# On Convergence Rates of Deep Nonparametric Regression under Covariate Shift

**Chenguang Duan** cgduan.math@whu.edu.cn
**Yuling Jiao** yulingjiaomath@whu.edu.cn
**Lican Kang** kanglican@whu.edu.cn
**Xiliang Lu** xllv.math@whu.edu.cn
**Jerry Zhijian Yang** zjyang.math@whu.edu.cn
*School of Mathematics and Statistics*
*Wuhan University*
*Wuhan, Hubei 430072, China*

## Abstract

Traditional machine learning and statistical modeling methodologies are rooted in a fundamental assumption: that both training and test data originate from the same underlying distribution. However, the practical reality often presents a challenge, as training and test distributions frequently manifest discrepancies or biases. In this work, we study covariate shift, a type of distribution mismatches, in the context of deep nonparametric regression. We thus formulate a two-stage pre-training reweighted framework relying on deep ReLU neural networks. Specifically, in the first pre-training stage, unlabeled data from both the source and target distributions is utilized to estimate the density ratio through deep logistic regression. Following this, a density ratio reweighting strategy is seamlessly integrated into deep nonparametric regression, incorporating the previously estimated density ratio. We rigorously establish convergence rates for the unweighted, reweighted, and pre-training reweighted estimators, illuminating the pivotal role played by the density-ratio reweighting strategy. Additionally, our analysis illustrates the efficacy of pre-training and provides valuable insights for practitioners, offering guidance for the judicious selection of the number of pre-training samples.

**Keywords:** Nonparametric regression, covariate shift, importance weighting, deep learning.

## 1 Introduction

Covariate shift (Quinoñero-Candela et al., 2009; Sugiyama and Kawanabe, 2012), a pervasive phenomenon within the domains of machine learning and statistical modeling, bears notable relevance across diverse domains, including computer vision, natural language processing, and medical image analysis, among others. It distinguishes itself from conventional machine learning and statistical modeling paradigms, where the traditional assumption posits that both training and testing data originate from the same underlying distribution. However, covariate shift manifests during the modeling process when the distribution of training data significantly deviates from that of the testing data. In simpler terms, covariate shift represents a scenario wherein the statistical properties of the data undergo substantial changes between the training and testing phases of a machine learning or statistical model. This phenomenon often leads to a deterioration in the model's generalization capability, as the model has primarily learned patterns from one distribution but is then tasked with making predictions on a distinctly different distribution. Consequently, many researchers have proposed an array of methods to address this intricate issue, seeking to mitigate the adverse effects of covariate shift on model performance. In the work of Kuang et al. (2021), a balanced-subsampled stable prediction algorithm is proposed, which is based on the fractional factorial design strategy. This algorithm is designed to address covariate balancing and ensure stable predictions. Duchi and Namkoong (2021) presented a distributionally robust stochastic optimization framework aimed at mitigating distribution shifts. It leverages the concept of $f$-divergence (Csiszár, 1967) to quantify the magnitude of distribution shift from the training distribution. Subsequently, the framework is implemented by incorporating the empirical distribution plug-in methodology. Krueger et al. (2021) introduced the risk extrapolation method, which performs as a form of robust optimization applied across a perturbation set that encompasses extrapolated domains. Additionally, they have devised a penalty function to tackle the variance in training risks, providing a simplified alternative. Dubois et al. (2021) explored covariate shift by focusing on the acquisition of optimal representations, ensuring that source risk minimizers generalize effectively across distributional shifts. It is paramount to underscore that density-ratio reweighting, referred to as importance weighting (Shimodaira, 2000; Huang et al., 2006; Sugiyama and Storkey, 2006; Sugiyama et al., 2007a,b; Bickel et al., 2007; Sugiyama et al., 2008; Bickel et al., 2009; Kanamori et al., 2009; Fang et al., 2020), emerges as a primary approach for addressing the intricacies of covariate shift. Some researchers have conducted extensive error analysis related to ratio-reweighting, as documented in the works of Cortes et al. (2008, 2010); Xu et al. (2022), while they imposed an upper bound assumption on the density ratio and Cortes et al. (2008, 2010) obtained sub-optimal rates. Recently, Ma et al. (2023) has explored the covariate shift problem within the framework of nonparametric regression over a reproducing kernel Hilbert space (RKHS), and endeavored to provide some theoretical insights. In these studies, the authors assume prior knowledge of the test distribution and directly employ the exact density ratio for the theoretical analysis. However, in practical scenarios, the exact density ratio is often unattainable. In this paper, we propose a pre-training strategy to initially estimate the density ratio using unlabeled data from both the source and target distributions, followed by the derivation of the pre-training reweighted estimator. Furthermore, covariate shift is also closely intertwined with out-of-distribution generalization, transfer learning, domain adaptation, and stable learning. We reserve this discussion for Section 2.

In comparison to RKHS, deep neural networks (DNNs) also stand out as a formidable technique employed within machine learning and statistics. The capabilities of DNNs extend far beyond their superficial appearance, encompassing a wide spectrum of applications and possibilities (Goodfellow et al., 2016). Recently, DNNs have catalyzed a surge of interest in the field of deep nonparametric regression, wherein these neural networks are harnessed to approximate underlying regression functions within the context of nonparametric regression (Stone, 1982; Gyorfi et al., 2002; Tsybakov, 2009). Numerous studies have contributed to our understanding of deep nonparametric regression. Remarkable works by Bauer and Kohler (2019); Schmidt-Hieber (2020); Nakada and Imaizumi (2020); Kohler and Langer (2021); Farrell et al. (2021); Chen et al. (2022); Kohler et al. (2022); Nakada and Imaizumi

(2020); Jiao et al. (2023), and others have illuminated various facets of this burgeoning field, elucidating intricate theoretical properties and unveiling innovative methodologies. However, it is noteworthy that existing literature has thus far overlooked the covariate shift phenomenon inherent in deep nonparametric regression. Building upon this research trajectory, our investigation embarks on a meticulous exploration of the covariate shift phenomenon in nonparametric regression utilizing DNNs. To the best of our knowledge, this work conducts the first effort to uncover and provide robust theoretical guarantees pertaining to the covariate shift phenomenon within the context of deep nonparametric regression.

### 1.1 Contributions

Covariate shift poses a significant challenge in practical applications. One effective strategy for mitigating the impact of covariate shift involves the density-ratio reweighting approach. Despite the empirical success of this strategy, rigorous theoretical analysis for deep nonparametric regression under covariate shift remains scarce. In this paper, we establish convergence rates for the unweighted and reweighted estimators in the case of bounded and unbounded density ratios, respectively. Both cases illustrate the provable advantage of the density ratio reweighting.

We also note that in practical scenarios, the exact density ratio is typically unattainable, leaving us with access to only unlabeled data from both the source and target distributions. To address this challenge, we propose a two-stage neural network-based methodology tailored to this specific scenario. In the initial pre-training stage, we leverage unlabeled data from both the source and target distributions to estimate the density ratio through pseudo-labeled deep logistic regression. Subsequently, we seamlessly integrate a density ratio reweighting strategy with the density ratio previously estimated into the framework of deep nonparametric regression, relying exclusively on labeled data originating from the source distribution. Further, we establish convergence rates for the aforementioned pre-training reweighted estimator under very mild conditions, which provides a comprehensive understanding of the role of pre-training and sheds light on how the sample size in pre-training precisely affects the downstream regression task.

In summary, the main theoretical contributions of this study can be outlined as follows.

(i) We present oracle inequalities for unweighted and reweighted estimators in the case of uniformly upper-bounded density ratio and unbounded density ratio with a finite second moment, aligning with Ma et al. (2023). These inequalities not only consider the generalization error as in the literature (Cortes et al., 2010), but also includes the approximation error. Additionally, the established generalization error bounds exhibit fast rates $\mathcal{O}(1/n)$, demonstrating a substantial enhancement over the slow rate $\mathcal{O}(1/\sqrt{n})$ originally posited by Cortes et al. (2010).

(ii) By strategically balancing approximation and generalization errors, we establish convergence rates for the aforementioned estimators. These obtained convergence rates are with respect to the boundedness or second moment of the density ratio and the sample size, which explicitly account for the impact of the shift between source and target distributions. To the best of our knowledge, we are the first to elucidate the convergence rates of deep nonparametric regression in the presence of covariate shift.

Moreover, comparing these rates between unweighted and reweighted estimators provides strong evidence of the benefits of density ratio reweighting.

(iii) We also propose convergence rates of the pre-training reweighted estimator, which reflect the combined influence of the number of unlabeled samples utilized in the pre-training phase and the number of labeled samples employed in subsequent deep nonparametric regression. Of independent interest, it is noteworthy that the rates of the density ratio estimator achieve minimax optimality. These theoretical results provide valuable insights for determining not only the sample size for nonparametric regression, but also the sample size required for the pre-training procedure.

### 1.2 Main Results

In this section, we expound upon the primary findings of this paper, specifically, the convergence rates pertaining to the unweighted estimator, reweighted estimator, and pre-training reweighted estimator within the context of the target distribution. These results are succinctly summarized in the following Table 1.

Table 1: A summary of convergence rates derived in this paper.

| Estimator | Extra Information | Convergence Rates | |
|---|---|---|---|
| UE[1] | None | $\mathcal{O}(\Lambda n^{-\frac{2\beta}{d+2\beta}})$ | Theorem 9 |
| | | $\mathcal{O}(V n^{-\frac{\beta}{d+2\beta}})$ | Corollary 12 |
| RE[2] | Density ratio | $\mathcal{O}(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}})$ | Theorem 16 |
| | | $\mathcal{O}(V^{\frac{4\beta}{d+4\beta}} n^{-\frac{2\beta}{d+4\beta}})$ | Corollary 21 |
| PT-RE[3] | Unlabeled data | $\mathcal{O}(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}}) + \mathcal{O}((\Lambda^{\frac{2\alpha}{d+2\alpha}} \lambda^{-1} \vee \lambda^{-\frac{d+4\alpha}{d+2\alpha}}) m^{-\frac{2\alpha}{d+2\alpha}})$ | Theorem 26 |

1. Unweighted estimator defined in (3).
2. Reweighted estimator defined in (5) or (7).
3. Pre-training Reweighted estimator defined in (11) and (12).

### 1.3 Preliminaries and Notations

In this section, we introduce the notations utilized throughout this paper and present some essential definitions.

**Notations** Let $P$ be a joint distribution over $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$. Denote by $P_X$ the marginal distribution of $X$ and $P_{Y|X}$ the conditional distribution of $Y$ given $X$. By a similar argument, we can define the marginal distribution $Q_X$ and conditional distribution $Q_{Y|X}$ for some joint distribution $Q$. Then by the definition it holds that $P(X,Y) = P_{Y|X}(Y|X)P_X(X)$ and $Q(X,Y) = Q_{Y|X}(Y|X)Q_X(X)$. The $L^2(P_X)$-norm of a function is defined as $\|f\|_{L^2(P_X)} = (\int_{\mathcal{X}} f(x)^2 dP_X)^{1/2}$. Denote by $\mathscr{L}(\mathcal{X})$ the set of measurable functions on $\mathcal{X}$. We write $A \lesssim B$ if there exists an absolute positive constant $c$ such that $A \leq cB$. We denote by $\mathbb{N} = \{0, 1, \cdots\}$ and $\mathbb{N}_+ = \{1, \cdots\}$.

**Definition 1 (Hölder space)** *Let $\mathcal{X} \subseteq [0,1]^d$ and $\beta = k + \lambda$ with $k \in \mathbb{N}$ and $\lambda \in (0,1]$. The Hölder class $\mathcal{H}^\beta(\mathcal{X})$ consists of all those functions mapping from $\mathcal{X}$ to $\mathbb{R}$ which have* *continuous derivatives up to order $k$ and whose $k$-th partial derivatives is $\lambda$-Hölder continuous on $\mathcal{X}$. Here a function $f : \mathcal{X} \to \mathbb{R}$ is called $\lambda$-Hölder continuous on $\mathcal{X}$, if*

$$|f|_{\mathcal{H}^\lambda(\mathcal{X})} := \sup_{x,y\in\mathcal{X},x\neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2^\lambda} < \infty.$$

*Then the $\mathcal{H}^\beta(\mathcal{X})$-norm is defined as*

$$\|f\|_{\mathcal{H}^\beta(\mathcal{X})} := \max_{\|\boldsymbol{m}\|_1\leq k} \|\partial^{\boldsymbol{m}} f\|_{L^\infty(\mathcal{X})} + \max_{\|\boldsymbol{m}\|_1=k} |\partial^{\boldsymbol{m}} f|_{\mathcal{H}^\lambda(\mathcal{X})},$$

*where $\partial^{\boldsymbol{m}} = \partial^{m_1}\cdots\partial^{m_d}$ with multi-index $\boldsymbol{m} = (m_1,\dots,m_d)^T \in \mathbb{N}^d$.*

**Definition 2 (ReLU DNNs)** *A neural network $\psi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is a function defined by*

$$\psi(x) = T_L(\phi(T_{L-1}(\cdots\phi(T_0(x))\cdots))),$$

*where the ReLU activation function $\phi(x) := \max\{0,x\}$ is applied component-wisely and $T_\ell(x) := A_\ell x + b_\ell$ is an affine transformation with $A_\ell \in \mathbb{R}^{N_{\ell+1}\times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 0,\dots,L$. In this paper, we consider the case $N_0 = d$ and $N_{L+1} = 1$. The numbers $W := \max\{N_1,\dots,N_L\}$ and $L$ are called the width and the depth of neural networks, respectively. Additionally, $S := \sum_{\ell=0}^{L} N_\ell N_{\ell+1} \leq LW^2$ is called the size of the neural network, representing the total number of parameters within the neural network. We denote by $\mathcal{N}(W,L)$ the set of functions implemented by ReLU neural networks with width at most $W$ and depth at most $L$.*

**Definition 3 (VC-dimension)** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\{\pm 1\}$. For any non-negative integer $m$, we define the growth function of $\mathcal{F}$ as*

$$\Pi_{\mathcal{F}}(m) = \max_{\{x_i\}_{i=1}^m\subseteq\mathcal{X}} \big|\{(f(x_1),\dots,f(x_m)) : f\in\mathcal{F}\}\big|.$$

*A set $\mathfrak{X} = \{x_i\}_{i=1}^m$ is said to be shattered by $\mathcal{F}$ when $|\{(f(x_1),\dots,f(x_m)) : f\in\mathcal{F}\}| = 2^m$. The Vapnik-Chervonenkis dimension of $\mathcal{F}$, denoted $\mathrm{VCdim}(\mathcal{F})$, is the size of the largest set that can be shattered by $\mathcal{F}$, that is, $\mathrm{VCdim}(\mathcal{F}) = \max\{m : \Pi_{\mathcal{F}}(m) = 2^m\}$. For a class $\mathcal{F}$ of real-valued functions, we define $\mathrm{VCdim}(\mathcal{F}) = \mathrm{VCdim}(\mathrm{sign}(\mathcal{F}))$.*

### 1.4 Organization

The remainder of this paper is structured as follows. Section 2 presents a review of related topics. Section 3 proposes three distinct estimators: the unweighted estimator, the reweighted estimator, and the pre-training reweighted estimator. Following this, Section 4 conducts an precisely analysis for the convergence rates of these estimators. Section 5 summarizes the conclusions of this work. Section A presents some supplemental definitions and lemmas for subsequent proofs. Comprehensive technical proofs of unweighted, reweighted, and pre-training reweighted estimators are delineated in Sections B, C and D, respectively. Lastly, the proof for the density ratio estimate is presented in Section E.

## 2 Related Topics

In this section, we discuss various research topics related to the covariate shift, prominently encompassing aspects such as out-of-distribution generalization, transfer learning, domain adaptation, and stable learning.

### 2.1 Out-of-Distribution Generalization

The out-of-distribution (OOD) generalization problem, as elucidated by Shen et al. (2021), represents a specific facet of the supervised learning paradigm wherein the test distribution, denoted as $Q$, exhibits a notable divergence from the training distribution, symbolized as $P$. It is imperative to emphasize that the test distribution remains unknown during the entirety of the training phase. Within the framework of OOD generalization, the observed distribution shift can be categorized into two principal types, namely concept shifts (Gama et al., 2014; Cai and Wei, 2021) and covariate shifts (Shimodaira, 2000). Concept shifts manifest as alterations in the conditional distribution, denoted as $P_{Y|X}$, resulting in a mismatch with the test distribution $Q_{Y|X}$. Conversely, covariate shifts are characterized by perturbations in the marginal distribution $P_X$, leading to incongruence with the test distribution $Q_X$.

### 2.2 Transfer Learning and Domain Adaptation

Transductive transfer learning constitutes a pivotal domain within the broader framework of transfer learning, operating under the invariance assumption that $P_{Y|X} = Q_{Y|X}$. Notably, this assumption has been expounded upon by Pan and Yang (2009); Zhuang et al. (2020). Transductive transfer learning can be dissected into two distinct scenarios, each with its unique characteristics. The first scenario pertains to instances where the feature spaces of the source and target distributions differ, i.e., $\mathcal{X}_P \neq \mathcal{X}_Q$. Conversely, the second scenario involves cases wherein the feature spaces in both the source and target domains remain identical, i.e., $\mathcal{X}_P = \mathcal{X}_Q$, while there exist disparities in the marginal probability distributions of the input data, specifically $P_X \neq Q_X$. This latter case is commonly referred to as domain adaptation, predicated on the assumption that prior knowledge regarding the test distribution is available, encompassing either the joint distribution $Q$ or the marginal distribution $Q_X$. This paper primarily addresses the latter scenario of domain adaptation, wherein the alignment of marginal distributions between the source and target domains becomes the focal point of consideration. Furthermore, it is noteworthy that ratio-reweighting emerges as a crucial mechanism within the domain adaptation framework, as acknowledged by Huang et al. (2006); Belkin et al. (2006); Sugiyama et al. (2008); Sun et al. (2011).

### 2.3 Stable Learning

In the context of machine learning, when provided with a training dataset $\{(X_i, Y_i)\}_{i=1}^n$, wherein $\{X_i\}_{i=1}^n$ are sampled from a single distribution defined over the feature space $\mathcal{X}$, the primary objective of stable learning is to formulate an estimator that exhibits consistent and uniformly excellent performance across the entirety of possible distributions over $\mathcal{X}$. To address this formidable challenge, Shen et al. (2020) introduced an important technique known as sample weighting. This technique serves as a pivotal tool in the pursuit of stable learning by assigning appropriate weights to individual data points, thus allowing the learning process to emphasize those instances that contribute significantly to achieving the desired uniformity of performance across diverse distributions over $\mathcal{X}$.

