $\tilde{C} > 0$

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

which follows from (25) by taking

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

The claim follows. ■

### C.3 Proof of Theorem 8

**Proof** Throughout we use the notation from the proof of Lemma 19. 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

$$\mathrm{B}(f_n, \|\cdot\|_{\ell_2}, \overline{r}),\ \mathrm{B}(f_n, \|\cdot\|_{\ell_2}, 2\overline{r})\backslash\mathrm{B}(f_n, \|\cdot\|_{\ell_2}, \overline{r}), \ldots, \mathrm{B}(f_n, \|\cdot\|_{\ell_2}, 2^l\overline{r})\backslash\mathrm{B}(f_n, \|\cdot\|_{\ell_2}, 2^{l-1}\overline{r}),$$

where

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

Next, if $\overline{r} > r^*$, the by Lemma 18, with probability at least $1 - le^{-\gamma}$, we have that

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

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

$$\hat{f} \in \mathrm{B}(f_n, \|\cdot\|_{\ell_2}, 2^j\overline{r})\backslash\mathrm{B}(f_n, \|\cdot\|_{\ell_2}, 2^{j-1}\overline{r}),$$

then by Lemma 19, with probability at least $1 - 4e^{-\gamma}$, we have that

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

provided that

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

and

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

both of which for all $j$ hold if

$$\overline{r} = 8\tilde{C}\left[F^{2.5}\sqrt{\frac{\gamma}{n}} + D\sqrt{\frac{N(\log N)^2}{n}} + D\sqrt{\frac{N\left[(\log N)^2 + 2\log n\right]}{n}} + N^{-s/d}F^{1.5}\right] + 2\sqrt{\frac{\tilde{C}F^{2.5}\gamma}{n}} + r^*. \tag{26}$$

Therefore, by Lemmas 18–19, with probability at least $1 - e^{-\gamma}$, we have that

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

which by the previous argument implies that

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

and continuing recursively we arrive at

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

The claim follows by noticing that Lemma 20 and (26) imply that

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

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

## Appendix D. Proof of Theorem 6

The proof is similar to that of Theorem 8 relying in the lemmas given next.

Just as in Lemma 15, we have the following result.

**Lemma 21** *Suppose that $\|f_n - f_\tau^*\|_\infty \leq c$ for a small enough constant $c$. Then*

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

*and*

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

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

Similarly to Lemmas 18–20, we have the following three lemmas.

**Lemma 22** *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 23** *If*

$$3\mathbb{E}\left(\sup_{f\in\mathcal{G}(L,p,S,F),\ \|f-f_n\|_{\ell_2}\leq r}\frac{1}{n}\sum_{i=1}^{n}\xi_i(f(X_i)-f_n(X_i))^2\right)\leq r^2, \tag{27}$$

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

$$\max\left\{4F\sqrt{\frac{\gamma}{n}},4F\sqrt{\frac{\gamma}{3n}}\right\}\leq r. \tag{28}$$

*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 24** *Suppose that* $\|\hat{f}-f_n\|\leq r_0$*, with* $r_0$ *satisfying (27)-(28) and Assumption 3 holds. Also, with the notation of Definition 5, suppose that for the class* $\mathcal{G}(L,p,S,F)$ *the parameters are chosen to satisfy*

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

*where*

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

*Then*

$$\begin{array}{rcl}\|\hat{f}-f_n\|_{\ell_2}^2 & \lesssim & r_0\sqrt{F\epsilon_n\log n\left[\log L+L\log n\right]}+r_0\sqrt{F\epsilon_n\log n\left[\log L+L\log n+\log(n)\right]}\\ & & +r_0F^2\sqrt{\frac{\gamma}{n}}+\frac{F^2\gamma}{n}+r_0F\max_{i=0,1,\ldots,q}n^{-\frac{\beta_i^*}{2\beta_i^*+t_i}}\end{array}$$

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

**Proof** Let

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

Prooceding as in the proof of Lemma 19, we obtain that

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

for some positive constant $C$.

