Journal of Machine Learning Research 25 (2024) 1-50 Submitted 6/24; Revised 12/24; Published 12/24

Deep Nonparametric Quantile Regression under Covariate Shift

Xingdong Feng FENG.XINGDONG@MAIL.SHUFE.EDU.CN
School of Statistics and Data Science &
Institute of Data Science and Statistics
Shanghai University of Finance and Economics
Shanghai, China
Xin He HE.XIN17@MAIL.SHUFE.EDU.CN
School of Statistics and Data Science
Shanghai University of Finance and Economics
Shanghai, China
Yuling Jiao YULINGJIAOMATH@WHU.EDU.CN
School of Artificial Intelligence
Hubei Key Laboratory of Computational Science
Wuhan University
Wuhan, China
Lican Kang KANGLICAN@WHU.EDU.CN
Institute for Math and AI
School of Mathematics and Statistics
Wuhan University
Wuhan, China
Caixing Wang WANG.CAIXING@STU.SUFE.EDU.CN
School of Statistics and Data Science
Shanghai University of Finance and Economics
Shanghai, China

Editor: Xiaotong Shen

Abstract

This work focuses on addressing the challenges posed by covariate shift in nonparametric quantile regression using deep neural networks. We propose a two-stage pre-training reweighted method that leverages importance weighting to mitigate the effects of distribution shift. In the first stage, density ratios are estimated with a neural network by minimizing least squares. In the second stage, a deep neural network estimator is obtained using pre-training weights. Theoretical analysis is provided, offering non-asymptotic error bounds for the unweighted, reweighted, and pre-training reweighted estimators. We consider scenarios with both bounded and unbounded density ratios. Notably, we employ a novel proof technique to bound the generalization error, characterized by the size and weights bound of ReLU neural networks. This enables us to establish fast rates of convergence under the adaptive self-calibration condition, distinguishing our approach from

∗. Xin He and Lican Kang are the corresponding authors. All authors contributed equally to this paper, and their names are listed in alphabetical order.

those relying on local Rademacher complexity techniques. Additionally, we derive the approximation error with weight bounds for ReLU neural networks approximating the Hölder class. Our theoretical findings provide valuable insights for the pre-training process and highlight the efficacy of reweighted techniques. Numerical experiments are conducted to further validate the theoretical findings and demonstrate the effectiveness of our proposed method.

Keywords: Deep neural networks, distribution shift, non-asymptotic error bounds, robust estimation

1. Introduction

Quantile regression (Koenker and Bassett Jr, 1978) occupies a pivotal and indispensable position within the domain of statistical modeling. It distinguishes itself substantially from traditional mean regression by offering a comprehensive portrayal of the response and covariates. The inherent capacity of quantile regression to estimate conditional quantiles of the response underscores its unique contribution to statistical modeling. This refined approach facilitates the exploration of how various quantiles of the response respond to variations in the covariate space, thereby providing a more profound understanding of the underlying data structure. Precisely, quantile regression not only complements but also elevates conventional mean regression models, particularly when addressing scenarios characterized by heteroscedasticity, outliers, or other manifestations of non-normal data features (Bondell et al., 2010; Wang et al., 2012).

Nonparametric quantile regression, a specific subset of quantile regression, proves invaluable when dealing with complex relationships among covariates and response that defy straightforward parametric characterization. Its primary objective revolves around the attainment of nonparametric estimates for the underlying regression function within the context of quantile regression. Notably, it distinguishes itself through its intrinsic flexibility, as it is capable of accommodating diverse data distributions without relying on predetermined functional forms. The research on nonparametric quantile regression has witnessed substantial development in the past decades (White, 1992; Koenker et al., 1994; He and Shi, 1994; He and Ng, 1999; Takeuchi et al., 2006; Sangnier et al., 2016). This extensive body of literature emphasizes the diverse methodological approaches grounded in reproducing kernels, smoothing splines, and shallow neural networks. Recent attention in statistical modeling has been drawn towards the application of deep neural networks, particularly within the framework of nonparametric estimation. Deep nonparametric quantile regression, a specialized domain within this domain, has also garnered significant interest (Padilla et al., 2022a; Madrid Padilla and Chatterjee, 2022; Shen et al., 2021).

Yet, the existing literature often lacks comprehensive exploration and effective solutions for the intricate challenge posed by distribution shift within the realm of quantile regression. Distribution shift refers to situations where there exists a substantial divergence between the distributions characterizing the training and testing data sets. This phenomenon is pervasive and manifests ubiquitously across practical modeling scenarios. Notably, instances of distribution shift materialize across a spectrum of real-world domains, including, but not limited to, image analysis (Taori et al., 2020; Guan and Liu, 2021), natural language processing (Jiang and Zhai, 2007; Hassan et al., 2013), and recommender systems (Carroll et al., 2022; Tan et al., 2016). As illustrated in Daume III and Marcu (2006); Torralba

and Efros (2011), a model that exhibits robust performance when trained on specific data may falter when exposed to new data characterized by distribution shift. This discrepancy arises from the fundamental principle that models glean insights and knowledge from the distribution of training data, and their effectiveness is contingent on the persistence of these patterns in the testing data. Therefore, distribution shift stands as a formidable challenge in statistics and machine learning.

Distribution shift encompasses various types, such as covariate shift, concept shift, marginal shift, conditional shift, label shift, domain shift, and mode shift. Among them, covariate shift (Quinonero-Candela et al., 2008; Sugiyama and Kawanabe, 2012) is a specific and significant manifestation within distribution shift, and it is prevalent in practical scenarios. Covariate shift arises when the distribution of the input covariates within the training data set deviates from that within the testing data set, while simultaneously maintaining an unaltered conditional distribution of the response given the input covariates. In essence, covariate shift entails alterations in the distribution of input covariates between the training and testing data sets, while the underlying relationship between these covariates and the response remains invariant. Addressing the intricacies posed by covariate shift necessitates the deployment of density-ratio reweighting techniques, often referred to as importance weighting. This approach has garnered significant attention and is well-studied in Shimodaira (2000); Huang et al. (2006); Sugiyama et al. (2007b); Bickel et al. (2009); Kanamori et al. (2009); Fang et al. (2020). Furthermore, some works provided related theoretical analysis, see Cortes et al. (2008, 2010); Xu et al. (2022), while the suboptimal convergence rate is given in Cortes et al. (2008, 2010). Recently, the covariate shift problem is studied in Ma et al. (2023) and Feng et al. (2023) within nonparametric regression based on a reproducing kernel Hilbert space framework, which also provided some theoretical insights into addressing the challenges induced by covariate shift. Nevertheless, these works rely on the unrealistic and overly stringent assumption that the density ratio is known, which is impractical since only observed data from the source and target distributions are available in real-world applications. Moreover, we want to emphasize that all the aforementioned methods only focus on the covariate shift problem in the context of mean regression. Despite its practical importance, the covariate shift problem in nonparametric quantile regression remains largely underexplored, particularly when the density ratio is unknown, which is quite challenging from both methodological and theoretical points of view.

In literature, estimating the true density ratio is a key challenge in covariate shift adaptation. A straightforward approach is to estimate the source and target densities separately using kernel density estimation (Sugiyama and Müller, 2005; Baktashmotlagh et al., 2014), followed by calculating their ratio. Yet, it is worthy pointing out that such a procedure is inefficient and time-consuming, especially under the high-dimensional case. An alternative approach in literature is to directly estimate the density ratio instead of estimating the densities individually. These works mainly focus on minimizing some discrepancy measures between distributions, including the kernel mean matching (Huang et al., 2006; Gretton et al., 2009), Kullback-Leibler divergence (Sugiyama et al., 2007a,b, 2012) and non-negative Bregman divergence (Kato and Teshima, 2021). To further improve computational efficiency, Kanamori et al. (2009); Sugiyama et al. (2012) formulate the direct importance estimation problem as a least-squares function fitting problem, which can be efficiently solved using a standard quadratic program. It performs comparably to the Kullback-Leibler importance

estimation method (Sugiyama et al., 2007b). We want to point out that although there exist some methods on nonparametric density estimation, the investigations on integrating the estimated density ratio into the problem of covariate shift is still lacking in literature, even for the mean regression. To the best of our knowledge, it is still unknown how to explore the effect of estimation accuracy of density ratio on the convergence rate of the final estimated quantile function when covariate shift occurs.

In this paper, we investigate the deep nonparametric quantile regression under covariate shift. We design a novel two-stage pre-training procedure, incorporating a reweighting mechanism. At the first stage, we obtain a deep estimator for the density ratio which is constructed based on the least-squares method (Kanamori et al., 2009; Sugiyama et al., 2012). Subsequently, in the second stage, we leverage the density ratio estimates acquired in the first phase to derive a pre-training reweighted estimator for the underlying regression function using deep neural networks. To more comprehensively underscore the merits of our approach, we further introduce two supplementary estimators for comparative analysis, which are referred to as the unweighted and reweighted estimators introduced in Section 3.2. The performance evaluation of the pre-training reweighted estimator with these two alternative estimators elucidates the superiority of our method, thereby illustrating the pivotal role assumed by pre-training and reweighting techniques in our framework.

In summary, the contributions of this work are outlined as follows.

(i) We introduce a novel two-stage pre-training reweighted method to address the issue of covariate shift in deep nonparametric quantile regression. We thus propose a pre-training reweighted estimator using deep neural networks for the underlying regression function. To our best knowledge, this is the first work to consider the estimated ratios for covariate shift problems.

(ii) We provide a non-asymptotic error analysis for unweighted, reweighted, and pre-training reweighted quantile regression estimators. This analysis is primarily achieved by decomposing the error into statistical and approximation errors, followed by establishing error bounds via trade-off considerations. Under the adaptive self-calibration condition, the resulting error bounds of order $\mathcal{O}\left(n^{-\frac{2\zeta}{d+2\zeta}}\right)$, where n is the sample size, d is the data dimension, and ζ is a smoothness parameter, attain minimax optimal rates in nonparametric regression. Furthermore, our theoretical development simultaneously explores scenarios involving both bounded and unbounded density ratios under some weaker conditions than those commonly considered in the existing literature. Our theoretical findings provide prior guidance for pre-training and underscore the significance of reweighted techniques.

(iii) Our technical novelty lies in an alternative proof method for bounding the statistical error, enabling us to obtain a sharp rate. This approach simplifies and differs from the conventional local Rademacher complexity techniques (Bartlett et al., 2005). Additionally, we also derive the approximation error with weight bounds for ReLU neural networks approximating the Hölder class.

1.1 Outlines

The rest of the paper is organized as follows. In Section 2, we introduce the standard nonparametric quantile regression and the definition of covariate shift. In Section 3, we formally formulate the deep nonparametric quantile estimation problem under covariate shift and propose a two-stage pre-training reweighted method involving the reweighted estimator with the pre-training density ratio. We also consider the naive unweighted estimator and the reweighted estimator with the true density ratio. In Section 4, we provide the details of error analysis for both unweighted, reweighted, and pre-training reweighted estimators, and establish the theoretical guarantees on these estimators under some conditions on the density ratio. Numerical experiments on synthetic examples are provided in Section 5. Concluding remarks are then given in Section 6. All the technical proofs of lemmas and theorems and additional numerical experiments are deferred to the Appendix.

1.2 Notations

For any $a, b \in \mathbb{R}$, we use $\lfloor a \rfloor$ to denote the largest integer less than a, and write $a \vee b = \max\{a, b\}$. We define $a \lesssim b$ and $a = \mathcal{O}(b)$ if there exists some positive constant C such that $a \leq Cb$, $a \gtrsim b$ and $a = \Omega(b)$ if there exists some positive constant C such that $a \geq Cb$. We use $\mathbb{N}_0$, $\mathbb{N}$ and $\mathbb{R}$ to denote the set of non-negative, strictly positive integers and real numbers, respectively. For a multi-index $\boldsymbol{s} = (s_1, \ldots, s_d) \in \mathbb{N}_0^d$, the symbol $\partial^{\boldsymbol{s}}$ denotes the partial differential operator $\partial^{\boldsymbol{s}} = (\frac{\partial}{\partial x_1})^{s_1} \ldots (\frac{\partial}{\partial x_d})^{s_d}$ and we use the convention that $\partial^{\boldsymbol{s}}$ is the identity operator when $\boldsymbol{s} = \boldsymbol{0}$. Let ν be a probability distribution over $\mathbb{R}^d$ and f be a measurable function from $\mathbb{R}^d$ to $\mathbb{R}$. We use $L_p(\nu) = \{f : \int |f(x)|^p \nu(dx) < \infty\}$ for $p \geq 1$ to denote the space of L_p-integrable functions with respect to ν, equipped with the norm $\|f\|_{L_p(\nu)} = \{\int |f(x)|^p \nu(dx)\}^{1/p}$.

2. Preliminaries

In this section, we introduce the standard nonparametric quantile regression model and the definition of covariate shift.

2.1 The Standard Nonparametric Quantile Regression

Given a univariate response $Y \in \mathcal{Y} \subseteq \mathbb{R}$ and a d-dimensional covariate vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathcal{X} \subseteq \mathbb{R}^d, d \in \mathbb{N}$, the τ-th quantile $Q_\tau(Y|\boldsymbol{X})$ of Y given $\boldsymbol{X}$ is

$$Q_\tau(Y|\boldsymbol{X}) = f_0(\boldsymbol{X}, \tau). \tag{1}$$

According to (1), we have $\mathbb{P}(Y - f_0(\boldsymbol{X}, \tau) \leq 0) = \tau$ for any given $\tau \in (0, 1)$. If we define the error term as $\varepsilon = Y - f_0(\boldsymbol{X}, \tau)$, then model (1) becomes

$$Y = f_0(\boldsymbol{X}, \tau) + \varepsilon, \tag{2}$$

where $\mathbb{P}(\varepsilon \leq 0 | \boldsymbol{X} = \boldsymbol{x}) = \tau$ for any $\boldsymbol{x} \in \mathcal{X}$. For notation simplicity, we suppress τ and denote $f_0(\cdot) := f_0(\cdot, \tau)$ since we only focus on one specific quantile level τ. Suppose that a random training sample $\mathbb{D} := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ is composed of independent copies of the

random pair $(\boldsymbol{X}, Y)$ drawn from some unknown joint distribution $P_{\boldsymbol{X},Y}$. The standard quantile regression (Koenker and Bassett Jr, 1978) estimates f_0 by

$$\widehat{f} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - f(\boldsymbol{X}_i)), \tag{3}$$

where $\rho_\tau(u) = u(\tau - \mathbf{I}\{u < 0\})$ denotes the τ-th quantile loss and $\mathcal{F}$ denotes a specific function class containing a set of measurable functions. It is known that the quantile loss $\rho_\tau(\cdot)$ is Lipschitz continuous with the Lipschitz constant $\lambda := \max\{\tau, 1-\tau\}$. This implies that $\big|\rho_\tau(u_1) - \rho_\tau(u_2)\big| \le \lambda|u_1 - u_2|$ for any $u_1, u_2 \in \mathbb{R}$.

Note that the optimization task (3) is the empirical version of the following learning problem, which aims to find a $\widetilde{f} : \mathbb{R}^d \to \mathbb{R}$ overall measurable functions satisfying

$$\widetilde{f} = \arg\min_{f} \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f_0(\boldsymbol{X}))\big], \tag{4}$$

where $\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}$ denotes the expectation with repect to $P_{\boldsymbol{X},Y}$. It is worthy pointing out that the sample version of the term $\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\rho_\tau(Y - f_0(\boldsymbol{X}))]$ can be regarded as a constant and thus is omitted in (3) and somewhere else in this paper. Moreover, under the assumption that the τ-th conditional quantile of ε given $\boldsymbol{X}$ is zero, the global minimizer $\widetilde{f}$ in (4) coincides with the underlying regression function f_0 in (2).

2.2 Phenomenon of Covariate Shift

In literature, it is commonly assumed that the source (training) and target (testing) data originate from the same distribution $P_{\boldsymbol{X},Y}$ defined over the joint space $\mathcal{X}\times\mathcal{Y}$. Moreover, the joint distribution can be decomposed as $P_{\boldsymbol{X},Y} = P_{Y|\boldsymbol{X}}P_{\boldsymbol{X}}$, where $P_{Y|\boldsymbol{X}}$ denotes the conditional distribution determined by the model given in (2) and $P_{\boldsymbol{X}}$ denotes the marginal distribution of $\boldsymbol{X}$. Yet, in many real applications, the source and target data may come from some different joint distributions. Specifically, we assume that the target data are drawn from some other joint distribution $Q_{\boldsymbol{X},Y}$, which is also defined over the joint space $\mathcal{X}\times\mathcal{Y}$ and can be decomposed as $Q_{\boldsymbol{X},Y} = Q_{Y|\boldsymbol{X}}Q_{\boldsymbol{X}}$. To be more precise, in this paper, we consider the special case that for the source and target data, the conditional distributions are the same that $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$ representing the invariance of the underlying regression model (2), while the marginal distributions of the source data $P_{\boldsymbol{X}}$ (source distribution) and the target data $Q_{\boldsymbol{X}}$ (target distribution) are significantly different. Note that assuming $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$ implicitly requires that the conditional distribution of ε given $\boldsymbol{X}$ remains invariant, regardless of whether $\boldsymbol{X}$ is generated from $P_{\boldsymbol{X}}$ or $Q_{\boldsymbol{X}}$. This scenario is often referred to as covariate shift.

Under the standard quantile regression without covariate shift, the prediction performance of the estimator $\widehat{f}$ in (3) can be evaluated under the $L_2(P_{\boldsymbol{X}})$ norm (Padilla et al., 2022b) that

$$\|\widehat{f} - f_0\|^2_{L_2(P_{\boldsymbol{X}})} = \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[(\widehat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\big],$$

where $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}$ is the expectation over $P_{\boldsymbol{X}}$, and it is well-known that $\widehat{f}$ is indeed a consistent estimator under the $L_2(P_{\boldsymbol{X}})$ norm. However, once the phenomenon of covariate shift occurs,

this metric becomes problematic in the sense that we primarily aim to construct an estimator whose prediction error for the target source is small under the $L_2(Q_{\boldsymbol{X}})$ norm that

$$\|\widehat{f} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})} = \mathbb{E}_{\boldsymbol{X}\sim Q_{\boldsymbol{X}}}\big[(\widehat{f}(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\big],$$

where $\mathbb{E}_{\boldsymbol{X}\sim Q_{\boldsymbol{X}}}$ is the expectation over $Q_{\boldsymbol{X}}$. It is thus clear that the evaluation of the discrepancy between the source and target distributions plays a crucial role in tackling the problem of covariate shift.

3. Method

In this section, we provide a deep nonparametric estimation procedure using feedforward neural networks (FNNs) with Rectified Linear Unit (ReLU) activation. We first give a brief introduction of ReLU FNNs and Hölder class in Section 3.1, and then propose the unweighted, reweighted, and the novel pre-training reweighted estimators in Sections 3.2 and 3.3, respectively.

3.1 ReLU FNNs and Hölder class

Recently, deep neural networks have attracted tremendous attention and received a variety of successful achievements in many applications (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Lu et al., 2021; Jiao et al., 2023). Neural network functions have demonstrated their effectiveness in approximating high-dimensional functions. In this paper, we aim to estimate the regression function within the function class of ReLU FNNs. Precisely, we use $\mathcal{F}_{d_1,d_2}(\mathcal{W},\mathcal{D},\mathcal{S},\mathcal{B})$ to denote the set of functions $\{\phi\}'s$ that can be parameterized by ReLU FNNs with width $\mathcal{W}$, depth $\mathcal{D}$, size $\mathcal{S}$, weights bound $\mathcal{B}$, and d_1, d_2 denote the input and output dimensions of the ReLU FNNs. Specifically, the ReLU FNN $\phi : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ can be expressed as

$$\phi(x) = A_{\mathcal{D}}\phi_{\mathcal{D}}(x) + b_{\mathcal{D}},$$

where $\phi_0(x) = x$, $\phi_{\ell+1}(x) = \sigma\left(A_\ell\phi_\ell(x) + b_\ell\right), \ell = 0,\ldots,\mathcal{D}-1$, and $A_\ell \in \mathbb{R}^{N_{\ell+1}\times N_\ell}$ denote the weight matrix, $N_\ell \in \mathbb{N}$ denotes the width of the ℓ-th hidden layer with $N_0 = d_1$ and $N_{\mathcal{D}+1} = d_2$, $b_\ell \in \mathbb{R}^{N_{\ell+1}}$ denotes the bias vector, and $\sigma(x) = \max(0,x)$ is the ReLU activation function defined for each element of x. Therefore, the parameters of the ReLU FNN $\phi(\cdot)$ can be denoted as

$$\theta = \big((A_0,b_0),\ldots,(A_{\mathcal{D}-1},b_{\mathcal{D}-1}),(A_{\mathcal{D}},b_{\mathcal{D}})\big).$$

Furthermore, we denote the number of non-zero elements and the largest absolute value of the parameters in θ as

$$\|\theta\|_0 = \sum_{\ell=0}^{\mathcal{D}} \|\mathrm{vec}\,(A_\ell)\|_0 + \|b_\ell\|_0\,,$$

and

$$\|\theta\|_\infty = \max\Big\{\max_{\ell\in\{0,\ldots,\mathcal{D}\}}\|\,\mathrm{vec}(A_\ell)\|_\infty, \max_{\ell\in\{0,\ldots,\mathcal{D}\}}\|b_\ell\|_\infty\Big\},$$

respectively. Note that $\mathrm{vec}(A)$ transforms the matrix A into the corresponding vector by concatenating the column vectors. Then, the width $\mathcal{W}$, the size $\mathcal{S}$, and the bound $\mathcal{B}$ are

defined as $\mathcal{W} = \max\{N_1, \ldots, N_{\mathcal{D}}\}$, $\mathcal{S} = \|\theta\|_0$, and $\|\theta\|_\infty \leq \mathcal{B}$, respectively. It is thus clear that the ReLU FNN may not be fully connected, and thus $\mathcal{S}$ can be much smaller than that of the fully connected case. In the rest of this paper, we focus on the ReLU FNNs $\mathcal{F}_{d_1,d_2}(\mathcal{W}, \mathcal{D}, \mathcal{S}, \mathcal{B})$ that $d_1 = d$ and $d_2 = 1$.