## 3 Problem Formulation

Let $\mathcal{X} \subseteq [0,1]^d$ $(d \geq 1)$ be the feature space and $\mathcal{Y} \subseteq \mathbb{R}$ be the response space. Consider the following nonparametric regression model

$$Y = f_0(X) + \xi, \tag{1}$$

where the response variable $Y \in \mathcal{Y}$ is associated with the covariate $X \in \mathcal{X}$. The underlying regression function is defined as $f_0 : \mathcal{X} \to \mathbb{R}$. Furthermore, $\xi$ represents a random noise term that is independent of $X$ and satisfies the condition $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2] < \infty$. It is obvious that $f_0(x) = \mathbb{E}[Y|X = x]$ for each $x \in \mathcal{X}$.

In the context of nonparametric regression, our focus centers on the observation of $n$ independent and identically distributed (i.i.d.) random pairs denoted as $\mathcal{D} := \{(X_i^P, Y_i^P)\}_{i=1}^n$. These pairs are drawn from a training distribution $P$ defined over $\mathcal{X} \times \mathcal{Y}$. Specifically, $\{X_i^P\}_{i=1}^n$ are sampled from $P_X$, and the conditional probability $P_{Y|X}$ is determined in accordance with (1). We introduce another distribution $Q$ as the test distribution over $\mathcal{X} \times \mathcal{Y}$. More precisely, the covariates of test samples are derived from $Q_X$, and the corresponding responses are also generated following the model (1), implying $Q_{Y|X} = P_{Y|X}$.

Within this context, $P_X$ is referred to as the source distribution for covariates, while $Q_X$ serves as the target distribution for covariates. In practical scenarios, it is common for the target distribution $Q_X$, on which a model is deployed, to exhibit divergence from the source distribution $P_X$. This phenomenon is commonly referred to as covariate shifts, and it can be attributed to various factors, including temporal or spatial data evolution or inherent biases introduced during the data collection process.

Our primary objective revolves around the development of an estimator denoted as $\widehat{f}$, which is constructed based on the observed data. This estimator is carefully designed with the primary aim of minimizing the $L^2(Q_X)$-risk, as defined below:

$$\|\widehat{f} - f_0\|_{L^2(Q_X)}^2 = \mathbb{E}_{X \sim Q_X}\left[(\widehat{f}(X) - f_0(X))^2\right] = \int_{\mathcal{X}} (\widehat{f}(x) - f_0(x))^2 q_X(x)dx, \tag{2}$$

where $q_X$ represents the probability density function associated with $Q_X$. In essence, our goal is to minimize this risk, which quantifies the expected squared difference between the estimator $\widehat{f}$ and the underlying regression function $f_0$ over the distribution $Q_X$.

### 3.1 Unweighted Estimators

When confronted with a situation where information about the target distribution is unavailable, a natural approach for constructing an estimator is to directly minimize the unweighted empirical risk over a hypothesis class $\mathcal{F}$, representing a set of measurable functions. This estimator, denoted as $\widehat{f}_{\mathcal{D}}$, is determined as follows:

$$\widehat{f}_{\mathcal{D}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \widehat{L}_{\mathcal{D}}(f) := \frac{1}{n} \sum_{i=1}^n (f(X_i^P) - Y_i^P)^2. \tag{3}$$

It is important to note that $\widehat{L}_{\mathcal{D}}(f)$ serves as a sample average approximation to the unweighted population risk $L(f)$, expressed as:

$$L(f) := \mathbb{E}_{(X^P,Y^P)\sim P}\big[(f(X^P) - Y^P)^2\big].$$

According to (1), it is straightforward to verify that $L(\widehat{f}_{\mathcal{D}}) = \|\widehat{f}_{\mathcal{D}} - f_0\|^2_{L^2(P_X)} + \mathbb{E}[\xi^2]$, which means that the minimizer of the unweighted population risk concurrently minimizes the $L^2(P_X)$-risk, as opposed to $L^2(Q_X)$-risk defined in (2).

### 3.2 Reweighted Estimators

When we have knowledge of the target distribution, a direct approach is available for minimizing the population risk with the exact weight. To facilitate this, we introduce the concept of the density ratio, as defined for the target distribution $Q_X$ and the source distribution $P_X$. We denote by $p_X$ the probability density function of $P_X$, then this density ratio, denoted as $\varrho(x) := q_X(x)/p_X(x)$, is also referred to as the importance weight (Cortes et al., 2010). It is worth noting that the density ratio $\varrho(\cdot)$ measures the discrepancy between $Q_X$ and $P_X$, which has been discussed in Cortes et al. (2010).

With the exact density ratio $\varrho$ at our disposal, our objective turns to minimizing the population reweighted risk, defined as:

$$L_{\varrho}(f) := \mathbb{E}_{(X^P,Y^P)\sim P}\big[\varrho(X^P)(f(X^P) - Y^P)^2\big]. \tag{4}$$

Minimizing this population reweighted risk is equivalent to minimizing the $L^2(Q_X)$-risk, that is, $L_{\varrho}(f) = \|f - f_0\|^2_{L^2(Q_X)} + \mathbb{E}[\xi^2]$.

Once the density ratio is uniformly upper-bounded, that is, $\|\varrho\|_{L^\infty(\mathcal{X})} < \infty$, a commonly employed approach to achieve this is through empirical reweighted risk minimization within a hypothesis class $\mathcal{F}$, resulting in the following estimator:

$$\widehat{f}_{\varrho,\mathcal{D}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{L}_{\varrho,\mathcal{D}}(f) := \frac{1}{n}\sum_{i=1}^{n} \varrho(X_i^P)(f(X_i^P) - Y_i^P)^2. \tag{5}$$

This approach leverages the density ratio $\varrho$ to reweight the contributions of individual samples in the empirical risk minimization, effectively adapting the learning process to account for the covariate shift.

In the case where the density ratio is unbounded, the empirical risk in (5) may diverge for some data sets. To address this issue, we introduce the truncation operator (Ma et al., 2023). For a positive truncation parameter $\beta_n$ depending upon $n$, denote by $T_{\beta_n}\varrho$ the truncated density ratio

$$T_{\beta_n}\varrho(x) = \begin{cases} \varrho(x), & \text{if } \varrho(x) \le \beta_n, \\ \beta_n, & \text{otherwise.} \end{cases} \tag{6}$$

Then the reweighted estimator with truncated density ratio is defined as

$$\widehat{f}_{T_{\beta_n}\varrho,\mathcal{D}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{L}_{T_{\beta_n}\varrho,\mathcal{D}}(f) := \frac{1}{n}\sum_{i=1}^{n} T_{\beta_n}\varrho(X_i^P)(f(X_i^P) - Y_i^P)^2, \tag{7}$$

where the truncation level $\beta_n$ is a hyper-parameter to be specified.

### 3.3 Pre-training Reweighted Estimators

However, the accessibility of exact density ratio frequently confronts inherent limitations in practical applications. Nonetheless, it is pertinent to note that a pragmatic solution presents itself in the form of deriving estimations for these density ratio functions. This can be achieved through the utilization of unlabeled data from both the source and target distributions. The methodology underpinning this estimation process aligns with the well-defined principles expounded upon in the following lemma.

**Lemma 4** *Let $p_X$ and $q_X$ be two probability density functions on $\mathcal{X}$. Then the density ratio $\varrho$ is given by $\varrho(x) = q_X(x)/p_X(x) = \exp(-u^*(x))$, where the function $u^*$ satisfies*

$$u^* = \operatorname*{argmin}_{u \in \mathscr{L}(\mathcal{X})} \Big\{ \mathbb{E}_{X\sim P_X}\big[\log(1+\exp(-u(X)))\big] + \mathbb{E}_{X\sim Q_X}\big[\log(1+\exp(u(X)))\big] \Big\}. \tag{8}$$

**Remark 5 (Pseudo-labels)** *Let $X^P$ and $X^Q$ be random variables distributed from $P_X$ and $Q_X$, respectively. We assign a pseudo-label $Z^P = +1$ for $X^P$ and $Z^Q = -1$ for $X^Q$, and construct a random variable pair $(X^\mu, Z^\mu)$ by*

$$\begin{cases} X^\mu = \epsilon X^P + (1-\epsilon) X^Q, \\ Z^\mu = \epsilon Z^P + (1-\epsilon) Z^Q, \end{cases} \tag{9}$$

*where $\epsilon$ is a random variable satisfying $\Pr\{\epsilon = 1\} = \Pr\{\epsilon = 0\} = 1/2$ and is independent of $X^P$, $Z^P$, $X^Q$, $Z^Q$. We denote by $\mu$ the joint distribution of $(X^\mu, Z^\mu)$ in (9), then the population logistic risk can be given by $L_{\mathrm{logit}}(u) = \mathbb{E}_{(X^\mu,Z^\mu)\sim\mu} \log(1+\exp(-Z^\mu u(X^\mu)))$, which is the objective function in (8).*

As shown in Lemma 4, we define the population pre-training risk as

$$L^{\mathrm{pre}}(u) := \mathbb{E}_{X^P\sim P_X}\big[\log(1+\exp(-u(X^P)))\big] + \mathbb{E}_{X^Q\sim Q_X}\big[\log(1+\exp(u(X^Q)))\big].$$

Let $\mathcal{S}_P := \{X_i^P\}_{i=1}^m$ and $\mathcal{S}_Q := \{X_i^Q\}_{i=1}^m$ represent the collections of unlabeled samples originating from $P_X$ and $Q_X$, respectively. We proceed by assigning pseudo-labels $Z_i^P = +1$ for $X_i^P$ and $Z_i^Q = -1$ for $X_i^Q$. Accordingly, we can construct a pseudo-labeled sample set $\mathcal{S} := \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ following the expressions:

$$\begin{cases} X_i^\mu = \epsilon_i X_i^P + (1-\epsilon_i) X_i^Q, \\ Z_i^\mu = \epsilon_i Z_i^P + (1-\epsilon_i) Z_i^Q, \end{cases} \tag{10}$$

where $\{\epsilon_i\}_{i=1}^m$ are random variables satisfying $\Pr\{\epsilon_i = 1\} = \Pr\{\epsilon_i = 0\} = 1/2$. Consequently, the empirical pre-training risk $\widehat{L}_{\mathcal{S}}^{\mathrm{pre}}$ is formulated as:

$$\widehat{L}_{\mathcal{S}}^{\mathrm{pre}}(u) := \frac{1}{m}\sum_{i=1}^m \log(1+\exp(-Z_i^\mu u(X_i^\mu))).$$

It is straightforward to verify that $\mathbb{E}_{\mathcal{S}} \widehat{L}_{\mathcal{S}}^{\mathrm{pre}}(u) = L^{\mathrm{pre}}(u)$ for each fixed $u \in \mathscr{L}(\mathcal{X})$. Subsequently, the minimization of the empirical pre-training risk $\widehat{L}_{\mathcal{S}}^{\mathrm{pre}}$ within a given hypothesis class $\mathcal{U}$ yields the following density ratio estimator:

$$\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}}), \quad \text{where } \widehat{u}_{\mathcal{S}} \in \operatorname*{argmin}_{u\in\mathcal{U}} \widehat{L}_{\mathcal{S}}^{\mathrm{pre}}(u). \tag{11}$$

By substituting the weight function $\varrho$ in (4) and (5) with $\widehat{\varrho}_{\mathcal{S}}$, we derive the population pre-training reweighted risk, defined as follows:

$$L_{\widehat{\varrho}_{\mathcal{S}}}(f) := \mathbb{E}_{(X^P,Y^P)\sim P}\big[\widehat{\varrho}_{\mathcal{S}}(X^P)(f(X^P)-Y^P)^2\big].$$

Then, the pre-training reweighted estimator is formulated as follows:

$$\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{L}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(f) := \frac{1}{n}\sum_{i=1}^{n} \widehat{\varrho}_{\mathcal{S}}(X_i^P)(f(X_i^P)-Y_i^P)^2. \tag{12}$$

We notice that in the first stage, $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$, is independent of the second stage data, $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$. This independence between the two stages is noteworthy.

In summary, we present the pre-training reweighted algorithm designed for regression tasks under the influence of covariate shift, as outlined in Algorithm 1. This algorithm is structured into two stages, each serving a crucial role in mitigating the challenges posed by covariate shift. The first stage, denoted as the unsupervised pre-training stage, involves the generation of pseudo-labels $\{Z_i^\mu\}_{i=1}^m$ for unlabeled data $\{X_i^\mu\}_{i=1}^m$, as per the methodology detailed in (10). Additionally, in this stage, we estimate the density ratio $\varrho$ by employing logistic regression on the unlabeled data augmented with these pseudo-labels. In the second stage, the supervised phase, we employ the pre-trained density ratio $\widehat{\varrho}_{\mathcal{S}}$ in conjunction with the labeled dataset $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ to estimate the underlying regression function $f_0$. This multi-stage approach encapsulates the essence of the pre-training reweighted algorithm, which strategically combines unsupervised and supervised learning paradigms to address the challenges posed by covariate shift in regression tasks.

**Algorithm 1** Two-stage life-cycle of pre-training reweighted regression.

**Input:** Unlabeled data for pre-training

- $\mathcal{S}_P = \{X_i^P\}_{i=1}^m$: Unlabeled data sampled from the source distribution $P_X$.
- $\mathcal{S}_Q = \{X_i^Q\}_{i=1}^m$: Unlabeled data sampled from the target distribution $Q_X$.

1: **Pre-training: Estimate density ratio.**
2: Construct pseudo-labeled sample set $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ by (10).
3: Minimize the empirical logistic risk based on $\mathcal{S}$:

$$\widehat{u}_{\mathcal{S}} \in \operatorname*{argmin}_{u\in\mathcal{U}} \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-Z_i^\mu u(X_i^\mu))).$$

**Output:** The density ratio estimator $\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}})$.

**Input:** The density ratio estimator $\widehat{\varrho}_{\mathcal{S}}$ and labeled data $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n \sim P^n$.

4: **Reweighted regression.**
5: Minimize the empirical pre-training reweighted risk based on $\mathcal{D}$:

$$\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \widehat{\varrho}_{\mathcal{S}}(X_i^P)(f(X_i^P)-Y_i^P)^2.$$

**Output:** The pre-training reweighted estimator $\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}$.

## 4 Convergence Rates Analysis

Up to now, we have introduced three estimators: unweighted estimator (3), reweighted estimator (5) or (7), and pre-training reweighted estimator (11)-(12). In this section, we aim to rigorously analyze the performance of these three estimators by addressing the following fundamental questions:

> *What are convergence rates of three estimators: unweighted estimator, reweighted estimator, and pre-training reweighted estimator in the presence of covariate shift? How do these convergence rates depend on the discrepancy between the source and target distribution?*

By focusing on these questions, we shall shed light on the provable advantage of ratio-reweighting and characterize the role of pre-training in deep nonparametric regression under covariate shift.

Subsequently, we introduce several fundamental assumptions. These assumptions encapsulate a sub-Gaussian constraint on the noise term $\xi$ in (1) (Assumption 1), a boundedness assumption applied to both the hypothesis $\mathcal{F}$ and the target function $f_0$ (Assumption 2), with the latter presumed to exhibit Hölder continuity (Assumption 3).

**Assumption 1 (Sub-Gaussian noise)** *The noise $\xi$ in (1) is a sub-Gaussian random variable with mean 0 and finite variance proxy $\sigma^2$, that is, its moment generating function satisfies*

$$\mathbb{E}[\exp(s\xi)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall\ s \in \mathbb{R}.$$

**Assumption 2 (Bounded hypothesis)** *There exists an absolute positive constant $B$, such that $\|f_0\|_{L^\infty(\mathcal{X})} \leq B$. Further, functions in hypothesis class $\mathcal{F}$ are also bounded, that is,* $\sup\{\|f\|_{L^\infty(\mathcal{X})} : f \in \mathcal{F}\} \leq B$.

**Assumption 3 (Hölder's continuity)** *The regression function $f_0$ in* (1) *is Hölder continuous, that is, $f_0 \in \mathcal{H}^\beta(\mathcal{X})$ for some $\beta > 0$.*

**Remark 6** *Assumptions 1, 2 and 3 are standard and very mild conditions in nonparametric regression, as extensively discussed in the literature (Stone, 1982; Gyorfi et al., 2002; Tsybakov, 2009; Bauer and Kohler, 2019; Nakada and Imaizumi, 2020; Schmidt-Hieber, 2020; Kohler and Langer, 2021; Farrell et al., 2021; Kohler et al., 2022; Jiao et al., 2023). It is worth noting that the upper-bound $B$ of functions in hypothesis class may be arbitrarily large and does not vary with the sample size $n$. In fact, this assumption can be removed through the technique of truncation, without affecting the subsequent proof, which can be found in Bauer and Kohler (2019); Kohler and Langer (2021); Kohler et al. (2022) for details.*

As discussed in Cortes et al. (2010); Ma et al. (2023), the density ratio reflects the discrepancy between the source and target distributions. In this paper, we consider cases that the density ratio is uniformly upper-bounded (Assumptions 4) or, alternatively, possesses a finite second moment (Assumptions 5).

**Assumption 4 (Density ratio with uniform upper-bound)** *The density ratio $\varrho$ has a finite uniform upper bound, that is, $\Lambda := \sup_{x\in\mathcal{X}} \varrho(x) < \infty$.*

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

**Remark 7** *Notice that the equality $V^2 = \chi^2(Q_X||P_X) + 1$ shows that the second moment of the density ratio is equal to $\chi^2$-divergence up to an absolute constant. Consequently, Assumption 5 is rendered equivalent to positing the finiteness of the $\chi^2$-divergence between the distributions $Q_X$ and $P_X$. Additionally, when Assumption 4 holds, it is easy to verify that Assumption 5 also holds with $V^2 = \mathbb{E}_{X\sim Q_X}[\varrho(X)] \leq \Lambda$.*

Assumption 4 introduces a constraint by enforcing a uniform upper-bound on the density ratio, a condition frequently explored in covariate shift studies (Cortes et al., 2010; Ma et al., 2023). However, this condition may not hold in practice. In this study, in the convergence analysis for unweighted and reweighted estimators, we extend our scrutiny to situations wherein the density ratio has a finite second moment (Assumptions 5), thereby incorporating a more comprehensive mathematical perspective into our exploration of the error analysis. The unbounded density ratio in importance weighting has been studies in Cortes et al. (2010); Ma et al. (2023), and our extension aligns with the methodology presented in the latter.

We can now proceed to establish convergence rates for these three estimators, as detailed in Sections 4.1, 4.2 and 4.3. This comprehensive exploration illuminates the intricate dynamics of these estimators within the context of covariate shift, providing valuable insights into their performance and utility.

