Deep Quantile Regression: Mitigating the Curse of Dimensionality Through Composition

Guohao Shen* Yuling Jiao† Yuanyuan Lin‡ Joel L. Horowitz§ Jian Huang¶

August 3, 2021

Abstract

$f_0 = h_q \circ \cdots \circ h_0$

This paper considers the problem of nonparametric quantile regression under the assumption that the target conditional quantile function is a composition of a sequence of low-dimensional functions. We study the nonparametric quantile regression estimator using deep neural networks to approximate the target conditional quantile function. For convenience, we shall refer to such an estimator as a deep quantile regression (DQR) estimator. We show that the DQR estimator achieves the nonparametric optimal convergence rate up to a logarithmic factor determined by the intrinsic dimension of the underlying compositional structure of the conditional quantile function, not the ambient dimension of the predictor. Therefore, DQR is able to mitigate the curse of dimensionality under the assumption that the conditional quantile function has a compositional structure. To establish these results, we analyze the approximation error of a composite function by neural networks and show that the error rate only depends on the dimensions of the component functions. We apply our general results to several important statistical models often used in mitigating the curse of dimensionality, including the single index, the additive, the projection pursuit, the univariate composite, and the generalized hierarchical interaction models. We explicitly describe the prefactors in the error bounds in terms of the dimensionality of the data and show that the prefactors depends on the dimensionality linearly or quadratically in these models. We also conduct extensive numerical experiments to evaluate the effectiveness of DQR and demonstrate that it outperforms a kernel-based method for nonparametric quantile regression.

N, L
$\downarrow$ determine
width & depth

Keywords: Approximation error; composite function; deep neural networks; nonparametric regression; non-asymptotic error bound.

[*] Equal contribution. Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China. Email: ghshen@link.cuhk.edu.hk

[†] Equal contribution. School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei Province, China 430072. Email: yulingjiaomath@whu.edu.cn

[‡] Department of Statistics, The Chinese University of Hong Kong, Hong Kong, China. Email: ylin@sta.cuhk.edu.hk

[§] Department of Economics, Northwestern University, Evanston, IL 60208, USA. Email: joelhorowitz@northwestern.edu

[¶] Department of Statistics and Actuarial Science, University of Iowa, IA 52242, USA. Email: jianhuang@uiowa.edu

1 Introduction

Consider a nonparametric regression model

$$\overset{\text{response}}{Y} = f_0(\overset{\text{dxn}}{X}) + \eta, \rightarrow \text{error.} \tag{1.1}$$

where $Y \in \mathbb{R}$ is a response variable, $X \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional vector of predictors, $f_0 : \mathcal{X} \to \mathbb{R}$ is an unknown regression function, and η is an error term that may depend on X. We consider the problem of nonparametric quantile regression under the assumption that the underlying regression function is a composition of a sequence of low-dimensional functions. We study the nonparametric quantile regression estimator using deep neural networks to approximate the target regression function. For convenience, we shall refer to such an estimator as a deep quantile regression (DQR) estimator.

Quantile regression (Koenker and Bassett, 1978; Koenker, 2005) is an important method in the toolkit for analyzing the relationship between a response Y and a predictor X. Unlike the least squares regression that models the conditional mean of Y given X, quantile regression estimates the conditional quantiles of Y given X. Thus quantile regression is able to describe the conditional distribution of Y given X. There is a rich literature on quantile regression, much of the work focus on the parametric case when the conditional quantile function is assumed to be a linear function of the predictor. The linear quantile regression has also been studied extensively in the context of regularized estimation and variable selection in the high-dimensional settings (Li and Zhu, 2008; Belloni et al., 2011, 2019; Wang et al., 2012; Zheng et al., 2015, 2018). In addition, there are many important studies on nonparametric quantile regression. Examples include the methods using smoothing splines (Koenker et al., 1994; He and Shi, 1994; He and Ng, 1999) and reproducing kernels (Takeuchi et al., 2006; Sangnier et al., 2016). These studies established the convergence rate of the nonparametric estimators and discussed related problems arising in quantile regression, including an approach to dealing with the quantile crossing problem and a method for incorporating prior qualitative knowledge such as monotonicity constraints in the conditional quantile function estimation. An early study on nonparametric quantile regression using shallow neural networks is White (1992). We refer to Koenker (2005) and the references therein for a detailed treatment of quantile regression. More discussions on nonparametric quantile regression related to this work are given in Section 8.

To give a snapshot of quantile regressions using deep neural networks compared with the traditional linear and the kernel quantile regressions, we look at the fitting of the univariate regression functions "Wave", when the error term follows a "Sine" distribution or conditionally follows a normal distribution $(\eta \mid X = x) \sim 0.5 \times \mathcal{N}(0, [\sin(\pi x)]^2)$. The functional form of the "Wave" function is given in Section 7. Figure 1 presents the fitting results using deep quantile regression (DQR), quantile regression in reproducing kernel Hilbert space (*kernel QR*) in Sangnier et al. (2016) and traditional linear quantile regression (*linear QR*) in Koenker and Bassett (1978) at the 0.25-th, the 0.50-th and the 0.75-th quantiles. Moreover, least squares regression using deep neural networks (DLS) is also compared with the above methods at the 0.50-th quantile. We see that *linear QR* fails when the model is nonlinear, while *kernel QR* and DQR yield acceptable fitting curves. In particular, DQR works best among the methods considered in this example.

Solid – true

Dash – estimate

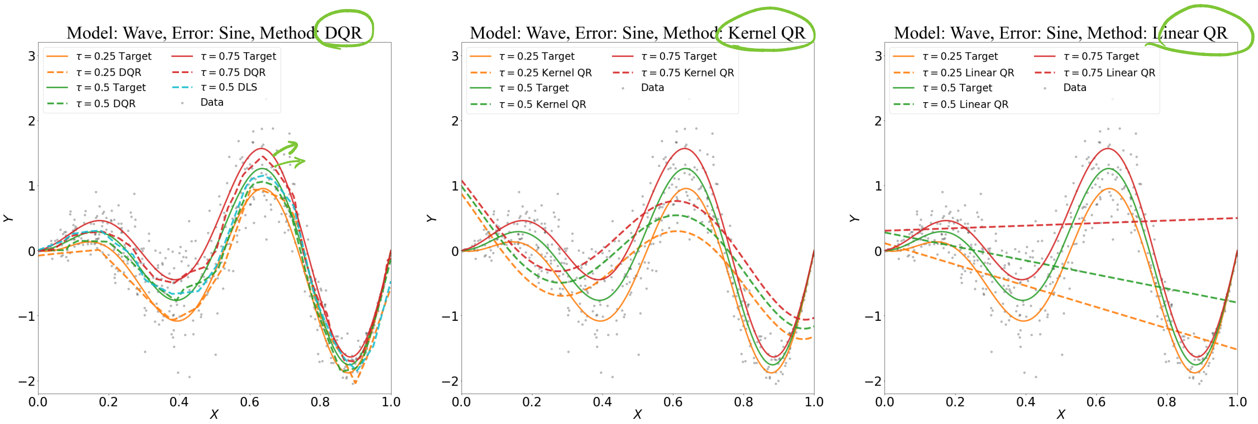

Figure 1: The fitted quantile curves by different methods under the univariate model "Wave" with "Sine" error. The training data is depicted as grey dots.The target quantile functions at the quantile levels τ =0.25 (yellow), 0.5 (green), 0.75 (red) are depicted as solid curves, and the estimated quantile functions are represented by dashed curves with the same color. From the left to right, the subfigures correspond to the methods: *DQR*, *kernel QR* and *linear QR*. The fitted *DLS* curve (in blue) is included in the left subfigure.