Furthermore, we introduce the definition of Hölder class, which is a generalization of Lipschitz continuity, and is commonly used to characterize the smoothness of functions (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Lu et al., 2021; Jiao et al., 2023).

Definition 1 (Hölder class)**.** *We denote the Hölder class* $\mathcal{H}^\zeta\left([0,1]^d, B\right)$ *as*

$$\mathcal{H}^\zeta([0,1]^d, B) := \Big\{ f : \mathbb{R}^d \to \mathbb{R},\ \max_{\|\boldsymbol{s}\|_1 \leq t} \|\partial^{\boldsymbol{s}} f\|_\infty \leq B,\ \max_{\|\boldsymbol{s}\|_1 = t} \sup_{x \neq y} \frac{|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(y)|}{\|x - y\|_\infty^s} \leq B \Big\},$$

for some $B, \zeta > 0$ *with* $\zeta = t + s$, $t \in \mathbb{N}_0$ *and* $s \in (0, 1]$*, and* $d \in \mathbb{N}$*.*

Here we assume that the underlying function f_0, defined in (2), is Hölder continuous with parameters B and ζ, and thus belongs to the Hölder class $\mathcal{H}^\zeta\left([0,1]^d, B\right)$ in Definition 1.

3.2 Unweighted and Reweighted Estimators

Based on the training sample $\mathbb{D} := \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$, traditional nonparametric quantile regression with a standard empirical risk minimization (ERM) framework leads to the naive unweighted estimator given by

$$\widehat{f}_{\mathbb{D}} \in \arg\min_{f \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - f(\boldsymbol{X}_i)), \tag{5}$$

where $\mathcal{G}$ denotes a function class of ReLU FNNs as defined in Section 3.1. Note that the right side of (5) is an unbiased estimator of $\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\rho_\tau(Y - f_0(\boldsymbol{X}))]$ in the absence of covariate shift. However, when covariate shift happens, this term becomes biased, potentially resulting in an inaccurate predictive estimator.

To tackle the issue of covariate shift, it is crucial to measure the discrepancy between the source and target distributions, and thus we introduce the concept of density ratio. Specifically, we denote $p_{\boldsymbol{X}}$ and $q_{\boldsymbol{X}}$ as the probability density functions of $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$, respectively. Then, the density ratio can be defined as

$$r(\boldsymbol{x}) = \frac{q_{\boldsymbol{X}}(\boldsymbol{x})}{p_{\boldsymbol{X}}(\boldsymbol{x})},$$

which quantifies the dissimilarity between $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$. Once the density ratio $r(\boldsymbol{x})$ is known, we can simply solve the covariate shift problem by considering the following reweighted ERM task that

$$\widehat{f}_{r,\mathbb{D}} \in \arg\min_{f \in \mathcal{J}} \frac{1}{n} \sum_{i=1}^n r(\boldsymbol{X}_i) \rho_\tau(f(\boldsymbol{X}_i) - Y_i), \tag{6}$$

where $\mathcal{J}$ denotes a function class of ReLU FNNs. Here, we use different notations of function class for theoretical simplicity. The reweighted procedure ensures an unbiased estimator

of $\mathbb{E}_{(\boldsymbol{X},Y)\sim Q_{\boldsymbol{X},Y}}[\rho_\tau(Y - f_0(\boldsymbol{X}))]$ since it is easy to verify that $\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[r(\boldsymbol{X})\rho_\tau(Y - f_0(\boldsymbol{X}))] = \mathbb{E}_{(\boldsymbol{X},Y)\sim Q_{\boldsymbol{X},Y}}[\rho_\tau(Y - f_0(\boldsymbol{X}))]$. Similar treatments can be also referred to Ma et al. (2023); Feng et al. (2023) under the reproducing kernel Hilbert space (RKHS) framework.

Yet, it is unrealistic to get the exact density ratio as the prior information in practice, and most existing studies (Ma et al., 2023; Feng et al., 2023) only investigate the statistical property of the reweighted estimator under the squared loss function and Lipschitz continuous loss functions with some known density ratios. Thus, it is still an open and fundamental problem to propose a reweighted estimator with the plug-in estimated ratio and investigate its statistical properties.

3.3 Pre-training Reweighted Estimator

In the scenario where access to the exact density ratio is not available, it is natural to consider an alternative problem akin to (6) with a plug-in density ratio estimator. Inspired by this, we propose a two-step pre-training reweighted estimator. In the first step, a density ratio estimator is obtained through least-squares density ratio fitting using a deep neural network (Kanamori et al., 2009; Sugiyama et al., 2012). Subsequently, we leverage this plug-in density ratio estimator to compute the final pre-training reweighted estimator. To the best of our knowledge, the established theoretical results in Section 4 are the first attempt in the literature that a pre-training density estimator is considered in the analysis instead of simply using the true density ratio.

In the first step, motivated by the fact that the exact density ratio r is the minimizer of $\min_u \frac{1}{2}\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[(u(\boldsymbol{X}) - r(\boldsymbol{X}))^2\right]$ overall measurable function u, we remove the negligible terms that are independent of $r(\boldsymbol{X})$, then the optimization becomes

$$r = \arg\min_u \; L(u) = \arg\min_u \frac{1}{2}\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[u(\boldsymbol{X})^2] - \mathbb{E}_{\boldsymbol{X}\sim Q_{\boldsymbol{X}}}[u(\boldsymbol{X})]. \tag{7}$$

Suppose that we obtain the extra unlabeled samples drawn from $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$, denoted by $\mathbb{S}_P := \{\boldsymbol{X}_i^P\}_{i=1}^m$ and $\mathbb{S}_Q := \{\boldsymbol{X}_i^Q\}_{i=1}^m$, respectively. Consequently, the empirical version of (7) can be formulated as

$$\widehat{r}_{\mathbb{S}} \in \arg\min_{u\in\mathcal{U}} \widehat{L}_{\mathbb{S}}(u) = \arg\min_{u\in\mathcal{U}} \frac{1}{2m}\sum_{i=1}^m u(\boldsymbol{X}_i^P)^2 - \frac{1}{m}\sum_{i=1}^m u(\boldsymbol{X}_i^Q), \tag{8}$$

where $\widehat{L}_{\mathbb{S}}(\cdot)$ denotes the empirical pre-training risk and $\mathcal{U}$ also denotes a function class of ReLU FNNs as defined in Section 3.1.

In the second step, by substituting the true density ratio r in (6) with its estimator $\widehat{r}_{\mathbb{S}}$ obtained from (8), we can obtain the final pre-training reweighted estimator over a hypothesis function class $\mathcal{M}$ defined as follows

$$\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} \in \arg\min_{f\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n \widehat{r}_{\mathbb{S}}(\boldsymbol{X}_i)\rho_\tau(f(\boldsymbol{X}_i) - Y_i). \tag{9}$$

Here we want to emphasize that the density ratio $\widehat{r}_{\mathbb{S}}$ is estimated by using the unlabeled samples $\mathbb{S}_P$ and $\mathbb{S}_Q$, which are independent of the data $\mathbb{D}$ used in solving (9). This independence assumption is a common presupposition in the context of covariate shift. It allows

us to obtain an accurate estimation for the density ratio r using a large volume of independent unlabeled data. Additionally, this assumption plays a crucial role in our theoretical analysis. We summarize our proposed two-step pre-training deep nonparametric quantile regression algorithm in Algorithm 1.

Algorithm 1 The two-step pre-training deep nonparametric quantile regression algorithm

1. **Input:** Unlabeled data $\mathbb{S}_P = \{\boldsymbol{X}_i^P\}_{i=1}^m$ from $P_{\boldsymbol{X}}$, $\mathbb{S}_Q = \{\boldsymbol{X}_i^Q\}_{i=1}^m$ from $Q_{\boldsymbol{X}}$ and labeled data $\mathbb{D} = \{(\boldsymbol{X}_j, Y_j)\}_{j=1}^n$ sampled from $P_{\boldsymbol{X},Y}$.
2. **Pre-train the density ratio:** Obtain $\widehat{r}_{\mathbb{S}}$ by solving (8) using $\mathbb{S}_P$ and $\mathbb{S}_Q$.
3. **Reweighted nonparametric quantile regression:** Obtain $\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}$ by solving (9) using $\widehat{r}_{\mathbb{S}}$ and $\mathbb{D}$.
4. **Return:** the pre-training reweighted estimate $\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}$.

Note that the first step of Algorithm 1 adopts unsupervised learning to estimate the density ratio r based on the unlabeled data $\mathbb{S}_P$ and $\mathbb{S}_Q$. The second step of Algorithm 1 integrates the estimated pre-trained density ratio $\widehat{r}_{\mathbb{S}}$ to estimate the underlying regression function f_0 by using the labeled data $\mathbb{D}$, and thus amalgamates information derived from the estimated density ratio and the labeled data. It is highlighted that the proposed algorithm encapsulates the essence of the pre-training reweighted algorithm, which strategically combines unsupervised and supervised learning paradigms to address the challenges of covariate shift under the regression setting.

It is worthy pointing out that the difficulty of estimating the density ratio r depends on the shapes of the distributions. Specifically, if $p_{\boldsymbol{X}}$ and $q_{\boldsymbol{X}}$ have different tail behaviors, for example, one is heavy-tailed and another is light-tailed, then the problem of estimating r may become challenging. This is largely due to the fact that in the tail regions, r may be very large or even unbounded. Then, the straightforward use of the original estimate of the density ratio may result in a significant increase in variance due to the potential presence of large quantities of $\widehat{r}_{\mathbb{S}}$ at some covariates. To address this issue, we consider the truncated density ratio estimator $\widehat{r}_{\xi,\mathbb{S}}$, that is

$$\widehat{r}_{\xi,\mathbb{S}} \in \arg\min_{u \in T_\xi \mathcal{U}} \widehat{L}_{\mathbb{S}}(u), \tag{10}$$

where $T_\xi \mathcal{U} = \{T_\xi u;\ u \in \mathcal{U}\}$ denotes a truncated set of $\mathcal{U}$ with

$$T_\xi u = \begin{cases} u, & \text{if } u \le \xi, \\ \xi, & \text{otherwise,} \end{cases}$$

and ξ is some previously chosen threshold. Note that the truncated function class $T_\xi \mathcal{U}$ can be realized by adding an extra activation function $\sigma'_\xi(x) = \min(x, \xi)$ to the original Relu FNN function class $\mathcal{U}$. Thus, the ratio estimator $\widehat{r}_{\mathbb{S}}$ in Algorithm 1 is replaced with the truncated counterpart $\widehat{r}_{\xi,\mathbb{S}}$ to obtain the following estimator

$$\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} \in \arg\min_{f \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^n \widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X}_i) \rho_\tau(f(\boldsymbol{X}_i) - Y_i). \tag{11}$$

Note that the estimation accuracies of $\widehat{r}_{\mathbb{S}}$ and $\widehat{r}_{\xi,\mathbb{S}}$ affect the performance of $\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}$ and $\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}$, respectively. In practice, obtaining unlabeled data is typically much cheaper and more readily available compared to labeled data, and thus an large amount of unlabeled data can usually be collected to ensure the estimation accuracy. For example, collecting response variables is laborious and costly in electronic health records (EHR) data sets, while the covariates are quite easy to obtain from the databases (Gronsbell and Cai, 2018). Numerically, we use $m = 1000$ unlabeled data points to guarantee accurate estimation of the density ratio in our all simulated examples, which yields satisfactory numerical performance.

4. Theoretical Analysis

In this section, we provide the theoretical analysis on the non-asymptotic error bounds for the unweighted (Section 4.1), reweighted (Section 4.2), and pre-training reweighted estimators (Section 4.3), respectively.

4.1 Theoretical Analysis for Unweighted Estimator

We establish the non-asymptotic error bounds for the unweighted estimator $\widehat{f}_{\mathbb{D}}$ defined in (5) (Theorems 7 and 8) under two broad scenarios of covariate shift where the density ratio is either bounded or unbounded. To start with, we define the ball centered at $f(\boldsymbol{x})$ with radius $c \geq 0$ that

$$\mathbf{B}(f(\boldsymbol{x}), c) = \big\{y : |y - f(\boldsymbol{x})| \leq c\big\},$$

and denote the conditional density function and the cumulative distribution function of Y given $\boldsymbol{X}$ as $p_{Y|\boldsymbol{X}}$ and $F_{Y|\boldsymbol{X}}(\cdot)$, respectively. The following technical assumptions are introduced for our theoretical development.

Assumption 1. *The density ratio r is well-defined, that is, the source density $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$ is required whenever the target density function $q_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.*

Assumption 2. *Whether $\boldsymbol{X}$ is distributed as specified by $P_{\boldsymbol{X}}$ or $Q_{\boldsymbol{X}}$, the τ-th quantile of ε given $\boldsymbol{X}$ is zero and ε given $\boldsymbol{X}$ shares the same conditional distribution.*

Assumption 3. *There exist some positive constants ξ_1, ξ_2 and κ such that for any $|\delta| \leq \xi_1$ and $y \in \mathbf{B}(f_0(\boldsymbol{x}), \xi_2)$, there holds*

$$\big|F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y + \delta) - F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y)\big| \geq \kappa|\delta|, \quad \textit{almost surely.}$$

Moreover, for some absolute positive constant $\tilde{c}$, there holds $\sup_{t\in\mathbb{R}} p_{Y|\boldsymbol{X}=\boldsymbol{x}}(t) \leq \tilde{c}$.

Assumption 1 requires the density ratio $r(\cdot) = q_{\boldsymbol{X}}(\cdot)/p_{\boldsymbol{X}}(\cdot)$ always exists, implying that the two densities must have overlapping support or the support of the source covariates must cover that of the target covariates. Similar assumption is commonly required in the literature of importance weighting (Cortes et al., 2010) and covariate shift (Ma et al., 2023; Feng et al., 2023). Assumption 2 ensures the exact equivalence of $\widetilde{f}$ in (4) to the underlying regression function f_0 in (2) and the requirement that $P_{Y|\boldsymbol{X}} = Q_{Y|\boldsymbol{X}}$. This assumption is naturally satisfied if ε is independent of $\boldsymbol{X}$. Assumption 3 can be regarded as an adaptive self-calibration condition governing the conditional distribution of Y given $\boldsymbol{X}$,

and significantly differs from the commonly assumed self-calibration condition in literature (Shen et al., 2021; Madrid Padilla and Chatterjee, 2022), where the specific case that $\xi_2 = 0$ is considered. Note that Assumption 3 requires the existence of a neighborhood around $y \in \mathbf{B}(f_0(\boldsymbol{x}), \xi_2)$ within which the perturbation of $F_{Y|\boldsymbol{X}}(\cdot)$ exceeds the variation in y. It is crucial to emphasize that Assumption 3 plays a vital role in deriving the refined error decomposition in Lemma 1, and consequently leads to the minimax optimal result in Theorem 7. By Assumption 3, the approximation error within our decomposition in Lemma 1 is quantified by the term $\inf_{f\in\mathcal{G}} \|f - f_0\|_\infty^2$. This is in sharp contrast to the commonly assumed self-calibration condition in literature, where the approximation error is quantified by the term $\inf_{f\in\mathcal{G}} \|f - f_0\|_\infty$ within the error decomposition using the Lipschitz continuity of the quantile loss, and thus leads to a suboptimal theoretical order of the error bound.

Remark 2. *Note that Assumption 1 does not hold if the two densities have no overlap supports since there will be regions where $p_{\boldsymbol{X}}(\boldsymbol{x}) = 0$ and $q_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, leading to a division by 0 and the failure of the definition. Under these cases, performing direct density ratio estimation becomes problematic, and some different treatments should be considered. If the non-overlapping supports are known in advance, one possible treatment is to restrict the estimation to the region where both densities have overlapping support. Therefore, we only estimate the density ratio in the overlapping region and assign the ratio to be zero where the region is outside of the support of $p_{\boldsymbol{X}}$. If the non-overlapping supports are unknown, the applied reweighted framework in this paper is no longer applicable, and some new framework needs to be developed (Mallinar et al., 2024), which may use some other divergence measures such as the total variation distance (Rényi, 1961; Liese and Vajda, 2006) to evaluate the difference between distributions. Yet, we want to point out that these methods and their theoretical investigation are out of this paper's scope, and we leave this interesting problem as further work.*

Remark 3. *It is worthy pointing out under the special case that $\xi_2 = 0$, Assumption 3 is similar to the assumptions required in Shen et al. (2021); Madrid Padilla and Chatterjee (2022), and is much weaker than assumptions required in He and Shi (1994); Belloni and Chernozhukov (2011); Padilla et al. (2022a). Precisely, Condition 2 in He and Shi (1994) assumes that the density function of Y is lower bounded by some positive constant; Condition D.1 in Belloni and Chernozhukov (2011) requires that the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ must be both continuously differentiable and bounded away from zero uniformly across all quantile levels and for any $x \in \mathcal{X}$; Assumption 2 in Padilla et al. (2022a) also assumes that the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ is upper-bounded by a positive constant.*

4.1.1 Error Decomposition and Estimates for Error Terms

In this section, we provide a novel error decomposition of the unweighted estimator including the approximation error and statistical error terms, which is important for the non-asymptotic error analysis. We define $f^* \in \mathcal{G}$ as the best approximation of f_0 within some function space $\mathcal{G}$ that is uniformly bounded by B with $B \geq 1$, where the approximation quality is measured by the distance of the $\|\cdot\|_\infty$ norm, namely,

$$f^* \in \arg\min_{f\in\mathcal{G}} \|f - f_0\|_\infty,$$

Together with Assumptions 2 and 3 density, the error $L_2(P_{\boldsymbol{X}})$ of the unweighted estimator $\widehat{f}_{\mathbb{D}}$ can be decomposed into two distinct components as stated in the following lemma.

Lemma 1. *Suppose that Assumptions 2 and 3 are satisfied, and the function space $\mathcal{G}$ is also uniformly bounded by B with $B \geq 1$. Then, the unweighted estimator $\widehat{f}_{\mathbb{D}}$ defined in* (5) *satisfies*

$$\|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(P_{\boldsymbol{X}})}^2 \lesssim B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right] + B^2 \inf_{f\in\mathcal{G}} \|f - f_0\|_\infty^2.$$

Lemma 1 decomposes the upper bound of $\|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(P_{\boldsymbol{X}})}^2$ into two terms, and it suffices to separately bound the statistical error term $\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))]$ and the approximation error term $\inf_{f\in\mathcal{G}} \|f - f_0\|_\infty$. For the approximation error term, we use the technical tools and results in Yarotsky (2017); Petersen and Voigtlaender (2018) to show that for any Hölder continuous function, it can be approximated arbitrarily well by a ReLU FNN with properly chosen depth, size, and weights bound in Theorem 4. Note that for mathematical simplicity, we assume that the support of $\boldsymbol{X}$ is $\mathcal{X} = [0,1]^d$, which can be easily relaxed.

Theorem 4. *Suppose that $f \in \mathcal{H}^\zeta([0,1]^d, B)$. For any $\epsilon \in (0,1)$, there exists a ReLU FNN function ψ with the depth $\mathcal{D} \leq \mathcal{O}(\log(1/\epsilon))$, size $\mathcal{S} \leq \mathcal{O}\left(\epsilon^{-d/\zeta}\log(1/\epsilon)\right)$, weights bound $\mathcal{B} \leq \mathcal{O}\left(B\epsilon^{-d/\zeta}\right)$ such that*

$$\|f - \psi\|_\infty \leq B\epsilon.$$

Theorem 4 establishes the upper bound of the approximation error which is related to the depth $\mathcal{D}$, size $\mathcal{S}$ and weights bound $\mathcal{B}$ of the considered ReLU FNN. Specifically, it demonstrates that the upper bound of the approximation error is of order $\mathcal{O}(\mathcal{S}^{-\zeta/d})$ with an omitted logarithmic term, which is consistent with the results in Yarotsky (2017); Petersen and Voigtlaender (2018). It is worthy pointing out that the proof of Theorem 4 partially follows the technical treatment as that in Yarotsky (2017) and also adopts the technical tools used in Petersen and Voigtlaender (2018). This deliberate choice facilitates the derivation of the weights bound, aligning with the rate presented in Petersen and Voigtlaender (2018). The completed proof of Theorem 4 is provided in Section A.1.1.