### 4.1 Unweighted Estimators

In this section, we consider convergence rates for the unweighted estimator (3). This encompasses two scenarios: one with a bounded density ratio and the other with an unbounded but second-moment-finite density ratio.

#### 4.1.1 Bounded density ratio

We first present an oracle inequality for the unweighted estimator in the following lemma.

**Lemma 8 (Oracle inequality of unweighted estimator)** *Suppose Assumptions 1, 2 and 4 hold. Let $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be an i.i.d. sample set drawn from $P$. Suppose that $\mathcal{F}$ is a hypothesis class and $\widehat{f}_{\mathcal{D}} \in \mathcal{F}$ is defined by (3). Then the following inequality holds for $n \geq \mathrm{VCdim}(\mathcal{F})$,*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\mathcal{D}} - f_0\|^2_{L^2(Q_X)}\Big] \lesssim \Lambda \inf_{f\in\mathcal{F}} \|f - f_0\|^2_{L^2(P_X)} + \Lambda(B^2 + \sigma^2)\frac{\mathrm{VCdim}(\mathcal{F})}{n\log^{-1}(en)}.$$

With Lemma 8, we set the hypothesis class $\mathcal{F}$ to be a ReLU DNN with a certain structure, leading to the derivation of the convergence rate for the unweighted estimator. This result is formally stated in the following theorem.

**Theorem 9 (Convergence rates of unweighted estimator)** *Suppose Assumptions 1, 2, 3 and 4 hold. Assume that $P_X$ is absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be an i.i.d. sample set drawn from $P$. Set the hypothesis class $\mathcal{F} = \mathcal{N}(W_{\mathcal{F}}, L_{\mathcal{F}})$ with width $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}} \log U_{\mathcal{F}})$ and depth $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}} \log N_{\mathcal{F}})$ satisfying $U_{\mathcal{F}} N_{\mathcal{F}} = \mathcal{O}(n^{\frac{d}{2d+4\beta}})$. Suppose $\widehat{f}_{\mathcal{D}} \in \mathcal{F}$ is defined by (3). Then the following inequality holds:*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \leq \mathcal{O}\Big(\Lambda n^{-\frac{2\beta}{d+2\beta}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 10 (Consistency)** *As shown in Theorem 9, the convergence rate is established as $\mathcal{O}(\Lambda n^{-\frac{2\beta}{d+2\beta}})$. This result demonstrates that the $L^2(\mathcal{X})$-risk of the unweighted estimator $\widehat{f}_{\mathcal{D}}$ is consistent in the sense that $\mathbb{E}\|\widehat{f}_{\mathcal{D}} - f_0\|_{L^2(Q_X)}^2 \to 0$ as $n \to \infty$, regardless the $L^\infty(\mathcal{X})$-norm bound $\Lambda$ on the density ratio.*

**Remark 11** *According to the definition of the density ratio, when $\Lambda = 1$, the scenario simplifies to the standard nonparametric regression. In this case, the rate $\mathcal{O}(n^{-\frac{2\beta}{d+2\beta}})$ aligns with the minimax optimal rate within nonparametric regression, as established in Stone (1982); Yang and Barron (1999); Gyorfi et al. (2002); Tsybakov (2009). Additionally, our theoretical findings correspond to those in deep nonparametric regression (Schmidt-Hieber, 2020; Kohler and Langer, 2021; Jiao et al., 2023). It is worth noting that our approach differs from the aforementioned literature in terms of the proof methodology employed in statistical error analysis. Specifically, we derive the statistical error primarily based on the offset Rademacher complexity (Liang et al., 2015). For a comprehensive elucidation of our methodology and detailed insights, please refer to Section B.*

#### 4.1.2 Unbounded density ratio

In the following corollary, we extend the theoretical findings in Section 4.1.1 to the case of unbounded density ratio.

**Corollary 12** *Suppose Assumptions 1, 2, 3 and 5 hold. Assume that $P_X$ is absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be an i.i.d. sample set drawn from $P$. Set the hypothesis class $\mathcal{F} = \mathcal{N}(W_{\mathcal{F}}, L_{\mathcal{F}})$ with width $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}} \log U_{\mathcal{F}})$ and depth $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}} \log N_{\mathcal{F}})$ satisfying $U_{\mathcal{F}} N_{\mathcal{F}} = \mathcal{O}(n^{\frac{d}{2d+4\beta}})$. Suppose $\widehat{f}_{\mathcal{D}} \in \mathcal{F}$ is defined by (3). Then the following inequality holds:*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \leq \mathcal{O}\Big(V n^{-\frac{\beta}{d+2\beta}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 13** *Similar to the rate of convergence for bounded density ratio (Theorem 9), the rate of convergence in Corollary 12 shows that the $L^2(\mathcal{X})$-risk of the unweighted estimator $\widehat{f}_{\mathcal{D}}$ is also consistent for unbounded density ratio. However, this rate of convergence is not minimax optimal with respect to $n$.*

### 4.2 Reweighted Estimators

In this section, we give convergence rates of the reweighted estimator (5), including both scenarios of bounded density ratios and unbounded ones with finite second moments.

#### 4.2.1 Bounded density ratio

The following lemma provides an oracle inequality for the reweighted estimator.

**Lemma 14 (Oracle inequality of reweighted estimator)** *Suppose Assumptions 1, 2 and 4 hold. Let* $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ *be an i.i.d. sample set drawn from* $P$*. Suppose that* $\mathcal{F}$ *is a hypothesis class and* $\widehat{f}_{\varrho,\mathcal{D}} \in \mathcal{F}$ *is defined by* (5)*. Then the following inequality holds for* $n \geq \mathrm{VCdim}(\mathcal{F})$*,*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\varrho,\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \lesssim \inf_{f\in\mathcal{F}} \|f - f_0\|_{L^2(Q_X)}^2 + \Lambda(B^2 + \sigma^2)\frac{\mathrm{VCdim}(\mathcal{F})}{n\log^{-1}(en)}.$$

**Remark 15** *The oracle inequality in Lemma 14 decomposes the* $L^2(Q_X)$*-risk into two parts: the first term is called the approximation error and the second one is the generalization error. Recall the generalization error bound derived in Cortes et al. (2010, Theorem 2)*

$$\sup_{f\in\mathcal{F}} \big\{L_\varrho(f) - \widehat{L}_{\varrho,\mathcal{D}}(f)\big\} \lesssim C\Big\{\frac{\Lambda\,\mathrm{VCdim}(\mathcal{F})}{m} + \sqrt{\frac{V^2\,\mathrm{VCdim}(\mathcal{F})}{m}}\Big\},$$

*where* $C$ *is a positive constant independent of* $\Lambda$*,* $V$*,* $m$ *and* $\mathrm{VCdim}(\mathcal{F})$*, and we ignore the logarithmic factors. It is apparent that our generalization error bound attaches the fast rate and is tighter than Cortes et al. (2010, Theorem 2).*

We now present the convergence rate for the reweighted estimator by specifying the hypothesis class $\mathcal{F}$ as a ReLU DNN with a certain structure.

**Theorem 16 (Convergence rates of reweighted estimator)** *Suppose Assumptions 1, 2, 3 and 4 hold. Assume that* $Q_X$ *is absolutely continuous with respect to the Lebesgue measure. Let* $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ *be an i.i.d. sample set drawn from* $P$*. Set the hypothesis class* $\mathcal{F} = \mathcal{N}(W_\mathcal{F}, L_\mathcal{F})$ *with width* $W_\mathcal{F} = \mathcal{O}(U_\mathcal{F}\log U_\mathcal{F})$ *and depth* $L_\mathcal{F} = \mathcal{O}(N_\mathcal{F}\log N_\mathcal{F})$ *satisfying* $U_\mathcal{F}N_\mathcal{F} = \mathcal{O}(\Lambda^{-\frac{d}{2d+4\beta}} n^{\frac{d}{2d+4\beta}})$*. Suppose* $\widehat{f}_{\varrho,\mathcal{D}} \in \mathcal{F}$ *is defined by* (5)*. Then the following inequality holds:*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\varrho,\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \leq \mathcal{O}\Big(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 17** *The convergence rate* $\mathcal{O}(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}})$ *of the reweighted estimator, as derived in Theorem 16, is notably tighter than the rate* $\mathcal{O}(\Lambda n^{-\frac{2\beta}{d+2\beta}})$ *of the unweighted estimator from Theorem 9. This discrepancy underscores the theoretical advantage of employing density-ratio reweighting in nonparametric regression under covariate shift.*

#### 4.2.2 Unbounded density ratio

To further address the scenario of unbounded density ratios, in the following lemma, we relate the $L^2(Q_X)$-risk with the $L^2(P_X)$-risk weighted by the truncated density ratio.

**Lemma 18 (Truncation error)** *Let $\beta_n$ be a positive truncation level. Suppose Assumptions 2 and 5 hold. Then for each $f \in \mathcal{F}$ the following inequality holds:*

$$\|f - f_0\|^2_{L^2(Q_X)} \leq \mathbb{E}_{X^P \sim P_X}\left[T_{\beta_n}\varrho(X^P)(f(X^P) - f_0(X^P))^2\right] + \frac{8B^2V^2}{\beta_n}.$$

With the aid of Lemmas 14 and 18, we can obtain the following oracle inequality for the reweighted estimator with a truncated density ratio.

**Corollary 19** *Suppose Assumptions 1, 2 and 5 hold. Let $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be an i.i.d. sample set drawn from $P$. Suppose that $\mathcal{F}$ is a hypothesis class and $\widehat{f}_{T_{\beta_n}\varrho,\mathcal{D}} \in \mathcal{F}$ is an empirical truncated risk minimizer defined by (7) with the truncation level*

$$\beta_n = \sqrt{\frac{V^2B^2}{B^2+\sigma^2}\frac{n}{\mathrm{VCdim}(\mathcal{F})}}.$$

*Then the following inequality holds for $n \geq \mathrm{VCdim}(\mathcal{F})$,*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\left[\|\widehat{f}_{T_{\beta_n}\varrho,\mathcal{D}} - f_0\|^2_{L^2(Q_X)}\right] \lesssim \inf_{f\in\mathcal{F}} \|f - f_0\|^2_{L^2(Q_X)} + V(B^2+\sigma^2)\sqrt{\frac{\mathrm{VCdim}(\mathcal{F})}{n\log^{-1}(en)}}.$$

**Remark 20** *The generalization error bound in Cortes et al. (2010, Theorem 3) is*

$$\sup_{f\in\mathcal{F}}\left\{L_\varrho(f) - \widehat{L}_{\varrho,\mathcal{D}}(f)\right\} \leq CV\Big(\frac{\mathrm{VCdim}(\mathcal{F})}{m}\Big)^{3/8}.$$

*This bound is larger than the generalization bound $\mathcal{O}(V(\frac{\mathrm{VCdim}(\mathcal{F})}{n})^{-1/2})$ derived in Corollary 19.*

The convergence rate can also be established in the following corollary by setting the hypothesis class $\mathcal{F}$ as a ReLU DNN and choosing an appropriate value for the truncated level $\beta_n$.

**Corollary 21** *Suppose Assumptions 1, 2, 3 and 5 hold. Assume that $Q_X$ is absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be an i.i.d. sample set drawn from $P$. Set the hypothesis class $\mathcal{F} = \mathcal{N}(W_{\mathcal{F}}, L_{\mathcal{F}})$ with width $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}}\log U_{\mathcal{F}})$ and depth $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}}\log N_{\mathcal{F}})$ satisfying $U_{\mathcal{F}}N_{\mathcal{F}} = \mathcal{O}(V^{-\frac{d}{d+4\beta}}n^{\frac{d}{2(d+4\beta)}})$. Suppose that $\widehat{f}_{T_{\beta_n}\varrho,\mathcal{D}} \in \mathcal{F}$ is an empirical truncated risk minimizer given by (7) with the truncation level $\beta_n = \mathcal{O}(V^{\frac{2d+4\beta}{d+4\beta}}n^{\frac{2\beta}{d+4\beta}})$. Then the following inequality holds:*

$$\mathbb{E}_{\mathcal{D}\sim P^n}\left[\|\widehat{f}_{T_{\beta_n}\varrho,\mathcal{D}} - f_0\|^2_{L^2(Q_X)}\right] \leq \mathcal{O}\Big(V^{\frac{4\beta}{d+4\beta}}n^{-\frac{2\beta}{d+4\beta}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 22** *In contrast to the findings presented in Corollary 12, the rate* $\mathcal{O}(V^{\frac{4\beta}{d+4\beta}} n^{-\frac{2\beta}{d+4\beta}})$ *in Corollary 21 is comparatively smaller than the corresponding rate for the unweighted estimator. However, when compared to the results of Theorem 16, this rate is not minimax optimal with respect to* $n$*.*

### 4.3 Pre-training Reweighted Estimators

In this section, we expound upon the cornerstone findings of this paper, focusing on the convergence rate analysis of the pre-training reweighted estimator (11) and (12). This analysis extends beyond the results of the two previous estimators, which makes it distinct from both. Consequently, additional assumptions are introduced, as outlined below (refer to Assumptions 6, 7 and 8). Specifically, one of the most crucial steps involves obtaining density ratio estimator, which is derived from a pseudo-labeled nonparametric logistic regression, as detailed in Lemma 24. Subsequently, we can ascertain the convergence rate of the pre-training reweighted estimator.

In addition to Assumption 4, we further make assumptions on the lower bound and regularity of the density ratio as follows, which provide guarantees for the validity of the density ratio estimate by pseudo-labeled logistic regression (11).

**Assumption 6 (Density ratio with uniform positive lower-bound)** *The density ratio* $\varrho$ *has an uniform positive lower bound, that is,* $\lambda := \inf_{x\in\mathcal{X}} \varrho(x) > 0$.

**Assumption 7** *The logarithm of the density ratio* $\varrho$ *is Hölder continuous, that is,* $\log \varrho \in \mathcal{H}^{\alpha}(\mathcal{X})$ *for some* $\alpha > 0$.

Since the target density ratio $\varrho$ satisfies Assumptions 4 and 6, it is natural to make the following assumption on the hypothesis for density ratio estimates.

**Assumption 8** *For each* $u \in \mathcal{U}$*, the inequality* $\log(1/\Lambda) \leq u(x) \leq \log(1/\lambda)$ *a.e. on* $\mathcal{X}$*.*

In light of the density ratio estimator denoted as $\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}})$, where $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ is defined in (11), we can establish an oracle inequality for the pre-training reweighted estimator in the following lemma.

**Lemma 23 (Oracle inequality of pre-training reweighted estimator)** *Suppose Assumptions 1, 2, 4, 6, 8 hold. Let* $\mathcal{S} = \{(X_i^{\mu}, Z_i^{\mu})\}_{i=1}^m$ *and* $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ *be two i.i.d. sample sets drawn from* $\mu$ *and* $P$*, respectively. Suppose that* $\mathcal{U}$ *is a hypothesis class and* $\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}})$ *with* $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ *defined in* (11)*, and suppose that* $\mathcal{F}$ *is a hypothesis class and* $\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} \in \mathcal{F}$ *is defined by* (12)*. Then the following inequality holds for* $m \geq \mathrm{VCdim}(\mathcal{U})$ *and* $n \geq \mathrm{VCdim}(\mathcal{F})$*,*

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}\sim\mu^m}\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \\
&\lesssim \frac{B^2}{\lambda}\mathbb{E}_{\mathcal{S}\sim\mu^m}\Big[\|\widehat{\varrho}_{\mathcal{S}} - \varrho\|_{L^2(P_X)}^2\Big] + \inf_{f\in\mathcal{F}}\|f - f_0\|_{L^2(Q_X)}^2 + \Lambda(B^2+\sigma^2)\frac{\mathrm{VCdim}(\mathcal{F})}{n\log^{-1}(en)}.
\end{aligned}
$$

As evidenced in Lemma 23, the determination of the convergence rate for the pre-training reweighted estimator necessitates the acquisition of the error bound pertaining to the density ratio estimator. This pivotal result is provided by the following lemma.

**Lemma 24 (Convergence rates of density-ratio estimator)** *Suppose Assumptions 4, 6, 7, 8 hold. Assume that $P_X$ and $Q_X$ are absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{S} = \{(X_i^{\mu}, Z_i^{\mu})\}_{i=1}^m$ be an i.i.d. sample set drawn from $\mu$. Set the hypothesis class $\mathcal{U}$ as $\mathcal{U} = \mathcal{N}(W_{\mathcal{U}}, L_{\mathcal{U}})$ with width $W_{\mathcal{U}} = \mathcal{O}(U_{\mathcal{U}} \log U_{\mathcal{U}})$ and depth $L_{\mathcal{U}} = \mathcal{O}(N_{\mathcal{U}} \log N_{\mathcal{U}})$ satisfying $U_{\mathcal{U}} N_{\mathcal{U}} = \mathcal{O}(n^{\frac{d}{2d+4\alpha}})$. Given $\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}})$ with $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ defined in (11), then the following inequality holds:*

$$\mathbb{E}_{\mathcal{S}\sim\mu^m}\Big[\|\widehat{\varrho}_{\mathcal{S}} - \varrho\|_{L^2(P_X)}^2\Big] \leq \mathcal{O}\Big(\big(\Lambda^{\frac{2\alpha}{d+2\alpha}} \vee \lambda^{-\frac{2\alpha}{d+2\alpha}}\big) m^{-\frac{2\alpha}{d+2\alpha}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 25** *In Lemma 24, we derive a convergence rate of the density-ratio estimator $\widehat{\varrho}_{\mathcal{S}}$ under very mild conditions, which is given by $\mathcal{O}(m^{-\frac{2\alpha}{d+2\alpha}})$, with constants and logarithmic terms omitted. The derivation of this error bound is mainly facilitated through the utilization of local complexity techniques (Bartlett et al., 2005), enabling it to achieve the minimax optimal rate (Yang, 1999; Farrell et al., 2021; Zhang et al., 2023). See Section E for the proof of Lemma 24.*

In view of Theorem 16 and Lemmas 23-24, we can derive following convergence rates for the pre-training reweighted estimator.