In classical nonparametric statistics, including nonparametric quantile regression, the complexity of a function such as regression function and density function is measured through smoothness in terms of the order of the derivatives. The rate of convergence in estimating such functions is determined by the dimension and the smoothness index (Stone, 1982). Specifically, under the assumption that the target function f_0 is in a Hölder class with a smoothness index $\beta > 0$ (β-Hölder smooth), i.e., all the partial derivatives up to order $\lfloor\beta\rfloor$ exist and the partial derivatives of order $\lfloor\beta\rfloor$ are $\beta-\lfloor\beta\rfloor$ Hölder continuous, where $\lfloor\beta\rfloor$ denotes the largest integer strictly smaller than β, the optimal convergence rate of the prediction error is $C_d n^{-\beta/(2\beta+d)}$ under mild conditions (Stone, 1982), where C_d is a prefactor independent of n but depending on d and other model parameters. When d is small, say, $d = 2$, assuming the target function has a continuous second derivative, the optimal rate of convergence is $C_d n^{-1/3}$. Therefore, in the low-dimensional settings, a sufficient degree of smoothness will overcome the adverse impact of the dimensionality on the convergence rate. Moreover, in low-dimensional models with a small d, the impact of C_d on the convergence rate is not significant. However, in high-dimensional models with a large d, the situation is completely different. First, the rate of convergence can be painfully slow, unless the function f_0 is assumed to have an extremely large smoothness index β. But such an assumption is not realistic in practice. Second, the impact of C_d can be substantial when d is large. For example, if the prefactor C_d depends on d exponentially, it can overwhelm the convergence rate $n^{-\beta/(2\beta+d)}$. Therefore, it is important to clearly describe how C_d depends on the dimensionality.

$f_0 \rightarrow \alpha$-th lipschitz continuity

minimax rate: $\left(\frac{\sigma^2}{n}\right)^{\frac{2\beta}{2\beta+d}}$

$\exp(d)$

① worse cov-rate

② Cd depend on d, need to describe clearly for high dimensional model

Recently, several authors carried out important and inspiring studies on the convergence properties of least squares nonparametric estimation using neural network approximation of the regression function (Bauer and Kohler, 2019; Schmidt-Hieber et al., 2020; Chen et al., 2019a; Kohler et al., 2019; Nakada and Imaizumi, 2019; Farrell et al., 2021). These studies show that deep neural network regression can achieve the minimax optimal rate of convergence up to a logarithmic factor for estimating the conditional mean regression function established by Stone (1982). However, nonparametric estimation using deep neural networks

① log N covering number

d t·

$f_0 = h_q \circ \cdots \circ \boxed{h_0}$

cannot escape the well-know problem of *curse of dimensionality* in high-dimensions without any conditions on the underlying model.

② It is clear that smoothness is not the right measure of the complexity of a function class in the high-dimensional settings, since smoothness does not help mitigate the curse of dimensionality. An effective approach to mitigating the curse of dimensionality is to consider functions with a compositional structure. Deep neural network modeling has achieved impressive success and often outperformed kernel based methods in many important applications with high-dimensional data, including speech recognition, image classification, object detection, drug discovery and genomics, among others (LeCun et al., 2015). Thus it is desirable to consider statistical models in a function class that can mitigate the curse of dimensionality and can be well approximated by deep neural networks. It has been shown that deep ReLU networks are solutions to regularized data fitting problems in the function space consisting of compositions of functions from the Banach spaces of second-order bounded variation (Parhi and Nowak, 2021). Using composite functions in nonparametric regression modeling has a long history in statistics. For example, the nonparametric additive model, which can be considered a composition of a linear function with a vector function whose components depend on only one of the variables, has been studied by many authors (Breiman and Friedman, 1985; Stone, 1985, 1986; Hastie and Tibshirani, 1990). Recently, more general composite functions for statistical modeling have been proposed in several interesting works (Horowitz and Mammen, 2007; Bauer and Kohler, 2019; Schmidt-Hieber et al., 2020). Under this assumption, the convergence rate $C_d n^{-\beta/(2\beta+d)}$ could be improved to $C_{d,d_*} n^{-\beta/(2\beta+d_*)}$ for some $d_* \ll d$, where C_{d,d_*} is a constant depending on (d_*, d), where d_* is the intrinsic dimension of the model. In these results, the convergence rate part is improved from $n^{-\beta/(2\beta+d)}$ to $n^{-\beta/(2\beta+d_*)}$. When $d_* \ll d$, the improvement is substantial. However, the prefactor C_{d,d_*} in the error bounds depends on d exponentially or are not clearly described in the aforementioned works (Stone, 1985, 1986; Horowitz and Mammen, 2007; Bauer and Kohler, 2019; Schmidt-Hieber et al., 2020). In a low-dimensional model with a small d, the impact of the prefactor on the overall error bound is not significant. However, in a high-dimensional model with a large d, the impact of the prefactor can be substantial, even overwhelm the convergence rate part (Ghorbani et al., 2020). Therefore, it is important to describe how the prefactor depends on the dimension d in the error bound.

combine several function together

In this paper, we establish non-asymptotic upper bounds for the excess risk and mean integrated squared error of the *DQR* estimator under the assumption that the target regression function is a composite function. A novel aspect of our work is that we clearly describe how the prefactors in the error bounds depend on the ambient dimension d and the dimensions of the low-dimensional component functions of the composite function. Our error bounds achieve the minimax optimal rates and significantly improve over the existing ones in the sense that their prefactors depend linearly or quadratically on the dimension d, instead of exponentially on d. This shows that *DQR* can mitigate the curse of dimensionality under the assumption that the target regression function belongs to the class of composite functions. These results are based on new approximation error bounds of composite functions by the neural networks, which may be of independent interest. Our main contributions are as follows.

1. We establish excess risk bounds for the proposed *DQR* estimator under the assumption

that the target conditional quantile function has a compositional structure with lower-dimensional component functions. With appropriately specified ReLU networks in terms of depth, width and size of the network, our *DQR* estimator achieves near optimal convergence rate up to a logarithmic factor under a heavy-tailed error (finite p-th moment for $p \geq 1$) and mild regular conditions on the joint distribution of the response and the predictor. Moreover, we show that DQR can mitigate the curse of dimensionality in the sense that the convergence rate of the error bound depends on the dimensions of the component functions, not the ambient dimension. We also show that the prefactors of the error bounds depend on the ambient dimension linearly or quadratically.

2. We derive novel approximation error results of composite functions using ReLU activated neural networks under the assumption that the component functions are Hölder continuous. This result shows that the curse of dimensionality can be mitigated through composition in the sense the approximate error rate depends on the intrinsic dimension of a composite functions, instead of the ambient dimension of the function. Equally importantly, the prefactor of the error bound is significantly improved in the sense that it depends on the dimensionality d polynomially instead of exponentially as in the existing results. This approximation result is the key building block in establishing the bounds for excess risk and mean integrated squared error for *DQR*.

3. We apply our general results to several important statistical models often used in mitigating the curse of dimensionality, including the single index, the additive, the projection pursuit, the univariate composite, and the generalized hierarchical interaction models. We show that *DQR* achieves the optimal convergence rate up to a logarithmic factor under these models. We also present the prefactors of the error bounds for these models.

