Journal of Machine Learning Research 27 (2026) 1-60 Submitted 4/23; Revised 3/25; Published 1/26

Error Analysis for Deep ReLU Feedforward Density-Ratio Estimation with Bregman Divergence

Siming Zheng simizheng4-c@my.cityu.edu.hk
School of Statistics and Data Science
Southeast University
Jiangsu, China

Guohao Shen guohao.shen@polyu.edu.hk
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China

Yuanyuan Lin[*] ylin@sta.cuhk.edu.hk
Department of Statistics and Data Science
The Chinese University of Hong Kong
Hong Kong, China

Jian Huang[*] j.huang@polyu.edu.hk
Departments of Data Science and AI, and Applied Mathematics
The Hong Kong Polytechnic University
Hong Kong, China

Editor: Kenji Fukumizu

Abstract

We consider the problem of density-ratio estimation using Bregman Divergence with Deep ReLU feedforward neural networks (BDD). We establish non-asymptotic error bounds for BDD density-ratio estimators, which are minimax optimal up to a logarithmic factor when the data distribution has finite support. As an application of our theoretical findings, we propose an estimator for the KL-divergence that is asymptotically normal, leveraging our convergence results for the deep density-ratio estimator and a data-splitting method. We also extend our results to cases with unbounded support and unbounded density ratios. Furthermore, we show that the BDD density-ratio estimator can mitigate the curse of dimensionality when data distributions are supported on an approximately low-dimensional manifold. Our results are applied to investigate the convergence properties of the telescoping density-ratio estimator proposed by Rhodes (2020). We provide sufficient conditions under which it achieves a lower error bound than a single-ratio estimator. Moreover, we conduct simulation studies to validate our main theoretical results and assess the performance of the BDD density-ratio estimator.

Keywords: Non-asymptotic error; Deep neural networks; Curse of dimensionality; KL-divergence estimation; Telescoping density-ratio estimator.

[*] Yuanyuan Lin and Jian Huang are the co-corresponding authors.

1. Introduction

We consider the problem of density-ratio estimation using deep neural networks. Let Z_q and $Z_p \in \mathcal{Z} = [0,1]^d$ be two random vectors with probability density functions q^* and p^*, respectively. It is assumed that q^* and p^* have the same support. Given independent and identically distributed (i.i.d) samples $S_q = \{Z_{q,i}\}_{i=1}^{n_q}$ from q^* and $S_p = \{Z_{p,j}\}_{j=1}^{n_p}$ from p^*, a basic problem in many statistical and machine learning tasks (Sugiyama et al., 2012a; Kato and Teshima, 2021) is to estimate the density ratio

$$R^*(z) = q^*(z)/p^*(z), z \in \mathcal{Z}.$$

Density ratio plays a key role in a wide range of applications, including discriminative analysis (Silverman, 1978; Cox and Ferry, 1991), covariate shift adaptation (Sugiyama et al., 2008; Tsuboi et al., 2009), two-sample testing (Qin, 1998; Sugiyama et al., 2011), energy-based modelling (Gutmann and Hyvärinen, 2012; Ceylan and Gutmann, 2018), generative learning (Goodfellow et al., 2014; Nowozin et al., 2016; Gao et al., 2022), and mutual information estimation (Moustakides and Basioti, 2019; Rhodes et al., 2020), among others. In this paper, we study the properties of density-ratio estimation with the Bregman Divergence using Deep neural networks (BDD).

A naive density-ratio estimator of R^* is the ratio of individual density estimators, that is, $\hat{q}/\hat{p}$, where $\hat{q}$ and $\hat{p}$ are the density estimators of q^* and p^*, respectively. However, such an estimator can be highly unstable. Moreover, density estimation itself is a difficult problem, especially in the high-dimensional settings. For example, kernel density estimators (Rosenblatt, 1956; Parzen, 1962) works well when $d \leq 3$, but deteriorate dramatically as d increases. To avoid the estimation of individual densities, various methods have been proposed to estimate the density ratio R^* directly, including the density matching approach (Sugiyama et al., 2008; Tsuboi et al., 2009; Yamada and Sugiyama, 2009; Nguyen et al., 2010; Yamada et al., 2010), the moment matching approach (Qin, 1998; Gretton et al., 2009; Kanamori et al., 2012b), the density-ratio fitting approach (Kanamori et al., 2009, 2012a), and the unified density-ratio matching approach under Bregman divergence framework (Sugiyama et al., 2012b). Impressive empirical successes of using deep neural networks in density-ratio estimation have been reported in some recent works (Moustakides and Basioti, 2019; Rhodes et al., 2020).

Several studies have established error bounds for density ratio estimation in reproducing kernel Hilbert space (Nguyen et al., 2010; Sugiyama et al., 2008; Kanamori et al., 2012a; Yamada et al., 2013). However, there is a lack of systematic analysis of the properties of the density-ratio estimation using deep neural networks. To the best of our knowledge, the only work is Kato and Teshima (2021), which studied the convergence properties of deep density-ratio estimation under a modified Bregman divergence criterion. This intriguing work makes a well-specified model assumption, that is, the target density ratio function is assumed to be bounded and belongs to a class of neural networks with bounded weights. Under such an assumption, the theoretical analysis is less challenging as there is no need to analyze the approximation error incurred by approximating a smooth function using deep neural networks, i.e., the approximation error is zero. Nevertheless, such an assumption could be unrealistic as the true density ratio may not conform to a neural network structure.