**Theorem 26 (Convergence rates of pre-training reweighted estimator)** *Suppose Assumptions 1, 2, 3, 4, 6, 7 and 8 hold. Assume that $P_X$ and $Q_X$ are absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{S} = \{(X_i^{\mu}, Z_i^{\mu})\}_{i=1}^m$ and $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ be two i.i.d. sample sets drawn from $\mu$ and $P$, respectively. Set the hypothesis class $\mathcal{U}$ as $\mathcal{U} = \mathcal{N}(W_{\mathcal{U}}, L_{\mathcal{U}})$ with width $W_{\mathcal{U}} = \mathcal{O}(U_{\mathcal{U}} \log U_{\mathcal{U}})$ and depth $L_{\mathcal{U}} = \mathcal{O}(N_{\mathcal{U}} \log N_{\mathcal{U}})$ satisfying $U_{\mathcal{U}} N_{\mathcal{U}} = \mathcal{O}(n^{\frac{d}{2d+4\alpha}})$. Further, set the hypothesis class $\mathcal{F} = \mathcal{N}(W_{\mathcal{F}}, L_{\mathcal{F}})$ with width $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}} \log U_{\mathcal{F}})$ and depth $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}} \log N_{\mathcal{F}})$ satisfying $U_{\mathcal{F}} N_{\mathcal{F}} = \mathcal{O}(\Lambda^{-\frac{d}{2d+4\beta}} n^{\frac{d}{2d+4\beta}})$. Given $\widehat{\varrho}_{\mathcal{S}} = \exp(-\widehat{u}_{\mathcal{S}})$ with $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ defined in (11), and $\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} \in \mathcal{F}$ defined in (12), then the following inequality holds:*

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}\sim\mu^m}\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \\ &\leq \mathcal{O}\Big(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}}\Big) + \mathcal{O}\Big(\big(\Lambda^{\frac{2\alpha}{d+2\alpha}} \lambda^{-1} \vee \lambda^{-\frac{d+4\alpha}{d+2\alpha}}\big) m^{-\frac{2\alpha}{d+2\alpha}}\Big).\end{aligned}$$

*If the pre-training sample size $m$ satisfies*

$$m \geq \mathcal{O}\Big(\big(\Lambda^{\frac{(\alpha-\beta)d}{\alpha(d+2\beta)}} \lambda^{-\frac{d+2\alpha}{2\alpha}} \wedge \Lambda^{-\frac{\beta(d+2\alpha)}{\alpha(d+2\beta)}} \lambda^{-\frac{d+4\alpha}{2\alpha}}\big) n^{\frac{\beta(d+2\alpha)}{\alpha(d+2\beta)}}\Big), \tag{13}$$

*then the following inequality holds:*

$$\mathbb{E}_{\mathcal{S}\sim\mu^m}\mathbb{E}_{\mathcal{D}\sim P^n}\Big[\|\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] \leq \mathcal{O}\Big(\Lambda^{\frac{2\beta}{d+2\beta}} n^{-\frac{2\beta}{d+2\beta}}\Big),$$

*where we ignore the logarithmic factors.*

**Remark 27** *Theorem 26 establishes that, subsequent to the pre-training operation, the resulting pre-training reweighted estimator can achieve the same nonparametric efficiency as that in Theorem 16. This holds true under the condition that the pre-training sample size $m$ satisfies* (13)*, which provides valuable guidance for selecting an appropriate unsupervised pre-training sample size. Of paramount significance is the observation that this condition is often straightforward to fulfill in practical applications. This is attributable to the fact that collecting unlabeled data is typically more cost-effective than acquiring labeled data in many practical seniors, rendering it a more feasible operation.*

**Remark 28** *Importantly, in the absence of consideration of upper and lower bounds on the density ratio, the convergence rate $\mathcal{O}(n^{-\frac{2\beta}{d+2\beta}})$ attains the minimax optimal rate in nonparametric regression. While the influence of upper and lower bounds on the density ratio is undoubtedly substantial, the optimal dependency in the convergence rates for this two-stage estimation problem remains elusive to our current understanding. Accordingly, we identify this as a topic meriting future exploration and intend to undertake further research in this direction.*

## 5 Conclusion

This study investigates nonparametric regression under covariate shift and then introduces a two-stage pre-training reweighted approach. We focus on three estimators based on deep ReLU neural networks: the unweighted estimator, reweighted estimator, and pre-training reweighted estimator. We establish rigorous convergence rates for these estimators, wherein our technical novelty lies in using offset complexity techniques for statistical error analysis, resulting in a fast rate of $\mathcal{O}(1/n)$. Especially, we consider two scenarios for both unweighted and reweighted estimators: one involving an upper-bounded density ratio and the other characterized by a density ratio with a finite second moment. These theoretical results shed light on the significance of density-ratio reweighting strategy and offer a priori guide for selecting the appropriate number of pre-training samples. To further explore the impact of the upper and lower bounds on the density ratio on the convergence rate carries substantial implications. We regard this as a direction for future research.

## Appendix A. Supplementary Definitions and Lemmas

**Definition 29 (Covering number)** *Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathfrak{X} = \{X_i\}_{i=1}^m \subseteq \mathcal{X}$. Define the $L^p(\mathfrak{X})$-norm as*

$$\|f\|_{L^p(\mathfrak{X})} = \Big(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)|^p\Big)^{1/p}, \quad \textit{for } 1 \leq p < \infty,$$

*and $\|f\|_{L^\infty(\mathfrak{X})} = \max_{1\leq i\leq m}|f(X_i)|$. A set $\mathcal{F}_\delta$ is called a $L^p(\mathfrak{X})$ $\delta$-cover of $\mathcal{F}$ if for each $f \in \mathcal{F}$, there exits $f_\delta \in \mathcal{F}_\delta$ such that $\|f - f_\delta\|_{L^p(\mathfrak{X})} \leq \delta$. Furthermore,*

$$N(\delta, \mathcal{F}, L^p(\mathfrak{X})) = \inf\big\{|\mathcal{F}_\delta| : \mathcal{F}_\delta \textit{ is a } L^p(\mathfrak{X}) \ \delta\textit{-cover of } \mathcal{F}\big\}$$

*is called the empirical $\delta$-covering number of $\mathcal{F}$ based the sample $\mathfrak{X} = \{X_i\}_{i=1}^n$.*

**Lemma 30** *Let $\xi_j$ be a $\sigma^2$-sub-Gaussian random variable for $j \in [N]$. Then*

$$\mathbb{E}\Big[\max_{1\leq j\leq N} \xi_j^2\Big] \leq 4\sigma^2(\log N + 1).$$

**Proof of Lemma 30** By Jensen's inequality, it is straightforward that

$$\exp\Big(\frac{\lambda}{2\sigma^2}\mathbb{E}\Big[\max_{1\leq j\leq N} \xi_j^2\Big]\Big) \leq \mathbb{E}\Big[\max_{1\leq j\leq N} \exp\Big(\frac{\lambda\xi_j^2}{2\sigma^2}\Big)\Big] \leq N\mathbb{E}\Big[\exp\Big(\frac{\lambda\xi_1^2}{2\sigma^2}\Big)\Big] \leq \frac{N}{\sqrt{1-\lambda}},$$

where the last inequality holds from Wainwright (2019, Theorem 2.6) for each $\lambda \in [0, 1)$. Letting $\lambda = 1/2$ yields the desired inequality. ■

**Proof of Lemma 4** Setting the first variation of

$$\int_{\mathcal{X}} \log(1 + \exp(-u(x)))p_X(x) + \log(1 + \exp(u(x)))q_X(x)dx$$

to zero yields

$$\frac{\exp(-u(x))}{1 + \exp(-u(x))}p_X(x) = \frac{\exp(u(x))}{1 + \exp(u(x))}q_X(x),$$

which completes the proof. ■

Technically, the convergence rate can be dissected into two fundamental components: approximation error and generalization error. Deep approximation theory has played a pivotal role in elucidating the capabilities of neural networks in approximating smooth functions. A comprehensive body of work, encompassing studies by Yarotsky (2018); Yarotsky and Zhevnerchuk (2020); Shen et al. (2019); Shen (2020); Lu et al. (2021); Petersen and Voigtlaender (2018); Jiao et al. (2023), has established the theoretical foundations for understanding the approximation capabilities of deep neural networks. These theoretical contributions have provided insights into the capacity of neural networks to represent complex functions effectively. To bound the generalization error, researchers have harnessed the tools of empirical process theory. The works of Van Der Vaart and Wellner (1996); Van de Geer and van de Geer (2000); Van der Vaart (2000); Bartlett et al. (2005); Giné and Nickl (2021) have been instrumental in this regard. These tools allow for the quantification of generalization errors in terms of the complexity of a hypothesis class, often measured using concepts like the VC-dimension, as outlined in Definition 3.

To facilitate our discussion, let us first introduce key findings from the approximation results (Lemma 31) presented in Jiao et al. (2023) and delve into the VC-dimension analysis of ReLU neural networks (Lemma 32), as explored in Bartlett et al. (2019).

**Lemma 31 (Jiao et al. (2023, Theorem 3.3))** *Let $\mathcal{X} \subseteq [0, 1]^d$ and let $\beta = s + r$ with $s \in \mathbb{N}$ and $r \in (0, 1]$. Assume that $\mu_X$ is absolutely continuous with respect to the Lebesgue measure. For each $U, N \in \mathbb{N}_+$, there exists a ReLU neural network class $\mathcal{N}(W, L)$ with width $W = \mathcal{O}((s + 1)^2 d^{s+1} U \log U)$ and depth $L = \mathcal{O}((s + 1)^2 N \log N)$ such that*

$$\sup_{f\in\mathcal{H}^\beta(\mathcal{X})} \inf_{\psi\in\mathcal{N}(W,L)} \|f - \psi\|_{L^2(\mu_X)}^2 \lesssim (s + 1)^4 d^{2s+\beta\vee 1}(UN)^{-4\beta/d}\|f\|_{\mathcal{H}^\beta(\mathcal{X})}^2.$$

**Lemma 32 (Bartlett et al. (2019, Theorem 7))** *For each $W, L \in \mathbb{N}_+$, the VC-dimension of a ReLU neural network class $\mathcal{N}(W, L)$ with width $W$ and depth $L$ is given by*

$$\mathrm{VCdim}(\mathcal{N}(W, L)) \lesssim W^2 L^2 \log(WL).$$

## Appendix B. Error Analysis for Unweighted Estimators in Section 4.1

### B.1 Bounded Density Ratio

**Proof of Lemma 8** According to Assumption 4, it is easy to show that

$$\|\widehat{f}_{\mathcal{D}} - f_0\|_{L^2(Q_X)}^2 = \mathbb{E}_{X^P \sim P_X}\big[(\widehat{f}_{\mathcal{D}}(X^P) - f_0(X^P))^2 \varrho(X^P)\big] \leq \Lambda \|\widehat{f}_{\mathcal{D}} - f_0\|_{L^2(P_X)}^2.$$

Define $R(f) = \mathbb{E}_{(X^P,Y^P)\sim P}[(f(X^P) - Y^P)^2]$. By setting $\varrho(x) \equiv 1$ in the Step 1 of the proof of Lemma 14, we find that

$$\mathbb{E}_{\mathcal{D}}\big[R(\widehat{f}_{\mathcal{D}}) - 3\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}})\big] \leq 24B^2 \frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n}, \tag{14}$$

where $\widehat{R}_{\mathcal{D}}(f) := \frac{1}{n}\sum_{i=1}^n (f(X_i^P) - f_0(X_i^P))^2$. Similarly, it holds from the Step 2 in the proof of Lemma 14 that

$$\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}})\Big] \leq 2 \inf_{f\in\mathcal{F}} \|f - f_0\|_{L^2(P_X)}^2 + 100\sigma^2 \frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n} + \frac{4B^2}{n}. \tag{15}$$

Combining (14) and (15) completes the proof. ■

**Proof of Theorem 9** According to Assumption 3 and Lemma 31, there exists $f \in \mathcal{F} = N(W_{\mathcal{F}}, L_{\mathcal{F}})$ such that

$$\inf_{f\in\mathcal{F}} \|f - f_0\|_{L^2(P_X)}^2 \leq C_1 (U_{\mathcal{F}} N_{\mathcal{F}})^{-4\beta/d},$$

where $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}} \log U_{\mathcal{F}})$, $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}} \log N_{\mathcal{F}})$ and $C_1$ is a constant only depending on $d$, $\beta$ and $\|f_0\|_{\mathcal{H}^\beta(\mathcal{X})}$. Using Lemma 32, we find that $\mathrm{VCdim}(\mathcal{F}) \leq C_2 U_{\mathcal{F}}^2 N_{\mathcal{F}}^2 (\log U_{\mathcal{F}} \log N_{\mathcal{F}})^2$, where the constant $C_2$ depends on $d$, $\beta$ and $B$. By setting $U_{\mathcal{F}} N_{\mathcal{F}} = \mathcal{O}(n^{\frac{d}{2d+4\beta}})$, we conclude the final result. ■

### B.2 Unbounded Density Ratio

We first establish the relationship between the $L^2(Q_X)$-risk and $L^2(P_X)$-risk under Assumption 5, as elucidated in the following lemma.

**Lemma 33** *Suppose Assumptions 2 and 5 hold. Then the following inequality holds:*

$$\|f - f_0\|_{L^2(Q_X)}^2 \leq 2VB \|f - f_0\|_{L^2(P_X)}$$

*for each $f \in \mathcal{F}$.*

**Proof of Lemma 33** By the definition of the density ratio, we have

$$\begin{aligned}
\|f - f_0\|_{L^2(Q_X)}^2 &= \mathbb{E}_{X^P}[\varrho(X^P)(f(X^P) - f_0(X^P))^2] \\
&\leq \mathbb{E}_{X^P}^{1/2}[\varrho^2(X^P)]\mathbb{E}_{X^P}^{1/2}[(f(X^P) - f_0(X^P))^4] \\
&\leq V\mathbb{E}_{X^P}^{1/2}[(f(X^P) - f_0(X^P))^4] \\
&\leq 2VB\|f - f_0\|_{L^2(P_X)},
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second and third inequalities are due to Assumptions 2 and 5. ∎

Subsequently, we obtain the convergence rate for the unweighted estimator in the context of an unbounded density ratio in Corollary 12.

**Proof of Corollary 12** Combining Lemma 33 with the proof of Theorem 9 gives the desired result. ∎

## Appendix C. Error Analysis for Reweighted Estimators

### C.1 Bounded Density Ratio

Recall that $\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n$ are i.i.d. drawn from the probability distribution $P$. We define the reweighted excess risk $R_\varrho$ by

$$R_\varrho(f) = \mathbb{E}_{X^P \sim P_X}\big[\varrho(X^P)(f(X^P) - f_0(X^P))^2\big],$$

and its empirical counterpart $\widehat{R}_{\varrho,\mathcal{D}}$ based on $\mathcal{D}$ can be given by

$$\widehat{R}_{\varrho,\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^n \varrho(X_i^P)(f(X_i^P) - f_0(X_i^P))^2.$$

It is easy to verify that $R_\varrho(f) = \|f - f_0\|_{L^2(Q_X)}^2$.

**Proof of Lemma 14** To begin with, it is straightforward that

$$\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{f}_{\varrho,\mathcal{D}} - f_0\|_{L^2(Q_X)}^2\Big] = \mathbb{E}_{\mathcal{D}}\Big[R_\varrho(\widehat{f}_{\varrho,\mathcal{D}}) - 3\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] + 3\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big]. \tag{16}$$

We now derive the upper bound of these two parts on the right hand of (16), respectively.

*Step 1. Symmetrization by a ghost sample.*

We define the function class $\mathcal{H} = \{x \mapsto h(x) = \varrho(x)(f(x) - f_0(x))^2 : f \in \mathcal{F}\}$. Since that $0 \le \varrho(x) \le \Lambda$ for each $x \in \mathcal{X}$, it is apparent that $0 \le h(x) \le 4\Lambda B^2$ for each $x \in \mathcal{X}$ and $h \in \mathcal{H}$. Then it is easy to show that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\big[R_\varrho(\widehat{f}_{\varrho,\mathcal{D}}) - 3\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\big] \le \mathbb{E}_{\mathcal{D}} \sup_{h\in\mathcal{H}} \Big\{\mathbb{E}_{X^P}[h(X^P)] - \frac{3}{n}\sum_{i=1}^n h(X_i^P)\Big\} \\
&\le \mathbb{E}_{\mathcal{D}} \sup_{h\in\mathcal{H}} \Big\{2\mathbb{E}_{X^P}[h(X^P)] - \frac{1}{4\Lambda B^2}\mathbb{E}_{X^P}[h^2(X^P)] - \frac{2}{n}\sum_{i=1}^n h(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h^2(X_i^P)\Big\},
\end{aligned}$$

where we used the fact that $h^2(x) \le 4\Lambda B^2 h(x)$ for each $x \in \mathcal{X}$ and $h \in \mathcal{H}$.

Let us introduce a ghost sample $\mathcal{D}' = \{(X_i^{P,\prime}, Y_i^{P,\prime})\}_{i=1}^n$ sampled from $P$, which is independent of $\mathcal{D}$. Let $\{\varepsilon_i\}_{i=1}^n$ be a set of Rademacher variables. Then replacing the expectation by the empirical mean based on the ghost sample $\mathcal{D}'$ yields

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}} \sup_{h\in\mathcal{H}} \Big\{2\mathbb{E}_{X^P}[h(X^P)] - \frac{1}{4\Lambda B^2}\mathbb{E}_{X^P}[h^2(X^P)] - \frac{2}{n}\sum_{i=1}^n h(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h^2(X_i^P)\Big\} \\
&= \mathbb{E}_{\mathcal{D}} \sup_{h\in\mathcal{H}} \Big\{2\mathbb{E}_{\mathcal{D}'}\Big[\frac{1}{n}\sum_{i=1}^n (h(X_i^{P,\prime}) - h(X_i^P))\Big] - \frac{1}{4\Lambda B^2}\mathbb{E}_{\mathcal{D}'}\Big[\frac{1}{n}\sum_{i=1}^n (h^2(X_i^{P,\prime}) + h^2(X_i^P))\Big]\Big\} \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'} \sup_{h\in\mathcal{H}} \Big\{\frac{2}{n}\sum_{i=1}^n (h(X_i^{P,\prime}) - h(X_i^P)) - \frac{1}{4\Lambda B^2}\frac{1}{n}\sum_{i=1}^n (h^2(X_i^{P,\prime}) + h^2(X_i^P))\Big\} \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\mathbb{E}_{\varepsilon} \sup_{h\in\mathcal{H}} \Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i(h(X_i^{P,\prime}) - h(X_i^P)) - \frac{1}{4\Lambda B^2}\frac{1}{n}\sum_{i=1}^n (h^2(X_i^{P,\prime}) + h^2(X_i^P))\Big\} \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon} \sup_{h\in\mathcal{H}} \Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h(X_i^P) - \frac{1}{4\Lambda B^2}\frac{1}{n}\sum_{i=1}^n h^2(X_i^P)\Big\},
\end{aligned}$$

where $\mathbb{E}_{\varepsilon}[\cdot]$ is the expectation conditional on $\mathcal{D}$ and $\mathcal{D}'$, and the inequality holds from Jensen's inequality. We note that $\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\sup_{h\in\mathcal{H}}\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h(X_i^P) - \frac{1}{4\Lambda B^2}\frac{1}{n}\sum_{i=1}^n h^2(X_i^P)\Big\}$ refers to the offset Rademacher complexity (Liang et al., 2015). Then we transform into bounding this offset Rademacher complexity to derive the upper bound.

