

ESTIMATING UNBOUNDED DENSITY RATIOS: APPLICATIONS IN ERROR CONTROL UNDER COVARIATE SHIFT

BY SHUNTUO XU^{1,a}, ZHOU YU^{1,b} AND JIAN HUANG^{*2,c}

¹*School of Statistics, East China Normal University, aoksword@163.com; ^bzyu@stat.ecnu.edu.cn*

²*Departments of Data Science and Artificial Intelligence, and Applied Mathematics
The Hong Kong Polytechnic University, ^cj.huang@polyu.edu.hk*

The density ratio is an important metric for evaluating the relative likelihood of two probability distributions, with extensive applications in statistics and machine learning. However, existing estimation theories for density ratios often depend on stringent regularity conditions, mainly focusing on density ratio functions with bounded domains and ranges. In this paper, we study density ratio estimators using loss functions based on least squares and logistic regression. We establish upper bounds on estimation errors with standard minimax optimal rates, up to logarithmic factors. Our results accommodate density ratio functions with unbounded domains and ranges. We apply our results to nonparametric regression and conditional flow models under covariate shift and identify the tail properties of the density ratio as crucial for error control across domains affected by covariate shift. We provide sufficient conditions under which loss correction is unnecessary and demonstrate effective generalization capabilities of a source estimator to any suitable target domain. Our simulation experiments support these theoretical findings, indicating that the source estimator can outperform those derived from loss correction methods, even when the true density ratio is known.

1. Introduction. The density ratio is a crucial metric for assessing the relative likelihood of two probability distributions. By comparing the densities of these distributions, the density ratio quantifies how one distribution differs from another. It has extensive applications across various areas, including nonparametric regression [55, 59], generative learning [14, 18, 21], change-point detection [36, 65], and reinforcement learning [8, 35]. In this paper, we study the theoretical properties of density ratio estimation using Bregman divergence. Our results accommodate density ratio functions with unbounded domains and ranges. We apply our results to covariate shift problems in the context of nonparametric regression and conditional distribution estimation using continuous flow models.

Let X^s and X^t represent two d -dimensional random vectors corresponding to the source and target domains, respectively. We assume that their probability measures are absolutely continuous with respect to the Lebesgue measure. As a consequence, X^s and X^t admit well-defined probability density functions, denoted by $p(x)$ and $q(x)$, respectively. Furthermore, assume that X^t is absolutely continuous with respect to X^s . To be more specific, by defining $\mathcal{X}^s = \{x : p(x) > 0\}$ and $\mathcal{X}^t = \{x : q(x) > 0\}$, we suppose that $\mathcal{X}^t \subset \mathcal{X}^s$. The density ratio is then defined as $r_0(x) = q(x)/p(x) \in [0, \infty)$, where we adopt the convention that $0/0 = 0$. Covariate shift occurs when $p(x) \neq q(x)$ but the conditional distribution of the corresponding response variable given the covariate remains constant across both domains.

In practice, we only observe samples $\{X_1^s, \dots, X_n^s\}$ and $\{X_1^t, \dots, X_n^t\}$ from the source and target domains. Therefore, the true density ratio remains unknown and must be estimated.

MSC2020 subject classifications: Primary 62G05, 62G08; secondary 68T07.

Keywords and phrases: Conditional flow models, deep neural network, local Hölder class, nonparametric regression, transfer learning.

Various density ratio estimation methods have been proposed in the literature [29, 47, 56, 58]. However, the theoretical analyses accompanying these estimators are notably limited. Most theoretical advancements were derived under stringent conditions [33, 42, 68], often assuming that the density ratio was bounded from above or below, which may not be satisfied in practice.

We observe that unbounded density ratios are quite common. For instance, consider a scenario where the source distribution is $\text{Gamma}(1, 1)$ and the target distribution is $\text{Gamma}(2, 1)$. It is clear that the density ratio $r_0(x) = x\mathbf{1}(x > 0)$ diverges as $x \rightarrow \infty$. In this example, both the domain and range of $r_0(x)$ are unbounded, and existing results on estimation error that assume a bounded density ratio do not apply to this simple case.

We study the estimation error of density ratio estimators when both the domain and range of $r_0(x)$ are allowed to be unbounded. The estimators we consider are established based on the Bregman divergence induced by certain differentiable and strictly convex functions [5, 27]. Particularly, we focus on two specific cases, including the least squares loss and the logistic regression-based loss. It is important to note that these two types of loss functions present distinct continuity patterns, therefore, different regularity conditions are required for their respective analyses. Our results show that even when the density ratio is not bounded from both above and below, the estimation can still achieve nearly minimax optimal results, up to factors of logarithms.

Recently, [12] established a theoretical guarantee in handling unbounded density ratio functions under mild moment conditions. They required the truncated density ratio function to belong to a Hölder class defined on $[0, 1]^d$. This limitation inadvertently restricts flexibility when dealing with unbounded covariate domains. Furthermore, the rationale behind this constraint appears to be mainly for technical convenience rather than based on practical considerations. In contrast, the local Hölder class we propose in Subsection 2.2 effectively addresses challenges associated with unbounded covariate domains while enhancing interpretability.

Furthermore, another significant deficiency of assuming a bounded density ratio relates to downstream tasks. When the density ratio is bounded by a universal constant, the estimation process may be redundant in a supervised learning task subject to covariate shift. An estimator derived solely based on source data can potentially generalize to the target domain without loss of efficacy, provided that $\sup_x r_0(x) \leq B$ for some constant $B > 0$, in the sense that the expected excess risk in the target domain exhibits the same convergence rate to that in the source domain. Specifically, [39] demonstrated such property in RKHS-based nonparametric regression. This raises a natural question: Is this still true for unbounded density ratios, and if so, under what conditions?

To address this question, the tail behavior of the density ratio $r_0(X^s)$ is crucial. In more detail, the assumption that $r_0(X^s)$ is sub-exponentially distributed, articulated in our analysis of density ratio estimation, enables us to concentrate on the region where $r_0(X^s) \leq c \cdot \log n$ with c being a constant. Contributions from the tail beyond this range have a negligible impact on the upper bound of the expected excess risk. This observation motivates us to rethink the necessity of loss correction (also termed as importance reweighting) through density ratios [10, 37, 39]. Surprisingly and interestingly, we discover that, in the absence of such correction, the expected excess risk of a source estimator in the target domain can still be effectively controlled by its counterpart in the source domain, provided that the tail of $r_0(X^s)$ is not excessively heavy. To substantiate this phenomenon, we develop a series of general results with progressively relaxed assumptions. Furthermore, we illustrate these results through two specific cases related to nonparametric regression and conditional distribution estimation using continuous flow models.

To summarize, this paper makes two significant contributions. Firstly, we extend the estimation theory for density ratios to include cases with unbounded domains and ranges. By

carefully analyzing the divergence pattern of $r_0(x)$ and the tail behavior of $r_0(x^s)$ over the source domain, we obtain nearly optimal results. Our estimators are implemented using deep neural networks, which are particularly well-suited for handling unbounded multivariate functions using the truncation technique. Secondly, in nonparametric regression and conditional flow models under covariate shift, we identify specific regularity conditions under which classical loss correction is unnecessary, while still maintaining control over the expected excess risk in the target domain. This finding suggests that a plug-in strategy can be effectively employed in downstream tasks within the target domain by using an estimator derived from the source domain, especially when the source data is significantly more accessible.

The remainder of this paper is organized as follows: In Section 2, we provide a brief introduction to density ratio estimation using Bregman divergence. Section 3 presents the theoretical results concerning density ratio estimators derived from two types of loss functions, which are special cases of Bregman divergence. In Section 4, we provide the conditions under which loss correction is unnecessary under covariate shift. In Section 5, we apply these results to nonparametric regression and conditional flow models under covariate shift. Section 6 includes several numerical experiments that support our theoretical findings. Finally, Section 7 offers a brief conclusion along with further discussions. Technical details are included in the Supplementary Materials.

2. Preliminaries. In this section, we present the preliminaries that will be important in the subsequent sections.

2.1. Bregman divergence. The Bregman divergence [5] quantifies the difference between two points based on a differentiable and strictly convex function φ . We present a formal definition of the Bregman divergence below.

DEFINITION 2.1 (Bregman divergence). Let $\varphi : \mathcal{X} \rightarrow \mathbb{R}$ be a differentiable and strictly convex function where $\mathcal{X} \subset \mathbb{R}$ is a convex set. The Bregman divergence associated with φ for two scalars x and y , denoted as $D_\varphi(x\|y)$, is defined by

$$D_\varphi(x\|y) = \varphi(x) - \varphi(y) - \varphi'(y)(x - y),$$

where φ' represents the derivative of φ .

Due to the convexity of φ , it follows that $D_\varphi(x\|y) = 0$ implies $x = y$ (see the Supplementary Materials for more details). When $\varphi(x) = \varphi_{\text{LS}}(x) = (x - 1)^2$, the Bregman divergence $D_\varphi(x\|y)$ coincides with the least squares loss, specifically expressed as $D_{\text{LS}}(x\|y) = (x - y)^2$. On the other hand, when $\varphi(x) = \varphi_{\text{LR}}(x) = x \log x - (x + 1) \log(x + 1)$, the Bregman divergence takes the form $D_\varphi(x\|y) = x \log x - (x + 1) \log(x + 1) + \log(y + 1) - x \log y + x \log(y + 1)$, which is associated with the logistic regression-based loss. This divergence will be referred to as $D_{\text{LR}}(x\|y)$ hereafter.

2.2. Local Hölder class. The Hölder continuous condition is basic for analyzing the risk convergence rate of estimators in density estimation [69] and nonparametric regression [19, 25]. Definition 2.2 provides a characterization of the Hölder class.

DEFINITION 2.2 (Hölder class). Let $\beta = s + r$ where $s = \lfloor \beta \rfloor \in \mathbb{N}$ and $r \in (0, 1]$. Here, $\lfloor \beta \rfloor$ denotes the integer strictly smaller than β and \mathbb{N} is the set of nonnegative integers. For

a finite constant $B > 0$, the Hölder class of functions defined on the d -dimensional unit hypercube and yielding a scalar as output, denoted by $\mathcal{H}^\beta([0, 1]^d, B)$, is defined as

$$\mathcal{H}^\beta([0, 1]^d, B) = \left\{ f : [0, 1]^d \rightarrow \mathbb{R}, \max_{\|\alpha\|_1 \leq s} \|\partial^\alpha f\|_\infty \leq B, \max_{\|\alpha\|_1 = s} \sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_2^r} \leq B \right\}.$$

Here, $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{N}^d$, $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$, and $\partial^\alpha = \partial^{\alpha_1} \dots \partial^{\alpha_d}$. In addition, we call β the smoothness index.

Furthermore, for $u \in \mathbb{R}_+$ where \mathbb{R}_+ denotes the set of positive real numbers, let $B_u : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a function of u . For arbitrary function f , let $f|_{\mathcal{X}}$ represent the function f constrained on a domain \mathcal{X} , such that

$$f|_{\mathcal{X}}(x) = \begin{cases} x, & x \in \mathcal{X}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the local Hölder class is defined in Definition 2.3, which is a natural extension for the original Hölder class.

DEFINITION 2.3 (Local Hölder class). A local Hölder class with smoothness index β and divergence regime B_u , denoted by $\mathcal{H}_{\text{Loc}}^\beta(\mathbb{R}^d, B_u)$, is defined as

$$\mathcal{H}_{\text{Loc}}^\beta(\mathbb{R}^d, B_u) = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R}, g(x) = f|_{[-u, u]^d}(2ux - u1_d) \in \mathcal{H}^\beta([0, 1]^d, B_u) \text{ for any } u > 0 \right\}.$$

Here, 1_d represents the d -dimensional vector with all entries equal to 1.

REMARK 1. Consider the function $f(x) = x^m$ for $x \in \mathbb{R}$, where m is an integer no less than 2. Let $s = m - 1$ and $r = 1$. For $u > 0$ and $x \in [-u, u]$, let $g(x) = f(2ux - u) = u^m(2x - 1)^m$. A straightforward calculation suggests that $f \in \mathcal{H}_{\text{Loc}}^m(\mathbb{R}^d, m!2^m u^m)$. Furthermore, it can be verified that $f \in \mathcal{H}_{\text{Loc}}^\ell(\mathbb{R}^d, m!2^m u^m)$ for any $\ell \geq m$.

2.3. Neural networks. A neural network implemented in the multi-layer perceptron (MLP) architecture comprises a series of linear transformations and nonlinear activations. While diverse innovative network architectures have been proposed, demonstrating impressive performance, such as convolutional neural networks [30], residual networks [20], and transformers [63], we focus on the most fundamental representation, MLP, in this paper. Mathematically, a function f implemented by an MLP with depth L can be expressed as

$$f(x) = \phi_{L+1} \circ \sigma_L \circ \phi_L \circ \dots \circ \sigma_1 \circ \phi_1(x),$$

where $\phi_i(x) = W_i x + b_i$ with W_i a matrix of d_i rows and d_{i-1} columns, b_i a d_i -dimensional vector, and we let all the activations be the rectified linear unit (relu), i.e., $\sigma_i(x) = \max(x, 0)$ functioning by element, for $i = 1, \dots, L + 1$. The width of a network is defined as $\max_{i=1, \dots, L} d_i$.

A truncated neural network function can be represented by a deeper neural network. To illustrate this, let us define a truncation operator, denoted by $T_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$ for $a < b$, expressed as

$$T_{a,b}(x) = \begin{cases} a, & x < a, \\ x, & a \leq x \leq b, \\ b, & x > b. \end{cases}$$

For the cases where $a < 0 < b$, it follows that

$$T_{a,b}(x) = \text{relu}(-\text{relu}(-x + b) + b) - \text{relu}(-\text{relu}(x - a) - a).$$

When $0 \leq a < b$, we have

$$T_{a,b}(x) = \text{relu}(-\text{relu}(-x + b) + b - a) + a.$$

This property facilitates the enforcement of boundedness within neural network functions, which is crucial in theoretical analysis where it is presumed that an estimator does not grow unrestrainedly.

For simplicity, we denote the function space consisting of elements implemented by MLPs with output dimension d_{out} , depth L , width M , number of parameters S , uniformly upper bounded by a scalar $\bar{\delta}$ and lower bounded by another scalar $\underline{\delta}$, as $\mathcal{F}_{\text{NN}}^{d_{\text{out}}}$. The values of $\bar{\delta}$ and $\underline{\delta}$, typically relied on the sample size, should be carefully determined to avoid sub-optimal or trivial results. When $d_{\text{out}} = 1$, we abbreviate $\mathcal{F}_{\text{NN}}^1$ to \mathcal{F}_{NN} .

2.4. Covariate shift. Covariate shift is a prevalent challenge in supervised learning [40], signifying that the training and testing data are collected from different domains, namely the source domain and the target domain. Mathematically, the source domain is represented by a random pair (X^s, Y^s) , where X^s denotes the covariate vector and Y^s is the response. Concurrently, the target domain is characterized by (X^t, Y^t) with Y^t remaining unobservable. Under covariate shift, the conditional distributions are assumed identical, specifically that $Y^s|X^s = x$ and $Y^t|X^t = x$ share the same distribution given x . However, the marginal distributions of X^s and X^t may differ, which is a key aspect.

Let us denote the observations from source domain as $\{(X_1^s, Y_1^s), \dots, (X_N^s, Y_N^s)\}$ and those from target domain as $\{X_1^t, \dots, X_n^t\}$. For a nonparametric quantity $\theta_0(X)$ of interest, an empirical estimate $\hat{\theta}_N(X)$ is presumed to be constructed, by solely utilizing the source data. It is well-established for numerous classical methods that $\hat{\theta}_N(X)$ yields a sound performance within the source domain, in the sense that $\mathbb{E}\|\hat{\theta}_N(X^s) - \theta_0(X^s)\|_2^2$ converges to zero and even achieves an optimal rate [62]. However, evaluating the performance of $\hat{\theta}_N(X)$ in the target domain, i.e., how $\mathbb{E}\|\hat{\theta}_N(X^t) - \theta_0(X^t)\|_2^2$ behaves, is not trivial.

Given that X^s and X^t possesses probability density functions, when the density ratio $r_0(x)$ is uniformly bounded, it becomes evident that $\hat{\theta}_N(X)$ generalizes effectively to the target domain, by noting that

$$\begin{aligned} \mathbb{E}\left\|\hat{\theta}_N(X^t) - \theta_0(X^t)\right\|_2^2 &= \mathbb{E}\left[\left\|\hat{\theta}_N(X^s) - \theta_0(X^s)\right\|_2^2 \cdot r_0(X^s)\right] \\ &\leq \|r_0\|_\infty \mathbb{E}\left\|\hat{\theta}_N(X^s) - \theta_0(X^s)\right\|_2^2. \end{aligned}$$

Here, $\|\cdot\|_\infty$ denotes the supremum norm of a function. Conversely, when $r_0(x)$ is unbounded, the situation becomes significantly more complicated, which necessitates a deeper investigation towards various types of discrepancies.

2.5. Flow-based generative learning. Beyond regression tasks, modern machine learning is rapidly advancing to explore entire data distributions through generative models [17, 22, 52]. Flow-based generative models, among the notable recent developments, recover the data distribution from a base distribution (typically Gaussian noise) using bijective transformations [43, 46] or continuous-time dynamics [34, 38]. In this paper, we adopt the conditional stochastic interpolation framework [1, 23] due to its generality.

Specifically, let $\mathcal{I}(y_0, y_1, \tau)$ be a continuous interpolant connecting an initial point y_0 and a terminal point y_1 as the time parameter τ spans from 0 to 1, such that $\mathcal{I}(y_0, y_1, 0) = y_0$

and $\mathcal{I}(y_0, y_1, 1) = y_1$. Suppose that (X, Y) is a random pair of interest, where $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$. We consider the interpolation,

$$Y_\tau = \mathcal{I}(\eta, Y, \tau),$$

where η represents the d_y -dimensional standard Gaussian distribution and is independent of (X, Y) . Let $v_0(x, y, \tau)$ be the time-dependent velocity field defined by

$$v_0(x, y, \tau) = \mathbb{E} [\partial_\tau \mathcal{I}(\eta, Y, \tau) | Y_\tau = y, X = x].$$

We denote the conditional probability density function of $Y_\tau | X = x$ by $\omega(x, \cdot, \tau)$ for $\tau \in [0, 1]$. A key property is that the family of densities $\{\omega(x, \cdot, \tau) : \tau \in [0, 1]\}$ solves the following transport equation with respect to $\{\rho(\cdot, \tau) : \mathbb{R}^{d_y} \rightarrow \mathbb{R}, \tau \in [0, 1]\}$ when x is fixed [23],

$$(1) \quad \partial_\tau \rho(y, \tau) + \nabla_y \cdot [v_0(x, y, \tau) \rho(y, \tau)] = 0.$$

Here, $\nabla_y \cdot$ denotes the divergence operator over the variable y . Note that the transport equation (1) corresponds tightly to the following ordinary differential equation (ODE),

$$(2) \quad dZ_\tau = v_0(x, Z_\tau, \tau) d\tau,$$

for any fixed value of x . Under some regularity conditions, the terminal distribution deduced by ODE (2) at $\tau = 1$, with initial condition $Z_0 \sim N(0, I_{d_y})$, is essentially the distribution of $Y | X = x$ [15].

3. Density ratio estimation: error analysis. In this section, we conduct an in-depth analysis of density ratio estimators by examining two specific instances of the Bregman divergence: the least squares loss and the logistic regression loss. We establish error bounds for density ratio estimators based on these two loss functions.

3.1. Estimation based on least squares loss. In this subsection, we focus on a specific type of density ratio estimators deduced by the least squares loss function. As illustrated in Subsection 2.1, letting $\varphi(x) = \varphi_{\text{LS}}(x) = (x - 1)^2$ leads to that $D_{\text{LS}}(x \| y) = (x - y)^2$. Hence, by assuming that $r_0 \in \mathcal{L}^2(X^s)$, we have

$$\begin{aligned} r_0 &= \underset{f \in \mathcal{L}^2(X^s)}{\operatorname{argmin}} \mathbb{E} D_{\text{LS}}(r_0(X^s) \| f(X^s)) \\ &= \underset{f \in \mathcal{L}^2(X^s)}{\operatorname{argmin}} \mathbb{E} [r_0(X^s)^2] + \mathbb{E} [f(X^s)^2] - 2\mathbb{E} [f(X^s)r_0(X^s)] \\ &= \underset{f \in \mathcal{L}^2(X^s)}{\operatorname{argmin}} \mathbb{E} [f(X^s)^2] - 2\mathbb{E} [f(X^t)]. \end{aligned}$$

Here, $\mathcal{L}^2(X^s)$ represents the function space comprising all square-integrable functions with respect to the distribution of X^s . We note that the minimizer for $\mathbb{E} D_{\text{LS}}(r_0(X^s) \| f(X^s))$ with respect to $f \in \mathcal{L}^2(X^s)$ is not unique. Nonetheless, every two minimizers are equivalent X^s -almost surely.

At the empirical level, given the observations $\{X_1^s, \dots, X_n^s\}$ and $\{X_1^t, \dots, X_n^t\}$, the least squares estimator, denoted as \hat{r}_{LS} , is constructed by

$$\hat{r}_{\text{LS}} = \underset{f \in \mathcal{F}_{\text{NN}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n f(X_i^s)^2 - \frac{2}{n} \sum_{i=1}^n f(X_i^t).$$

To evaluate the estimation error of \hat{r}_{LS} , we concentrate on the expected excess risk, defined as

$$\mathcal{R}^j(\hat{r}_{\text{LS}}) = \mathbb{E} [r_0(X^j) - \hat{r}_{\text{LS}}(X^j)]^2,$$

for $j = s, t$.

Generally, $\mathcal{R}^j(\hat{r}_{\text{LS}})$ can be decomposed into two components, namely the stochastic error term and the approximation error term [19]. Both of these components are intricately related to the richness of the hypothesis space \mathcal{F}_{NN} . Recall that every function in \mathcal{F}_{NN} is bounded from above by $\bar{\delta}$ and from below by $\underline{\delta}$. In this scenario, it is appropriate to set $\underline{\delta}$ to 0. However, the selection of $\bar{\delta}$ presents a more complex challenge. Since the true function r_0 may not be bounded, it is essential for $\bar{\delta}$ to increase with the sample size n . Consequently, the divergence pattern of $\bar{\delta} = \bar{\delta}_n$ becomes critical in ensuring that the hypothesis space adequately captures the complexity of r_0 as more data are available. On one hand, $\bar{\delta}_n$ should not be excessively large, as this may compromise control over stochastic error. On the other hand, an overly small value of $\bar{\delta}_n$ could hinder the ability of \mathcal{F}_{NN} to densely approximate the true function r_0 . Here, we present Theorem 3.1, which achieves a nuanced balance.

THEOREM 3.1. *Assume that*

- (i) $r_0(x) \in \mathcal{H}_{\text{Loc}}^{\beta_r}(\mathbb{R}^d, B_u)$ with $\beta_r > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0, m \geq 0$;
- (ii) $r_0(X^s)$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Then, given the hyper-parameters L set to $\mathcal{O}(n^{d/(2d+4\beta_r)} \log n)$, M set to $\mathcal{O}(1)$, $\underline{\delta}$ set to 0 and $\bar{\delta} = \bar{\delta}_n$ set to $(\log n)^{1+\kappa}$ with an arbitrarily fixed $\kappa \in (0, 1]$, for $n \geq 2$, we have

$$\begin{aligned}\mathcal{R}^s(\hat{r}_{\text{LS}}) &\leq c^* n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(8+4\kappa)\vee(2m)}, \\ \mathcal{R}^t(\hat{r}_{\text{LS}}) &\leq c^{**} n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(8+4\kappa)\vee(2m)+1},\end{aligned}$$

where c^ and c^{**} are constants not depending on n .*

In Theorem 3.1, we make two critical assumptions on r_0 . Firstly, we require that the function $r_0(x)$ can be bounded by a universal polynomial function within any compact region. This assumption encompasses not only the trivial case where $r_0(x)$ is uniformly bounded, but also permits the density ratio to diverge across \mathbb{R}^d at a polynomial rate. The divergence rate m plays a secondary role in determining the final upper bound of $\mathcal{R}^j(\hat{r}_{\text{LS}})$ for $j = s, t$. While a more rapid rate yields a larger upper bound, the rate m influences only the logarithmic term, thereby ensuring that the overall upper bound remains controllable and (nearly) optimal [54, 60].

Secondly, we suppose that $r_0(X^s)$ follows a sub-exponential distribution. This condition serves two purposes. On one hand, it ensures that $r_0(X^s)$ is square-integrable and is less stringent than conditions imposed in other works that require boundedness on $r_0(X^s)$. Additionally, our finding enhances the result presented in [12], where a (nearly) minimax optimal rate was attained under the condition that the square of the density ratio was sub-exponential. On the other hand, this condition allows us to set $\bar{\delta}_n$ to $(\log n)^{1+\kappa}$ for $\kappa \in (0, 1]$, as we only need to consider values no greater than $c \cdot \log n$ for a sub-exponentially distributed random variable, where c is a constant. Meanwhile, we assume that $\|X^s\|_\infty$ is also sub-exponentially distributed for similar reasons, facilitating the application of approximation theorems derived on compact sets [48, 51]. Note that the condition of sub-exponential distribution for $r_0(X^s)$ could potentially be relaxed if the divergence rate m is known. The proof of Theorem 3.1 is presented in the Supplementary Materials.

3.2. Estimation based on logistic regression loss. In this subsection, we consider the logistic regression-based loss function with $\varphi_{\text{LR}}(x) = x \log x - (x+1) \log(x+1)$. The behavior of the range of $r_0(x)$ is crucial in this scenario. Recall that $\mathcal{X}^s = \{x : p(x) > 0\}$ denotes

the domain of X^s and $\mathcal{X}^t = \{x : q(x) > 0\}$ represents the domain of X^t . It is clear that r_0 equals to 0 over $\mathcal{X}^s \setminus \mathcal{X}^t$. While it is reasonable to assign $\varphi_{\text{LR}}(0) = 0$ so that the divergence $D_{\text{LR}}(r_0(x) \| f(x))$ is well-defined on $\mathcal{X}^s \setminus \mathcal{X}^t$, we note that the estimation efficiency may be compromised when $P(X^s \in \mathcal{X}^s \setminus \mathcal{X}^t) > 0$. This inefficiency arises from an overwhelming approximation error within the region $\mathcal{X}^s \setminus \mathcal{X}^t$. To achieve a fast rate, we therefore assume that $P(X^s \in \mathcal{X}^s \setminus \mathcal{X}^t) = 0$. It is important to note that this condition was not imposed in the previous analysis, where we set $\underline{\delta}$ to 0. Contrastively, $\underline{\delta} = \underline{\delta}_n$ will vanish slowly in the sequel to align with the regularity of function $\varphi_{\text{LR}}(x)$.

Let $\mathcal{L}_{\text{LR}}(X^s) = \{h : \mathbb{R}^d \rightarrow \mathbb{R}_+ \cup \{0\}, \mathbb{E}[h(X^s)^2] < \infty, \mathbb{E}[h(X^s)^{-2} \mathbf{1}(X^s \in \mathcal{X}^t)] < \infty \text{ and } h(x) > 0 \text{ for any } x \in \mathcal{X}^t\}$. At the population level, we have Lemma 3.2 which justifies the validity of the logistic regression-based loss; see the Supplementary Materials for its proof. We note that in Lemma 3.2, the function space $\mathcal{L}_{\text{LR}}(X^s)$ can be relaxed to $\tilde{\mathcal{L}} = \{h : \mathbb{R}^d \rightarrow \mathbb{R}_+ \cup \{0\}, h \text{ is measurable and } h(x) > 0 \text{ for any } x \in \mathcal{X}^t\}$.

LEMMA 3.2. *Suppose that $r_0 \in \mathcal{L}_{\text{LR}}(X^s)$. Then, it holds that $r_0(X^s) = f^*(X^s)$ a.s. X^s , where f^* is a minimizer of $\mathbb{E}D_{\text{LR}}(r_0(X^s) \| f(X^s))$ with respect to $f \in \mathcal{L}_{\text{LR}}(X^s)$.*

As a consequence, by assuming $r_0 \in \mathcal{L}_{\text{LR}}(X^s)$, we have

$$\begin{aligned} r_0 &= \underset{f \in \mathcal{L}_{\text{LR}}(X^s)}{\operatorname{argmin}} \mathbb{E}D_{\text{LR}}(r_0(X^s) \| f(X^s)) \\ &= \underset{f \in \mathcal{L}_{\text{LR}}(X^s)}{\operatorname{argmin}} \mathbb{E}[\log(f(X^s) + 1) - r_0(X^s) \log f(X^s) + r_0(X^s) \log(f(X^s) + 1)] \\ &= \underset{f \in \mathcal{L}_{\text{LR}}(X^s)}{\operatorname{argmin}} \mathbb{E}[\log(f(X^s) + 1) - \log f(X^t) + \log(f(X^t) + 1)]. \end{aligned}$$

Thus, at the sample level, the estimator is given by

$$\hat{r}_{\text{LR}} = \underset{f \in \mathcal{F}_{\text{NN}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \log(f(X_i^s) + 1) + \frac{1}{n} \sum_{i=1}^n [-\log f(X_i^t) + \log(f(X_i^t) + 1)].$$

The loss function above is similar to the logistic loss. Here, the lower bounded $\underline{\delta}$ should not be set to 0, which otherwise may result in the loss being undefined. The determination of both $\bar{\delta}$ and $\underline{\delta}$ relies on the continuity of φ_{LR} . Specifically, let x and y be two positive scalars lying in the interval $[a, b]$ with $a > 0$. Then, it can be verified that

$$(3) \quad \frac{1}{2b(b+1)}(x-y)^2 \leq D_{\text{LR}}(x \| y) \leq \frac{1}{2a(a+1)}(x-y)^2.$$

The first inequality in Eqn. (3) illustrates the relationship between the estimation error and Bregman divergence, indicating that b (associated with $\bar{\delta}$) is supposed to be small. Meanwhile, the second inequality in Eqn. (3) connects Bregman divergence with the approximation error, suggesting that a^{-1} (corresponding to $\underline{\delta}^{-1}$) should also remain small. Through a careful balancing of these parameters, we obtain the error bounds of \hat{r}_{LR} in Theorem 3.3.

THEOREM 3.3. *Assume that*

- (i) $r_0(x) \in \mathcal{H}_{\text{Loc}}^{\beta_r}(\mathbb{R}^d, B_u)$ with $\beta_r > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0$, $m \geq 0$;
- (ii) $r_0(X^s)$, $r_0(X^s)^{-1} \mathbf{1}(X^s \in \mathcal{X}^t)$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Then, given the hyper-parameters L set to $\mathcal{O}(n^{d/(2d+4\beta_r)} \log n)$, M set to $\mathcal{O}(1)$, $\underline{\delta} = \underline{\delta}_n$ set to $(\log n)^{-1-\kappa}$ and $\bar{\delta} = \bar{\delta}_n$ set to $(\log n)^{1+\kappa}$ with an arbitrarily fixed $\kappa \in (0, 1]$, for $n \geq 3$, we have

$$\begin{aligned}\mathcal{R}^s(\hat{r}_{\text{LR}}) &\leq c^* n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(11+7\kappa)\vee(2m+3+3\kappa)}, \\ \mathcal{R}^t(\hat{r}_{\text{LR}}) &\leq c^{**} n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(11+7\kappa)\vee(2m+3+3\kappa)+1},\end{aligned}$$

where c^* and c^{**} are constants not depending on n .

The assumptions in Theorem 3.3 closely parallel those outlined in Theorem 3.1, while we introduce an additional requirement concerning the sub-exponential restriction of $r_0(X^s)^{-1} \mathbf{1}(X^s \in \mathcal{X}^t)$, which is well-defined under the convention $0/0 = 0$. This condition constrains the rate at which $r_0(X^s)$ approaches 0, which ensures that the region $r_0(X^s) < c(\log n)^{-1}$ remains sufficiently small for some constant c . In addition, we note that these assumptions may need modification depending on the formulation of the estimator. For instance, in cases where the estimator is expressed as $\exp(f)$ for $f \in \mathcal{F}_{\text{NN}}$, attention should be given to the properties of $\log r_0$.

The upper bounds established in Theorem 3.3 have a slower convergence rate compared to those derived in Theorem 3.1. This discrepancy can be attributed to two factors. Firstly, in this subsection, we work on the surrogate loss deduced by φ_{LR} , which incurs a certain cost when converting the Bregman divergence into squared loss. Secondly, the logarithm terms enhances sharpness of φ_{LR} with respect to its derivative, leading to an increased stochastic error. The proof of Theorem 3.3 is given in the Supplementary Materials.

4. Error control under covariate shift. In order to tackle covariate shift, a typical strategy is the density ratio correction [53, 57, 70]. Specifically, let us consider the population-level loss function associated with the parameter of interest, θ , in the target domain, denoted as $\mathbb{E}\ell(X^t, Y^t, \theta)$. It is direct to observe that $\mathbb{E}\ell(X^t, Y^t, \theta) = \mathbb{E}[r_0(X^s)\ell(X^s, Y^s, \theta)]$. This relationship indicates that the original loss $\ell(x, y, \theta)$ can be adjusted to a corrected loss $r_0(x)\ell(x, y, \theta)$, which, when integrated with source data, essentially serves as the desired loss function for the target domain.

In situations where *the density ratio is known* and has a finite second moment, [39] showed that an estimator derived from a corrected loss function can achieve a minimax optimal rate. However, in practice, the true density ratio is typically unknown, necessitating the use of a density ratio estimator for correction. Consequently, the estimation error of θ is also influenced by the estimation error of the density ratio, especially in transfer learning problems where source data is often much more accessible than target data [32, 66, 67]. For example, [45] showed that density ratio correction can be fragile, exhibiting high variance and sensitivity to the methods used for density ratio estimation, even in cases of seemingly minor shifts [9]. Therefore, it is crucial to explore the conditions under which density ratio correction can be avoided.

The generalization capacity of a source estimator in the target domain was examined by [49] in their Lemma 12. Although they provided a clear upper bound, deriving explicit conditions to ensure that the convergence rate remains optimal is still a nontrivial task. Moreover, their hypothesis space was limited to the class of all 1-Lipschitz functions, and their analysis was focused on the nonparametric regression setting. Based on the analyses in Section 3, we show that the sub-exponential property ensures effective generalization between source and target domains.

In this subsection, we outline some regularity conditions that facilitate the control for the performance of $\hat{\theta}_N(X)$ in the target domain based on its performance in the source domain, allowing for the presence of slowly divergent factors. We first present the following generic lemma; see the Supplementary Materials for its proof.

LEMMA 4.1. Suppose that U_1, \dots, U_n, U are d -dimensional random vectors, with $\|U_i\|_\infty \leq \xi_i$ almost surely for $i = 1, \dots, n$. Assume that $\|U\|_\infty$ has a finite fourth moment, and V is a random variable such that $\mathbb{E} \exp(\varsigma |V|) < \infty$ for some positive constant ς . Let $\gamma_n = \mathbb{E} \|U_n - U\|_2^2$. Then, for $n \geq 2$, we have

$$\mathbb{E} (\|U_n - U\|_2^2 | V) \leq c_1 \gamma_n \log n + \frac{c_2 d(\xi_n^2 + 1)}{n},$$

where c_1 and c_2 are constants not depending on n .