For the statistical error, we use the empirical process techniques (Vaart and Wellner, 2023; Van der Vaart, 2000; Van de Geer and van de Geer, 2000; Vershynin, 2018) to derive the error bound in Theorem 6, which is characterized by the complexity of the considered function class $\mathcal{G}$, including measures such as the covering number (Anthony et al., 1999).

Definition 5. *(Covering number) For $\delta > 0$, the covering number $\mathcal{N}\big(\delta, \mathbf{F}, \tilde{d}\big)$ related to a semi-metric $\tilde{d}$ on the set $\mathbf{F}$ is defined as*

$$\mathcal{N}(\delta, \mathbf{F}, \tilde{d}) = \min_\kappa \Big\{ \text{ there are } g_1, \ldots, g_\kappa \text{ such that } \min_{1\leq j\leq\kappa} \tilde{d}\,(f, g_j) \leq \delta, \text{ for any } f \text{ in } \mathbf{F} \Big\}.$$

With the help of empirical process techniques, the explicit upper bound of the statistical error is provided in the following theorem.

Theorem 6. *Suppose that Assumption 3 holds and the considered function class $\mathcal{G}$ is a ReLU FNN as defined in Section 3.1 and $\|f^*\|_\infty \leq B$, we have*

$$\mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[\rho_\tau(Y-\widehat{f}_{\mathbb{D}}(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\right] \leq \mathcal{O}\Big(\frac{B\mathcal{S}\mathcal{D}\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\Big),$$

where $\mathbb{E}_{\mathbb{D}}$ is the expectation taken over the training sample $\mathbb{D}$.

Theorem 6 provides the upper bound of the statistical error, which is related to the depth $\mathcal{D}$, size $\mathcal{S}$ and weights bound $\mathcal{B}$ of the considered ReLU FNN. Note that if we keep the other terms fixed and omit the logarithmic term, the rate of the upper bound becomes $\mathcal{O}\left(\frac{\mathcal{S}\mathcal{D}}{n}\right)$, and the fastest rate $\mathcal{O}\left(\frac{1}{n}\right)$ can also be achieved if we further assume $\mathcal{S}$ and $\mathcal{D}$ are fixed. We want to emphasize that an innovative proof technique is adopted for analyzing the statistical error, where we first derive the tail probability of $\rho_\tau(Y-\widehat{f}_{\mathbb{D}}(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))$ by using the Bernstein's inequality, and then characterize the statistical error by utilizing the covering number under the norm $\|\cdot\|_\infty$. This innovative technical treatment is motivated by the properties of the quantile loss, which exhibits Lipschitz continuity, boundedness, and local strong convexity over the target function. Consequently, we can establish a tight bound for the covering number of the ReLU FNN under the norm $\|\cdot\|_\infty$ and leverage the local convexity of $\rho_\tau(\cdot)$, which yields the sharp bound of $\mathcal{O}\left(\frac{\mathcal{S}\mathcal{D}}{n}\right)$ for the statistical error in Theorem 6. Clearly, the technical tools adopted in our theoretical analysis significantly differ from and simplify the commonly used local Rademacher complexity techniques (Bartlett et al., 2005). More details are provided in Section A.1.2

Now we are ready to investigate the theoretical performance of the unweighted estimator under covarite shift. Specifically, we establish the upper bound of $\|\widehat{f}_{\mathbb{D}}-f_0\|_{L_2(Q_{\boldsymbol{X}})}$ under two different scenarios where the density ratio is either uniformly bounded or unbounded but has a finite second moment. We first introduce the uniformly bounded assumption for the density ratio below.

Assumption 4 (Uniformly bounded)**.** *The density ratio $r(\boldsymbol{x}) = q_{\boldsymbol{X}}(\boldsymbol{x})/p_{\boldsymbol{X}}(\boldsymbol{x})$ is upper-bounded, that is, $\Gamma := \sup_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) < \infty$.*

Assumption 4 is a commonly used condition in the literature of covariate shift (Cortes et al., 2010; Ma et al., 2023; Feng et al., 2023). Note that Γ is typically assumed to be larger than 1 since the equality indicates that $P_{\boldsymbol{X}}$ and $Q_{\boldsymbol{X}}$ are the same. Under Assumption 4, it is easy to verify that $\|\widehat{f}_{\mathbb{D}}-f_0\|^2_{L_2(Q_{\boldsymbol{X}})} \leq \Gamma\|\widehat{f}_{\mathbb{D}}-f_0\|^2_{L_2(P_{\boldsymbol{X}})}$. Then, along with the approximation and statistical errors established in Theorems 4 and 6, we can provide the non-asymptotic error bound for $\widehat{f}_{\mathbb{D}}$ with the proper choices of parameters, including the size $\mathcal{S}$, depth $\mathcal{D}$, and weights boundedness $\mathcal{B}$ of the considered ReLU FNN class.

Theorem 7. *Suppose that Assumptions 2-4 are satisfied. If the function class $\mathcal{G}$ is a ReLU FNN function class bounded by $B \geq 1$ with the size $\mathcal{S} = \mathcal{O}\left(n^{\frac{d}{d+2\zeta}}\log n\right)$ and the depth $\mathcal{D} = \mathcal{O}(\log n)$, and weights boundedness $\mathcal{B} = \mathcal{O}\left(Bn^{\frac{d}{d+2\zeta}}\right)$, then we have*

$$\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{\mathbb{D}}-f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \leq \mathcal{O}\Big(\Gamma B^2 n^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\Big).$$

Theorem 7 establishes the non-asymptotic error bound for the unweighted estimator $\widehat{f}_{\mathbb{D}}$ under the uniformly bounded ratio case, and the obtained bound concurs with the minimax optimal rate in the literature of nonparametric regression (Stone, 1982; Gyorfi et al., 2002; Tsybakov, 2009). This implies that even when the phenomenon of covariate shift occurs, the unweighted estimator is still optimal when the density ratio is uniformly bounded. The proof sketch of Theorem 7 is provided below, which may give some intuitions on why the minimax optimal rate can still be achieved under the bounded case. Note that in the proof of Theorem 7, we first show that the convergence rate of $\widehat{f}_{\mathbb{D}}$ with respect to the source distribution aligns with the minimax lower bound in the literature of nonparametric regression (Stone, 1982; Tsybakov, 2009), i.e., $\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(P_{\boldsymbol{X}})} \leq \mathcal{O}\left(B^2 n^{-\frac{2\zeta}{d+2\zeta}} (\log n)^3\right)$. Then, when the covariate shift occurs and the density ratio is assumed to be bounded by $\sup_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) \leq \Gamma$, we have

$$\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})} \leq \Gamma \mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(P_{\boldsymbol{X}})} \leq \mathcal{O}\left(\Gamma B^2 n^{-\frac{2\zeta}{d+2\zeta}} (\log n)^3\right).$$

This implies that the convergence rate of $\widehat{f}_{\mathbb{D}}$ with respect to the target distribution is also minimax optimal (ignoring the constants and the log terms). Moreover, since the density ratio is assumed to be uniformly bounded, it implies that the shift between the source and target distribution may not be too large, and thus may be well controlled. Therefore, the unweighted estimator may still achieve good prediction performance. However, such a uniform boundedness condition is somewhat restrictive in practice, and the following assumption introduces a relaxation, assuming that the density ratio has a finite second moment.

Assumption 5 (Finite second moment)**.** *The density ratio r has a finite second moment with respect to $P_{\boldsymbol{X}}$, that is, $V := \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^2(\boldsymbol{X})] < \infty$.*

Note that Assumption 5 is much weaker than Assumption 4, since the finite second moment is implied by the uniformly bounded condition with $V^2 = \Gamma$ (Ma et al., 2023; Feng et al., 2023). To see this, we have $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^2(\boldsymbol{X})] = \mathbb{E}_{\boldsymbol{X}\sim Q_{\boldsymbol{X}}}[r(\boldsymbol{X})] \leq \Gamma$. Interestingly, these two conditions on the density ratio are related to Rényi divergence $D_l(q_{\boldsymbol{X}}||p_{\boldsymbol{X}})$ (Rényi, 1961) between the source and target densities $p_{\boldsymbol{X}}$ and $q_{\boldsymbol{X}}$. Specifically, the conditions on Assumptions 4 and 5 are equivalent to requiring $D_\infty(q_{\boldsymbol{X}}||p_{\boldsymbol{X}})$ and $D_2(q_{\boldsymbol{X}}||p_{\boldsymbol{X}})$ to be bounded (Cortes et al., 2010), respectively. The following theorem establishes the non-asymptotic error bound for the unweighted estimator $\widehat{f}_{\mathbb{D}}$ under the second moment bounded case.

Theorem 8. *Suppose that Assumptions 2, 3, and 5 are satisfied. If the function class $\mathcal{G}$ is a ReLU FNN function class bounded by $B \geq 1$ with the size $\mathcal{S} = \mathcal{O}\left(n^{\frac{d}{d+2\zeta}} \log n\right)$ and the depth $\mathcal{D} = \mathcal{O}(\log n)$, and the weights bound $\mathcal{B} = \mathcal{O}\left(Bn^{\frac{d}{d+2\zeta}}\right)$, then we have*

$$\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \leq \mathcal{O}\left(VB^2 n^{-\frac{\zeta}{d+2\zeta}} (\log n)^{\frac{3}{2}}\right).$$

Although the unweighted estimator $\widehat{f}_{\mathbb{D}}$ is still consistent with the true quantile function f_0, it is clear that its convergence rate of order $\mathcal{O}(n^{-\frac{\zeta}{d+2\zeta}})$ is suboptimal compared to the minimax rate in Stone (1982); Gyorfi et al. (2002); Tsybakov (2009). Consequently, some additional reweighted adjustments are required to tackle this issue.

4.2 Theoretical Analysis for Reweighted Estimator

In this part, we give the non-asymptotic error bounds for the reweighted estimator using the exact density ratio. The following theorem provides the non-asymptotic error bound for the reweighted estimator $\widehat{f}_{r,\mathbb{D}}$ defined in (6) under the case that the density ratio is uniformly bounded.

Theorem 9. *Suppose that all the assumptions in Lemma 1 as well as Assumptions 1 and 4 are satisfied. If the function class $\mathcal{J}$ is a ReLU FNN function class bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}\Big(n^{\frac{d}{d+2\zeta}} \log n\Big)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and the weights bound $\mathcal{B} = \mathcal{O}\Big(Bn^{\frac{d}{d+2\zeta}}\Big)$, then we have*

$$\mathbb{E}_{\mathbb{D}}\big[\|\widehat{f}_{r,\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(\Gamma^2 B^2 n^{-\frac{2\zeta}{d+2\zeta}} (\log n)^3\Big).$$

In Theorem 9, we obtain a convergence rate of order $\mathcal{O}(\Gamma^2 B^2 n^{-2\zeta/(d+2\zeta)}(\log n)^3)$, which includes an additional term Γ compared to the rate in Theorem 7. This is due to the influence of the density ratio in the reweighted estimation procedure in (6) and in fact, the derived rate in Theorem 9 is exactly the same as that in Theorem 7 if Γ is a constant.

When the density ratio is unbounded but has a finite second moment, as noted in Section 4.3, we also apply a truncated density ratio. Specifically, let $\xi > 0$ be the pre-specified truncation level, then the truncated density ratio is defined as

$$T_\xi r(\boldsymbol{x}) = \begin{cases} r(\boldsymbol{x}), & \text{if } r(\boldsymbol{x}) \leq \xi, \\ \xi, & \text{otherwise.} \end{cases}$$

Thus, the reweighted estimator with a truncated density ratio can be obtained by solving the following optimization task

$$\widehat{f}_{T_\xi r,\mathbb{D}} \in \arg\min_{f \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} T_\xi r(\boldsymbol{X}_i) \rho_\tau\big(f(\boldsymbol{X}_i) - Y_i\big), \tag{12}$$

where $\mathcal{K}$ refers to a certain hypothesis class of measurable functions. The non-asymptotic error bound for the reweighted estimator $\widehat{f}_{T_\xi r,\mathbb{D}}$ is established in the following theorem with an appropriate choice of the truncated level ξ.

Theorem 10. *Suppose that all the assumptions in Lemma 1 as well as Assumptions 1 and 5 are satisfied. If the function class $\mathcal{K}$ is a ReLU FNN function class bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}\Big(n^{\frac{d}{d+6\zeta}} \log n\Big)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and the weights bound $\mathcal{B} = \mathcal{O}\Big(Bn^{\frac{d}{d+6\zeta}}\Big)$, then we have*

$$\mathbb{E}_{\mathbb{D}}\Big[\|\widehat{f}_{T_\xi r,\mathbb{D}} - f_0\|^2_{L^2(Q_{\boldsymbol{X}})}\Big] \leq \mathcal{O}\Big(V^{\frac{4}{3}} B^2 n^{-\frac{2\zeta}{d+6\zeta}} \log n\Big).$$

Note that the obtained convergence rate of order $\mathcal{O}(n^{-\frac{2\zeta}{d+6\zeta}})$ is suboptimal. In the proof of Theorem 10, we divide the statistical error with the truncated ratio into two terms, the

error with the true density ratio and their difference. With the similar argument in the proof of Theorem 6 and Markov inequality, we can show that the upper bounds of the two terms are of orders $\mathcal{O}(\frac{B\mathcal{S}\mathcal{D}\xi^2\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n})$ and $\mathcal{O}(\frac{B^2V^2}{\xi})$, respectively. The trade-off between the two upper bounds leads to such a suboptimal rate. However, it can be further improved with extra conditions on the density ratio, which is shown in the next assumption.

Assumption 6. *There exists some constant $\delta > 0$ such that the density ratio r has a finite $(1+\delta)$-th moment with respect to $P_{\boldsymbol{X}}$, that is, $U := \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^{1+\delta}(\boldsymbol{X})] < \infty$.*

Note that Assumption 6 is more general than Assumption 5, where the latter assumption is a special case of Assumption 6 when $\delta = 1$. We can then obtain the convergence rate of $\widehat{f}_{T_\xi r,\mathbb{D}}$ in the following Corollary.

Corollary 11. *Suppose that all the assumptions in Lemma 1 as well as Assumptions 1 and 6 are satisfied. If the function class $\mathcal{K}$ is a ReLU FNN function class bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}\left(n^{\frac{d}{d+(2+4/\delta)\zeta}}\log n\right)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and weights bound $\mathcal{B} = \mathcal{O}\left(Bn^{\frac{d}{d+(2+4/\delta)\zeta}}\right)$, it follows that*

$$\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{T_\xi r,\mathbb{D}} - f_0\|^2_{L^2(Q_{\boldsymbol{X}})}\right] \leq \mathcal{O}\left(U^{2\delta/(2+\delta)}B^2 n^{-\frac{2\zeta}{d+(2+4/\delta)\zeta}}\log n\right).$$

When $\delta = 1$, corresponding to the second moment bounded case, the convergence rate obtained in Corollary 11 reduces to $\mathcal{O}(n^{-\frac{2\zeta}{d+6\zeta}})$, which coincides with that of Theorem 10. When $\delta \to \infty$, the convergence achieves the minimax optimal rate of $\mathcal{O}(n^{-\frac{2\zeta}{d+2\zeta}})$ in nonparametric regression.

4.3 Theoretical Analysis for Pre-training Reweighted Estimator

In this part, we present the most pivotal result of this paper, focusing on the convergence rate of the pre-training reweighted estimator. The established results are closely aligned with, yet extend beyond, the findings of both the unweighted and reweighted estimators. Furthermore, we concurrently consider the cases where the density ratio is either bounded or unbounded but with certain bounded moments.

For the uniformly bounded case, as elucidated in Algorithm 1, obtaining the pre-training reweighted estimator necessitates the preliminary acquisition of a density ratio estimator, a process grounded in least-squares density ratio fitting. This density ratio estimator is subsequently incorporated into the empirical reweighted risk to derive the pre-training reweighted estimator. To reach this ultimate result, we need to introduce some additional conditions on the density ratio, as outlined in Assumptions 7 and 8. To proceed, we first establish the non-asymptotic error bound of the density ratio estimator, as expounded in Theorem 12. Subsequently, we derive the non-asymptotic error bound of the pre-training reweighted estimator, as detailed in Theorem 13.

Assumption 7. *The density ratio r is Hölder continuous that $r \in \mathcal{H}^\alpha([0,1]^d, \Gamma)$ for some positive α.*

Assumption 7 is a standard condition restricting the underlying function space of the density ratio r, which is important to establish the estimation consistency of the pre-training weight $\widehat{r}_{\mathbb{S}}$.

Theorem 12. *Suppose that Assumptions 1, 4 and 7 are satisfied. If the function class $\mathcal{U}$ is a ReLU FNN function space bounded by Γ, with the size $\mathcal{S} = \mathcal{O}\big(m^{\frac{d}{d+2\alpha}} \log m\big)$, the depth $\mathcal{D} = \mathcal{O}(\log m)$, and the weights bound $\mathcal{B} = \mathcal{O}\big(\Gamma m^{\frac{d}{d+2\alpha}}\big)$, then we have*

$$\mathbb{E}_{\mathbb{S}}\big[\|\widehat{r}_{\mathbb{S}} - r\|^2_{L_2(P_{\boldsymbol{X}})}\big] \leq \mathcal{O}\big(\Gamma^3 m^{-\frac{2\alpha}{d+2\alpha}} (\log m)^3\big).$$

Theorem 12 establishes a non-asymptotic error bound for the density-ratio estimator $\widehat{r}_{\mathbb{S}}$ under the uniformly bounded case. Ignoring the fixed and log terms, the obtained convergence rate becomes $\mathcal{O}\big(m^{-\frac{2\alpha}{d+2\alpha}}\big)$, which concurs with the minimax optimal rate in the literature of nonparametric estimation.

With the obtained convergence rate of $\widehat{r}_{\mathbb{S}}$ and the following lower bounded assumption, we are now ready to give the non-asymptotic error bound of the final pre-training reweighted estimator $\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}$ defined in (9).

Assumption 8. *The density ratio r is bounded away from zero, that is, $\Upsilon := \inf_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) > 0$.*

Assumption 8 further requires that the density ratio is uniformly lower bounded, which excludes some extreme cases.

Theorem 13. *Suppose that Assumptions 1-4, 7 and 8 are satisfied. If the function class $\mathcal{M}$ is a ReLU FNN function class bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}\big(n^{\frac{d}{d+2\zeta}} \log n\big)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and the weights bound $\mathcal{B} = \mathcal{O}\big(Bn^{\frac{d}{d+2\zeta}}\big)$, we have*

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2\Gamma^2 n^{\frac{-2\zeta}{2\zeta+d}} (\log n)^3\Big) + \mathcal{O}\Big(\frac{B^2\Gamma^3}{\Upsilon} m^{-\frac{2\alpha}{d+2\alpha}} (\log m)^3\Big).$$

Moreover, if $m \geq \Omega\Big(\big(\frac{\Gamma}{\Upsilon}\big)^{\frac{d+2\alpha}{2\alpha}} n^{\frac{\zeta(d+2\alpha)}{\alpha(d+2\zeta)}}\Big)$, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2\Gamma^2 n^{-\frac{2\zeta}{d+2\zeta}} (\log n)^3\Big).$$

Theorem 13 establishes the convergence rates of the pre-training reweighted estimator $\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}$ under the uniformly bounded case, which consists of two terms representing the estimation accuracies of the underlying function and the density ratio, respectively. Clearly, Theorem 13 provides valuable theoretical insights for determining an appropriate pre-training sample size m, i.e., $m \geq \mathcal{O}\big(\big(\frac{\Gamma}{\Upsilon}\big)^{\frac{d+2\alpha}{2\alpha}} n^{\frac{\zeta(d+2\alpha)}{\alpha(d+2\zeta)}}\big)$, where the obtained rate is of the same order as that of the reweighted estimator in Theorem 4. This further illustrates the effectiveness of the combination of pre-training and reweighted techniques in our proposed method.

As the density ratio is unbounded, the following assumption is needed to establish the sharp rate for the truncated density ratio estimator.

Assumption 9. *There exists some constant $\delta > 0$ such that the density ratio r has a finite $(2+\delta)$-th moment with respect to $P_{\boldsymbol{X}}$, that is, $\Xi := \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^{2+\delta}(\boldsymbol{X})] < \infty$.*

Although Assumption 9 is slightly stronger than Assumption 5, it is necessary to provide a sharp bound for the difference between the true density ratio r and its truncated version $T_\xi r$ in terms of $L_2(P_{\boldsymbol{X}})$, denoted as $\|T_\xi r - r\|^2_{L_2(P_{\boldsymbol{X}})}$. More details regarding this issue are deferred to Appendix A.4. Now we are ready to establish the convergence rate of the truncated density ratio estimator under Assumption 9.