Let $\delta \in (0, 4\Lambda B^2)$ and $\mathcal{H}_\delta$ be a $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathcal{H}$ satisfying $|N_\delta| = N(\delta, \mathcal{H}, L^\infty(\mathcal{D}))$. Then for each $h \in \mathcal{H}$, there exists $h_\delta \in \mathcal{H}_\delta$ such that $\max_{1\leq i\leq n}|h(X_i^P) - h_\delta(X_i^P)| \leq \delta$. Consequently, it follows from Hölder's inequality that

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(X_i^P) \leq \frac{1}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) + \frac{1}{n}\sum_{i=1}^n |\varepsilon_i||h(X_i^P) - h_\delta(X_i^P)| \leq \frac{1}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) + \delta,$$

and

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^n h^2(X_i^P) &\leq -\frac{1}{n}\sum_{i=1}^n h_\delta^2(X_i^P) + \frac{1}{n}\sum_{i=1}^n |h(X_i^P) + h_\delta(X_i^P)||h_\delta(X_i^P) - h(X_i^P)| \\
&\leq -\frac{1}{n}\sum_{i=1}^n h_\delta^2(X_i^P) + 8\Lambda B^2\delta.
\end{aligned}$$

Hence we find that

$$\begin{aligned}
&\mathbb{E}_{\varepsilon} \sup_{h\in\mathcal{H}} \Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h^2(X_i^P)\Big\} \\
&\leq \mathbb{E}_{\varepsilon} \max_{h_\delta\in\mathcal{H}_\delta} \Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big\} + 4\delta.
\end{aligned} \tag{17}$$

By Hoeffding's inequality (Mohri et al., 2018, Theorem D.2), the conditional probability can be bounded as follows

$$\begin{aligned}
&\Pr\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) > t + \frac{1}{4\Lambda B^2}\frac{1}{n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big|\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n\Big\} \\
&= \Pr\Big\{\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) > \frac{nt}{2} + \frac{1}{8\Lambda B^2}\sum_{i=1}^n h_\delta^2(X_i^P)\Big|\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n\Big\} \\
&\leq \exp\Big(-\frac{(\frac{nt}{2} + \frac{1}{8\Lambda B^2}\sum_{i=1}^n h_\delta^2(X_i^P))^2}{2\sum_{i=1}^n h_\delta^2(X_i^P)}\Big) \leq \exp\Big(-\frac{nt}{8\Lambda B^2}\Big).
\end{aligned}$$

As a consequence, it follows that for each $T > 0$,

$$\begin{aligned}
&\mathbb{E}_\varepsilon \max_{h_\delta\in\mathcal{H}_\delta}\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big\} \\
&\leq \int_0^\infty \Pr\Big\{\max_{h_\delta\in\mathcal{H}_\delta}\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big\} > t\Big|\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n\Big\}dt \\
&\leq T + |\mathcal{H}_\delta|\int_T^\infty \Pr\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) > t + \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big|\mathcal{D} = \{(X_i^P, Y_i^P)\}_{i=1}^n\Big\}dt \\
&\leq T + \frac{8\Lambda B^2}{n}|\mathcal{H}_\delta|\exp\Big(-\frac{nT}{8\Lambda B^2}\Big)
\end{aligned}$$

Letting $T = \frac{8\Lambda B^2}{n}\log|\mathcal{H}_\delta|$ gives that

$$\mathbb{E}_\varepsilon \max_{h_\delta\in\mathcal{H}_\delta}\Big\{\frac{2}{n}\sum_{i=1}^n \varepsilon_i h_\delta(X_i^P) - \frac{1}{4\Lambda B^2 n}\sum_{i=1}^n h_\delta^2(X_i^P)\Big\} \leq \frac{8\Lambda B^2}{n}(\log|\mathcal{H}_\delta| + 1). \tag{18}$$

It remains to estimate the covering number $|\mathcal{H}_\delta|$. Noticing

$$|h(x) - h'(x)| = |\varrho(x)||(f(x) - f_0(x))^2 - (f'(x) - f_0(x))^2| \leq 4\Lambda B|f(x) - f'(x)|,$$

we find that

$$\log N(\delta, \mathcal{H}, L^\infty(\mathcal{D})) \leq \log N\Big(\frac{\delta}{4\Lambda B}, \mathcal{F}, L^\infty(\mathcal{D})\Big) \leq \mathrm{VCdim}(\mathcal{F})\log\Big(\frac{4e\Lambda B^2 n}{\delta}\Big), \tag{19}$$

where the last inequality is owing to Anthony et al. (1999, Theorem 12.2). By setting $\delta = 4\Lambda B^2/n$, it holds from (17), (18) and (19) that

$$\mathbb{E}_\mathcal{D}\big[R_\varrho(\widehat{f}_{\varrho,\mathcal{D}}) - 3\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\big] \leq 24\Lambda B^2\frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n}. \tag{20}$$

*Step 2. Estimate of empirical excess risk.*

For each function $f : \mathcal{X} \to \mathbb{R}$, we connect the empirical risk of it with its empirical excess risk by

$$\widehat{R}_{\varrho,\mathcal{D}}(f) = \widehat{L}_{\varrho,\mathcal{D}}(f) + \frac{2}{n}\sum_{i=1}^n \varrho(X_i^P)\xi_i(f(X_i^P) - f_0(X_i^P)) - \frac{1}{n}\sum_{i=1}^n \varrho(X_i^P)\xi_i^2.$$

Plugging the reweighted empirical risk minimizer $\widehat{f}_{\varrho,\mathcal{D}}$ and taking expectation with respect to $\mathcal{D}$ on both sides of the equality implies that for each $f \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] &= \mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] - \sigma^2 + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\varrho(X_i^P)\xi_i\widehat{f}_{\varrho,\mathcal{D}}(X_i^P)\Big] \\ &\leq \mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{\varrho,\mathcal{D}}(f)\Big] - \sigma^2 + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\varrho(X_i^P)\xi_i\widehat{f}_{\varrho,\mathcal{D}}(X_i^P)\Big] \\ &= \|f - f_0\|_{L^2(Q_X)}^2 + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\varrho(X_i^P)\xi_i\widehat{f}_{\varrho,\mathcal{D}}(X_i^P)\Big],\end{aligned}$$

which deduces

$$\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] \leq \inf_{f\in\mathcal{F}}\|f - f_0\|_{L^2(Q_X)}^2 + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\varrho(X_i^P)\xi_i\widehat{f}_{\varrho,\mathcal{D}}(X_i^P)\Big]. \tag{21}$$

Define $\widehat{g}_{\mathcal{D}}(x) = \varrho(x)\widehat{f}_{\varrho,\mathcal{D}}(x)$ and $g_0 = \varrho(x)f_0(x)$ for each $x \in \mathcal{X}$. In addition, define the function class $\mathcal{G} = \{x \mapsto g(x) = \varrho(x)f(x) : f \in \mathcal{F}\}$. Let $\delta \in (0, \Lambda B)$ and $\mathcal{G}_\delta$ be a $L^\infty(\mathcal{D})$ $\delta$-cover of $\mathcal{G}$ with $|\mathcal{G}_\delta| = N(\delta, \mathcal{G}, L^\infty(\mathcal{D}))$. Suppose that $g_\delta$ is a function in $\mathcal{G}_\delta$ such that $\max_{1\leq i\leq n}|\widehat{g}_{\mathcal{D}}(X_i^P) - g_\delta(X_i^P)| \leq \delta$. Then we find that

$$\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{g}_{\mathcal{D}}(X_i^P) - g_\delta(X_i^P))\Big] \leq \delta\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\Big] \leq \delta\sigma,$$

where the last inequality is due to Hölder's inequality. Consequently, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{g}_{\mathcal{D}}(X_i^P)\Big] &= \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{g}_{\mathcal{D}}(X_i^P) - g_0(X_i^P))\Big] \\ &\leq \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\xi_i(g_\delta(X_i^P) - g_0(X_i^P))\Big] + \delta\sigma \\ &\leq \mathbb{E}_{\mathcal{D}}\Big[\frac{\|\widehat{g}_{\mathcal{D}} - g_0\|_{L^2(\mathcal{D})} + \delta}{\sqrt{n}}\psi(g_\delta)\Big] + \delta\sigma \\ &\leq \Big(\mathbb{E}_{\mathcal{D}}^{1/2}\Big[\|\widehat{g}_{\mathcal{D}} - g_0\|_{L^2(\mathcal{D})}^2\Big] + \delta\Big)\frac{1}{\sqrt{n}}\mathbb{E}_{\mathcal{D}}^{1/2}\Big[\psi^2(g_\delta)\Big] + \delta\sigma \\ &\leq \frac{1}{4}\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{g}_{\mathcal{D}} - g_0\|_{L^2(\mathcal{D})}^2\Big] + \frac{2}{n}\mathbb{E}_{\mathcal{D}}\Big[\psi^2(g_\delta)\Big] + \frac{1}{4}\delta^2 + \delta\sigma. \end{aligned} \tag{22}$$

Here, the first and second inequalities are from the definition of covering, and

$$\psi(g_\delta) := \frac{\sum_{i=1}^{n}\xi_i(g_\delta(X_i^P) - g_0(X_i^P))}{\sqrt{n}\|g_\delta - g_0\|_{L^2(\mathcal{D})}},$$

the third inequality holds from Cauchy-Schwarz inequality, while the last one is owing to the AM-GM inequality $ab \leq a^2/4 + b^2$. Observe that for each fixed $g_\delta$, the random variable $\psi(g_\delta)$ is sub-Gaussian with variance proxy $\sigma^2$. Then using Lemma 30 gives that

$$\mathbb{E}_{\xi}\Big[\psi^2(g_\delta)\Big] \leq \mathbb{E}_{\xi}\Big[\max_{g\in\mathcal{G}_\delta}\psi^2(g)\Big] \leq 4\sigma^2(\log|\mathcal{G}_\delta| + 1). \tag{23}$$

We now estimate the covering number $|\mathcal{G}_\delta|$. Using the fact that

$$|g(x) - g'(x)| = |\varrho(x)||f(x) - f'(x)| \leq \Lambda |f(x) - f'(x)|,$$

we implies for $n \geq \mathrm{VCdim}(\mathcal{F})$,

$$\log N(\delta, \mathcal{G}, L^\infty(\mathcal{D})) \leq \log N\Big(\frac{\delta}{\Lambda}, \mathcal{F}, L^\infty(\mathcal{D})\Big) \leq \mathrm{VCdim}(\mathcal{F}) \log\Big(\frac{e\Lambda B n}{\delta}\Big), \tag{24}$$

where the last inequality is due to Anthony et al. (1999, Theorem 12.2). Combining (22), (23) and (24), and setting $\delta = \Lambda B/n$ gives

$$\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n \xi_i \widehat{g}_{\mathcal{D}}(X_i^P)\Big] \leq \frac{1}{4}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] + 25\sigma^2 \frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n} + \frac{\Lambda(B^2+\sigma^2)}{n}. \tag{25}$$

Using (21) and (25) yields

$$\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\varrho,\mathcal{D}}(\widehat{f}_{\varrho,\mathcal{D}})\Big] \leq 2 \inf_{f\in\mathcal{F}} \|f - f_0\|^2_{L^2(Q_X)} + 100\sigma^2 \frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n} + \frac{4\Lambda(B^2+\sigma^2)}{n}. \tag{26}$$

Combining (16), (20) and (26) completes the proof. ■

**Proof of Theorem 16** According to Assumption 3 and Lemma 31, there exists $f \in \mathcal{F} = N(W_{\mathcal{F}}, L_{\mathcal{F}})$ such that

$$\inf_{f\in\mathcal{F}} \|f - f_0\|^2_{L^2(Q_X)} \leq C_1 (U_{\mathcal{F}} N_{\mathcal{F}})^{-4\beta/d},$$

where $W_{\mathcal{F}} = \mathcal{O}(U_{\mathcal{F}} \log U_{\mathcal{F}})$, $L_{\mathcal{F}} = \mathcal{O}(N_{\mathcal{F}} \log N_{\mathcal{F}})$ and $C_1$ is a constant only depending on $d$, $\beta$ and $\|f_0\|_{\mathcal{H}^\beta(\mathcal{X})}$. Using Lemma 32, we find that $\mathrm{VCdim}(\mathcal{F}) \leq C_2 U_{\mathcal{F}}^2 N_{\mathcal{F}}^2 (\log U_{\mathcal{F}} \log N_{\mathcal{F}})^2$, where the constant $C_2$ depends on $d$, $\beta$ and $B$. By setting $U_{\mathcal{F}} N_{\mathcal{F}} = \mathcal{O}(\Lambda^{-\frac{d}{2d+4\beta}} n^{\frac{d}{2d+4\beta}})$, we conclude the final result. ■

### C.2 Unbounded Density Ratio

**Proof of Lemma 18** It is apparent from the definition of density ratio and Assumption 5 that $\mathbb{E}_{X^Q}[\varrho(X^Q)] = \mathbb{E}_{X^P}[\varrho^2(X^P)] = V^2$. Then it follows that

$$\begin{aligned}
R_\varrho(f) - R_{T_\tau\varrho}(f) &= \mathbb{E}_{X^Q}\big[\mathbb{1}\{\varrho(X^Q) \geq \tau\}(f(X^Q) - f_0(X^Q))^2\big] \\
&\quad + \mathbb{E}_{X^P}\big[\mathbb{1}\{\varrho(X^P) < \tau\}\varrho(X^P)(f(X^P) - f_0(X^P))^2\big] \\
&\quad - \mathbb{E}_{X^P}\big[T_\tau\varrho(X^P)(f(X^P) - f_0(X^P))^2\big] \\
&= \mathbb{E}_{X^Q}\big[\mathbb{1}\{\varrho(X^Q) \geq \tau\}(f(X^Q) - f_0(X^Q))^2\big] \\
&\quad + \tau\mathbb{E}_{X^P}\big[\mathbb{1}\{\varrho(X^P) \geq \tau\}(f(X^P) - f_0(X^P))^2\big] \\
&\leq 4B^2 \Pr\{\varrho(X^Q) \geq \tau\} + 4B^2\tau \Pr\{\varrho(X^P) \geq \tau\} \\
&\leq 4B^2 \frac{\mathbb{E}_{X^Q}[\varrho(X^Q)]}{\beta_n} + 4B^2\tau \frac{\mathbb{E}_{X^P}[\varrho^2(X^P)]}{\tau^2} \leq \frac{8B^2V^2}{\tau},
\end{aligned}$$

where the first inequality is due to Assumption 2, the second inequality holds from the Markov's inequality and Chebyshev's inequality. ■

**Proof of Corollary 19** To begin with, we estimate the truncated reweighting risk of the estimator $R_{T_\tau\varrho}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})$. Using a same technique as that used in the Step 3 of the proof of Lemma 14, it follows that

$$\mathbb{E}_{\mathcal{D}}\Big[R_{T_\tau\varrho}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})-3\widehat{R}_{T_\tau\varrho,\mathcal{D}}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})\Big]\leq 24\tau B^2\frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n}. \tag{27}$$

For the empirical truncated reweighting risk $\widehat{R}_{T_\tau\varrho,\mathcal{D}}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})$ and the definition of $\widehat{f}_{T_\tau\varrho,\mathcal{D}}$, it follows that for each $f\in\mathcal{F}$

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{T_\tau\varrho,\mathcal{D}}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})\Big]\\
&=\mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{T_\tau\varrho,\mathcal{D}}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})\Big]+2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n T_\tau\varrho(X_i^P)\xi_i\widehat{f}_{T_\tau\varrho,\mathcal{D}}(X_i^P)\Big]-\mathbb{E}_{X^P\sim P_X}\Big[T_\tau\varrho(X^P)\Big]\sigma^2\\
&\leq\mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{T_\tau\varrho,\mathcal{D}}(f)\Big]+2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n T_\tau\varrho(X_i^P)\xi_i\widehat{f}_{T_\tau\varrho,\mathcal{D}}(X_i^P)\Big]-\mathbb{E}_{X^P\sim P_X}\Big[T_\tau\varrho(X^P)\Big]\sigma^2\\
&=L_{T_\tau\varrho}(f)+2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n T_\tau\varrho(X_i^P)\xi_i\widehat{f}_{T_\tau\varrho,\mathcal{D}}(X_i^P)\Big]-\mathbb{E}_{X^P\sim P_X}\Big[T_\tau\varrho(X^P)\Big]\sigma^2\\
&=R_{T_\tau\varrho}(f)+2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n T_\tau\varrho(X_i^P)\xi_i\widehat{f}_{T_\tau\varrho,\mathcal{D}}(X_i^P)\Big]\\
&\leq R_{\varrho}(f)+2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n T_\tau\varrho(X_i^P)\xi_i\widehat{f}_{T_\tau\varrho,\mathcal{D}}(X_i^P)\Big],
\end{aligned}$$

where the last inequality follows from $T_\tau\varrho(x)\leq\varrho(x)$ for each $x\in\mathcal{X}$. Using a similar argument as (26) in the Step 3 of the proof of Lemma 14, we have

$$\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{T_\tau\varrho,\mathcal{D}}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})\Big]\quad\leq 2\inf_{f\in\mathcal{F}}\|f-f_0\|_{L^2(Q_X)}^2+100\sigma^2\frac{\mathrm{VCdim}(\mathcal{F})\log(e\tau n)}{n}+\frac{4B^2}{n}. \tag{28}$$

Finally, we can relate $L^2(Q_X)$-risk to the truncated reweighting risk by Lemma 18, that is,

$$\|\widehat{f}_{T_\tau\varrho,\mathcal{D}}-f_0\|_{L^2(Q_X)}^2\leq R_{T_\tau\varrho}(\widehat{f}_{T_\tau\varrho,\mathcal{D}})+\frac{8B^2V^2}{\tau}. \tag{29}$$

Combining (27), (28) and (29) yields the desired result. ■

**Proof of Corollary 21** This proof can be complete by using Corollary 19 and Lemma 31. ■

## Appendix D. Error Analysis for Pre-training Reweighted Estimators

**Proof of Lemma 23** It is straightforward to verify that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}-f_0\|^2_{L^2(Q_X)}\Big]&=\mathbb{E}_{\mathcal{D}}\Big[R_{\varrho}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})-R_{\widehat{\varrho}_{\mathcal{S}}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big]\\&\quad+\mathbb{E}_{\mathcal{D}}\Big[R_{\widehat{\varrho}_{\mathcal{S}}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})-3\widehat{R}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big]+3\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big].\end{aligned}\tag{30}$$