REMARK 2. The existence of a positive scalar ς such that $\mathbb{E} \exp(\varsigma |V|) < \infty$ is satisfied when V is sub-exponentially distributed; see, e.g., [64].

Lemma 4.1 illustrates that when the tail of random variable V is not excessively heavy, the impact of multiplying by $|V|$ is minimal, resulting in only a sacrifice of $\log n$. By leveraging Lemma 4.1 within the framework of covariate shift, we derive Corollary 4.2. We note that the condition requiring $r_0(X^s)$ to be sub-exponentially distributed can be satisfied when $X^s \sim Ga(\alpha_1, \lambda)$ and $X^t \sim Ga(\alpha_2, \lambda)$ with $0 < \alpha_2 - \alpha_1 \leq 1$.

COROLLARY 4.2. Suppose that

- (i) $\|\hat{\theta}_N(X^s)\|_\infty \leq \xi_N$ almost surely for every $N \geq 1$;
- (ii) $\|\theta_0(X^s)\|_\infty$ has a finite fourth moment;
- (iii) $r_0(X^s)$ is sub-exponentially distributed.

Then, for $N \geq 2$, we have

$$\begin{aligned} \mathbb{E} \left\| \hat{\theta}_N(X^t) - \theta_0(X^t) \right\|_2^2 &= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \right] \\ &\leq c_1 \mathbb{E} \left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \log N + \frac{c_2 d_\theta (\xi_N^2 + 1)}{N}, \end{aligned}$$

where c_1, c_2 are constants not depending on N , and d_θ represents the dimensionality of $\theta_0(X)$.

While the condition that $r_0(X^s)$ follows a sub-exponential distribution is appealing, it is susceptible to violation under certain circumstances. For instance, consider the case where X^s and X^t are drawn from normal distributions, specifically $N(\mu_1, 1)$ and $N(\mu_2, 1)$ with $\mu_1 \neq \mu_2$. In this scenario, the density ratio can be expressed as $r_0(X^s) = \exp((\mu_2 - \mu_1)X^s - (\mu_2^2 - \mu_1^2)/2)$, which clearly indicates a significant departure from the characteristics of a sub-exponential distribution. To effectively tackle this challenge, it is crucial to analyze the divergence pattern of $r_0(x)$ alongside the tail property of X^s , as shown in the following two propositions. Their proofs can be found in Supplementary Materials.

PROPOSITION 4.3. Suppose that

- (i) $\|\hat{\theta}_N(X^s)\|_\infty \leq \xi_N$ almost surely for every $N \geq 1$;
- (ii) $\|\theta_0(X^s)\|_\infty$ has a finite eighth moment;
- (iii) there exists a dominant function $G(u) = c(u^m + 1)$ with constants $c > 0$ and $m \geq 0$ such that $r_0(x) \leq G(\|x\|_\infty)$;
- (iv) $r_0(X^s)$ has a finite second moment;
- (v) there exists a positive constant ς such that $\mathbb{E} \exp(\varsigma \|X^s\|_\infty) < \infty$.

Then, for $N \geq 2$, we have

$$\begin{aligned} & \mathbb{E} \left\| \hat{\theta}_N(X^t) - \theta_0(X^t) \right\|_2^2 \\ &= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \right] \\ &\leq c_1 (\log N)^m \mathbb{E} \left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 + \frac{c_2 d_\theta (\xi_N^2 + 1)}{N}, \end{aligned}$$

where c_1, c_2 are constants not depending on N , and d_θ represents the dimensionality of $\theta_0(X)$.

PROPOSITION 4.4. Suppose that

- (i) $\|\hat{\theta}_N(X^s)\|_\infty \leq \xi_N$ almost surely for every $N \geq 1$;
- (ii) $\|\theta_0(X^s)\|_\infty$ has a finite eighth moment;
- (iii) there exists a dominant function $G(u) = c \exp(mu)$ with constants $c > 0, m \geq 0$ such that $r_0(x) \leq G(\|x\|_\infty)$;
- (iv) $r_0(X^s)$ has a finite second moment;
- (v) there exists a positive constant ς such that $\mathbb{E} \exp(\varsigma \|X^s\|_\infty^2) < \infty$.

Then, for $N \geq 2$, we have

$$\begin{aligned} & \mathbb{E} \left\| \hat{\theta}_N(X^t) - \theta_0(X^t) \right\|_2^2 \\ &= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \right] \\ &\leq c_1 \exp \left\{ c_2 (\log N)^{1/2} \right\} \mathbb{E} \left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 + \frac{c_3 d_\theta (\xi_N^2 + 1)}{N}, \end{aligned}$$

where c_1, c_2, c_3 are constants not depending on N , and d_θ represents the dimensionality of $\theta_0(X)$.

REMARK 3. For any positive scalar ζ , we have $\exp\{(\log N)^{1/2}\} = o(N^\zeta)$. The existence of a positive scalar ς such that $\mathbb{E} \exp(\varsigma \|X^s\|_\infty^2) < \infty$ can be satisfied when $\|X^s\|_\infty$ is sub-Gaussian distributed; see, e.g., [64].

Propositions 4.3 – 4.4 extend the result presented in Corollary 4.2 to the scenarios where $r_0(X^s)$ may not be sub-exponentially distributed and instead possesses only a second finite moment. By assuming various divergence patterns of $r_0(x)$ as well as the tail properties of X^s , we derive distinct upper bounds for the expected excess risk in the target domain. Specifically, when $r_0(x)$ diverges according to a polynomial rate, the excess risk in target domain is shown to differ from that in source domain by a factor that is a polynomial function of logarithm order. Furthermore, if $r_0(x)$ diverges at a more rapid rate, we require that X^s exhibits greater concentration. In this case, the difference in excess risks between the source and target domains becomes more pronounced, which is larger than any polynomial function of logarithm order but smaller than any positive power of N .

When the density ratio is unbounded, [39] considered a reweighted RKHS least squares estimator using the truncated density ratio as weights. They demonstrated that this reweighted RKHS estimator is nearly optimal in the target domain under appropriate conditions. However, their estimator assumes that the density ratio is known, which is often not the case in practice, as the density ratio typically needs to be estimated. In contrast to [39], our analysis

indicates that it is possible to construct a nearly optimal estimator without relying on the density ratio. In such scenarios, an estimator based solely on source domain data can still generalize effectively to the target domain. For a class of parametric models, [16] showed that the classical maximum likelihood estimator, using only source data without any modifications, achieves minimax optimality for covariate shift if the parametric model is correctly specified. Their results also hold without requiring any boundedness condition on the density ratio.

5. Nonparametric regression and conditional flow models. In this section, we apply the results from Section 4 to two important scenarios: nonparametric regression and conditional distribution estimation using flow models under covariate shift.

5.1. Nonparametric regression. Over the past few decades, nonparametric regression has emerged as an active area of research in statistical learning, with extensive studies established based on methods such as splines [61], reproducing kernels [7] and neural networks [48]. In this subsection, we aim to elucidate the generalization capacity of the regression estimator derived from source domain when applied to target domain. Specifically, the regression task is framed in a general case, where both the covariate domain and true conditional mean function may be unbounded, and the response can be multi-dimensional.

Consider the following models

$$Y^j = f_0(X^j) + \varepsilon^j, \quad \text{for } j = s, t.$$

Here, f_0 represents the unknown regression function of interest, and ε^j denotes the noise term with $\mathbb{E}(\varepsilon^j | X^j) = 0$ and $\text{Var}(\varepsilon^j | X^j) = \Xi_0$ for some positive semi-definite matrix Ξ_0 and for $j = s, t$; additionally, ε^s and ε^t are identically distributed. Suppose that the covariate vector is d_x -dimensional and the response vector is d_y -dimensional. Our estimation paradigm concentrate on the source domain. Given source data $\{(X_1^s, Y_1^s), \dots, (X_N^s, Y_N^s)\}$, the sample estimator is given by

$$\hat{f}_N^s = \underset{f \in \mathcal{F}_{\text{NN}}^{d_y}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \|Y_i^s - f(X_i^s)\|_2^2.$$

We note that $\mathcal{F}_{\text{NN}}^{d_y}$ is a neural network function class such that $f : \mathbb{R}^{d_x} \rightarrow [\underline{\delta}, \bar{\delta}]^{d_y}$ for any $f \in \mathcal{F}_{\text{NN}}^{d_y}$.

Theorem 5.1 delineates the estimation error, founded on the expected excess risk, associated with the source estimator \hat{f}_N^s in both source and target domains. It is noticeable that, \hat{f}_N^s attains, within the source domain, a (nearly) standard minimax optimal convergence rate [54], while offering an remarkably similar rate in the target domain, with only a logarithmic factor as the compromise. The proof of Theorem 5.1 is present in Supplementary Materials.

THEOREM 5.1. *Assume that*

- (i) $e_j^\top f_0(x) \in \mathcal{H}_{\text{Loc}}^{\beta_f}(\mathbb{R}^d, B_u)$ with $\beta_f > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0$, $m \geq 0$, and for any $j \in \{1, \dots, d_y\}$, where e_j denotes a d_y -dimensional one-hot vector with the j -th component equal to 1 and all other components equal to 0;
- (ii) $\|Y^s\|_\infty$, $r_0(X^s)$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Then, given the hyper-parameters L set to $\mathcal{O}(N^{d_x/(2d_x+4\beta_f)} \log N)$, M set to $\mathcal{O}(1)$, $\bar{\delta} = \bar{\delta}_N$ set to $(\log N)^{1+\kappa}$ with an arbitrarily fixed $\kappa \in (0, 1]$, and $\underline{\delta} = \underline{\delta}_N$ set to $-(\log N)^{1+\kappa}$, for

$N \geq 2$, we have

$$\begin{aligned}\mathbb{E} \left\| \hat{f}_N^s(X^s) - f_0(X^s) \right\|_2^2 &\leq c^* N^{-\frac{2\beta_f}{d_x+2\beta_f}} (\log N)^{(8+4\kappa)\vee(2m)}, \\ \mathbb{E} \left\| \hat{f}_N^s(X^t) - f_0(X^t) \right\|_2^2 &\leq c^{**} N^{-\frac{2\beta_f}{d_x+2\beta_f}} (\log N)^{(8+4\kappa)\vee(2m)+1},\end{aligned}$$

where c^* and c^{**} are constants not depending on N .

We note that Theorem 5.1 is a direct application of Corollary 4.2, under the assumption that the density ratio is distributed sub-exponentially. Furthermore, when $r_0(X^s)$ exhibits different patterns, as demonstrated in Propositions 4.3 and 4.4, analogous results can be obtained with appropriate modifications.

In recent years, there has been considerable effort dedicated to the error analysis of non-parametric regression using deep neural network models [4, 11, 25, 41, 48]. These studies typically rely on the crucial assumption that the regression function belongs to a uniformly bounded Hölder class defined on a bounded domain. This assumption simplifies the analysis by ensuring that the function's behavior is well-controlled across its entire domain. In contrast, our results, as presented in Theorem 5.1, relax this assumption by only requiring that the regression function belongs to a local Hölder class. This allows for unbounded domains and ranges, which is more realistic for scenarios where data may not be neatly confined within bounded limits. Handling unbounded functions requires a careful analysis of the tail behavior of the relevant distributions and functions, making the analysis more technically challenging. Our results significantly enhance the understanding of deep neural networks' performance in nonparametric regression tasks.

5.2. Conditional flow models. In this subsection, we focus on the task of learning a conditional distribution using generative flow models, as described in Section 2.5. We consider a specific stochastic interpolant,

$$(4) \quad Y_\tau^s = a_\tau \eta + b_\tau Y^s,$$

where a_τ and b_τ are continuously differentiable with respect to $\tau \in [0, 1]$, satisfying the boundary conditions $a_0 = b_1 = 1$ and $a_1 = b_0 = 0$, and η denotes the d_y -dimensional standard Gaussian random vector.

We define the deduced velocity field as

$$v_0(x, y, \tau) = \mathbb{E} \left(\dot{a}_\tau \eta + \dot{b}_\tau Y^s \mid Y_\tau^s = y, X^s = x \right).$$

Let $\mathcal{L}^2(X^s, Y^s) = \{f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [0, 1], \mathbb{E} \|f(X^s, Y_\tau^s, \tau)\|_2^2 < \infty \text{ for any } \tau \in [0, 1]\}$. Clearly, at the population level, it holds that

$$v_0 = \underset{f \in \mathcal{L}^2(X^s, Y^s)}{\operatorname{argmin}} \int_0^1 \mathbb{E} \left\| \dot{a}_\tau \eta + \dot{b}_\tau Y^s - f(X^s, Y_\tau^s, \tau) \right\|_2^2 d\tau,$$

provided that $\|Y^s\|_2$ has a finite second moment. At the empirical level, given source observations $\{(X_1^s, Y_1^s), \dots, (X_N^s, Y_N^s)\}$, we independently sample N random vectors $\{\eta_1, \dots, \eta_N\}$ from the Gaussian distribution $N(0, I_{d_y})$, and N random values $\{\tau_1, \dots, \tau_N\}$ from the uniform distribution $U(0, 1)$. Then, the empirical estimator of v_0 is constructed by

$$\hat{v}_N^s = \underset{f \in \mathcal{F}_{NN}^{d_y}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^N \left\| \dot{a}_{\tau_i} \eta_i + \dot{b}_{\tau_i} Y_i^s - f(X_i^s, Y_{i,\tau_i}^s, \tau_i) \right\|_2^2,$$

where $Y_{i,\tau_i}^s = a_{\tau_i} \eta_i + b_{\tau_i} Y_i^s$.

With the estimate \hat{v}_N^s and for any fixed $x \in \mathcal{X}^s$, an ODE, with respect to $\tau \in [0, 1]$, is established as follows,

$$d\hat{Z}_\tau = \hat{v}_N^s(x, \hat{Z}_\tau, \tau) d\tau, \quad \hat{Z}_0 \sim N(0, I_{d_y}).$$

Intuitively, the distribution of \hat{Z}_1 is an approximation of the conditional distributions of $Y^s|X^s = x$ and $Y^t|X^t = x$. To quantify this approximation more concretely, we employ the 2-Wasserstein distance as a criterion for measuring the discrepancy of two distributions (see, e.g., [26, 50] for its definition). We use $W_2^2(\rho_1 \| \rho_2)$ to denote the squared 2-Wasserstein distance for two probability density functions ρ_1 and ρ_2 . Supposing that $Y^s|X^s = x$ admits a conditional density function denoted as $\rho_{0,x}$, we denote the density function of \hat{Z}_1 , when given x , as $\hat{\rho}_x^s$. Then, the estimation errors for source and target domains are defined respectively as

$$\begin{aligned} \mathcal{E}^s &= \mathbb{E} [W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s)] = \int \mathbb{E} [W_2^2(\rho_{0,x} \| \hat{\rho}_x^s)] p(x) dx, \\ \mathcal{E}^t &= \mathbb{E} [W_2^2(\rho_{0,X^t} \| \hat{\rho}_{X^t}^s)] = \int \mathbb{E} [W_2^2(\rho_{0,x} \| \hat{\rho}_x^s)] q(x) dx. \end{aligned}$$

While the relationship between \mathcal{E}^s and \mathcal{E}^t is not immediately evident, Lemma 5.2 reveals that they exhibits similar property compared to the squared loss demonstrated in Corollary 4.2. The proof of Lemma 5.2 is present in the Supplementary Materials.

LEMMA 5.2. *Assume that*

- (i) *the solution of ODE (2), with standard Gaussian initialization, is unique such that Z_1 given x follows the distribution of $Y^s|X^s = x$ for all $x \in \mathcal{X}^s$;*
- (ii) *$\|Y^s\|_2$ has a finite fourth moment;*
- (iii) *$r_0(X^s)$ is sub-exponentially distributed.*

Then, for $N \geq 2$, we have

$$\mathcal{E}^t \leq c_1 \mathcal{E}^s \log N + \frac{c_2 d_y [\max(\bar{\delta}^2, \underline{\delta}^2) + 1]}{N},$$

where c_1 and c_2 are constants not depending on N .

REMARK 4. Assumption (i) in Lemma 5.2 can be satisfied by some regularity conditions on the probability structure of (X^s, Y^s) and on the continuity of v_0 ; see, e.g., [6, 13, 23].

REMARK 5. The result present in Lemma 5.2 is not restricted to the specific interpolant (4).

We now proceed to present explicit upper bounds for both \mathcal{E}^s and \mathcal{E}^t . Essentially, the estimation error of conditional density function heavily relies on the estimation error of the velocity field. As in previous analyses, it is pivotal to specify the continuity pattern of the underlying function, which is v_0 in this case. Here, we adopt a Sobolev-type function class, which facilitates the application of Grönwall's inequality. The classical Sobolev space is defined as follows.

DEFINITION 5.3 (Sobolev space). Let $\beta \in \mathbb{N}$. The Sobolev space $\mathcal{W}^{\beta,\infty}(\Omega)$ is defined by

$$\mathcal{W}^{\beta,\infty}(\Omega) = \left\{ f : \Omega \rightarrow \mathbb{R}, \|f\|_\infty < \infty, \|D^\alpha f\|_\infty < \infty \text{ for all } \alpha \in \mathbb{N}^d \text{ with } \|\alpha\|_1 \leq \beta \right\}.$$

Furthermore, for any $f \in \mathcal{W}^{\beta,\infty}(\Omega)$, we define the Sobolev norm $\|\cdot\|_{\mathcal{W}^{\beta,\infty}(\Omega)}$ as

$$\|f\|_{\mathcal{W}^{\beta,\infty}(\Omega)} = \max_{0 \leq \|\alpha\|_1 \leq \beta} \|D^\alpha f\|_\infty.$$

Then, we introduce the local and time-space version of $\mathcal{W}^{\beta,\infty}(\Omega)$, denoted as $\mathcal{W}_{\text{Gen}}^{\beta,\infty}(\mathbb{R}^d, B_u)$, which is defined by

$$\begin{aligned} \mathcal{W}_{\text{Gen}}^{\beta,\infty}(\mathbb{R}^d, B_u) = & \left\{ f : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}, \right. \\ & g(x, \tau) = f|_{[-u, u]^d \times [0, 1]}(2ux - u1_d, \tau) \in \mathcal{W}^{\beta,\infty}([0, 1]^{d+1}) \\ & \left. \text{with } \|g\|_{\mathcal{W}^{\beta,\infty}([0, 1]^{d+1})} \leq B_u \text{ for any } u > 0 \right\}. \end{aligned}$$

It is worthy noting that this function class is chosen for simplicity. Recent studies indicate that a general velocity field may exhibit singular behavior at $\tau = 1$ [15, 24]; however, this aspect falls outside the scope of this paper and deserves a more thorough investigation.

To be compatible with the Lipschitz continuity inherent in Sobolev space, the modified neural network class $\mathcal{F}_{\text{NN}, \Lambda}^d = \mathcal{F}_{\text{NN}}^d \cap \mathcal{F}_{\text{Lip}, \Lambda}^d$ is employed, where

$$\mathcal{F}_{\text{Lip}, \Lambda}^d = \{f : \Omega \rightarrow \mathbb{R}^d, \|f(z_1) - f(z_2)\|_2 \leq \Lambda \|z_1 - z_2\|_2, \text{ for any } z_1, z_2 \in \Omega\}.$$

Furthermore, the depth and width of $\mathcal{F}_{\text{NN}, \Lambda}^d$ correspond to the depth and width of $\mathcal{F}_{\text{NN}}^d$. Theorem 5.4 establishes the sample convergence of the conditional density estimation error; see the Supplementary Materials for its proof.

THEOREM 5.4. *Assume that*

- (i) *the solution of ODE (2), with standard Gaussian initialization, is unique such that Z_τ given x follows the distribution of $Y_\tau^s | X^s = x$ for all $x \in \mathcal{X}^s$ and $\tau \in [0, 1]$;*
- (ii) *$e_j^\top v_0 \in \mathcal{W}_{\text{Gen}}^{1,\infty}(\mathbb{R}^{d_x+d_y}, B_u)$ with $B_u \leq c(u^m + 1)$ for some universal constants $c > 0$, $m \in [0, 1]$, and for any $j \in \{1, \dots, d_y\}$, where e_j denotes a d_y -dimensional one-hot vector with the j -th component equal to 1 and all other components equal to 0;*
- (iii) *$\|Y^s\|_\infty$ and $\|X^s\|_\infty$ are sub-Gaussian random variables;*
- (iv) *$r_0(X^s)$ is sub-exponentially distributed.*

Then, given the hyper-parameters L set to $\mathcal{O}(N^{(d_x+d_y+1)/[2(d_x+d_y+1)+4]} \log N)$, M set to $\mathcal{O}(1)$, $\bar{\delta} = \bar{\delta}_N$ set to $(\log N)^{(1+\kappa)/2}$ with an arbitrarily fixed $\kappa \in (0, 1)$, $\underline{\delta} = \underline{\delta}_N$ set to $-(\log N)^{(1+\kappa)/2}$ and $\Lambda = \Lambda_N$ set to $(\log N)^{(1+\kappa)/2}$, for $N \geq 2$, we have

$$\begin{aligned} \mathcal{E}^s &\leq c^* N^{-\frac{2}{d_x+d_y+3}} (\log N)^{6+2\kappa} \exp\left(1 + 2(\log N)^{(1+\kappa)/2}\right), \\ \mathcal{E}^t &\leq c^{**} N^{-\frac{2}{d_x+d_y+3}} (\log N)^{7+2\kappa} \exp\left(1 + 2(\log N)^{(1+\kappa)/2}\right) \end{aligned}$$

where c^ and c^{**} are constants not depending on N .*

The unique-solution assumption in Theorem 5.4 is an enhancement of assumption (i) in Lemma 5.2. This refinement stems from the examination of the whole dynamics concerning the induced ODEs, whose validity is also contingent on the probability structure of (X^s, Y^s) . Furthermore, we assume the divergent rate of each component in v_0 does not exceed a linear rate. Therefore, the Lipschitz constant $(\log N)^{(1+\kappa)/2}$ with $\kappa \in (0, 1)$ is sufficient for performing the approximation, while simultaneously ensuring the convergence of \mathcal{E}^s . Notably, $\|Y^s\|_\infty$ and $\|X^s\|_\infty$ are assumed to exhibit sub-Gaussian behavior, as tighter concentration

TABLE 1
Averages and standard deviations (shown in the brackets) of mean squared errors between the true density ratio values and predicted values, based on results from 100 replications.

Domain	Sample size	$d = 1$	$d = 2$	$d = 5$
Source	200	0.099 (0.206)	0.508 (0.447)	4.032 (1.753)
	500	0.049 (0.065)	0.302 (0.226)	3.979 (1.706)
	1000	0.022 (0.023)	0.211 (0.178)	2.434 (1.044)
	1500	0.021 (0.051)	0.149 (0.091)	1.801 (1.033)
	2000	0.013 (0.023)	0.130 (0.063)	1.263 (0.560)
	3000	0.013 (0.019)	0.117 (0.107)	1.351 (0.852)
Target	200	0.390 (0.789)	2.337 (1.343)	64.130 (26.729)
	500	0.226 (0.316)	1.574 (0.906)	57.939 (25.431)
	1000	0.095 (0.112)	1.240 (0.896)	49.817 (25.358)
	1500	0.089 (0.261)	1.116 (0.699)	43.566 (19.931)
	2000	0.053 (0.098)	1.017 (0.478)	39.076 (18.381)
	3000	0.051 (0.082)	0.926 (0.739)	37.078 (19.932)

is essential for the approximation process. We emphasize that Theorem 5.4 aims to offer a concrete instance of the controllability for \mathcal{E}^t given the convergence of \mathcal{E}^s . The convergence rates of both \mathcal{E}^s and \mathcal{E}^t can potentially be improved through a more nuanced investigation into the continuity properties of v_0 .

6. Simulation studies. To practically justify the theoretical findings, we here present some empirical results from simulation experiments, demonstrating the consistency of our density ratio estimators and the risk controllability under covariate shift. In particular, we concentrate on the scenarios where the source and target covariates follow gamma distributions, thereby fulfilling or surpassing the sub-exponential assumptions. For clarity of notation, given a d -dimensional vector x , its j -th entry is denoted as $x_{(j)}$ for $j = 1, \dots, d$.

6.1. Performance of density ratio estimators. Regarding the source covariate X^s and the target covariate X^t , let $X_{(j)}^s$ independently follows $Ga(j, 2)$ and let $X_{(j)}^t$ independently follows $Ga(j + 1, 2)$, for $j = 1, \dots, d$. It is straightforward to verify that the true density ratio function can be expressed as $r_0(x) = 2^d(d!)^{-1} \prod_{j=1}^d x_{(j)}$. Hence, $r_0(x) \in \mathcal{H}_{Loc}^{\beta_r}(\mathbb{R}^d, 2^{2d}(d!)^{-1}u^d)$ for any $\beta_r \geq d$, and $\|X^s\|_\infty$ is sub-exponentially distributed. Notably, $r_0(X^s)$ is a sub-exponential random variable when $d = 1$ while exhibiting a heavier tail for larger $d \geq 2$.

We adopted the least squares loss for estimation. The neural network was designed with $\lfloor (\log n)/2 \rfloor$ hidden layers, where n represented the sample size. Each hidden layer contained 64 neurons. We simply set $\kappa = 0.5$. The training algorithm was implemented using Pytorch framework [44] along with the Adam optimizer [28]. We specified a learning rate at 1e-4 and a batch size of 100. The sample size was varied among $\{200, 500, 1000, 1500, 2000, 3000\}$; the dimension d was set to 1, 2 and 5, with the number of training iterations being 1000, 2000 and 5000, respectively. After obtaining an estimator, we evaluated its performance by calculating the mean squared loss within both source and target domains based on 1000 testing samples per domain. To enhance robustness and reliability, we conducted 100 replications for each (n, d) combination.

Table 1 and Figure 1 illustrate the performance of density ratio estimators across source and target domains. A consistent inverse relationship between sample size and mean squared errors, with the latter asymptotically approaching zero as the former increased, was noticeable. In the univariate cases ($d = 1$), the estimation errors in the target domain exhibited a magnitude approximately four times those of the source domain. As the dimensionality

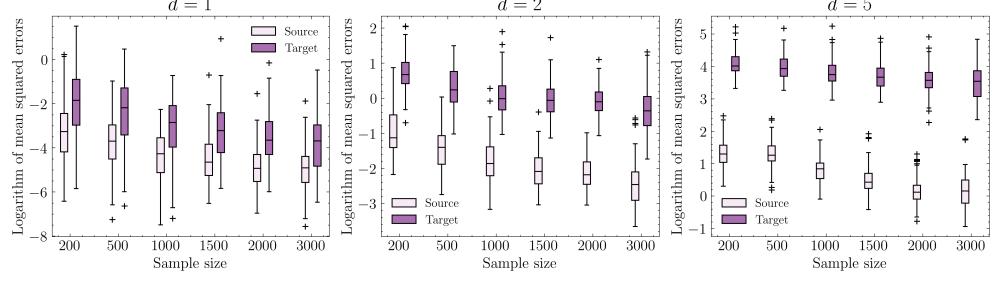


FIG 1. Boxplots of logarithm of mean square errors between the true density ratio values and predicted values, based on results from 100 replications.

increased, the disparity in estimation errors between the two domains became more pronounced.

6.2. Sufficiency of source estimators for covariate shift. In this subsection, we consider the regression model $Y = f_0(X) + \nu\varepsilon$ where $\nu > 0$ and

$$f_0(X) = \begin{pmatrix} f_{01}(X) \\ f_{02}(X) \end{pmatrix} = \begin{pmatrix} \sin(\pi(X_{(1)} - X_{(2)})) \log(1 + X_{(3)}^2) \\ \exp(-X_{(2)}) \mathbb{1}(X_{(4)} > 2) \end{pmatrix},$$

$$\varepsilon = \begin{pmatrix} W \\ -W \end{pmatrix}, \quad W \sim N(0, 1).$$

The covariate X was drawn from either the source domain (X^s) or the target domain (X^t), as defined in Subsection 6.1 with the dimensionality $d = 5$. The parameter ν was assigned values of 0.1, 0.2 and 1, corresponding to low, moderate, and high noise levels, respectively. Such levels were calibrated relative to the variances of $f_{01}(X^s)$ and $f_{02}(X^s)$, with the moderate noise level (that is, 0.2²) approximating the variance of $f_{02}(X^s)$ and the high noise level approximating that of $f_{01}(X^s)$. This setup generated both source and target datasets. Specifically, for training, we sampled n_{11} observations of covariates and responses from the source domain, and $n_{12} = 500$ observations containing only covariates from the target domain. Here, n_{11} varied among {500, 1000, 1500, 2000, 2500, 3000}. Additionally, we generated $n_2 = 1000$ testing data in the form $(X, f_0(X))$ for each domain.

By minimizing the least squares loss and utilizing merely source data, we obtained the source estimator. The neural network architecture remained identical to that described in Subsection 6.1, while maintaining κ to 0.5. For the regression task, we employed a learning rate of 1e-3, and carefully selected the number of iterations through cross validation among the candidate list {1000, 2000, 3000, 4000, 5000}. Subsequently, as benchmarks, we performed two types of loss correction methods, namely the estimated density ratio correction (EDRC) and the oracle density ratio correction (ODRC). For EDRC, we first conducted the density ratio estimation using n_{11} source covariates and n_{12} target covariates. Then, a corrected least squares loss based on this estimated ratio was applied to construct an estimator for f_0 (see Section 4). For ODRC, we corrected the least squares loss using the oracle density ratio. For both correction methods, we maintained the same neural network architecture and determined the optimal number of training iterations for estimating f_0 through cross validation.

We assessed the performance through the mean squared error. To be more specific, we recorded the mean squared errors between true and predicted conditional mean values for the source estimator in both source and target domains using testing data. In addition, we calculated mean squared errors in the target domain with respect to estimators derived from EDRC and ODRC methods. To summarize, we obtained four risk measures, namely the

TABLE 2

The averages and standard deviations (indicated in brackets) of the mean squared errors between the true conditional mean values and the predicted values, based on results from 100 replications. SERS denotes the source estimator risk in the source domain, while SERT represents the source estimator risk in the target domain. EDRC stands for the estimated density ratio correction, and ODRC refers to the oracle density ratio correction.

Noise level	Sample size (n_{11})	SERS	SERT	EDRC	ODRC
$\nu = 0.1$	500	0.201 (0.061)	0.623 (0.162)	2.080 (0.375)	0.819 (0.184)
	1000	0.110 (0.034)	0.391 (0.092)	1.157 (0.338)	0.530 (0.119)
	1500	0.071 (0.024)	0.265 (0.068)	0.829 (0.280)	0.394 (0.094)
	2000	0.059 (0.021)	0.220 (0.053)	0.572 (0.151)	0.322 (0.084)
	2500	0.054 (0.022)	0.193 (0.040)	0.485 (0.121)	0.283 (0.063)
	3000	0.032 (0.013)	0.119 (0.042)	0.332 (0.106)	0.193 (0.056)
$\nu = 0.2$	500	0.246 (0.060)	0.698 (0.164)	2.142 (0.384)	0.926 (0.186)
	1000	0.133 (0.035)	0.439 (0.099)	1.250 (0.355)	0.651 (0.130)
	1500	0.092 (0.025)	0.308 (0.069)	0.859 (0.280)	0.432 (0.089)
	2000	0.074 (0.021)	0.256 (0.056)	0.611 (0.150)	0.355 (0.083)
	2500	0.063 (0.021)	0.218 (0.045)	0.539 (0.132)	0.323 (0.067)
	3000	0.045 (0.014)	0.150 (0.037)	0.371 (0.104)	0.235 (0.061)
$\nu = 1.0$	500	0.959 (0.155)	2.000 (0.363)	3.691 (0.616)	2.435 (0.402)
	1000	0.454 (0.079)	1.036 (0.184)	2.586 (0.440)	1.710 (0.309)
	1500	0.321 (0.048)	0.773 (0.124)	2.039 (0.410)	1.394 (0.288)
	2000	0.264 (0.037)	0.658 (0.093)	1.725 (0.306)	1.100 (0.199)
	2500	0.243 (0.039)	0.606 (0.085)	1.551 (0.354)	0.979 (0.183)
	3000	0.210 (0.042)	0.480 (0.092)	1.344 (0.400)	0.851 (0.157)

source estimator risk in source domain, the source estimator risk in target domain, the EDRC estimator risk in target domain and the ODRC estimator risk in target domain. Furthermore, for each combination of (n_{11}, ν) , we ran 100 replications.

It is evident from Table 2 and Figure 2 that the source estimator's risk in the target domain decreased commensurately with its risk in the source domain, demonstrating the risk controllability of the source estimator in the target domain. Notably, the source estimator significantly outperformed the estimator based on EDRC method. It is particularly surprising and interesting that even with access to the true density ratio, the ODRC estimator showed its weakness compared to the source estimator, which lacked this additional information. This observation substantiates the fragility of density ratio correction [32] from the empirical perspective.

7. Conclusion. In this paper, we address the problem of density ratio estimation, allowing density ratios with unbounded domains and ranges. We develop a rigorous theoretical framework for density ratio estimators based on Bregman divergences, including least squares and logistic regression loss. Our findings contribute significantly to the existing literature on the estimation theory of density ratios.

To demonstrate the applications of our results in unbounded density ratio estimation, we study nonparametric regression and conditional flow models under covariate shift. We discover that the tail properties of the density ratio are crucial for ensuring risk transferability across different domains. Based on suitable tail conditions and divergent patterns of the density ratio function, we show that the source estimator is nearly optimal in the target domain. Our numerical results support these theoretical insights, particularly highlighting that the source estimator can outperform estimators derived from loss correction methods, even when the true density ratio is known.

We caution that, in general, the near-optimality of the source estimator cannot be guaranteed without explicitly accounting for covariate shift. For instance, in constrained RKHS-based nonparametric regression, [39] showed that there exists a specific pair of random elements $((X^s, Y^s), (X^t, Y^t))$ characterized by a particular probability structure, such that a

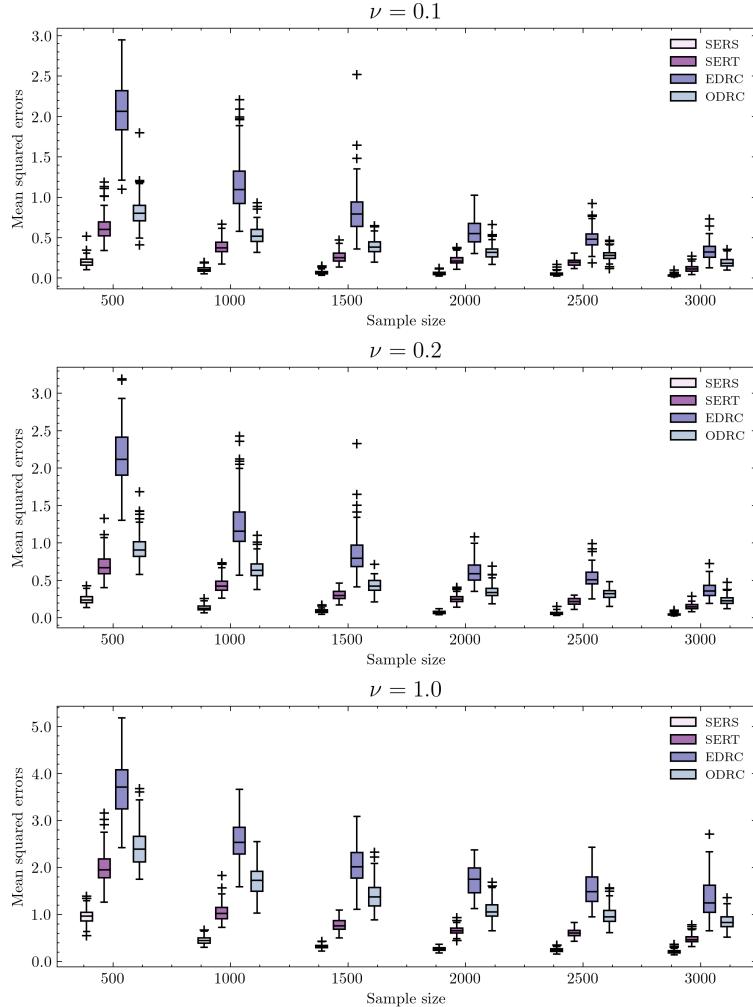


FIG 2. Boxplots of mean square errors between the true conditional mean values and predicted values, based on results from 100 replications. SERS means the source estimator risk in source domain and SERT represents the source estimator risk in target domain.

kernel regression estimator for the conditional mean has a slower convergence rate in the target domain compared to the source domain. However, if the density ratio is unknown, constructing an optimal estimator in this constrained kernel regression setting requires further investigation.