Theorem 14. *Suppose that Assumptions 1 and 9 are satisfied, and $T_\xi r \in \mathcal{H}^\alpha([0,1]^d, \xi)$ holds for some $\alpha > 0$. If the function space $\mathcal{U}$ is a ReLU FNN function class bounded by ξ, with the size $\mathcal{S} = \mathcal{O}\Big(m^{\frac{d}{d+(2+6/\delta)\alpha}}\Big)$, the depth $\mathcal{D} = \mathcal{O}(\log m)$, weights bound $\mathcal{B} = \mathcal{O}\Big(m^{\frac{\delta d+2\alpha}{\delta d+(6+2\delta)\alpha}}\Big)$ for the ReLU DNNs, and the truncation level $\xi = \mathcal{O}\Big(m^{\frac{2\alpha}{\delta d+(6+2\delta)\alpha}}\Big)$, then we have*

$$\mathbb{E}_{\mathbb{S}}\Big[\|\widehat{r}_{\xi,\mathbb{S}} - r\|^2_{L_2(P_{\boldsymbol{X}})}\Big] \leq \mathcal{O}\Big(m^{-\frac{2\alpha}{d+(2+6/\delta)\alpha}}(\log m)^2\Big).$$

Theorem 14 ensures that under some technical assumptions, the truncated density ratio estimator can achieve a sharp rate of convergence, even in the case of unbounded densities. Particularly, as $\delta \to \infty$, this convergence rate approaches the minimax optimal nonparametric rate of $\mathcal{O}(m^{-\frac{2\alpha}{d+2\alpha}})$.

Remark 15. *If we further strengthen Assumption 9 such that the square of the density ratio is sub-exponential with respect to $P_{\boldsymbol{X}}$, i.e., $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[\exp(\sigma r^2(\boldsymbol{X}))\big] < \infty$, where σ is some positive constant, the convergence rate of $\widehat{r}_{\xi,\mathbb{S}}$ in Theorem 14 can be improved to be minimax optimal, which corresponds to the case where $\delta \to \infty$. It is worth noting that under the nonparametric regression setting, the finite $(2+\delta)$-th moment condition and the sub-exponential tail condition usually lead to different non-asymptotic error bounds (Han and Wellner, 2019; Schmidt-Hieber, 2020; Farrell et al., 2021).*

Finally, we obtain the convergence rate for this pre-training reweighted estimator $\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}$ in the following theorem.

Theorem 16. *Suppose that Assumptions 1-3, 8 and 9 are satisfied. If the function space $\mathcal{M}$ is ReLU FNN function class bounded by $B \geq 1$, the size $\mathcal{S} = \mathcal{O}(n^{\frac{d}{d+(2+4/\delta)\zeta}})$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, weights bound $\mathcal{B} = \mathcal{O}(Bn^{\frac{d}{d+(2+4/\delta)\zeta}})$, and the truncation level $\xi = \mathcal{O}(n^{\frac{2\zeta}{\delta d+(4+2\delta)\zeta}})$, then we have*

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \lesssim B^2 n^{-\frac{2\zeta}{d+(2+4/\delta)\zeta}}(\log n)^2 + \frac{B^2 m^{-\frac{2\alpha}{d+(2+6/\delta)\alpha}}(\log m)^2}{\Upsilon}.$$

Moreover, if $m \geq \Omega\Big(n^{\frac{[\delta d+(6+2\delta)\alpha]\zeta}{[\delta d+(4+2\delta)\zeta]\alpha}}\Big)$, we obtain that

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2 n^{-\frac{2\zeta}{d+(2+4/\delta)\zeta}}(\log n)^2\Big).$$

Theorem 16 provides a non-asymptotic error bound for $\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}$, which approaches the minimax optimal rate in the field of nonparametric regression when $\delta \to \infty$. As shown in Corollary 11, some sharp upper bounds can be similarly obtained if the density ratio r is known previously under a weaker condition than Assumption 5 commonly used in the

literature. As discussed in Remark 15, if we further assume that r^2 is sub-exponential, these bounds can achieve the minimax optimal rate. Notably, this result is particularly appealing as it is established by considering the pre-training density ratio estimator. To the best of our knowledge, Theorem 16 is the first result for investigation on the problem of covariate shift in the context of nonparametric quantile regression.

5. Numerical Experiments

In this section, we evaluate the numerical performance of the pre-training reweighted estimators (*PWDQR*) in (9) and (11) for both bounded and unbounded ratios. We compare it to some state-of-the-art methods, including the reweighted estimator (*WDQR*) as defined in (6) and 12 for the bounded and unbounded ratios, respectively, and the unweighted estimator (*DQR*) as defined in (5). The implementation details of all the considered methods are provided below.

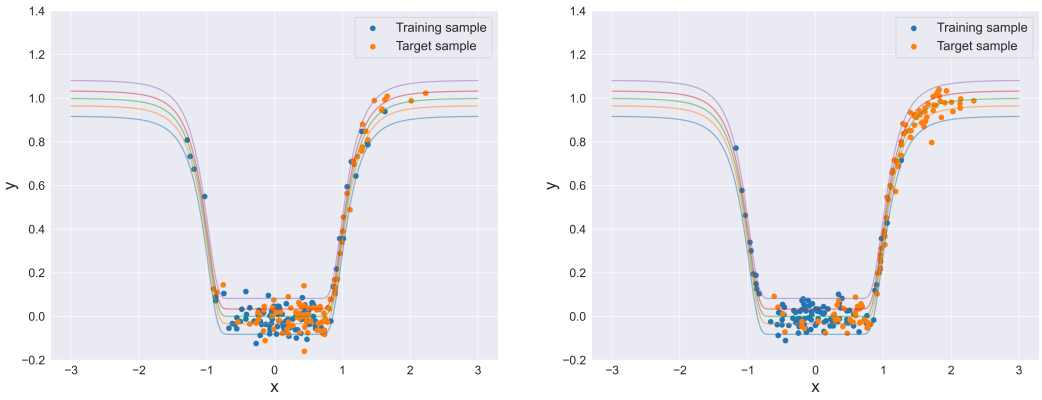
- DQR: we implement it in *Pytorch* using the stochastic gradient descent (*SGD*) (Bottou, 2012) with Nesterov momentum of 0.9 and initial learning rate of 0.1 with rate decay 0.5. We consider the fixed width neural network consisting of ReLU activated multilayer perceptrons with three hidden layers.

- WDQR: the implementation details of WDQR are almost identical to those of DQR, except for using the weighted quantile loss function instead of the unweighted one, where the true density ratio r is plugged in.

- PWDQR: the implementation details of PWDQR is exactly the same as that of WDQR, except that we replace the true density ratio with the (truncated) pre-trained density ratio. Specifically, it is similar to WDQR with a pre-trained density ratio $\widehat{r}_{\mathbb{S}}$ instead of the given truth. For the estimation of $\widehat{r}_{\mathbb{S}}$, we solve (8) by a neural network using Pytorch, which consists of ReLU activated multilayer perceptrons with two hidden layers. The optimization algorithm is *Adam* (Kingma and Ba, 2017) with a learning rate 10^{-4}. For the truncated pre-trained density ratio $\widehat{r}_{\xi,\mathbb{S}}$, we additionally use an activate function $\sigma'_{\xi}(x) = \min(x, \xi)$ for the last layer of the neural network.

We consider two generating scenarios including the univariate and multivariate cases, and two covariate shift settings with bounded and unbounded density ratios, respectively. In this simulation study, we only use the truncated ratio for the unbounded case in WDQR and PWDQR, and the truncated levels ξ for WDQR and PWDQR are suggested to be $c_1(n/\log n)^{1/3}$ and $c_2 \log m$ for some constants $c_1, c_2 > 0$ as indicated in the theorems, respectively.

For each simulated scenario, we generate the training data $\{\boldsymbol{X}_i^{tr}, Y_i^{tr}\}_{i=1}^{n_{tr}}$ with sample size n_{tr} from the source distribution to train those three nonparametric quantile regression models at five quantile levels $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. To evaluate each model, we generate the target data $\{\boldsymbol{X}_i^{ta}, Y_i^{ta}\}_{i=1}^{n_{ta}}$ with sample size n_{ta} from the target distribution. For notation simplicity, we denote $\widehat{f}_{n_{tr}}^{\tau}$ and f_0^{τ} as the estimated and true quantile functions at the specific quantile level $\tau \in (0, 1)$, respectively. We evaluate the performance of these



(a) Bounded case (b) Unbounded case

Figure 1: The quantile curves of model (13) for the bounded case (a) and unbounded case (b), where the training and target data are represented as blue and orange dots, respectively, and the target quantile functions at levels $\tau = 0.05$ (blue line), $\tau = 0.25$ (orange line), $\tau = 0.5$ (green line), $\tau = 0.75$ (red line), $\tau = 0.95$ (purple line) are plotted as curves.

methods based on two norms between $\widehat{f}^{\tau}_{n_{tr}}$ and f^{τ}_0 as given by

$$L_1: \ \|\widehat{f}^{\tau}_{n_{tr}} - f^{\tau}_0\|_{L^1(\nu)} = \frac{1}{n_{ta}} \sum_{i=1}^{n_{ta}} \left| \widehat{f}^{\tau}_{n_{tr}}(\boldsymbol{X}^{ta}_i) - f^{\tau}_0(\boldsymbol{X}^{ta}_i) \right|,$$

and

$$L_2: \ \|\widehat{f}^{\tau}_{n_{tr}} - f^{\tau}_0\|_{L^2(\nu)} = \left\{ \frac{1}{n_{ta}} \sum_{i=1}^{n_{ta}} \left(\widehat{f}^{\tau}_{n_{tr}}(\boldsymbol{X}^{ta}_i) - f^{\tau}_0(\boldsymbol{X}^{ta}_i) \right)^2 \right\}^{1/2}.$$

To estimate the pre-training density ratio, we also independently generate extra training data $\{\widetilde{\boldsymbol{X}}^{tr}_i, \widetilde{Y}^{tr}_i\}_{i=1}^m$ and target data $\{\widetilde{\boldsymbol{X}}^{ta}_i, \widetilde{Y}^{ta}_i\}_{i=1}^m$ with the same sample size m. In our study, we fix $n_{ta} = 10000$ and $m = 1000$, and we report the averaged L_1 and the square of L_2 distances together with their corresponding standard errors over 100 independent repetitions under different scenarios.

5.1 Univariate Model

We generate the data from the following univariate model

$$Y = e^{-\frac{1}{X^6}} + \sigma\varepsilon, \tag{13}$$

where $\varepsilon \sim N(0,1)$ and $\sigma = 0.05$. Here, the true quantile function at quantile level τ is $f^{\tau}_0(X) = e^{-\frac{1}{X^6}} + \sigma\Phi^{-1}(\tau)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random error ε.

The source and target covariates are drawn from the normal distributions with mean μ_1 and variance σ_1^2, and mean μ_2 and variance σ_2^2, respectively. According to Cortes et al.

Table 1: Averaged L_1 and L_2^2 errors ($\times 10^{-1}$) based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for Model (13) with the bounded density ratio.

Sample size		$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	0.545(0.219)	0.086(0.126)	0.264(0.065)	0.021(0.009)
	WDQR	0.541(0.332)	0.084(0.118)	0.259(0.107)	0.020(0.019)
	PWDQR	0.555(0.243)	0.087(0.110)	0.270(0.088)	0.025(0.044)
0.25	DQR	0.249(0.108)	0.023(0.027)	0.140(0.035)	0.006(0.003)
	WDQR	0.249(0.088)	0.022(0.016)	0.138(0.049)	0.006(0.004)
	PWDQR	0.255(0.173)	0.031(0.091)	0.146(0.012)	0.007(0.004)
0.5	DQR	0.213(0.081)	0.019(0.012)	0.129(0.028)	0.004(0.002)
	WDQR	0.211(0.104)	0.017(0.014)	0.128(0.039)	0.004(0.003)
	PWDQR	0.224(0.179)	0.025(0.037)	0.131(0.039)	0.005(0.003)
0.75	DQR	0.245(0.084)	0.023(0.012)	0.145(0.036)	0.004(0.002)
	WDQR	0.241(0.179)	0.021(0.034)	0.145(0.049)	0.005(0.003)
	PWDQR	0.269(0.288)	0.029(0.067)	0.147(0.052)	0.006(0.004)
0.95	DQR	0.634(0.354)	0.106(0.321)	0.275(0.054)	0.014(0.005)
	WDQR	0.627(0.626)	0.103(0.423)	0.274(0.088)	0.015(0.007)
	PWDQR	0.684(0.697)	0.126(0.589)	0.276(0.094)	0.016(0.009)

(2010) and Feng et al. (2023), we can show that the density ratio $r(\cdot)$ is uniformly bounded if and only if $\sigma_1^2 \geq \sigma_2^2$, and second moment bounded if and only if $\sigma_2^2 \geq \sigma_1^2 \geq \sigma_2^2/2$. Consequently, we choose $\mu_1 = 0, \sigma_1^2 = 0.4, \mu_2 = 0.5, \sigma_2^2 = 0.3$ for the uniform bounded case and $\mu_1 = 0, \sigma_1^2 = 0.3, \mu_2 = 1, \sigma_2^2 = 0.5$ for the second moment bounded case, respectively. Figure 1 shows the univariate data generation model in the bounded case and unbounded case and their corresponding conditional quantile curves at $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$. Performance of different estimators based on L_1 and L_2^2 of prediction errors are summarized in Tables 1-2 for those cases with bounded and unbounded ratios, respectively.

From the results in Table 1, we observe that the prediction errors of the estimator DQR are very close to those of the estimator WDQR in all scenarios, aligning with our theoretical findings that the weighted and unweighted estimators can both achieve the minimax optimal rate for the uniformly bounded case. As the sample size n increases, the performance of PWDQR tends to coincide with that of WDQR, which is expected since the pre-trained weight function converges to its true counterpart when n is large. For the case with unbounded ratios, as shown in Table 2, both WDQR and PWDQR significantly outperform DQR in all scenarios. This phenomenon is consistent with our theoretical findings. Specifically, the unweighted estimator is sub-optimal when the importance ratio is second moment bounded. In contrast, both weighted and pre-trained weighted estimators can achieve the optimal rates. These results further validate the necessity and effectiveness of the reweighted procedure and our pre-training algorithm under covariate shift.

5.2 Multivariate Model

In this section, we consider the following additive multivariate model

$$Y = \sin(2\pi X_1) + 0.5e^{X_2} + 1.5|(X_3 - 0.4)(X_3 - 0.6)| + \sigma X_2 \varepsilon, \tag{14}$$

Table 2: Averaged L_1 and L_2^2 errors ($\times 10^{-1}$) based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for Model (13) with the unbounded density ratio.

Sample size		$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	1.420(0.738)	0.610(0.689)	0.678(0.634)	0.185(0.062)
	WDQR	1.192(0.561)	0.369(0.450)	0.618(0.237)	0.100(0.107)
	PWDQR	1.276(0.624)	0.377(0.482)	0.620(0.198)	0.103(0.064)
0.25	DQR	1.018(0.446)	0.443(0.374)	0.475(0.185)	0.096(0.085)
	WDQR	0.731(0.489)	0.237(0.428)	0.408(0.138)	0.066(0.061)
	PWDQR	0.809(0.321)	0.299(0.455)	0.428(0.157)	0.068(0.061)
0.5	DQR	0.912(0.396)	0.427(0.361)	0.402(0.185)	0.095(0.112)
	WDQR	0.779(0.505)	0.298(0.464)	0.374(0.143)	0.066(0.064)
	PWDQR	0.783(0.336)	0.306(0.445)	0.396(0.162)	0.067(0.065)
0.75	DQR	0.928(0.398)	0.455(0.384)	0.520(0.207)	0.101(0.123)
	WDQR	0.762(0.534)	0.330(0.499)	0.410(0.161)	0.072(0.049)
	PWDQR	0.798(0.405)	0.345(0.460)	0.443(0.180)	0.073(0.070)
0.95	DQR	1.527(0.591)	0.672(0.542)	0.714(0.208)	0.154(0.077)
	WDQR	1.389(0.814)	0.501(0.599)	0.594(0.283)	0.116(0.096)
	PWDQR	1.427(0.545)	0.479(0.542)	0.634(0.243)	0.132(0.162)

where $\varepsilon \sim t(3)$ and $\sigma = 0.1$. Here, the true quantile function at quantile level τ is $f_0^\tau(\boldsymbol{X}) = \sin(2\pi X_1) + 0.5e^{X_2} + 1.5|(X_3 - 0.4)(X_3 - 0.6)| + \sigma X_2 F_t^{-1}(\tau, 3)$, where $F_t^{-1}(\cdot, 3)$ is the cumulative distribution function of the Student's t random error ε with degrees 3.

We assume that three covariates X_1, X_2 and X_3 are independent, and X_2, X_3 are generated from the uniform distribution on $[0, 1]$ for both source and target distributions. The source and target data of X_1 are drawn from Beta distributions with parameters (α_1, β_1) and (α_2, β_2), respectively. It is easy to verify that the importance ratio $r(\boldsymbol{X})$ is uniformly bounded if and only if $\alpha_2 \geq \alpha_1$ and $\beta_2 \geq \beta_1$, and second moment bounded if and only if $\alpha_2 < \alpha_1, 2\alpha_2 \geq \alpha_1, 2\beta_2 \geq \beta_1$ or $\beta_2 < \beta_1, 2\alpha_2 \geq \alpha_1, 2\beta_2 \geq \beta_1$. In our study, we choose $\alpha_1 = 2.5, \beta_1 = 1.5, \alpha_2 = 3, \beta_2 = 4$ for the uniformly bounded case and $\alpha_1 = 4, \beta_1 = 1, \alpha_2 = 3, \beta_2 = 6$ for the second moment bounded case, respectively. Results on the performance of different estimators are summarized in Tables 3-4 for those cases with the bounded and unbounded ratios, respectively.

As indicated in Tables 3-4, the conclusions for the multivariate model are very similar to those of the univariate model. The errors of the three estimators are very close for the uniformly bounded case, and both WDQR and PWDQR have a better performance than DQR for the second moment bounded case.

6. Conclusion and Discussion

In this work, we leverage deep nonparametric quantile regression as the foundational framework to systematically explore and illuminate the phenomenon of covariate shift within quantile regression. We propose a two-stage method, leading to the development of a pre-training reweighted estimator for the target quantile function. In our theoretical analysis, we present rigorous non-asymptotic error bounds for unweighted, reweighted, and pre-training reweighted estimators. This analysis simultaneously considers scenarios in which the den-

Table 3: Averaged L_1 and L_2^2 errors based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for model (14) under the case with bounded density ratio.

	Sample size	$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	0.348(0.064)	0.156(0.057)	0.215(0.031)	0.059(0.017)
	WDQR	0.348(0.074)	0.154(0.062)	0.213(0.023)	0.058(0.011)
	PWDQR	0.350(0.069)	0.158(0.063)	0.219(0.024)	0.063(0.012)
0.25	DQR	0.149(0.031)	0.038(0.017)	0.103(0.017)	0.020(0.005)
	WDQR	0.150(0.054)	0.037(0.015)	0.102(0.012)	0.019(0.004)
	PWDQR	0.151(0.030)	0.041(0.015)	0.104(0.013)	0.021(0.004)
0.5	DQR	0.114(0.013)	0.028(0.006)	0.078(0.007)	0.014(0.003)
	WDQR	0.112(0.017)	0.027(0.007)	0.076(0.005)	0.013(0.002)
	PWDQR	0.120(0.015)	0.027(0.006)	0.077(0.005)	0.014(0.003)
0.75	DQR	0.145(0.028)	0.038(0.012)	0.102(0.016)	0.019(0.005)
	WDQR	0.144(0.032)	0.036(0.010)	0.101(0.013)	0.019(0.004)
	PWDQR	0.149(0.025)	0.039(0.010)	0.103(0.012)	0.019(0.004)
0.95	DQR	0.320(0.052)	0.141(0.025)	0.215(0.027)	0.059(0.007)
	WDQR	0.318(0.060)	0.139(0.050)	0.213(0.023)	0.059(0.010)
	PWDQR	0.321(0.054)	0.145(0.054)	0.214(0.024)	0.060(0.010)

Table 4: Averaged L_1 and L_2^2 errors based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for model (14) under the case with unbounded density ratio.

	Sample size	$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	0.390(0.093)	0.217(0.098)	0.290(0.037)	0.124(0.029)
	WDQR	0.345(0.165)	0.187(0.242)	0.265(0.038)	0.094(0.024)
	PWDQR	0.363(0.362)	0.195(0.262)	0.270(0.049)	0.099(0.045)
0.25	DQR	0.267(0.043)	0.143(0.055)	0.216(0.037)	0.108(0.044)
	WDQR	0.246(0.064)	0.109(0.060)	0.161(0.020)	0.051(0.014)
	PWDQR	0.254(0.134)	0.118(0.152)	0.165(0.027)	0.053(0.018)
0.5	DQR	0.245(0.050)	0.147(0.072)	0.209(0.047)	0.122(0.057)
	WDQR	0.209(0.047)	0.095(0.040)	0.146(0.021)	0.054(0.019)
	PWDQR	0.219(0.071)	0.100(0.064)	0.150(0.029)	0.055(0.023)
0.75	DQR	0.272(0.070)	0.174(0.089)	0.239(0.063)	0.150(0.070)
	WDQR	0.254(0.053)	0.129(0.055)	0.173(0.034)	0.069(0.025)
	PWDQR	0.263(0.072)	0.132(0.075)	0.182(0.039)	0.072(0.031)
0.95	DQR	0.426(0.107)	0.302(0.138)	0.369(0.083)	0.243(0.101)
	WDQR	0.386(0.127)	0.274(0.167)	0.295(0.053)	0.135(0.043)
	PWDQR	0.397(0.136)	0.285(0.178)	0.308(0.057)	0.142(0.054)

sity ratio is either bounded or unbounded under some weaker moment conditions than those commonly considered in the existing literature. Importantly, we introduce a novel approach to constrain generalization error, simplify the tools of local Rademacher complexities, and derive an approximation error with weights bound for ReLU neural networks approximating the Hölder continuous function class. Furthermore, our theoretical insights provide valuable prior guidance in the selection of an appropriate sample size for the pre-training strategy. These results underscore the importance of the pre-training and reweighting techniques in mitigating the challenges posed by the covariate shift phenomenon. Several directions for future work are worth exploring. One possible future work is to further investigate the impact of lower and upper bounds of the density ratio on convergence rates. Most recently, generative adversarial networks and diffusion models have both been proven effective in generating high-quality samples from complex distributions, which may provide promising alternatives. Thus, another possible further work to integrate generative models within the proposed framework is to adopt these methods to model the conditional distribution directly, or use the generative model as a pre-processing step.