However, defining $V=\prod_{l=0}^{L+1}(p_l+1)$, then Lemma 5 in Schmidt-Hieber (2020) and (29) imply that

$$\begin{array}{rcl}\mathbb{E}_\xi\left(\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\xi_ig(x_i,y_i)\right) & \leq & C\inf_{0<\alpha<r_0}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^{2r_0}\sqrt{\log\mathcal{N}(\delta,\mathcal{G}(L,p,S,F),\|\cdot\|_\infty)}d\delta\right\}\\ & \leq & 4C\alpha+\frac{24Cr_0}{\sqrt{n}}\sqrt{(S+1)\log\left(2\alpha^{-1}(L+1)V^2\right)},\end{array} \tag{30}$$

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

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

(30) implies

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

Hence, as in (22), we obtain that

$$\begin{aligned}\|\hat{f}-f_n\|_{\ell_2}^2 &\le \frac{2\sqrt{F}}{c_\tau}\Bigg[24r_0C\sqrt{\frac{F(S+1)\log((L+1)V^2)}{n}}+\frac{144\sqrt{F}r_0C}{\sqrt{n}}\sqrt{(S+1)\log\left(2(L+1)V^2n\right)}+\\ &\quad +4r_0F^2\sqrt{\frac{\gamma}{n}}+\frac{100F^2\gamma}{3n}+\|f_n-f_\tau^*\|_\infty\|\hat{f}-f_n\|_{\ell_2}F\Bigg],\end{aligned} \tag{31}$$

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

Furthermore, by Equation (26) in the proof of Theorem 1 in Schmidt-Hieber (2020) and the argument therein, we have that

$$\|f_n-f_\tau^*\|_\infty \le C'\max_{i=0,1,\ldots,q} c^{-\frac{\beta_i^*}{t_i}}n^{-\frac{\beta_i^*}{2\beta_i^*+t_i}}, \tag{32}$$

for positive constants $c$ and $C'$. Hence, combining (31) with (32) we arrive at

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

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

**Lemma 25** *Let $r^*$ be defined as*

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

*for $\{\xi_i\}_{i=1}^n$ Rademacher variables independent of $\{(x_i,y_i)_{i=1}^n\}$. Then under the conditions of Lemma 24,*

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

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

**Proof** Proceeding as in the proof of Lemma 20, we obtain that for a constant $C > 0$,

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

where the second inequality follows from Lemma 5 in Schmidt-Hieber (2020), 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

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

■

## Appendix E. Plots of scenarios from Section 4.1

Here we visualize the performances of different approaches and true quantile functions under scenarios from Section 4.1. The plots are shown in Figures 2–4. There, we can see that Q-Network is a better estimate for the quantile functions in general, compared to the other methods.

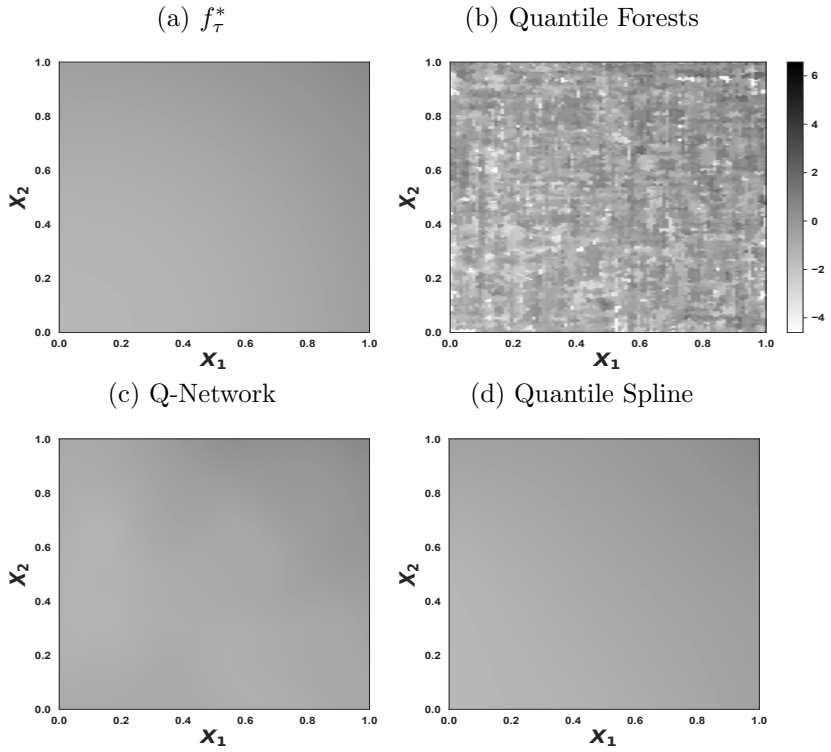


Figure 2: One instance of the true quantile function with $\tau = 0.25$ and its corresponding estimates obtained from different methods. Here $n = 10000$ and the data are generated under Scenario 2.