Several other directions merit exploration. Beyond the covariate shift problem addressed in this work, our density estimation results have potential applications in areas where density ratios are crucial, such as transfer learning, optimal transport methods for generative learning [14], mutual information estimation, and propensity score estimation [31]. Moreover, the techniques developed in this work for handling density ratios with unbounded domains and ranges could be adapted and extended to other settings where unbounded functions arise, such as score-based generative models.

Appendix

In the Appendix, we provide proofs of the results presented in the paper, along with additional technical details.

APPENDIX A: AUXILIARY LEMMAS

A.1. Regularity of Bregman divergence.

LEMMA A.1. *Let $\varphi : \mathcal{X} \rightarrow \mathbb{R}$ be a differentiable and strictly convex function where $\mathcal{X} \subset \mathbb{R}$ is a convex set. Then, the deduced Bregman divergence $D_\varphi(x\|y) = 0$ implies $x = y$.*

PROOF OF LEMMA A.1. Suppose that there exist distinct $x, y \in \mathcal{X}$ such that $D_\varphi(x\|y) = 0$. By the definition of $D_\varphi(x\|y)$, we have

$$(5) \quad \varphi(x) = \varphi(y) + \varphi'(y)(x - y).$$

For arbitrary $t \in (0, 1)$, the strictly convexity of φ indicates that

$$(6) \quad \varphi(y + t(x - y)) = \varphi(tx + (1 - t)y) < t\varphi(x) + (1 - t)\varphi(y)$$

Combining Eqns. (5) and (6), we obtain

$$\varphi(y + t(x - y)) < \varphi(y) + t\varphi'(y)(x - y).$$

However, this contradicts to the fact that φ is a convex function. Therefore, distinct x, y can not yield that $D_\varphi(x\|y) = 0$. When x equals to y , it is straightforward to verify that $D_\varphi(x\|y) = 0$. This completes the proof. \square

A.2. Approximation properties of neural networks.

LEMMA A.2 (Theorem 3.3 in [25]). *Assume $g \in \mathcal{H}^\beta([0, 1]^d, B)$. For any $S_1, S_2 \in \mathbb{N}_+$, there exists a function f implemented by ReLU feedforward neural network with depth $L = 21(\lfloor \beta \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d$, width $M = 38(\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, such that*

$$|f(x) - g(x)| \leq 18B(\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + (\beta \vee 1)/2} (S_1 S_2)^{-2\beta/d},$$

for all $x \in [0, 1]^d \setminus \Omega([0, 1]^d, K, \delta)$. Here, \mathbb{N}_+ denotes the set of positive integers, $\lceil a \rceil$ means the smallest integer no less than a , $a \vee b = \max(a, b)$, and

$$\Omega([0, 1]^d, K, \delta) = \bigcup_{i=1}^d \left\{ x = (x_1, \dots, x_d)^\top : x_i \in \bigcup_{k=1}^{K-1} (k/K - \delta, k/K) \right\},$$

where $K = \lceil (S_1 S_2)^{2/d} \rceil$ and δ is an arbitrary scalar in $(0, 1/(3K))$.

LEMMA A.3 (Corollary B.2 in [15]). *Given any $g \in \mathcal{W}^{1,\infty}((0, 1)^d)$ with $\|g\|_{\mathcal{W}^{1,\infty}((0,1)^d)} < \infty$, for any $S_1, S_2 \in \mathbb{N}_+$, there exists a function f implemented by a deep ReLU network with depth $\mathcal{O}(d^2 S_1 \log S_1)$ and width $\mathcal{O}(2^d d S_2 \log S_2)$ such that $\|f\|_{\mathcal{W}^{1,\infty}((0,1)^d)} \leq C_1 \|g\|_{\mathcal{W}^{1,\infty}((0,1)^d)}$ and*

$$|f(x) - g(x)| \leq C_2 \|g\|_{\mathcal{W}^{1,\infty}((0,1)^d)} (S_1 S_2)^{-2/d},$$

for all $x \in (0, 1)^d$. Here, C_1 and C_2 are constants depending only on d .

A.3. Complexity of neural network function space. Let \mathcal{D} be a subset of \mathbb{R}^d . Given a positive real number ϵ , a set \mathcal{C} is called an ϵ -covering set of \mathcal{D} with respect to the supremum norm if $\mathcal{C} \subset \mathcal{D}$ and for any $x \in \mathcal{D}$, there exists $z \in \mathcal{C}$ such that $\|x - z\|_\infty < \epsilon$. Then, the minimal cardinality of all possible \mathcal{C} is termed the covering number of \mathcal{D} , denoted as $\mathcal{N}(\epsilon, \|\cdot\|_\infty, \mathcal{D})$.

Furthermore, consider a function space \mathcal{F} whose elements are defined on \mathcal{X} . Given an integer n and $\mathcal{D}_n = \{x_1, \dots, x_n\} \subset \mathcal{X}^n$, the covering number of \mathcal{F} constrained on \mathcal{D}_n is defined as the covering number of $\mathcal{F}|_{\mathcal{D}_n}$, where

$$\mathcal{F}|_{\mathcal{D}_n} = \left\{ (f(x_1), \dots, f(x_n))^\top : f \in \mathcal{F} \right\}.$$

Then, the covering number of \mathcal{F} with respect to n , denoted as $\mathcal{N}_n(\epsilon, \|\cdot\|_\infty, \mathcal{F})$, is defined as $\sup_{\mathcal{D}_n} \mathcal{N}(\epsilon, \|\cdot\|_\infty, \mathcal{F}|_{\mathcal{D}_n})$.

LEMMA A.4 (Theorem 12.2 in [2]). *Let \mathcal{F} be a set of real functions that map from a domain \mathcal{X} to a bounded interval $[0, B]$. Denote the pseudo-dimension of \mathcal{F} as $\text{Pdim}(\mathcal{F})$. Then, for $n \geq \text{Pdim}(\mathcal{F})$ and $B \geq \epsilon$, we have*

$$\mathcal{N}_n(\epsilon, \|\cdot\|_\infty, \mathcal{F}) \leq \left(\frac{eBn}{\epsilon \text{Pdim}(\mathcal{F})} \right)^{\text{Pdim}(\mathcal{F})}.$$

LEMMA A.5 (Theorem 7 in [3]). *Let \mathcal{F}_{NN} be a neural network function class with depth L and number of parameters S . Then, there exists a universal constant C such that*

$$\text{Pdim}(\mathcal{F}_{\text{NN}}) \leq CSL \log S.$$

A.4. Stochastic error analysis.

THEOREM A.6. *Let Z be a random vector supported on $\mathcal{Z} \subset \mathbb{R}^d$, and let $\mathcal{D}_n = \{Z_1, \dots, Z_n\}$ be a random sample of Z , whose elements are independent and identically distributed. Let \mathcal{F}_n be a nonrandom function space with elements real-valued. For a functional $g : \mathcal{F}_n \times \mathcal{Z} \rightarrow \mathbb{R}$, assume there exist some universal sequences $\xi_n, \zeta_n > 0$, such that (i) $\sup_{f \in \mathcal{F}_n, Z \in \mathcal{Z}} |g(f, Z)| \leq \xi_n$, (ii) $\mathbb{E}[g(f, Z)^2] \leq \zeta_n \mathbb{E}[g(f, Z)]$ for all $f \in \mathcal{F}_n$, where the expectation are taken with respect to Z , and (iii) there exists an integer $N > 0$ such that for any $n \geq N$, $\zeta_n \geq \eta$ for some constant $\eta \geq 1$, and $\xi_n \leq \gamma \zeta_n$ for some constant $\gamma > 0$. Then, for $n \geq N$, we have*

$$\begin{aligned} & \mathbb{P} \left(\exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \epsilon \{ \alpha + \beta + \mathbb{E}[g(f, Z)] \} \right) \\ & \leq 14 \mathcal{N}_n(\varpi \epsilon \beta, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}) \exp \left(- \frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40(\xi_n^2 \vee \zeta_n)(1+\epsilon)} \right). \end{aligned}$$

where $\alpha, \beta > 0$, $0 < \epsilon \leq 1/2$, $\varpi = (6\eta - 2)/(30\eta + 3\gamma\eta)$, and $\omega = 400/(\gamma + 60)^2$.

REMARK 6. Theorem A.6 is a generalization of Theorem 11.4 in [19]. Particularly, the case that $Z = (X, Y)$ and $g(f, Z) = |f(X) - Y|^2 - |\mathbb{E}(Y|X) - Y|^2$ represents classical nonparametric regression. The proof of Theorem A.6 can be found in Appendix C.

A.5. Grönnwall's inequality.

LEMMA A.7 (Lemma 37 in [24]). *Suppose that $f(t)$ is a scalar-output function defined on the interval $[a, b]$, satisfying $df(t)/dt \leq \alpha f(t) + g(t)$ with some constant $\alpha \geq 0$. Then, we have*

$$f(b) \leq e^{\alpha(b-a)} f(a) + \int_a^b e^{\alpha(b-t)} g(t) dt.$$

APPENDIX B: PROOFS OF MAIN RESULTS

For a sub-exponentially distributed random variable X , there exists a universal constant $\varsigma > 0$ such that $\mathbb{E} \exp(\varsigma |X|) < \infty$, where ς is termed the sub-exponential index. Throughout the subsequent proofs, we let ς be a uniform sub-exponential index without loss of generality, due to the finite number of associated sub-exponentially distributed random variables we will handle.

B.1. Proof of Theorem 3.1. For any square-integrable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, define

$$\begin{aligned} L(f) &= \mathbb{E}_P[f(X)^2 - 2f(X)r_0(X)], \\ L_n(f) &= \frac{1}{n} \sum_{i=1}^n f(X_i^s)^2 - \frac{2}{n} \sum_{i=1}^n f(X_i^t). \end{aligned}$$

Here, $\mathbb{E}_P[h(X)] \equiv \mathbb{E}[h(X^s)]$ for any X^s -integrable function h , where the expectation is taken with respect to X^s . Analogously, $\mathbb{E}_Q[h(X)] \equiv \mathbb{E}[h(X^t)]$ for any X^t -integrable function h .

LEMMA B.1. *Assume that $r_0(X^s)$ is a square-integrable. Then, $\mathcal{R}^s(\hat{r}_{\text{LS}}) \leq \mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] + 2 \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P[f(X) - r_0(X)]^2$.*

PROOF OF LEMMA B.1. For any $f \in \mathcal{F}_{\text{NN}} \subset \mathcal{L}^2(X^s)$, we have

$$\begin{aligned} \mathcal{R}^s(\hat{r}_{\text{LS}}) &= \mathbb{E}[L(\hat{r}_{\text{LS}}) - L(r_0)] \\ &\leq \mathbb{E}[L(\hat{r}_{\text{LS}}) - L(r_0)] + 2\mathbb{E}[L_n(f) - L_n(\hat{r}_{\text{LS}})] \\ &= \mathbb{E}[L(\hat{r}_{\text{LS}}) - L(r_0)] + 2\mathbb{E}[L_n(f) - L_n(r_0) + L_n(r_0) - L_n(\hat{r}_{\text{LS}})] \\ &= \mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] + 2[L(f) - L(r_0)] \\ &= \mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] + 2\mathbb{E}_P[f(X) - r_0(X)]^2. \end{aligned}$$

This indicates that $\mathcal{R}^s(\hat{r}_{\text{LS}}) \leq \mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] + 2 \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P[f(X) - r_0(X)]^2$. \square

LEMMA B.2. *Assume that $r_0(X^s)$ is sub-exponentially distributed. Let $\bar{\delta} = \bar{\delta}_n = (\log n)^{1+\kappa}$, with an arbitrarily fixed $\kappa \in (0, 1]$, and let $\underline{\delta} = 0$. Then, for sufficiently large n and $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, it follows that*

$$\mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] \leq \frac{c^* S L \log S (\log n)^{5+4\kappa}}{n},$$

where c^* is a constant not depending on S, L and n .

PROOF OF LEMMA B.2. Let $T_i = (X_i^s, X_i^t)$ for $i = 1, \dots, n$, $\mathcal{D}_n = \{T_1, \dots, T_n\}$, and $T = (X^s, X^t)$ be an independent copy of T_1 . Firstly, we have

$$\begin{aligned}\mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] &= \mathbb{E}_{\mathcal{D}_n}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] \\ &= \mathbb{E}_{\mathcal{D}_n}\{L(\hat{r}_{\text{LS}}) - L(r_0) - 2[L_n(\hat{r}_{\text{LS}}) - L_n(r_0)]\} \\ &= \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_T[g(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g(\hat{r}_{\text{LS}}, T_i)\right\},\end{aligned}$$

where $g(f, T) = [f(X^s)^2 - r_0(X^s)^2] - 2[f(X^t) - r_0(X^t)]$ for $f \in \mathcal{F}_{\text{NN}}$. Next, we introduce a truncation step. For any $\iota_n > 0$ and any $f \in \mathcal{F}_{\text{NN}}$, define

$$g_{\iota_n}(f, T) = [f(X^s)^2 - r_0(X^s)^2]\mathbb{1}(r_0(X^s) \leq \iota_n) - 2[f(X^t) - r_0(X^t)]\mathbb{1}(r_0(X^t) \leq \iota_n).$$

Then, it follows that

$$\begin{aligned}|g(f, T) - g_{\iota_n}(f, T)| &\leq |f(X^s)^2 - r_0(X^s)^2|\mathbb{1}(r_0(X^s) > \iota_n) + 2|f(X^t) - r_0(X^t)|\mathbb{1}(r_0(X^t) > \iota_n) \\ &\leq [\bar{\delta}_n^2 + r_0(X^s)^2]\mathbb{1}(r_0(X^s) > \iota_n) + 2[\bar{\delta}_n + r_0(X^t)]\mathbb{1}(r_0(X^t) > \iota_n).\end{aligned}$$

Taking expectation, we have

$$\begin{aligned}&\mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] \\ &= \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_T[g(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g(\hat{r}_{\text{LS}}, T_i)\right\} \\ &\leq \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LS}}, T_i)\right\} + 3\mathbb{E}_P\{[\bar{\delta}_n^2 + r_0(X)^2]\mathbb{1}(r_0(X) > \iota_n)\} \\ &\quad + 6\mathbb{E}_Q\{[\bar{\delta}_n + r_0(X)]\mathbb{1}(r_0(X) > \iota_n)\}.\end{aligned}$$

Specifically, observe that

$$\begin{aligned}&\mathbb{E}_P\{[\bar{\delta}_n^2 + r_0(X)^2]\mathbb{1}(r_0(X) > \iota_n)\} \\ &\leq \bar{\delta}_n^2 \mathbb{E}_P[\exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_n/2) + \mathbb{E}_P[r_0(X)^2 \mathbb{1}(r_0(X) > \iota_n)] \\ &\leq \bar{\delta}_n^2 \mathbb{E}_P[\exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_n/2) + \frac{16}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_Q\{[\bar{\delta}_n + r_0(X)]\mathbb{1}(r_0(X) > \iota_n)\} \\ &= \mathbb{E}_P\{[\bar{\delta}_n + r_0(X)]r_0(X)\mathbb{1}(r_0(X) > \iota_n)\} \\ &\leq \frac{2}{\varsigma} \bar{\delta}_n \mathbb{E}_P[\exp(r_0(X))] \exp(-\varsigma \iota_n/2) + \frac{16}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2).\end{aligned}$$

Here, we have applied the inequalities $a \leq \exp(a)$ and $\mathbb{1}(a > 0) \leq \exp(a)$ for $a \in \mathbb{R}$. As a consequence, we obtain

$$\begin{aligned}&\mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] \\ &\leq \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LS}}, T_i)\right\} + c_1(\bar{\delta}_n^2 + \bar{\delta}_n + 1) \exp(-\varsigma \iota_n/2),\end{aligned}$$

where c_1 is a constant which depends only on ς and $\mathbb{E}_P[\exp(\varsigma r_0(X))]$. Recall that $\bar{\delta}_n = (\log n)^{1+\kappa}$. Setting ι_n to $(2\varsigma^{-1} \log n) \vee 1$ yields

$$\begin{aligned} & \mathbb{E}[L(\hat{r}_{\text{LS}}) - 2L_n(\hat{r}_{\text{LS}}) + L(r_0)] \\ & \leq \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LS}}, T_i) \right\} + c_1[(\log n)^{1+\kappa} + 1]^2 n^{-1}. \end{aligned}$$

Then, we proceed to verify the conditions in Theorem A.6. Notice that

$$\sup_{f \in \mathcal{F}_{\text{NN}}, T \in \mathbb{R}^{2d}} |g_{\iota_n}(f, T)| \leq \bar{\delta}_n^2 + \iota_n^2 + 2\bar{\delta}_n + 2\iota_n \leq 6\bar{\delta}_n^2 = 6(\log n)^{2+2\kappa},$$

whenever $\bar{\delta}_n \geq \iota_n$. Furthermore, for any $f \in \mathcal{F}_{\text{NN}}$,

$$\begin{aligned} & \mathbb{E}[g_{\iota_n}(f, T)] \\ & = \mathbb{E}_P \{ [f(X)^2 - r_0(X)^2] \mathbb{1}(r_0(X) \leq \iota_n) \} - 2\mathbb{E}_Q \{ [f(X) - r_0(X)] \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & = \mathbb{E}_P \{ [f(X)^2 - r_0(X)^2] \mathbb{1}(r_0(X) \leq \iota_n) \} - 2\mathbb{E}_P \{ [f(X) - r_0(X)] r_0(X) \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & = \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \}, \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E}[g_{\iota_n}(f, T)^2] \\ & = \mathbb{E}_P \{ [f(X)^2 - r_0(X)^2]^2 \mathbb{1}(r_0(X) \leq \iota_n) \} + 4\mathbb{E}_Q \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & \quad - 4\mathbb{E}_P \{ [f(X)^2 - r_0(X)^2] \mathbb{1}(r_0(X) \leq \iota_n) \} \mathbb{E}_Q \{ [f(X) - r_0(X)] \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & \leq 2(\bar{\delta}_n^2 + \iota_n^2) \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & \quad + 4\iota_n \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & \quad + 4\iota_n(\bar{\delta}_n + \iota_n) \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \} \\ & \leq 16\bar{\delta}_n^2 \mathbb{E}[g_{\iota_n}(f, T)] \\ & = 16(\log n)^{2+2\kappa} \mathbb{E}[g_{\iota_n}(f, T)], \end{aligned}$$

provided that n is sufficiently large such that $\bar{\delta}_n \geq \iota_n$. Hence, Theorem A.6 suggests that, for sufficiently large n such that $\bar{\delta}_n \geq \iota_n$, with $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, and for arbitrary $t > 0$, we have

$$\begin{aligned} & \mathbb{P}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LS}}, T_i) \geq t \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] - \frac{1}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LS}}, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LS}}, T)] \right\} \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_n} \left(\exists f \in \mathcal{F}_{\text{NN}} : \mathbb{E}[g_{\iota_n}(f, Z)] - \frac{1}{n} \sum_{i=1}^n g_{\iota_n}(f, Z_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{\iota_n}(f, T)] \right\} \right) \\ & \leq 14\mathcal{N}_n(c_2 t, \|\cdot\|_\infty, \{g_{\iota_n}(f, \cdot) : \mathcal{X}^t \times \mathcal{X}^s \rightarrow \mathbb{R}, f \in \mathcal{F}_{\text{NN}}\}) \exp\left(-\frac{nt}{c_3(\log n)^{4+4\kappa}}\right), \end{aligned}$$

where c_2, c_3 are universal constants and $\mathcal{X}^s, \mathcal{X}^t$ represents the domain of X^s, X^t , respectively. Subsequently, we bound the covering number. Fix $\{x_1^s, \dots, x_n^s\} \subset (\mathcal{X}^s)^n$ and $\{x_1^t, \dots, x_n^t\} \subset (\mathcal{X}^t)^n$. Let $\mathcal{C} = \{x_1^s, \dots, x_n^s, x_1^t, \dots, x_n^t\}$, and let $h^\sharp = \{h_1, \dots, h_k\}$ be an

ϵ -covering set of $\mathcal{F}_{\text{NN}|\mathcal{C}}$ where $h_i = f_i|_{\mathcal{C}}$ for some $f_i \in \mathcal{F}_{\text{NN}}(i = 1, \dots, k)$, such that for any $f \in \mathcal{F}_{\text{NN}}$, there exists $h^* = f_{|\mathcal{C}}^* \in h^\sharp$ satisfying $\|h^* - f_{|\mathcal{C}}\|_\infty < \epsilon$. This indicates

$$\begin{aligned} & |g_{\ell_n}(f, (x_i^s, x_i^t)) - g_{\ell_n}(f^*, (x_i^s, x_i^t))| \\ & \leq |f(x_i^s)^2 - f^*(x_i^s)^2| + 2|f(x_i^t) - f^*(x_i^t)| \\ & \leq 2(\bar{\delta}_n + 1)\epsilon. \end{aligned}$$

Therefore,

$$\mathcal{N}_n(c_2 t, \|\cdot\|_\infty, \{g_{\ell_n}(f, \cdot) : \mathcal{X}^t \times \mathcal{X}^s \rightarrow \mathbb{R}, f \in \mathcal{F}_{\text{NN}}\}) \leq \mathcal{N}_{2n}(c_2 t / [2(\bar{\delta}_n + 1)], \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}).$$

Then, with Lemma A.4 and Lemma A.5, for sufficiently large n with $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$ and any $a_n \geq 1/n$, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\ell_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\ell_n}(\hat{r}_{\text{LS}}, T_i) \right\} \\ & \leq a_n + 14 \int_{a_n}^{\infty} \mathcal{N}_{2n}(c_2 t / [2(\bar{\delta}_n + 1)], \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \exp\left(-\frac{nt}{c_3(\log n)^{4+4\kappa}}\right) dt \\ & \leq a_n + 14 \mathcal{N}_{2n}(c_2 a_n / [2(\bar{\delta}_n + 1)], \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \int_{a_n}^{\infty} \exp\left(-\frac{nt}{c_3(\log n)^{4+4\kappa}}\right) dt \\ & \leq a_n + 14 (c_4 n^2 \bar{\delta}_n^2)^{c_5 S L \log S} \cdot \frac{c_3 (\log n)^{4+4\kappa}}{n} \exp\left(-\frac{na_n}{c_3(\log n)^{4+4\kappa}}\right), \end{aligned}$$

where c_4 and c_5 are universal constants. Choose

$$a_n = \frac{c_3 c_5 (\log n)^{4+4\kappa}}{n} S L \log S \log (c_4 n^2 \bar{\delta}_n^2).$$

For sufficiently large n , we have

$$\mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\ell_n}(\hat{r}_{\text{LS}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\ell_n}(\hat{r}_{\text{LS}}, T_i) \right\} \leq \frac{c_6 S L \log S (\log n)^{5+4\kappa}}{n},$$

where c_6 is a constant not depending on S, L and n . This completes the proof. \square

LEMMA B.3. Assume that

- (i) $r_0(x) \in \mathcal{H}_{\text{Loc}}^{\beta_r}(\mathbb{R}^d, B_u)$ with $\beta_r > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0, m \geq 0$;
- (ii) $r_0(X^s)$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Suppose that the depth L and width M of \mathcal{F}_{NN} are expressed as

$$L = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d + 3,$$

$$M = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil,$$

for any $S_1, S_2 \in \mathbb{N}_+$. Let $\bar{\delta} = \bar{\delta}_n = (\log n)^{1+\kappa}$, with an arbitrarily fixed $\kappa \in (0, 1]$, and $\underline{\delta} = 0$. Then, for sufficiently large n , it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P[f(X) - r_0(X)]^2 \\ & \leq c^* \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 + \frac{(\log n)^{2+2\kappa}}{n} \right\}, \end{aligned}$$

where c^* is a constant not depending on S_1, S_2 and n .

PROOF OF LEMMA B.3. For any $\iota_n > 0$, observe that

$$\begin{aligned} & \mathbb{E}_P[f(X) - r_0(X)]^2 \\ &= \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(\|X\|_\infty \leq \iota_n)\} + \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(\|X\|_\infty > \iota_n)\}. \end{aligned}$$

On the one hand, it follows that

$$\begin{aligned} & \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(\|X\|_\infty > \iota_n)\} \\ &\leq 2\mathbb{E}_P\{f(X)^2 \mathbf{1}(\|X\|_\infty > \iota_n)\} + 2\mathbb{E}_P\{r_0(X)^2 \mathbf{1}(\|X\|_\infty > \iota_n)\} \\ &\leq 2\bar{\delta}_n^2 \mathbb{E}_P[\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_n / 2) \\ &\quad + \frac{32}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X) / 2) \exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_n / 2) \\ &\leq 2\bar{\delta}_n^2 \mathbb{E}_P[\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_n / 2) \\ &\quad + \frac{32}{\varsigma^2} \{\mathbb{E}_P[\exp(\varsigma r_0(X))] \mathbb{E}_P[\exp(\varsigma \|X\|_\infty)]\}^{1/2} \exp(-\varsigma \iota_n / 2) \\ &\leq c_1 (\bar{\delta}_n^2 + 1) \exp(-\varsigma \iota_n / 2), \end{aligned}$$

where c_1 is a constant which depends only on $\varsigma, \mathbb{E}_P[\exp(\varsigma \|X\|_\infty)]$ and $\mathbb{E}_P[\exp(\varsigma r_0(X))]$. On the other hand, we first notice that

$$\begin{aligned} & \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\ &= \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\ &\quad + \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(r_0(X) > \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\ &\leq \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\ &\quad + 2\mathbb{E}_P[f(X)^2 \mathbf{1}(r_0(X) > \iota_n)] + 2\mathbb{E}_P[r_0(X)^2 \mathbf{1}(r_0(X) > \iota_n)] \\ &\leq \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\ &\quad + 2\bar{\delta}_n^2 \mathbb{E}_P[\exp(\varsigma r_0(X) / 2)] \exp(-\varsigma \iota_n / 2) + \frac{32}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n / 2) \\ &\leq \mathbb{E}_P\{[f(X) - r_0(X)]^2 \mathbf{1}(r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n)\} + c_2 (\bar{\delta}_n^2 + 1) \exp(-\varsigma \iota_n / 2), \end{aligned}$$

where c_2 is a constant which depends only on ς and $\mathbb{E}_P[\exp(\varsigma r_0(X))]$. Then, we focus on $\{x : \|x\|_\infty \leq \iota_n\} = [-\iota_n, \iota_n]^d$. Let $r_0^*(x) = r_0(2\iota_n x - \iota_n 1_d)$ for $x \in [0, 1]^d$. Lemma A.2 demonstrates that for any $S_1, S_2 \in \mathbb{N}_+$, there exists a function f^* implemented by a ReLU network with depth $L^* = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d$, width $M^* = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, such that

$$|f^*(x) - r_0^*(x)| \leq 18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d},$$

for all $x \in [0, 1]^d \setminus \Omega([0, 1]^d, K, \Delta)$. Here,

$$\Omega([0, 1]^d, K, \Delta) = \bigcup_{i=1}^d \left\{ x = (x_1, \dots, x_d)^\top : x_i \in \bigcup_{k=1}^{K-1} (k/K - \Delta, k/K) \right\},$$

where $K = \lceil (S_1 S_2)^{2/d} \rceil$ and Δ is an arbitrary scalar in $(0, 1/(3K)]$. Let $f^\dagger(x) = f^*((x + \iota_n 1_d)/(2\iota_n))$ for $x \in [-\iota_n, \iota_n]^d$. We obtain that

$$|f^\dagger(x) - r_0(x)| \leq 18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d},$$

for all $x \in [-\iota_n, \iota_n]^d \setminus \Omega^\dagger$, where $\Omega^\dagger = \{x : (x + \iota_n \mathbf{1}_d) / (2\iota_n) \in \Omega([0, 1]^d, K, \Delta)\}$. Furthermore, note that

$$f^\dagger(x) = f^* \left(\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) = f^* \left(\text{relu} \left(\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) - \text{relu} \left(-\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) \right),$$

which is implemented by a neural network with ReLU activations, depth $L^\dagger = L^* + 1$, and width $M^\dagger = M^*$. In addition, let

$$f^\ddagger(x) = \begin{cases} \bar{\delta}_n, & f^\dagger(x) > \bar{\delta}_n, \\ f^\dagger(x), & 0 \leq f^\dagger(x) \leq \bar{\delta}_n, \\ 0, & f^\dagger(x) < 0. \end{cases}$$

A straightforward calculation shows that

$$f^\ddagger(x) = \text{relu}(-\text{relu}(-f^\dagger(x) + \bar{\delta}_n) + \bar{\delta}_n),$$

indicating that $f^\ddagger(x)$ can be implemented by a ReLU network with depth $L = L^* + 3$ and width $M = M^*$. Due to the arbitrariness of Δ , when $\bar{\delta}_n \geq \iota_n$, it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \mathbb{1}(\|X\|_\infty \leq \iota_n) \} \\ & \leq \mathbb{E}_P \{ [f^\ddagger(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \mathbb{1}(\|X\|_\infty \leq \iota_n) \} \\ & \leq \left[18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} \right]^2. \end{aligned}$$

Recall that $\bar{\delta}_n = (\log n)^{1+\kappa}$, and set ι_n to $(2\varsigma^{-1} \log n) \vee 1$. We conclude that for sufficiently large n satisfying $\bar{\delta}_n \geq \iota_n$ and $\log n \geq \varsigma/2$, it holds that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P [f(X) - r_0(X)]^2 \\ & \leq \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P \{ [f(X) - r_0(X)]^2 \mathbb{1}(r_0(X) \leq \iota_n) \mathbb{1}(\|X\|_\infty \leq \iota_n) \} \\ & \quad + (c_1 + c_2)(\bar{\delta}_n^2 + 1) \exp(-\varsigma \iota_n/2) \\ & \leq c_3 \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 + \frac{(\log n)^{2+2\kappa}}{n} \right\}, \end{aligned}$$

where c_3 is a constant not depending on S_1, S_2 and n . \square

PROOF OF THEOREM 3.1. To commence, we notice that Lemma B.1, Lemma B.2 and Lemma B.3 indicate

$$\begin{aligned} \mathcal{R}^s(\hat{r}_{\text{LS}}) & \leq \frac{c_1 S L \log S (\log n)^{5+4\kappa}}{n} \\ & \quad + c_2 \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 + \frac{(\log n)^{2+2\kappa}}{n} \right\}, \end{aligned}$$

where c_1, c_2 are constants not depending on S, L, S_1, S_2 and n , and S_1, S_2 satisfy the conditions that the network depth $L = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d + 3$, network width $M = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, for sufficiently large n and $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$. Therefore, by letting $S_1 = \mathcal{O}(n^{d/(2d+4\beta_r)})$ and $S_2 = \mathcal{O}(1)$, we obtain

$$M = \mathcal{O}(1), \quad L = \mathcal{O}\left(n^{\frac{d}{2d+4\beta_r}} \log n\right), \quad S = \mathcal{O}(M^2 L) = \mathcal{O}\left(n^{\frac{d}{2d+4\beta_r}} \log n\right),$$

yielding

$$\mathcal{R}^s(\hat{r}_{\text{LS}}) \leq c_3 n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(8+4\kappa)\vee(2m)},$$

where c_3 is a constant not depending on n , for $n \geq 2$. Furthermore, observe that

$$\begin{aligned} \mathcal{R}^t(\hat{r}_{\text{LS}}) &= \mathbb{E} [\hat{r}_{\text{LS}}(X^t) - r_0(X^t)]^2 \\ &= \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \right\} \end{aligned}$$

Similar truncation operation suggests that, for any $\iota_n > 0$, we have

$$\begin{aligned} \mathcal{R}^t(\hat{r}_{\text{LS}}) &= \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \right\} \\ &= \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbb{1}(r_0(X^s) \leq \iota_n) \right\} \\ &\quad + \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n) \right\} \\ &\leq \iota_n \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 \right\} \\ &\quad + \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n) \right\} \\ &= \iota_n \mathcal{R}^s(\hat{r}_{\text{LS}}) + \mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n) \right\}. \end{aligned}$$

Subsequently, note that

$$\begin{aligned} &\mathbb{E} \left\{ [\hat{r}_{\text{LS}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n) \right\} \\ &\leq 2\mathbb{E} [\hat{r}_{\text{LS}}(X^s)^2 r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n)] + 2\mathbb{E} [r_0(X^s)^3 \mathbb{1}(r_0(X^s) > \iota_n)] \\ &\leq 2\bar{\delta}_n^2 \mathbb{E} [r_0(X^s) \mathbb{1}(r_0(X^s) > \iota_n)] + 2\mathbb{E} [r_0(X^s)^3 \mathbb{1}(r_0(X^s) > \iota_n)] \\ &\leq 4\varsigma^{-1} \bar{\delta}_n^2 \mathbb{E} [\exp(\varsigma r_0(X^s))] \exp(-\varsigma \iota_n/2) + 432\varsigma^{-3} \mathbb{E} [\exp(\varsigma r_0(X^s))] \exp(-\varsigma \iota_n/2) \\ &\leq c_4 (\bar{\delta}_n^2 + 1) \exp(-\varsigma \iota_n/2) \\ &= c_4 [(\log n)^{2+2\kappa} + 1] \exp(-\varsigma \iota_n/2), \end{aligned}$$

where c_4 is a constant depending only on ς and $\mathbb{E}[\exp(\varsigma r_0(X^s))]$. Hence, by taking $\iota_n = (2\varsigma^{-1} \log n) \vee 1$, we obtain that

$$\mathcal{R}^t(\hat{r}_{\text{LS}}) \leq [(2\varsigma^{-1} \log n) \vee 1] \mathcal{R}^s(\hat{r}_{\text{LS}}) + \frac{c_4 [(\log n)^{2+2\kappa} + 1]}{n}.$$

This completes the proof. \square

B.2. Proof of Lemma 3.2.