Acknowledgments

We would like to express our sincere gratitude to the editor, the action editor, and two anonymous reviewers for their constructive comments, which have greatly contributed to the significant improvement of the manuscript. This research was partly supported by the National Natural Science Foundation of China grant (12371441, 12371270, 12001356, U24A2002), Shanghai Science and Technology Development Funds (23JC1402100), Natural Science Foundation of Shanghai (24ZR1421400), Shanghai Research Center for Data Science and Decision Technology, and the Fundamental Research Funds for the Central Universities.

Appendix A. Proof of the Main Results

In Section A.1, we first prove the error decomposition including approximation error and statistical error and provide the tight estimates of each error bound in Sections A.1.1 and A.1.2, respectively. Based on the error decomposition, in Section A.2, we provide the proofs of the non-asymptotic error bounds for the unweighted estimator under the uniformly bounded case and second moment bounded case in Sections A.2.1 and A.2.2, respectively. In Section A.3, with a slightly modified error decomposition in Lemma 5 and estimates for the statistical error term in Lemma 6, we succeed in deriving the non-asymptotic error bounds for the reweighted estimator under the uniformly bounded case and second moment bounded case in Sections A.3.1 and A.3.2, respectively. In Section A.4, we provide the proofs of the non-asymptotic error bounds for the pre-training density ratio and the pre-training reweighted estimator under the uniformly bounded case and exponential moment bounded case in Sections A.4.1 and A.4.2, respectively.

A.1 Proof of the Error Terms

In this section, we give the proof of Lemma 1, which indicates the error decomposition of the $L_2(P_{\boldsymbol{X}})$ norm of the difference between the true quantile function f_0 and its estimates $\widehat{f}_{\mathbb{D}}$.

Proof. Recall the Knight's identity (Belloni and Chernozhukov, 2011) that for any $u, v \in \mathbb{R}$, there holds

$$\rho_\tau(u-v)-\rho_\tau(u)=-v(\tau-\mathbf{I}\{u\le 0\})+\int_0^v(\mathbf{I}\{u\le z\}-\mathbf{I}\{u\le 0\})dz.$$

Then, for any $f\in\mathcal{G}$, by taking $u=Y-f^*(\boldsymbol{X})$ and $v=f(\boldsymbol{X})-f^*(\boldsymbol{X})$, we have

$$\begin{aligned}
&\rho_\tau\big(Y-f(\boldsymbol{X})\big)-\rho_\tau\big(Y-f^*(\boldsymbol{X})\big)=-\big(f(\boldsymbol{X})-f^*(\boldsymbol{X})\big)\big(\tau-\mathbf{I}\{Y\le f^*(\boldsymbol{X})\}\big)\\
&+\int_0^{f(\boldsymbol{X})-f^*(\boldsymbol{X})}\Big[\mathbf{I}\big\{Y\le f^*(\boldsymbol{X})+z\big\}-\mathbf{I}\big\{Y\le f^*(\boldsymbol{X})\big\}\Big]dz\\
&=-\big(f(\boldsymbol{X})-f^*(\boldsymbol{X})\big)\big(\tau-\mathbf{I}\{Y\le f_0(\boldsymbol{X})\}\big)-\big(f(\boldsymbol{X})-f^*(\boldsymbol{X})\big)\big(\mathbf{I}\{Y\le f_0(\boldsymbol{X})\}\\
&\quad-\mathbf{I}\{Y\le f^*(\boldsymbol{X})\}\big)+\int_0^{f(\boldsymbol{X})-f^*(\boldsymbol{X})}\big[\mathbf{I}\{Y\le f^*(\boldsymbol{X})+z\}-\mathbf{I}\{Y\le f^*(\boldsymbol{X})\}\big]dz.
\end{aligned}$$

By taking expectations, we note that $\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}[(\tau - \mathbf{I}\{Y \le f_0(\boldsymbol{X})\}) \mid \boldsymbol{X}] = 0$ due to the fact that $\mathbb{P}(Y \le f_0(\boldsymbol{X}) \mid \boldsymbol{X}) = \tau$, and using Fubini's theorem, there holds

$$
\begin{aligned}
&\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big] \\
=&\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[- \big(f(\boldsymbol{X}) - f^*(\boldsymbol{X})\big)\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}\big[(\tau - \mathbf{I}\{Y \le f_0(\boldsymbol{X})\}) \mid \boldsymbol{X}\big]\Big] \\
&- \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[\big(f(\boldsymbol{X}) - f^*(\boldsymbol{X})\big)\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}\big[(\mathbf{I}\{Y \le f_0(\boldsymbol{X})\} - \mathbf{I}\{Y \le f^*(\boldsymbol{X})\}) \mid \boldsymbol{X}\big]\Big] \\
&+ \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[\int_0^{f(\boldsymbol{X})-f^*(\boldsymbol{X})} \big(\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}[\mathbf{I}\{Y \le f^*(\boldsymbol{X}) + z\} \mid \boldsymbol{X}] \\
&- \mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}[\mathbf{I}\{Y \le f^*(\boldsymbol{X})\} \mid \boldsymbol{X}]\big)dz\Big] \\
\ge& - C_1\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f(\boldsymbol{X}) - f^*(\boldsymbol{X})||f_0(\boldsymbol{X}) - f^*(\boldsymbol{X})|\big] + C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[D^2(f(\boldsymbol{X}) - f^*(\boldsymbol{X}))\big] \\
\ge& - C_1\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big]}\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f_0(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big]} \\
&+ C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[D^2(f(\boldsymbol{X}) - f^*(\boldsymbol{X}))\big],
\end{aligned}
$$

where C_1, C_2 are two absolute positive constants and $D^2(t) := \min\{|t|, |t|^2\}$, $t \in \mathbb{R}$, the first inequality follows from Assumptions 2 and 3 and Lemma 13 in Madrid Padilla and Chatterjee (2022), and the last inequality follows from the Cauchy-Schwarz inequality. Consequently, for any $\beta > 0$, there holds

$$
\begin{aligned}
&C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[D^2(f(\boldsymbol{X}) - f^*(\boldsymbol{X}))\big] \le \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big] \\
&+ C_1\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big]}\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f_0(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big]} \\
\le&\frac{C_1}{4\beta}\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big] + C_1\beta\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[|f_0(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\big] \\
&+ \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big].
\end{aligned}
$$

Note that $f^* \in \mathcal{G}$, we have $\|f^*\|_\infty \le B$ with $B \ge 1$, then for any $\|f\|_\infty \le B$, we have

$$
\begin{aligned}
D^2((f(\boldsymbol{X}) - f^*(\boldsymbol{X})) &= \min\{|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|, |f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2\} \\
&\ge 2B\min\{\frac{|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|}{2B}, \frac{|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2}{4B^2}\} \\
&= \frac{|f(\boldsymbol{X}) - f^*(\boldsymbol{X})|^2}{2B},
\end{aligned}
$$

where the last equality follows from $|f(\boldsymbol{X}) - f^*(\boldsymbol{X})| \le \|f\|_\infty + \|f^*\|_\infty \le 2B$.

By setting $\beta = \frac{C_1 B}{C_2}$, there holds

$$
\begin{aligned}
\|f - f^*\|_{L_2(P_{\boldsymbol{X}})}^2 \le& \frac{4B}{C_2}\Big(\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big]\Big) \\
&+ \frac{2C_1^2B^2}{C_2^2}\|f_0 - f^*\|_{L_2(P_{\boldsymbol{X}})}^2.
\end{aligned}
$$

Note that $\widehat{f}_{\mathbb{D}} \in \mathcal{G}$, then $\|\widehat{f}_{\mathbb{D}}\|_\infty \leq B$, it follows directly by the triangle inequality that

$$\begin{aligned}
&\big\|\widehat{f}_{\mathbb{D}} - f_0\big\|^2_{L_2(P_{\boldsymbol{X}})} \leq 2\big\|\widehat{f}_{\mathbb{D}} - f^*\big\|^2_{L_2(P_{\boldsymbol{X}})} + 2\big\|f^* - f_0\big\|^2_{L_2(P_{\boldsymbol{X}})} \\
\lesssim& B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big] + B^2\|f^* - f_0\|^2_{L_2(P_{\boldsymbol{X}})} \\
\lesssim& B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\big] + B^2\|f^* - f_0\|^2_{\infty}.
\end{aligned}$$

This completes the proof of Lemma 1. ■

A.1.1 Proof of Estimates for the Approximation Error

We give the proof of the upper bound for the approximation error as stated in Theorem 4. This proof aligns with the methodology employed in previous works (Yarotsky, 2017; Petersen and Voigtlaender, 2018). Diverging from the approach in Yarotsky (2017), our contribution lies in deriving bounds for the weights through the techniques introduced in Petersen and Voigtlaender (2018). To start with, we first introduce two preliminary lemmas.

Lemma 2. *Given $M > 0$ and $\delta \in (0,1)$, there exists a ReLU neural network h satisfies:*

(i). for all $x, y \in [-M, M]$, we have $|xy - h(x,y)| \leq \delta$;

(ii). if $x = 0$ or $y = 0$, then $h(x,y) = 0$;

(iii). the depth $\mathcal{D}$ and the size $\mathcal{S}$ in h are less than $c_1 \ln(\frac{1}{\delta}) + c_2$, and weights bound $\mathcal{B}$ is not larger than $\frac{c_3}{\delta}$, where c_1 is an absolute constant, and c_2, c_3 are two constants depending on M.

The proof of Lemma 2 can be completed by combining Proposition 3 in Yarotsky (2017) and Lemma A.3 in Petersen and Voigtlaender (2018), and thus is omitted here.

Lemma 3. *Let $f_1 \in \mathcal{F}_{d_1,k_1}(\mathcal{W}_1, \mathcal{D}_1, \mathcal{S}_1, \mathcal{B}_1)$ and $f_2 \in \mathcal{F}_{d_2,k_2}(\mathcal{W}_1, \mathcal{D}_1, \mathcal{S}_1, \mathcal{B}_1)$, and then the following statements hold.*

(i). **(Composition)** *If $k_1 = d_2$, then $f_2 \circ f_1 \in \mathcal{F}_{d_1,k_2}(\max\{\mathcal{W}_1, \mathcal{W}_2\}, \mathcal{D}_1 + \mathcal{D}_2, \mathcal{S}_1 + \mathcal{S}_2, \mathcal{B}_1 \cdot \mathcal{B}_2 \max\{\mathcal{W}_1, \mathcal{W}_2\})$. Moreover, if $A \in \mathbb{R}^{d_2 \times d_1}$, $b \in \mathbb{R}^{d_2}$ and define the function $f(\boldsymbol{x}) = f_2(A\boldsymbol{x} + b)$ for $\boldsymbol{x} \in \mathbb{R}^{d_1}$, then there holds $f \in \mathcal{F}_{d_1,k_2}(\mathcal{W}_2, \mathcal{D}_2, \mathcal{S}_2, d_2\mathcal{B}_2\|(A,b)\|_\infty)$.*

(ii). **(Parallelization)** *If $d_1 = d_2$, denote $f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$, then there holds $f \in \mathcal{F}_{d_1,k_1+k_2}(\mathcal{W}_1 + \mathcal{W}_2, \max\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{S}_1 + \mathcal{S}_2, \max\{\mathcal{B}_1, \mathcal{B}_2\})$.*

(iii). **(Linear Combination)** *Let $c_1, c_2 \in \mathbb{R}$. If $d_1 = d_2$ and $k_1 = k_2$, then we have $c_1 f_1 + c_2 f_2 \in \mathcal{F}_{d_1,k_1}(\mathcal{W}_1 + \mathcal{W}_2, \max\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{S}_1 + \mathcal{S}_2, \max\{\mathcal{B}_1, \mathcal{B}_2, |c_1|\mathcal{B}_1 + |c_2|\mathcal{B}_2\})$.*

Proof. To start with, we first denote f_i as ReLU neural networks with parameters

$$\theta_i = \big((A_0^{(i)}, b_0^{(i)}), \ldots, (A_{\mathcal{D}_i}^{(i)}, b_{\mathcal{D}_i}^{(i)})\big), \text{ fori} = 1, 2.$$

For (i), without loss of generality, we assume that $\mathcal{W}_1 = \mathcal{W}_2$ and then, $f_2 \circ f_1$ can be parameterized by

$$\big((A_0^{(1)}, b_0^{(1)}), \ldots, (A_{\mathcal{D}_1-1}^{(1)}, b_{\mathcal{D}_1-1}^{(1)}), (A_0^{(2)}A_{\mathcal{D}_1}^{(1)}, A_0^{(2)}b_{\mathcal{D}_1}^{(1)} + b_0^{(2)}), (A_1^{(2)}, b_1^{(2)}), \ldots, (A_{\mathcal{D}_2}^{(2)}, b_{\mathcal{D}_2}^{(2)})\big).$$

Note that it holds true that

$$\begin{aligned}\left\|(A_0^{(2)}A_{\mathcal{D}_1}^{(1)}, A_0^{(2)}b_{\mathcal{D}_1}^{(1)} + b_0^{(2)})\right\|_\infty &\le \max\left\{\left\|(A_0^{(2)}A_{\mathcal{D}_1}^{(1)})\right\|_\infty, \left\|(A_0^{(2)}b_{\mathcal{D}_1}^{(1)} + b_0^{(2)})\right\|_\infty\right\}\\ &\le \mathcal{B}_1\mathcal{B}_2\mathcal{W}_1.\end{aligned}$$

Then, we have $f_2 \circ f_1 \in \mathcal{F}_{d_1,k_2}\big(\max\{\mathcal{W}_1,\mathcal{W}_2\}, \mathcal{D}_1+\mathcal{D}_2, \mathcal{S}_1+\mathcal{S}_2, \mathcal{B}_1\cdot\mathcal{B}_2\max\{\mathcal{W}_1,\mathcal{W}_2\}\big)$. Following a similar treatment, we can conclude that $f \in \mathcal{F}_{d_1,k_2}(\mathcal{W}_2, \mathcal{D}_2, \mathcal{S}_2, d_2\mathcal{B}_2\|(A,b)\|_\infty)$ for the function $f(\boldsymbol{x}) = f_2(A\boldsymbol{x}+b)$ since it is a composition of f_2 with $f_1(\boldsymbol{x}) = A\boldsymbol{x}+b$, which can be regarded as a neural network with depth zero.

For (ii), without loss of generality, we assume that $\mathcal{D}_1 = \mathcal{D}_2$ and then, f can be parameterized by the parameters $((A_0,b_0),\ldots,(A_{\mathcal{D}_1},b_{\mathcal{D}_1}))$ with

$$A_\ell = \begin{pmatrix} A_\ell^{(1)} & \mathbf{0} \\ \mathbf{0} & A_\ell^{(2)}\end{pmatrix} \text{ and } b_\ell = \begin{pmatrix} b_\ell^{(1)} \\ b_\ell^{(2)}\end{pmatrix}.$$

Then, the derived result directly follows from

$$\|(A_\ell,b_\ell)\|_\infty = \left\|\begin{pmatrix} A_\ell^{(1)} & \mathbf{0} & b_\ell^{(1)} \\ \mathbf{0} & A_\ell^{(2)} & b_\ell^{(2)}\end{pmatrix}\right\|_\infty = \max\left\{\|(A_\ell^{(1)},b_\ell^{(1)})\|_\infty, \|(A_\ell^{(2)},b_\ell^{(2)})\|_\infty\right\}.$$

For (iii), directly replacing the matrix $(A_{\mathcal{D}_1}, b_{\mathcal{D}_1})$ in (ii) with $(c_1A_{\mathcal{D}_1}^{(1)}, c_2A_{\mathcal{D}_2}^{(2)}, c_1b_{\mathcal{D}_1}^{(1)} + c_2b_{\mathcal{D}_2}^{(2)})$, the derived result directly follows from

$$\begin{aligned}\left\|(c_1A_{\mathcal{D}_1}^{(1)}, c_2A_{\mathcal{D}_1}^{(2)}, c_1b_{\mathcal{D}_1}^{(1)} + c_2b_{\mathcal{D}_2}^{(2)})\right\|_\infty &\le |c_1|\left\|(A_{\mathcal{D}_1}^{(1)}, b_{\mathcal{D}_1}^{(1)})\right\|_\infty + |c_2|\left\|(A_{\mathcal{D}_1}^{(2)}, b_{\mathcal{D}_1}^{(2)})\right\|_\infty\\ &\le |c_1|\mathcal{B}_1 + |c_2|\mathcal{B}_2.\end{aligned}$$

Thus, we have

$$c_1f_1 + c_2f_2 \in \mathcal{F}_{d_1,k_1}(\mathcal{W}_1+\mathcal{W}_2, \max\{\mathcal{D}_1,\mathcal{D}_2\}, \mathcal{S}_1+\mathcal{S}_2, \max\{\mathcal{B}_1,\mathcal{B}_2,|c_1|\mathcal{B}_1+|c_2|\mathcal{B}_2\}).$$

This completes the proof of Lemma 3. ■

Combining Lemmas 2 and 3, we are now ready to complete the proof of Theorem 4.

Proof. We denote the function $\psi(\cdot)$ by

$$\psi(t) := \sigma(1-|t|) = \sigma(1-\sigma(t)-\sigma(-t)) \in [0,1], \quad t \in \mathbb{R}.$$

Notice that ψ is a two-layer neural network contained in $\mathcal{F}(2,2,6,1)$. Let $N \in \mathbb{N}$, for any $\mathbf{n} = (n_1,\ldots,n_d)^\top \in \{0,1,\ldots,N\}^d$, we define

$$\psi_{\mathbf{n}}(\boldsymbol{x}) := \prod_{i=1}^d \psi\left(Nx_i - n_i\right), \quad \boldsymbol{x} = (x_1,\ldots,x_d)^\top \in \mathbb{R}^d.$$

Then, $\psi_{\mathbf{n}}$ is supported on $\left\{\boldsymbol{x}\in\mathbb{R}^d : \left\|\boldsymbol{x} - \frac{\mathbf{n}}{N}\right\|_\infty \le \frac{1}{N}\right\}$, and $(N+1)^d$ functions $\{\psi_{\mathbf{n}}\}_{\mathbf{n}}$ form a partition of unity of the domain $[0,1]^d$, i.e.,

$$\sum_{\mathbf{n}\in\{0,1,\ldots,N\}^d} \psi_{\mathbf{n}}(\boldsymbol{x}) = \prod_{i=1}^d \sum_{n_i=0}^N \psi\left(Nx_i - n_i\right) \equiv 1, \quad \boldsymbol{x} \in [0,1]^d.$$

Let $c_{\mathbf{n},\mathbf{s}} := \partial^{\mathbf{s}} f\left(\frac{\mathbf{n}}{N}\right)/\mathbf{s}!$ be the Taylor coefficients of f at $\frac{\mathbf{n}}{N}$, where $\mathbf{s} := (s_1, \ldots, s_d) \in \mathbb{N}^d$ with $\|\mathbf{s}\|_1 \leq t = \lfloor \zeta \rfloor$. We denote by $p_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}) := \psi_{\mathbf{n}}(\boldsymbol{x})\left(\boldsymbol{x} - \frac{\mathbf{n}}{N}\right)^{\mathbf{s}}$. Then, it follows that $p_{\mathbf{n},\mathbf{s}}$ is supported on $\left\{\boldsymbol{x} \in \mathbb{R}^d : \left\|\boldsymbol{x} - \frac{\mathbf{n}}{N}\right\|_\infty \leq \frac{1}{N}\right\}$. Denote by

$$p(\boldsymbol{x}) = \sum_{\mathbf{n} \in \{0,1,\ldots,N\}^d} \sum_{\|\mathbf{s}\|_1 \leq t} c_{\mathbf{n},\mathbf{s}} p_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}).$$