4. We bridge the gap between the excess risk and the mean integrated squared error of the *DQR* estimator under mild conditions. We do not require the bounded support condition on the conditional distribution of the response given the predictor as in the existing literature. The mean integrated squared error of our *DQR* estimator is shown to converge at the near optimal rate up to a logarithmic factor, inheriting the properties of the corresponding excess risk. The convergence rate of the mean integrated squared error of the *DQR* estimator is determined by the dimensions of the component functions and the prefactor depends polynomially on the widest layer of the composite functions.

The remainder of this paper is organized as follows. In Section 2 we describe the deep quantile regression problem, the deep neural networks used in the estimation and the assumption on the compositional structure of the conditional quantile function. In Section 3 we provide a high level description of our main results and the overall approach we take to establish these results. In Section 4 we present non-asymptotic bounds on the excess risk and mean integrated squared error of the *DQR* estimator. Section 5 includes applications of our general error bounds to several important models in nonparametric statistics. In Section 6 we present a result on the approximation error of composite functions using deep neural networks. In Section 7 we present simulation results demonstrating that *DQR* outperforms a kernel nonparametric quantile regression method based on vector-valued reproducing kernel Hilbert space (RKHS) (Sangnier et al., 2016). Section 8 contains discussions on the related work. Concluding remarks are given in Section 9. Proofs and additional simulation results are given in the appendix.

2 Deep quantile regression

In this section, we present the basic setup of nonparametric regression. We describe the structure of the feedforward neural networks to be used in the estimation and define the compositional structure for the target conditional quantile function.

For a given quantile level $\tau \in (0,1)$, the quantile check loss function is defined by

$$\rho_\tau(x) = x\{\tau - I(x \leq 0)\}, \ x \in \mathbb{R}.$$

$I_{x>0}\, \tau \cdot |x| + I_{x\leq 0}\, (1-\tau)\cdot |x|$

For a possibly random function $f : \mathbb{R}^d \to \mathbb{R}$, let $Z \equiv (X, Y)$ be a random vector independent of f. We define the risk of f under the loss function $\rho_\tau(\cdot)$ by

$$\mathcal{R}^\tau(f) = \mathbb{E}_Z\{\rho_\tau(Y - f(X))\}.$$

At the population level, the nonparametric quantile estimation is to find a measurable function $f^* : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f^* := \arg\min_f \mathcal{R}^\tau(f) = \arg\min_f \mathbb{E}_Z\{\rho_\tau(Y - f(X))\},$$

where $\mathbb{E}_Z$ means that the expectation is taken with respect to the distribution of Z. If the conditional τ-th quantile of η given X is 0 and $\mathbb{E}(|\eta||X = x) < \infty$ for all $x \in \mathcal{X}$, then the true regression function f_0 is the optimal solution f^* on $\mathcal{X}$.

In applications, when only a random sample $S \equiv \{(X_i, Y_i)\}_{i=1}^n$ is available, we consider the empirical risk

$$\mathcal{R}_n^\tau(f) = \frac{1}{n}\sum_{i=1}^n \rho_\tau(Y_i - f(X_i)). \tag{2.1}$$

Our goal is to construct an estimator of f_0 within a certain class of functions $\mathcal{F}_n$ by minimizing the empirical risk, that is,

$$\hat{f}_n \in \arg\min_{f\in\mathcal{F}_n} \mathcal{R}_n^\tau(f), \tag{2.2}$$

where $\hat{f}_n$ is called the empirical risk minimizer (ERM). We choose $\mathcal{F}_n$ to be a function class consisting of deep neural networks (DNN). We will also refer to $\hat{f}_n$ as a deep quantile regression (DQR) estimator below.

2.1 Deep neural networks

We set the function class $\mathcal{F}_n$ to be $\mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$, a class of feedforward neural networks $f_\phi : \mathbb{R}^d \to \mathbb{R}$ with parameter ϕ, depth $\mathcal{D}$, width $\mathcal{W}$, size $\mathcal{S}$, number of neurons $\mathcal{U}$ and f_ϕ satisfying $\|f_\phi\|_\infty \leq \mathcal{B}$ for some $0 < B < \infty$, where $\|f\|_\infty$ is the supreme norm of a function $f : \mathbb{R}^d \to \mathbb{R}$. Note that the network parameters may depend on the sample size n, but the dependence is omitted in the notation for simplicity. A brief description of multilayer perceptrons (MLPs), the commonly used feedforward neural networks, are given below. The architecture of a MLP can be expressed as a composition of a series of functions

$$f_\phi(x) = \mathcal{L}_\mathcal{D} \circ \sigma \circ \mathcal{L}_{\mathcal{D}-1} \circ \sigma \circ \cdots \circ \sigma \circ \mathcal{L}_1 \circ \sigma \circ \mathcal{L}_0(x), \;\; x \in \mathbb{R}^d,$$

where $\sigma(x) = \max(0, x)$ is the rectified linear unit (ReLU) activation function (defined for each component of x if x is a vector) and

$$\mathcal{L}_i(x) = W_i x + b_i, \quad i = 0, 1, \ldots, \mathcal{D},$$

where $W_i \in \mathbb{R}^{d_{i+1} \times d_i}$ is a weight matrix, d_i is the width (the number of neurons or computational units) of the i-th layer, and $b_i \in \mathbb{R}^{d_{i+1}}$ is the bias vector in the i-th linear transformation $\mathcal{L}_i$.

output

Such a network f_ϕ has $\mathcal{D}$ hidden layers and $(\mathcal{D}+1)$ layers in total. We use a $(\mathcal{D}+1)$-vector $(w_0, w_1, \ldots, w_\mathcal{D})^\top$ to describe the width of each layer; particularly in nonparametric regression problems, $w_0 = d$ is the dimension of the input and $w_\mathcal{D} = 1$ is the dimension of the response . The width $\mathcal{W}$ is defined as the maximum width of hidden layers, i.e., $\mathcal{W} = \max\{w_1, \ldots, w_\mathcal{D}\}$; the size $\mathcal{S}$ is defined as the total number of parameters in the network f_ϕ, i.e., $\mathcal{S} = \sum_{i=0}^{\mathcal{D}}\{w_{i+1} \times (w_i + 1)\}$; the number of neurons $\mathcal{U}$ is defined as the number of computational units in hidden layers, i.e., $\mathcal{U} = \sum_{i=1}^{\mathcal{D}} w_i$. For an MLP $\mathcal{F}_{\mathcal{D},\mathcal{U},\mathcal{W},\mathcal{S},\mathcal{B}}$, its parameters satisfy the simple relationship

input

$$\max\{\mathcal{W}, \mathcal{D}\} \leq \mathcal{S} \leq \mathcal{W}(\mathcal{D}+1) + (\mathcal{W}^2 + \mathcal{W})(\mathcal{D}-1) + \mathcal{W} + 1 = O(\mathcal{W}^2\mathcal{D}).$$

2.2 Structured composite functions

Let the target quantile regression function $f_0 : \mathbb{R}^d \to \mathbb{R}$ be a d-dimensional function. We assume that f_0 is a composition of a series of functions $h_i, i = 0 \ldots, q$, i.e.,

$$f_0 = h_q \circ \cdots \circ h_0,$$

$t_i \to h_i$

where $h_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$. Here $d_0 = d$ and $d_{q+1} = 1$. For each h_i, denote by $h_i = (h_{ij})^\top_{j=1,\ldots,d_{i+1}}$ the components of h_i and let t_i be the maximal number of variables on which each of h_{ij} the depends on. Note that $t_i \leq d_i$ and each h_{ij} is a t_i-variate function for $j = 1, \ldots, d_i$.