PROOF OF LEMMA 3.2. Given $r_0 \in \mathcal{L}_{\text{LR}}(X^s)$, we first claim that for any function $f \in \mathcal{L}_{\text{LR}}(X^s)$, we have $\mathbb{E}D_{\text{LR}}(r_0(X^s)\|f(X^s)) < \infty$. In fact, it follows that

$$\begin{aligned} 0 &\leq \mathbb{E}D_{\text{LR}}(r_0(X^s)\|f(X^s)) \\ &= \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)] \\ &= \mathbb{E}_P \left(\{r_0(X) \log r_0(X) - [r_0(X) + 1] \log(r_0(X) + 1) + \log(f(X) + 1) \right. \\ &\quad \left. - r_0(X) \log f(X) + r_0(X) \log(f(X) + 1)\} \mathbf{1}(X \in \mathcal{X}^t) \right) \\ &\leq \mathbb{E}_P \left(\{ |r_0(X) \log r_0(X)| + [r_0(X) + 1] \log(r_0(X) + 1) + \log(f(X) + 1) \right. \\ &\quad \left. + |r_0(X) \log f(X)| + r_0(X) \log(f(X) + 1)\} \mathbf{1}(X \in \mathcal{X}^t) \right). \end{aligned}$$

Recall that $\mathbb{E}_P h(X) \equiv \mathbb{E}h(X^s)$ for any X^s -integrable function h . Then, observe that for any scalars $x, y \in \mathbb{R}_+$, we have $\max(\log x, \log(x+1)) \leq x$ and $|x \log y| \leq xy^{-1} + xy$. Hence, it holds that

$$\begin{aligned} 0 &\leq \mathbb{E}D_{\text{LR}}(r_0(X^s)\|f(X^s)) \\ &\leq \mathbb{E}_P \left(\{ |r_0(X) \log r_0(X)| + [r_0(X) + 1] \log(r_0(X) + 1) + \log(f(X) + 1) \right. \\ &\quad \left. + |r_0(X) \log f(X)| + r_0(X) \log(f(X) + 1)\} \mathbf{1}(X \in \mathcal{X}^t) \right) \\ &\leq \mathbb{E}_P \left(\{ 1 + r_0(X)^2 + [r_0(X) + 1]r_0(X) + f(X) \right. \\ &\quad \left. + r_0(X)f(X)^{-1} + r_0(X)f(X) \} \mathbf{1}(X \in \mathcal{X}^t) \right) \\ &= \mathbb{E}_P \{ 1 + r_0(X)^2 + [r_0(X) + 1]r_0(X) + f(X) \\ &\quad + r_0(X)f(X)^{-1} + r_0(X)f(X) \} < \infty. \end{aligned}$$

Next, note that $\mathbb{E}D_{\text{LR}}(r_0(X^s)\|r_0(X^s)) = 0$. Therefore, for any minimizer f^* of $\mathbb{E}D_{\text{LR}}(r_0(X^s)\|f(X^s))$ with respect to $f \in \mathcal{L}_{\text{LR}}(X^s)$, we have $\mathbb{E}D_{\text{LR}}(r_0(X^s)\|f^*(X^s)) = 0$, which indicates that

$$D_{\text{LR}}(r_0(X^s)\|f^*(X^s)) = 0 \text{ a.s. } X^s.$$

Let the set $\mathcal{A} = \{x \in \mathcal{X}^t : D_{\text{LR}}(r_0(x)\|f^*(x)) = 0\}$. Notice that $\varphi''_{\text{LR}}(x) = [x(x+1)]^{-1} > 0$ for any $x \in \mathbb{R}_+$, thus φ_{LR} is strictly convex over \mathbb{R}_+ . By Lemma A.1, we obtain that

$$1 = \mathbb{P}(X^s \in \mathcal{A}) \leq \mathbb{P}(r_0(X^s) = f^*(X^s)).$$

Consequently, it follows that $r_0(X^s) = f^*(X^s)$ a.s. X^s . \square

B.3. Proof of Theorem 3.3.

For any function $f \in \mathcal{L}_{\text{LR}}(X^s)$, define

$$\begin{aligned} J(f) &= \mathbb{E}_P [\log(f(X) + 1) - r_0(X) \log f(X) + r_0(X) \log(f(X) + 1)], \\ J_n(f) &= \frac{1}{n} \sum_{i=1}^n \log(f(X_i^s) + 1) + \frac{1}{n} \sum_{i=1}^n [-\log f(X_i^t) + \log(f(X_i^t) + 1)]. \end{aligned}$$

Here, $\mathbb{E}_P[h(X)] \equiv \mathbb{E}[h(X^s)]$ for any X^s -integrable function h , where the expectation is taken with respect to X^s . Analogously, $\mathbb{E}_Q[h(X)] \equiv \mathbb{E}[h(X^t)]$ for any X^t -integrable function h .

LEMMA B.4. Assume that $r_0(X^s)$ and $r_0(X^s)^{-1}\mathbb{1}(X^s \in \mathcal{X}^t)$ are square-integrable. Then, it follows that

$$\begin{aligned} & \mathbb{E}D_{\text{LR}}(r_0(X^s)\|\hat{r}_{\text{LR}}(X^s)) \\ & \leq \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] + 2 \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P D_{\text{LR}}(r_0(X)\|f(X)). \end{aligned}$$

PROOF OF LEMMA B.4. Given that $r_0(X^s)$ and $r_0(X^s)^{-1}\mathbb{1}(X^s \in \mathcal{X}^t)$ are square-integrable, it is evident that $r_0 \in \mathcal{L}_{\text{LR}}(X^s)$. In addition, setting $\underline{\delta}$ to arbitrary positive scalar implies $\mathcal{F}_{\text{NN}} \subset \mathcal{L}_{\text{LR}}(X^s)$. Subsequently, for any $f \in \mathcal{F}_{\text{NN}}$, we have

$$\begin{aligned} \mathbb{E}D_{\text{LR}}(r_0(X^s)\|\hat{r}_{\text{LR}}(X^s)) &= \mathbb{E}[J(\hat{r}_{\text{LR}}) - J(r_0)] \\ &\leq \mathbb{E}[J(\hat{r}_{\text{LR}}) - J(r_0)] + 2\mathbb{E}[J_n(f) - J_n(\hat{r}_{\text{LR}})] \\ &= \mathbb{E}[J(\hat{r}_{\text{LR}}) - J(r_0)] + 2\mathbb{E}[J_n(f) - J_n(r_0) + J_n(r_0) - J_n(\hat{r}_{\text{LR}})] \\ &= \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] + 2[J(f) - J(r_0)] \\ &= \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] + 2\mathbb{E}_P D_{\text{LR}}(r_0(X)\|f(X)). \end{aligned}$$

Take the infimum on both sides with respect to $f \in \mathcal{F}_{\text{NN}}$ and we complete the proof. \square

LEMMA B.5. Assume that $r_0(X^s)$ and $r_0(X^s)^{-1}\mathbb{1}(X^s \in \mathcal{X}^t)$ are sub-exponentially distributed random variables. Let $\bar{\delta} = \bar{\delta}_n = (\log n)^{1+\kappa}$ and $\underline{\delta} = \underline{\delta}_n = (\log n)^{-1-\kappa}$ for arbitrarily fixed constant $\kappa \in (0, 1]$. Then, for sufficiently large n and $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, it follows that

$$\mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] \leq \frac{c^* S L \log S (\log n)^{6+5\kappa}}{n},$$

where c^* is a constant not depending on S, L and n .

PROOF OF LEMMA B.5. Let $T_i = (X_i^s, X_i^t)$ for $i = 1, \dots, n$, $\mathcal{D}_n = \{T_1, \dots, T_n\}$, and $T = (X^s, X^t)$ be an independent copy of T_1 . Firstly, we have

$$\begin{aligned} \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] &= \mathbb{E}_{\mathcal{D}_n}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] \\ &= \mathbb{E}_{\mathcal{D}_n}\{J(\hat{r}_{\text{LR}}) - J(r_0) - 2[J_n(\hat{r}_{\text{LR}}) - J_n(r_0)]\} \\ &= \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_T[g(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g(\hat{r}_{\text{LR}}, T_i)\right\}, \end{aligned}$$

where

$$\begin{aligned} g(f, T) &= \log(f(X^s) + 1) - \log(r_0(X^s) + 1) \\ &\quad + \log r_0(X^t) - \log f(X^t) \\ &\quad + \log(f(X^t) + 1) - \log(r_0(X^t) + 1), \end{aligned}$$

for $f \in \mathcal{F}_{\text{NN}}$. Next, we introduce a truncation step. For any $\iota_n \geq 1$ and any $f \in \mathcal{F}_{\text{NN}}$, define

$$\begin{aligned} g_{\iota_n}(f, T) &= [\log(f(X^s) + 1) - \log(r_0(X^s) + 1)]\mathbb{1}(\iota_n^{-1} \leq r_0(X^s) \leq \iota_n) \\ &\quad + [\log r_0(X^t) - \log f(X^t)]\mathbb{1}(\iota_n^{-1} \leq r_0(X^t) \leq \iota_n) \\ &\quad + [\log(f(X^t) + 1) - \log(r_0(X^t) + 1)]\mathbb{1}(\iota_n^{-1} \leq r_0(X^t) \leq \iota_n). \end{aligned}$$

Then, it follows that

$$\begin{aligned}
|g(f, T) - g_{\iota_n}(f, T)| &\leq |\log(f(X^s) + 1) - \log(r_0(X^s) + 1)| \mathbb{1}(r_0(X^s) > \iota_n) \\
&\quad + |\log(f(X^s) + 1) - \log(r_0(X^s) + 1)| \mathbb{1}(r_0(X^s) < \iota_n^{-1}) \\
&\quad + |\log r_0(X^t) - \log f(X^t)| \mathbb{1}(r_0(X^t) > \iota_n) \\
&\quad + |\log r_0(X^t) - \log f(X^t)| \mathbb{1}(r_0(X^t) < \iota_n^{-1}) \\
&\quad + |\log(f(X^t) + 1) - \log(r_0(X^t) + 1)| \mathbb{1}(r_0(X^t) > \iota_n) \\
&\quad + |\log(f(X^t) + 1) - \log(r_0(X^t) + 1)| \mathbb{1}(r_0(X^t) < \iota_n^{-1}) \\
&\leq [\log(\bar{\delta}_n + 1) + \log(r_0(X^s) + 1)] \mathbb{1}(r_0(X^s) > \iota_n) \\
&\quad + [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X^s) < \iota_n^{-1}) \\
&\quad + [\log r_0(X^t) + \log \bar{\delta}_n] \mathbb{1}(r_0(X^t) > \iota_n) \\
&\quad + [-\log r_0(X^t) + \log \bar{\delta}_n] \mathbb{1}(r_0(X^t) < \iota_n^{-1}) \\
&\quad + [\log(\bar{\delta}_n + 1) + \log(r_0(X^t) + 1)] \mathbb{1}(r_0(X^t) > \iota_n) \\
&\quad + [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X^t) < \iota_n^{-1}).
\end{aligned}$$

Taking expectation, we have

$$\begin{aligned}
&\mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] \\
&= \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g(\hat{r}_{\text{LR}}, T_i) \right\} \\
&\leq \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LR}}, T_i) \right\} \\
&\quad + 3\mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + \log(r_0(X) + 1)] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\quad + 3\mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&\quad + 3\mathbb{E}_Q \{ [\log r_0(X) + \log \bar{\delta}_n] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\quad + 3\mathbb{E}_Q \{ [-\log r_0(X) + \log \bar{\delta}_n] \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&\quad + 3\mathbb{E}_Q \{ [\log(\bar{\delta}_n + 1) + \log(r_0(X) + 1)] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\quad + 3\mathbb{E}_Q \{ [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X) < \iota_n^{-1}) \}.
\end{aligned}$$

Specifically, for $n \geq 3$, observe that firstly,

$$\begin{aligned}
&\mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + \log(r_0(X) + 1)] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + r_0(X)] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \log(\bar{\delta}_n + 1) \mathbb{E}_P[\exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_n/2) + \mathbb{E}_P \{ r_0(X) \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \log(\bar{\delta}_n + 1) \mathbb{E}_P[\exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_n/2) + \frac{2}{\varsigma} \mathbb{E}_P \{ \exp(\varsigma r_0(X)) \} \exp(-\varsigma \iota_n/2),
\end{aligned}$$

secondly,

$$\begin{aligned}
& \mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&= [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}_P [\mathbb{1}(X \in \mathcal{X}^t) \mathbb{1}(r_0(X) < \iota_n^{-1})] \\
&= [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}_P [\mathbb{1}(X \in \mathcal{X}^t) \mathbb{1}(r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t) > \iota_n)] \\
&\leq [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}[\exp(\varsigma r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t)/2)] \exp(-\varsigma \iota_n/2),
\end{aligned}$$

thirdly,

$$\begin{aligned}
& \mathbb{E}_Q \{ [\log r_0(X) + \log \bar{\delta}_n] \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \mathbb{E}_Q \{ [r_0(X) + \log \bar{\delta}_n] \mathbb{1}(r_0(X) > \iota_n) \} \\
&= \mathbb{E}_P \{ [r_0(X) + \log \bar{\delta}_n] r_0(X) \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \frac{2}{\varsigma} \log(\bar{\delta}_n) \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2) + \frac{16}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2),
\end{aligned}$$

fourthly,

$$\begin{aligned}
& \mathbb{E}_Q \{ [-\log r_0(X) + \log \bar{\delta}_n] \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&= \mathbb{E}_P \{ [-\log r_0(X) + \log \bar{\delta}_n] r_0(X) \mathbb{1}(X \in \mathcal{X}^t) \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&\leq (1 + \log \bar{\delta}_n) \mathbb{E}_P [\mathbb{1}(X \in \mathcal{X}^t) \mathbb{1}(r_0(X) < \iota_n^{-1})] \\
&= (1 + \log \bar{\delta}_n) \mathbb{E}_P [\mathbb{1}(X \in \mathcal{X}^t) \mathbb{1}(r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t) > \iota_n)] \\
&\leq (1 + \log \bar{\delta}_n) \mathbb{E}_P[\exp(\varsigma r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t)/2)] \exp(-\varsigma \iota_n/2),
\end{aligned}$$

fifthly,

$$\begin{aligned}
& \mathbb{E}_Q \{ [\log(\bar{\delta}_n + 1) + \log(r_0(X) + 1)] \mathbb{1}(r_0(X) > \iota_n) \} \\
&= \mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + \log(r_0(X) + 1)] r_0(X) \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \mathbb{E}_P \{ [\log(\bar{\delta}_n + 1) + r_0(X)] r_0(X) \mathbb{1}(r_0(X) > \iota_n) \} \\
&\leq \frac{2}{\varsigma} \log(\bar{\delta}_n + 1) \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2) + \frac{16}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2),
\end{aligned}$$

and lastly,

$$\begin{aligned}
& \mathbb{E}_Q \{ [\log(\bar{\delta}_n + 1) + 1] \mathbb{1}(r_0(X) < \iota_n^{-1}) \} \\
&= [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}_P [r_0(X) \mathbb{1}(r_0(X) < \iota_n^{-1})] \\
&\leq [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}_P [\mathbb{1}(r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t) > \iota_n)] \\
&\leq [\log(\bar{\delta}_n + 1) + 1] \mathbb{E}_P[\exp(\varsigma r_0(X)^{-1} \mathbb{1}(X \in \mathcal{X}^t)/2)] \exp(-\varsigma \iota_n/2).
\end{aligned}$$

Here, we have applied the inequalities $a \leq \exp(a)$ and $\mathbb{1}(a > 0) \leq \exp(a)$ for $a \in \mathbb{R}$, as well as that $-a \log a \leq 1$ for $a \in \mathbb{R}_+$. Consequently, we obtain

$$\begin{aligned}
& \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] \\
&\leq \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LR}}, T_i) \right\} + c_1(1 + \log \bar{\delta}_n) \exp(-\varsigma \iota_n/2),
\end{aligned}$$

where c_1 is a constant which depends only on ς , $\mathbb{E}_P\{\exp(\varsigma r_0(X))\}$ and $\mathbb{E}_P[\exp(\varsigma r_0(X)^{-1}\mathbf{1}(X \in \mathcal{X}^t))]$. Recall that $\bar{\delta}_n = (\log n)^{1+\kappa}$ with $\kappa \in (0, 1]$. Setting ι_n to $(2\varsigma^{-1}\log n) \vee 1$ yields

$$\begin{aligned} & \mathbb{E}[J(\hat{r}_{\text{LR}}) - 2J_n(\hat{r}_{\text{LR}}) + J(r_0)] \\ & \leq \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LR}}, T_i) \right\} + c_1(1 + \log n)n^{-1}. \end{aligned}$$

Then, we proceed to verify the conditions in Theorem A.6. Notice that

$$\begin{aligned} \sup_{f \in \mathcal{F}_{\text{NN}}, T \in \mathbb{R}^{2d}} |g_{\iota_n}(f, T)| & \leq 3[\log(\bar{\delta}_n + 1) + \log(\iota_n + 1)] \\ & \leq c_2(1 + \log n), \end{aligned}$$

where c_2 is a constant which depends only on ς and κ . Furthermore, for any $f \in \mathcal{F}_{\text{NN}}$,

$$\mathbb{E}[g_{\iota_n}(f, T)] = \mathbb{E}_P[D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)].$$

The smoothness of φ_{LR} then implies that for sufficiently large n such that $\bar{\delta}_n \geq \iota_n$, we have

$$\mathbb{E}[g_{\iota_n}(f, T)] \geq \frac{1}{\bar{\delta}_n(\bar{\delta}_n + 1)} \mathbb{E}_P \{ [r_0(X) - f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \}.$$

Hence, it follows that

$$\begin{aligned} & \mathbb{E}[g_{\iota_n}(f, T)^2] \\ & \leq 3\mathbb{E}_P \{ [\log(f(X) + 1) - \log(r_0(X) + 1)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \quad + 3\mathbb{E}_Q \{ [\log r_0(X) - \log f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \quad + 3\mathbb{E}_Q \{ [\log(f(X) + 1) - \log(r_0(X) + 1)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & = 3\mathbb{E}_P \{ [\log(f(X) + 1) - \log(r_0(X) + 1)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \quad + 3\mathbb{E}_P \{ [\log r_0(X) - \log f(X)]^2 r_0(X) \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \quad + 3\mathbb{E}_P \{ [\log(f(X) + 1) - \log(r_0(X) + 1)]^2 r_0(X) \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \leq 3(1 + \bar{\delta}_n^2 \iota_n + \iota_n) \mathbb{E}_P \{ [r_0(X) - f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \} \\ & \leq 18(\log n)^{5+5\kappa} \mathbb{E}[g_{\iota_n}(f, T)], \end{aligned}$$

provided that $\bar{\delta}_n \geq \iota_n$. Therefore, Theorem A.6 suggests that, for sufficiently large n such that $\bar{\delta}_n \geq \iota_n$, with $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, and for arbitrary $t > 0$, we have

$$\begin{aligned} & \mathbb{P}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LR}}, T_i) \geq t \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] - \frac{1}{n} \sum_{i=1}^n g_{\iota_n}(\hat{r}_{\text{LR}}, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{\iota_n}(\hat{r}_{\text{LR}}, T)] \right\} \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_n} \left(\exists f \in \mathcal{F}_{\text{NN}} : \mathbb{E}[g_{\iota_n}(f, T)] - \frac{1}{n} \sum_{i=1}^n g_{\iota_n}(f, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{\iota_n}(f, T)] \right\} \right) \\ & \leq 14\mathcal{N}_n(c_3 t, \|\cdot\|_\infty, \{g_{\iota_n}(f, \cdot) : \mathcal{X}^t \times \mathcal{X}^s \rightarrow \mathbb{R}, f \in \mathcal{F}_{\text{NN}}\}) \exp \left(-\frac{nt}{c_4(\log n)^{5+5\kappa}} \right), \end{aligned}$$

where c_3, c_4 are universal constants and $\mathcal{X}^s, \mathcal{X}^t$ represents the domain of X^s, X^t , respectively. Subsequently, we bound the covering number. Fix $\{x_1^s, \dots, x_n^s\} \subset (\mathcal{X}^s)^n$ and $\{x_1^t, \dots, x_n^t\} \subset (\mathcal{X}^t)^n$. Let $\mathcal{C} = \{x_1^s, \dots, x_n^s, x_1^t, \dots, x_n^t\}$, and let $h^\sharp = \{h_1, \dots, h_k\}$ be an ϵ -covering set of $\mathcal{F}_{\text{NN}|\mathcal{C}}$ where $h_i = f_{i|\mathcal{C}}$ for some $f_i \in \mathcal{F}_{\text{NN}}(i=1, \dots, k)$, such that for any $f \in \mathcal{F}_{\text{NN}}$, there exists $h^* = f_{|\mathcal{C}}^* \in h^\sharp$ satisfying $\|h^* - f_{|\mathcal{C}}\|_\infty < \epsilon$. This indicates

$$\begin{aligned} & |g_{\ell_n}(f, (x_i^s, x_i^t)) - g_{\ell_n}(f^*, (x_i^s, x_i^t))| \\ & \leq |f(x_i^s) - f^*(x_i^s)| + \bar{\delta}_n |f(x_i^t) - f^*(x_i^t)| + |f(x_i^t) - f^*(x_i^t)| \\ & \leq (\bar{\delta}_n + 2)\epsilon. \end{aligned}$$

Therefore,

$$\mathcal{N}_n(c_3 t, \|\cdot\|_\infty, \{g_{\ell_n}(f, \cdot) : \mathcal{X}^t \times \mathcal{X}^s \rightarrow \mathbb{R}, f \in \mathcal{F}_{\text{NN}}\}) \leq \mathcal{N}_{2n}(c_3 t / (\bar{\delta}_n + 2), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}).$$

Then, with Lemma A.4 and Lemma A.5, for sufficiently large n with $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$ and any $a_n \geq 1/n$, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\ell_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\ell_n}(\hat{r}_{\text{LR}}, T_i) \right\} \\ & \leq a_n + 14 \int_{a_n}^{\infty} \mathcal{N}_{2n}(c_3 t / (\bar{\delta}_n + 2), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \exp\left(-\frac{nt}{c_4(\log n)^{5+5\kappa}}\right) dt \\ & \leq a_n + 14 \mathcal{N}_{2n}(c_3 a_n / (\bar{\delta}_n + 2), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \int_{a_n}^{\infty} \exp\left(-\frac{nt}{c_4(\log n)^{5+5\kappa}}\right) dt \\ & \leq a_n + 14 (c_5 n^2 \bar{\delta}_n^2)^{c_6 S L \log S} \cdot \frac{c_4(\log n)^{5+5\kappa}}{n} \exp\left(-\frac{na_n}{c_4(\log n)^{5+5\kappa}}\right), \end{aligned}$$

where c_5 and c_6 are universal constants. Choose

$$a_n = \frac{c_4 c_6 (\log n)^{5+5\kappa}}{n} S L \log S \log (c_5 n^2 \bar{\delta}_n^2).$$

For sufficiently large n , we have

$$\mathbb{E}_{\mathcal{D}_n} \left\{ \mathbb{E}_T[g_{\ell_n}(\hat{r}_{\text{LR}}, T)] - \frac{2}{n} \sum_{i=1}^n g_{\ell_n}(\hat{r}_{\text{LR}}, T_i) \right\} \leq \frac{c_7 S L \log S (\log n)^{6+5\kappa}}{n},$$

where c_7 is a constant not depending on S, L and n . This completes the proof. \square

LEMMA B.6. *Assume that*

- (i) $r_0(x) \in \mathcal{H}_{\text{Loc}}^{\beta_r}(\mathbb{R}^d, B_u)$ with $\beta_r > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0, m \geq 0$;
- (ii) $r_0(X^s), r_0(X^s)^{-1} \mathbf{1}(X^s \in \mathcal{X}^t)$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Suppose that the depth L and width M of \mathcal{F}_{NN} are expressed as

$$L = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d + 3,$$

$$M = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil,$$

for any $S_1, S_2 \in \mathbb{N}_+$. Let $\bar{\delta} = \bar{\delta}_n = (\log n)^{1+\kappa}$ and $\underline{\delta} = \underline{\delta}_n = (\log n)^{-1-\kappa}$ for arbitrarily fixed $\kappa \in (0, 1]$. Then, for sufficiently large n , it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P D_{\text{LR}}(r_0(X) \| f(X)) \\ & \leq c^* \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 (\log n)^{1+\kappa} + \frac{(\log n)^2}{n} \right\}, \end{aligned}$$

where c^* is a constant not depending on S_1, S_2 and n .

PROOF OF LEMMA B.6. For any $\iota_n \geq 1$, observe that

$$\begin{aligned} \mathbb{E}_P D_{\text{LR}}(r_0(X) \| f(X)) &= \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t)] \\ &= \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)] \\ &\quad + \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) < \iota_n^{-1})] \\ &\quad + \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) > \iota_n)]. \end{aligned}$$

On the one hand, for $n \geq 3$, it follows that

$$\begin{aligned} & \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) < \iota_n^{-1})] \\ &= \mathbb{E}_P \left(\{r_0(X) \log r_0(X) - [r_0(X) + 1] \log(r_0(X) + 1) + \log(f(X) + 1) \right. \\ &\quad \left. - r_0(X) \log f(X) + r_0(X) \log(f(X) + 1)\} \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) < \iota_n^{-1}) \right) \\ &\leq 3 \log(\bar{\delta}_n + 1) \mathbb{E}_P [\mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) < \iota_n^{-1})] \\ &\leq 3 \log(\bar{\delta}_n + 1) \mathbb{E}_P [\mathbf{1}(r_0(X)^{-1} \mathbf{1}(X \in \mathcal{X}^t) > \iota_n)] \\ &\leq 3 \log(\bar{\delta}_n + 1) \mathbb{E}_P [\exp(\varsigma r_0(X)^{-1} \mathbf{1}(X \in \mathcal{X}^t)/2)] \exp(-\varsigma \iota_n/2), \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E}_P [D_{\text{LR}}(r_0(X) \| f(X)) \mathbf{1}(X \in \mathcal{X}^t) \mathbf{1}(r_0(X) > \iota_n)] \\ &= \mathbb{E}_P \left(\{r_0(X) \log r_0(X) - [r_0(X) + 1] \log(r_0(X) + 1) + \log(f(X) + 1) \right. \\ &\quad \left. - r_0(X) \log f(X) + r_0(X) \log(f(X) + 1)\} \mathbf{1}(r_0(X) > \iota_n) \right) \\ &\leq \mathbb{E}_P \left\{ [1 + 2r_0(X) + 2r_0(X)^2 + \log(\bar{\delta}_n + 1) + 2r_0(X) \log(\bar{\delta}_n + 1)] \mathbf{1}(r_0(X) > \iota_n) \right\} \\ &\leq \mathbb{E}_P \left\{ \left[1 + \frac{4}{\varsigma} \exp(\varsigma r_0(X)/2) + \frac{32}{\varsigma^2} \exp(\varsigma r_0(X)/2) + \log(\bar{\delta}_n + 1) \right. \right. \\ &\quad \left. \left. + \frac{4}{\varsigma} \exp(\varsigma r_0(X)/2) \log(\bar{\delta}_n + 1) \right] \exp(\varsigma r_0(X)/2) \right\} \exp(-\varsigma \iota_n/2) \\ &\leq c_1 [1 + \log(\bar{\delta}_n + 1)] \exp(-\varsigma \iota_n/2), \end{aligned}$$

where c_1 is a constant which depends only on ς and $\mathbb{E}_P[\exp(\varsigma r_0(X))]$. On the other hand, we first notice that

$$\begin{aligned}
& \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)] \\
&= \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty \leq \iota_n)] \\
&\quad + \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty > \iota_n)] \\
&\leq \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty \leq \iota_n)] \\
&\quad + \mathbb{E}_P [\{\log(f(X) + 1) - r_0(X) \log f(X) + r_0(X) \log(f(X) + 1)\}] \\
&\quad \cdot \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty > \iota_n) \\
&\leq \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty \leq \iota_n)] \\
&\quad + 3\iota_n \log(\bar{\delta}_n + 1)\mathbb{E}_P \mathbf{1}(\|X\|_\infty > \iota_n) \\
&\leq \mathbb{E}_P [D_{\text{LR}}(r_0(X)\|f(X))\mathbf{1}(X \in \mathcal{X}^t)\mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty \leq \iota_n)] \\
&\quad + 3\iota_n \log(\bar{\delta}_n + 1)\mathbb{E}_P [\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_n / 2) \\
&\leq \frac{1}{2}(\bar{\delta}_n \vee \iota_n)\mathbb{E}_P \{[r_0(X) - f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n)\mathbf{1}(\|X\|_\infty \leq \iota_n)\} \\
&\quad + 3\iota_n \log(\bar{\delta}_n + 1)\mathbb{E}_P [\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_n / 2).
\end{aligned}$$

Then, we focus on the region $\{x : \|x\|_\infty \leq \iota_n\} = [-\iota_n, \iota_n]^d$. Let $r_0^*(x) = r_0(2\iota_n x - \iota_n \mathbf{1}_d)$ for $x \in [0, 1]^d$. Lemma A.2 demonstrates that for any $S_1, S_2 \in \mathbb{N}_+$, there exists a function f^* implemented by a ReLU network with depth $L^* = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d$, width $M^* = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, such that

$$|f^*(x) - r_0^*(x)| \leq 18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d},$$

for all $x \in [0, 1]^d \setminus \Omega([0, 1]^d, K, \Delta)$. Here,

$$\Omega([0, 1]^d, K, \Delta) = \bigcup_{i=1}^d \left\{ x = (x_1, \dots, x_d)^\top : x_i \in \bigcup_{k=1}^{K-1} (k/K - \Delta, k/K) \right\},$$

where $K = \lceil (S_1 S_2)^{2/d} \rceil$ and Δ is an arbitrary scalar in $(0, 1/(3K)]$. Let $f^\dagger(x) = f^*((x + \iota_n \mathbf{1}_d)/(2\iota_n))$ for $x \in [-\iota_n, \iota_n]^d$. We obtain that

$$|f^\dagger(x) - r_0(x)| \leq 18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d},$$

for all $x \in [-\iota_n, \iota_n]^d \setminus \Omega^\dagger$, where $\Omega^\dagger = \{x : (x + \iota_n \mathbf{1}_d)/(2\iota_n) \in \Omega([0, 1]^d, K, \Delta)\}$. Furthermore, note that

$$f^\dagger(x) = f^* \left(\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) = f^* \left(\text{relu} \left(\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) - \text{relu} \left(-\frac{x + \iota_n \mathbf{1}_d}{2\iota_n} \right) \right),$$

which is implemented by a neural network with ReLU activations, depth $L^\dagger = L^* + 1$, and width $M^\dagger = M^*$. In addition, let

$$f^\ddagger(x) = \begin{cases} \bar{\delta}_n, & f^\dagger(x) > \bar{\delta}_n, \\ f^\dagger(x), & \underline{\delta}_n \leq f^\dagger(x) \leq \bar{\delta}_n, \\ \underline{\delta}_n, & f^\dagger(x) < \underline{\delta}_n. \end{cases}$$

A straightforward calculation shows that

$$f^\ddagger(x) = \text{relu}(-\text{relu}(-f^\dagger(x) + \bar{\delta}_n) + \bar{\delta}_n - \underline{\delta}_n) + \underline{\delta}_n,$$

indicating that $f^\ddagger(x)$ can be implemented by a ReLU network with depth $L = L^* + 3$ and width $M = M^*$. Due to the arbitrariness of Δ , when $\bar{\delta}_n \geq \iota_n$, it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P \{ [r_0(X) - f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n) \} \\ & \leq \mathbb{E}_P \left\{ [r_0(X) - f^\ddagger(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n) \right\} \\ & \leq \left[18c(\iota_n^m + 1)(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} \right]^2. \end{aligned}$$

Recall that $\bar{\delta}_n = (\log n)^{1+\kappa}$, and set ι_n to $(2\varsigma^{-1} \log n) \vee 1$. We conclude that for sufficiently large n satisfying $\bar{\delta}_n \geq \iota_n$ and $\log n \geq \varsigma/2$, it holds that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P D_{\text{LR}}(r_0(X) \| f(X)) \\ & \leq \frac{1}{2} (\bar{\delta}_n \vee \iota_n) \inf_{f \in \mathcal{F}_{\text{NN}}} \mathbb{E}_P \{ [r_0(X) - f(X)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X) \leq \iota_n) \mathbf{1}(\|X\|_\infty \leq \iota_n) \} \\ & \quad + c_2 \iota_n \log(\bar{\delta}_n + 1) \exp(-\varsigma \iota_n/2) \\ & \leq c_3 \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 (\log n)^{1+\kappa} + \frac{(\log n)^2}{n} \right\}, \end{aligned}$$

where c_2, c_3 are constants not depending on S_1, S_2 and n . \square

PROOF OF THEOREM 3.3. To commence, we notice that Lemmas B.4, B.5 and B.6 indicate

$$\begin{aligned} & \mathbb{E} D_{\text{LR}}(r_0(X^s) \| \hat{r}_{\text{LR}}(X^s)) \\ & \leq \frac{c_1 S L \log S (\log n)^{6+5\kappa}}{n} \\ & \quad + c_2 \left\{ \left[(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + (\beta_r \vee 1)/2} (S_1 S_2)^{-2\beta_r/d} (\log n)^m \right]^2 (\log n)^{1+\kappa} + \frac{(\log n)^2}{n} \right\}, \end{aligned}$$

where c_1, c_2 are constants not depending on S, L, S_1, S_2 and n , and S_1, S_2 satisfy the conditions that the network depth $L = 21(\lfloor \beta_r \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d + 3$, network width $M = 38(\lfloor \beta_r \rfloor + 1)^2 d^{\lfloor \beta_r \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, for sufficiently large n and $n \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$. Therefore, by letting $S_1 = \mathcal{O}(n^{d/(2d+4\beta_r)})$ and $S_2 = \mathcal{O}(1)$, we obtain

$$M = \mathcal{O}(1), \quad L = \mathcal{O}\left(n^{\frac{d}{2d+4\beta_r}} \log n\right), \quad S = \mathcal{O}(M^2 L) = \mathcal{O}\left(n^{\frac{d}{2d+4\beta_r}} \log n\right),$$

yielding

$$\mathbb{E} D_{\text{LR}}(r_0(X^s) \| \hat{r}_{\text{LR}}(X^s)) \leq c_3 n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(9+5\kappa) \vee (2m+1+\kappa)},$$

where c_3 is a constant not depending on n , for $n \geq 3$. Furthermore, note that for any $\iota_n \geq 1$, it follows that

$$\begin{aligned} \mathcal{R}^s(\hat{r}_{\text{LR}}) &= \mathbb{E} [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \\ &= \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X^s) \leq \iota_n) \right\} \\ &\quad + \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(X^s \in \mathcal{X}^t) \mathbf{1}(r_0(X^s) < \iota_n^{-1}) \right\} \\ &\quad + \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(r_0(X^s) > \iota_n) \right\}. \end{aligned}$$

Specifically, on one hand, the smoothness of φ_{LR} demonstrates that

$$\begin{aligned} & \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(\iota_n^{-1} \leq r_0(X^s) \leq \iota_n) \right\} \\ & \leq 2(\bar{\delta}_n \vee \iota_n)[(\bar{\delta}_n \vee \iota_n) + 1] \mathbb{E} [D_{\text{LR}}(r_0(X^s) \| \hat{r}_{\text{LR}}(X^s)) \mathbf{1}(\iota_n^{-1} \leq r_0(X^s) \leq \iota_n)] \\ & \leq 2(\bar{\delta}_n \vee \iota_n)[(\bar{\delta}_n \vee \iota_n) + 1] \mathbb{E} D_{\text{LR}}(r_0(X^s) \| \hat{r}_{\text{LR}}(X^s)). \end{aligned}$$

On the other hand, observe that

$$\begin{aligned} & \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(X^s \in \mathcal{X}^t) \mathbf{1}(r_0(X^s) < \iota_n^{-1}) \right\} \\ & \leq 2(1 + \bar{\delta}_n^2) \mathbb{E}_P [\mathbf{1}(r_0(X) \mathbf{1}(X \in \mathcal{X}^t) < \iota_n^{-1})] \\ & \leq 2(1 + \bar{\delta}_n^2) \mathbb{E}_P \mathbf{1}(r_0(X)^{-1} \mathbf{1}(X \in \mathcal{X}^t) > \iota_n) \\ & \leq 2(1 + \bar{\delta}_n^2) \mathbb{E} [\exp(\varsigma r_0(X)^{-1} \mathbf{1}(X \in \mathcal{X}^t)/2)] \exp(-\varsigma \iota_n/2), \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E} \left\{ [r_0(X^s) - \hat{r}_{\text{LR}}(X^s)]^2 \mathbf{1}(r_0(X^s) > \iota_n) \right\} \\ & \leq 2 \mathbb{E} [r_0(X^s)^2 \mathbf{1}(r_0(X^s) > \iota_n)] + 2\bar{\delta}_n^2 \mathbb{E} \mathbf{1}(r_0(X^s) > \iota_n) \\ & \leq \frac{32}{\varsigma^2} \mathbb{E}_P [\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_n/2) + 2\bar{\delta}_n^2 \mathbb{E}_P [\exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_n/2). \end{aligned}$$

Hence, by letting $\iota_n = (2\varsigma^{-1} \log n) \vee 1$, we have for $n \geq 3$,

$$\mathcal{R}^s(\hat{r}_{\text{LR}}) \leq c_4 n^{-\frac{2\beta_r}{d+2\beta_r}} (\log n)^{(11+7\kappa) \vee (2m+3+3\kappa)},$$

where c_4 is a constant not depending on n . In addition, we note that

$$\begin{aligned} \mathcal{R}^t(\hat{r}_{\text{LR}}) &= \mathbb{E} [\hat{r}_{\text{LR}}(X^t) - r_0(X^t)]^2 \\ &= \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \right\}. \end{aligned}$$

Similar truncation operation suggests that, for any $\varrho_n > 0$, we have

$$\begin{aligned} \mathcal{R}^t(\hat{r}_{\text{LR}}) &= \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \right\} \\ &= \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbf{1}(r_0(X^s) \leq \varrho_n) \right\} \\ &\quad + \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbf{1}(r_0(X^s) > \varrho_n) \right\} \\ &\leq \varrho_n \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 \right\} \\ &\quad + \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbf{1}(r_0(X^s) > \varrho_n) \right\} \\ &= \varrho_n \mathcal{R}^s(\hat{r}_{\text{LR}}) + \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbf{1}(r_0(X^s) > \varrho_n) \right\}. \end{aligned}$$

Subsequently, note that

$$\begin{aligned} & \mathbb{E} \left\{ [\hat{r}_{\text{LR}}(X^s) - r_0(X^s)]^2 r_0(X^s) \mathbf{1}(r_0(X^s) > \varrho_n) \right\} \\ & \leq 2\bar{\delta}_n^2 \mathbb{E}_P [r_0(X) \mathbf{1}(r_0(X) > \varrho_n)] + 2\mathbb{E}_P [r_0(X)^3 \mathbf{1}(r_0(X) > \varrho_n)] \\ & \leq \frac{4}{\varsigma} \bar{\delta}_n^2 \mathbb{E}_P [\exp(\varsigma r_0(X))] \exp(-\varsigma \varrho_n/2) + \frac{432}{\varsigma^3} \mathbb{E}_P [\exp(\varsigma r_0(X))] \exp(-\varsigma \varrho_n/2) \\ & \leq c_5 [(\log n)^{2+2\kappa} + 1] \exp(-\varsigma \varrho_n/2), \end{aligned}$$

where c_5 is a constant which depends only on ς and $\mathbb{E}_P[\exp(\varsigma r_0(X))]$. Hence, by taking $\varrho_n = (2\varsigma^{-1} \log n) \vee 1$, we obtain that

$$\mathcal{R}^t(\hat{r}_{\text{LR}}) \leq [(2\varsigma^{-1} \log n) \vee 1] \mathcal{R}^s(\hat{r}_{\text{LR}}) + \frac{c_5[(\log n)^{2+2\kappa} + 1]}{n}.$$

This completes the proof. \square

B.4. Proof of Lemma 4.1.