Using Taylor expansion in Lemma A.8 in Petersen and Voigtlaender (2018), we have

$$\begin{aligned}
|f(\boldsymbol{x}) - p(\boldsymbol{x})| &= \left| \sum_{\mathbf{n}} \psi_{\mathbf{n}}(\boldsymbol{x}) f(\boldsymbol{x}) - \sum_{\mathbf{n}} \psi_{\mathbf{n}}(\boldsymbol{x}) \sum_{\|\mathbf{s}\|_1 \leq t} c_{\mathbf{n},\mathbf{s}} \left(\boldsymbol{x} - \frac{\mathbf{n}}{N}\right)^{\mathbf{s}} \right| \\
&\leq \sum_{\mathbf{n}} \psi_{\mathbf{n}}(\boldsymbol{x}) \left| f(\boldsymbol{x}) - \sum_{\|\mathbf{s}\|_1 \leq t} c_{\mathbf{n},\mathbf{s}} \left(\boldsymbol{x} - \frac{\mathbf{n}}{N}\right)^{\mathbf{s}} \right| \\
&= \sum_{\mathbf{n}: \left\|\boldsymbol{x} - \frac{\mathbf{n}}{N}\right\|_\infty < \frac{1}{N}} \left| f(\boldsymbol{x}) - \sum_{\|\mathbf{s}\|_1 \leq t} c_{\mathbf{n},\mathbf{s}} \left(\boldsymbol{x} - \frac{\mathbf{n}}{N}\right)^{\mathbf{s}} \right| \\
&\leq \sum_{\mathbf{n}: \left\|\boldsymbol{x} - \frac{\mathbf{n}}{N}\right\|_\infty < \frac{1}{N}} B d^t \left\|\boldsymbol{x} - \frac{\mathbf{n}}{N}\right\|_\infty^\zeta \\
&\leq B 2^d d^t N^{-\zeta}.
\end{aligned}$$

Setting

$$N = \left(\frac{\epsilon}{2^{d+1} d^t}\right)^{-1/\zeta} \tag{15}$$

for some $\epsilon > 0$, then $|f(\boldsymbol{x}) - p(\boldsymbol{x})| \leq \frac{B\epsilon}{2}$. Hence, it remains to construct a neural network approximating $p(x)$ with the approximation error $\frac{B\epsilon}{2}$. Equivalently, it aims to construct neural networks approximating the product $p_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}) = \psi_{\mathbf{n}}(\boldsymbol{x})\left(\boldsymbol{x} - \frac{\mathbf{n}}{N}\right)^{\mathbf{s}}$. Let $\delta > 0$, then we can recursively define

$$f_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}) = h\left(\psi(Nx_1 - n_1), h(\psi(Nx_2 - n_2), \ldots, h(x_1 - n_1/N, \ldots), \ldots\right),$$

where h is defined in Lemma 2. Using similar arguments in the proof of Theorem 1 in Yarotsky (2017), it holds that $f_{\mathbf{n},\mathbf{s}}$ can be implemented by a ReLU network with the depth and size not larger than $c_1(d+t)\ln(1/\delta)$ for some constant $c_1 = c_1(d,t)$, and

$$|f_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}) - p_{\mathbf{n},\mathbf{s}}(\boldsymbol{x})| \leq (d+t)\delta. \tag{16}$$

Consequently, we establish the desired neural network

$$\tilde{f}(\boldsymbol{x}) = \sum_{\mathbf{n} \in \{0,1,\ldots,N\}^d} \sum_{\|\mathbf{s}\|_1 \leq t} c_{\mathbf{n},\mathbf{s}} f_{\mathbf{n},\mathbf{s}}(\boldsymbol{x}).$$

Therefore, for any $\boldsymbol{x} \in [0,1]^d$, we have

$$\begin{aligned}
|p(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| &= \left| \sum_n \sum_{\|s\|_1 \le t} c_{n,s} p_{n,\boldsymbol{s}}(\boldsymbol{x}) - \sum_n \sum_{\|s\|_1 \le t} c_{n,s} f_{n,\boldsymbol{s}}(\boldsymbol{x}) \right| \\
&\le \sum_n \sum_{\|s\|_1 \le t} |c_{n,s}| |p_{n,s}(\boldsymbol{x}) - f_{n,s}(\boldsymbol{x})| \\
&\le B \sum_{n: \left\|\boldsymbol{x} - \frac{n}{N}\right\|_\infty < \frac{1}{N}} \sum_{\|s\|_1 \le t} |p_{n,s}(\boldsymbol{x}) - \phi_{n,s}(\boldsymbol{x})| \\
&\le B(t+1)d^t \sum_{n: \left\|\boldsymbol{x} - \frac{n}{N}\right\|_\infty < \frac{1}{N}} |p_{n,s}(\boldsymbol{x}) - \phi_{n,s}(\boldsymbol{x})| \\
&\le B(t+1)d^t 2^d \max_{n: \left\|\boldsymbol{x} - \frac{n}{N}\right\|_\infty < \frac{1}{N}} |p_{n,s}(\boldsymbol{x}) - \phi_{n,s}(\boldsymbol{x})| \\
&\le B(t+1)d^{t+1}2^{d+1}\delta,
\end{aligned}$$

where the third inequality follows from $\sum_{\|s\|_1 \le t} 1 = \sum_{j=0}^t \sum_{\|s\|_1 = j} 1 \le \sum_{j=0}^t d^j \le (t+1)d^t$, and the last inequality holds by (16). Setting

$$\delta = \frac{\epsilon}{(t+1)d^{t+1}2^{d+2}}, \tag{17}$$

we have

$$|f(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| \le |f(\boldsymbol{x}) - p(\boldsymbol{x})| + |p(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})| \le B\epsilon.$$

Moreover, $\tilde{f}$ is a liner combination of $d^t(N+1)^d$ neural networks $f_{\mathbf{n},\mathbf{s}}$, we can conclude that the neural network $\tilde{f}$ has not more than $c_1 \ln(1/\delta) + 1$ layers and $d^t(N+1)^d(c_1 \ln(1/\delta) + 1)$ weights by Theorem 1 of Yarotsky (2017). With δ given by (17) and N given by (15), it holds that $\tilde{f}$ has the depth at most $c_2(\ln(1/\epsilon) + 1)$ and at most $c_2\epsilon^{-d/\zeta}(\ln(1/\epsilon) + 1)$ weights, where $c_2 = c_2(d, \zeta)$ is a constant depending d, ζ. Using Lemma 3 and the techniques in proof of Theorem A.9 in Petersen and Voigtlaender (2018), the weights of $\tilde{f}$ can be upper-bounded by $\mathcal{B} = c_3 B \epsilon^{-d/\zeta}$, where $c_3 = c_3(d, \zeta)$ is a constant depending on d, ζ. This completes the proof. ■

A.1.2 Proof of Estimates for the Statistical Error

In this section, we give the proof of Theorem 6. We first introduce the following lemma bounding the covering number of ReLU FNNs in the uniform norm.

Lemma 4. *Let $\mathcal{F}$ be the ReLU FNNs with width $\mathcal{W}$, depth $\mathcal{D}$, and size $\mathcal{S}$. Assume that the parameters of $\mathcal{F}$ are bounded by a constant $\mathcal{B} > 0$, then for each $\delta > 0$,*

$$\log \mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty) \le \mathcal{O}(\mathcal{S}\mathcal{D} \log(\mathcal{B}\mathcal{W}\mathcal{D}/\delta)).$$

Proof. For each $\phi_\theta \in \mathcal{F}$, we have

$$\phi_\theta(x) = (A_{\mathcal{D}}\sigma(\cdot) + b_{\mathcal{D}}) \circ \ldots (A_2\sigma(\cdot) + b_2) \circ (A_1 x + b_1),$$

where $x \in [0,1]^d$, and $\|\theta\|_\infty \leq \mathcal{B}$. For different parameters θ and $\tilde{\theta}$, we denote

$$\phi_\theta(x) = (A_\mathcal{D}\sigma(\cdot) + b_\mathcal{D}) \circ \ldots (A_2\sigma(\cdot) + b_2) \circ (A_1 x + b_1) = \phi_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x),$$
$$\phi_{\tilde{\theta}}(x) = \left(\tilde{A}_\mathcal{D}\sigma(\cdot) + \tilde{b}_\mathcal{D}\right) \circ \ldots \left(\tilde{A}_2\sigma(\cdot) + \tilde{b}_2\right) \circ \left(\tilde{A}_1 x + \tilde{b}_1\right) = \tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \tilde{\phi}_1 \circ \tilde{\phi}_0(x).$$

By adding and subtracting one item, it follows that

$$\begin{aligned}
|\phi_\theta(x) - \phi_{\tilde{\theta}}(x)| = &|\phi_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \tilde{\phi}_1 \circ \tilde{\phi}_0(x)| \\
\leq &|\phi_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x)| \\
&+ |\tilde{\phi}_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x)| \\
&+ \cdots + |\tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \tilde{\phi}_1 \circ \phi_0(x)| \\
&+ |\tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \tilde{\phi}_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \tilde{\phi}_{\mathcal{D}-1} \ldots \tilde{\phi}_1 \circ \tilde{\phi}_0(x)|.
\end{aligned}$$

Therefore, it suffices to bound $\mathcal{D}+1$ terms on the right hand of the above inequality. We provide detailed proofs for obtaining the upper bound of the first term, and the remaining $\mathcal{D}$ terms can be controlled in a similar manner. Through elementary algebraic calculations, we have

$$\begin{aligned}
&|\phi_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x)| \\
=&|(A_\mathcal{D} - \tilde{A}_\mathcal{D})\sigma(g_{\mathcal{D}-1}(x)) + b_\mathcal{D} - \tilde{b}_\mathcal{D}| \\
\leq&(\|g_{\mathcal{D}-1}(x)\|_1 + 1)\|\theta - \tilde{\theta}\|_\infty \\
\leq&(\mathcal{W}\|g_{\mathcal{D}-1}(x)\|_\infty + 1)\|\theta - \tilde{\theta}\|_\infty,
\end{aligned} \tag{18}$$

where $g_{\mathcal{D}-1}(x) := A_{\mathcal{D}-1}\sigma(g_{\mathcal{D}-2}(x)) + b_{\mathcal{D}-1}$. By mathematical recursion, we can bound $\|g_{\mathcal{D}-1}(x)\|_\infty$ as follows:

$$\begin{aligned}
\|g_{\mathcal{D}-1}(x)\|_\infty &= \|A_{\mathcal{D}-1}\sigma(g_{\mathcal{D}-2}(x)) + b_{\mathcal{D}-1}\|_\infty \\
&\leq \|A_{\mathcal{D}-1}\sigma(g_{\mathcal{D}-2}(x))\|_\infty + \|b_{\mathcal{D}-1}\|_\infty \\
&\leq \mathcal{W}\mathcal{B}\|g_{\mathcal{D}-2}(x)\|_\infty + \mathcal{W}\mathcal{B} \\
&\leq (\mathcal{W}\mathcal{B})^{\mathcal{D}-1} + \ldots + \mathcal{W}\mathcal{B} \\
&\leq (\mathcal{D}-1)(\mathcal{W}\mathcal{B})^{\mathcal{D}-1}.
\end{aligned} \tag{19}$$

Combining (18) and (19) gives that

$$\begin{aligned}
&|\phi_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x) - \tilde{\phi}_\mathcal{D} \circ \phi_{\mathcal{D}-1} \ldots \phi_1 \circ \phi_0(x)| \\
=&|(A_\mathcal{D} - \tilde{A}_\mathcal{D})\sigma(g_{\mathcal{D}-1}(x)) + b_\mathcal{D} - \tilde{b}_\mathcal{D}| \\
\leq&(\|g_{\mathcal{D}-1}(x)\|_1 + 1)\|\theta - \tilde{\theta}\|_\infty \\
\leq&(\mathcal{W}\|g_{\mathcal{D}-1}(x)\|_\infty + 1)\|\theta - \tilde{\theta}\|_\infty \\
\leq&\mathcal{D}(\mathcal{W}\mathcal{B})^{\mathcal{D}}\|\theta - \tilde{\theta}\|_\infty.
\end{aligned}$$

Hence, we can conclude that

$$|\phi_\theta(x) - \phi_{\tilde{\theta}}(x)| \leq \bar{L}\|\theta - \tilde{\theta}\|_\infty,$$

where $\bar{L} = \mathcal{O}\left(\mathcal{D}^2(\mathcal{B}\mathcal{W})^{\mathcal{D}}\right)$. Utilizing Lemma 5.13 and Problem 5.5 in van Handel (2016), for each $\delta > 0$, we have

$$\log \mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty) \leq \mathcal{O}(\mathcal{S}\mathcal{D}\log(\mathcal{B}\mathcal{W}\mathcal{D}/\delta)).$$

■

With Lemma 4, we can prove Theorem 6 as follows.

Proof. For each $f \in \mathcal{G}$, we denote

$$L(f) := \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right],$$
$$\widehat{L}_{\mathbb{D}}(f) := \frac{1}{n}\sum_{i=1}^{n}\left(\rho_\tau(Y_i - f(\boldsymbol{X}_i)) - \rho_\tau(Y_i - f^*(\boldsymbol{X}_i))\right).$$

Then, for each $f \in \mathcal{G}$, we have

$$\begin{aligned} L(\widehat{f}_{\mathbb{D}}) = L(\widehat{f}_{\mathbb{D}}) - L(f^*) &= L(f^*) - 2\widehat{L}_{\mathbb{D}}(\widehat{f}_{\mathbb{D}}) + L(\widehat{f}_{\mathbb{D}}) + 2\widehat{L}_{\mathbb{D}}(\widehat{f}_{\mathbb{D}}) - 2L(f^*) \\ &\leq L(f^*) - 2\widehat{L}_{\mathbb{D}}(\widehat{f}_{\mathbb{D}}) + L(\widehat{f}_{\mathbb{D}}) + 2\widehat{L}_{\mathbb{D}}(f) - 2L(f^*). \end{aligned}$$

Taking expectations followed by taking infimum about f over $\mathcal{G}$ in the above equation, we have

$$\mathbb{E}_{\mathbb{D}}\left[L(\widehat{f}_{\mathbb{D}})\right] \leq \mathbb{E}_{\mathbb{D}}\left[L(f^*) - 2\widehat{L}_{\mathbb{D}}(\widehat{f}_{\mathbb{D}}) + L(\widehat{f}_{\mathbb{D}})\right].$$

Therefore, it remains to derive the upper bound of the statistical error

$$\mathcal{E}_{\text{sta}} := \mathbb{E}_{\mathbb{D}}\left[L(f^*) - 2\widehat{L}_{\mathbb{D}}(\widehat{f}_{\mathbb{D}}) + L(\widehat{f}_{\mathbb{D}})\right].$$

Denote by $\tilde{\ell}(f;\omega) := \rho_\tau(y - f(x)) - \rho_\tau(y - f^*(x))$ with $\omega := (x, y)$. It is easy to check that $\tilde{\ell}(f;\omega)$ is Lipschitz continuous over f, i.e., for each ω, we have

$$|\tilde{\ell}(f_1;\omega) - \tilde{\ell}(f_2;\omega)| \leq \lambda|f_1(x) - f_2(x)|, \text{ with } \lambda = 1.$$

Let $\mathbb{D}' := \{U_1', \ldots, U_n'\}$ be an i.i.d ghost sample independent of $\mathbb{D} = \{U_1, \ldots, U_n\}$ with $U_i := (X_i, Y_i)$, $i = 1, \ldots, n$. Let $G(f;U) := \mathbb{E}_{\mathbb{D}'}\left[\tilde{\ell}\left(f;U'\right) - 2\tilde{\ell}(f;U)\right]$. Then, we have

$$\mathcal{E}_{\text{sta}} = \mathbb{E}_{\mathbb{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(-2\tilde{\ell}\left(\widehat{f}_{\mathbb{D}};U_i\right) + \mathbb{E}_{\mathbb{D}'}\left[\tilde{\ell}\left(\widehat{f}_{\mathbb{D}};U_i'\right)\right]\right)\right] = \mathbb{E}_{\mathbb{D}}\left[\frac{1}{n}\sum_{i=1}^{n}G\left(\widehat{f}_{\mathbb{D}};U_i\right)\right].$$

Let $\mathcal{N}(\mathcal{G}, \delta, \|\cdot\|_\infty)$ be the covering number of $\mathcal{G}$ with cover $\mathcal{C} = \{f_1, \ldots, f_{\mathcal{N}}\}$. $\forall f \in \mathcal{G}$, there exists a $\tilde{f} \in \mathcal{C}$ such that

$$|\tilde{\ell}(f;\omega) - \tilde{\ell}(\tilde{f};\omega)| \leq \lambda\|f - \tilde{f}\|_\infty \leq \lambda\delta,$$
$$G(f;\omega) \leq G(\tilde{f};\omega) + 3\lambda\delta.$$

By Assumption 3, it follows that L is strongly convex at f^*, i.e., for each f,

$$L(f) - L(f^*) \geq c\,\|f - f^*\|_{L_2(P_X)}^2,$$

where $c > 0$ is an absolute constant. For each $\tilde{f} \in \mathcal{C}$, we have

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} G\left(\tilde{f}; U_i\right) > t\right] &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbb{D}'}\left[\tilde{\ell}(\tilde{f}; U_i')\right] - \frac{2}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i) > t\right] \\ =\mathbb{P}\left[\mathbb{E}_{\mathbb{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i) > t/2 + \mathbb{E}_{\mathbb{D}}\left[\frac{1}{2n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i)\right]\right].\end{aligned} \tag{20}$$

It is easy to show that $|\tilde{\ell}(\tilde{f}; U_i)| \leq 2\lambda B$, and $|\tilde{\ell}(\tilde{f}; U_i) - \mathbb{E}\tilde{\ell}(\tilde{f}; U_i)| \leq 4\lambda B := b$. Denote by $\sigma^2 := Var(\tilde{\ell}(\tilde{f}, U_i))$, then we have

$$\sigma^2 \leq \mathbb{E}[\tilde{\ell}(\tilde{f}; U_i)^2] \leq \lambda^2 \|\tilde{f} - f^*\|_{L_2(P_X)}^2 \leq \frac{\lambda^2}{c}\mathbb{E}[\tilde{\ell}(\tilde{f}; U_i)].$$

Thus, we have $\mathbb{E}[\tilde{\ell}(\tilde{f}; U_i)] \geq \frac{c\sigma^2}{\lambda^2}$. Let $v := \frac{t}{2} + \frac{c\sigma^2}{2\lambda^2} = \frac{t}{2} + \frac{2Bc\sigma^2}{b\lambda}$, then we have $\sigma^2 \leq \frac{b\lambda v}{2Bc}$, and $v \geq t/2$. By Bernstein's inequality and (20), we have

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} G\left(\tilde{f}; U_i\right) > t\right] &\leq \mathbb{P}\left[\mathbb{E}_{\mathbb{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i) > t/2 + \frac{c\sigma^2}{2\lambda^2}\right] \\ &= \mathbb{P}\left[\mathbb{E}_{\mathbb{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(\tilde{f}; U_i) > v\right] \\ &\leq \exp(-nv^2/2(\sigma^2 + bv)) \\ &\leq \exp(-nv/b(2 + \lambda/Bc)) \\ &\leq \exp(-nt/(8\lambda B + 4\lambda^2/c)).\end{aligned}$$

Hence, $\forall t > 3\lambda\delta$, we have

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} G\left(\widehat{f}_{\mathbb{D}}; U_i\right) > t\right] &\leq \mathbb{P}\left[\max_{f\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n} G\left(f; U_i\right) > t\right] \\ &\leq \mathbb{P}\left[\max_{\tilde{f}\in\mathcal{C}}\frac{1}{n}\sum_{i=1}^{n} G\left(\tilde{f}; U_i\right) > t - 3\lambda\delta\right] \\ &\leq \mathcal{N}\exp\left(-\frac{n(t-3\lambda\delta)}{8\lambda B + 4\lambda^2/c}\right).\end{aligned}$$

By Lemma 4 and setting $a = 3\lambda\delta + c_0$ with $\delta = 1/n$ and $c_0 = (8\lambda B + 4\lambda^2/c)\log\mathcal{N}/n$ yield that

$$\begin{aligned}\mathcal{E}_{\text{sta}} &\leq a + \int_a^{\infty}\mathcal{N}\exp\left(-\frac{n(t-3\lambda\delta)}{8\lambda B + 4\lambda^2/c}\right)\mathrm{d}t \\ &\leq a + \mathcal{N}\exp\left(-\frac{c_0 n}{8\lambda B + 4\lambda^2/c}\right)\frac{4\lambda B + 4\lambda^2/c}{n} \\ &\leq \frac{\left(8\lambda B + 4\lambda^2/c\right)(\log\mathcal{N} + 1) + 3\lambda}{n} \\ &\leq \mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right).\end{aligned}$$

This completes the proof. ■

A.2 Proof of the Main Results for the Unweighted Estimators

A.2.1 Proof of Error Bounds for the Unweighted Estimators Under the Uniformly Bounded Case

In this section, we give the proof of Theorem 7, which provides the non-asymptotic error bound for the unweighted estimators under the uniformly bounded case. The proof highly relies on the nice property that $\|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(Q_{\boldsymbol{X}})} \leq \Gamma \|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(P_{\boldsymbol{X}})}$ given the assumption that $\sup_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) \leq \Gamma$.

Proof. By Lemma 1 and taking the expectation on the training data $\mathbb{D}$, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(P_{\boldsymbol{X}})}\right] \lesssim\ & B\mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right] \\ & + B^2 \inf_{f\in\mathcal{G}} \|f - f_0\|^2_\infty\end{aligned}$$

For the statistical error term, from Theorem 6, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[\rho_\tau(Y - \widehat{f}_{\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right] \\ \leq&\mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right) \leq \mathcal{O}\left(Bn^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\right),\end{aligned}$$

where the second inequality is from the fact that the function class $\mathcal{G}$ is a ReLU FNN $\mathcal{F}$ bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}(n^{\frac{d}{d+2\zeta}}\log n)$ and the depth $\mathcal{D} = \mathcal{O}(\log n)$, and weights bound $\mathcal{B} = \mathcal{O}(Bn^{\frac{d}{d+2\zeta}})$.