For each $f:\mathcal{X}\to[-B,B]$, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}R_{\varrho}(f)-R_{\widehat{\varrho}_{\mathcal{S}}}(f)&=\mathbb{E}_{X^P\sim P_X}\big[(\varrho(X^P)-\widehat{\varrho}_{\mathcal{S}}(X^P))(f(X^P)-f_0(X^P))^2\big]\\&\leq\mathbb{E}^{1/2}_{X^P\sim P_X}\big[(\varrho(X^P)-\widehat{\varrho}_{\mathcal{S}}(X^P))^2\big]\mathbb{E}^{1/2}_{X^P\sim P_X}\big[(f(X^P)-f_0(X^P))^4\big]\\&\leq\frac{2B^2}{\lambda}\|\varrho-\widehat{\varrho}_{\mathcal{S}}\|^2_{L^2(P_X)}+\frac{\lambda}{2}\|f-f_0\|^2_{L^2(P_X)}\\&\leq\frac{2B^2}{\lambda}\|\varrho-\widehat{\varrho}_{\mathcal{S}}\|^2_{L^2(P_X)}+\frac{1}{2}\|f-f_0\|^2_{L^2(Q_X)},\end{aligned}\tag{31}$$

where the second inequality follows from Assumption 2 and the weighted Cauchy-Schwarz inequality $ab\leq\tau a+\frac{b}{4\tau}$ with $\tau=2B^2/\lambda$, and the last one is due to Assumption 6. By substituting $\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}$ into (31) and taking expectation, we obtain the following inequality

$$\mathbb{E}_{\mathcal{D}}\Big[R_{\varrho}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})-R_{\widehat{\varrho}_{\mathcal{S}}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big]\leq\frac{2B^2}{\lambda}\|\varrho-\widehat{\varrho}_{\mathcal{S}}\|^2_{L^2(P_X)}+\frac{1}{2}\mathbb{E}_{\mathcal{D}}\big[\|\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}-f_0\|^2_{L^2(Q_X)}\big].\tag{32}$$

We next consider the second and third terms in (30). Using a same technique as that used in the Step 3 of the proof of Lemma 14, it follows from Assumption 8 that

$$\mathbb{E}_{\mathcal{D}}\Big[R_{\widehat{\varrho}_{\mathcal{S}}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})-3\widehat{R}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big]\leq 24\Lambda B^2\frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n}.\tag{33}$$

Finally, we estimate the last term in (30). To this end, we next relate the reweighted population loss with the reweighted $L^2$-risk by

$$\begin{aligned}L_{\widehat{\varrho}_{\mathcal{S}}}(f)&=\mathbb{E}_{(X^P,Y^P)\sim P}\Big[\widehat{\varrho}_{\mathcal{S}}(X^P)(f(X^P)-f_0(X^P)-\xi)^2\Big]\\&=R_{\widehat{\varrho}_{\mathcal{S}}}(f)+\mathbb{E}_{X^P\sim P_X}\Big[\widehat{\varrho}_{\mathcal{S}}(X^P)\Big]\sigma^2.\end{aligned}\tag{34}$$

Similarly, their empirical counterparts are related by

$$\widehat{L}_{\widehat{\varrho}_{\mathcal{S}}}(f)=\widehat{R}_{\widehat{\varrho}_{\mathcal{S}}}(f)-\frac{2}{n}\sum_{i=1}^n\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i(f(X_i^P)-f_0(X_i^P))+\frac{1}{n}\sum_{i=1}^n\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i^2.\tag{35}$$

Then it follows for each $f \in \mathcal{F}$ that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big] \\
&= \mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big] + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(X_i^P)\Big] - \mathbb{E}_{X^P\sim P_X}\Big[\widehat{\varrho}_{\mathcal{S}}(X^P)\Big]\sigma^2 \\
&\leq \mathbb{E}_{\mathcal{D}}\Big[\widehat{L}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(f)\Big] + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(X_i^P)\Big] - \mathbb{E}_{X^P\sim P_X}\Big[\widehat{\varrho}_{\mathcal{S}}(X^P)\Big]\sigma^2 \\
&= L_{\widehat{\varrho}_{\mathcal{S}}}(f) + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(X_i^P)\Big] - \mathbb{E}_{X^P\sim P_X}\Big[\widehat{\varrho}_{\mathcal{S}}(X^P)\Big]\sigma^2 \\
&= \Big(R_{\widehat{\varrho}_{\mathcal{S}}}(f) - R_{\varrho}(f)\Big) + R_{\varrho}(f) + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(X_i^P)\Big] \\
&\leq \frac{2B^2}{\lambda}\|\varrho - \widehat{\varrho}_{\mathcal{S}}\|_{L^2(P_X)}^2 + \frac{3}{2}\|f - f_0\|_{L^2(Q_X)}^2 + 2\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(X_i^P)\Big],
\end{aligned}$$

where the first equality holds from (35) and the fact that $\mathbb{E}_{\xi}[\sum_{i=1}^{n}\widehat{\varrho}_{\mathcal{S}}(X_i^P)\xi_i f_0(X_i^P)] = 0$, the first inequality holds since $\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}$ is the minimizer of $\widehat{L}_{\widehat{\varrho}_{\mathcal{S}}}(f)$ over $\mathcal{F}$, the third equality is due to (34), and the last inequality is from (31). Using a similar argument as (26) in the Step 3 of the proof of Lemma 14, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}}(\widehat{f}_{\widehat{\varrho}_{\mathcal{S}},\mathcal{D}})\Big] \leq \frac{4B^2}{\lambda}\|\widehat{\varrho}_{\mathcal{S}} - \varrho\|_{L^2(P_X)}^2 + 3\inf_{f\in\mathcal{F}}\|f - f_0\|_{L^2(Q_X)}^2 \\
+ 100\sigma^2\frac{\mathrm{VCdim}(\mathcal{F})\log(en)}{n} + \frac{4\Lambda(B^2+\sigma^2)}{n}.
\end{aligned} \tag{36}$$

Combining (30), (32), (33) and (36) completes the proof. ∎

## Appendix E. Error Analysis for Density Ratio Estimators

### E.1 Supplementary Materials about Local Rademacher Complexity

**Definition 34 (Rademacher complexity)** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and $\mathfrak{X} = \{X_i\}_{i=1}^{m} \subseteq \mathcal{X}$ be a sample drawn from $\mu_X^m$. Let $\varepsilon = \{\varepsilon_i\}_{i=1}^{m}$ be independent Rademacher variables. Then the empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample $\mathfrak{X}$ is defined as*

$$\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F}) = \mathbb{E}_{\varepsilon}\Big[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i f(X_i)\Big],$$

*where $\mathbb{E}_{\varepsilon}[\cdot]$ is the expectation with respect to $\varepsilon$ conditional on $\mathfrak{X}$. The Rademacher complexity of $\mathcal{F}$ is the expectation of the empirical Rademacher complexity over all samples drawn according to $\mu_X^m$, that is, $\mathfrak{R}_m = \mathbb{E}_{\mathfrak{X}}[\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F})]$.*

**Lemma 35 (Bartlett et al. (2005, Lemma A.4))** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[-B, B]$ for some positive constant $B$. Then with probability at least $1-\delta$, the following inequality holds*

$$\mathfrak{R}_m(\mathcal{F}) \leq 2\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F}) + \frac{2B\log(1/\delta)}{n}.$$

**Lemma 36 (Bartlett et al. (2005, Theorem 2.1))** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[-B, B]$ for some positive constant $B$. Assume that there exists some $r > 0$ such that*

$$\mathcal{F} \subseteq \big\{f \in \mathcal{F} : \mathbb{E}_X[f^2(X)] \leq r\big\}.$$

*Then for each $\delta \in (0, 1)$ with probability at least $1-\delta$, the following inequality holds*

$$\sup_{f\in\mathcal{F}} \Big(\frac{1}{m}\sum_{i=1}^m f(X_i) - \mathbb{E}_X[f(X)]\Big) \leq 3\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{2r\log(1/\delta)}{m}} + \frac{14}{3}\frac{B\log(1/\delta)}{m}.$$

**Lemma 37 (Bartlett et al. (2005, Corollary 2.2))** *Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[-B, B]$ for some positive constant $B$. For each $\delta \in (0, 1)$ and each $r$ satisfy*

$$r \geq 12B\mathfrak{R}_m\Big(\big\{f \in \mathcal{F}, \mathbb{E}_X[f^2(X)] \leq r\big\}\Big) + \frac{12B^2\log(1/\delta)}{m},$$

*the following holds with probability at least $1-\delta$*

$$\Big\{f \in \mathcal{F} : \mathbb{E}_X[f^2(X)] \leq r\Big\} \subseteq \Big\{f \in \mathcal{F} : \frac{1}{m}\sum_{i=1}^m f^2(X_i) \leq 2r\Big\}.$$

**Proof of Lemma 37** Note that $\mathbb{E}_X[f^2(X)] \leq r$ implies $\mathbb{E}[f^4(X)] \leq B^2\mathbb{E}[f^2(X)] \leq B^2 r$. Then applying Lemma 36 gives that for each $f \in \mathcal{F}$ satisfying $\mathbb{E}_X[f^2(X)] \leq r$, the following holds with probability at least $1-\delta$

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m f^2(X_i) &\leq \mathbb{E}_X[f^2(X)] + 3\mathfrak{R}_m\Big(\big\{f^2 : f \in \mathcal{F}, \mathbb{E}_X[f^2(X)] \leq r\big\}\Big) \\ &\qquad + \sqrt{\frac{2B^2 r\log(1/\delta)}{m}} + \frac{14}{3}\frac{B^2\log(1/\delta)}{m} \\ &\leq r + 6B\mathfrak{R}_m\Big(\big\{f \in \mathcal{F}, \mathbb{E}_X[f^2(X)] \leq r\big\}\Big) + \frac{r}{2} + \frac{17}{3}\frac{B^2\log(1/\delta)}{m} \leq 2r,\end{aligned}$$

where the second inequality is due to Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) and Cauchy-Schwarz inequality $ab \leq a^2/4 + b^2$, while the last inequality is owing to the assumption. ■

**Definition 38 (Sub-root function)** *A function $\psi : [0,\infty) \to [0,\infty)$ is sub-root if it is non-negative, non-decreasing and if $r \mapsto \psi(r)/\sqrt{r}$ is non-increasing for $r > 0$.*

**Lemma 39 (Bartlett et al. (2005, Lemma 3.2))** *If $\psi : [0,\infty) \to [0,\infty)$ is a nontrivial sub-root function, then it is continuous on $[0,\infty)$ and the equality $\psi(r) = r$ has a unique positive solution. Moreover, if we denote the solution by $r^*$, then for each $r > 0$, the inequality $r \geq \psi(r)$ holds if and only if $r^* \leq r$.*

**Lemma 40 (Bartlett et al. (2005, Theorem 3.3))** *Suppose the following conditions hold:*

*(i) Let $\mathcal{F}$ be a class of functions taking values in $[-B, B]$.*

*(ii) There are some functional $T : \mathcal{F} \to [0, +\infty)$ and some positive constant $\nu$ such that $\mathbb{E}_X[f^2(X)] \leq T(f) \leq \nu \mathbb{E}_X[f(X)]$ for each $f \in \mathcal{F}$.*

*(iii) Let $\psi$ be a sub-root function and let $r^*$ be the fixed point of $\psi$, satisfying*

$$\psi(r) \geq \nu \mathfrak{R}_m\Big(\big\{f \in \mathcal{F} : T(f) \leq r\big\}\Big).$$

*Then for each $\delta \in (0,1)$, the following holds with probability at least $1-\delta$*

$$\mathbb{E}_X\big[f(X)\big] \leq \frac{2}{m}\sum_{i=1}^m f(X_i) + \frac{1408}{\nu} r^* + \frac{(22B + 52\nu)\log(1/\delta)}{m}, \tag{37}$$

*for each $f \in \mathcal{F}$. Also, the following holds with probability at least $1-\delta$*

$$\frac{1}{m}\sum_{i=1}^m f(X_i) \leq \frac{3}{2}\mathbb{E}_X\big[f(X)\big] + \frac{1408}{\nu} r^* + \frac{(22B + 52\nu)\log(1/\delta)}{m}, \tag{38}$$

*for each $f \in \mathcal{F}$.*

To compute the local Rademacher complexities, we introduce Lemma 41, which is a simplified version of Dudley's integral bound (Srebro and Sridharan, 2010, Theorem 2.1). This lemma is also inspired by Lemma 5.7 in van Handel (2016).

**Lemma 41 (Lipschitz maximal inequality)** *Let $\mathcal{F}$ be a class of functions. The following inequality holds for each $r > 0$*

$$\widehat{\mathfrak{R}}_{\mathcal{X}}\Big(\Big\{f \in \mathcal{F} : \frac{1}{m}\sum_{i=1}^m f^2(X_i) \leq r\Big\}\Big) \leq \inf_{\varepsilon > 0}\left\{2\varepsilon + \sqrt{\frac{2r \log N(\varepsilon, \mathcal{F}, L^\infty(\mathcal{X}))}{m}}\right\}.$$

Before the proof of Lemma 41, we first introduce Massart's lemma as preparation.

**Lemma 42 (Massart's lemma)** *Let $\mathcal{F}$ be a class of functions satisfying $|\mathcal{F}| < \infty$. The following inequality holds for each $r > 0$*

$$\widehat{\mathfrak{R}}_{\mathcal{X}}\Big(\Big\{f \in \mathcal{F} : \frac{1}{m}\sum_{i=1}^m f^2(X_i) \leq r\Big\}\Big) \leq \sqrt{\frac{2r \log |\mathcal{F}|}{m}}.$$

**Proof of Lemma 42** Let $\{\varepsilon_i\}_{i=1}^m$ be a set of independent Rademacher random variables. For the fixed sample $\mathcal{X} = \{X_i\}_{i=1}^n$, $\{\varepsilon_i f(X_i)\}_{i=1}^m$ are random variables satisfying $-f(X_i) \leq \varepsilon_i f(X_i) \leq f(X_i)$ and $\mathbb{E}[\varepsilon_i f(X_i)|X_i] = 0$ for $i \in [n]$. Then it follows from Hoeffding's lemma (Mohri et al., 2018, Lemma D.1) that for each $i \in [n]$,

$$\mathbb{E}\big[\exp(\lambda\varepsilon_i f(X_i))\big|X_i\big] \leq \exp\Big(\frac{\lambda^2 f^2(X_i)}{2}\Big).$$

Consequently, we have

$$\begin{aligned}
\mathbb{E}\Big[\exp\Big(\lambda\sum_{i=1}^m \varepsilon_i f(X_i)\Big)\Big|\mathcal{X}\Big] &= \mathbb{E}\Big[\prod_{i=1}^m \exp(\lambda\varepsilon_i f(X_i))\Big|\mathcal{X}\Big] = \prod_{i=1}^m \mathbb{E}\big[\exp(\lambda\varepsilon_i f(X_i))\big|X_i\big] \\
&\leq \prod_{i=1}^m \exp\Big(\frac{\lambda^2 f^2(X_i)}{2}\Big) = \exp\Big(\frac{\lambda^2\sum_{i=1}^m f^2(X_i)}{2}\Big).
\end{aligned}$$

Furthermore, it follows from Jensen's inequality that

$$\begin{aligned}
&\exp\Big(\lambda\mathbb{E}\Big[\max_{f\in\mathcal{F}}\sum_{i=1}^m \varepsilon_i f(X_i)\Big|\mathcal{X}\Big]\Big) \\
&\leq \mathbb{E}\Big[\exp\Big(\lambda\max_{f\in\mathcal{F}}\sum_{i=1}^m \varepsilon_i f(X_i)\Big)\Big|\mathcal{X}\Big] = \mathbb{E}\Big[\max_{f\in\mathcal{F}}\exp\Big(\lambda\sum_{i=1}^m \varepsilon_i f(X_i)\Big)\Big|\mathcal{X}\Big] \\
&\leq \mathbb{E}\Big[\sum_{f\in\mathcal{F}}\exp\Big(\lambda\sum_{i=1}^m \varepsilon_i f(X_i)\Big)\Big|\mathcal{X}\Big] \leq |\mathcal{F}|\exp\Big(\frac{\lambda^2\sum_{i=1}^m f^2(X_i)}{2}\Big).
\end{aligned}$$

Taking the logarithm of both sides of the inequality yields

$$\mathbb{E}\Big[\max_{f\in\mathcal{F}}\sum_{i=1}^m \varepsilon_i f(X_i)\Big|\mathcal{X}\Big] \leq \frac{\log|\mathcal{F}|}{\lambda} + \frac{\lambda\sum_{i=1}^m f^2(X_i)}{2}.$$

Setting $\lambda^2 = 2\log|\mathcal{F}|(\sum_{i=1}^m f^2(X_i))^{-1}$ gives

$$\widehat{\mathfrak{R}}_{\mathcal{X}}(\mathcal{F}) \leq \mathbb{E}\Big[\max_{f\in\mathcal{F}}\sum_{i=1}^m \varepsilon_i f(X_i)\Big|\mathcal{X}\Big] \leq \sqrt{\frac{2(\frac{1}{m}\sum_{i=1}^m f^2(X_i))\log|\mathcal{F}|}{m}},$$

which completes the proof. ∎

**Proof of Lemma 41** Denote $\mathcal{F}^r = \{f\in\mathcal{F} : \frac{1}{m}\sum_{i=1}^m f^2(X_i) \leq r\}$. Let $\mathcal{F}_\varepsilon^r$ be an $L^\infty(\mathcal{X})$ $\varepsilon$-cover of $\mathcal{F}^r$ such that $|\mathcal{F}_\varepsilon^r| = N(\varepsilon, \mathcal{F}^r, L^\infty(\mathcal{X}))$, which means, for each $f\in\mathcal{F}^r$ there exists $f_\varepsilon \in \mathcal{F}_\varepsilon^r$ such that $\max_{1\leq i\leq n}|f(X_i) - f_\varepsilon(X_i)| \leq \varepsilon$. For $f_\varepsilon \in \mathcal{F}_\varepsilon$, if $\frac{1}{m}\sum_{i=1}^m f_\varepsilon^2(X_i) \leq r$, we define $\tilde{f}_\varepsilon = f_\varepsilon$. If $\frac{1}{m}\sum_{i=1}^m f_\varepsilon^2(X_i) > r$, let $\tilde{f}_\varepsilon$ be the nearest element of $f_\varepsilon$ in $\mathcal{F}^r$, that is,

$$\tilde{f}_\varepsilon \in \operatorname*{argmin}_{f\in\mathcal{F}^r}\Big(\max_{1\leq i\leq n}|f(X_i) - f_\varepsilon(X_i)|\Big).$$