PROOF OF LEMMA 4.1. For any $\iota_n > 0$, observe that

$$\begin{aligned} \mathbb{E}(\|U_n - U\|_2^2 | V) &= \mathbb{E}(\|U_n - U\|_2^2 | V | \mathbf{1}(|V| \leq \iota_n)) + \mathbb{E}(\|U_n - U\|_2^2 | V | \mathbf{1}(|V| > \iota_n)) \\ &\leq \iota_n \gamma_n + \mathbb{E}(\|U_n - U\|_2^2 | V | \mathbf{1}(|V| > \iota_n)) \\ &\leq \iota_n \gamma_n + 4\varsigma^{-1} \mathbb{E}(\|U_n - U\|_2^2 \exp(\varsigma |V|/4) \mathbf{1}(|V| > \iota_n)) \\ &\leq \iota_n \gamma_n + 4\varsigma^{-1} \mathbb{E}(\|U_n - U\|_2^2 \exp(\varsigma |V|/2)) \exp(-\varsigma \iota_n/4), \end{aligned}$$

where we have applied the inequalities that $a \leq \exp(a)$ and $\mathbf{1}(a > 0) \leq \exp(a)$. Let $U_{n,(j)}$ and $U_{(j)}$ be the j -th entry of U_n and U , respectively, for $j = 1, \dots, d$. Note that

$$\begin{aligned} \mathbb{E}(\|U_n - U\|_2^2 \exp(\varsigma |V|/2)) &= \sum_{j=1}^d \mathbb{E}\left[\left(U_{n,(j)} - U_{(j)}\right)^2 \exp(\varsigma |V|/2)\right] \\ &\leq \sum_{j=1}^d \left[\mathbb{E}\left(U_{n,(j)} - U_{(j)}\right)^4 \mathbb{E} \exp(\varsigma |V|)\right]^{1/2} \\ &\leq \sum_{j=1}^d \left[8\mathbb{E}\left(U_{n,(j)}^4 + U_{(j)}^4\right) \mathbb{E} \exp(\varsigma |V|)\right]^{1/2} \\ &\leq d \left[8(\xi_n^4 + \mathbb{E}\|U\|_\infty^4) \mathbb{E} \exp(\varsigma |V|)\right]^{1/2} \\ &\leq c_3 d (\xi_n^2 + 1), \end{aligned}$$

where c_3 is a constant only depending on $\mathbb{E}\|U\|_\infty^4$ and $\mathbb{E} \exp(\varsigma |V|)$. Therefore, we have

$$\mathbb{E}(\|U_n - U\|_2^2 | V) \leq \iota_n \gamma_n + 4c_3 \varsigma^{-1} d (\xi_n^2 + 1) \exp(-\varsigma \iota_n/4).$$

Let $\iota_n = 4\varsigma^{-1} \log n$. Then, for $n \geq 2$, it follows that

$$\mathbb{E}(\|U_n - U\|_2^2 | V) \leq 4\varsigma^{-1} \gamma_n \log n + \frac{4c_3 \varsigma^{-1} d (\xi_n^2 + 1)}{n}.$$

This completes the proof. \square

B.5. Proof of Proposition 4.3.

PROOF OF PROPOSITION 4.3. For any $\iota_N > 0$, observe that

$$\begin{aligned}
& \mathbb{E} \left\| \hat{\theta}_N(X^t) - \theta_0(X^t) \right\|_2^2 \\
&= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \right] \\
&= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty \leq \iota_N) \right] \\
&\quad + \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right] \\
&\leq G(\iota_N) \mathbb{E} \left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \\
&\quad + \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right].
\end{aligned}$$

Let $\hat{\theta}_{N,(j)}(X^s)$ and $\theta_{0,(j)}(X^s)$ be the j -th component of $\hat{\theta}_N(X^s)$ and $\theta_0(X^s)$, respectively, for $j = 1, \dots, k$. By using Cauchy-Schwarz inequality twice, we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right] \\
&= \sum_{j=1}^k \mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right\} \\
&\leq \sum_{j=1}^k \mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^2 \cdot r_0(X^s) \exp(\varsigma \|X^s\|_\infty / 4) \right\} \exp(-\varsigma \iota_N / 4) \\
&\leq \sum_{j=1}^k \left(\mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^4 \exp(\varsigma \|X^s\|_\infty / 2) \right\} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N / 4) \\
&\leq \sum_{j=1}^k \left(\left\{ \mathbb{E} \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^8 \mathbb{E} \exp(\varsigma \|X^s\|_\infty) \right\}^{1/2} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N / 4) \\
&\leq k \left(\left\{ 128 \left(\xi_N^8 + \mathbb{E} \|\theta_0(X^s)\|^8 \right) \mathbb{E} \exp(\varsigma \|X^s\|_\infty) \right\}^{1/2} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N / 4) \\
&\leq c_3 k (\xi_N^2 + 1) \exp(-\varsigma \iota_N / 4),
\end{aligned}$$

where c_3 is a constant only depending on $\mathbb{E} \|\theta_0(X^s)\|^8$, $\mathbb{E} \exp(\varsigma \|X^s\|_\infty)$ and $\mathbb{E} [r_0(X^s)^2]$. Hence, let $\iota_N = 4\varsigma^{-1} \log N$ and we obtain the result. \square

B.6. Proof of Proposition 4.4.

PROOF OF PROPOSITION 4.4. For any $\iota_N > 0$, observe that

$$\begin{aligned} & \mathbb{E} \left\| \hat{\theta}_N(X^t) - \theta_0(X^t) \right\|_2^2 \\ &= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \right] \\ &= \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty \leq \iota_N) \right] \\ &\quad + \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right] \\ &\leq G(\iota_N) \mathbb{E} \left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \\ &\quad + \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right]. \end{aligned}$$

Let $\hat{\theta}_{N,(j)}(X^s)$ and $\theta_{0,(j)}(X^s)$ be the j -th component of $\hat{\theta}_N(X^s)$ and $\theta_0(X^s)$, respectively, for $j = 1, \dots, k$. By using Cauchy-Schwarz inequality twice, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \hat{\theta}_N(X^s) - \theta_0(X^s) \right\|_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right] \\ &= \sum_{j=1}^k \mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^2 \cdot r_0(X^s) \mathbf{1}(\|X^s\|_\infty > \iota_N) \right\} \\ &\leq \sum_{j=1}^k \mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^2 \cdot r_0(X^s) \exp(\varsigma \|X^s\|_\infty^2 / 4) \right\} \exp(-\varsigma \iota_N^2 / 4) \\ &\leq \sum_{j=1}^k \left(\mathbb{E} \left\{ \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^4 \exp(\varsigma \|X^s\|_\infty^2 / 2) \right\} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N^2 / 4) \\ &\leq \sum_{j=1}^k \left(\left\{ \mathbb{E} \left[\hat{\theta}_{N,(j)}(X^s) - \theta_{0,(j)}(X^s) \right]_2^8 \mathbb{E} \exp(\varsigma \|X^s\|_\infty^2) \right\}^{1/2} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N^2 / 4) \\ &\leq k \left(\left\{ 128 \left(\xi_N^8 + \mathbb{E} \|\theta_0(X^s)\|^8 \right) \mathbb{E} \exp(\varsigma \|X^s\|_\infty^2) \right\}^{1/2} \mathbb{E} [r_0(X^s)^2] \right)^{1/2} \exp(-\varsigma \iota_N^2 / 4) \\ &\leq c_3 k (\xi_N^2 + 1) \exp(-\varsigma \iota_N^2 / 4), \end{aligned}$$

where c_3 is a constant only depending on $\mathbb{E} \|\theta_0(X^s)\|^8$, $\mathbb{E} \exp(\varsigma \|X^s\|_\infty^2)$ and $\mathbb{E} [r_0(X^s)^2]$. Hence, let $\iota_N = 2(\varsigma^{-1} \log N)^{1/2}$ and we obtain the result. \square

B.7. Proof of Theorem 5.1. For any X^s -square-integrable function $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$ such that $\mathbb{E} \|f(X^s)\|_2^2 < \infty$, define

$$K^{\text{reg}}(f) = \mathbb{E}_P \|Y - f(X)\|_2^2,$$

$$K_N^{\text{reg}}(f) = \frac{1}{N} \sum_{i=1}^N \|Y_i^s - f(X_i^s)\|_2^2.$$

Here, $\mathbb{E}_P[h(X, Y)] \equiv \mathbb{E}[h(X^s, Y^s)]$ for any (X^s, Y^s) -integrable function h , where the expectation is taken with respect to (X^s, Y^s) .

LEMMA B.7. *Assume that $\|Y^s\|_\infty$ and $\|f_0(X^s)\|_\infty$ attain a finite second moment. Then,*

$$\begin{aligned} & \mathbb{E}\|\hat{f}_N^s(X^s) - f_0(X^s)\|_2^2 \\ & \leq \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] + 2 \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P\|f(X) - f_0(X)\|_2^2. \end{aligned}$$

PROOF OF LEMMA B.7. Given that $\|Y^s\|_\infty$ and $\|f_0(X^s)\|_\infty$ have a finite second moment, we have $\mathbb{E}\|Y^s\|_2^2 < \infty$ and $\mathbb{E}\|f_0(X^s)\|_2^2 < \infty$. For any $f \in \mathcal{F}_{\text{NN}}^{d_y}$, we have

$$\begin{aligned} & \mathbb{E}\|\hat{f}_N^s(X^s) - f_0(X^s)\|_2^2 \\ & = \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - K^{\text{reg}}(f_0)] \\ & \leq \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - K^{\text{reg}}(f_0)] + 2\mathbb{E}[K_N^{\text{reg}}(f) - K_N^{\text{reg}}(\hat{f}_N^s)] \\ & = \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - K^{\text{reg}}(f_0)] + 2\mathbb{E}[K_N^{\text{reg}}(f) - K_N^{\text{reg}}(f_0) + K_N^{\text{reg}}(f_0) - K_N^{\text{reg}}(\hat{f}_N^s)] \\ & = \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] + 2[K^{\text{reg}}(f) - K^{\text{reg}}(f_0)] \\ & = \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] + 2\mathbb{E}_P\|f(X) - f_0(X)\|_2^2. \end{aligned}$$

This indicates that $\mathcal{R}^s(\hat{f}_N^s) \leq \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] + 2 \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P\|f(X) - f_0(X)\|_2^2$. \square

LEMMA B.8. *Assume that $\|Y^s\|_\infty$ is sub-exponentially distributed. Let $\bar{\delta} = \bar{\delta}_N = (\log N)^{1+\kappa}$, with an arbitrarily fixed $\kappa \in (0, 1]$, and let $\underline{\delta} = \underline{\delta}_N = -(\log N)^{1+\kappa}$. Then, for sufficiently large N and $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, it follows that*

$$\mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] \leq \frac{c^* S L \log S (\log N)^{5+4\kappa}}{N},$$

where c^* is a constant not depending on S, L and N .

PROOF OF LEMMA B.8. Let $T_i = (X_i^s, Y_i^s)$ for $i = 1, \dots, N$, $\mathcal{D}_N = \{T_1, \dots, T_N\}$, and $T = (X^s, Y^s)$ be an independent copy of T_1 . Firstly, we have

$$\begin{aligned} & \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] \\ & = \mathbb{E}_{\mathcal{D}_N}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] \\ & = \mathbb{E}_{\mathcal{D}_N}\{K^{\text{reg}}(\hat{f}_N^s) - K^{\text{reg}}(f_0) - 2[K_N^{\text{reg}}(\hat{f}_N^s) - K_N^{\text{reg}}(f_0)]\} \\ & = \mathbb{E}_{\mathcal{D}_N}\left\{\mathbb{E}_T[g(\hat{f}_N^s, T)] - \frac{2}{N} \sum_{i=1}^N g(\hat{f}_N^s, T_i)\right\}, \end{aligned}$$

where $g(f, T) = \|Y^s - f(X^s)\|_2^2 - \|Y^s - f_0(X^s)\|_2^2$ for $f \in \mathcal{F}_{\text{NN}}^{d_y}$. For a d_y -dimensional vector v , denote its j -th component as $v_{(j)}$; additionally, we denote the j -th output coordinate of a function $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$ as $f_{(j)}$, with $j \in \{1, \dots, d_y\}$. Furthermore, for any measurable function $h : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$, define

$$g_j(h, T) = [Y_{(j)}^s - h(X^s)]^2 - [Y_{(j)}^s - f_0(X^s)_{(j)}]^2, \quad \text{for } j = 1, \dots, d_y.$$

It is then clear that $g(f, T) = \sum_{j=1}^{d_y} g_j(f_{(j)}, T)$. Hence, we obtain

$$\begin{aligned} & \mathbb{E}[K^{\text{reg}}(\hat{f}_N^s) - 2K_N^{\text{reg}}(\hat{f}_N^s) + K^{\text{reg}}(f_0)] \\ &= \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g(\hat{f}_N^s, T)] - \frac{2}{N} \sum_{i=1}^N g(\hat{f}_N^s, T_i) \right\} \\ &= \sum_{j=1}^{d_y} \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{f}_{N,(j)}^s, T_i) \right\}. \end{aligned}$$

Subsequently, let us fix an arbitrary $j \in \{1, \dots, d_y\}$. For any $\iota_N > 0$, we let $U = Y_{(j)}^s \mathbf{1}(\|Y^s\|_\infty \leq \iota_N)$ and $V = \mathbb{E}[Y_{(j)}^s \mathbf{1}(\|Y^s\|_\infty \leq \iota_N) | X^s]$. Then, for any measurable function $h : \mathbb{R}^{d_x} \rightarrow \mathbb{R}$, define

$$g_{j,\iota_N}(h, T) = [U - h(X^s)]^2 - (U - V)^2 = [V - h(X^s)][2U - h(X^s) - V].$$

It follows that

$$\begin{aligned} & |g_j(f_{(j)}, T) - g_{j,\iota_N}(f_{(j)}, T)| \\ &= \left| [Y_{(j)}^s - f(X^s)_{(j)}]^2 - [U - f(X^s)_{(j)}]^2 - [Y_{(j)}^s - f_0(X^s)_{(j)}]^2 + (U - V)^2 \right| \\ &\leq \left| (Y_{(j)}^s - U)[Y_{(j)}^s + U - 2f(X^s)_{(j)}] \right| \\ &\quad + \left| [Y_{(j)}^s - U - f_0(X^s)_{(j)} + V][Y_{(j)}^s + U - f_0(X^s)_{(j)} - V] \right| \\ &\leq \left| Y_{(j)}^s \mathbf{1}(\|Y^s\|_\infty > \iota_N)[Y_{(j)}^s + U - 2f(X^s)_{(j)}] \right| \\ &\quad + \left| Y_{(j)}^s \mathbf{1}(\|Y^s\|_\infty > \iota_N)[Y_{(j)}^s + U - f_0(X^s)_{(j)} - V] \right| \\ &\quad + \left| \mathbb{E}[Y_{(j)}^s \mathbf{1}(\|Y^s\|_\infty > \iota_N) | X^s][Y_{(j)}^s + U - f_0(X^s)_{(j)} - V] \right| \\ &\leq (2\|Y^s\|_\infty + \|f_0(X^s)\|_\infty + 3\iota_N + 2\bar{\delta}_N) \|Y^s\|_\infty \mathbf{1}(\|Y^s\|_\infty > \iota_N) \\ &\quad + (\|Y^s\|_\infty + \|f_0(X^s)\|_\infty + 2\iota_N) \mathbb{E}[\|Y^s\|_\infty \mathbf{1}(\|Y^s\|_\infty > \iota_N) | X^s]. \end{aligned}$$

Taking expectation, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{f}_{N,(j)}^s, T_i) \right\} \\ &\leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \right\} \\ &\quad + 3\mathbb{E}_P [(2\|Y\|_\infty + \|f_0(X)\|_\infty + 3\iota_N + 2\bar{\delta}_N) \|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\ &\quad + 3\mathbb{E}_P \{(\|Y\|_\infty + \|f_0(X)\|_\infty + 2\iota_N) \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N) | X]\}. \end{aligned}$$

Specifically, observe that

$$\begin{aligned}
& \mathbb{E}_P \left[(2\|Y\|_\infty + \|f_0(X)\|_\infty + 3\iota_N + 2\bar{\delta}_N) \|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N) \right] \\
&= 2\mathbb{E}_P [\|Y\|_\infty^2 \mathbf{1}(\|Y\|_\infty > \iota_N)] + \mathbb{E}_P [\|f_0(X)\|_\infty \|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\
&\quad + (3\iota_N + 2\bar{\delta}_N) \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\
&\leq \frac{32}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \exp(-\varsigma\iota_N/2) + \frac{4}{\varsigma} \mathbb{E}_P[\|f_0(X)\|_\infty \exp(\varsigma\|Y\|_\infty/2)] \exp(-\varsigma\iota_N/4) \\
&\quad + \frac{2}{\varsigma} (3\iota_N + 2\bar{\delta}_N) \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \exp(-\varsigma\iota_N/2) \\
&\leq \frac{32}{\varsigma^2} \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \exp(-\varsigma\iota_N/2) \\
&\quad + \frac{4}{\varsigma} \left\{ \mathbb{E}_P(\|f_0(X)\|_\infty^2) \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \right\}^{1/2} \exp(-\varsigma\iota_N/4) \\
&\quad + \frac{2}{\varsigma} (3\iota_N + 2\bar{\delta}_N) \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \exp(-\varsigma\iota_N/2),
\end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}_P \{ (\|Y\|_\infty + \|f_0(X)\|_\infty + 2\iota_N) \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N) | X] \} \\
&= \mathbb{E}_P \{ \|Y\|_\infty \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N) | X] + \mathbb{E}_P [\|f_0(X)\|_\infty \|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\
&\quad + 2\iota_N \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \} \\
&\leq \{ \mathbb{E}_P(\|Y\|_\infty^2) \mathbb{E}_P[\|Y\|_\infty^2 \mathbf{1}(\|Y\|_\infty > \iota_N)] \}^{1/2} + \mathbb{E}_P [\|f_0(X)\|_\infty \|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\
&\quad + 2\iota_N \mathbb{E}_P [\|Y\|_\infty \mathbf{1}(\|Y\|_\infty > \iota_N)] \\
&\leq \frac{4}{\varsigma} \left\{ \mathbb{E}_P(\|Y\|_\infty^2) \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \right\}^{1/2} \exp(-\varsigma\iota_N/4) \\
&\quad + \frac{4}{\varsigma} \left\{ \mathbb{E}_P(\|f_0(X)\|_\infty^2) \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \right\}^{1/2} \exp(-\varsigma\iota_N/4) \\
&\quad + \frac{4}{\varsigma} \iota_N \mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)] \exp(-\varsigma\iota_N/2).
\end{aligned}$$

Here, we have applied the inequalities $a \leq \exp(a)$ and $\mathbf{1}(a > 0) \leq \exp(a)$ for $a \in \mathbb{R}$. Note that $\mathbb{E}_P\|f_0(X)\|_\infty^2 < \infty$ since $\|Y^s\|_\infty$ has a finite second moment and $f_0(X^s) = \mathbb{E}(Y^s | X^s)$. As a consequence, we obtain

$$\begin{aligned}
& \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{f}_{N,(j)}^s, T_i) \right\} \\
&\leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \right\} + c_1(\iota_N + \bar{\delta}_N + 1) \exp(-\varsigma\iota_N/4),
\end{aligned}$$

where c_1 is a constant which depends only on ς and $\mathbb{E}_P[\exp(\varsigma\|Y\|_\infty)]$. Recall that $\bar{\delta}_N = (\log N)^{1+\kappa}$. Set ι_N to $(4\varsigma^{-1} \log N) \vee 1$. Then, for sufficiently large N such that $\bar{\delta}_N \geq \iota_N$, it

holds that

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{f}_{N,(j)}^s, T_i) \right\} \\ & \leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \right\} + 3c_1(\log N)^{1+\kappa} N^{-1}. \end{aligned}$$

Furthermore, we proceed to verify the conditions in Theorem A.6. Notice that

$$\sup_{f \in \mathcal{F}_{\text{NN}}^{d_y}, T \in \mathbb{R}^{d_x+d_y}} |g_{j,\iota_N}(f_{(j)}, T)| \leq 6\iota_N^2 + 2\bar{\delta}_N^2 \leq 8\bar{\delta}_N^2 = 8(\log N)^{2+2\kappa},$$

whenever $\bar{\delta}_N \geq \iota_N$. Furthermore, for any $f \in \mathcal{F}_{\text{NN}}^{d_y}$,

$$\begin{aligned} \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)] &= \mathbb{E}\{[V - f(X^s)_{(j)}][2U - f(X^s)_{(j)} - V]\} \\ &= \mathbb{E}[V - f(X^s)_{(j)}]^2, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[g_{j,\iota_N}(f, T)^2] &= \mathbb{E}\{[V - f(X^s)_{(j)}]^2[2U - f(X^s)_{(j)} - V]^2\} \\ &\leq (3\iota_N + \bar{\delta}_N)^2 \mathbb{E}[V - f(X^s)_{(j)}]^2 \\ &\leq 16\bar{\delta}_N^2 \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)] \\ &= 16(\log N)^{2+2\kappa} \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)], \end{aligned}$$

provided that N is sufficiently large such that $\bar{\delta}_N \geq \iota_N$. Hence, Theorem A.6 suggests that, for sufficiently large N such that $\bar{\delta}_N \geq \iota_N$, with $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, and for arbitrary $t > 0$, we have

$$\begin{aligned} & \mathbb{P}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \geq t \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{1}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] \right\} \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_N} \left(\exists h \in \mathcal{F}_{\text{NN}} : \mathbb{E}[g_{j,\iota_N}(h, T)] - \frac{1}{N} \sum_{i=1}^N g_{j,\iota_N}(h, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{j,\iota_N}(h, T)] \right\} \right) \\ & \leq 14\mathcal{N}_N \left(c_2 t, \|\cdot\|_\infty, \{g_{j,\iota_N}(h, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}, h \in \mathcal{F}_{\text{NN}}\} \right) \exp \left(-\frac{Nt}{c_3(\log N)^{4+4\kappa}} \right), \end{aligned}$$

where c_2, c_3 are universal constants. Subsequently, we bound the covering number. Fix $\{x_1, \dots, x_N\} \subset (\mathcal{R}^{d_x})^N$ and $\{y_1, \dots, y_N\} \subset (\mathcal{R}^{d_y})^N$. Let $\mathcal{C} = \{x_1, \dots, x_N\}$, and let $w^\sharp = \{w_1, \dots, w_k\}$ be an ϵ -covering set of $\mathcal{F}_{\text{NN}|C}$ where $w_i = h_{i|\mathcal{C}}$ for some $h_i \in \mathcal{F}_{\text{NN}}$ ($i = 1, \dots, k$), such that for any $h \in \mathcal{F}_{\text{NN}}$, there exists $w^* = h_{|\mathcal{C}}^* \in w^\sharp$ satisfying $\|w^* - h_{|\mathcal{C}}\|_\infty < \epsilon$. This indicates

$$\begin{aligned} & |g_{j,\iota_N}(h, (x_i, y_i)) - g_{j,\iota_N}(h^*, (x_i, y_i))| \\ & \leq |h^*(x_i) - h(x_i)| \cdot |2y_{i,(j)} \mathbb{1}(\|y_i\|_\infty \leq \iota_N) - h^*(x_i) - h(x_i)| \\ & \leq 2(\iota_N + \bar{\delta}_N)\epsilon \\ & \leq 4\bar{\delta}_N\epsilon, \end{aligned}$$

whenever $\bar{\delta}_N \geq \iota_N$. Therefore,

$$\mathcal{N}_N(c_2 t, \|\cdot\|_\infty, \{g_{j,\iota_N}(h, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}, h \in \mathcal{F}_{\text{NN}}\}) \leq \mathcal{N}_N(c_2 t / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}).$$

Then, with Lemma A.4 and Lemma A.5, for sufficiently large N with $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$ and any $a_N \geq 1/N$, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \right\} \\ & \leq a_N + 14 \int_{a_N}^\infty \mathcal{N}_N(c_2 t / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \exp\left(-\frac{Nt}{c_3(\log N)^{4+4\kappa}}\right) dt \\ & \leq a_N + 14 \mathcal{N}_N(c_2 a_N / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \int_{a_N}^\infty \exp\left(-\frac{Nt}{c_3(\log N)^{4+4\kappa}}\right) dt \\ & \leq a_N + 14 (c_4 N^2 \bar{\delta}_N^2)^{c_5 S L \log S} \cdot \frac{c_3(\log N)^{4+4\kappa}}{N} \exp\left(-\frac{Na_N}{c_3(\log N)^{4+4\kappa}}\right), \end{aligned}$$

where c_4 and c_5 are universal constants. Choose

$$a_N = \frac{c_3 c_5 (\log N)^{4+4\kappa}}{N} S L \log S \log(c_4 N^2 \bar{\delta}_N^2).$$

For sufficiently large N , we have

$$\mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{f}_{N,(j)}^s, T_i) \right\} \leq \frac{c_6 S L \log S (\log N)^{5+4\kappa}}{N},$$

where c_6 is a constant not depending on S, L and N . Noticing the arbitrariness of j , we complete the proof. \square

LEMMA B.9. *Assume that*

- (i) $e_j^\top f_0 \in \mathcal{H}_{\text{Loc}}^{\beta_f}(\mathbb{R}^d, B_u)$ with $\beta_f > 0$ and $B_u \leq c(u^m + 1)$ for some universal constants $c > 0$, $m \geq 0$, and for any $j \in \{1, \dots, d_y\}$, where e_j denotes a d_y -dimensional one-hot vector with the j -th component equal to 1 and all other components equal to 0;
- (ii) $\|f_0(X^s)\|_\infty$ and $\|X^s\|_\infty$ are sub-exponentially distributed random variables.

Suppose that the depth L and width M of $\mathcal{F}_{\text{NN}}^{d_y}$ are expressed as

$$L = 21(\lfloor \beta_f \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d_x + 3,$$

$$M = 38d_y(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil,$$

for any $S_1, S_2 \in \mathbb{N}_+$. Let $\bar{\delta} = \bar{\delta}_N = (\log N)^{1+\kappa}$, with an arbitrarily fixed $\kappa \in (0, 1]$, and let $\underline{\delta} = \underline{\delta}_N = -(\log N)^{-1-\kappa}$. Then, for sufficiently large N , it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P \|f(X) - f_0(X)\|_2^2 \\ & \leq c^* \left\{ \left[(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x} (\log N)^m \right]^2 + \frac{(\log N)^{2+2\kappa}}{N} \right\}, \end{aligned}$$

where c^* is a constant not depending on S_1, S_2 and N .