In this work, we establish error bounds for BDD density-ratio estimators. Compared with the existing works, we do not make the well-specified model assumption and allow the target density ratio to be in a general class of smooth functions. We also relax the boundedness assumption on the target density ratio. Such a boundedness assumption is usually assumed in error analysis in nonparametric statistics. Moreover, we apply our results to construct an estimator for statistical inference for the Kullback-Liebler divergence and study the theoretical properties of the telescoping density-ratio estimator (Rhodes et al., 2020). Our contributions are as follows:

1. We establish non-asymptotic error bounds for the density-ratio estimator using deep ReLU feedforward neural networks under the Bregman divergence under a bounded support assumption (BD, Bregman, 1967), and provide a neural network architecture for the estimator to achieve minimax optimal rate up to a logarithmic factor.
2. We extend our results to the case of unbounded support and demonstrate that the proposed deep BDD density-ratio estimator effectively mitigates the curse of dimensionality when the data lies on an approximately low-dimensional manifold. Furthermore, we establish an extension of our result to unbounded density ratios.
3. We apply our results to two important problems: (a) we propose an asymptotically normal distributed estimator of the Kullback-Leibler (KL) divergence and illustrate the results through simulation studies; (b) we study the convergence properties of the telescoping density-ratio estimator (Rhodes et al., 2020) with a mixing chain of intermediate samples, and demonstrate the advantages over single-ratio estimators under certain conditions.
4. We conduct simulation studies to numerically validate our main theoretical results, evaluate the performance of our proposed KL estimator, and examine the impact of certain tuning parameters on the proposed mixing telescoping density-ratio estimator (mTRE). We also compare the mTRE to the convolution-based telescoping density-ratio estimator (cTRE) from Rhodes (2020). The numerical comparisons highlight the advantages of mTRE over cTRE.

The rest of the paper is organized as follows. We first describe the problem of density-ratio estimation with Bregman divergence and deep ReLU feedforward neural networks (BDD) in Section 2. We present error bounds for BDD density-ratio estimators in Section 3, where we discuss the optimality of the BDD density-ratio estimator and apply our results to the problem of estimating the Kullback-Leibler (KL) divergence. In Section 4, we show how the BDD density-ratio estimator can mitigate the curse of dimensionality under proper assumptions and provide extended results to deal with unbounded target density ratio or target density ratio with unbounded support. As an application, we apply our theory to study the convergence properties of the telescoping density-ratio estimator. In Section 5, we conduct numerical experiments to validate our main theoretical results and evaluate the performance of the BDD density-ratio estimator. Section 6 contains a theoretical comparison to the density-ratio estimation results in a related work. Concluding remarks are given in Section 7.

Notation. Let $n = \min\{n_q, n_p\}$ be the smaller sample size of the two samples $S_q = \{Z_{q,i}\}_{i=1}^{n_q}$ and $S_p = \{Z_{p,j}\}_{j=1}^{n_p}$. In addition, $\|\cdot\|_\infty$ denotes the sup-norm on some specific domain, and C, C_0 are generic constants that may vary from place to place. For any measurable function f, we denote $\|f\|_{\max} := \max\{\|f\|_p, \|f\|_q\}$ and $\|f\|_{n_p,n_q} = \max\{\|f\|_{p,n_p}, \|f\|_{q,n_q}\}$, where $\|f\|_k^2 = E_{h^*} f^2(Z)$ and $\|f\|_{h,n_h}^2 = E_{n_h} f^2(Z) = (1/n_h)\sum_{t=1}^{n_h} f^2(Z_{h,t})$, $h = p, q$. In the rest of the paper, $\mathbb{I}\{\cdot\}$ denotes the indicator function.

2. Deep density-ratio estimation with Bregman divergence

In this section, we first present the density-ratio estimation problem using the Bregman divergence (BD, Bregman, 1967; Sugiyama et al., 2012a,b) and then describe the structure of the deep neural networks to be used in density-ratio estimation.

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a first-order continuously differentiable and strictly convex function. Define

$$\Delta_\psi(x, y) = \psi(x) - \psi(y) - \psi'(y)(x - y), \tag{1}$$

where ψ' is the derivative of ψ. Then, the convexity of ψ implies that $\Delta_\psi(x, y) \geq 0$ and the equality holds if and only if $x = y$. It follows that $E_{Z\sim p^*}\Delta_\psi(R^*(Z), R(Z)) \geq 0$ and the equality holds if and only if $R = R^*$ almost everywhere with respect to the probability measure with density p^*. Therefore, the target density-ratio $R^* = q^*/p^*$ can be characterized as a minimizer:

$$R^* \in \underset{R \text{ nonnegative and measurable}}{\operatorname{argmin}} E_{p^*}\Delta_\psi(R^*, R),$$

where to simplify the notation without causing confusion, we have written $E_{p^*}\Delta_\psi(R^*, R) = E_{Z\sim p^*}\Delta_\psi(R^*(Z), R(Z))$. We will use similar notation below. By the definition of $\Delta_\psi(x, y)$,

$$E_{p^*}\Delta_\psi(R^*, R) = E_{p^*}[\psi'(R)R - \psi(R)] - E_{p^*}[\psi'(R)R^*] + E_{p^*}[\psi(R^*)].$$