Many well-known important models in semiparametric and nonparametric statistics have a compositional structure. Examples include the single index model (Härdle et al., 1993; Horowitz and Härdle, 1996), the additive model (Stone, 1985, 1986; Hastie and Tibshirani, 1990), the projection pursuit model (Friedman and Stuetzle, 1981), the interaction model (Stone, 1994), the composite regression model (Horowitz and Mammen, 2007), and the generalized hierarchical interaction model (Bauer and Kohler, 2019). We consider the bounds for the excess risk of *DQR* under these models in Section 5.

In this work, we focus on the quantile regression models in which the conditional quantile function has a compositional structure. This is the key condition we use to mitigate the curse of dimensionality. We will only assume the Hölder continuity on the component functions of the composite conditional quantile function. A function $h : [a_1, b_1]^{d_1} \to [a_2, b_2]^{d_2}$ is said to be Hölder continuous with order α and Hölder constant λ if there exist $\alpha \in (0, 1]$ and $\lambda \geq 0$ such that

$$\|h(x) - h(y)\|_2 \leq \lambda \|x - y\|_2^{\alpha} \tag{2.3}$$

for any $x, y \in [a_1, b_1]^{d_1}$.

We now describe the assumptions on the target regression function f_0 in detail below.

f_0 have such de component

Assumption 1 (Structured target regression function with continuous components)**.** *The target quantile regression function $f_0 = h_q \circ \cdots \circ h_0$ is a composition of a series of functions $h_i, i = 0 \ldots, q$, where $h_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}}$ with $d_0 = d$ and $d_{q+1} = 1$. For each $h_i = (h_{ij})^{\top}_{j=1,\ldots,d_{i+1}}$ $(i = 0, \ldots, q)$, its components $h_{ij} : [a_i, b_i]^{t_i} \to [a_{i+1}, b_{i+1}]$ $(j = 1, \ldots, d_{i+1})$ are Hölder continuous functions with order $\alpha_i \in [0, 1]$ and constant $\lambda_i \geq 0$, where t_i is the maximal number of variables on which each of h_{ij} depends on $(t_i \leq d_i)$. Let $J \subset \{0, \ldots, q\}$ be a set consisting of the indices of linear transformation layers of f_0 (if any) and $J^c := \{0, \ldots, q\} \backslash J$ denote the complement of J.*

We will show that, if the target regression function f_0 satisfies Assumption 1, the *DQR* estimator can automatically adapt to the compositional structure and circumvent the curse of dimensionality.

3 A high-level description of the results

In this section, we present a high-level description of our approach, the non-asymptotic bounds for the excess risk and the mean integrated squared error of the *DQR* estimator. Detailed statements of the results and the assumptions are given in the Sections 4-6 below.

For a *DQR* estimator $\hat{f}_n \in \mathcal{F}_n$ defined in (2.2), we evaluate its quality via the *excess risk*, defined as the difference between the risks of $\hat{f}_n$ and f_0,

$$\mathcal{R}^{\tau}(\hat{f}_n) - \mathcal{R}^{\tau}(f_0) = \mathbb{E}_Z \rho_{\tau}(\hat{f}_n(X) - Y) - \mathbb{E}_Z \rho_{\tau}(f_0(X) - Y).$$

We first establish an upper bound on the excess risk, which is the starting point of our error analysis.

Lemma 1. *For any random sample $S = \{(X_i, Y_i)_{i=1}^n\}$, the excess risk of the DQR estimator $\hat{f}_n$ satisfies*

$$\mathcal{R}^{\tau}(\hat{f}_n) - \mathcal{R}^{\tau}(f_0) \leq 2 \sup_{f \in \mathcal{F}_n} |\mathcal{R}^{\tau}(f) - \mathcal{R}^{\tau}_n(f)| + \inf_{f \in \mathcal{F}_n} \mathcal{R}^{\tau}(f) - \mathcal{R}^{\tau}(f_0), \tag{3.1}$$

where $\mathcal{R}^{\tau}_n$ is defined in (2.1).

The excess risk of the *DQR* estimator is bounded above by the sum of two terms: the stochastic error $2 \sup_{f \in \mathcal{F}_n} |\mathcal{R}^{\tau}(f) - \mathcal{R}^{\tau}_n(f)|$ and the approximation error $\inf_{f \in \mathcal{F}_n} \mathcal{R}^{\tau}(f) - \mathcal{R}(f_0)$. It is interesting to note that the upper bound no longer depends on the *DQR* estimator itself, but the function class $\mathcal{F}_n$, the loss function ρ_{τ} and the random sample S.

(1) The stochastic error $2 \sup_{f \in \mathcal{F}_n} |\mathcal{R}^{\tau}(f) - \mathcal{R}^{\tau}_n(f)|$ can be analyzed using the empirical process theory (Van der Vaart and Wellner, 1996; Anthony and Bartlett, 1999; Bartlett et al.,

2019). A key step is to calculate the complexity measure of $\mathcal{F}_n$ in terms of its covering number. The details are given in Section 4.

2 The approximation error term $\inf_{f\in\mathcal{F}_n}\mathcal{R}^\tau(f)-\mathcal{R}^\tau(f_0)$ measures the approximation error of the function class $\mathcal{F}_n$ for f_0 under the loss function ρ_τ. To utilize the approximation theories of neural networks, we need to relate $\inf_{f\in\mathcal{F}_n}\mathcal{R}^\tau(f)-\mathcal{R}^\tau(f_0)$ to the quantity $\inf_{f\in\mathcal{F}_n}\|f-f_0\|$ for some functional norm $\|\cdot\|$. The power of neural network functions approximating high-dimensional functions have been studied by many authors, some recent works include Yarotsky (2017, 2018); Shen et al. (2019, 2020), among others. For a composite function f_0 under Assumption 1, we derive new approximation results in Section 6.

To clearly describe how the error bounds depend on various parameters, including the network parameters such as depth, width and size of the network, as well as the model parameters such as the intrinsic and ambient dimensions of the model, we present general expressions of the stochastic errors and the approximation errors, which constitute the upper bounds for the excess risk and the mean integrated squared error (MISE), in Theorems 1 and 2 in Section 4 below. The network parameters, similar to the bandwidth in kernel nonparametric regression or density estimation, can be tuned as a function of the sample size and the model dimension to obtain the best trade-off between the stochastic error and the approximation error, and therefore achieve the best overall error rate. An appealing aspect of our results is that they clearly and explicitly describe how the prefactors in the error bounds depend on the network parameters and the dimensionality of the model. Explicit expressions of the bounds for the excess risk and the MISE are presented in Corollaries 2 and 3 in Section 4.

In Section 5, we consider several well-known semiparametric and nonparametric models that are widely used to mitigate the curse of dimensionality, including the single index model, the additive model, the projection pursuit model, the interaction model, the univariate composite regression model, and the generalized hierarchical interaction model. We derive explicit expressions of the error bounds when the underlying conditional quantile function takes the form of these well-known models

As can be seen in Corollary 2 for the excess risk of the DQR estimator and the error bounds for the models considered in Section 5, based on appropriately specified network parameters (depth, width and size of the network), we have the following upper bound for the excess risk,

$$\mathbb{E}\big\{\mathcal{R}^\tau(\hat{f}_\phi)-\mathcal{R}^\tau(f_0)\big\}\leq C_0C_{d,d^*}(\log n)^2 n^{-\left(1-\frac{1}{p}\right)\frac{2\alpha^*}{2\alpha^*+t^*}}, \tag{3.2}$$

where C_0 is a constant only depending on the model parameters such as the smoothness index of the underlying conditional quantile function, C_{d,d^*} is the prefactor depending on d, the dimension of the predictor; and d^*, determined by the dimensions of the component functions in the composite function. The convergence rate part of the error bound (3.2), $n^{-(1-1/p)2\alpha^*/(2\alpha^*+t^*)}$, is determined by the number of moments p of the response Y (see Assumption 2 below), the smoothness index of the composite function α^*, and the intrinsic dimension of the model t^*. If Y has sub-exponential tail probabilities, we can set $p=\infty$. The bound for the mean integrated squared error of the DQR estimator has a form similar to (3.2), see Corollary 3.