PROOF OF LEMMA B.9. For any $\iota_N > 0$, observe that

$$\begin{aligned} & \mathbb{E}_P \|f(X) - f_0(X)\|_2^2 \\ &= \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|X\|_\infty \leq \iota_N)\} + \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|X\|_\infty > \iota_N)\}. \end{aligned}$$

For clarity, denote the j -th output coordinate of a function $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$ as $f_{(j)}$, with $j \in \{1, \dots, d_y\}$. On the one hand, it follows that

$$\begin{aligned} & \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|X\|_\infty > \iota_N)\} \\ &= \sum_{j=1}^{d_y} \mathbb{E}_P \{[f_{(j)}(X) - f_{0,(j)}(X)]^2 \mathbf{1}(\|X\|_\infty > \iota_N)\} \\ &\leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\mathbf{1}(\|X\|_\infty > \iota_N)] + 2d_y \mathbb{E}_P [\|f_0(X)\|_\infty^2 \mathbf{1}(\|X\|_\infty > \iota_N)] \\ &\leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_N / 2) \\ &\quad + 2d_y \mathbb{E}_P [\|f_0(X)\|_\infty^2 \exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_N / 2) \\ &\leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\exp(\varsigma \|X\|_\infty / 2)] \exp(-\varsigma \iota_N / 2) \\ &\quad + 2d_y \{\mathbb{E}_P [\|f_0(X)\|_\infty^4] \mathbb{E}_P [\exp(\varsigma \|X\|_\infty)]\}^{1/2} \exp(-\varsigma \iota_N / 2) \\ &\leq c_1 (\bar{\delta}_N^2 + 1) \exp(-\varsigma \iota_N / 2), \end{aligned}$$

where c_1 is a constant which depends only on $d_y, \varsigma, \mathbb{E}_P[\exp(\varsigma \|X\|_\infty)]$ and $\mathbb{E}_P[\|f_0(X)\|_\infty^4]$. On the other hand, we first notice that

$$\begin{aligned} & \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|X\|_\infty \leq \iota_N)\} \\ &= \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbf{1}(\|X\|_\infty \leq \iota_N)\} \\ &\quad + \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|f_0(X)\|_\infty > \iota_N) \mathbf{1}(\|X\|_\infty \leq \iota_N)\} \\ &\leq \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbf{1}(\|X\|_\infty \leq \iota_N)\} \\ &\quad + 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\mathbf{1}(\|f_0(X)\|_\infty > \iota_N)] + 2d_y \mathbb{E}_P [\|f_0(X)\|_\infty^2 \mathbf{1}(\|f_0(X)\|_\infty > \iota_N)] \\ &\leq \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbf{1}(\|X\|_\infty \leq \iota_N)\} \\ &\quad + 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\exp(\varsigma \|f_0(X)\|_\infty / 2)] \exp(-\varsigma \iota_N / 2) \\ &\quad + \frac{32}{\varsigma^2} d_y \mathbb{E}_P [\exp(\varsigma \|f_0(X)\|_\infty)] \exp(-\varsigma \iota_N / 2) \\ &\leq \mathbb{E}_P \{\|f(X) - f_0(X)\|_2^2 \mathbf{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbf{1}(\|X\|_\infty \leq \iota_N)\} + c_2 (\bar{\delta}_N^2 + 1) \exp(-\varsigma \iota_N / 2), \end{aligned}$$

where c_2 is a constant which depends only on d_y, ς and $\mathbb{E}_P[\exp(\varsigma \|f_0(X)\|_\infty)]$. Then, we focus on $\{x : \|x\|_\infty \leq \iota_N\} = [-\iota_N, \iota_N]^{d_x}$. Fix an arbitrary $j \in \{1, \dots, d_y\}$. Let $h_j(x) = f_{0,(j)}(2\iota_N x - \iota_N \mathbf{1}_{d_x})$ for $x \in [0, 1]^{d_x}$. Lemma A.2 demonstrates that for any $S_1, S_2 \in \mathbb{N}_+$, there exists a function h_j^* implemented by a ReLU network with depth $L^* = 21(\lfloor \beta_f \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d_x$, width $M^* = 38(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, such that

$$|h_j^*(x) - h_j(x)| \leq 18c(\iota_N^m + 1)(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x},$$

for all $x \in [0, 1]^{d_x} \setminus \Omega([0, 1]^{d_x}, K, \Delta)$. Here,

$$\Omega([0, 1]^{d_x}, K, \Delta) = \bigcup_{i=1}^{d_x} \left\{ x = (x_1, \dots, x_{d_x})^\top : x_i \in \bigcup_{k=1}^{K-1} (k/K - \Delta, k/K) \right\},$$

where $K = \lceil (S_1 S_2)^{2/d_x} \rceil$ and Δ is an arbitrary scalar in $(0, 1/(3K)]$. Let $h_j^\dagger(x) = h_j^*((x + \iota_N \mathbf{1}_{d_x})/(2\iota_N))$ for $x \in [-\iota_N, \iota_N]^{d_x}$. We obtain that

$$|h_j^\dagger(x) - f_{0,(j)}(x)| \leq 18c(\iota_N^m + 1)(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x},$$

for all $x \in [-\iota_N, \iota_N]^{d_x} \setminus \Omega^\dagger$, where $\Omega^\dagger = \{x : (x + \iota_N \mathbf{1}_{d_x})/(2\iota_N) \in \Omega([0, 1]^{d_x}, K, \Delta)\}$. Furthermore, note that

$$h_j^\dagger(x) = h_j^* \left(\frac{x + \iota_N \mathbf{1}_{d_x}}{2\iota_N} \right) = h_j^* \left(\text{relu} \left(\frac{x + \iota_N \mathbf{1}_{d_x}}{2\iota_N} \right) - \text{relu} \left(-\frac{x + \iota_N \mathbf{1}_{d_x}}{2\iota_N} \right) \right),$$

which is implemented by a neural network with ReLU activations, depth $L^\dagger = L^* + 1$, and width $M^\dagger = M^*$. In addition, let

$$h_j^\ddagger(x) = \begin{cases} \bar{\delta}_N, & h_j^\dagger(x) > \bar{\delta}_N, \\ h_j^\dagger(x), & \underline{\delta}_N \leq h_j^\dagger(x) \leq \bar{\delta}_N, \\ \underline{\delta}_N, & h_j^\dagger(x) < \underline{\delta}_N. \end{cases}$$

A straightforward calculation shows that

$$h_j^\ddagger(x) = \text{relu}(-\text{relu}(-h_j^\dagger(x) + \bar{\delta}_N) + \bar{\delta}_N) - \text{relu}(-\text{relu}(h_j^\dagger(x) - \underline{\delta}_N) - \underline{\delta}_N),$$

indicating that $h_j^\ddagger(x)$ can be implemented by a ReLU network with depth $L^\ddagger = L^* + 3$ and width $M^\ddagger = M^*$. Due to the arbitrariness of Δ , when $\bar{\delta}_N \geq \iota_N$, it follows that

$$\begin{aligned} & \mathbb{E}_P \left\{ [h_j^\ddagger(X) - f_{0,(j)}(X)]^2 \mathbb{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbb{1}(\|X\|_\infty \leq \iota_N) \right\} \\ & \leq \left[18c(\iota_N^m + 1)(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x} \right]^2. \end{aligned}$$

Let $f^\ddagger(x) = (h_1^\ddagger(x), \dots, h_{d_y}^\ddagger(x))^\top$. It is straightforward to verify that f^\ddagger can be implemented by a ReLU network in $\mathcal{F}_{\text{NN}}^{d_y}$ with depth $L = L^* + 3$ and width $M = d_y M^*$. Hence, we have

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P \left\{ \|f(X) - f_0(X)\|_2^2 \mathbb{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbb{1}(\|X\|_\infty \leq \iota_N) \right\} \\ & \leq \mathbb{E}_P \left\{ \|f^\ddagger(X) - f_0(X)\|_2^2 \mathbb{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbb{1}(\|X\|_\infty \leq \iota_N) \right\} \\ & = \sum_{j=1}^{d_y} \mathbb{E}_P \left\{ [h_j^\ddagger(X) - f_{0,(j)}(X)]^2 \mathbb{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbb{1}(\|X\|_\infty \leq \iota_N) \right\} \\ & \leq d_y \left[18c(\iota_N^m + 1)(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x} \right]^2, \end{aligned}$$

provided that $\bar{\delta}_N \geq \iota_N$. Recall that $\bar{\delta}_N = (\log N)^{1+\kappa}$, and set ι_N to $(2\varsigma^{-1} \log N) \vee 1$. We conclude that for sufficiently large N satisfying $\bar{\delta}_N \geq \iota_N$, it holds that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P \|f(X) - f_0(X)\|_2^2 \\ & \leq \inf_{f \in \mathcal{F}_{\text{NN}}^{d_y}} \mathbb{E}_P \left\{ \|f(X) - f_0(X)\|_2^2 \mathbb{1}(\|f_0(X)\|_\infty \leq \iota_N) \mathbb{1}(\|X\|_\infty \leq \iota_N) \right\} \\ & \quad + (c_1 + c_2)(\bar{\delta}_N^2 + 1) \exp(-\varsigma \iota_N/2) \\ & \leq c_3 \left\{ \left[(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x} (\log N)^m \right]^2 + \frac{(\log N)^{2+2\kappa}}{N} \right\}, \end{aligned}$$

where c_3 is a constant not depending on S_1, S_2 and N . \square

PROOF OF THEOREM 5.1. To commence, we notice that $\|f_0(X^s)\|_\infty$ is sub-exponentially distributed provided that $\|Y^s\|_\infty$ is a sub-exponential random variable (consider Jensen's inequality). Then, Lemma B.7, Lemma B.8 and Lemma B.9 indicate

$$\begin{aligned} & \mathbb{E}\|\hat{f}_N^s(X^s) - f_0(X^s)\|_2^2 \\ & \leq \frac{c_1 S L \log S (\log N)^{5+4\kappa}}{N} \\ & \quad + c_2 \left\{ \left[(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + (\beta_f \vee 1)/2} (S_1 S_2)^{-2\beta_f/d_x} (\log N)^m \right]^2 + \frac{(\log N)^{2+2\kappa}}{N} \right\}, \end{aligned}$$

where c_1, c_2 are constants not depending on S, L, S_1, S_2 and N , and S_1, S_2 satisfy the conditions that the network depth $L = 21(\lfloor \beta_f \rfloor + 1)^2 S_1 \lceil \log_2(8S_1) \rceil + 2d_x + 3$, network width $M = 38d_y(\lfloor \beta_f \rfloor + 1)^2 d_x^{\lfloor \beta_f \rfloor + 1} S_2 \lceil \log_2(8S_2) \rceil$, for sufficiently large N and $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$. Therefore, by letting $S_1 = \mathcal{O}(N^{d_x/(2d_x+4\beta_f)})$ and $S_2 = \mathcal{O}(1)$, we obtain

$$M = \mathcal{O}(1), \quad L = \mathcal{O}\left(N^{\frac{d_x}{2d_x+4\beta_f}} \log N\right), \quad S = \mathcal{O}(M^2 L) = \mathcal{O}\left(N^{\frac{d_x}{2d_x+4\beta_f}} \log N\right),$$

yielding

$$\mathbb{E}\|\hat{f}_N^s(X^s) - f_0(X^s)\|_2^2 \leq c_3 N^{-\frac{2\beta_f}{d_x+2\beta_f}} (\log N)^{(8+4\kappa) \vee (2m)},$$

where c_3 is a constant not depending on N , for $N \geq 2$. Furthermore, observe that

$$\mathbb{E}\|\hat{f}_N^s(X^t) - f_0(X^t)\|_2^2 = \mathbb{E}\left\{\|\hat{f}_N^s(X^s) - f_0(X^s)\|_2^2 \cdot r_0(X^s)\right\}.$$

As $\|\hat{f}_N^s(X^s)\|_\infty$ is bounded by $\bar{\delta}_N = (\log N)^{1+\kappa}$, $\|f_0(X^s)\|_\infty$ is a sub-exponential random variable (hence it possesses a finite fourth moment), and $r_0(X^s)$ is presumed to be sub-exponentially distributed, by Corollary 4.2, we conclude that

$$\mathbb{E}\|\hat{f}_N^s(X^t) - f_0(X^t)\|_2^2 \leq c_4 N^{-\frac{2\beta_f}{d_x+2\beta_f}} (\log N)^{(8+4\kappa) \vee (2m)+1} + \frac{c_5 d_y [(\log N)^{2+2\kappa} + 1]}{N},$$

for $N \geq 2$, where c_4, c_5 are constants which do not depend on N . This completes the proof. \square

B.8. Proof of Lemma 5.2. In this subsection, we abbreviate $\mathbb{E}[h(X^s, Y^s)]$ to $\mathbb{E}_P[h(X, Y)]$ for any (X^s, Y^s) -integrable function h , whenever the expectation is taken with respect to (X^s, Y^s) .

PROOF OF LEMMA 5.2. We first observe that

$$\begin{aligned} \mathcal{E}^t &= \int \mathbb{E}\left[W_2^2(\rho_{0,x} \| \hat{\rho}_x^s)\right] q(x) dx \\ &= \int \mathbb{E}\left[W_2^2(\rho_{0,x} \| \hat{\rho}_x^s)\right] p(x) r_0(x) dx \\ &= \mathbb{E}\left[W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s)\right]. \end{aligned}$$

For any $\iota_N > 0$, it follows that

$$\begin{aligned} \mathcal{E}^t &= \mathbb{E}\left[W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s)\right] \\ &= \mathbb{E}\left[W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s) \mathbf{1}(r_0(X^s) \leq \iota_N)\right] \\ &\quad + \mathbb{E}\left[W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s) \mathbf{1}(r_0(X^s) > \iota_N)\right] \\ &\leq \iota_N \mathcal{E}^s + \mathbb{E}\left[W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s) \mathbf{1}(r_0(X^s) > \iota_N)\right]. \end{aligned}$$

Fix any $x \in \mathcal{X}^s$. Consider the following two ODEs

$$\begin{aligned} dZ_\tau &= v_0(x, Z_\tau, \tau) d\tau, \quad Z_0 \sim N(0, I_{d_y}), \\ d\hat{Z}_\tau &= \hat{v}_N^s(x, \hat{Z}_\tau, \tau) d\tau, \quad \hat{Z}_0 \sim N(0, I_{d_y}). \end{aligned}$$

We denote the particles at time $\tau \in [0, 1]$ as $Z_\tau(x, z)$ and $\hat{Z}_\tau(x, z)$ given $Z_0 = z$ and $\hat{Z}_0 = z$, respectively. Then, it follows that

$$\begin{aligned} W_2^2(\rho_{0,x} \| \hat{\rho}_x^s) &\leq \int \left\| Z_1(x, z) - \hat{Z}_1(x, z) \right\|_2^2 \cdot \lambda(z) dz \\ &\leq 2 \int \|Z_1(x, z)\|_2^2 \cdot \lambda(z) dz + 2 \int \left\| \hat{Z}_1(x, z) \right\|_2^2 \cdot \lambda(z) dz \\ &= 2\mathbb{E} (\|Y^s\|_2^2 | X^s = x) + 2 \int \left\| \hat{Z}_1(x, z) \right\|_2^2 \cdot \lambda(z) dz, \end{aligned}$$

where $\lambda(\cdot)$ represents the density function of d_y -dimensional standard Gaussian distribution. Furthermore, define

$$\begin{aligned} H_1(x) &= \mathbb{E} (\|Y^s\|_2^2 | X^s = x), \\ H_2(x, \tau) &= \int \left\| \hat{Z}_\tau(x, z) \right\|_2^2 \cdot \lambda(z) dz, \quad \text{for } \tau \in [0, 1]. \end{aligned}$$

Observe that

$$\begin{aligned} \frac{d}{d\tau} H_2(x, \tau) &= 2 \int \left\langle \hat{v}_N^s(x, \hat{Z}_\tau(x, z), \tau), \hat{Z}_\tau(x, z) \right\rangle \cdot \lambda(z) dz \\ &\leq \int \left\| \hat{v}_N^s(x, \hat{Z}_\tau(x, z), \tau) \right\|_2^2 \cdot \lambda(z) dz + \int \left\| \hat{Z}_\tau(x, z) \right\|_2^2 \cdot \lambda(z) dz \\ &\leq d_y \max(\bar{\delta}^2, \underline{\delta}^2) + H_2(x, \tau). \end{aligned}$$

Let $\delta^* = \max(\bar{\delta}^2, \underline{\delta}^2)$. By Lemma A.7, we obtain

$$\int \left\| \hat{Z}_1(x, z) \right\|_2^2 \cdot \lambda(z) dz = H_2(x, 1) \leq d_y e(\delta^* + 1).$$

Hence, we conclude that $W_2^2(\rho_{0,x} \| \hat{\rho}_x^s) \leq 2H_1(x) + 2d_y e(\delta^* + 1)$ for any $x \in \mathcal{X}^s$. As a consequence, it holds that

$$\begin{aligned} \mathcal{E}^t &\leq \iota_N \mathcal{E}^s + \mathbb{E} [W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s) r_0(X^s) \mathbf{1}(r_0(X^s) > \iota_N)] \\ &\leq \iota_N \mathcal{E}^s + 2\mathbb{E}_P[H_1(X) r_0(X) \mathbf{1}(r_0(X) > \iota_N)] + 2d_y e(\delta^* + 1) \mathbb{E}_P[r_0(X^s) \mathbf{1}(r_0(X^s) > \iota_N)] \\ &\leq \iota_N \mathcal{E}^s + \frac{8}{\varsigma} \mathbb{E}_P[H_1(X) \exp(\varsigma r_0(X)/2)] \exp(-\varsigma \iota_N/4) \\ &\quad + \frac{4d_y e}{\varsigma} (\delta^* + 1) \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_N/2) \\ &\leq \iota_N \mathcal{E}^s + \frac{8}{\varsigma} \left\{ \mathbb{E}_P[H_1(X)^2] \mathbb{E}_P[\exp(\varsigma r_0(X))] \right\}^{1/2} \exp(-\varsigma \iota_N/4) \\ &\quad + \frac{4d_y e}{\varsigma} (\delta^* + 1) \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_N/2). \end{aligned}$$

Here, $H_1(X^s)$ attains a finite second moment since $\|Y^s\|_2$ is presumed to have a finite fourth moment. Set ι_N to $(4\varsigma^{-1} \log N) \vee 1$. We have

$$\begin{aligned} \mathcal{E}^t &\leq_{\iota_N} \mathcal{E}^s + \frac{8}{\varsigma} \left\{ \mathbb{E}_P[H_1(X)^2] \mathbb{E}_P[\exp(\varsigma r_0(X))] \right\}^{1/2} \exp(-\varsigma \iota_N / 4) \\ &\quad + \frac{4d_y e}{\varsigma} (\delta^* + 1) \mathbb{E}_P[\exp(\varsigma r_0(X))] \exp(-\varsigma \iota_N / 2) \\ &\leq [(4\varsigma^{-1} \log N) \vee 1] \mathcal{E}^s + \frac{8}{\varsigma} \left\{ \mathbb{E}_P[H_1(X)^2] \mathbb{E}_P[\exp(\varsigma r_0(X))] \right\}^{1/2} N^{-1} \\ &\quad + \frac{4d_y e}{\varsigma} (\delta^* + 1) \mathbb{E}_P[\exp(\varsigma r_0(X))] N^{-2}. \end{aligned}$$

This completes the proof. \square

B.9. Proof of Theorem 5.4. For any function $f \in \mathcal{L}^2(X^s, Y^s)$, define

$$\begin{aligned} K^{\text{gen}}(f) &= \int_0^1 \mathbb{E}_P \left\| \dot{a}_\tau \eta + \dot{b}_\tau Y - f(X, Y_\tau, \tau) \right\|_2^2 d\tau, \\ K_N^{\text{gen}}(f) &= \frac{1}{N} \sum_{i=1}^N \left\| \dot{a}_{\tau_i} \eta_i + \dot{b}_{\tau_i} Y_i^s - f(X_i^s, Y_{i,\tau_i}^s, \tau_i) \right\|_2^2. \end{aligned}$$

Here, $\mathbb{E}_P[h(X, Y, \eta, Y_\tau)] \equiv \mathbb{E}[h(X^s, Y^s, \eta, Y_\tau^s)]$ for any (X^s, Y^s, η) -integrable function h , where the expectation is taken with respect to (X^s, Y^s, η) for any nonrandom $\tau \in [0, 1]$.

LEMMA B.10. Assume that $\|Y^s\|_\infty$ attains a finite second moment. Then,

$$\begin{aligned} &\int_0^1 \mathbb{E} \|\hat{v}_N^s(X^s, Y_\tau^s, \tau) - v_0(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \\ &\leq \mathbb{E} [K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] \\ &\quad + 2 \inf_{f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}} \int_0^1 \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 d\tau. \end{aligned}$$

PROOF OF LEMMA B.10. Recall that $Y_\tau^s = a_\tau \eta + b_\tau Y^s$ and $v_0(X^s, Y_\tau^s, \tau) = \mathbb{E}(\dot{a}_\tau \eta + \dot{b}_\tau Y^s | Y_\tau^s, X^s)$. For any $\tau \in [0, 1]$, we have

$$\mathbb{E} \|v_0(X^s, Y_\tau^s, \tau)\|_2^2 \leq \mathbb{E}_P \|\dot{a}_\tau \eta + \dot{b}_\tau Y^s\|_2^2 \leq 2\dot{a}_\tau^2 \mathbb{E} \|\eta\|_2^2 + 2\dot{b}_\tau^2 \mathbb{E}_P \|Y^s\|_2^2.$$

Given that $\|Y^s\|_\infty$ has a finite second moment, we obtain $\mathbb{E}_P \|Y^s\|_2^2 < \infty$ and hence $v_0 \in \mathcal{L}^2(X^s, Y^s)$. Furthermore, as a_τ and b_τ are presumed to be continuously differentiable, notice that

$$\int_0^1 \mathbb{E} \|v_0(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \leq \int_0^1 \left(2\dot{a}_\tau^2 \mathbb{E} \|\eta\|_2^2 + 2\dot{b}_\tau^2 \mathbb{E}_P \|Y^s\|_2^2 \right) d\tau < \infty.$$

Next, for any $f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}$, it follows that

$$\begin{aligned}
& \int_0^1 \mathbb{E} \|\hat{v}_N^s(X^s, Y_\tau^s, \tau) - v_0(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \\
&= \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - K^{\text{gen}}(v_0)] \\
&\leq \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - K^{\text{gen}}(v_0)] + 2\mathbb{E}[K_N^{\text{gen}}(f) - K_N^{\text{gen}}(\hat{v}_N^s)] \\
&= \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - K^{\text{gen}}(v_0)] + 2\mathbb{E}[K_N^{\text{gen}}(f) - K_N^{\text{gen}}(v_0) + K_N^{\text{gen}}(v_0) - K_N^{\text{gen}}(\hat{v}_N^s)] \\
&= \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] + 2[K^{\text{gen}}(f) - K^{\text{gen}}(v_0)] \\
&= \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] + 2 \int_0^1 \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 d\tau.
\end{aligned}$$

Taking infimum with respect to $f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}$ on the both sides, we obtain the result. \square

LEMMA B.11. *Assume that $\|Y^s\|_\infty$ is a sub-Gaussian random variable. Let $\bar{\delta} = \bar{\delta}_N = (\log N)^{(1+\kappa)/2}$, with an arbitrarily fixed $\kappa \in (0, 1)$, and let $\underline{\delta} = \underline{\delta}_N = -(\log N)^{(1+\kappa)/2}$. Then, for sufficiently large N and $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$, it follows that*

$$\mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] \leq \frac{c^* S L \log S (\log N)^{3+2\kappa}}{N},$$

where c^* is a constant not depending on S, L and N .

PROOF OF LEMMA B.11. Let $T_i = (X_i^s, Y_i^s, \eta_i, \tau_i)$ for $i = 1, \dots, N$, $\mathcal{D}_N = \{T_1, \dots, T_N\}$, and $T = (X^s, Y^s, \eta, \tau)$ be an independent copy of T_1 . Firstly, we have

$$\begin{aligned}
& \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] \\
&= \mathbb{E}_{\mathcal{D}_N}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] \\
&= \mathbb{E}_{\mathcal{D}_N}\{K^{\text{gen}}(\hat{v}_N^s) - K^{\text{gen}}(v_0) - 2[K_N^{\text{gen}}(\hat{v}_N^s) - K_N^{\text{gen}}(v_0)]\} \\
&= \mathbb{E}_{\mathcal{D}_N}\left\{\mathbb{E}_T[g(\hat{v}_N^s, T)] - \frac{2}{N} \sum_{i=1}^N g(\hat{v}_N^s, T_i)\right\},
\end{aligned}$$

where $g(f, T) = \|\dot{a}_\tau \eta + \dot{b}_\tau Y^s - f(X^s, Y_\tau^s, \tau)\|_2^2 - \|\dot{a}_\tau \eta + \dot{b}_\tau Y^s - v_0(X^s, Y_\tau^s, \tau)\|_2^2$ for $f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}$. For a d_y -dimensional vector v , denote its j -th component as $v_{(j)}$; additionally, we denote the j -th output coordinate of a function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [0, 1] \rightarrow \mathbb{R}^{d_y}$ as $f_{(j)}$, with $j \in \{1, \dots, d_y\}$. Furthermore, for any measurable function $h : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [0, 1] \rightarrow \mathbb{R}$, define

$$g_j(h, T) = [\dot{a}_\tau \eta_{(j)} + \dot{b}_\tau Y_{(j)}^s - h(X^s, Y_\tau^s, \tau)]^2 - [\dot{a}_\tau \eta_{(j)} + \dot{b}_\tau Y_{(j)}^s - v_0(X^s, Y_\tau^s, \tau)_{(j)}]^2,$$

for $j = 1, \dots, d_y$. It is then clear that $g(f, T) = \sum_{j=1}^{d_y} g_j(f_{(j)}, T)$. Hence, we obtain

$$\begin{aligned}
& \mathbb{E}[K^{\text{gen}}(\hat{v}_N^s) - 2K_N^{\text{gen}}(\hat{v}_N^s) + K^{\text{gen}}(v_0)] \\
&= \mathbb{E}_{\mathcal{D}_N}\left\{\mathbb{E}_T[g(\hat{v}_N^s, T)] - \frac{2}{N} \sum_{i=1}^N g(\hat{v}_N^s, T_i)\right\} \\
&= \sum_{j=1}^{d_y} \mathbb{E}_{\mathcal{D}_N}\left\{\mathbb{E}_T[g_j(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{v}_{N,(j)}^s, T_i)\right\}.
\end{aligned}$$

Subsequently, let us fix an arbitrary $j \in \{1, \dots, d_y\}$. For any $\iota_N \geq 1$, we let

$$\begin{aligned} U_0 &= \dot{a}_\tau \eta_{(j)} + \dot{b}_\tau Y_{(j)}^s, \\ U_1 &= (\dot{a}_\tau \eta_{(j)} + \dot{b}_\tau Y_{(j)}^s) \mathbb{1}(\|\eta\|_\infty + \|Y^s\|_\infty \leq \iota_N), \\ U_2 &= \|\eta\|_\infty + \|Y^s\|_\infty, \\ V_0 &= v_0(X^s, Y_\tau^s, \tau)_{(j)}, \\ V_1 &= \mathbb{E} \left[(\dot{a}_\tau \eta_{(j)} + \dot{b}_\tau Y_{(j)}^s) \mathbb{1}(\|\eta\|_\infty + \|Y^s\|_\infty \leq \iota_N) \middle| Y_\tau^s, X^s, \tau \right]. \end{aligned}$$

Then, for any measurable function $h : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [0, 1] \rightarrow \mathbb{R}$, define

$$\begin{aligned} g_{j, \iota_N}(h, T) &= [U_1 - h(X^s, Y_\tau^s, \tau)]^2 - (U_1 - V_1)^2 \\ &= [V_1 - h(X^s, Y_\tau^s, \tau)][2U_1 - h(X^s, Y_\tau^s, \tau) - V_1]. \end{aligned}$$

Since η is a standard Gaussian random vector, $\|\eta\|_\infty^2$ follows a sub-exponential distribution. To see this, note that $\|\eta\|_\infty^2 = \max(\eta_{(1)}^2, \dots, \eta_{(d_y)}^2)$ and $\eta_{(j)}^2$ follows a Gamma distribution with a shape parameter 1/2 and a rate parameter 1/2 for all $j \in \{1, \dots, d_y\}$. Hence, for any $\omega \in (0, 1/2)$, we have $\mathbb{E}[\exp(\omega \eta_{(j)}^2)] = (1 - 2\omega)^{-1/2} < \infty$ for all $j \in \{1, \dots, d_y\}$. This indicates that

$$\begin{aligned} \mathbb{E}[\exp(\omega d_y^{-1} \|\eta\|_\infty^2)] &\leq \mathbb{E} \left[\exp \left(\frac{1}{d_y} \sum_{j=1}^{d_y} \omega \eta_{(j)}^2 \right) \right] \\ &\leq \frac{1}{d_y} \sum_{j=1}^{d_y} \mathbb{E}[\exp(\omega \eta_{(j)}^2)] = (1 - 2\omega)^{-1/2} < \infty. \end{aligned}$$

On the other hand, we have assumed that $\|Y^s\|_\infty$ is a sub-Gaussian random variable. Consequently, there exists a constant ς (which possibly associates with d_y) such that $\mathbb{E}[\exp(\varsigma \|\eta\|_\infty^2)] < \infty$ and $\mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2)] < \infty$. Hence, observe that

$$\mathbb{E}[\exp(\varsigma U_2^2/2)] \leq \mathbb{E}_P[\exp(\varsigma \|\eta\|_\infty^2 + \varsigma \|Y\|_\infty^2)] = \mathbb{E}[\exp(\varsigma \|\eta\|_\infty^2)] \mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2)] < \infty.$$

Additionally, as a_τ and b_τ are presumed to be continuously differentiable over $[0, 1]$, we let $\gamma = \max_{\xi \in [0, 1]} \max(|\dot{a}_\xi|, |\dot{b}_\xi|)$. Then, it follows that

$$\begin{aligned} &|g_j(f_{(j)}, T) - g_{j, \iota_N}(f_{(j)}, T)| \\ &\leq |[U_0 - f(X^s, Y_\tau^s, \tau)_{(j)}]^2 - [U_1 - f(X^s, Y_\tau^s, \tau)_{(j)}]^2| + |(U_0 - V_0)^2 - (U_1 - V_1)^2| \\ &\leq |(U_1 - U_0)[U_0 + U_1 - 2f(X^s, Y_\tau^s, \tau)_{(j)}]| \\ &\quad + |(U_0 - U_1 + V_1 - V_0)(U_0 + U_1 - V_0 - V_1)| \\ &\leq \gamma U_2 \mathbb{1}(U_2 > \iota_N) (\gamma U_2 + \gamma \iota_N + 2\bar{\delta}_N) \\ &\quad + \gamma U_2 \mathbb{1}(U_2 > \iota_N) [\gamma U_2 + 2\gamma \iota_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \\ &\quad + \gamma \mathbb{E}[U_2 \mathbb{1}(U_2 > \iota_N) | Y_\tau^s, X^s, \tau] [\gamma U_2 + 2\gamma \iota_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \\ &= \gamma U_2 [2\gamma U_2 + 3\gamma \iota_N + 2\bar{\delta}_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \mathbb{1}(U_2 > \iota_N) \\ &\quad + \gamma \mathbb{E}[U_2 \mathbb{1}(U_2 > \iota_N) | Y_\tau^s, X^s, \tau] [\gamma U_2 + 2\gamma \iota_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty]. \end{aligned}$$

Taking expectation, we have

$$\begin{aligned}
& \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{v}_{N,(j)}^s, T_i) \right\} \\
& \leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T_i) \right\} \\
& \quad + 3\gamma \mathbb{E} \{ U_2 [2\gamma U_2 + 3\gamma\iota_N + 2\bar{\delta}_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \mathbf{1}(U_2 > \iota_N) \} \\
& \quad + 3\gamma \mathbb{E} \{ \mathbb{E}[U_2 \mathbf{1}(U_2 > \iota_N) | Y_\tau^s, X^s, \tau] [\gamma U_2 + 2\gamma\iota_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \}.
\end{aligned}$$

Specifically, observe that

$$\begin{aligned}
& \mathbb{E} \{ U_2 [2\gamma U_2 + 3\gamma\iota_N + 2\bar{\delta}_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \mathbf{1}(U_2 > \iota_N) \} \\
& \leq \frac{8\gamma}{\varsigma} \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) + \frac{4}{\varsigma} (3\gamma\iota_N + 2\bar{\delta}_N) \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) \\
& \quad + \frac{8}{\varsigma} \mathbb{E}[\|v_0(X^s, Y_\tau^s, \tau)\|_\infty \exp(\varsigma U_2^2/4)] \exp(-\varsigma\iota_N^2/8) \\
& \leq \frac{8\gamma}{\varsigma} \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) + \frac{4}{\varsigma} (3\gamma\iota_N + 2\bar{\delta}_N) \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) \\
& \quad + \frac{8}{\varsigma} \{ \mathbb{E}[\|v_0(X^s, Y_\tau^s, \tau)\|_\infty^2] \mathbb{E}[\exp(\varsigma U_2^2/2)] \}^{1/2} \exp(-\varsigma\iota_N^2/8) \\
& \leq \frac{8\gamma}{\varsigma} \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) + \frac{4}{\varsigma} (3\gamma\iota_N + 2\bar{\delta}_N) \mathbb{E}[\exp(\varsigma U_2^2/2)] \exp(-\varsigma\iota_N^2/4) \\
& \quad + \frac{8\gamma}{\varsigma} \{ \mathbb{E}(U_2^2) \mathbb{E}[\exp(\varsigma U_2^2/2)] \}^{1/2} \exp(-\varsigma\iota_N^2/8),
\end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E} \{ \mathbb{E}[U_2 \mathbf{1}(U_2 > \iota_N) | Y_\tau^s, X^s, \tau] [\gamma U_2 + 2\gamma\iota_N + \|v_0(X^s, Y_\tau^s, \tau)\|_\infty] \} \\
& \leq \gamma \{ \mathbb{E}(U_2^2) \mathbb{E}[U_2^2 \mathbf{1}(U_2 > \iota_N)] \}^{1/2} + 2\gamma\iota_N \mathbb{E}[U_2 \mathbf{1}(U_2 > \iota_N)] \\
& \quad + \mathbb{E}[\|v_0(X^s, Y_\tau^s, \tau)\|_\infty U_2 \mathbf{1}(U_2 > \iota_N)] \\
& \leq \frac{2\gamma}{\varsigma^{1/2}} \{ \mathbb{E}(U_2^2) \mathbb{E}[\exp(\varsigma U_2^2/2)] \}^{1/2} \exp(-\varsigma\iota_N^2/8) + \frac{8\gamma}{\varsigma} \iota_N \mathbb{E}[\exp(\varsigma U_2^2/2) \exp(-\varsigma\iota_N^2/4)] \\
& \quad + \frac{8}{\varsigma} \mathbb{E}[\|v_0(X^s, Y_\tau^s, \tau)\|_\infty \exp(\varsigma U_2^2/4)] \exp(-\varsigma\iota_N^2/8) \\
& \leq \frac{2\gamma}{\varsigma^{1/2}} \{ \mathbb{E}(U_2^2) \mathbb{E}[\exp(\varsigma U_2^2/2)] \}^{1/2} \exp(-\varsigma\iota_N^2/8) + \frac{8\gamma}{\varsigma} \iota_N \mathbb{E}[\exp(\varsigma U_2^2/2) \exp(-\varsigma\iota_N^2/4)] \\
& \quad + \frac{8\gamma}{\varsigma} \{ \mathbb{E}(U_2^2) \mathbb{E}[\exp(\varsigma U_2^2/2)] \}^{1/2} \exp(-\varsigma\iota_N^2/8).
\end{aligned}$$

Here, we have applied the inequalities $a \leq \exp(a)$ and $\mathbf{1}(a > 0) \leq \exp(a)$ for $a \in \mathbb{R}$. As a consequence, we obtain

$$\begin{aligned}
& \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{v}_{N,(j)}^s, T_i) \right\} \\
& \leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T_i) \right\} + c_1(\iota_N + \bar{\delta}_N + 1) \exp(-\varsigma\iota_N^2/8),
\end{aligned}$$

where c_1 is a constant which depends only on γ, ς and $\mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2)]$. Recall that $\bar{\delta}_N = (\log N)^{(1+\kappa)/2}$. Set ι_N to $(8\varsigma^{-1} \log N)^{1/2} \vee 1$. Then, for sufficiently large N such that $\bar{\delta}_N \geq \iota_N$, it holds that

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_j(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_j(\hat{v}_{N,(j)}^s, T_i) \right\} \\ & \leq \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T_i) \right\} + 3c_1(\log N)^{(1+\kappa)/2} N^{-1}. \end{aligned}$$

Furthermore, we proceed to verify the conditions in Theorem A.6. Notice that

$$\sup_{f \in \mathcal{F}_{NN,\Lambda}^{d_y}, T \in \mathbb{R}^{d_x+2d_y} \times [0,1]} |g_{j,\iota_N}(f_{(j)}, T)| \leq 6\gamma^2 \iota_N^2 + 2\bar{\delta}_N^2 \leq 8\bar{\delta}_N^2 = 8(\log N)^{1+\kappa},$$

whenever $\bar{\delta}_N \geq (\gamma \vee 1)\iota_N$. Furthermore, for any $f \in \mathcal{F}_{NN,\Lambda}^{d_y}$,

$$\begin{aligned} \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)] &= \mathbb{E}\{[V_1 - f(X^s, Y_\tau^s, \tau)_{(j)}][2U_1 - f(X^s, Y_\tau^s, \tau)_{(j)} - V_1]\} \\ &= \mathbb{E}[V_1 - f(X^s, Y_\tau^s, \tau)_{(j)}]^2, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[g_{j,\iota_N}(f, T)^2] &= \mathbb{E}\{[V_1 - f(X^s, Y_\tau^s, \tau)_{(j)}]^2[2U_1 - f(X^s, Y_\tau^s, \tau)_{(j)} - V_1]^2\} \\ &\leq (3\gamma\iota_N + \bar{\delta}_N)^2 \mathbb{E}[V_1 - f(X^s, Y_\tau^s, \tau)_{(j)}]^2 \\ &\leq 16\bar{\delta}_N^2 \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)] \\ &= 16(\log N)^{1+\kappa} \mathbb{E}[g_{j,\iota_N}(f_{(j)}, T)], \end{aligned}$$

provided that N is sufficiently large such that $\bar{\delta}_N \geq (\gamma \vee 1)\iota_N$. Hence, Theorem A.6 suggests that, for sufficiently large N such that $\bar{\delta}_N \geq (\gamma \vee 1)\iota_N$, with $N \geq \text{Pdim}(\mathcal{F}_{NN})$, and for arbitrary $t > 0$, we have

$$\begin{aligned} & \mathbb{P}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T_i) \geq t \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] - \frac{1}{N} \sum_{i=1}^N g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{j,\iota_N}(\hat{v}_{N,(j)}^s, T)] \right\} \right\} \\ & \leq \mathbb{P}_{\mathcal{D}_N} \left(\exists h \in \mathcal{F}_{NN} : \mathbb{E}[g_{j,\iota_N}(h, T)] - \frac{1}{N} \sum_{i=1}^N g_{j,\iota_N}(h, T_i) \geq \frac{1}{2} \left\{ \frac{t}{2} + \frac{t}{2} + \mathbb{E}_T[g_{j,\iota_N}(h, T)] \right\} \right) \\ & \leq 14\mathcal{N}_N \left(c_2 t, \|\cdot\|_\infty, \{g_{j,\iota_N}(h, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \times [0,1] \rightarrow \mathbb{R}, h \in \mathcal{F}_{NN} \} \right) \exp \left(-\frac{Nt}{c_3(\log N)^{2+2\kappa}} \right), \end{aligned}$$

where c_2, c_3 are universal constants. Subsequently, we bound the covering number. Fix $\{x_1, \dots, x_N\} \subset (\mathcal{R}^{d_x})^N$, $\{y_1, \dots, y_N\} \subset (\mathcal{R}^{d_y})^N$, $\{z_1, \dots, z_N\} \subset (\mathcal{R}^{d_y})^N$ and $\{\xi_1, \dots, \xi_N\} \in [0,1]^N$. Let $\mathcal{C} = \{(x_1, \zeta_1, \xi_1), \dots, (x_N, \zeta_N, \xi_N)\}$ where $\zeta_i = a_{\xi_i} z_i + b_{\xi_i} y_i$ for $i = 1, \dots, N$. Let $w^\sharp = \{w_1, \dots, w_k\}$ be an ϵ -covering set of $\mathcal{F}_{NN|\mathcal{C}}$ where $w_i = h_{i|\mathcal{C}}$ for some $h_i \in \mathcal{F}_{NN}$ ($i = 1, \dots, k$), such that for any $h \in \mathcal{F}_{NN}$, there exists $w^* = h_{|\mathcal{C}}^* \in w^\sharp$ satisfying

$\|w^* - h|_{\mathcal{C}}\|_\infty < \epsilon$. This indicates

$$\begin{aligned} & |g_{j,\ell_N}(h, (x_i, y_i, z_i, \xi_i)) - g_{j,\ell_N}(h^*, (x_i, y_i, z_i, \xi_i))| \\ & \leq |h^*(x_i, \zeta_i, \xi_i) - h(x_i, \zeta_i, \xi_i)| \cdot (2\gamma\ell_N + 2\bar{\delta}_N) \\ & \leq 2(\gamma\ell_N + \bar{\delta}_N)\epsilon \\ & \leq 4\bar{\delta}_N\epsilon, \end{aligned}$$

whenever $\bar{\delta}_N \geq (\gamma \vee 1)\ell_N$. Therefore,

$$\begin{aligned} & \mathcal{N}_N(c_2 t, \|\cdot\|_\infty, \{g_{j,\ell_N}(h, \cdot) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \times [0, 1] \rightarrow \mathbb{R}, h \in \mathcal{F}_{\text{NN}}\}) \\ & \leq \mathcal{N}_N(c_2 t / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}). \end{aligned}$$

Then, with Lemma A.4 and Lemma A.5, for sufficiently large N with $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$ and any $a_N \geq 1/N$, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\ell_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\ell_N}(\hat{v}_{N,(j)}^s, T_i) \right\} \\ & \leq a_N + 14 \int_{a_N}^\infty \mathcal{N}_N(c_2 t / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \exp\left(-\frac{Nt}{c_3(\log N)^{2+2\kappa}}\right) dt \\ & \leq a_N + 14 \mathcal{N}_N(c_2 a_N / (4\bar{\delta}_N), \|\cdot\|_\infty, \mathcal{F}_{\text{NN}}) \int_{a_N}^\infty \exp\left(-\frac{Nt}{c_3(\log N)^{2+2\kappa}}\right) dt \\ & \leq a_N + 14 (c_4 N^2 \bar{\delta}_N^2)^{c_5 S L \log S} \cdot \frac{c_3(\log N)^{2+2\kappa}}{N} \exp\left(-\frac{Na_N}{c_3(\log N)^{2+2\kappa}}\right), \end{aligned}$$

where c_4 and c_5 are universal constants. Choose

$$a_N = \frac{c_3 c_5 (\log N)^{2+2\kappa}}{N} S L \log S \log(c_4 N^2 \bar{\delta}_N^2).$$

For sufficiently large N , we have

$$\mathbb{E}_{\mathcal{D}_N} \left\{ \mathbb{E}_T[g_{j,\ell_N}(\hat{v}_{N,(j)}^s, T)] - \frac{2}{N} \sum_{i=1}^N g_{j,\ell_N}(\hat{v}_{N,(j)}^s, T_i) \right\} \leq \frac{c_6 S L \log S (\log N)^{3+2\kappa}}{N},$$

where c_6 is a constant not depending on S, L and N . Noticing the arbitrariness of j , we complete the proof. \square

LEMMA B.12. Assume that

- (i) $e_j^\top v_0 \in \mathcal{W}_{\text{Gen}}^{1,\infty}(\mathbb{R}^{d_x+d_y}, B_u)$ with $B_u \leq c(u^m + 1)$ for some universal constants $c > 0$, $m \in [0, 1]$, and for any $j \in \{1, \dots, d_y\}$, where e_j denotes a d_y -dimensional one-hot vector with the j -th component equal to 1 and all other components equal to 0;
- (ii) $\|Y^s\|_\infty$ and $\|X^s\|_\infty$ follow sub-Gaussian distributions.