Since $E_{p^*}[\psi'(R)R^*] = E_{q^*}[\psi'(R)]$ by the definition of R^*, we have

$$E_{p^*}\Delta_\psi(R^*, R) = E_{p^*}[\psi'(R)R - \psi(R)] - E_{q^*}[\psi'(R)] + E_{p^*}[\psi(R^*)]. \tag{2}$$

Now since the last term on the right side in (2) $E_{p^*}[\psi(R^*)]$ is independent of R, we have

$$R^* \in \underset{R \text{ nonnegative and measurable}}{\operatorname{argmin}} \big\{E_{p^*}[\psi'(R)R - \psi(R)] - E_{q^*}[\psi'(R)]\big\}. \tag{3}$$

Hence, for any measurable function $R : \mathcal{Z} \to \mathbb{R}$, the BD score induced by ψ for estimating the target density-ratio $R^* = q^*/p^*$ is

$$\mathcal{B}_\psi(R) = E_{p^*}[\psi'(R)R - \psi(R)] - E_{q^*}[\psi'(R)]. \tag{4}$$

Then, R^* is the minimizer of $\mathcal{B}_\psi(R)$ over all nonnegative measurable functions.

Because a density ratio is always nonnegative, a nonnegative constraint is needed when defining the density ratio as a minimizer, as in (3). This makes the minimization problem more difficult to solve. To avoid the non-negative constraint of the density ratio, we first consider the log-density ratio $D^* := \log R^*$. Then the nonnegativity constraint is no longer needed and by (3), we have

$$D^* \in \underset{D \text{ measurable}}{\operatorname{argmin}} \mathcal{B}_\psi(\exp(D)).$$

In practice, the estimation of R^* can be based on an empirical version of $\mathcal{B}_\psi$ when random samples from p^* and q^* are available. Suppose we have samples $\{Z_{q,i}\}_{i=1}^{n_q}$ i.i.d. as q^* and $\{Z_{p,j}\}_{j=1}^{n_p}$ i.i.d. as p^*. Then the BDD estimator of D^* is given by

$$\widehat{D} \in \underset{D \in \mathcal{F}_n}{\operatorname{argmin}} \widehat{\mathcal{B}}_\psi(e^D), \tag{5}$$

where $\mathcal{F}_n$ is a class of neural network functions and $\widehat{\mathcal{B}}_\psi(e^D)$ is an empirical version of $\mathcal{B}_\psi(e^D)$ defined in (4), which can be written as

$$\widehat{\mathcal{B}}_\psi(e^D) = \frac{1}{n_p}\sum_{j=1}^{n_p} \mathcal{L}_1(D(Z_{p,j})) + \frac{1}{n_q}\sum_{i=1}^{n_q} \mathcal{L}_2(D(Z_{q,i})),$$

where

$$\mathcal{L}_1(t) = \psi'(e^t)e^t - \psi(e^t) \text{ and } \mathcal{L}_2(t) = -\psi'(e^t). \tag{6}$$

The density-ratio estimator is $\widehat{R} = \exp(\widehat{D})$.

We take the function class $\mathcal{F}_n$ to be $\mathcal{F}_{M,\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S}}$, a class of ReLU activated feedforward neural networks (FNNs) $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ with parameter $\boldsymbol{\theta}$, depth $\mathcal{D}$, width $\mathcal{W}$, size $\mathcal{S}$, number of neurons $\mathcal{U}$. We require that $\|f_{\boldsymbol{\theta}}\|_\infty \le M$ for some $0 \le M \le \infty$. There are $\mathcal{D}$ hidden layers and $(\mathcal{D}+1)$ layers in total. The width $\mathcal{W}$ is the maximum width of the hidden layers; the number of neurons $\mathcal{U}$ is defined as the number of neurons of $f_{\boldsymbol{\theta}}$; the size $\mathcal{S}$ is the total number of parameters in the network. Note that $\mathcal{D}, \mathcal{W}, \mathcal{U}, \mathcal{S}$ may depend on n, but we suppress the dependence for notational simplicity. We write $\mathcal{F}_{M,\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S}}$ as $\mathcal{F}_{\text{FNN}}$ for brevity.

We use ReLU activation function as it is convenient to bound the pseudo dimension of ReLU-activated networks. But our results are also valid for neural networks with piecewise linear activation functions, as piecewise linear activation functions can be expressed as a linear combination of ReLU activation functions.

3. Main theoretical results

In this section, we study the error bounds for the deep logarithmic density-ratio estimator for bounded density ratio with finite support. The bounds for the density-ratio estimator follows directly based on the properties of the exponential function. We also show that deep density-ratio estimator can mitigate the curse of dimensionality when data is supported on an approximate low-dimensional manifold.

3.1 General error bounds

To state our assumptions and results, we need the definitions of μ-smoothness, σ-strong convexity and pseudo dimension.