Explicit expressions for C_{d,d^*} in (3.2) are given in Corollaries 2 and 3, as well as for the examples in Section 5. For example, for the single index model (5.1), the additive model (5.2) and the additive model with an unknown link function (5.3), $C_{d,d^*} = d^2 \log d$. For the interaction model (5.4), $C_{d,d^*} = (Kdd^*)^2 \log(Kdd^*)$, where K is the number of component functions and d^* is the dimension of the component functions in the model. For the projection pursuit model (5.5), $C_{d,d^*} = (\max\{K,d\})^2 \log(\max\{K,d\})$, where K is the number of component functions in the model. For the univariate composite model (5.6) and the generalized hierarchical interaction model (5.8), the forms of C_{d,d^*} are more complicated, they are given in Section 5.

These results demonstrate that *DQR* with deep neural networks can significantly attenuate the curse of dimensionality when the underlying conditional quantile function takes the form of one of these models, even though the construction of the *DQR* estimator does not use the specific structure of these models.

4 Non-asymptotic error bounds

In this section, we present non-asymptotic error bounds for the *DQR* estimator, including bounds for the excess risk upper bounds in section 4.1 and bounds for mean integrated squared error in 4.2. The bounds are determined by a trade-off between the stochastic error and the approximation error.

4.1 Excess risk bounds

For analyzing the stochastic error of the *DQR* estimator, we make the following assumption.

Assumption 2. *(i) The conditional τ-th quantile of η given $X = x$ is 0 and $\mathbb{E}(|\eta||X = x) < \infty$ for almost every $x \in \mathcal{X}$. (ii) The support of covariates $\mathcal{X}$ is a bounded compact set in $\mathbb{R}^d$, and without loss of generality $\mathcal{X} = [0,1]^d$. (iii) The response variable Y has a finite p-th moment for some $p > 1$, i.e., there exists a finite constant $M > 0$ such that $\mathbb{E}|Y|^p \leq M$.*

Note that throughout the paper, we focus on the case when $\mathcal{X} = [0,1]^d$. In the nonparametric regression problems, we can always first transform the predictors to a bounded region.

largest breakpoint

For a class $\mathcal{F}$ of functions: $\mathcal{X} \to \mathbb{R}$, its pseudo dimension, denoted by $\mathrm{Pdim}(\mathcal{F})$, is defined to be the largest integer m for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for any $(b_1, \ldots, b_m) \in \{0,1\}^m$ there exists $f \in \mathcal{F}$ such that $\forall i : f(x_i) > y_i \iff b_i = 1$ (Anthony and Bartlett, 1999; Bartlett et al., 2019). For a class of real-valued functions generated by neural networks, pseudo dimension is a natural measure of its complexity. In particular, if $\mathcal{F}$ is the class of functions generated by a neural network with a fixed architecture and fixed activation functions, we have $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathcal{F})$ (Theorem 14.1 in Anthony and Bartlett (1999)), where $\mathrm{VCdim}(\mathcal{F})$ is the VC dimension of $\mathcal{F}$. In our results, we require the sample size n to be greater than the pseudo dimension of the class of neural networks considered.

For a given sequence $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $\mathcal{F}_\phi|_x = \{(f(x_1), \ldots, f(x_n) : f \in \mathcal{F}_\phi\} \subset \mathbb{R}^n$. For a positive number δ, let $\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{F}_\phi|_x)$ be the covering number of $\mathcal{F}_\phi|_x$ under the norm $\|\cdot\|_\infty$ with radius δ. Define the uniform covering number $\mathcal{N}_n(\delta, \|\cdot\|_\infty, \mathcal{F}_\phi)$ to be the maximum over all $x \in \mathcal{X}$ of the covering number $\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{F}_\phi|_x)$, i.e.,

$$\mathcal{N}_n(\delta, \|\cdot\|_\infty, \mathcal{F}_\phi) = \max\{\mathcal{N}(\delta, \|\cdot\|_\infty, \mathcal{F}_\phi|_x) : x \in \mathcal{X}\}. \tag{4.1}$$

① We give an upper bound of the stochastic error in the following lemma. $f^*_{\beta_n}$ $Y \geq \beta_n$. β_n -

Lemma 2. *Consider the d-variate nonparametric regression model in (1.1) with an unknown regression function f_0. Let $\mathcal{F}_\phi = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ be a class of feedforward neural networks with a continuous piecewise-linear activation function of finite pieces and $\hat{f}_\phi \in \arg\min_{f\in\mathcal{F}_\phi} R^\tau_n(f)$ be the empirical risk minimizer over $\mathcal{F}_\phi$. Assume that Assumption 2 holds and $\|f_0\|_\infty \leq \mathcal{B}$ for $\mathcal{B} \geq 1$. Then, for $2n \geq Pdim(\mathcal{F}_\phi)$ and any $\tau \in (0, 1)$,* $p\uparrow$ tighter bound

→ upperbound

$$\sup_{f\in\mathcal{F}_\phi} |\mathcal{R}^\tau(f) - \mathcal{R}^\tau_n(f)| \leq c_0 \frac{\max\{\tau, 1-\tau\}\mathcal{B}}{n^{1-1/p}} \log \mathcal{N}_{2n}(n^{-1}, \|\cdot\|_\infty, \mathcal{F}_\phi), \tag{4.2}$$

where $c_0 > 0$ is a constant independent of $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$. Moreover,

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C_0 \frac{\max\{\tau, 1-\tau\}\mathcal{B}\mathcal{S}\mathcal{D}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + 2 \inf_{f\in\mathcal{F}_\phi} \{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\}, \tag{4.3}$$

where $C_0 > 0$ is a constant independent of $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{W}$ and $\mathcal{D}$.

Remark 1. *The denominator $n^{1-1/p}$ in (4.2) and (4.3) can be improved to n if the response Y is assumed to be sub-exponentially distributed, i.e., there exists a constant $\sigma_Y > 0$ such that $\mathbb{E}\exp(\sigma_Y|Y|) < \infty$. This corresponds to the case that $p = +\infty$.*

The stochastic error is bounded by a term determined by the metric entropy of $\mathcal{F}_\phi$ in (4.2), which is measured by the covering number of $\mathcal{F}_\phi$. To obtain (4.3), we further bound the covering number of $\mathcal{F}_\phi$ by its pseudo dimension (VC dimension). According to Bartlett et al. (2019), the pseudo dimension (VC dimension) of $\mathcal{F}_\phi$ with piecewise-linear activation function can be further contained and expressed in terms of its parameters $\mathcal{D}$ and $\mathcal{S}$, i.e., $\text{Pdim}(\mathcal{F}_\phi) = O(\mathcal{S}\mathcal{D}\log(\mathcal{S}))$. This leads to the upper bound for the prediction error by the sum of the stochastic error and the approximation error of $\mathcal{F}_\phi$ to f_0 in (4.3).

② To derive an upper bound for the approximation error $\inf_{f\in\mathcal{F}_\phi}\{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\}$, we first bound it in terms of $\inf_{f\in\mathcal{F}_\phi} \|f - f_0\|$ for some functional norm $\|\cdot\|$. In the following, we let ν denote the marginal distribution of X and define $\|f - f_0\|_{L^p(\nu)} := \{\mathbb{E}|f(X) - f_0(X)|^p\}^{1/p}$ for $p \in (0, \infty)$.