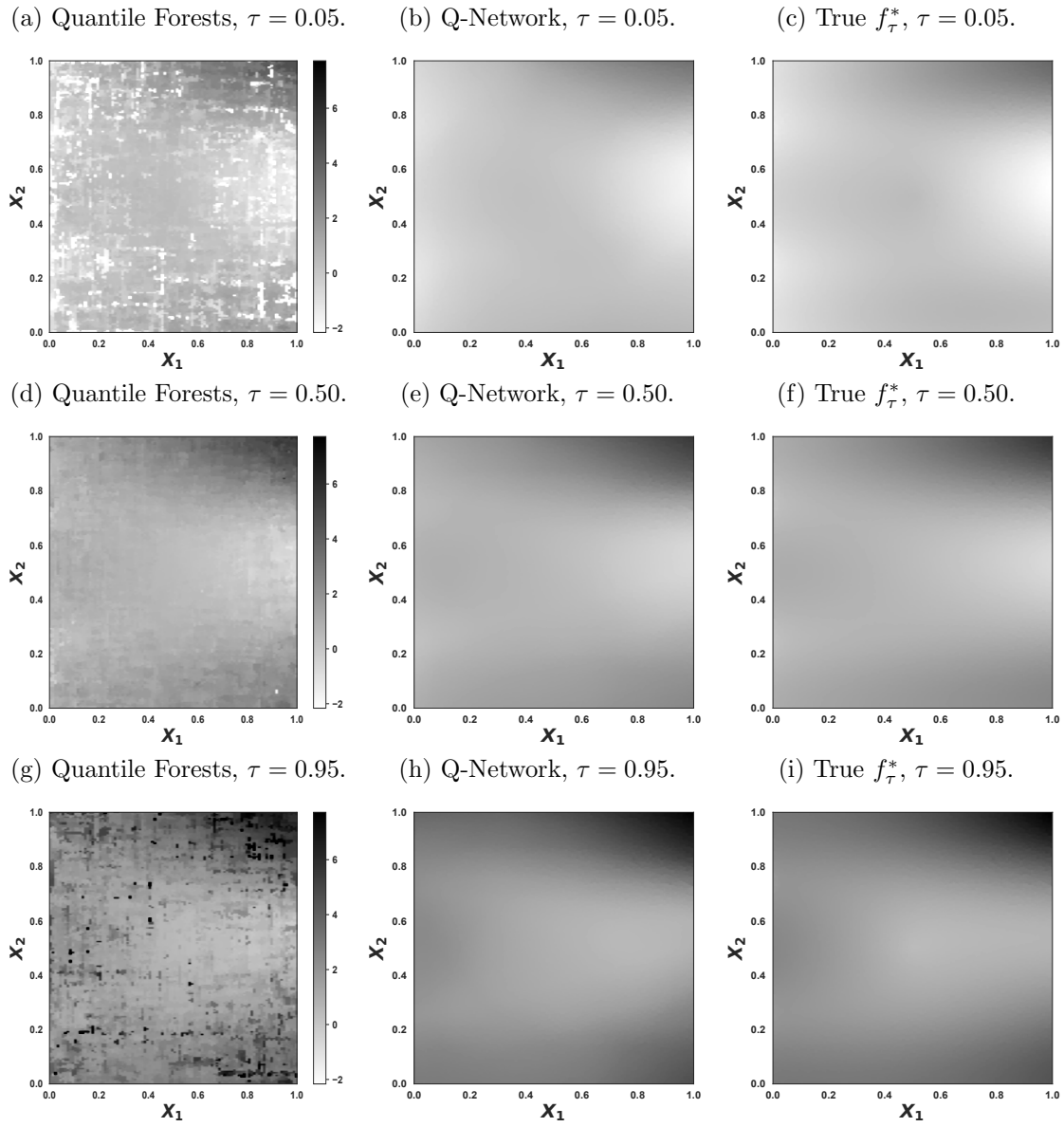


Figure 3: One instance of the true quantile function with $\tau \in \{0.05, 0.50, 0.95\}$ and its corresponding estimates based on Q-Network and Quantile Forests. Here $n = 10000$ and the data are generated under Scenario 1.