For the approximation error term, this can be bounded by using Theorem 4.

By combining these two error terms, we have

$$\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(P_{\boldsymbol{X}})} \leq \mathcal{O}\left(B^2 n^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\right). \tag{21}$$

Then by Assumption 4 that $\sup_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) \leq \Gamma$, there holds

$$\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})} \leq \Gamma\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|^2_{L_2(P_{\boldsymbol{X}})} \leq \mathcal{O}\left(\Gamma B^2 n^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\right).$$

This completes the proof of Theorem 7. ■

A.2.2 Proof of Error Bounds for the Unweighted Estimators Under the Bounded Second Moment Case

In this section, we give the proof of Theorem 8 , which provides the non-asymptotic error bound for the unweighted estimators under the bounded second moment case. In this case, the property that $\|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(Q_{\boldsymbol{X}})} \leq V^2\|\widehat{f}_{\mathbb{D}} - f_0\|_{L_2(P_{\boldsymbol{X}})}$ does not hold for the bounded second moment case. By using the Cauchy-Schwarz inequality, we can only obtain a suboptimal rate.

Proof. By the definition of the density ratio r, for each $f \in \mathcal{F}$, we have

$$\begin{aligned}\|f - f_0\|_{L^2(Q_{\boldsymbol{X}})}^2 &= \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[r(\boldsymbol{X})(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^2\right] \\ &\leq \left\{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[r^2(\boldsymbol{X})\right]\right\}^{\frac{1}{2}} \left\{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^4\right]\right\}^{\frac{1}{2}} \\ &\leq V^{\frac{1}{2}} \left\{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[(f(\boldsymbol{X}) - f_0(\boldsymbol{X}))^4\right]\right\}^{\frac{1}{2}} \\ &\leq (4B^2V)^{\frac{1}{2}} \left\{\|f - f_0\|_{L^2(P_{\boldsymbol{X}})}^2\right\}^{\frac{1}{2}},\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second inequality is from Assumption 5 that $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^2(\boldsymbol{X})] \leq V^2$, and the last inequality is due to that fact that $\|f\|_\infty \leq B$ and $\|f_0\|_\infty \leq B$. Combining this result with (21), we have

$$\mathbb{E}_{\mathbb{D}}\|\widehat{f}_{\mathbb{D}} - f_0\|_{L^2(Q_{\boldsymbol{X}})}^2 \leq (4B^2V)^{\frac{1}{2}}\mathbb{E}_{\mathbb{D}}\left\{\|f - f_0\|_{L^2(P_{\boldsymbol{X}})}^2\right\}^{\frac{1}{2}} \leq \mathcal{O}\left(B^2V^{\frac{1}{2}}n^{-\frac{\zeta}{d+2\zeta}}(\log n)^{\frac{3}{2}}\right).$$

This completes the proof. ■

A.3 Proof of the Main results for the Reweighted estimators

A.3.1 Proof of Error Bounds for the Reweighted Estimators Under the Uniformly Bounded Case

In this section, we give the proof of Theorem 9, which provides the non-asymptotic error bound for the reweighted estimators under the uniformly bounded case. This proof follows a similar approach to that in Theorem 7. Specifically, it incorporates a modified error decomposition as detailed in Lemma 5 and includes estimates for the statistical error term from Lemma 6.

Lemma 5. *Suppose that Assumptions 2 and 3 are satisfied, and the function space $\mathcal{J}$ is also uniformly bounded by B with $B \geq 1$. Then, the reweighted estimator $\widehat{f}_{r,\mathbb{D}}$ defined in* (6) *satisfies*

$$\begin{aligned}\|\widehat{f}_{r,\mathbb{D}} - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2 \lesssim\ & B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})\left(\rho_\tau(Y - \widehat{f}_{r,\mathbb{D}}(\boldsymbol{X})) - \rho_\tau\left(Y - f^*(\boldsymbol{X})\right)\right)\right] \\ &+ B^2\|f_0 - f^*\|_\infty^2.\end{aligned}$$

Proof. By the Knight identity and Fubini's theorem, we can employ the same proof procedure as detailed in Lemma 1 in Section 4.1.1, which shows that

$$
\begin{aligned}
&\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})(\rho_\tau(Y-f(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X})))\right]\\
=&\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[-r(\boldsymbol{X})\big(f(\boldsymbol{X})-f^*(\boldsymbol{X})\big)\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}\big[(\tau-\mathbf{I}\{Y\le f_0(\boldsymbol{X})\})\mid\boldsymbol{X}\big]\Big]\\
-&\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[r(\boldsymbol{X})\big(f(\boldsymbol{X})-f^*(\boldsymbol{X})\big)\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}\big[(\mathbf{I}\{Y\le f_0(\boldsymbol{X})\}-\mathbf{I}\{Y\le f^*(\boldsymbol{X})\})\mid\boldsymbol{X}\big]\Big]\\
&+\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\Big[\int_0^{f(\boldsymbol{X})-f^*(\boldsymbol{X})} r(\boldsymbol{X})\big(\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}[\mathbf{I}\{Y\le f^*(\boldsymbol{X})+z\}\mid\boldsymbol{X}]\\
&-\mathbb{E}_{Y|\boldsymbol{X}\sim P_{Y|\boldsymbol{X}}}[\mathbf{I}\{Y\le f^*(\boldsymbol{X})\}\mid\boldsymbol{X}]\big)dz\Big]\\
\ge&-C_1\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f(\boldsymbol{X})-f^*(\boldsymbol{X})||f_0(\boldsymbol{X})-f^*(\boldsymbol{X})|\big]\\
&+C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})D^2(f(\boldsymbol{X})-f^*(\boldsymbol{X}))\big]\\
\ge&-C_1\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]}\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f_0(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]}\\
&+C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})D^2(f(\boldsymbol{X})-f^*(\boldsymbol{X}))\big],
\end{aligned}
$$

where C_1, C_2 are two absolute positive constants. Then, for any $\beta>0$,

$$
\begin{aligned}
&C_2\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})D^2(f(\boldsymbol{X})-f^*(\boldsymbol{X}))\big]\\
\le&C_1\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]}\sqrt{\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f_0(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]}\\
&+\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[r(\boldsymbol{X})\big(\rho_\tau(Y-f(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\big)\big]\\
\le&\frac{C_1}{4\beta}\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]+C_1\beta\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[r(\boldsymbol{X})|f_0(\boldsymbol{X})-f^*(\boldsymbol{X})|^2\big]\\
&+\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[r(\boldsymbol{X})\big(\rho_\tau(Y-f(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\big)\big].
\end{aligned}
$$

By setting $\beta=\frac{C_1B}{C_2}$ and applying the inequality $D^2(f(\boldsymbol{X})-f^*(\boldsymbol{X}))\ge\frac{|f(\boldsymbol{X})-f^*(\boldsymbol{X})|^2}{2B}$, which holds almost surely for $B\ge1$, there holds

$$
\begin{aligned}
\|f-f^*\|_{L_2(Q_{\boldsymbol{X}})}^2\le&\frac{4B}{C_2}\Big(\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\big[r(\boldsymbol{X})\big(\rho_\tau(Y-f(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\big)\big]\Big)\\
&+\frac{2C_1^2B^2}{C_2^2}\|f_0-f^*\|_{L_2(Q_{\boldsymbol{X}})}^2.
\end{aligned}
$$

Using the triangle inequality yields that

$$
\begin{aligned}
&\|\widehat{f}_{r,\mathbb{D}}-f_0\|_{L_2(Q_{\boldsymbol{X}})}^2\\
&\le2\|\widehat{f}_{r,\mathbb{D}}-f^*\|_{L_2(Q_{\boldsymbol{X}})}^2+2\|f^*-f_0\|_{L_2(Q_{\boldsymbol{X}})}^2\\
&\lesssim B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})\left(\rho_\tau(Y-\widehat{f}_{r,\mathbb{D}}(\boldsymbol{X}))-\rho_\tau\left(Y-f^*(\boldsymbol{X})\right)\right)\right]+B^2\|f_0-f^*\|_{L_2(Q_{\boldsymbol{X}})}^2\\
&\lesssim B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})\left(\rho_\tau(Y-\widehat{f}_{r,\mathbb{D}}(\boldsymbol{X}))-\rho_\tau\left(Y-f^*(\boldsymbol{X})\right)\right)\right]+B^2\|f_0-f^*\|_{\infty}^2.
\end{aligned}
$$

This completes the proof. ■

Lemma 6. *Given the reweighted estimator $\widehat{f}_{r,\mathbb{D}}$ in (6) and the considered function class $\mathcal{J}$ is ReLU FNN $\mathcal{F}$, we have*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})\big(\rho_\tau(Y-\widehat{f}_{r,\mathbb{D}}(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\big)\right]\\ \leq&\mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\Gamma^2\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right).\end{aligned}$$

Proof. This proof is similar to that of Theorem 6, requiring solely the redefinition of L, $\widehat{L}_{\mathbb{D}}$, and $\tilde{\ell}$. Specifically, we reformulate them as

$$\begin{aligned}L(f)&:=\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})(\rho_\tau(Y-f(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X})))\right],\\ \widehat{L}_{\mathbb{D}}(f)&:=\frac{1}{n}\sum_{i=1}^{n}r(X_i)\left(\rho_\tau(Y_i-f(X_i))-\rho_\tau(Y_i-f^*(X_i))\right),\end{aligned}$$

where $f\in\mathcal{J}$. Then, we define $\tilde{\ell}(f;\omega):=r(x)\left(\rho_\tau(y-f(x))-\rho_\tau(y-f^*(x))\right)$ with $\omega:=(x,y)$. It is easy to check that $\tilde{\ell}(f;\omega)$ is Lipschitz continuous over f, i.e., for each ω, we have

$$|\tilde{\ell}(f_1;\omega)-\tilde{\ell}(f_2;\omega)|\leq r(x)|f_1(x)-f_2(x)|\leq\lambda|f_1(x)-f_2(x)|,\text{ with }\lambda=\Gamma.$$

Using Assumption 3, we can conclude that L is strongly convex at f^*, i.e., for each f,

$$L(f)-L(f^*)\geq c\left\|f-f^*\right\|_{L_2(Q_{\boldsymbol{X}})}^2,$$

where $c>0$ is an absolute constant. Therefore, employing analogous arguments to the proof of Theorem 6 gives the desired upper bound. This completes the proof. ■

With Theorem 4, and Lemmas 5-6, we proceed the proof of Theorem 9 as given below.

Proof. By Lemma 1 and taking the expectation on the training data $\mathbb{D}$, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{\mathbb{D}}-f_0\|_{L_2(P_{\boldsymbol{X}})}^2\right]\lesssim&B\mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})\big(\rho_\tau(Y-\widehat{f}_{\mathbb{D}}(\boldsymbol{X}))-\rho_\tau(Y-f^*(\boldsymbol{X}))\big)\right]\\&+B^2\inf_{f\in\mathcal{G}}\|f-f_0\|_\infty^2.\end{aligned}$$

We set the function class $\mathcal{J}$ to be a ReLU FNN bounded by $B\geq 1$, with the size $\mathcal{S}=\mathcal{O}(n^{\frac{d}{d+2\zeta}}\log n)$ and the depth $\mathcal{D}=\mathcal{O}(\log n)$, and weights bound $\mathcal{B}=\mathcal{O}(Bn^{\frac{d}{d+2\zeta}})$. Using Lemma 6 and Theorem 4, then we obtain the desired bound.

■

A.3.2 Proof of Error Bounds for the Reweighted Estimators Under the Bounded Second Moment Case

In this section, we give the proof of Theorem 10, which provides the non-asymptotic error bound for the truncated reweighted estimators in the case of a bounded second moment. We consider a more general assumption that $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}[r^{1+\delta}(\boldsymbol{X})]=U<\infty$, for some $\delta\geq 0$, and second moment bounded is a special case when $\delta=1$. **Proof.** We denote by

$$\begin{aligned}\mathcal{R}_r^*(f)&:=\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[r(\boldsymbol{X})(\rho_\tau(f(\boldsymbol{X})-Y)-\rho_\tau(f^*(\boldsymbol{X})-Y))\right],\\ \mathcal{R}_{T_\xi r}^*(f)&:=\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left[T_\xi r(\boldsymbol{X})(\rho_\tau(f(\boldsymbol{X})-Y)-\rho_\tau(f^*(\boldsymbol{X})-Y))\right],\end{aligned}$$

where $f \in \mathcal{K}$. Then, we have

$$\begin{aligned}
&\mathcal{R}_r^*(f) - \mathcal{R}_{T_\xi r}^*(f) \\
&= \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}} \left[(r(\boldsymbol{X}) - \xi) I(r(\boldsymbol{X}) \geq \xi)(\rho_\tau(f(\boldsymbol{X}) - Y) - \rho_\tau(f^*(\boldsymbol{X}) - Y))\right] \\
&= \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}} \left[r(\boldsymbol{X}) I(r(\boldsymbol{X}) \geq \xi)(\rho_\tau(f(\boldsymbol{X}) - Y) - \rho_\tau(f^*(\boldsymbol{X}) - Y))\right] \\
&\quad - \mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}} \left[\xi I(r(\boldsymbol{X}) \geq \xi)(\rho_\tau(f(\boldsymbol{X}) - Y) - \rho_\tau(f^*(\boldsymbol{X}) - Y))\right] \\
&\leq 2B\mathbb{E}_{X\sim P_{\boldsymbol{X}}} \left[r(\boldsymbol{X}) \frac{r^\delta(\boldsymbol{X})}{\xi^\delta}\right] + 2B\xi \mathbb{E}_{X\sim P_{\boldsymbol{X}}} \left[\frac{r^{1+\delta}(\boldsymbol{X})}{\xi^{1+\delta}}\right] \\
&\leq \frac{4BU}{\xi^\delta},
\end{aligned}$$

where the inequality holds by the $1+\delta$ moment bounded assumption and Markov inequality. According to the proof of Lemma 5, for each $f \in \mathcal{K}$,

$$\begin{aligned}
\|f - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2 &\lesssim B\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}} \left[r(\boldsymbol{X}) \left(\rho_\tau(Y - f(\boldsymbol{X})) - \rho_\tau\left(Y - f^*(\boldsymbol{X})\right)\right)\right] \\
&\quad + B^2 \|f_0 - f^*\|_{L_2(Q_{\boldsymbol{X}})}^2 \\
&\lesssim B\mathcal{R}_{T_\xi r}^*(f) + \frac{B^2 U}{\xi^\delta} + B^2 \|f_0 - f^*\|_\infty^2.
\end{aligned}$$

Similar to the proof of Lemma 6, it holds that $\mathbb{E}_{\mathbb{D}}[\mathcal{R}_{T_\xi r}^*(\widehat{f}_{T_\xi r,\mathbb{D}})] \leq \mathcal{O}\big(\frac{B\mathcal{S}\mathcal{D}\xi^2 \log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\big)$. Setting $\xi = \left(\frac{nU}{\mathcal{S}\mathcal{D}\log(n\mathcal{B}\mathcal{W}\mathcal{D})}\right)^{\frac{1}{2+\delta}}$ yields that

$$\mathbb{E}_{\mathbb{D}} \|\widehat{f}_{T_\xi r,\mathbb{D}} - f_0\|_{L_2(Q_X)}^2 \lesssim U^{2\delta/(2+\delta)} B^2 (\mathcal{S}\mathcal{D}\log(n\mathcal{B}\mathcal{W}\mathcal{D})/n)^{1/(1+\delta)} + B^2 \|f_0 - f^*\|_\infty^2.$$

Using Theorem 4 and setting the function class $\mathcal{K}$ as a ReLU FNN bounded by $B \geq 1$, with the size $\mathcal{S} = \mathcal{O}\left(n^{\frac{d}{d+(2+4/\delta)\zeta}} \log n\right)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and weights bound $\mathcal{B} = \mathcal{O}\left(Bn^{\frac{d}{d+(2+4/\delta)\zeta}}\right)$, it follows that

$$\mathbb{E}_{\mathbb{D}}\left[\|\widehat{f}_{T_\xi r,\mathbb{D}} - f_0\|_{L^2(Q_{\boldsymbol{X}})}^2\right] \leq \mathcal{O}\left(U^{2\delta/(2+\delta)} B^2 n^{-\frac{2\zeta}{d+(2+4/\delta)\zeta}} \log n\right).$$

Let $\delta = 1$ and $U = V^2$, we complete the proof.

■

A.4 Proof of the Main results for the Pre-training Reweighted estimators

A.4.1 Proof of Error Bounds for the Pre-training Reweighted Estimators Under the Uniformly Bounded Case

In this section, we give the proof of Theorem 12 and Theorem 13, which provide the non-asymptotic error bound for the pre-training density ratio and the pre-training reweighted estimators under the uniformly bounded case.

Proof. For each $u \in \mathcal{U}$, we have

$$\begin{aligned} L(\widehat{r}_{\mathbb{S}}) - L(r) &= L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\mathbb{S}}) + L(\widehat{r}_{\mathbb{S}}) + 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\mathbb{S}}) - 2L(r) \\ &\leq L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\mathbb{S}}) + L(\widehat{r}_{\mathbb{S}}) + 2\widehat{L}_{\mathbb{S}}(u) - 2L(r). \end{aligned}$$

Next, we bound the statistical error $\mathcal{E}_{\text{sta}} = \mathbb{E}_{\mathbb{S}}\left[L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\mathbb{S}}) + L(\widehat{r}_{\mathbb{S}})\right]$ and approximation error $2\widehat{L}_{\mathbb{S}}(u) - 2L(r)$. This approximation error can be controlled by Theorem 4. Therefore, it remains to derive the upper bound of the statistical error. This can be done by following the proof of Theorem 6. We denote by $\ell(u; x) := \frac{1}{2}u^2(x_p) - u(x_q)$, with $x := (x_p, x_q)$. Then, we define $\tilde{\ell}(u; x) := \ell(u; x) - \ell(r; x)$. It is easy to check that $\tilde{\ell}(u; x)$ is Lipschitz continuous over u, i.e., for each x, we have

$$|\tilde{\ell}(u_1; x) - \tilde{\ell}(u_2; x)| \leq \lambda \left(|u_1(x_p) - u_2(x_p)| + |u_1(x_q) - u_2(x_q)|\right), \text{ with } \lambda = \Gamma.$$

Moreover, L is strongly convex at r, i.e., for each u,

$$L(u) - L(r) \geq c \left\|u - r\right\|_{L_2(P_X)}^2, \ c := 1.$$

Then, using the similar arguments in the proof of Theorem 6 yields that

$$\mathcal{E}_{\text{sta}} \leq \frac{\left(32\lambda\Gamma + 8(2+\Gamma)\lambda^2/c\right)(\log \mathcal{N} + 1)}{m} \lesssim \frac{\Gamma^3 \mathcal{S}\mathcal{D}\log(m\mathcal{B}\mathcal{W}\mathcal{D})}{m}.$$

Then, setting the size $\mathcal{S} = \mathcal{O}\left(m^{\frac{d}{d+2\alpha}} \log m\right)$, the depth $\mathcal{D} = \mathcal{O}(\log m)$, and weights bound $\mathcal{B} = \mathcal{O}\left(\Gamma m^{\frac{d}{d+2\alpha}}\right)$ gives the desired result.

■

By using Theorem 12, we are ready to prove Theorem 13.

Proof.

Using Lemma 5, we can similarly get

$$\begin{aligned} &\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|_{L_2(Q_X)}^2\right] \\ &\lesssim B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[r(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\ &+ B^2\|f^* - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2 \\ &\lesssim B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[(r(\boldsymbol{X}) - \widehat{r}_{\mathbb{S}}(\boldsymbol{X}))(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\ &\quad + B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\ &\quad + B^2\|f^* - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2. \end{aligned} \tag{22}$$

For any $\beta > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[(r(\boldsymbol{X}) - \widehat{r}_{\mathbb{S}}(\boldsymbol{X}))(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
&\leq \mathbb{E}_{\mathbb{S}}\left[\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left|(r(\boldsymbol{X}) - \widehat{r}_{\mathbb{S}}(\boldsymbol{X}))\right|^2/2\beta\right] \\
&\quad + \mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left|\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right|^2\beta/2\right] \\
&\leq \mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\mathbb{S}}\|^2_{L_2(P_{\boldsymbol{X}})}\right]/2\beta + \mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\frac{\beta}{2\Upsilon}\left\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f^*\right\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \\
&\leq \mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\mathbb{S}}\|^2_{L_2(P_{\boldsymbol{X}})}\right]/2\beta + \frac{\beta}{\Upsilon}\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\left\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\right\|^2_{L_2(Q_{\boldsymbol{X}})}\right] + \frac{\beta}{\Upsilon}\|f_0 - f^*\|^2_{L_2(Q_{\boldsymbol{X}})}. \qquad (23)
\end{aligned}$$

Combining (22)-(23) and setting $\beta = \frac{\Upsilon}{2B}$ yield that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \\
&\lesssim B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
&\quad + B^2\|f^* - f_0\|^2_\infty + \frac{B^2}{\Upsilon}\mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\mathbb{S}}\|^2_{L_2(P_{\boldsymbol{X}})}\right]. \qquad (24)
\end{aligned}$$

In (24), we can deduce that $\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))]$ $\leq \mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\Gamma^2\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right)$ by employing similar arguments in Lemma 6, $\|f^* - f_0\|^2_\infty$ can be bounded by Theorem 4, and $\mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\mathbb{S}}\|^2_{L_2(P_{\boldsymbol{X}})}\right]$ can be bounded using Theorem 12. Therefore, setting the size $\mathcal{S} = \mathcal{O}\left(n^{\frac{d}{d+2\zeta}}\log n\right)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, and weights bound $\mathcal{B} = \mathcal{O}\left(Bn^{\frac{d}{d+2\zeta}}\right)$, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \leq \mathcal{O}\left(B^2\Gamma^2 n^{\frac{-2\zeta}{2\zeta+d}}(\log n)^3\right) + \mathcal{O}\left(\frac{B^2\Gamma^3 m^{-\frac{2\alpha}{d+2\alpha}}(\log m)^3}{\Upsilon}\right).$$

Moreover, if $m \geq \Omega\Big(\left(\frac{\Gamma}{\Upsilon}\right)^{\frac{d+2\alpha}{2\alpha}} n^{\frac{\zeta(d+2\alpha)}{\alpha(d+2\zeta)}}\Big)$, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2\Gamma^2 n^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\Big).$$

This completes the proof.