Lemma 3. *Assume that Assumption 2 (i) holds. Let f_0 be the target function defined in (1.1) and $\mathcal{R}^\tau(f_0)$ be its risk. Then, we have*

$$\inf_{f\in\mathcal{F}_\phi} \{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\} \leq \max\{\tau, 1-\tau\} \inf_{f\in\mathcal{F}_\phi} \mathbb{E}|f(X) - f_0(X)| = \max\{\tau, 1-\tau\} \inf_{f\in\mathcal{F}_\phi} \|f - f_0\|_{L^1(\nu)},$$

where ν denotes the marginal distribution of X.

As a consequence of Lemma 3, we only need to give upper bounds on the approximation error $\inf_{f\in\mathcal{F}_\phi} \|f - f_0\|_{L^1(\nu)}$ to give the overall bounds on the excess risk of the ERM $\hat{f}_\phi$ defined in (2.2). Furthermore, if the conditional distributions of error given covariates satisfy proper conditions and the risk function $\mathcal{R}(\cdot)$ has a local quadratic approximation around f_0, the convergence rate results can be further improved.

excess risk for close part .

Assumption 3 (Local quadratic bound of the excess risk). *There exist some constants $c_\tau^0 = c_\tau^0(\tau, X, \eta, f_0) > 0$ and $\delta_\tau^0 = \delta_\tau^0(\tau, X, \eta, f_0) > 0$ which may depend on τ, X, η and f_0 such that*

$$\boxed{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0) \leq c_\tau^0 \|f - f_0\|^2_{L^2(\nu)},}$$

for any f satisfying $\|f - f_0\|_{L^\infty(\mathcal{X}^0)} \leq \delta_\tau^0$, where $\mathcal{X}^0$ is any subset of $\mathcal{X}$ such that $P(X \in \mathcal{X}^0) = P(X \in \mathcal{X})$.

Remark 2. *Assumption 3 is generally satisfied when the conditional density of η given $X = x$ is positive in a neighborhood of its τ-th conditional quantile.*

By Lemma 3 and Assumption 3, a sharper bound for the approximation error improves over that of Lemma 3 can be obtained and presented in the next lemma.

Lemma 4. *Assume that Assumption 2 (i) and 3 hold, let f_0 be the target function defined in (1.1) and $\mathcal{R}^\tau(f_0)$ be its risk, then we have*

$$\inf_{f \in \mathcal{F}_\phi} \{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\} \leq c_\tau \inf_{f \in \mathcal{F}_\phi} \|f - f_0\|^2_{L^2(\nu)},$$

replace $L^1(\nu)$ by $L^2(\nu)$

where $c_\tau \geq \max\left\{c_\tau^0, \max\{\tau, 1-\tau\}/\delta_\tau^0\right\} > 0$ is a constant, ν denotes the marginal probability measure of X and $\mathcal{F}_\phi = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ denotes the class of feedforward neural networks with parameters $\mathcal{D}, \mathcal{W}, \mathcal{U}, \mathcal{S}$ and $\mathcal{B}$.

Remark 3. *We establish the error bounds for approximating a composite function using deep neural networks in Theorem 3 in Section 6. Theorem 3 can be used to bound the approximation error term $\inf_{f \in \mathcal{F}_\phi} \|f - f_0\|_{L^2(\nu)}$ in Lemmas 3 and 4, which leads to the bound for the approximation error in Theorem 1 below.*

nonlinear part related to N_i, L_i intrinsic parameter

Before stating the results for the excess risk bounds, we specify the network parameters. For any given $N_i, L_i \in \mathbb{N}^+, i \in J^c$, we set the function class $\mathcal{F}_\phi = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ consisting of ReLU multi-layer perceptrons with width no more than $\mathcal{W}$ and depth $\mathcal{D}$, where

$$\mathcal{W} = \max_{i=0,\ldots,q} d_i \max\{4t_i \lfloor N_i^{1/t_i} \rfloor + 3t_i, 12N_i + 8\}, \tag{4.4}$$

$$\mathcal{D} = \sum_{i \in J^c} (12L_i + 15) + 2|J|. \tag{4.5}$$

Here recall $J \subset \{0, \ldots, q\}$ is a set collecting the indices of linear layers of f_0 (if any) and $J^c := \{0, \ldots, q\} \backslash J$ denotes the complement of J.

Theorem 1 (Non-asymptotic excess risk bound). *Under model (1.1), suppose that Assumptions 1 and 2 hold, ν is absolutely continuous with respect to the Lebesgue measure, and $\|f_0\|_\infty \leq \mathcal{B}$ for some $\mathcal{B} \geq 1$. Suppose the network parameters of the function class $\mathcal{F}_\phi$ are specified as in (4.4) and (4.5). Then, for $2n \geq Pdim(\mathcal{F}_\phi)$, the excess risk of the DQR estimator $\hat{f}_\phi$ satisfies*

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C \frac{\lambda_\tau \mathcal{B} \mathcal{S} \mathcal{D} \log(\mathcal{S}) \log(n)}{n^{1-1/p}} + 2\lambda_\tau \sum_{i \in J^c} C_i^* \lambda_i^* t_i^* (N_i L_i)^{-2\alpha_i^*/t_i},$$

($L_1(\nu)$)

where $\lambda_\tau = \max\{\tau, 1-\tau\}$ *and* $C > 0$ *is a constant which does not depend on* $n, d, \tau, \mathcal{B}$, $\mathcal{S}$, $\mathcal{D}$, C_i^*, λ_i^*, α_i^*, N_i *or* L_i, *and* $C_i^* = 18^{\Pi_{j=i+1}^q \alpha_j}$, $\lambda_i^* = \Pi_{j=i}^q \lambda_j^{\Pi_{k=j+1}^q \alpha_k}$, $\alpha_i^* = \Pi_{j=i}^q \alpha_j$ *and* $t_i^* = (\Pi_{j=i}^q \sqrt{t_j}^{\Pi_{k=j}^q \alpha_k})/\sqrt{t_i}^{\alpha_i}$.

Additionally if Assumption 3 also holds, we have

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C\frac{\lambda_\tau \mathcal{BSD}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + \underline{2c_\tau\Big[\sum_{i\in J^c} C_i^*\lambda_i^* t_i^*(N_iL_i)^{-2\alpha_i^*/t_i}\Big]^2},$$

L(V)

where $c_\tau > 0$ *is a constant defined in Lemma 4 and* $C > 0$ *is a constant not depending on* $n, d, \tau, \mathcal{B}$, $\mathcal{S}$, $\mathcal{D}$, C_i^*, λ_i^*, α_i^*, N_i *or* L_i.

1

Remark 4. *In Theorem 1, the bounds for the excess risk are explicitly expressed in terms of the network parameters* $\mathcal{D}$ *and* $\mathcal{S}$ *and the parameters* N_i *and* L_i. *, which determine the width and the depth of the network as specified in (4.4) and (4.5). The dependence of the bounds on the dimensions of the functions* (d, t_j) *and the Hölder constants* (α_j, λ_j) *for the functions is also explicitly described. These constants are given and determined by the underlying model, so we cannot change them. The constants* C *and* c_τ *are independent of all the above parameters, in particular, they do not depend on the dimensions* (d, t_j).

constant Indie with dim

2

bias-variance trade off

Theorem 1 gives a general expression of the upper bound for the excess risk. This bound clearly describes how the bounds depend on various parameters. The parameters that can be changed or tuned are the network parameters given in terms of N_i and L_i. We note that the stochastic error term increases with (N_i, L_i), while the approximation error term decreases with (N_i, L_i). Thus we can select (N_i, L_i) to balance these two error terms, which lead to the best error bound. We will present an explicit expression of the risk bound in Corollary 2 below. First, we state a simpler bound assuming that all the component functions in the composition are Lipschitz continuous with $\alpha_i = 1, i = 0, 1, \ldots, q$.