Definition 1 (μ-smoothness) *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be μ-smooth over a set $\mathcal{A} \subseteq \mathbb{R}$ if it is differentiable over $\mathcal{A}$ and its first-order derivative f' satisfies*

$$|f'(x) - f'(y)| \le \mu|x - y|, \ \forall\ x, y \in \mathcal{A}, \tag{7}$$

where $0 \le \mu < \infty$. The constant μ is called the smoothness parameter.

Definition 2 (σ-strong convexity) *A function $f : \mathbb{R} \to \mathbb{R}$ is called σ-strongly convex if the domain $dom(f)$ of f is convex and for any $x, y \in dom(f)$ and $\lambda \in [0, 1]$, f satisfies*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\sigma}{2}\lambda(1 - \lambda)(x - y)^2, \tag{8}$$

where $0 \leq \sigma < \infty$. The constant σ is called the strong convexity (SC) parameter.

Definition 3 (Pseudo dimension) *For a function class $\mathcal{F}$, its pseudo dimension $Pdim(\mathcal{F})$, is the largest integer B such that there exists $(x_1, x_2, \dots, x_B, y_1, y_2, \dots, y_B) \in \mathcal{Z}^B \times \mathbb{R}^B$ such that for any $(r_1, r_2, \dots, r_B) \in \{0, 1\}^B$, there exists an $f \in \mathcal{F}$ satisfying for any $i \in \{1, 2, \dots, B\} : f(x_i) > y_i \Leftrightarrow r_i = 1$ (Anthony and Bartlett, 1999; Bartlett et al., 2019).*

For any measurable function class $\mathcal{F}$, by the definition of VC dimension, $\text{VCdim}(\mathcal{F}) \leq \text{Pdim}(\mathcal{F})$. If $\mathcal{F}$ is the class of functions generated by ReLU FNNs, it follows from Theorem 14.1 of Anthony and Bartlett (1999) that $\text{Pdim}(\mathcal{F}) \leq \text{VCdim}(\mathcal{F})$. Hence, for the function class $\mathcal{F}$ generated by ReLU FNNs, $\text{Pdim}(\mathcal{F}) = \text{VCdim}(\mathcal{F})$.

We make the following assumptions in this paper.

Assumption 4 *The function ψ is μ-smooth & σ-strongly convex, that is, it satisfies (7) and (8).*

Table 1 includes some commonly-used ψ's that satisfy Assumption 4.

Assumption 5 *There exists a constant $0 < M < \infty$ such that $\|D^*\|_\infty \leq M, \|D\|_\infty \leq M$ for every $D \in \mathcal{F}_{\text{FNN}}$.*

Assumption 5 assumes that the target density ratio is bounded. Such an assumption is often made in nonparametric statistics for avoiding technical difficulties associated with dealing with unbounded functions. We will partially relax this assumption below. The finite M in Assumption 5 can be relaxed to $M = \mathcal{O}(\log \log n)$ at a small price of an extra logarithm term in the error bounds. The boundedness of a network can be achieved by clipping operation. For example, let $T_M(t) = -M\mathbb{I}\{t< -M\} + t\mathbb{I}\{-M \leq t \leq M\} + M\mathbb{I}\{t>M\}$ be the truncation function taking values in $[-M, M]$, where M is the clipping level. Simple algebra shows that we can write $T_M(t) = \sigma(t) - \sigma(\sigma(t) - M) - \{\sigma(-t) - \sigma(\sigma(-t) - M)\}$, where $\sigma(t) = \max(0, t)$ is the ReLU activation function. Therefore, $T_M(t)$ can be computed by a ReLU network with depth 2 and width 4. Hence, through network concatenation, we can ensure that the boundedness assumption is satisfied through clipping. When the clipping level is not less than $\|D^*\|_\infty$, the capacity of the clipped FNNs to approximating D^* is not impacted.

Table 1: Commonly-used Loss Functions ψ. LS: least squares; LR: logistic regression; LK: Kullback-Liebler. C is the crude prefactor in Theorem 6 when $M \geq 1$.

Name	$\psi(c)$	Domain	μ	σ	C
LS	$(c-1)^2$	$\mathbb{R}$	2	2	$\tilde{C}Me^{5M}$
LR	$c\log c - (c+1)\log(c+1)$	$[a,b]$	$\frac{1}{a(a+1)}$	$\frac{1}{b(b+1)}$	$\frac{\tilde{C}Me^{5M}b(b+1)}{a(a+1)}$
LK	$c\log c - c$	$[a,b]$	$\frac{1}{a}$	$\frac{1}{b}$	$\frac{\tilde{C}Me^{5M}b}{a}$