Then it is apparent that for each $f\in\mathcal{F}^r$,

$$\max_{1\leq i\leq n}|f_\varepsilon(X_i) - \tilde{f}_\varepsilon(X_i)| \leq \max_{1\leq i\leq n}|f_\varepsilon(X_i) - f(X_i)| \leq \varepsilon.$$

According to the triangular inequality, for each $f \in \mathcal{F}^r$ satisfying $\max_{1\leq i\leq n} |f(X_i) - f_\varepsilon(X_i)| \leq \varepsilon$, it holds that

$$\max_{1\leq i\leq n} |f(X_i) - \tilde{f}_\varepsilon(X_i)| \leq \max_{1\leq i\leq n} |f(X_i) - f_\varepsilon(X_i)| + \max_{1\leq i\leq n} |f_\varepsilon(X_i) - \tilde{f}_\varepsilon(X_i)| \leq 2\varepsilon.$$

Hence $\tilde{\mathcal{F}}_\varepsilon^r = \{\tilde{f}_\varepsilon : f_\varepsilon \in \mathcal{F}_\varepsilon^r\}$ is an $L^\infty(\mathfrak{X})$ $(2\varepsilon)$-cover of $\mathcal{F}^r$ satisfying $|\tilde{\mathcal{F}}_\varepsilon^r| = N(\varepsilon, \mathcal{F}^r, L^\infty(\mathfrak{X}))$, and $\frac{1}{m}\sum_{i=1}^m \tilde{f}_\varepsilon^2(X_i) \leq r$ for each $\tilde{f}_\varepsilon \in \tilde{\mathcal{F}}_\varepsilon^r$. Then it is straightforward that

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F}^r) &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}^r} \frac{1}{m}\sum_{i=1}^m \varepsilon_i(f(X_i) - \tilde{f}_\varepsilon(X_i))\Big|\mathfrak{X}\Big] + \mathbb{E}\Big[\sup_{\tilde{f}_\varepsilon\in\tilde{\mathcal{F}}_\varepsilon^r} \frac{1}{m}\sum_{i=1}^m \varepsilon_i\tilde{f}_\varepsilon(X_i)\Big|\mathfrak{X}\Big] \\
&\leq \sup_{f\in\mathcal{F}^r}\max_{1\leq i\leq n} |f(X_i) - \tilde{f}_\varepsilon(X_i)| + \widehat{\mathfrak{R}}_{\mathfrak{X}}(\tilde{\mathcal{F}}_\varepsilon^r) \leq 2\varepsilon + \widehat{\mathfrak{R}}_{\mathfrak{X}}(\tilde{\mathcal{F}}_\varepsilon^r),
\end{aligned}$$

where the first inequality holds from Hölder's inequality. Combining this with Lemma 42 and noting that $N(\varepsilon, \mathcal{F}^r, L^\infty(\mathfrak{X})) \leq N(\varepsilon, \mathcal{F}, L^\infty(\mathfrak{X}))$ yield the desired result. ■

**Lemma 43** *Let $\mathcal{F}$ be a class of functions, and $f^*$ be a function may depending on $\mathfrak{X}$. Then it follows that*

$$\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F}) = \widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F} - f^*),$$

*where $\mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$.*

**Proof of Lemma 43** It is straightforward that

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F} - f^*) &= \mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \varepsilon_i(f(X_i) - f^*(X_i))\Big] \\
&= \mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \varepsilon_i f(X_i)\Big] - \mathbb{E}_\varepsilon\Big[\frac{1}{m}\sum_{i=1}^m \varepsilon_i f^*(X_i)\Big] = \widehat{\mathfrak{R}}_{\mathfrak{X}}(\mathcal{F}),
\end{aligned}$$

which completes the proof. ■

**Definition 44 (Star-hull)** *Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. The star-hull of $\mathcal{F}$ around $f^* : \mathcal{X} \to \mathbb{R}$ is defined by*

$$\mathrm{star}(\mathcal{F}, f^*) = \big\{f^* + \alpha(f - f^*) : f \in \mathcal{F}, \alpha \in [0,1]\big\}.$$

Notice that making a class star-hull increases the complexities. However, this increase is moderate as shown in the following lemma.

**Lemma 45 (Lemma 4.5 in Mendelson (2002))** *Let $f^* : \mathcal{X} \to [-B, B]$ and $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[-B, B]$ for some positive constant $B$. Then the following inequality holds for each $\varepsilon > 0$*

$$\log N(\varepsilon, \mathrm{star}(\mathcal{F}, f^*), L^\infty(\mathfrak{X})) \leq \log N(\varepsilon/2, \mathcal{F}, L^\infty(\mathfrak{X})) + \log(4B/\varepsilon).$$

**Proof of Lemma 45** Let $\mathcal{F}_\varepsilon$ be an $L^\infty(\mathcal{X})$ $(\varepsilon/2)$-cover of $\mathcal{F}$ such that

$$N = |\mathcal{F}_\varepsilon| = N(\varepsilon/2, \mathcal{F}, L^\infty(\mathcal{X})).$$

Denote $\mathcal{F}_\varepsilon = \{f_j\}_{j=1}^N$. Without loss of generality, we assume that $|f_j(X_i)| \leq B$ for each $i \in [m]$. Then for each $f \in \mathcal{F}$ there exists $j \in [N]$ such that $\max_{1\leq i\leq n} |f(X_i) - f_j(X_i)| \leq \varepsilon/2$. Denote by $I(f^*, f_j)$ the segment between $f^*$ and $f_j$:

$$I(f^*, f_j) = \big\{f^* + \alpha(f_j - f^*) : \alpha \in [0,1]\big\}.$$

Furthermore, we construct an $L^\infty(\mathcal{X})$ $(\varepsilon/(4B))$-cover of it by

$$I_\varepsilon(f^*, f_j) = \Big\{f^* + \alpha_k(f_j - f^*) : \alpha_k = \frac{k\varepsilon}{4B}, k = 1, \ldots, \Big\lfloor \frac{4B}{\varepsilon} \Big\rfloor\Big\}.$$

Observe that $\cup_{j=1}^N I(f^*, f_j)$ is an $L^\infty(\mathcal{X})$ $\varepsilon$-cover of $\mathrm{star}(\mathcal{F}, f^*)$. Indeed, it holds that

$$\begin{aligned}
&\max_{1\leq i\leq n} |f^*(X_i) + \alpha(f(X_i) - f^*(X_i)) - f^*(X_i) - \alpha_k(f_j(X_i) - f^*(X_i))| \\
&\leq |\alpha - \alpha_k| \max_{1\leq i\leq n} |f^*(X_i)| + \alpha \max_{1\leq i\leq n} |f(X_i) - f_j(X_i)| + |\alpha - \alpha_k| \max_{1\leq i\leq n} |f_j(X_i)| \\
&\leq \frac{\varepsilon}{4B}B + \frac{\varepsilon}{2} + \frac{\varepsilon}{4B}B = \varepsilon.
\end{aligned}$$

Therefore, it follows that $N(\varepsilon, \mathrm{star}(\mathcal{F}, f^*), L^\infty(\mathcal{X})) \leq N(\varepsilon/2, \mathcal{F}, L^\infty(\mathcal{X}))4B/\varepsilon$, which completes the proof. ■

### E.2 Oracle Inequality of Density Ratio Estimator

To begin with, we have $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ i.i.d. drawn from the probability distribution $\mu$, which denotes the distribution of $(X^\mu, Z^\mu)$ defined in Remark 5. Let $\mu_X$ be the marginal distribution of $X^\mu$. It is easy to verify that $2\|u\|_{L^2(\mu_X)}^2 = \|u\|_{L^2(P_X)}^2 + \|u\|_{L^2(Q_X)}^2$. Let $u^* = -\log \varrho$ and define the pre-training excess risk $R^{\mathrm{pre}}(u)$ by $R^{\mathrm{pre}}(u) = L^{\mathrm{pre}}(u) - L^{\mathrm{pre}}(u^*)$.

**Lemma 46** *Suppose Assumptions 4, 6 and 8 hold. Let $u^* = -\log \varrho$. Then it follows that*

$$\frac{1}{2}\min\Big\{\frac{\Lambda}{(1+\Lambda)^2}, \frac{\lambda}{(1+\lambda)^2}\Big\}\|u - u^*\|_{L^2(\mu_X)}^2 \leq L^{\mathrm{pre}}(u) - L^{\mathrm{pre}}(u^*) \leq \frac{1}{8}\|u - u^*\|_{L^2(\mu_X)}^2.$$

**Proof of Lemma 46** Denote $\ell_{\mathrm{logit}}(v, z) = \log(1 + \exp(-zv))$. According to Taylor's expansion, we find that

$$\begin{aligned}
&\ell_{\mathrm{logit}}(u(X), Z) - \ell_{\mathrm{logit}}(u^*(X), Z) \\
&= -\frac{u(X) - u^*(X)}{1 + \exp(Zu^*(X))} + \frac{1}{2}\frac{\exp(Z\theta(X))}{(1 + \exp(Z\theta(X)))^2}(u(X) - u^*(X))^2,
\end{aligned}$$

where $\theta(x) = cu(x) + (1-c)u^*(x)$ for some $c \in [0,1]$. Taking expectation with respect to $(X, Z) \sim \mu$ yields

$$L^{\mathrm{pre}}(u) - L^{\mathrm{pre}}(u^*) = \mathbb{E}_{(X,Z)\sim\mu}\Big[\frac{1}{2}\frac{\exp(Z\theta(X))}{(1 + \exp(Z\theta(X)))^2}(u(X) - u^*(X))^2\Big],$$

where we used the fact that $u^*$ is the minimizer of $L^{\mathrm{pre}}(\cdot)$ over $\mathscr{L}(\mathcal{X})$. According to Assumptions 4 and 8, the following inequality holds for each $x \in \mathcal{X}$

$$\min\left\{\frac{\Lambda}{(1+\Lambda)^2}, \frac{\lambda}{(1+\lambda)^2}\right\} \leq \frac{\exp(\theta(x))}{(1+\exp(\theta(x)))^2} \leq \frac{1}{4},$$

which completes the proof. ■

**Lemma 47** *Suppose Assumptions 4, 6 and 8 hold. Let* $u^* = -\log \varrho$. *Let* $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ *be an i.i.d. sample set drawn from* $\mu$. *Let* $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ *be defined in* (11)*. Define the distribution-dependent measurement functional*

$$T(u) = \mathbb{E}_{X^\mu \sim \mu_X}\big[(u(X^\mu) - u^*(X^\mu))^2\big].$$

*Assume that* $\psi$ *ia a sub-root function for which*

$$\psi(r) \geq \nu \mathfrak{R}_m\Big(\big\{u \in \mathcal{U} : T(u) \leq r\big\}\Big).$$

*Further, let* $r^*$ *be the fixed point of* $\psi$*. Then for each* $\delta \in (0,1)$ *the following inequality holds with probability as least* $1 - 2\delta$

$$R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq 3 \inf_{u \in \mathcal{U}} R^{\mathrm{pre}}(u) + \frac{4224}{\nu} r^* + \frac{(132M + 156\nu)\log(1/\delta)}{m},$$

*where* $\nu = \frac{1}{2}\max\{\frac{(1+\Lambda)^2}{\Lambda}, \frac{(1+\lambda)^2}{\lambda}\}$ *and* $M = \max\{\log(1+1/\lambda), \log(1+\Lambda)\}$.

**Proof of Lemma 47** For simplicity of notations, we define

$$g(u,x,z) = \ell_{\mathrm{logit}}(u(x),z) - \ell_{\mathrm{logit}}(u^*(x),z), \quad \ell_{\mathrm{logit}}(v,z) = \log(1+\exp(-zv)).$$

Notice that the pre-training excess risk satisfies $R^{\mathrm{pre}}(u) = \mathbb{E}_{(X,Z)\sim\mu}[g(u,X,Z)]$. Since $\ell_{\mathrm{logit}}(u(x),z)$ is 1-Lipschitz with respect to $u(x)$ for each $z \in \{\pm 1\}$, which deduces that the following inequality holds for each $(x,z) \in \mathcal{X} \times \{\pm 1\}$

$$|g(u,x,z)| = |\ell_{\mathrm{logit}}(u(x),z) - \ell_{\mathrm{logit}}(u^*(x),z)| \leq |u(x) - u^*(x)|.$$

As a consequence, it holds that

$$\mathbb{E}_{(X^\mu,Z^\mu)\sim\mu}\big[g^2(u,X^\mu,Z^\mu)\big] \leq \|u - u^*\|_{L^2(\mu_X)}^2 = T(u), \tag{39}$$

where $T(u)$ is called the measurement functional. On the other hand, using Lemma 46 yields

$$T(u) \leq \nu R^{\mathrm{pre}}(u) = \nu \mathbb{E}_{(X^\mu,Z^\mu)\sim\mu}\big[g(u,X^\mu,Z^\mu)\big]. \tag{40}$$

By Ledoux-Talagrand contraction inequality Ledoux and Talagrand (1991), we have

$$\mathfrak{R}_m\Big(\big\{g \circ u : u \in \mathcal{U}, T(u) \leq r\big\}\Big) \leq \mathfrak{R}_m\Big(\big\{u \in \mathcal{U} : T(u) \leq r\big\}\Big), \tag{41}$$

which implies that

$$\psi(r) \geq \nu \mathfrak{R}_m\Big(\big\{g \circ u : u \in \mathcal{U}, T(u) < r\big\}\Big).$$

By applying (37) in Lemma 40 to the function $g \circ u$, (39), (40) and (41) deduce that the following inequality holds with probability as least $1-\delta$ for each $u \in \mathcal{U}$

$$R^{\mathrm{pre}}(u) \leq 2\widehat{R}_{\mathcal{S}}^{\mathrm{pre}}(u) + \frac{1408}{\nu} r^* + \frac{(44M + 52\nu)\log(1/\delta)}{m},$$

where we used the fact that $|g(u,x,z)| \leq |u(x) - u^*(x)| \leq 2M$ for each $(x,z) \in \mathcal{X} \times \{\pm 1\}$ from Assumptions 4 and 8. Since $\widehat{u}_{\mathcal{S}}$ is the minimizer of $\widehat{R}_{\mathcal{S}}^{\mathrm{pre}}(\cdot)$ over $\mathcal{U}$, we have that with probability as least $1-\delta$ for each $u \in \mathcal{U}$,

$$R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq 2\widehat{R}_{\mathcal{S}}^{\mathrm{pre}}(u) + \frac{1408}{\nu} r^* + \frac{(44M + 52\nu)\log(1/\delta)}{m}, \tag{42}$$

Further, using (38) in Lemma 40 gives that the following inequality holds with with probability as least $1-\delta$ for each $u \in \mathcal{U}$,

$$\widehat{R}_{\mathcal{S}}^{\mathrm{pre}}(u) \leq \frac{3}{2} R^{\mathrm{pre}}(u) + \frac{1408}{\nu} r^* + \frac{(44M + 52\nu)\log(1/\delta)}{m}. \tag{43}$$

Combining (43) and (42) yields the desired result. ∎

The results in Lemma 47 use distribution-dependent measures of complexity of the class. By a similar technique as the proof of Bartlett et al. (2005, Lemma 3.4), we next provide error bounds which can be identified directly from the sample set, without a priori information.

**Lemma 48** *Suppose Assumptions 4, 6 and 8 hold. Let $u^* = -\log \varrho$. Let $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ be an i.i.d. sample set drawn from $\mu$. Let $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ be defined in* (11)*. Define the data-dependent measurement functional*

$$\widehat{T}_{\mathcal{S}}(u) = \frac{1}{m}\sum_{i=1}^m (u(X_i^\mu) - u^*(X_i^\mu))^2$$

*Assume that $\widehat{\psi}_{\mathcal{S}}$ ia a sub-root function for which*

$$\widehat{\psi}_{\mathcal{S}}(r) \geq 2(12M + \nu)\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U}, u^*) : \widehat{T}_{\mathcal{S}}(u) \leq 2r\big\}\Big) + \frac{(36M^2 + 2\nu M)\log(1/\delta)}{m}.$$

*Further, let $\widehat{r}_{\mathcal{S}}^*$ be the fixed point of $\widehat{\psi}_{\mathcal{S}}$ and $r \geq \widehat{r}_{\mathcal{S}}^*$. Then for each $\delta \in (0,1)$ the following inequality holds with probability at least $1 - 4\delta$*

$$R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq 3 \inf_{u \in \mathcal{U}} R^{\mathrm{pre}}(u) + \frac{4224}{\nu} r + \frac{(132M + 156\nu)\log(1/\delta)}{m},$$

*where $\nu = \frac{1}{2}\max\{\frac{(1+\Lambda)^2}{\Lambda}, \frac{(1+\lambda)^2}{\lambda}\}$ and $M = \max\{\log(1+1/\lambda), \log(1+\Lambda)\}$.*

**Proof of Lemma 48** We use the same definition of the measurement functional $T(u)$ as in Lemma 47. The proof is divided into two steps.