Suppose that the depth L and width M of $\mathcal{F}_{\text{NN}, \Lambda}^{d_y}$ satisfy

$$L \leq C_1(d_x + d_y + 1)^2 S_1 \log S_1 + 3,$$

$$M \leq C_2 2^{d_x+d_y+1} d_y (d_x + d_y + 1) S_2 \log S_2,$$

for any $S_1, S_2 \in \mathbb{N}_+$, where C_1 and C_2 are universal constants. Let $\bar{\delta} = \bar{\delta}_N = (\log N)^{(1+\kappa)/2}$, with an arbitrarily fixed $\kappa \in (0, 1)$, and let $\underline{\delta} = \underline{\delta}_N = -(\log N)^{-(1+\kappa)/2}$, $\Lambda = \Lambda_N = (\log N)^{(1+\kappa)/2}$. Then, for sufficiently large N , it follows that

$$\begin{aligned} & \inf_{f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}} \int_0^1 \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 d\tau \\ & \leq c^* \left\{ \left[(S_1 S_2)^{-2/(d_x + d_y + 1)} (\log N)^{m/2} \right]^2 + \frac{(\log N)^{1+\kappa}}{N} \right\}, \end{aligned}$$

where c^* is a constant not depending on S_1, S_2 and N .

PROOF OF LEMMA B.12. Let $\gamma_1 = \max_{\tau \in [0, 1]} \max(|a_\tau|, |b_\tau|) \vee 1$ and

$$\gamma_2 = \max_{\tau \in [0, 1]} \max(|\dot{a}_\tau|, |\dot{b}_\tau|),$$

both of which are well-defined as we presume that a_τ and b_τ are continuously differentiable over $[0, 1]$. For any $\iota_N \geq 1$ and $\tau \in [0, 1]$, observe that

$$\begin{aligned} & \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \\ & = \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty < \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) < \gamma_1^{-1} \iota_N) \right\} \\ & \quad + \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty \geq \iota_N) \right\} \\ & \quad + \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty < \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N) \right\}. \end{aligned}$$

As we assume that $\|Y^s\|_\infty$ and $\|X^s\|_\infty$ follow sub-Gaussian distributions, $\|Y^s\|_\infty^2$ and $\|X^s\|_\infty^2$ are sub-exponentially distributed random variables. This indicates the existence of a constant ς such that $\mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2)] < \infty$, $\mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2)] < \infty$ and $\mathbb{E}[\exp(\varsigma \|\eta\|_\infty^2)] < \infty$. Hence, we have

$$\begin{aligned} \mathbb{E}_P[\exp(\varsigma(\|\eta\|_\infty + \|Y\|_\infty)^2/2)] & \leq \mathbb{E}_P[\exp(\varsigma \|\eta\|_\infty^2 + \varsigma \|Y\|_\infty^2)] \\ & \leq \mathbb{E}[\exp(\varsigma \|\eta\|_\infty^2)] \mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2)]. \end{aligned}$$

For clarity, denote the j -th output coordinate of a function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times [0, 1] \rightarrow \mathbb{R}^{d_y}$ as $f_{(j)}$, with $j \in \{1, \dots, d_y\}$. It then follows that

$$\begin{aligned} & \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty \geq \iota_N) \right\} \\ & = \sum_{j=1}^{d_y} \mathbb{E}_P \left\{ [f_{(j)}(X, Y_\tau, \tau) - v_{0,(j)}(X, Y_\tau, \tau)]^2 \mathbf{1}(\|X\|_\infty \geq \iota_N) \right\} \\ & \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P[\mathbf{1}(\|X\|_\infty \geq \iota_N)] + 2d_y \mathbb{E}_P[\|v_0(X, Y_\tau, \tau)\|_\infty^2 \mathbf{1}(\|X\|_\infty \geq \iota_N)] \\ & \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2/2)] \exp(-\varsigma \iota_N^2/2) \\ & \quad + 2d_y \mathbb{E}_P[\|v_0(X, Y_\tau, \tau)\|_\infty^2 \exp(\varsigma \|X\|_\infty^2/2)] \exp(-\varsigma \iota_N^2/2) \\ & \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2/2)] \exp(-\varsigma \iota_N^2/2) \\ & \quad + 2d_y \left\{ \mathbb{E}_P[\|v_0(X, Y_\tau, \tau)\|_\infty^4] \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2)] \right\}^{1/2} \exp(-\varsigma \iota_N^2/2) \\ & \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2/2)] \exp(-\varsigma \iota_N^2/2) \\ & \quad + 8d_y \gamma_2^2 \left\{ \mathbb{E}_P(\|\eta\|_\infty^4 + \|Y\|_\infty^4) \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2)] \right\}^{1/2} \exp(-\varsigma \iota_N^2/2), \end{aligned}$$

and

$$\begin{aligned}
& \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty < \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N) \right\} \\
& \leq \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N) \right\} \\
& \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N)] \\
& \quad + 2d_y \mathbb{E}_P [\|v_0(X, Y_\tau, \tau)\|_\infty^2 \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N)] \\
& \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\exp(\varsigma(\|\eta\|_\infty + \|Y\|_\infty)^2/4)] \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4) \\
& \quad + 8d_y \gamma_2^2 \left\{ \mathbb{E}_P (\|\eta\|_\infty^4 + \|Y\|_\infty^4) \mathbb{E}_P [\exp(\varsigma(\|\eta\|_\infty + \|Y\|_\infty)^2/2)] \right\}^{1/2} \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4) \\
& \leq 2d_y \bar{\delta}_N^2 \mathbb{E}_P [\exp(\varsigma \|\eta\|_\infty^2/2)] \mathbb{E}_P [\exp(\varsigma \|Y\|_\infty^2/2)] \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4) \\
& \quad + 8d_y \gamma_2^2 \left\{ \mathbb{E}_P (\|\eta\|_\infty^4 + \|Y\|_\infty^4) \mathbb{E}_P [\exp(\varsigma \|\eta\|_\infty^2)] \mathbb{E}_P [\exp(\varsigma \|Y\|_\infty^2)] \right\}^{1/2} \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4).
\end{aligned}$$

Hence, we obtain that

$$\begin{aligned}
& \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty \geq \iota_N) \right\} \\
& \quad + \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty < \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \geq \gamma_1^{-1} \iota_N) \right\} \\
& \leq c_1 (\bar{\delta}_N^2 + 1) \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4),
\end{aligned}$$

where c_1 is a constant which depends only on $d_y, \gamma_2, \varsigma, \mathbb{E}_P[\exp(\varsigma \|X\|_\infty^2)]$ and $\mathbb{E}_P[\exp(\varsigma \|Y\|_\infty^2/2)]$. Therefore,

$$\begin{aligned}
& \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \\
& \leq \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty < \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) < \gamma_1^{-1} \iota_N) \right\} \\
& \quad + c_1 (\bar{\delta}_N^2 + 1) \exp(-\varsigma \gamma_1^{-2} \iota_N^2/4).
\end{aligned}$$

Next, let us focus on the region

$$\{(x, y, \tau) : \|x\|_\infty < \iota_N, \|y\|_\infty < \iota_N, \tau \in (0, 1)\} = (-\iota_N, \iota_N)^{d_x+d_y} \times (0, 1).$$

Fix an arbitrary $j \in \{1, \dots, d_y\}$. Let

$$h_j(x, y, \tau) = v_{0,(j)}(2\iota_N x - \iota_N \mathbf{1}_{d_x}, 2\iota_N y - \iota_N \mathbf{1}_{d_y}, \tau),$$

for $(x, y, \tau) \in (0, 1)^{d_x} \times (0, 1)^{d_y} \times (0, 1)$. By the assumption that each component of v_0 belongs to $\mathcal{W}_{\text{Gen}}^{1,\infty}(\mathbb{R}^{d_x+d_y}, B_u)$ with $B_u \leq c(u^m + 1)$, we have $\|h_j\|_{\mathcal{W}^{1,\infty}((0, 1)^{d_x+d_y+1})} \leq c(\iota_N^m + 1)$, where the constants $c > 0$, $m \in [0, 1]$. Lemma A.3 demonstrates that for any $S_1, S_2 \in \mathbb{N}_+$, there exists a function h_j^* implemented by a ReLU network with depth $L^* \leq (d_x + d_y + 1)^2 S_1 \log S_1$, width $M^* \leq 2^{d_x+d_y+1} (d_x + d_y + 1) S_2 \log S_2$, such that $\|h_j^*\|_{\mathcal{W}^{1,\infty}((0, 1)^{d_x+d_y+1})} \leq c_2(\iota_N^m + 1)$ and

$$|h_j^*(x, y, \tau) - h_j(x, y, \tau)| \leq c_3(\iota_N^m + 1)(S_1 S_2)^{-2/(d_x+d_y+1)},$$

for all $(x, y, \tau) \in (0, 1)^{d_x} \times (0, 1)^{d_y} \times (0, 1)$, where c_2 and c_3 are constants which depend only on d_x and d_y . Let

$$h_j^\dagger(x, y, \tau) = h_j^* \left((x + \iota_N \mathbf{1}_{d_x}) / (2\iota_N), (y + \iota_N \mathbf{1}_{d_y}) / (2\iota_N), \tau \right),$$

for $(x, y, \tau) \in (-\iota_N, \iota_N)^{d_x} \times (-\iota_N, \iota_N)^{d_y} \times (0, 1)$. We obtain that

$$|h_j^\dagger(x, y, \tau) - v_{0,(j)}(x, y, \tau)| \leq c_3(\iota_N^m + 1)(S_1 S_2)^{-2/(d_x+d_y+1)},$$

for all $(x, y, \tau) \in (-\iota_N, \iota_N)^{d_x} \times (-\iota_N, \iota_N)^{d_y} \times (0, 1)$. Furthermore, note that

$$\begin{aligned} h_j^\dagger(x, y, \tau) &= h_j^*(\frac{x + \iota_N 1_{d_x}}{2\iota_N}, \frac{y + \iota_N 1_{d_y}}{2\iota_N}, \tau) \\ &= h_j^*\left(\text{relu}\left(\frac{x + \iota_N 1_{d_x}}{2\iota_N}\right) - \text{relu}\left(-\frac{x + \iota_N 1_{d_x}}{2\iota_N}\right), \right. \\ &\quad \text{relu}\left(\frac{y + \iota_N 1_{d_y}}{2\iota_N}\right) - \text{relu}\left(-\frac{y + \iota_N 1_{d_y}}{2\iota_N}\right), \\ &\quad \left. \text{relu}(\tau) - \text{relu}(-\tau)\right), \end{aligned}$$

which is implemented by a neural network with ReLU activations, depth $L^\dagger = L^* + 1$, width $M^\dagger = M^*$ and Lipschitz constant no more than $c_2(d_x + d_y + 1)(\iota_N^m + 1)$. In addition, let

$$h_j^\ddagger(x, y, \tau) = \begin{cases} \bar{\delta}_N, & h_j^\dagger(x, y, \tau) > \bar{\delta}_N, \\ h_j^\dagger(x, y, \tau), & \underline{\delta}_N \leq h_j^\dagger(x, y, \tau) \leq \bar{\delta}_N, \\ \underline{\delta}_N, & h_j^\dagger(x, y, \tau) < \underline{\delta}_N. \end{cases}$$

A straightforward calculation shows that

$$h_j^\ddagger(x, y, \tau) = \text{relu}(-\text{relu}(-h_j^\dagger(x, y, \tau) + \bar{\delta}_N) + \bar{\delta}_N) - \text{relu}(-\text{relu}(h_j^\dagger(x, y, \tau) - \underline{\delta}_N) - \underline{\delta}_N),$$

indicating that $h_j^\ddagger(x, y, \tau)$ can be implemented by a ReLU network with depth $L^\ddagger = L^* + 3$, width $M^\ddagger = M^*$ and Lipschitz constant no more than $2c_2(d_x + d_y + 1)(\iota_N^m + 1)$. When $\bar{\delta}_N \geq c_2(\iota_N^m + 1)$, it follows that for any $\tau \in (0, 1)$,

$$\begin{aligned} &\mathbb{E}_P \left\{ [h_j^\ddagger(X, Y_\tau, \tau) - v_{0,(j)}(X, Y_\tau, \tau)]^2 \mathbb{1}(\|X\|_\infty < \iota_N) \mathbb{1}((\|\eta\|_\infty + \|Y\|_\infty) < \gamma_1^{-1} \iota_N) \right\} \\ &\leq \left[c_3(\iota_N^m + 1)(S_1 S_2)^{-2/(d_x + d_y + 1)} \right]^2. \end{aligned}$$

Let $f^\ddagger(x, y, \tau) = (h_1^\ddagger(x, y, \tau), \dots, h_{d_y}^\ddagger(x, y, \tau))^\top$. It is straightforward to verify that f^\ddagger can be implemented by a ReLU network in $\mathcal{F}_{\text{NN}, \Lambda}^{d_y}$ with depth $L = L^* + 3$, width $M = d_y M^*$ and Lipschitz constant no more than $2c_2 d_y (d_x + d_y + 1)(\iota_N^m + 1)$. Hence, we have for any $\tau \in (0, 1)$,

$$\begin{aligned} &\mathbb{E}_P \left\{ \|f^\ddagger(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbb{1}(\|X\|_\infty < \iota_N) \mathbb{1}((\|\eta\|_\infty + \|Y\|_\infty) < \gamma_1^{-1} \iota_N) \right\} \\ &= \sum_{j=1}^{d_y} \mathbb{E}_P \left\{ [h_j^\ddagger(X, Y_\tau, \tau) - v_{0,(j)}(X, Y_\tau, \tau)]^2 \mathbb{1}(\|X\|_\infty < \iota_N) \mathbb{1}((\|\eta\|_\infty + \|Y\|_\infty) < \gamma_1^{-1} \iota_N) \right\} \\ &\leq d_y \left[c_3(\iota_N^m + 1)(S_1 S_2)^{-2/(d_x + d_y + 1)} \right]^2, \end{aligned}$$

provided that $\bar{\delta}_N \geq c_2(\iota_N^m + 1)$ and $\Lambda_N \geq 2c_2 d_y (d_x + d_y + 1)(\iota_N^m + 1)$. Recall that $\bar{\delta}_N = (\log N)^{(1+\kappa)/2}$, $\Lambda_N = (\log N)^{(1+\kappa)/2}$ and set ι_N to $[2\zeta^{-1/2} \gamma_1^{-1} (\log N)^{1/2}] \vee 1$. We conclude that for sufficiently large N satisfying $\bar{\delta}_N \geq c_2(\iota_N^m + 1)$, $\Lambda_N \geq 2c_2 d_y (d_x + d_y + 1)(\iota_N^m + 1)$

and $\log N \geq (\varsigma\gamma_1^2/4) \vee 1$, it holds that

$$\begin{aligned}
& \inf_{f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}} \int_0^1 \mathbb{E}_P \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 d\tau \\
& \leq \inf_{f \in \mathcal{F}_{\text{NN}, \Lambda}^{d_y}} \int_0^1 \mathbb{E}_P \left\{ \|f(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty \leq \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \leq \gamma_1^{-1} \iota_N) \right\} d\tau \\
& \quad + c_1 (\bar{\delta}_N^2 + 1) \exp(-\varsigma\gamma_1^{-2} \iota_N^2/4) \\
& \leq \int_0^1 \mathbb{E}_P \left\{ \|f^\dagger(X, Y_\tau, \tau) - v_0(X, Y_\tau, \tau)\|_2^2 \mathbf{1}(\|X\|_\infty \leq \iota_N) \mathbf{1}((\|\eta\|_\infty + \|Y\|_\infty) \leq \gamma_1^{-1} \iota_N) \right\} d\tau \\
& \quad + c_1 (\bar{\delta}_N^2 + 1) \exp(-\varsigma\gamma_1^{-2} \iota_N^2/4) \\
& \leq c_4 \left\{ \left[(S_1 S_2)^{-2/(d_x+d_y+1)} (\log N)^{m/2} \right]^2 + \frac{(\log N)^{1+\kappa}}{N} \right\},
\end{aligned}$$

where c_4 is a constant not depending on S_1, S_2 and N . This completes the proof. \square

PROOF OF THEOREM 4.8. To commence, Lemma B.10, Lemma B.11 and Lemma B.12 indicate

$$\begin{aligned}
& \int_0^1 \mathbb{E} \|\hat{v}_N^s(X^s, Y_\tau^s, \tau) - v_0(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \\
& \leq \frac{c_1 S L \log S (\log N)^{3+2\kappa}}{N} + c_2 \left\{ \left[(S_1 S_2)^{-2/(d_x+d_y+1)} (\log N)^{m/2} \right]^2 + \frac{(\log N)^{1+\kappa}}{N} \right\},
\end{aligned}$$

where c_1, c_2 are constants not depending on S, L, S_1, S_2 and N , and S_1, S_2 satisfy the conditions that the network depth $L \leq c_3(d_x + d_y + 1)^2 S_1 \log S_1 + 3$, network width $M \leq c_4 2^{d_x+d_y+1} d_y (d_x + d_y + 1) S_2 \log S_2$ for some universal constants c_3 and c_4 , when N is sufficiently large and $N \geq \text{Pdim}(\mathcal{F}_{\text{NN}})$. Therefore, by letting $S_1 = \mathcal{O}(N^{(d_x+d_y+1)/(2(d_x+d_y+1)+4)})$ and $S_2 = \mathcal{O}(1)$, we obtain

$$M = \mathcal{O}(1), \quad L = \mathcal{O}\left(N^{\frac{(d_x+d_y+1)}{2(d_x+d_y+1)+4}} \log N\right), \quad S = \mathcal{O}(M^2 L) = \mathcal{O}\left(N^{\frac{(d_x+d_y+1)}{2(d_x+d_y+1)+4}} \log N\right),$$

yielding

$$\int_0^1 \mathbb{E} \|\hat{v}_N^s(X^s, Y_\tau^s, \tau) - v_0(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \leq c_5 N^{-\frac{2}{d_x+d_y+3}} (\log N)^{6+2\kappa},$$

where c_5 is a constant not depending on N , for $N \geq 2$.

Next, we proceed to tackle the conditional density estimation error. Fix any $x \in \mathcal{X}^s$. Consider the following two ODEs

$$\begin{aligned}
dZ_\tau &= v_0(x, Z_\tau, \tau) d\tau, \quad Z_0 \sim N(0, I_{d_y}), \\
d\hat{Z}_\tau &= \hat{v}_N^s(x, \hat{Z}_\tau, \tau) d\tau, \quad \hat{Z}_0 \sim N(0, I_{d_y}).
\end{aligned}$$

We denote the particles at time $\tau \in [0, 1]$ as $Z_\tau(x, z)$ and $\hat{Z}_\tau(x, z)$ given $Z_0 = z$ and $\hat{Z}_0 = z$, respectively. Note that

$$W_2^2(\rho_{0,x} \|\hat{\rho}_x^s) \leq \int \|Z_1(x, z) - \hat{Z}_1(x, z)\|_2^2 \lambda(z) dz,$$

where $\lambda(\cdot)$ represents the density function of the d_y -dimensional standard Gaussian distribution. For $\tau \in [0, 1]$, define

$$H_\tau(x) = \int \|Z_\tau(x, z) - \hat{Z}_\tau(x, z)\|_2^2 \lambda(z) dz.$$

Then, it follows that

$$\begin{aligned} \frac{\partial}{\partial \tau} H_\tau(x) &= 2 \int \left\langle v_0(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, \hat{Z}_\tau(x, z), \tau), Z_\tau(x, z) - \hat{Z}_\tau(x, z) \right\rangle \lambda(z) dz \\ &= 2 \int \left\langle v_0(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, Z_\tau(x, z), \tau), Z_\tau(x, z) - \hat{Z}_\tau(x, z) \right\rangle \lambda(z) dz \\ &\quad + 2 \int \left\langle \hat{v}_N^s(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, \hat{Z}_\tau(x, z), \tau), Z_\tau(x, z) - \hat{Z}_\tau(x, z) \right\rangle \lambda(z) dz. \end{aligned}$$

Specifically, we first observe that

$$\begin{aligned} &2 \int \left\langle v_0(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, Z_\tau(x, z), \tau), Z_\tau(x, z) - \hat{Z}_\tau(x, z) \right\rangle \lambda(z) dz \\ &\leq \int \|v_0(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, Z_\tau(x, z), \tau)\|_2^2 \lambda(z) dz + H_\tau(x) \\ &= \mathbb{E}_{Y_\tau^s | X^s=x} \|v_0(x, Y_\tau^s, \tau) - \hat{v}_N^s(x, Y_\tau^s, \tau)\|_2^2 + H_\tau(x). \end{aligned}$$

In addition, the Lipschitz continuity of \hat{v}_N^s and Cauchy-Schwarz inequality suggest that

$$\begin{aligned} &2 \int \left\langle \hat{v}_N^s(x, Z_\tau(x, z), \tau) - \hat{v}_N^s(x, \hat{Z}_\tau(x, z), \tau), Z_\tau(x, z) - \hat{Z}_\tau(x, z) \right\rangle \lambda(z) dz \\ &\leq 2\Lambda_N H_\tau(x). \end{aligned}$$

Hence, we conclude that

$$\frac{\partial}{\partial \tau} H_\tau(x) \leq (1 + 2\Lambda_N) H_\tau(x) + \mathbb{E}_{Y_\tau^s | X^s=x} \|v_0(x, Y_\tau^s, \tau) - \hat{v}_N^s(x, Y_\tau^s, \tau)\|_2^2.$$

By Lemma A.7, we have

$$W_2^2(\rho_{0,x} \| \hat{\rho}_x^s) = H_1(x) \leq \exp(1 + 2\Lambda_N) \int_0^1 \mathbb{E}_{Y_\tau^s | X^s=x} \|v_0(x, Y_\tau^s, \tau) - \hat{v}_N^s(x, Y_\tau^s, \tau)\|_2^2 d\tau.$$

Therefore, it follows that

$$\begin{aligned} \mathcal{E}^s &= \mathbb{E} [W_2^2(\rho_{0,X^s} \| \hat{\rho}_{X^s}^s)] \leq \exp(1 + 2\Lambda_N) \int_0^1 \mathbb{E} \|v_0(X^s, Y_\tau^s, \tau) - \hat{v}_N^s(X^s, Y_\tau^s, \tau)\|_2^2 d\tau \\ &\leq c_5 N^{-\frac{2}{d_x+d_y+3}} (\log N)^{6+2\kappa} \exp(1 + 2(\log N)^{(1+\kappa)/2}), \end{aligned}$$

for $N \geq 2$. Furthermore, by Lemma 4.6, we conclude that

$$\mathcal{E}^t \leq c_6 N^{-\frac{2}{d_x+d_y+3}} (\log N)^{7+2\kappa} \exp(1 + 2(\log N)^{(1+\kappa)/2}),$$

where c_6 is a constant not depending on N , for $N \geq 2$. This completes the proof. \square

APPENDIX C: PROOF OF THEOREM A.6

LEMMA C.1 (Lemma 11.2 in [19]). *Let V_1, \dots, V_n be independent and identically distributed random variables, $0 \leq V_i \leq B$, $0 < \alpha < 1$, and $\nu > 0$. Then,*

$$(7) \quad \mathbb{P} \left\{ \frac{\left| \frac{1}{n} \sum_{i=1}^n V_i - \mathbb{E}(V_1) \right|}{\nu + \frac{1}{n} \sum_{i=1}^n V_i + \mathbb{E}(V_1)} > \alpha \right\} \leq \mathbb{P} \left\{ \frac{\left| \frac{1}{n} \sum_{i=1}^n V_i - \mathbb{E}(V_1) \right|}{\nu + \mathbb{E}(V_1)} > \alpha \right\} < \frac{B}{4\alpha^2 \nu n}.$$

REMARK 7. If we substitute B , the upper bound of V_i , to B_n varying with n , the right-hand side of Eqn. (7) should be modified to $B_n/(4\alpha^2 \nu n)$ and the proof stays the same.

LEMMA C.2. *Let Z, Z_1, \dots, Z_n be independent and identically distributed random vectors with dimensions of d , and let \mathcal{H} be a set of nonrandom functions $h : \mathbb{R}^d \rightarrow [0, A_n]$, where $A_n > 0$ is a nonrandom sequence. Assume $\alpha > 0$, $0 < \epsilon < 1$. Then, for $n \geq 1$, we have*

$$(8) \quad \mathbb{P} \left\{ \sup_{h \in \mathcal{H}} \frac{\frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)]}{\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \mathbb{E}[h(Z)]} > \epsilon \right\} \leq 4\mathbb{E} \left[\mathcal{N} \left(\frac{\alpha \epsilon}{5}, \|\cdot\|_\infty, \mathcal{H}|_{\mathcal{D}_n} \right) \right] \exp \left(-\frac{3\epsilon^2 \alpha n}{40A_n} \right),$$

where $\mathcal{D}_n = \{Z_1, \dots, Z_n\}$.

REMARK 8. Lemma C.2 is a generalization of Theorem 11.6 in [19].

PROOF OF LEMMA C.2. The proof contains four steps.

Step 1. Substitution for the expectation by an empirical mean. Draw a pseudo-sample $\mathcal{D}'_n = \{Z'_1, \dots, Z'_n\}$ as an independent copy of \mathcal{D}_n . Let $h^* \in \mathcal{G}$ be a function satisfying

$$\frac{1}{n} \sum_{i=1}^n h^*(Z_i) - \mathbb{E}[h^*(Z)] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z_i) + \mathbb{E}[h^*(Z)] \right\},$$

if there exists any such function; otherwise, let h^* be an arbitrary element of \mathcal{H} . We note that the conditions

$$\frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \mathbb{E}[h(Z)] \right\},$$

and

$$\frac{1}{n} \sum_{i=1}^n h(Z'_i) - \mathbb{E}[h(Z)] \leq \frac{\epsilon}{4} \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h(Z'_i) + \mathbb{E}[h(Z)] \right\},$$

imply

$$\frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon\alpha}{4} + \frac{\epsilon}{n} \sum_{i=1}^n h(Z_i) - \frac{\epsilon}{4n} \sum_{i=1}^n h(Z'_i) + \frac{3\epsilon}{4} \mathbb{E}[h(Z)],$$

which is equivalent to

$$\begin{aligned} & \left(1 - \frac{5\epsilon}{8} \right) \left(\frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \\ & > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) + \frac{3\epsilon}{4} \mathbb{E}[h(Z)]. \end{aligned}$$

We further obtain that

$$\frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right),$$

since $0 < 1 - 5\epsilon/8 < 1$ and $\mathbb{E}[h(Z)] \geq 0$. As a result, it then follows that

$$\begin{aligned} & \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \right\} \\ & \geq \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n h^*(Z_i) - \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z_i) + \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) \right) \right\} \\ & \geq \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n h^*(Z_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\}, \right. \\ & \quad \left. \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] \leq \frac{\epsilon}{4} \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\} \right) \\ & = \mathbb{E} \left\{ \mathbb{1} \left(\frac{1}{n} \sum_{i=1}^n h^*(Z_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\} \right) \right. \\ & \quad \times \mathbb{P} \left. \left(\frac{1}{n} \sum_{i=1}^n h^*(Z'_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] \leq \frac{\epsilon}{4} \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\} \right| \mathcal{D}_n \right). \end{aligned}$$

Lemma C.1 yields that

$$\begin{aligned} (9) \quad & \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n h^*(Z'_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] > \frac{\epsilon}{4} \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z'_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\} \middle| \mathcal{D}_n \right) \\ & < \frac{A_n}{4(\epsilon/4)^2 \alpha n} = \frac{4A_n}{\epsilon^2 \alpha n}. \end{aligned}$$

Therefore, for $n > 8A_n/(\epsilon^2 \alpha)$, the probability in Eqn. (9) is no less than $1/2$, and we conclude that

$$\begin{aligned} & \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \right\} \\ & \geq \frac{1}{2} \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n h^*(Z_i) - \mathbb{E}[h^*(Z)|\mathcal{D}_n] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h^*(Z_i) + \mathbb{E}[h^*(Z)|\mathcal{D}_n] \right\} \right) \\ & = \frac{1}{2} \mathbb{P} \left(\exists h : \frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)] > \epsilon \left\{ \alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \mathbb{E}[h(Z)] \right\} \right). \end{aligned}$$

This proves

$$\begin{aligned} & \mathbb{P} \left(\exists h : \frac{\frac{1}{n} \sum_{i=1}^n h(Z_i) - \mathbb{E}[h(Z)]}{\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \mathbb{E}[h(Z)]} > \epsilon \right) \\ & \leq 2 \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \right\}, \end{aligned}$$

when $n > 8A_n/(\epsilon^2\alpha)$. For $n \leq 8A_n/(\epsilon^2\alpha)$, on the other hand, the right-hand side of Eqn. (8) exceeds one, and hence the assertion holds true trivially.

Step 2. Introduction of Rademacher random variables. Let U_1, \dots, U_n be independent Rademacher random variables which are uniformly distributed over $\{-1, 1\}$, meanwhile independent of $\mathcal{D}_n \cup \mathcal{D}'_n$. Importantly, note that \mathcal{D}_n and \mathcal{D}'_n are interchangeable with respect to corresponding components while their joint distribution remains invariant. Therefore, we have

$$\begin{aligned} & \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n h(Z_i) - \frac{1}{n} \sum_{i=1}^n h(Z'_i) > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \right\} \\ &= \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i [h(Z_i) - h(Z'_i)] > \frac{3\epsilon}{8} \left(2\alpha + \frac{1}{n} \sum_{i=1}^n [h(Z_i) + h(Z'_i)] \right) \right\} \\ &\leq \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(Z_i) > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) \right) \right\} \\ &\quad + \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(Z'_i) < -\frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(Z'_i) \right) \right\} \\ &\leq 2\mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(Z_i) > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(Z_i) \right) \right\}. \end{aligned}$$

Here, we note that $-U_i$ is identically distributed as U_i .