Define the best in class approximation of D^* in $\mathcal{F}_{\mathrm{FNN}}$ as $D_{\mathrm{NN}} \in \operatorname{argmin}_{D\in\mathcal{F}_{\mathrm{FNN}}} \|D - D^*\|_{\max}$. Denote

$$\xi_n = \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}. \tag{9}$$

Theorem 6 *Suppose Assumptions 4 and 5 are satisfied. When* $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$, *with probability at least* $1-\exp(-\gamma)$,

$$\|\widehat{D} - D^*\|_{\max} \leq C\left(\xi_n + \|D_{\mathrm{NN}} - D^*\|_{\max} + \sqrt{\frac{\gamma}{n}}\right),$$

and

$$\|\widehat{D} - D^*\|_{n_p,n_q} \leq 2C\left(\xi_n + \|D_{\mathrm{NN}} - D^*\|_{\max} + \sqrt{\frac{\gamma}{n}}\right),$$

for a constant

$$C = 2\max\left\{\frac{3072(C_1+C_2)}{c_0}, 3120M, \sqrt{\frac{\max\{6C_0, 124(C_1+C_2)M, 5(C_1+C_2)\}}{c_0}}\right\} + \log_2 M + 1,$$

where $C_1 = 2e^{2M}\mu, C_2 = e^M\mu$, $c_0 = \sigma e^{-3M}/2$ *and* $C_0 = \mu e^{3M}/2$. *When* $\sigma \leq \mu$ *and* $M \geq 1$, C *can be replaced by a simpler but crude bound* $\tilde{C}Me^{5M}\mu/\sigma$, *where* $\tilde{C}$ *is some universal positive constant.*

Our analysis techniques for the proof of Theorem 6 can be applied to other problems. In the proof, we develop a novel localization technique to handle two interactive empirical processes w.r.t the two involved samples as in Theorem 6. Moreover, our analysis framework is flexible to accommodate other neural network structures, such as sparse neural networks in Schmidt-Hieber (2020) and Kato and Teshima (2021). We study extensions of Theorem 6 and Theorem 10 to accommodate sparse neural networks in Appendix D.

We have the following corollary for the expected error.

Corollary 7 *Under the conditions of Theorem 6, there exists a constant* C *depending only on* (μ, σ, M), *such that*

$$E_{S_p,S_q}\|\widehat{D} - D^*\|_{\max}^2 \leq C\left(\xi_n^2 + \|D_{\mathrm{NN}} - D^*\|_{\max}^2\right),$$

and

$$E_{S_p,S_q}\|\widehat{D}-D^*\|^2_{n_p,n_q} \leq 2C\left(\xi_n^2+\|D_{\mathrm{NN}}-D^*\|^2_{\max}\right).$$

Under Assumption 5, to derive a nonasymptotic error bound for the log-density ratio estimator $\widehat{R}$, we note that

$$E_{S_p,S_q}\|\widehat{R}-R^*\|^2_{\max} \leq e^{2M}E_{S_p,S_q}\|\widehat{D}-D^*\|^2_{\max}.$$

Thus an upper bound for $\widehat{D}$ can induce an upper bound for $\widehat{R}$. Furthermore, under mild conditions, there exist some positive constants c and C such that

$$cE_{S_p,S_q}\|\widehat{R}-R^*\|^2_{\max} \leq E_{S_p,S_q}\|\widehat{R}-R^*\|^2_2 \leq CE_{S_p,S_q}\|\widehat{R}-R^*\|^2_{\max}.$$

where $\|f\|_2^2=\int_{\mathcal{Z}}f^2(z)dz$. With this review, the upper bounds for $E_{S_p,S_q}\|\widehat{D}-D^*\|^2_{\max}$ in Corollary 7 can provide general bounds for the L_2-type error associated with the density ratio estimation. The L_2-type error bounds are broadly considered in theoretical analysis across diverse fields, such as conformal prediction under covariate shift (Proposition 1 in Lei and Candès (2021), Theorem 2 in Yang et al. (2024) and Theorem 3 in Ai and Ren (2024)), two-sample conditional distribution test (Assumption 2 in Hu and Lei (2024)), double reinforcement learning for efficient off-policy evaluation (Theorem 12 in Kallus and Uehara (2020)).

Here we provide an example on conformal prediction under covariate shift to illustrate its usefulness. It is known that density ratio estimation plays an important role in conformal prediction under covariate shift (Tibshirani et al., 2019), where a source labeled dataset $S_p^Y=\{(Z_{p,i},Y_{p,i})\}_{i=1}^{n_p}$ from a distribution $P_{Z,Y}=P_Z\times P_{Y|Z}$ and an unlabeled target dataset $S_q=\{Z_{q,i}\}_{i=1}^{n_q}$ from the target distribution $Q_{Z,Y}=Q_Z\times P_{Y|Z}$ are observed, P_Z,Q_Z are the marginal distributions of Z in the source and target distribution respectively, and $P_{Y|Z}$ is the conditional distribution of Y given Z that is assumed the same across the source and target distributions. Given a miscoverage rate $0<\alpha<1$, the goal of conformal prediction under covariate shift is to construct a distribution-free prediction interval $\hat{C}(\cdot)\subseteq\mathbb{R}$ based on the observed datasets S_p^Y and S_q, such that for $(Z_0,Y_0)\sim Q_{Z,Y}$,

$$P\{Y_0\in\hat{C}(Z_0)\}\geq 1-\alpha+o(1),$$

where $o(1)$ means converging to zero as $\min\{n_q,n_p\}$ tends to ∞. Here $o(1)$ is involved as it is impossible to construct prediction intervals that are both nontrivial and distribution-free with coverage at least $(1-\alpha)\%$ when the ratio of the source and target covariate densities is unknown and need to be be estimated; see Lemma 1 in Qiu et al. (2023) and Theorem 1 in Yang et al. (2024). In this context, a nontrivial interval should have a finite length. To construct such a prediction interval, weighted conformal prediction methods based on density ratio estimation have been proposed and further developed by Tibshirani et al. (2019), Lei and Candès (2021), Ai and Ren (2024), etc. Under certain regularity conditions, Proposition 1 with $r=2$ in Lei and Candès (2021) tells that for a prediction interval $\hat{C}(z)\subseteq\mathbb{R}$ provided by weighted split conformal prediction with a density ratio estimator $\hat{R}$ based on $S_p=\{Z_{p,i}\}_{i=1}^{n_p}$ and S_q, the following relationship holds: For $(Z_0,Y_0)\sim Q_{Z,Y}$,