Corollary 1. *Under model (1.1), suppose Assumptions 1 and 2 hold and all* $h_{ij} : D_{ij} \to \mathbb{R}$ *in Theorem 3 are Lipschitz continuous functions* ($\alpha_i = 1$ *for* $i = 0, \ldots, q$) *with Lipschitz constants* $\lambda_i \geq 0$. *Given any* $N, L \in \mathbb{N}^+$, *for* $i \in J^c$, *we set the same shape for each subnetwork with* $N_i = N \in \mathbb{N}^+$ *and* $L_i = L \in \mathbb{N}^+$, *and for* $j \in J$, *we set the 3-layer subnetwork with width* $(d_j, 2d_j, d_{j+1})$ *according to Lemma 9. Suppose the network parameters of the function class* $\mathcal{F}_\phi$ *are specified as in (4.4) and (4.5). Then, for* $2n \geq Pdim(\mathcal{F}_\phi)$, *the excess risk of the DQR estimator* $\hat{f}_\phi$ *satisfies*

$\alpha_i = 1$, from theorem 1

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C\frac{\lambda_\tau \mathcal{BSD}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + 36\lambda_\tau \sum_{i\in J^c} \Pi_{k=i+1}\underset{\lambda_k}{\sqrt{t_k}}(N_iL_i)^{-2/t_i},$$

where $\lambda_\tau = \max\{\tau, 1-\tau\}$ *and* $C > 0$ *is a constant independent of* $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{D}, N$ *or* L. *Additionally if Assumption 3 also holds, we have*

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C\frac{\lambda_\tau \mathcal{BSD}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + 648c_\tau\Big[\sum_{i\in J^c} \Pi_{k=i+1}\underset{\lambda_k}{\sqrt{t_k}}(N_iL_i)^{-2/t_i}\Big]^2,$$

where $c_\tau > 0$ *is a constant defined in Assumption 3 and* $C > 0$ *is a constant independent of* $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{D}, N$ *or* L.

Remark 5. *The* $\log(n)$ *factor in the stochastic error of the upper bound in Theorem 1 and Corollary 1 is due to the truncation technique used in the proof. Power of log factors,* $(logn)^k$ *for some* $k \in \mathbb{N}^+$*, are commonly seen in the results of related work, e.g., Bauer and Kohler (2019); Schmidt-Hieber et al. (2020) and Farrell et al. (2021). By properly setting the network size* $\mathcal{S}$ *or depth* $\mathcal{D}$ *to have order* $O(n^c/(\log n)^k)$ *for some constant* $c > 0$ *and* $k \in \mathbb{N}^+$*, the final convergence rate of the excess risk could be made optimal. However, this will make the selection of the network parameters more complicated. Therefore, we will not do so in this paper. The rate of convergence is (nearly) optimal up to a logarithmic factor* $(\log n)^2$.

We now present an explicit risk bound for the *DQR* estimators with three sets of network parameters with different depth and width. All these three different specifications of the network parameters lead to the same risk bound.

Corollary 2. *Under model (1.1), suppose that Assumptions 1-3 hold,* ν *is absolutely continuous with respect to the Lebesgue measure,* $\|f_0\|_\infty \leq \mathcal{B}$ *for some* $\mathcal{B} \geq 1$ *and* $2n \geq Pdim(\mathcal{F}_\phi)$*. Let* $(\alpha^*, t^*) = \arg\min_{(\alpha_i^*, t_i), i \in J^c}\{\alpha_i^*/t_i\}$, $\lambda^* = \max_{i=0,\ldots,q}\lambda_i^*$ *and* $d^* = \max_{i=0,\ldots,q} t_i^*$*, where* α_i^*, λ_i^* *and* t_i^* *are defined in Theorem 1. Suppose the network parameters of the function class* $\mathcal{F}_\phi$ *are specified as follows:*

put into 4.4 and 4.5

1. *(Deep and fixed width MLP) Let* $N_i = 1$ *and* $L_i = \lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor$*. The corresponding width, depth and size of the networks satisfy:*

$$\mathcal{W}_1 = \max_{i=0,\ldots,q} d_i \max\{7t_i, 20\},$$
$$\mathcal{D}_1 = (12\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor + 15)|J^c| + 2|J|,$$
$$\mathcal{S}_1 \leq \mathcal{W}_1^2\mathcal{D}_1 \leq \max_{i=0,\ldots,q}(20d_it_i)^2 \times 29q\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor.$$

2. *(Deep and wide MLP) Let* $N_i = \lfloor n^{(1-1/p)t^*/(8\alpha^*+4t^*)}\rfloor$ *and* $L_i = \lfloor n^{(1-1/p)t^*/(8\alpha^*+4t^*)}\rfloor$*. The corresponding width, depth and size of the networks satisfy:*

$$\mathcal{W}_2 = \max_{i=0,\ldots,q} d_i \max\{4t_i\lfloor\lfloor n^{(1-1/p)t^*/(8\alpha^*+4t^*)}\rfloor^{1/t_i}\rfloor + 3t_i, 12\lfloor n^{(1-1/p)t^*/(8\alpha^*+4t^*)}\rfloor + 8\},$$
$$\mathcal{D}_2 = (12\lfloor n^{(1-1/p)t^*/(8\alpha^*+4t^*)}\rfloor + 15)|J^c| + 2|J|,$$
$$\mathcal{S}_2 \leq \mathcal{W}_2^2\mathcal{D}_2 \leq \max_{i=0,\ldots,q}(20d_it_i)^2 \times 29q\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor^{3/2}.$$

3. *(Fixed depth and wide MLP) Let* $N_i = \lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor$ *and* $L_i = 1$*. The corresponding width, depth and size of the networks satisfy:*

$$\mathcal{W}_3 = \max_{i=0,\ldots,q} d_i \max\{4t_i\lfloor\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor^{1/t_i}\rfloor + 3t_i, 12\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor + 8\},$$
$$\mathcal{D}_3 = 27|J^c| + 2|J|,$$
$$\mathcal{S}_3 \leq \mathcal{W}_3^2\mathcal{D}_3 \leq \max_{i=0,\ldots,q}(20d_it_i)^2 \times 29q\lfloor n^{(1-1/p)t^*/(4\alpha^*+2t^*)}\rfloor^2.$$

Then, the excess risk satisfies

$$\mathbb{E}\{\mathcal{R}^\tau(\hat{f}_\phi) - \mathcal{R}^\tau(f_0)\} \leq C_0 C_{d,d^*} (\log n)^2 n^{-\left(1-\frac{1}{p}\right)\frac{2\alpha^*}{2\alpha^*+t^*}}, \tag{4.6}$$

where $C_{d,d^*} = (d^*)^2(\max_{i=0,\ldots,q} d_i t_i)^2 \log(\max_{i=0,\ldots,q} d_i t_i)$, $C_0 = c\lambda_\tau c_\tau \mathcal{B} q^2 \log(q)(\lambda^*)^2$. *Here* c *is a universal constant not depending on any parameters.*

In Corollary 2, three sets of different network parameters lead to the same risk bound. Therefore, generally the choice of network parameters is not unique to achieve a desired risk bound. Although the three sets of network parameters given in Corollary 2 yield the same risk bound, the sizes of the networks are different. As can be seen from the expressions of the network sizes $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$, we have, on the logarithmic scale,

$$\log \mathcal{S}_1 : \log \mathcal{S}_2 : \log \mathcal{S}_3 = 1 : \frac{3}{2} : 2.$$

different size
same bound

Therefore, the deep and fixed width network in the first network specification with width $\mathcal{W}_1$ and depth $\mathcal{D}_1$ is the most efficient design among the three network structures in the sense that it has the smallest network size. Corollary 2 shows that deep networks have advantages over shallow ones in the sense that deep networks achieve the same risk bound with a smaller network size. More detailed discussions on the relationship between convergence rate and network structure can be found in Jiao et al. (2021).