■

A.4.2 Proof of Error Bounds for the Pre-training Reweighted Estimators under the $2+\delta$ Moment Bounded Case

In this section, we give the proof of Theorem 14 and Theorem 16 density, which provide the non-asymptotic error bound for the pre-training density ratio and the pre-training reweighted estimators under the $2+\delta$ moment bounded case. The proof of Theorem 14 is similar to that of Theorem 12, except for an extra error term for the truncation $\|T_\xi r -$

$r\|^2_{L_2(P_{\boldsymbol{X}})}$ which can be well bounded with Assumption 9. And with Theorem 14, the proof of Theorem 16 is exactly the same as that of Theorem 13.

Proof. For each $u \in T_\xi \mathcal{U}$, we have

$$\begin{aligned} L(\widehat{r}_{\xi,\mathbb{S}}) - L(r) &= L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\xi,\mathbb{S}}) + L(\widehat{r}_{\xi,\mathbb{S}}) + 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\xi,\mathbb{S}}) - 2L(r) \\ &\leq L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\xi,\mathbb{S}}) + L(\widehat{r}_{\xi,\mathbb{S}}) + 2\widehat{L}_{\mathbb{S}}(u) - 2L(r). \end{aligned}$$

Next, we bound the statistical error $\mathcal{E}_{\text{sta}} = \mathbb{E}_{\mathbb{S}}\left[L(r) - 2\widehat{L}_{\mathbb{S}}(\widehat{r}_{\xi,\mathbb{S}}) + L(\widehat{r}_{\xi,\mathbb{S}})\right]$ and approximation error $2\widehat{L}_{\mathbb{S}}(u) - 2L(r)$. Using the triangle inequality, the approximation error satisfies

$$\inf_{u\in T_\xi\mathcal{U}} \mathbb{E}_{\mathbb{S}}[\widehat{L}_{\mathbb{S}}(u) - L(r)] \leq 2 \inf_{u\in T_\xi\mathcal{U}} \|u - T_\xi r\|^2_{L_2(P_{\boldsymbol{X}})} + 2\|T_\xi r - r\|^2_{L_2(P_{\boldsymbol{X}})}. \tag{25}$$

On the right hand of (25), the first term can be bounded by Theorem 4. In terms of the second term, it can be bounded by using Assumption 9. Specifically, there holds that

$$\begin{aligned} \|T_\xi r - r\|^2_{L_2(P_{\boldsymbol{X}})} &\leq \|rI(r > \xi)\|^2_{L_2(P_{\boldsymbol{X}})} \\ &\leq \mathbb{E}_{X\sim P_{\boldsymbol{X}}}\left[r^2(\boldsymbol{X})\frac{r^\delta(\boldsymbol{X})}{\xi^\delta}\right] = \frac{\mathbb{E}_{X\sim P_{\boldsymbol{X}}}[r^{2+\delta}(\boldsymbol{X})]}{\xi^\delta} \\ &\leq \frac{\Xi}{\xi^\delta}, \end{aligned}$$

where the last inequality follows from Assumption 9. Moreover, similar to the argument of the proof of Theorem 12, the statistical error $\mathcal{E}_{\text{sta}}$ can be bounded by $\mathcal{O}\left(\frac{\xi^3\mathcal{S}\mathcal{D}\log(m\mathcal{B}\mathcal{W}\mathcal{D})}{m}\right)$. Then, by setting the size $\mathcal{S} = \mathcal{O}\left(m^{\frac{\delta d}{\delta d+(6+2\delta)\alpha}}\right)$, the depth $\mathcal{D} = \mathcal{O}(\log m)$, weights bound $\mathcal{B} = \mathcal{O}\left(\xi m^{\frac{\delta d}{\delta d+(6+2\delta)\alpha}}\right)$ for the ReLU DNNs, and the truncation level $\xi = \mathcal{O}\left(m^{\frac{2\alpha}{\delta d+(6+2\delta)\alpha}}\right)$, we obtain the desired result.

If we further assume that the square of the density ratio r^2 is sub-exponential with respect to $P_{\boldsymbol{X}}$, i.e., $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left[\exp(\sigma r^2(\boldsymbol{X}))\right] < \infty$, for some positive constant σ, then

$$\begin{aligned} \|T_\xi r - r\|^2_{L_2(P_{\boldsymbol{X}})} &\leq \|rI(r > \xi)\|^2_{L_2(P_{\boldsymbol{X}})} \\ &\leq \frac{2}{\sigma}\mathbb{E}_{X\sim P_{\boldsymbol{X}}}\left[\frac{\sigma r^2(\boldsymbol{X})}{2}\exp\left(\frac{\sigma}{2}(r^2(\boldsymbol{X}) - \xi^2)\right)\right] \\ &\leq \frac{2}{\sigma}\exp\left(-\frac{\sigma}{2}\xi^2\right)\mathbb{E}_{X\sim P_{\boldsymbol{X}}}\left[\exp\left(\sigma r^2(\boldsymbol{X})\right)\right] \\ &\lesssim \exp\left(-\frac{\sigma}{2}\xi^2\right), \end{aligned}$$

Then, setting the size $\mathcal{S} = \mathcal{O}\left(m^{\frac{d}{d+2\alpha}}\log m\right)$, the depth $\mathcal{D} = \mathcal{O}(\log m)$, weights bound $\mathcal{B} = \mathcal{O}\left(\xi m^{\frac{d}{d+2\alpha}}\right)$ for the ReLU DNNs, and the truncation level $\xi = \mathcal{O}\left(\sqrt{\log m}\right)$ gives the convergence rate of $\mathcal{O}(m^{-\frac{2\alpha}{d+2\alpha}}(\log m)^{\frac{9}{2}})$. ■

Proof. Following a similar procedure in the proof of Theorem 13, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|_{L_2(Q_X)}^2\right] \\
\lesssim & B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[r(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
&+ B^2\|f^* - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2 \\
\lesssim & B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[(r(\boldsymbol{X}) - \widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X}))(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
&+ B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
&+ B^2\|f^* - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2.
\end{aligned}$$

For the first part of the above equation, for any $\beta > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[(r(\boldsymbol{X}) - \widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X}))(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
\leq & \mathbb{E}_{\mathbb{S}}\left[\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\left|(r(\boldsymbol{X}) - \widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X}))\right|^2/2\beta\right] \\
&+ \mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}\left|\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X}))\right|^2\beta/2\right] \\
\leq & \mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\xi,\mathbb{S}}\|_{L_2(P_{\boldsymbol{X}})}^2\right]/2\beta + \mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\frac{\beta}{2\Upsilon}\left\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f^*\right\|_{L_2(Q_{\boldsymbol{X}})}^2\right] \\
\leq & \mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\xi,\mathbb{S}}\|_{L_2(P_{\boldsymbol{X}})}^2\right]/2\beta + \frac{\beta}{\Upsilon}\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\left\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\right\|_{L_2(Q_{\boldsymbol{X}})}^2\right] + \frac{\beta}{\Upsilon}\|f_0 - f^*\|_{L_2(Q_{\boldsymbol{X}})}^2,
\end{aligned}$$

where the second inequality follows from the fact that $\rho_\tau(\cdot)$ is Lipschitz continuous and Assumption 4.9 that $\Upsilon = \inf_{\boldsymbol{x}\in\mathcal{X}} r(\boldsymbol{x}) > 0$.

Plugging in this result and setting $\beta = \frac{\Upsilon}{2B}$ yield that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2\right] \lesssim B^2\|f^* - f_0\|_\infty^2 + \frac{B^2}{\Upsilon}\mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\xi,\mathbb{S}}\|_{L_2(P_{\boldsymbol{X}})}^2\right] \\
&+ B\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))].
\end{aligned}$$

Note that $\|\widehat{r}_{\xi,\mathbb{S}}\|_\infty \leq \xi$, then we employ the similar arguments in Lemma 6 and obtain that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y - \widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y - f^*(\boldsymbol{X})))] \\
\leq & \mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\xi^2\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right).
\end{aligned}$$

Set the size $\mathcal{S} = \mathcal{O}(n^{\frac{\delta d}{\delta d+(4+2\delta)\zeta}})$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, weights bound $\mathcal{B} = \mathcal{O}(Bn^{\frac{\delta d}{\delta d+(4+2\delta)\zeta)}})$, and the truncation level $\xi = \mathcal{O}(n^{\frac{2\zeta}{\delta d+(4+2\delta)\zeta}})$. By Theorem 4, we can get

$$\inf_{f\in\mathcal{M}}\|f - f_0\|_\infty^2 \leq \mathcal{O}(Bn^{-\frac{2\delta\zeta}{\delta d+(4+2\delta)\zeta}}).$$

At last, the term $\mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\xi,\mathbb{S}}\|_{L_2(P_{\boldsymbol{X}})}^2\right]$ can be bounded using Theorem 14. Therefore, by combining the three error terms, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|_{L_2(Q_{\boldsymbol{X}})}^2\right] \leq \mathcal{O}\left(B^2n^{-\frac{2\delta\zeta}{\delta d+(4+2\delta)\zeta}}(\log n)^2\right) + \mathcal{O}\left(\frac{B^2m^{-\frac{2\delta\alpha}{\delta d+(6+2\delta)\alpha}}(\log m)^2}{\Upsilon}\right).$$

Moreover, if $m \geq \Omega\Big(n^{\frac{[\delta d+(6+2\delta)\alpha]\zeta}{[\delta d+(4+2\delta)\zeta]\alpha}}\Big)$, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2 n^{-\frac{2\delta\zeta}{\delta d+(4+2\delta)\zeta}}(\log n)^2\Big).$$

If we further assume that the square of the density ratio r^2 is sub-exponential with respect to $P_{\boldsymbol{X}}$, i.e., $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{X}}}\big[\exp(\sigma r^2(\boldsymbol{X}))\big] < \infty$, for some positive constant σ, then

$$\begin{aligned}&\mathbb{E}_{\mathbb{S},\mathbb{D}}\mathbb{E}_{(\boldsymbol{X},Y)\sim P_{\boldsymbol{X},Y}}[\widehat{r}_{\xi,\mathbb{S}}(\boldsymbol{X})(\rho_\tau(Y-\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}}(\boldsymbol{X})) - \rho_\tau(Y-f^*(\boldsymbol{X})))]\\ \leq&\mathcal{O}\left(\frac{B\mathcal{S}\mathcal{D}\xi^2\log(n\mathcal{B}\mathcal{W}\mathcal{D})}{n}\right) \leq \mathcal{O}\left(Bn^{-\frac{2\zeta}{d+2\zeta}}(\log n)^3\log m\right),\end{aligned}$$

with the size $\mathcal{S} = \mathcal{O}(n^{\frac{d}{d+2\zeta}}\log n)$, the depth $\mathcal{D} = \mathcal{O}(\log n)$, weights bound $\mathcal{B} = \mathcal{O}(Bn^{\frac{d}{d+2\zeta}})$, and the truncation level $\xi = \mathcal{O}(\sqrt{\log m})$. By Theorem 4, we can get

$$\inf_{f\in\mathcal{M}}\|f - f_0\|^2_\infty \leq \mathcal{O}(Bn^{-\frac{2\zeta}{d+2\zeta}}).$$

At last, the term $\mathbb{E}_{\mathbb{S}}\left[\|r - \widehat{r}_{\xi,\mathbb{S}}\|^2_{L_2(P_{\boldsymbol{X}})}\right]$ can be bounded using Theorem 17. Therefore, by combining the three error terms, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\left[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\right] \leq \mathcal{O}\left(B^2 n^{\frac{-2\zeta}{d+2\zeta}}(\log n)^3\log m\right) + \mathcal{O}\left(\frac{B^2 m^{-\frac{2\alpha}{d+2\alpha}}(\log m)^{9/2}}{\Upsilon}\right).$$

Moreover, if $m \geq \Omega\Big(n^{\frac{(d+2\alpha)\zeta}{(d+2\zeta)\alpha}}\Big)$, we have

$$\mathbb{E}_{\mathbb{S},\mathbb{D}}\big[\|\widehat{f}_{\widehat{r}_{\xi,\mathbb{S}},\mathbb{D}} - f_0\|^2_{L_2(Q_{\boldsymbol{X}})}\big] \leq \mathcal{O}\Big(B^2 n^{-\frac{2\zeta}{d+2\zeta}}(\log n)^4\Big).$$

This completes the proof. ■

Appendix B. Additional Numerical Experiments

In this part, we provide some additional numerical results. Specifically, the generating scheme is the same as that the generating scheme is the same as that in Section 5.1, except that we consider two different scenarios where the covariates are drawn from different distribution families. The first case is that the target covariate is drawn from $N(0,1)$, the source covariate is drawn from the standard Cauchy distribution, and clearly, the density ratio is uniformly bounded under this case. The second case is that the target covariate is drawn from a Pareto distribution with scale parameter 0.2 and shape parameter 2, the source covariate is drawn from Student's t-distribution with 3 degrees of freedom, and clearly, the density ratio is unbounded density ratios with the bounded second moment. The averaged performances of all the estimators are summarized in terms of L_1 and L_2^2 prediction errors under different scenarios in Tables 5 and 6.

It is thus clear from Tables 5 and 6 that the obtained numerical results are consistent with those presented in Section 5. Under the uniformly bounded case, the averaged errors

Table 5: Averaged L_1 and L_2^2 errors ($\times 10^{-1}$) based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for Model (13) with the bounded density ratio.

	Sample size	$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	1.374(0.169)	0.246(0.062)	1.102(0.104)	0.153(0.027)
	WDQR	1.373(0.192)	0.246(0.068)	1.092(0.093)	0.151(0.022)
	PWDQR	1.428(0.232)	0.254(0.091)	1.195(0.194)	0.162(0.045)
0.25	DQR	0.584(0.080)	0.049(0.013)	0.540(0.049)	0.044(0.007)
	WDQR	0.562(0.071)	0.050(0.012)	0.539(0.047)	0.044(0.007)
	PWDQR	0.598(0.094)	0.053(0.032)	0.570(0.072)	0.049(0.012)
0.5	DQR	0.449(0.023)	0.031(0.004)	0.421(0.010)	0.028(0.002)
	WDQR	0.446(0.025)	0.032(0.004)	0.419(0.013)	0.027(0.002)
	PWDQR	0.468(0.043)	0.034(0.009)	0.429(0.031)	0.029(0.005)
0.75	DQR	0.607(0.079)	0.055(0.013)	0.555(0.044)	0.043(0.007)
	WDQR	0.605(0.079)	0.055(0.012)	0.553(0.050)	0.044(0.008)
	PWDQR	0.613(0.083)	0.057(0.031)	0.571(0.063)	0.047(0.015)
0.95	DQR	1.333(0.154)	0.215(0.043)	1.044(0.094)	0.137(0.021)
	WDQR	1.334(0.180)	0.217(0.054)	1.047(0.091)	0.141(0.022)
	PWDQR	1.459(0.213)	0.233(0.078)	1.121(0.176)	0.151(0.045)

Table 6: Averaged L_1 and L_2^2 errors ($\times 10^{-1}$) based on testing data with the corresponding standard deviations in brackets for DQR, WDQR and PWDQR for Model (13) with the unbounded density ratio.

	Sample size	$n = 512$		$n = 2048$	
τ	Method	L_1	L_2^2	L_1	L_2^2
0.05	DQR	1.332(0.184)	0.305(0.262)	1.089(0.112)	0.157(0.034)
	WDQR	1.205(0.344)	0.184(0.112)	0.987(0.109)	0.124(0.023)
	PWDQR	1.223(0.210)	0.189(0.052)	1.009(0.143)	0.128(0.029)
0.25	DQR	0.586(0.090)	0.130(0.260)	0.551(0.049)	0.061(0.055)
	WDQR	0.515(0.055)	0.046(0.007)	0.508(0.043)	0.043(0.009)
	PWDQR	0.549(0.102)	0.048(0.015)	0.525(0.060)	0.043(0.008)
0.5	DQR	0.460(0.039)	0.112(0.260)	0.423(0.159)	0.044(0.057)
	WDQR	0.424(0.021)	0.031(0.007)	0.409(0.011)	0.028(0.006)
	PWDQR	0.433(0.029)	0.033(0.006)	0.418(0.034)	0.031(0.009)
0.75	DQR	0.585(0.075)	0.131(0.259)	0.542(0.055)	0.061(0.061)
	WDQR	0.523(0.087)	0.046(0.015)	0.514(0.055)	0.042(0.011)
	PWDQR	0.539(0.103)	0.048(0.016)	0.521(0.059)	0.043(0.012)
0.95	DQR	1.257(0.152)	0.269(0.259)	1.046(0.097)	0.153(0.067)
	WDQR	1.234(0.394)	0.211(0.138)	1.037(0.115)	0.135(0.027)
	PWDQR	1.238(0.219)	0.215(0.056)	1.110(0.353)	0.156(0.053)

of all the three estimators are very similar, even when the source and target distributions belong to different families. Under the unbounded case, where the target distribution has heavier tails than the source distribution, both WDQR and PWDQR significantly outperform DQR. These numerical results further validate our theoretical findings as discussed in Section 4.

References

Martin Anthony, Peter L Bartlett, and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*, volume 9. Cambridge University Press Cambridge, 1999.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2481–2488, 2014.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82 – 130, 2011.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *The Journal of Machine Learning Research*, 10:2137–2155, 2009.

Howard D Bondell, Brian J Reich, and Huixia Wang. Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pages 2686–2708. PMLR, 2022.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, pages 38–53. Springer, 2008.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in Neural Information Processing Systems*, 23, 2010.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 33:11996–12007, 2020.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. Towards a unified analysis of kernel-based methods under covariate shift. *Advances in Neural Information Processing Systems*, 36:73839–73851, 2023.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3:131–160, 2009.

Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):579–594, 2018.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Laszlo Gyorfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer, 2002.

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

Ali Hassan, Robert Damper, and Mahesan Niranjan. On acoustic emotion recognition: compensating for covariate shift. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1458–1468, 2013.

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. *Journaltitle of Nonparametric Statistics*, 3(3-4):299–308, 1994.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19:601–608, 2006.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, 2007.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics*, 51(2):691 – 716, 2023.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Masahiro Kato and Takeshi Teshima. Non-negative Bregman divergence minimization for deep direct density ratio estimation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5320–5333. PMLR, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46:33–50, 1978.

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

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. Risk bounds for quantile trend filtering. *Biometrika*, 109(3):751–768, 2022.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34543–34585. PMLR, 2024.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with ReLU networks: Estimators and minimax rates. *The Journal of Machine Learning Research*, 23(1):11251–11292, 2022a.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with ReLU networks: Estimators and minimax rates. *The Journal of Machine Learning Research*, 23(1):11251–11292, 2022b.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. Mit Press, 2008.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 547–562, 1961.

Maxime Sangnier, Olivier Fercoq, and Florence d'Alché Buc. Joint quantile regression in vector-valued RKHSs. *Advances in Neural Information Processing Systems*, 29:3700–3708, 2016.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Guohao Shen, Yuling Jiao, Yuanyuan Lin, Joel L Horowitz, and Jian Huang. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. MIT press, 2012.

Masashi Sugiyama and Klaus-Robert Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (35):985–1005, 2007a.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems*, 20:1433–1440, 2007b.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.

Ichiro Takeuchi, Quoc Le, Timothy Sears, Alexander Smola, et al. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264, 2006.

Yong Kiam Tan, Xinxing Xu, and Yong Liu. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 17–22, 2016.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

AW van der Vaart and Jon A Wellner. *Empirical Processes*. Springer, 2023.

Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Ramon van Handel. *Probability in High Dimension*. APC 550 Lecture Notes, Princeton University. 12 2016.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Huixia Judy Wang, Deyuan Li, and Xuming He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500): 1453–1464, 2012.

Halbert White. Nonparametric estimation of conditional quantiles using neural networks. In *Computing Science and Statistics*, pages 190–199. Springer, 1992.

Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, and Peng Cui. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 24803–24829. PMLR, 2022.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.