$$1-\alpha-e_{n_p,n_q}\leq P\{Y_0\in\hat{C}(Z_0)\}\leq 1-\alpha+\frac{c}{\sqrt{m}}+e_{n_p,n_q},$$

where c is a positive constant, m is the size of the calibration set, $e_{n_p,n_q} = E_{S_p,S_q}\|\widehat{R}-R^*\|_{p,1}$ and $\|f\|_{p,1} := E_{p^*}|f(Z)|$. Clearly,

$$e_{n_p,n_q} = E_{S_p,S_q}\|\widehat{R}-R^*\|_{p,1} \leq E_{S_p,S_q}\|\widehat{R}-R^*\|_{\max} \leq \left(E_{S_p,S_q}\|\widehat{R}-R^*\|_{\max}^2\right)^{\frac{1}{2}}.$$

Hence, the bound on $E_{S_p,S_q}\|\widehat{R}-R^*\|_{\max}^2$ offers an assessment of the extent to which the coverage of the constructed prediction interval deviates from the target coverage level of $1-\alpha$ in the average sense.

3.2 Exact non-asymptotic error bounds

In this subsection, we focus on target density ratio belonging to the Hölder function class, a commonly-used assumption for density functions (Devroye and Lugosi, 2001; Wasserman, 2006; Tsybakov, 2008; Kandasamy et al., 2015). We apply the results in Theorem 6 and Corollary 7 to analyze the theoretical properties of the BDD estimator under Hölder smoothness condition. In particular, we will prove that the BDD estimate achieves the minimax optimal convergence rate for the class of Hölder continuous density ratios. We first give the definition of a Hölder class.

Definition 8 (Hölder class) *A Hölder class $\mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$ consists of function $f : [0,1]^d \to \mathbb{R}$ satisfying*

$$\max_{\|\boldsymbol{\alpha}\|_1 \leq k} \|\partial^{\boldsymbol{\alpha}} f\|_\infty \leq M, \max_{\|\boldsymbol{\alpha}\|_1 = k} \max_{x \neq y} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(y)|}{\|x-y\|_2^a} \leq M,$$

where $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^d \alpha_i$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1}\partial^{\alpha_2}\cdots\partial^{\alpha_d}$ for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^{+d}$.

By Corollary 7, it suffices to bound the stochastic error $\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n/n$ and the approximation error $\|D_{\mathrm{NN}} - D^*\|_{\max}^2$. It follows from Theorem 6 in Bartlett et al. (2019) that, for $\mathcal{F}_{\mathrm{FNN}} = \mathcal{F}_{M,\mathcal{D},\mathcal{W},\mathcal{U},\mathcal{S}}$, there exists a universal constant C_2 such that $\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \leq C_2\mathcal{S}\mathcal{D}\log\mathcal{S}$ and the stochastic error can be well controlled then. As for the approximation error $\|D_{\mathrm{NN}}-D^*\|_{\max}^2$, we use Theorem 3.3 in Jiao et al. (2023a) to bound it. For convenience, we include this result in the following lemma. We specify the width $\mathcal{W}$ and depth $\mathcal{D}$ as follows. For any $K, L \in \mathbb{N}^+$,

$$\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} L\lceil\log_2(8L)\rceil, \tag{10}$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 K\lceil\log_2(8K)\rceil, \tag{11}$$

where $\lfloor a\rfloor$ is the largest integer no greater than a and $\lceil a\rceil$ is the smallest integer no less than a.

Lemma 9 (Approximation error) *Assume $f \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$. Then there exists a function ϕ_0 implemented by a ReLU network with width $\mathcal{W}$ and depth $\mathcal{D}$ specified in (10) and (11) such that*

$$\sup_{x\in[0,1]^d\setminus H_{B,\delta}} |f - \phi_0| \leq 18MC_\beta(KL)^{-\frac{2\beta}{d}},$$

where $C_\beta = (\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}$, $H_{B,\delta} = \cup_{i=1}^{d}\{x = [x_1, \ldots, x_d] : x_i \in \cup_{b=1}^{B-1} (b/B - \delta, b/B)\}$ *for* $B = \lceil (KL)^{2/d} \rceil, \delta \in (0, 1/(3B)]$ *and* $a \vee b = \max(a, b)$.

Furthermore, if

$$\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} 3^d L \lceil \log_2(8L) \rceil,$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 K \lceil \log_2(8K) \rceil + 2d,$$

then

$$\sup_{x\in[0,1]^d} |f - \phi_0| \leq 19 M C_\beta (KL)^{-\frac{2\beta}{d}}.$$

In this uniform approximation result, the width $\mathcal{W}$ *is required to depend on d exponentially.*

The following theorem gives an error bound for $\widehat{D}$ when D^* is a Hölder function.