4.2 Mean integrated squared error

The empirical risk minimization quantile estimator typically results in an estimator $\hat{f}_n$ for which its risk $\mathcal{R}^\tau(\hat{f}_n)$ is close to optimal risk $\mathcal{R}^\tau(f_0)$ in expectation or with high probability. However, small excess risk in general only implies in a weak sense that the ERM $\hat{f}_n$ is close to f_0 (Remark 3.18, Steinwart (2007)). Hence, in this subsection, we bridge the gap between the excess risk and the mean integrated squared error (MISE) of the estimated conditional quantile function. To this end, we need the following condition on the conditional distribution of Y given X.

Assumption 4. *There exist constants* $\gamma > 0$ *and* $\kappa > 0$ *such that for any* $|\delta| \leq \gamma$,

$$\left|P_{Y|X}(f_0(x) + \delta \mid x) - P_{Y|X}(f_0(x) \mid x)\right| \geq \kappa|\delta|,$$

for all $x \in \mathcal{X}$ *up to a* ν*-negligible set, where* $P_{Y|X}(\cdot|x)$ *denotes the conditional distribution function of* Y *given* $X = x$.

Remark 6. *A similar condition is assumed by Padilla and Chatterjee (2021) in studying nonparametric quantile trend filtering. This condition is weaker than Condition 2.1 in He and Shi (1994) and condition D.1 in Belloni et al. (2011), which require the conditional density of* Y *given* $X = x$ *to be bounded below near its* τ*-th quantile.*

Under Assumption 4, the self-calibration condition can be established as stated below. This will lead to a bound on the MISE of the estimated quantile function based on a bound for the excess risk.

Lemma 5 (Self-calibration)**.** *Suppose that Assumption 2 (i) and Assumption 4 hold. For any $f : \mathcal{X} \to \mathbb{R}$, denote $\Delta^2(f, f_0) = \mathbb{E}\big[\min\{|f(X) - f_0(X)|^2, |f(X) - f_0(X)|\}\big]$ where κ and $\gamma > 0$ are defined in Assumption 4. Then we have*

$$\Delta^2(f, f_0) \leq c_{\kappa,\gamma}\big\{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\big\},$$

for any $f : \mathcal{X} \to \mathbb{R}$ where $c_{\kappa,\gamma} = \max\{2/\kappa, 4/(\kappa\gamma)\}$. More exactly, for $f : \mathcal{X} \to \mathbb{R}$ satisfying $|f(x) - f_0(x)| \leq \gamma$ for $x \in \mathcal{X}$ up to a ν-negligible set, we have

$$\|f - f_0\|^2_{L^2(\nu)} \leq \frac{2}{\kappa}\big\{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\big\},$$

otherwise we have

$$\|f - f_0\|_{L^1(\nu)} \leq \frac{4}{\kappa\gamma}\big\{\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)\big\}.$$

Remark 7. *Similar self-calibration conditions can be found in Christmann and Steinwart (2007); Steinwart et al. (2011); Lv et al. (2018) and Padilla et al. (2020). A general result is obtained in Steinwart et al. (2011) under the so-called τ-quantile of t-average type assumption on the joint distribution P, where $\|f - f_0\|_{L^r(\nu)}$ is upper bounded by the q-th root of excess risk $\mathcal{R}^\tau(f) - \mathcal{R}^\tau(f_0)$ for $t \in (0, \infty]$, $q \in [1, \infty)$ and $r = tq/(t + 1)$. However, those assumptions on the joint distribution P generally require that the conditional distribution of Y given X is bounded, which may not be applicable to models with heavy-tailed response as in our setting, see, e.g., Assumption 2.*

Theorem 2 (Non-asymptotic bound for mean integrated squared error)**.** *Under model (1.1), suppose that Assumptions 1, 2 and 4 hold, ν is absolutely continuous with respect to the Lebesgue measure, and $\|f_0\|_\infty \leq \mathcal{B}$ for some $\mathcal{B} \geq 1$. Then, given any $N_i, L_i \in \mathbb{N}^+, i \in J^c$, for the function class of ReLU multi-layer perceptrons $\mathcal{F}_\phi = \mathcal{F}_{\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S},\mathcal{B}}$ with width no larger than $\mathcal{W} = \max_{i=0,\ldots,q} d_i \max\{4t_i\lfloor N_i^{1/t_i}\rfloor + 3t_i, 12N_i + 8\}$ and depth $\mathcal{D} = \sum_{i\in J^c}(12L_i + 15) + 2|J|$, for $2n \geq Pdim(\mathcal{F}_\phi)$, the MISE of the DQR estimator $\hat{f}_\phi$ satisfies*

$$\mathbb{E}\big\{\Delta^2(\hat{f}_\phi, f_0)\big\} \leq c_{\kappa,\gamma}\lambda_\tau\Big[C\frac{\mathcal{B}\mathcal{S}\mathcal{D}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + 2\sum_{i\in J^c} C_i^*\lambda_i^* t_i^*(N_iL_i)^{-2\alpha_i^*/t_i}\Big],$$

where $c_{\kappa,\gamma} = \max\{4/(\kappa\gamma), 2/\kappa\}$ and $\Delta^2(\cdot,\cdot)$ are defined in Lemma 5, $\lambda_\tau = \max\{\tau, 1 - \tau\}$ and $C > 0$ is a constant not depending on $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{D}, C_i^*, \lambda_i^*, \alpha_i^*, N_i$ or L_i, and $C_i^* = 18^{\Pi_{j=i+1}^q \alpha_j}$, $\lambda_i^* = \Pi_{j=i}^q \lambda_j^{\Pi_{k=j+1}^q \alpha_k}$, $\alpha_i^* = \Pi_{j=i}^q \alpha_j$ and $t_i^* = (\Pi_{j=i}^q \sqrt{t_j}^{\Pi_{k=j}^q \alpha_k})/\sqrt{t_i}^{\alpha_i}$. Additionally if Assumption 3 also holds, we have

$$\mathbb{E}\|\hat{f}_\phi - f_0\|_{L^*(\nu)} \leq c_{\kappa,\gamma}\Big[C\frac{\lambda_\tau\mathcal{B}\mathcal{S}\mathcal{D}\log(\mathcal{S})\log(n)}{n^{1-1/p}} + 2c_\tau\Big\{\sum_{i\in J^c} C_i^*\lambda_i^* t_i^*(N_iL_i)^{-2\alpha_i^*/t_i}\Big\}^2\Big],$$

where $c_\tau > 0$ is a constant defined in Assumption 3 and $C > 0$ is a constant independent of $n, d, \tau, \mathcal{B}, \mathcal{S}, \mathcal{D}, C_i^*, \lambda_i^*, \alpha_i^*, N_i$ or L_i.

Similar to Corollary 2, we have the following corollary for the MISE of the *DQR* estimator.