Step 3. Conditioning and covering. Given $Z_i = z_i$ for $i = 1, \dots, n$, and consider

$$\mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(z_i) > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \right\}.$$

Let $\delta > 0$ and let \mathcal{C}_δ be a δ -covering set of \mathcal{H} constrained on $\{z_1, \dots, z_n\}$ with respect to the supremum norm. For any $h \in \mathcal{H}$, there exists a vector $h^\sharp = (\bar{h}(z_1), \dots, \bar{h}(z_n))^\top \in \mathcal{C}_\delta$, such that $\max_{i=1, \dots, n} |h(z_i) - \bar{h}(z_i)| < \epsilon$, thereby indicating

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n U_i h(z_i) &= \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) + \frac{1}{n} \sum_{i=1}^n U_i [h(z_i) - \bar{h}(z_i)] \\ &\leq \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) + \max_{i=1, \dots, n} |h(z_i) - \bar{h}(z_i)| \leq \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) + \delta, \end{aligned}$$

and

$$\frac{1}{n} \sum_{i=1}^n h(z_i) \geq \frac{1}{n} \sum_{i=1}^n \bar{h}(z_i) - \frac{1}{n} \sum_{i=1}^n |h(z_i) - \bar{h}(z_i)| \geq \frac{1}{n} \sum_{i=1}^n \bar{h}(z_i) - \delta.$$

As a result, we have

$$\begin{aligned} & \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(z_i) > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \right\} \\ & \leq \mathbb{P} \left\{ \exists h^\sharp \in \mathcal{C}_\delta : \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) + \delta > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n \bar{h}(z_i) - \delta \right) \right\} \\ & \leq |\mathcal{C}_\delta| \max_{h^\sharp \in \mathcal{C}_\delta} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) > \frac{3\epsilon\alpha}{8} - \frac{3\epsilon\delta}{8} - \delta + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i) \right\}. \end{aligned}$$

Specifying $\delta = \epsilon\alpha/5$ deduces that

$$\frac{3\epsilon\alpha}{8} - \frac{3\epsilon\delta}{8} - \delta = \frac{3\epsilon\alpha}{8} - \frac{3\epsilon\alpha}{40} - \frac{\epsilon\alpha}{5} = \frac{\epsilon\alpha}{10}.$$

By choosing $\mathcal{C}_{\epsilon\alpha/5}$ as an $\epsilon\alpha/5$ -covering set of minimal size, we obtain

$$\begin{aligned} & \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{1}{n} \sum_{i=1}^n U_i h(z_i) > \frac{3\epsilon}{8} \left(\alpha + \frac{1}{n} \sum_{i=1}^n h(z_i) \right) \right\} \\ & \leq \mathcal{N} \left(\frac{\epsilon\alpha}{5}, \|\cdot\|_\infty, \mathcal{H}_{|\{z_1, \dots, z_n\}} \right) \max_{h^\sharp \in \mathcal{C}_{\epsilon\alpha/5}} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) > \frac{\epsilon\alpha}{10} + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i) \right\}. \end{aligned}$$

Step 4. Leveraging the Hoeffding's inequality. Note that given fixed $z_1, \dots, z_n, U_1 \bar{h}(z_1), \dots, U_n \bar{h}(z_n)$ are independent random variables with mean zero and absolute bound $\bar{h}(z_1), \dots, \bar{h}(z_n)$ (recall that $\bar{h} \in [0, A_n]$). Therefore, Hoeffding's inequality suggests that

$$\begin{aligned} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) > \frac{\epsilon\alpha}{10} + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i) \right\} & \leq \exp \left\{ -\frac{2n^2 [\frac{\epsilon\alpha}{10} + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i)]^2}{4 \sum_{i=1}^n \bar{h}(z_i)^2} \right\} \\ & \leq \exp \left\{ -\frac{2n^2 [\frac{\epsilon\alpha}{10} + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i)]^2}{4A_n \sum_{i=1}^n \bar{h}(z_i)} \right\} \\ & = \exp \left\{ -\frac{9\epsilon^2}{128A_n} \frac{[\frac{4\alpha}{15} + \sum_{i=1}^n \bar{h}(z_i)]^2}{\sum_{i=1}^n \bar{h}(z_i)} \right\}. \end{aligned}$$

Note that for any $a, y > 0$, $(a + y)^2/y \geq 4a$, implying

$$\frac{[\frac{4\alpha}{15} + \sum_{i=1}^n \bar{h}(z_i)]^2}{\sum_{i=1}^n \bar{h}(z_i)} \geq \frac{16\alpha n}{15}.$$

Hence, it concludes that

$$\begin{aligned} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n U_i \bar{h}(z_i) > \frac{\epsilon\alpha}{10} + \frac{3\epsilon}{8n} \sum_{i=1}^n \bar{h}(z_i) \right\} & \leq \exp \left\{ -\frac{9\epsilon^2}{128A_n} \frac{[\frac{4\alpha}{15} + \sum_{i=1}^n \bar{h}(z_i)]^2}{\sum_{i=1}^n \bar{h}(z_i)} \right\} \\ & \leq \exp \left\{ -\frac{3\epsilon^2 \alpha n}{40A_n} \right\}, \end{aligned}$$

which completes the proof. \square

PROOF OF THEOREM A.6. The proof is composed of six steps.

Step 1. Symmetrization. We commence by replacing $\mathbb{E}[g(f, Z)]$ through an empirical mean deduced by a pseudo-sample $\mathcal{D}'_n = \{Z'_1, \dots, Z'_n\}$ independent of \mathcal{D}_n . Consider a function $f^* \in \mathcal{F}_n$ depending on \mathcal{D}_n , such that

$$\mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f^*, Z) | \mathcal{D}_n],$$

if such a function exists; otherwise, we let f^* be an arbitrary element in \mathcal{F}_n . Then, Chebychev's inequality implies

$$\begin{aligned} & \mathbb{P} \left\{ \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z'_i) > \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \middle| \mathcal{D}_n \right\} \\ & \leq \frac{\text{Var}[g(f^*, Z) | \mathcal{D}_n]}{n \{ \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \}^2} \\ & \leq \frac{\zeta_n \mathbb{E}[g(f^*, Z) | \mathcal{D}_n]}{n \{ \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \}^2} \\ & \leq \frac{\zeta_n}{\epsilon^2(\alpha + \beta)n}, \end{aligned}$$

where the last inequality stems from $x/(a+x)^2 \leq 1/(4a)$ for $x \geq 0$ and $a > 0$. Thus, for $n > 8\zeta_n/[\epsilon^2(\alpha + \beta)]$, we have

$$\mathbb{P} \left\{ \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z'_i) \leq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \middle| \mathcal{D}_n \right\} > \frac{7}{8},$$

yielding that

$$\begin{aligned} & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f, Z)] \right\} \\ & \geq \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n g(f^*, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f^*, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \right\} \\ & \geq \mathbb{P} \left\{ \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f^*, Z) | \mathcal{D}_n], \right. \\ & \quad \left. \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z'_i) \leq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \right\} \\ & = \mathbb{E} \left(\mathbf{1} \left(\mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \right) \right. \\ & \quad \left. \mathbb{P} \left\{ \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z'_i) \leq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \right\} \right) \\ & \geq \frac{7}{8} \mathbb{P} \left\{ \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] - \frac{1}{n} \sum_{i=1}^n g(f^*, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f^*, Z) | \mathcal{D}_n] \right\} \\ & = \frac{7}{8} \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f, Z)] \right\}. \end{aligned}$$

To conclude, for $n > 8\zeta_n/[\epsilon^2(\alpha + \beta)]$, we have

$$\begin{aligned} & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g(f, Z)] \right\} \\ & \leq \frac{8}{7} \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f, Z)] \right\}. \end{aligned}$$

Step 2. Randomization for $\mathbb{E}[g(f, Z)]$. By introducing additional conditions, we notice that (10)

$$\begin{aligned} & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f, Z)] \right\} \\ & \leq \mathbb{P} \left(\exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f, Z)], \right. \\ & \quad \left. \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \mathbb{E}[g(f, Z)^2] \leq \epsilon \left\{ \alpha + \beta + \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 + \mathbb{E}[g(f, Z)^2] \right\} \right. \\ & \quad \left. \frac{1}{n} \sum_{i=1}^n g(f, Z'_i)^2 - \mathbb{E}[g(f, Z)^2] \leq \epsilon \left\{ \alpha + \beta + \frac{1}{n} \sum_{i=1}^n g(f, Z'_i)^2 + \mathbb{E}[g(f, Z)^2] \right\} \right) \\ & \quad + 2 \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{\frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \mathbb{E}[g(f, Z)^2]}{\alpha + \beta + \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 + \mathbb{E}[g(f, Z)^2]} > \epsilon \right\}. \end{aligned}$$

Then, Lemma C.2 verifies that

$$\begin{aligned} & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{\frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \mathbb{E}[g(f, Z)^2]}{\alpha + \beta + \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 + \mathbb{E}[g(f, Z)^2]} > \epsilon \right\} \\ & \leq 4 \mathbb{E} \left[\mathcal{N} \left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n} \right) \right] \exp \left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2} \right). \end{aligned}$$

Next, we focus on the first probability on the right-hand side of Eqn. (10). The second inequality inside the probability demonstrates that

$$(1 + \epsilon) \mathbb{E}[g(f, Z)^2] \geq (1 - \epsilon) \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \epsilon(\alpha + \beta),$$

which equals to

$$\frac{1}{2\zeta_n} \mathbb{E}[g(f, Z)^2] \geq \frac{1 - \epsilon}{2\zeta_n(1 + \epsilon)} \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \frac{\epsilon(\alpha + \beta)}{2\zeta_n(1 + \epsilon)},$$

while the third inequality is processed in the same manner. By the assumption that $\mathbb{E}[g(f, Z)^2] \leq \zeta_n \mathbb{E}[g(f, Z)]$ for all $f \in \mathcal{F}_n$, the first probability on the right-hand side of Eqn. (10) is bounded by

$$\begin{aligned} & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) \right. \\ & \quad \left. + \frac{\epsilon}{2} \left[\frac{1 - \epsilon}{2\zeta_n(1 + \epsilon)} \frac{1}{n} \sum_{i=1}^n g(f, Z_i)^2 - \frac{\epsilon(\alpha + \beta)}{2\zeta_n(1 + \epsilon)} + \frac{1 - \epsilon}{2\zeta_n(1 + \epsilon)} \frac{1}{n} \sum_{i=1}^n g(f, Z'_i)^2 - \frac{\epsilon(\alpha + \beta)}{2\zeta_n(1 + \epsilon)} \right] \right\}. \end{aligned}$$

This shows

$$\begin{aligned}
(11) \quad & \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n g(f, Z'_i) - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g(f, Z)] \right\} \\
& \leq \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n [g(f, Z'_i) - g(f, Z_i)] \geq \frac{\epsilon}{2}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{2\zeta_n(1 + \epsilon)} \right. \\
& \quad \left. + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n [g(f, Z_i)^2 + g(f, Z'_i)^2] \right\} \\
& \quad + 8\mathbb{E} \left[\mathcal{N} \left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n} \right) \right] \exp \left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2} \right).
\end{aligned}$$

Step 3. Introduction of Rademacher random variables. Let U_1, \dots, U_n be independent Rademacher random variables which are uniformly distributed over $\{-1, 1\}$, meanwhile independent of $\mathcal{D}_n \cup \mathcal{D}'_n$. We note that \mathcal{D}_n and \mathcal{D}'_n are interchangeable with respect to corresponding components while their joint distribution remains invariant. As a consequence, the first probability on the right-hand side of Eqn. (11) is equivalent to

$$\begin{aligned}
& \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \frac{1}{n} \sum_{i=1}^n U_i [g(f, Z'_i) - g(f, Z_i)] \geq \frac{\epsilon}{2}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{2\zeta_n(1 + \epsilon)} \right. \\
& \quad \left. + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n [g(f, Z_i)^2 + g(f, Z'_i)^2] \right\},
\end{aligned}$$

which is further bounded by

$$\begin{aligned}
& \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \left| \frac{1}{n} \sum_{i=1}^n U_i g(f, Z'_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\zeta_n(1 + \epsilon)} + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n g(f, Z'_i)^2 \right\} \\
& \quad + \mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \left| \frac{1}{n} \sum_{i=1}^n U_i g(f, Z_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\zeta_n(1 + \epsilon)} + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n g(f, Z_i)^2 \right\} \\
& = 2\mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \left| \frac{1}{n} \sum_{i=1}^n U_i g(f, Z_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\zeta_n(1 + \epsilon)} + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n g(f, Z_i)^2 \right\}.
\end{aligned}$$

Step 4. Conditioning and Covering. Given $Z_i = z_i$ for $i = 1, \dots, n$, consider

$$\mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \left| \frac{1}{n} \sum_{i=1}^n U_i g(f, z_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\zeta_n(1 + \epsilon)} + \frac{\epsilon(1 - \epsilon)}{4\zeta_n(1 + \epsilon)n} \sum_{i=1}^n g(f, z_i)^2 \right\}.$$

Let $\delta > 0$ and let \mathcal{C}_δ be a δ -covering set of $\{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}$ constrained on $\{z_1, \dots, z_n\}$ with respect to the supremum norm. For any $f \in \mathcal{F}$, there exists a vector $h^\sharp = (h(z_1), \dots, h(z_n))^\top \in \mathcal{C}_\delta$, such that $\max_{i=1, \dots, n} |g(f, z_i) - h(z_i)| < \epsilon$, thereby indicat-

ing

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n U_i g(f, z_i) \right| &= \left| \frac{1}{n} \sum_{i=1}^n U_i h(z_i) + \frac{1}{n} \sum_{i=1}^n U_i [g(f, z_i) - h(z_i)] \right| \\
&\leq \left| \frac{1}{n} \sum_{i=1}^n U_i h(z_i) \right| + \frac{1}{n} \sum_{i=1}^n |g(f, z_i) - h(z_i)| \\
&< \left| \frac{1}{n} \sum_{i=1}^n U_i h(z_i) \right| + \delta,
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n g(f, z_i)^2 &= \frac{1}{n} \sum_{i=1}^n h(z_i)^2 + \frac{1}{n} \sum_{i=1}^n [g(f, z_i)^2 - h(z_i)^2] \\
&= \frac{1}{n} \sum_{i=1}^n h(z_i)^2 + \frac{1}{n} \sum_{i=1}^n [g(f, z_i) - h(z_i)] [g(f, z_i) + h(z_i)] \\
&\geq \frac{1}{n} \sum_{i=1}^n h(z_i)^2 - \frac{2\xi_n}{n} \sum_{i=1}^n |g(f, z_i) - h(z_i)| \\
&> \frac{1}{n} \sum_{i=1}^n h(z_i)^2 - 2\delta\xi_n.
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
&\mathbb{P} \left\{ \exists f \in \mathcal{F}_n : \left| \frac{1}{n} \sum_{i=1}^n U_i g(f, z_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\xi_n(1 + \epsilon)} + \frac{\epsilon(1 - \epsilon)}{4\xi_n(1 + \epsilon)n} \sum_{i=1}^n g(f, z_i)^2 \right\} \\
&\leq \mathbb{P} \left\{ \exists h^\sharp \in \mathcal{C}_\delta : \left| \frac{1}{n} \sum_{i=1}^n U_i h(z_i) \right| + \delta \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\xi_n(1 + \epsilon)} \right. \\
&\quad \left. + \frac{\epsilon(1 - \epsilon)}{4\xi_n(1 + \epsilon)} \left[\frac{1}{n} \sum_{i=1}^n h(z_i)^2 - 2\delta\xi_n \right] \right\} \\
&\leq |\mathcal{C}_\delta| \max_{h^\sharp \in \mathcal{C}_\delta} \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n U_i h(z_i) \right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\xi_n(1 + \epsilon)} - \delta - \frac{\epsilon(1 - \epsilon)\delta\xi_n}{2\xi_n(1 + \epsilon)} \right. \\
&\quad \left. + \frac{\epsilon(1 - \epsilon)}{4\xi_n(1 + \epsilon)n} \sum_{i=1}^n h(z_i)^2 \right\}.
\end{aligned}$$

Next we set $\delta = (6\eta - 2)\epsilon\beta/(30\eta + 3\gamma\eta) = \varpi\epsilon\beta$. Then, when $n \geq N$, we have $\zeta_n \geq \eta$, $\xi_n \leq \gamma\xi_n$, and for $0 < \epsilon \leq 1/2$,

$$\begin{aligned}
\frac{\epsilon\beta}{4} - \frac{\epsilon^2\beta}{4\xi_n(1 + \epsilon)} - \delta - \frac{\epsilon(1 - \epsilon)\delta\xi_n}{2\xi_n(1 + \epsilon)} &= \epsilon\beta \left(\frac{1}{4} - \frac{\epsilon}{4\xi_n(1 + \epsilon)} - \varpi - \frac{\epsilon(1 - \epsilon)\varpi\xi_n}{2\xi_n(1 + \epsilon)} \right) \\
&\geq \epsilon\beta \left(\frac{1}{4} - \frac{1}{12\xi_n} - \varpi - \frac{\varpi\xi_n}{10\xi_n} \right) \\
&\geq \epsilon\beta \left(\frac{1}{4} - \frac{1}{12\eta} - \varpi - \frac{\varpi\gamma}{10} \right) \geq 0.
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned} & \mathbb{P}\left\{\exists f \in \mathcal{F}_n : \left|\frac{1}{n} \sum_{i=1}^n U_i g(f, z_i)\right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\zeta_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\zeta_n(1+\epsilon)n} \sum_{i=1}^n g(f, z_i)^2\right\} \\ & \leq |\mathcal{C}_{\varpi\epsilon\beta}| \max_{h^\sharp \in \mathcal{C}_{\varpi\epsilon\beta}} \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n U_i h(z_i)\right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\zeta_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\zeta_n(1+\epsilon)n} \sum_{i=1}^n h(z_i)^2\right\}. \end{aligned}$$

Step 5. Leveraging the Bernstein's inequality. Firstly, we note that

$$\frac{1}{n} \sum_{i=1}^n \text{Var}[U_i h(z_i)] = \frac{1}{n} \sum_{i=1}^n h(z_i)^2 \text{Var}(U_i) = \frac{1}{n} \sum_{i=1}^n h(z_i)^2.$$

Hence, we have

$$\begin{aligned} & \mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n U_i h(z_i)\right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\zeta_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\zeta_n(1+\epsilon)n} \sum_{i=1}^n h(z_i)^2\right\} \\ & = \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n V_i\right| \geq A_1 + A_2\sigma^2\right), \end{aligned}$$

where

$$\begin{aligned} V_i &= U_i h(z_i), \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}[U_i h(z_i)] \\ A_1 &= \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\zeta_n(1+\epsilon)}, \quad A_2 = \frac{\epsilon(1-\epsilon)}{4\zeta_n(1+\epsilon)}. \end{aligned}$$

Observe that V_1, \dots, V_n are independent random variables satisfying $|V_i| \leq |h(z_i)| \leq \xi_n$ ($i = 1, \dots, n$), and that $A_1, A_2 > 0$ for $n \geq N$. By Bernstein's inequality, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n V_i\right| \geq A_1 + A_2\sigma^2\right) &\leq 2 \exp\left(-\frac{n(A_1 + A_2\sigma^2)^2}{2\sigma^2 + 2(A_1 + A_2\sigma^2)\frac{\xi_n}{3}}\right) \\ &= 2 \exp\left(-\frac{nA_2^2}{\frac{2}{3}\xi_n A_2} \cdot \frac{(\frac{A_1}{A_2} + \sigma^2)^2}{\frac{A_1}{A_2} + (1 + \frac{3}{\xi_n A_2})\sigma^2}\right) \\ &= 2 \exp\left(-\frac{3nA_2}{2\xi_n} \cdot \frac{(\frac{A_1}{A_2} + \sigma^2)^2}{\frac{A_1}{A_2} + (1 + \frac{3}{\xi_n A_2})\sigma^2}\right). \end{aligned}$$

It is easy to verify that for arbitrary $a, b, u > 0$, it follows that

$$\frac{(a+u)^2}{a+bu} \geq \frac{4a}{b^2}[(b-1) \vee 0].$$

Then, by letting $a = A_1/A_2$, $b = 1 + 3/(\xi_n A_2)$, $u = \sigma^2$, we obtain

$$\frac{3nA_2}{2\xi_n} \cdot \frac{(\frac{A_1}{A_2} + \sigma^2)^2}{\frac{A_1}{A_2} + (1 + \frac{3}{\xi_n A_2})\sigma^2} \geq \frac{3nA_2}{2\xi_n} \cdot \frac{4\frac{A_1}{A_2}}{(1 + \frac{3}{\xi_n A_2})^2} \cdot \frac{3}{\xi_n A_2} = \frac{18nA_1A_2}{(\xi_n A_2 + 3)^2}.$$

In addition, notice that for $n \geq N$ and $0 < \epsilon \leq 1/2$,

$$A_1 = \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\zeta_n(1+\epsilon)} > \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4(1+\epsilon)} = \frac{\epsilon\alpha}{4} \left(1 - \frac{\epsilon}{1+\epsilon}\right) \geq \frac{\epsilon\alpha}{6},$$

which results in

$$\begin{aligned} \frac{18nA_1A_2}{(\xi_n A_2 + 3)^2} &= 18n \cdot \frac{\epsilon\alpha}{6} \cdot \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)} \cdot \frac{1}{\left[\frac{\xi_n\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)} + 3\right]^2} \\ &\geq 18n \cdot \frac{\epsilon\alpha}{6} \cdot \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)} \cdot \frac{1}{\left(\frac{\gamma}{20} + 3\right)^2} \geq \frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\xi_n(1+\epsilon)}, \end{aligned}$$

where $\omega = 400/(\gamma + 60)^2$. To conclude, it follows that

$$\begin{aligned} &\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i h(z_i)\right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\xi_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)n} \sum_{i=1}^n h(z_i)^2\right\} \\ &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n V_i\right| \geq A_1 + A_2\sigma^2\right) \leq 2\exp\left(-\frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\xi_n(1+\epsilon)}\right). \end{aligned}$$

Step 6. Conclusion. We have shown that, for $n > (8\xi_n/[\epsilon^2(\alpha + \beta)]) \vee N$, it follows that

$$\begin{aligned} &\mathbb{P}\left(\exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n}\sum_{i=1}^n g(f, Z_i) \geq \epsilon\{\alpha + \beta + \mathbb{E}[g(f, Z)]\}\right) \\ &\leq \frac{8}{7}\mathbb{P}\left\{\exists f \in \mathcal{F}_n : \frac{1}{n}\sum_{i=1}^n g(f, Z'_i) - \frac{1}{n}\sum_{i=1}^n g(f, Z_i) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g(f, Z)]\right\} \\ &\leq \frac{8}{7}\left(\mathbb{P}\left\{\exists f \in \mathcal{F}_n : \frac{1}{n}\sum_{i=1}^n [g(f, Z'_i) - g(f, Z_i)] \geq \frac{\epsilon}{2}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{2\xi_n(1+\epsilon)}\right.\right. \\ &\quad \left.\left.+ \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)n} \sum_{i=1}^n [g(f, Z_i)^2 + g(f, Z'_i)^2]\right\}\right. \\ &\quad \left.+ 8\mathbb{E}\left[\mathcal{N}\left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n}\right)\right] \exp\left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2}\right)\right) \\ &\leq \frac{8}{7}\left(2\mathbb{P}\left\{\exists f \in \mathcal{F}_n : \left|\frac{1}{n}\sum_{i=1}^n U_i g(f, Z_i)\right| \geq \frac{\epsilon}{4}(\alpha + \beta) - \frac{\epsilon^2(\alpha + \beta)}{4\xi_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)n} \sum_{i=1}^n g(f, Z_i)^2\right\}\right. \\ &\quad \left.+ 8\mathbb{E}\left[\mathcal{N}\left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n}\right)\right] \exp\left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2}\right)\right) \\ &\leq \frac{8}{7}\left\{2\mathbb{E}\left(\left|\mathcal{C}_{\varpi\epsilon\beta}\right| \max_{h^\sharp \in \mathcal{C}_{\varpi\epsilon\beta}} \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i h(Z_i)\right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{4\xi_n(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{4\xi_n(1+\epsilon)n} \sum_{i=1}^n h(Z_i)^2\right\}\right)\right. \\ &\quad \left.+ 8\mathbb{E}\left[\mathcal{N}\left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n}\right)\right] \exp\left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2}\right)\right\} \\ &\leq \frac{8}{7}\left\{2\mathbb{E}\left[2|\mathcal{C}_{\varpi\epsilon\beta}| \exp\left(-\frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\xi_n(1+\epsilon)}\right)\right]\right. \\ &\quad \left.+ 8\mathbb{E}\left[\mathcal{N}\left(\frac{(\alpha + \beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n}\right)\right] \exp\left(-\frac{3\epsilon^2(\alpha + \beta)n}{40\xi_n^2}\right)\right\}. \end{aligned}$$

While conditioning on $Z_i = z_i$ for $i = 1, \dots, n$, we choose the $\varpi\epsilon\beta$ -covering set of minimal size, which yields

$$\begin{aligned} & \mathbb{P} \left(\exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \epsilon \{\alpha + \beta + \mathbb{E}[g(f, Z)]\} \right) \\ & \leq \frac{32}{7} \mathbb{E} [\mathcal{N}(\varpi\epsilon\beta, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n})] \exp \left(-\frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\zeta_n(1+\epsilon)} \right) \\ & \quad + \frac{64}{7} \mathbb{E} \left[\mathcal{N} \left(\frac{(\alpha+\beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}_{|\mathcal{D}_n} \right) \right] \exp \left(-\frac{3\epsilon^2(\alpha+\beta)n}{40\xi_n^2} \right) \\ & \leq \frac{32}{7} \mathcal{N}_n(\varpi\epsilon\beta, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}) \exp \left(-\frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\zeta_n(1+\epsilon)} \right) \\ & \quad + \frac{64}{7} \mathcal{N}_n \left(\frac{(\alpha+\beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\} \right) \exp \left(-\frac{3\epsilon^2(\alpha+\beta)n}{40\xi_n^2} \right). \end{aligned}$$

Observe that for $n \geq N$, we have $\eta \geq 1, \gamma > 0$,

$$\varpi = \frac{6\eta - 2}{30\eta + 3\gamma\eta} \leq \frac{6\eta}{30\eta} = \frac{1}{5}, \quad \omega = \frac{400}{(\gamma + 60)^2} \leq \frac{1}{9},$$

and

$$\begin{aligned} \frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\zeta_n(1+\epsilon)} & \geq \frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40\zeta_n(1+\epsilon)} \geq \frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40(\xi_n^2 \vee \zeta_n)(1+\epsilon)}, \\ \frac{3\epsilon^2(\alpha+\beta)n}{40\xi_n^2} & \geq \frac{3\epsilon^2\alpha n}{40(\xi_n^2 \vee \zeta_n)} \geq \frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40(\xi_n^2 \vee \zeta_n)(1+\epsilon)}. \end{aligned}$$

Consequently, it follows that

$$\begin{aligned} & \mathbb{P} \left(\exists f \in \mathcal{F}_n : \mathbb{E}[g(f, Z)] - \frac{1}{n} \sum_{i=1}^n g(f, Z_i) \geq \epsilon \{\alpha + \beta + \mathbb{E}[g(f, Z)]\} \right) \\ (12) \quad & \leq \frac{32}{7} \mathcal{N}_n(\varpi\epsilon\beta, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}) \exp \left(-\frac{3\epsilon^2(1-\epsilon)\alpha\omega n}{4\zeta_n(1+\epsilon)} \right) \\ & \quad + \frac{64}{7} \mathcal{N}_n \left(\frac{(\alpha+\beta)\epsilon}{5}, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\} \right) \exp \left(-\frac{3\epsilon^2(\alpha+\beta)n}{40\xi_n^2} \right) \\ & \leq 14 \mathcal{N}_n(\varpi\epsilon\beta, \|\cdot\|_\infty, \{g(f, \cdot) : \mathcal{Z} \rightarrow \mathbb{R}, f \in \mathcal{F}_n\}) \exp \left(-\frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40(\xi_n^2 \vee \zeta_n)(1+\epsilon)} \right), \end{aligned}$$

for $n > (8\zeta_n/[\epsilon^2(\alpha+\beta)]) \vee N$. When $N \leq n \leq 8\zeta_n/[\epsilon^2(\alpha+\beta)]$, on the other hand, we note

$$\exp \left(-\frac{27\epsilon^2(1-\epsilon)\alpha\omega n}{40(\xi_n^2 \vee \zeta_n)(1+\epsilon)} \right) \geq \exp \left(-\frac{3}{5} \right) \geq \frac{1}{14},$$

demonstrating that the last right-hand side of Eqn. (12) exceeds one, and hence the inequality holds trivially, which completes the proof. \square

REFERENCES

- [1] ALBERGO, M. S. and VANDEN-EIJNDEN, E. (2023). Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*.
- [2] ANTHONY, M. and BARTLETT, P. L. (1999). Neural Network Learning: Theoretical Foundations. *Ai Magazine* **22** 99-100.

- [3] BARTLETT, P. L., HARVEY, N. J. A., LIAW, C. and MEHRABIAN, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research* **20**.
- [4] BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285.
- [5] BREGMAN, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* **7** 200–217.
- [6] BRIS, C. L. and LIONS, P.-L. (2008). Existence and uniqueness of solutions to Fokker–Planck type equations with irregular coefficients. *Communications in Partial Differential Equations* **33** 1272–1317.
- [7] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics* **7** 331–368.
- [8] CHEN, J. and JIANG, N. (2022). Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence* 378–388. PMLR.
- [9] CORTES, C., MANSOUR, Y. and MOHRI, M. (2010). Learning bounds for importance weighting. *Advances in neural information processing systems* **23**.
- [10] FANG, T., LU, N., NIU, G. and SUGIYAMA, M. (2020). Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems* **33** 11996–12007.
- [11] FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213.
- [12] FENG, X., HE, X., JIAO, Y., KANG, L. and WANG, C. (2024). Deep nonparametric quantile regression under covariate shift. *Journal of Machine Learning Research* **25** 1–50.
- [13] GAO, Y., HUANG, J., and JIAO, Y. (2024). Gaussian interpolation flows. *Journal of Machine Learning Research* **25** 1–52.
- [14] GAO, Y., HUANG, J., JIAO, Y., LIU, J., LU, X. and YANG, Z. (2022). Deep generative learning via Euler particle transport. In *Mathematical and Scientific Machine Learning* 336–368. PMLR.
- [15] GAO, Y., HUANG, J., JIAO, Y. and ZHENG, S. (2024). Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*.
- [16] GE, J., TANG, S., FAN, J., MA, C. and JIN, C. (2023). Maximum likelihood estimation is all you need for well-specified covariate shift. *International Conference on Learning Representations*.
- [17] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems* **27**.
- [18] GROVER, A. and ERMON, S. (2018). Boosted generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**.
- [19] GYÖRFI, L., KOHLER, M., KRZYZAK, A., WALK, H. et al. (2002). *A Distribution-Free Theory of Nonparametric Regression* **1**. Springer.
- [20] HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778.
- [21] HENG, A., ANSARI, A. F. and SOH, H. (2024). Generative modeling with flow-guided density ratio learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 250–267. Springer.
- [22] HO, J., JAIN, A. and ABBEEL, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33** 6840–6851.
- [23] HUANG, D., HUANG, J., LI, T. and SHEN, G. (2023). Conditional stochastic interpolation for generative learning. *arXiv preprint arXiv:2312.05579*.
- [24] JIAO, Y., LAI, Y., WANG, Y. and YAN, B. (2024). Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*.
- [25] JIAO, Y., SHEN, G., LIN, Y. and HUANG, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *The Annals of Statistics* **51** 691–716.
- [26] KANTOROVICH, L. V. (1960). Mathematical methods of organizing and planning production. *Management science* **6** 366–422.
- [27] KATO, M. and TESHIMA, T. (2021). Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning* 5320–5333. PMLR.
- [28] KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [29] KREMER, J., GIESEKE, F., PEDERSEN, K. S. and IGEL, C. (2015). Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing* **12** 67–72.
- [30] LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86** 2278–2324.

- [31] LEI, L. and CANDÈS, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83** 911–938.
- [32] LI, F., LAM, H. and PRUSTY, S. (2020). Robust importance weighting for covariate shift. In *International conference on artificial intelligence and statistics* 352–362. PMLR.
- [33] LIN, Z., DING, P. and HAN, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica* **91** 2187–2217.
- [34] LIPMAN, Y., CHEN, R. T. Q., BEN-HAMU, H., NICKEL, M. and LE, M. (2023). Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.
- [35] LIU, Q., LI, L., TANG, Z. and ZHOU, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems* **31**.
- [36] LIU, S., YAMADA, M., COLLIER, N. and SUGIYAMA, M. (2013). Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* **43** 72–83.
- [37] LIU, T. and TAO, D. (2015). Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence* **38** 447–461.
- [38] LIU, X., GONG, C. and QIANG LIU (2022). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *NeurIPS 2022 Workshop on Score-Based Methods*.
- [39] MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics* **51** 738–761.
- [40] NAIR, N. G., SATPATHY, P., CHRISTOPHER, J. et al. (2019). Covariate shift: A review and analysis on classifiers. In *2019 Global Conference for Advancement in Technology (GCAT)* 1–6. IEEE.
- [41] NAKADA, R. and IMAIZUMI, M. (2020). Adaptive approximation and estimation of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research* **21** 1–38.
- [42] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* **56** 5847–5861.
- [43] PAPAMAKIROS, G., NALISNICK, E., REZENDE, D. J., MOHAMED, S. and LAKSHMINARAYANAN, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22** 1–64.
- [44] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**.
- [45] REZAEI, A., LIU, A., MEMARRAST, O. and ZIEBART, B. D. (2021). Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence* **35** 9419–9427.
- [46] REZENDE, D. and MOHAMED, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning* 1530–1538. PMLR.
- [47] RHODES, B., XU, K. and GUTMANN, M. U. (2020). Telescoping density-ratio estimation. *Advances in neural information processing systems* **33** 4905–4916.
- [48] SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of statistics* **48** 1875–1897.
- [49] SCHMIDT-HIEBER, J. and ZAMOLODCHIKOV, P. (2024). Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli* **30** 1845–1877.
- [50] SHEN, J., QU, Y., ZHANG, W. and YU, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* **32**.
- [51] SHEN, Z. (2020). Deep network approximation characterized by number of neurons. *Communications in Computational Physics* **28**.
- [52] SONG, Y. and ERMON, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32**.
- [53] STOJANOV, P., GONG, M., CARBONELL, J. and ZHANG, K. (2019). Low-dimensional density ratio estimation for covariate shift correction. In *The 22nd international conference on artificial intelligence and statistics* 3449–3458. PMLR.
- [54] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics* 1040–1053.
- [55] SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. and KAWANABE, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems* **20**.
- [56] SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* **64** 1009–1044.

- [57] SUGIYAMA, M., SUZUKI, T., NAKAJIMA, S., KASHIMA, H., VON BÜNAU, P. and KAWANABE, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* **60** 699–746.
- [58] SUZUKI, T., SUGIYAMA, M. and TANAKA, T. (2009). Mutual information approximation via maximum likelihood estimation of density ratio. In *2009 IEEE International Symposium on Information Theory* 463–467. IEEE.
- [59] TSUBOI, Y., KASHIMA, H., HIDO, S., BICKEL, S. and SUGIYAMA, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing* **17** 138–155.
- [60] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer New York, NY.
- [61] VAN DE GEER, S. (1990). Estimating a regression function. *The Annals of Statistics* 907–924.
- [62] VAN DER VAART, A. W., WELLNER, J. A., VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence*. Springer.
- [63] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. and POLOSKHIN, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17* 6000–6010. Curran Associates Inc., Red Hook, NY, USA.
- [64] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* **48**. Cambridge University press.
- [65] WANG, X., BORSOI, R. A., RICHARD, C. and CHEN, J. (2023). Change point detection with neural online density-ratio estimator. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1–5. IEEE.
- [66] WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *Journal of Big data* **3** 1–40.
- [67] WU, J., ZOU, D., BRAVERMAN, V., GU, Q. and KAKADE, S. (2022). The power and limitation of pretraining-finetuning for linear regression under covariate shift. *Advances in Neural Information Processing Systems* **35** 33041–33053.
- [68] YAMADA, M., SUZUKI, T., KANAMORI, T., HACHIYA, H. and SUGIYAMA, M. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural computation* **25** 1324–1370.
- [69] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599.
- [70] ZHANG, Y.-J., ZHANG, Z.-Y., ZHAO, P. and SUGIYAMA, M. (2024). Adapting to continuous covariate shift via online density ratio estimation. *Advances in Neural Information Processing Systems* **36**.