Theorem 10 (Non-asymptotic error bound for $\widehat{D}$) *Suppose that Assumptions 4 and 5 are satisfied,* $D^* \in \mathcal{H}^\beta([0, 1]^d, M)$ *with* $\beta = k + a$ *where* $k \in \mathbb{N}^+$ *and* $a \in (0, 1]$, *and* $\mathcal{F}_{\mathrm{FNN}}$ *is the function class of ReLU DNNs with width* $\mathcal{W}$ *and depth* $\mathcal{D}$ *specified in (10) and (11). Then, for* $M \geq 1$ *and* $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$, *we have*

$$E_{S_p,S_q} \|\widehat{D} - D^*\|^2_{\max} \leq C \left(\xi_n^2 + C_1 (KL)^{-\frac{4\beta}{d}} \right),$$

where $C_1 = (\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee 1)}$ *and the constant* C *depends only on* (μ, σ, M).

Furthermore, if

$$\mathcal{W} = 114(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1},$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \left\lceil n^{\frac{d}{2(d+2\beta)}} \log_2\left(8 n^{\frac{d}{2(d+2\beta)}}\right) \right\rceil,$$

then

$$E_{S_p,S_q} \|\widehat{D} - D^*\|^2_{\max} \leq C_0 (\lfloor\beta\rfloor + 1)^9 d^{2\lfloor\beta\rfloor+(\beta\vee 3)} n^{-\frac{2\beta}{d+2\beta}} \log^3 n, \tag{12}$$

where the constant C_0 *depends only on* (μ, σ, M).

The convergence rate in (12) is optimal (up to some logarithmic factor). This can be seen by considering a density estimation problem with i.i.d observations $S_q^{(1)} = \{Z_{q,i}^{(1)}\}_{i=1}^{m_q}$ from an underlying unknown density q_1 on $[0, 1]^d$. To estimate q_1, we sample referencing observations $S_p^{(1)} = \{Z_{p,j}^{(1)}\}_{j=1}^{m_p}$ with $m_p \geq m_q$, from a uniform distribution $\mathrm{Unif}([0, 1]^d)$ whose density $p_1 \equiv 1$. Thus, estimating the density ratio q_1/p_1 is equivalent to estimating q_1. According to (5), we obtain the estimator $\hat{q}_1$ of q_1. If $\log q_1 \in \mathcal{H}^\beta([0, 1]^d, M)$ where $\beta = k + a$ with $k \in \mathbb{N}^+$ and $a \in (0, 1]$, a neural estimator based on the network structure specified in Theorem 10 satisfies

$$E_{S_p^{(1)}, S_q^{(1)}} \|\hat{q}_1 - q_1\|^2_{\max} \leq C_0 (\lfloor\beta\rfloor + 1)^9 d^{2\lfloor\beta\rfloor+(\beta\vee 3)} m_q^{-\frac{2\beta}{d+2\beta}} \log^3 m_q. \tag{13}$$

Tsybakov (2008) showed that for a density belonging to the Hölder function class, the optimal minimax rate of the density estimation is $O_p\left(m_q^{-2\beta/(d+2\beta)}\right)$. Hence, our estimator achieves the optimal minimax rate up to some logarithmic factor.

In addition, the existing error bounds usually contain a prefactor depending on the dimension d exponentially, e.g. 2^d (Devroye and Lugosi, 1996). Such a prefactor can be very large even for a moderately large d, which severely degrades the quality of an error bound. The prefactors in our results depend on d only polynomially and are much smaller than those in the existing bounds.

Theorem 10 is not a simple application of Lemma 9 (Theorem 3.3, Jiao et al., 2023b), as it requires meticulous calculations and careful balance between the approximation error and stochastic error to achieve the optimal convergence rate. In particular, the selection of the network configuration specified in Theorem 10 is intricate and the calculation of the associated stochastic error is rather complicated. For more details on the stochastic error bound in this scenario, please refer to (43) in the Appendix.

Remark 1 *In Appendix B we provide some examples of p^* and q^* such that $D^* = \log(q^*/p^*) \in \mathcal{H}^\beta([0,1]^d, M)$.*

3.3 An application to the estimation of KL divergence

In this subsection, we apply our main results to the problem of estimating the Kullback-Leibler (KL) divergence. The KL divergence occupies a central position in information theory and statistics. For example, in the asymptotic analysis of hypothesis testing, the KL divergence controls the rates at which error probabilities diminish. We propose an estimator of the KL divergence that is asymptotically normal based on our convergence results for BDD density-ratio estimator and data-splitting.