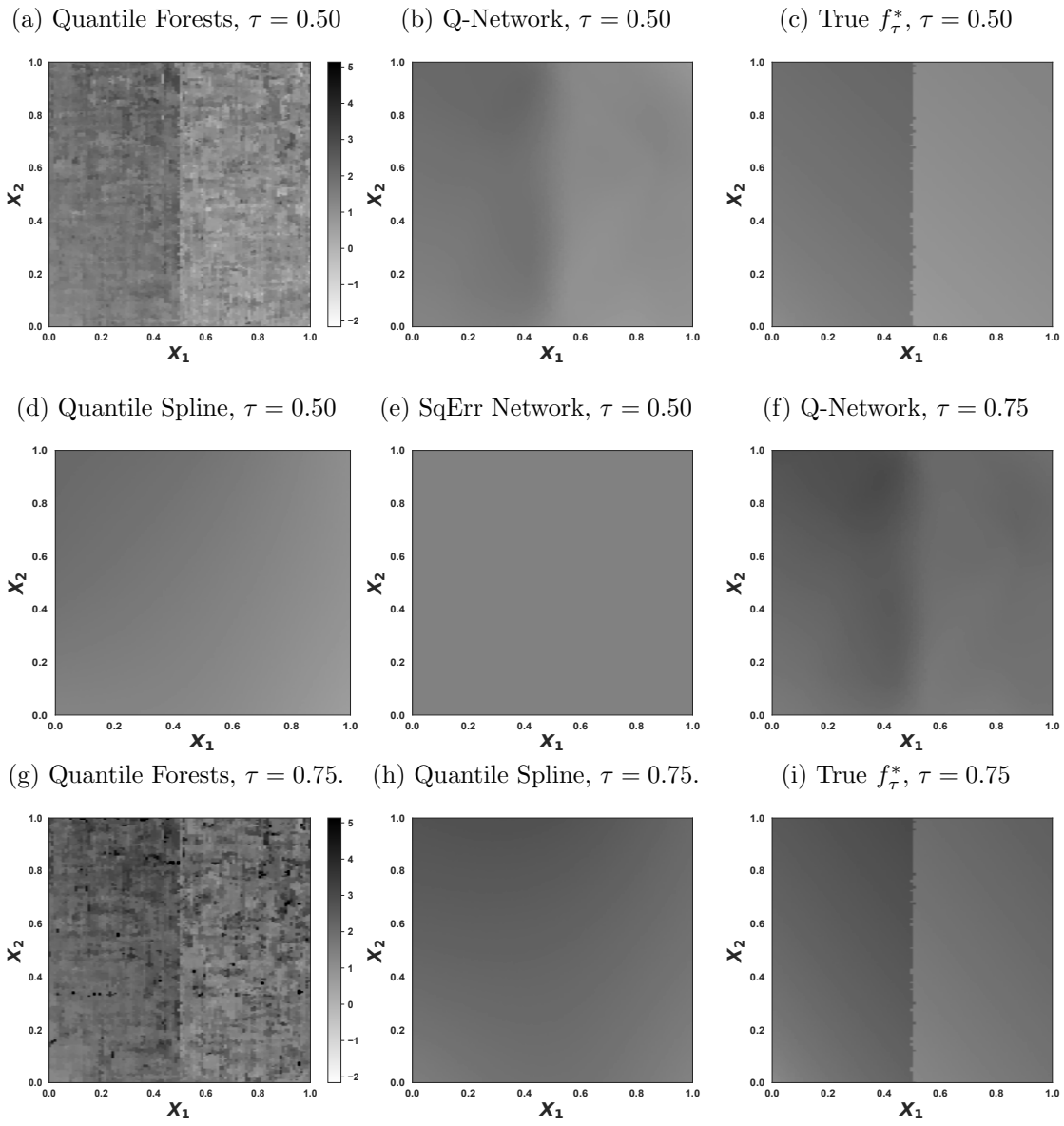


Figure 4: One instance of the true quantile function with $\tau \in \{0.50, 0.75\}$ and of the corresponding estimates obtained with the different methods. Here $n = 10000$ and the data are generated under Scenario 3.

## 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. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Alexandre Belloni and Victor Chernozhukov. $\ell_1$-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.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

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.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

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. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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.