*Step 1. Construct a non-trivial sub-root function.*

Let $\psi$ be a sub-root function satisfying

$$\psi(r) \geq \nu\mathfrak{R}_m\Big(\big\{u \in \mathcal{U} : T(u) \leq r\big\}\Big), \tag{44}$$

and

$$\psi(r) \geq 12M\mathfrak{R}_m\Big(\big\{u \in \mathcal{U} : T(u) \leq r\big\}\Big) + \frac{12M^2\log(1/\delta)}{m}. \tag{45}$$

To this end, we set $\psi$ as

$$\psi(r) = (12M + \nu)\mathfrak{R}_m\Big(\big\{u \in \mathrm{star}(\mathcal{U}, u^*) : T(u) \leq r\big\}\Big) + \frac{12M^2\log(1/\delta)}{m}. \tag{46}$$

Now we show that $\psi$ is a sub-root function. By Jensen's inequality, we find that the Rademacher complexity is non-negative and thus $\psi$ is non-negative. Furthermore, $\psi$ is non-decreasing since the following holds for $r \leq r'$

$$\big\{u \in \mathrm{star}(\mathcal{U}, u^*) : T(u) \leq r\big\} \subseteq \big\{u \in \mathrm{star}(\mathcal{U}, u^*) : T(u) \leq r'\big\}.$$

It remains to show that for each $0 < r_1 < r_2$, it holds that $\psi(r_1) \geq \sqrt{r_1/r_2}\psi(r_2)$. According to (46) and Lemma 43, we only need to verify

$$\begin{aligned}&\mathfrak{R}_m\Big(\big\{v \in \mathrm{star}(\mathcal{U} - u^*, 0) : \mathbb{E}_{X^\mu\sim\mu_X}\big[v^2(X^\mu)\big] \leq r_1\big\}\Big) \\ &\geq \sqrt{\frac{r_1}{r_2}}\mathfrak{R}_m\Big(\big\{v \in \mathrm{star}(\mathcal{U} - u^*, 0) : \mathbb{E}_{X^\mu\sim\mu_X}\big[v^2(X^\mu)\big] \leq r_2\big\}\Big).\end{aligned} \tag{47}$$

Fix a sample set $\{X_i\}_{i=1}^m$ drawn from $\mu_X$ and a set of Rademacher variables $\{\varepsilon_i\}_{i=1}^m$, define

$$\eta = \sup\Big\{\frac{1}{m}\sum_{i=1}^m \varepsilon_i v(X_i^\mu) : v \in \mathrm{star}(\mathcal{U} - u^*, 0), \mathbb{E}_{X\sim\mu_X}\big[v^2(X^\mu)\big] \leq r_2\Big\}.$$

Let $\{v_k\}_{k=1}^\infty \subseteq \{v \in \mathrm{star}(\mathcal{U} - u^*, 0), \mathbb{E}_{X\sim\mu_X}[v^2(X^\mu)] \leq r_2\}$ be a sequence of functions for which

$$\eta = \lim_{k\to\infty}\frac{1}{n}\sum_{i=1}^m \varepsilon_i v_k(X_i^\mu) \quad \text{and} \quad \eta \geq \frac{1}{m}\sum_{i=1}^m \varepsilon_i v_k(X_i^\mu), \quad k \geq 1$$

Since $\mathbb{E}_{X^\mu\sim\mu_X}[v_k^2(X^\mu)] \leq r_2$, we find that $\mathbb{E}_{X^\mu\sim\mu_X}[(\sqrt{r_1/r_2}v_k(X^\mu))^2] \leq r_1$. Moreover, by the definition of star-hull, we find that $\sqrt{r_1/r_2}v_k \in \mathrm{star}(\mathcal{U} - u^*, 0)$ for each $k \geq 1$. Thus the following inequality holds for each $k \geq 1$

$$\sup\Big\{\frac{1}{m}\sum_{i=1}^m \varepsilon_i v(X_i^\mu) : v \in \mathrm{star}(\mathcal{U} - u^*, 0), \mathbb{E}_{X^\mu\sim\mu_X}[v^2(X^\mu)] \leq r_1\Big\} \geq \sqrt{\frac{r_1}{r_2}}\frac{1}{m}\sum_{i=1}^m \varepsilon_i v_k(X_i^\mu).$$

Then taking limitation as $k \to \infty$, we find that

$$\sup\Big\{\frac{1}{m}\sum_{i=1}^m \varepsilon_i v(X_i^\mu) : v \in \mathrm{star}(\mathcal{U} - u^*, 0), \mathbb{E}_{X^\mu\sim\mu_X}[v^2(X^\mu)] \leq r_1\Big\} \geq \sqrt{\frac{r_1}{r_2}}\eta.$$

Taking expectation with respect to the sample set $\{X_i^\mu\}_{i=1}^m$ and Rademacher variables $\{\varepsilon_i\}_{i=1}^n$ yields (47). Hence we conclude that $\psi(r)$ is a non-trivial sub-root function.
*Step 2. Data-dependent error bounds.*

Let $r^*$ be the fixed point of the sub-root function $\psi$ defined as (46). Since $\psi$ satisfies (44), using (47) implies that for $\delta \in (0,1)$ the following inequality holds with probability as least $1-2\delta$

$$R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq 3 \inf_{u\in\mathcal{U}} R^{\mathrm{pre}}(u) + \frac{4224}{\nu} r^* + \frac{(132M+156\nu)\log(1/\delta)}{m}. \tag{48}$$

Notice that (48) uses distribution-dependent measures of complexity of the function class. We next establish distribution-free bounds, which only depend on a sample set $\mathcal{S}$. Since that $\psi$ satisfies (45), applying Lemma 37 gives that with probability at least $1-\delta$,

$$\big\{u \in \mathrm{star}(\mathcal{U},u^*) : T(u) \leq r^*\big\} \subseteq \big\{u \in \mathrm{star}(\mathcal{U},u^*) : \widehat{T}_{\mathcal{S}}(u) \leq 2r^*\big\},$$

which means that the following inequality holds with probability at least $1-\delta$

$$\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : T(u) \leq r^*\big\}\Big) \leq \widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : \widehat{T}_{\mathcal{S}}(u) \leq 2r^*\big\}\Big). \tag{49}$$

In addition, we find from Lemma 35 that the following inequality holds with probability at least $1-\delta$

$$\begin{aligned} &\mathfrak{R}_m\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : T(u) \leq r^*\big\}\Big) \\ &\leq 2\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : T(u) \leq r^*\big\}\Big) + \frac{2M\log(1/\delta)}{m}. \end{aligned} \tag{50}$$

Combining (46), (49) and (50) obtains that the following inequality holds with probability at least $1-2\delta$

$$\begin{aligned} \psi(r^*) &\leq 2(12M+\nu)\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : T(u) \leq r^*\big\}\Big) + \frac{(36M^2+2\nu M)\log(1/\delta)}{m} \\ &\leq 2(12M+V)\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U},u^*) : \widehat{T}_{\mathcal{S}}(u) \leq 2r^*\big\}\Big) + \frac{(36M^2+2\nu M)\log(1/\delta)}{m} \\ &\leq \widehat{\psi}_{\mathcal{S}}(r^*). \end{aligned}$$

As a consequence, we have $r^* = \psi(r^*) \leq \widehat{\psi}_{\mathcal{S}}(r^*)$, which deduces $r^* \leq \widehat{r}_{\mathcal{S}}^*$ with probability $1-2\delta$ from Lemma 39. Recalling that (48) holds with probability at least $1-2\delta$ yields the desired result. ■

**Lemma 49** *Suppose Assumptions 4, (6) and (8) hold. Let $u^* = -\log\varrho$. Let $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ be an i.i.d. sample set drawn from $\mu$. Let $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ be defined as (11). Then the following inequality holds*

$$\mathbb{E}_{\mathcal{S}}\big[R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}})\big] \leq \inf_{u\in\mathcal{U}} R^{\mathrm{pre}}(u) + \frac{M^2+\nu^2}{\nu}\frac{\mathrm{VCdim}(\mathcal{U})\log(em)}{m},$$

*where $\nu = \frac{1}{2}\max\{\frac{(1+\Lambda)^2}{\Lambda}, \frac{(1+\lambda)^2}{\lambda}\}$ and $M = \max\{\log(1+1/\lambda), \log(1+\Lambda)\}$.*

**Proof of Lemma 49**

We divide the proof into two steps.

*Step 1. Oracle inequality with high-probability statement.*

Using Lemmas 41 and 45 and setting $\varepsilon = M/m$, we find that for $n \geq \mathrm{VCdim}(\mathcal{F})$,

$$\begin{aligned}\widehat{\mathfrak{R}}_{\mathcal{S}}\Big(\big\{u \in \mathrm{star}(\mathcal{U}, u^*) : \widehat{T}_{\mathcal{S}}(u) \leq 2r\big\}\Big) &\leq \frac{2M}{m} + 2\sqrt{\frac{\log\{4mN(M/(2m), \mathcal{U}, L^\infty(\mathcal{S}))\}}{m}}\sqrt{r} \\ &\leq \frac{2M}{m} + 4\sqrt{\frac{\mathrm{VCdim}(\mathcal{U})\log(em)}{m}}\sqrt{r},\end{aligned}$$

where the second inequality is due to Anthony et al. (1999, Theorem 12.2). Define the sub-root function $\widehat{\psi}_{\mathcal{S}}(r)$ by $\widehat{\psi}_{\mathcal{S}}(r) = a\sqrt{r} + b$, where

$$a = 8(12M + \nu)\sqrt{\frac{\mathrm{VCdim}(\mathcal{U})\log(em)}{m}} \quad \text{and} \quad b = \frac{(60M^2 + 4\nu M)\log(1/\delta)}{m}.$$

By setting $r = 4a^2 + 2b$, we find that $\widehat{\psi}_{\mathcal{S}}(r) \leq r$. Combining this with Lemma 39 implies $\widehat{r}_{\mathcal{S}}^* \leq r$. Then using Lemma 48 yields

$$\Pr(R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) > \varepsilon(\delta, n)) \leq 4\delta,$$

where

$$\varepsilon(\delta, n) = 3\inf_{u \in \mathcal{U}} R^{\mathrm{pre}}(u) + \frac{4224(4a^2 + 2b)}{\nu} + \frac{(132M + 156\nu)\log(1/\delta)}{m}.$$

*Step 2. Convergence rates of the density ratio estimator.*

Since $g(u, x, z) \leq 2M$ for each $u \in \mathcal{U}$ and $(x, z) \in \mathcal{X} \times \{\pm 1\}$, it follows that $R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq 2M$, and consequently,

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}})\Big] \\ &= \mathbb{E}_{\mathcal{S}}\Big[R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \cdot \mathbb{1}(R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) > \varepsilon(\delta, n))\Big] + \mathbb{E}_{\mathcal{S}}\Big[R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \cdot \mathbb{1}(R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) \leq \varepsilon(\delta, n))\Big] \\ &\leq 2M \cdot \Pr\{R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) > \varepsilon(\delta, n)\} + \varepsilon(\delta, n) \cdot (1 - \Pr\{R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}}) > \varepsilon(\delta, n)\}) \\ &\leq 8M\delta + \varepsilon(\delta, n),\end{aligned}$$

Setting $\delta = 1/m$ completes the proof. ∎

**Lemma 50 (Oracle inequality of density-ratio estimator)** *Suppose Assumptions 4, 6 and 8 hold. Let $\mathcal{S} = \{(X_i^\mu, Z_i^\mu)\}_{i=1}^m$ be an i.i.d. sample set drawn from (10). Suppose that $\mathcal{U}$ is a hypothesis class and $\widehat{u}_{\mathcal{S}} \in \mathcal{U}$ is defined by (11). Then the following inequality holds for $m \geq \mathrm{VCdim}(\mathcal{U})$,*

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}\sim\mu^m}\Big[\|\widehat{\varrho}_{\mathcal{S}} - \varrho\|_{L^2(P_X)}^2 + \|\widehat{\varrho}_{\mathcal{S}} - \varrho\|_{L^2(Q_X)}^2\Big] \\ &\leq \nu \inf_{u \in \mathcal{U}}\Big(\|u + \log\varrho\|_{L^2(P_X)}^2 + \|u + \log\varrho\|_{L^2(Q_X)}^2\Big) + (M^2 + \nu^2)\frac{\mathrm{VCdim}(\mathcal{U})}{m\log^{-1}(em)},\end{aligned}$$

*where $\nu = \frac{1}{2}\max\{\frac{(1+\Lambda)^2}{\Lambda}, \frac{(1+\lambda)^2}{\lambda}\}$ and $M = \max\{\log(1 + 1/\lambda), \log(1 + \Lambda)\}$.*

**Proof of Lemma 50** Observe that

$$\begin{aligned}\|\widehat{\varrho}_{\mathcal{S}}-\varrho\|_{L^2(\mu_X)}^2&=\mathbb{E}_{X^\mu\sim\mu_X}\big[(\widehat{\varrho}_{\mathcal{S}}(X^\mu)-\varrho(X^\mu))^2\big]\\&=\mathbb{E}_{X^\mu\sim\mu_X}\big[(\exp(-\widehat{u}_{\mathcal{S}}(X^\mu))-\exp(-u(X^\mu)))^2\big]\\&\leq\mathbb{E}_{X^\mu\sim\mu_X}\big[(\widehat{u}_{\mathcal{S}}(X^\mu)-u(X^\mu))^2\big]=\|\widehat{u}_{\mathcal{S}}-u\|_{L^2(\mu_X)}^2.\end{aligned}$$

Then according to Lemma 49, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\big[\|\widehat{u}_{\mathcal{S}}-u^*\|_{L^2(\mu_X)}^2\big]&\leq 4\nu\mathbb{E}_{\mathcal{S}}\big[R^{\mathrm{pre}}(\widehat{u}_{\mathcal{S}})\big]\\&\leq\frac{\nu}{2}\inf_{u\in\mathcal{U}}\|u-u^*\|_{L^2(\mu_X)}^2+4(M^2+\nu^2)\frac{\mathrm{VCdim}(\mathcal{U})}{m\log^{-1}(em)},\end{aligned}$$

where we used Lemma 46. Combining the above two inequalities yields the desired result. ■

### E.3 Convergence rates of Density Ratio Estimator

**Proof of Lemma 24** According to Assumption 7 and Lemma 31, there exists $u\in\mathcal{U}=N(W_{\mathcal{U}},L_{\mathcal{U}})$ such that

$$\inf_{u\in\mathcal{U}}\|u+\log\varrho\|_{L^2(Q_X)}^2+\|u+\log\varrho\|_{L^2(P_X)}^2\leq C_1(U_{\mathcal{U}}N_{\mathcal{U}})^{-4\alpha/d},$$

where $W_{\mathcal{U}}=\mathcal{O}(U_{\mathcal{U}}\log U_{\mathcal{U}})$, $L_{\mathcal{U}}=\mathcal{O}(N_{\mathcal{U}}\log N_{\mathcal{U}})$ and $C_1$ is a constant only depending on $d$, $\alpha$ and $\|\log\varrho\|_{\mathcal{H}^\alpha(\mathcal{X})}$. Using Lemma 32, we find that $\mathrm{VCdim}(\mathcal{U})\leq C_2U_{\mathcal{U}}^2N_{\mathcal{U}}^2(\log U_{\mathcal{U}}\log N_{\mathcal{U}})^2$, where the constant $C_2$ depends on $d$ and $\alpha$. By setting

$$U_{\mathcal{U}}N_{\mathcal{U}}=\mathcal{O}\Big(\nu^{\frac{d}{2d+4\alpha}}(M^2+\nu^2)^{-\frac{d}{2d+4\alpha}}m^{\frac{d}{2d+4\alpha}}\Big),$$

we conclude that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\big[\|\widehat{\varrho}_{\mathcal{S}}-\varrho^*\|_{L^2(\mu_X)}^2\big]&\leq\mathcal{O}\Big(\nu^{-\frac{2\alpha}{d+2\alpha}}(M^2+\nu^2)^{\frac{2\alpha}{d+2\alpha}}m^{-\frac{2\alpha}{d+2\alpha}}\Big)\\&\leq\mathcal{O}\Big(\nu^{\frac{2\alpha}{d+2\alpha}}m^{-\frac{2\alpha}{d+2\alpha}}\Big)\leq\mathcal{O}\Big(\max\big\{\Lambda^{\frac{2\alpha}{d+2\alpha}},\lambda^{-\frac{2\alpha}{d+2\alpha}}\big\}m^{-\frac{2\alpha}{d+2\alpha}}\Big),\end{aligned}$$

where we used the fact that $\nu\geq M$ and $\Lambda\geq 1$, $\lambda\leq 1$. This completes the proof. ■

## References


Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural Network Learning: Theoretical Foundations*, volume 9. Cambridge university press Cambridge, 1999.

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

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-Tight VC-Dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Benedikt Bauer and Michael Kohler. On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of machine learning research*, 7(11), 2006.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative Learning for Differing Training and Test Distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative Learning under Covariate Shift. *Journal of Machine Learning Research*, 10(9), 2009.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric Regression on Low-Dimensional Manifolds using Deep ReLU Networks: Function Approximation and Statistical Recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 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. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Imre Csiszár. On Information-Type Measure of Difference of Probability Distributions and Indirect Observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Yann Dubois, Yangjun Ruan, and Chris J Maddison. Optimal Representations for Covariate Shifts. In *NeurIPS 2021 workshop on distribution shifts: connecting methods and applications*, 2021.

John C Duchi and Hongseok Namkoong. Learning Models with Uniform Performance via Distributionally Robust Optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking Importance Weighting for Deep Learning under Distribution Shift. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11996–12007. Curran Associates, Inc., 2020.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1):181–213, 2021.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A Survey on Concept Drift Adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge university press, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

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

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting Sample Selection Bias by Unlabeled Data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

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.

Michael Kohler and Sophie Langer. On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a Function of Low Local Dimensionality by Deep Neural Networks. *IEEE transactions on information theory*, 68 (6):4032–4042, 2022.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-Of-Distribution Generalization via Risk Extrapolation (REX). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

Kun Kuang, Hengtao Zhang, Runze Wu, Fei Wu, Yueting Zhuang, and Aijun Zhang. Balance-Subsampled Stable Prediction Across Unknown Test Data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3):1–21, 2021.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with Square Loss: Localization through Offset Rademacher Complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

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.

Shahar Mendelson. Improving the Sample Complexity using Global Data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21 (174):1–38, 2020.

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Philipp Petersen and Felix Voigtlaender. Optimal Approximation of Piecewise Smooth Functions using Deep ReLU Neural Networks. *Neural Networks*, 108:296–330, 2018.

Joaquin Quinoñero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset Shift in Machine Learning*. Mit Press, 2009.

Johannes Schmidt-Hieber. Nonparametric Regression using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. Stable Learning via Sample Reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34. AAAI Press, Palo Alto, California USA, 2020. doi: 10.1609/aaai.v34i04.6024.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey, 2021.

Zuowei Shen. Deep Network Approximation Characterized by Number of Neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear Approximation via Compositions. *Neural Networks*, 119:74–84, 2019.

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.

Nathan Srebro and Karthik Sridharan. Note on Refined Dudley Integral Covering Number Bound. 2010.

Charles J Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, pages 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 Amos J Storkey. Mixture Regression for Covariate Shift. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, 8(5), 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. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007b.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct Importance Estimation for Covariate Shift Adaptation. *Annals of the Institute of Statistical Mathematics*, 60:43–76, Jan 2008. ISSN 0018-9219. doi: 10.1007/s10463-008-0197-x.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. A Two-Stage Weighting Framework for Multi-Source Domain Adaptation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

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

Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

Aad W Van Der Vaart and Jon A Wellner. Weak Convergence and Empirical Processes: with Applications to Statistics. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

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

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge university press, 2019.

Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, and Peng Cui. A Theoretical Analysis on Independence-Driven Importance Weighting for Covariate-Shift Generalization. In *International Conference on Machine Learning*, pages 24803–24829. PMLR, 2022.

Yuhong Yang. Minimax nonparametric classification – Part I. Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999. doi: 10.1109/18.796368.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. doi: 10.1214/aos/1017939142.

Dmitry Yarotsky. Optimal Approximation of Continuous Functions by Very Deep ReLU Networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The Phase Diagram of Approximation Rates for Deep Neural Networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Zihan Zhang, Lei Shi, and Ding-Xuan Zhou. Classification with deep neural networks and logistic loss, 2023.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.