For two distributions with densities q and p, their Kullback-Leibler (KL) divergence is defined as

$$\mathrm{KL}(q||p) := \int_{\mathcal{Z}} q(z) \log \frac{q(z)}{p(z)} dz.$$

The KL divergence between two probability distributions plays an essential role in information theory, machine learning and statistics. Based on the variational representation of KL divergence, Nguyen et al. (2010) proposed two KL divergence estimators by maximizing some empirical objective functions over certain function class $\mathcal{R}$. The following three conditions are assumed in Nguyen et al. (2010):

1. The target density-ratio $R^* \in \mathcal{R}$, where $\mathcal{R}$ is the function class actually used in computation. This assumption implies that there is no approximation error;

2. There exist constants $0 < K_1 \leq K_2 < \infty$ such that for any $R \in \mathcal{R}$, it holds that $K_1 \leq \inf_{z\in\mathcal{Z}} R(z) \leq \sup_{z\in\mathcal{Z}} R(z) \leq K_2$;

3. The bracketing entropy of $\mathcal{R}$ satisfies $H_{[]}(\delta, \mathcal{R}, \|\cdot\|_{q^*}) = O(\delta^{-r})$ for some constant $r > 0$.

Nguyen et al. (2010) proved that their estimators are $\sqrt{n}$-consistent. However, their results are not applicable if the target R^* does not belong to the function class $\mathcal{R}$ used in the computation.

Suppose we have random samples $\{Z_{q,i}\}_{i=1}^{n_q}$ i.i.d. q^* and $\{Z_{p,i}\}_{i=1}^{n_q}$ i.i.d. p^*, and an estimator of the log-density ratio $\widehat{D}$ as defined in (5). Then a usual estimator of $\mathrm{KL}(q^*||p^*)$ is $\widehat{\mathrm{KL}}(q^*||p^*) = (1/n_q)\sum_{i=1}^{n_q} \widehat{D}(Z_{q,i})$. It follows from Theorem 6 and the proof of Theorem

10 that

$$\left|\widehat{\mathrm{KL}}(q^*||p^*) - \mathrm{KL}(q^*||p^*)\right| = O_p\left(n^{-\frac{\beta}{d+2\beta}} \log^{\frac{3}{2}} n\right). \tag{14}$$

A detailed proof of (14) is provided in Appendix A. Note that we only assume $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ instead of $D^* \in \mathcal{F}_{\mathrm{FNN}}$, but we can still show that the estimated KL-divergence $\widehat{\mathrm{KL}}(q^*||p^*)$ converges to the true KL-divergence $\mathrm{KL}(q^*||p^*)$ after a careful analysis of the approximation error induced by neural networks. However, in high-dimensional settings, $\widehat{\mathrm{KL}}(q^*||p^*)$ generally does not have a usual asymptotically normal distribution with root-n rate of convergence and its asymptotic distribution is unknown, this is because the convergence rate of $\widehat{D}$ may not reach $o(n^{-1/4})$, typically required for estimating a smooth functional of a nonparametric estimator (Bickle et al., 1998). Therefore, we cannot use this estimator directly for making statistical inference, such as constructing confidence intervals for $\mathrm{KL}(q^*||p^*)$.

Suppose we have observations $\{\tilde{Z}_{q,1}, \ldots, \tilde{Z}_{q,m}\}$ from q^* that are independent of the samples used in estimating D^*. For example, these independent observations can be obtained by splitting the existing samples $\{Z_{q,1}, \ldots, Z_{q,n_q}\}$ from q^* into a training set and a 'test set'. A new estimator for $\mathrm{KL}(q^*||p^*)$ can be defined based on the independent data:

$$\widetilde{\mathrm{KL}}_m(q^*||p^*) = \frac{1}{m}\sum_{i=1}^{m} \widehat{D}(\tilde{Z}_{q,i}). \tag{15}$$

Theorem 11 *Assume $m = o(n^{2\beta/(d+2\beta)} \log^3 n) \to \infty$ as $n \to \infty$, $E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{\max} = O(n^{-2\beta/(d+2\beta)} \log^3 n)$ and $E_{q^*} D^{*2}(Z) < \infty$. Then,*

$$\sqrt{m}\left(\widetilde{\mathrm{KL}}_m(q^*\|p^*) - \mathrm{KL}(q^*||p^*)\right) \to N(0, \sigma^2)$$

in distribution, where $\sigma^2 = E_{q^*} D^{*2}(Z) - E^2_{q^*} D^*(Z)$. The variance σ^2 can be consistently estimated by $\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^m \widehat{D}^2(\tilde{Z}_{q,i}) - \left[\frac{1}{m}\sum_{i=1}^m \widehat{D}(\tilde{Z}_{q,i})\right]^2$.

Confidence intervals for $\mathrm{KL}(q^*\|p^*)$ can be constructed based on this asymptotic normality result. We provide a simulation experiment to illustrate Theorem 11 in subsection 5.2. An important feature of Theorem 11 is that the convergence rate is root-m, where m is the size of the test sample. The requirement that $m = o(n^{2\beta/(d+2\beta)} \log^3 n) \to \infty$ is due to the fact that the optimal convergence rate of the density-ratio estimator is $O(n^{-2\beta/(d+2\beta)} \log^3 n)$ (Theorem 10). In practice, the prior knowledge in β can help decide the test sample size m. For example, in semi-parametric inference problems, $\beta > d/2$ is often assumed to establish asymptotic normality (Bickle et al., 1998); if we also assume $\beta > d/2$, then $m = o(n^{1/2}) \to \infty$ satisfies the requirement for m in Theorem 11. In subsection 5.2, we conduct a simulation study to illustrate Theorem 11.

Remark 2 *It can be easily shown that results similar to Theorem 11 hold for the f-divergence, which is $\mathcal{D}_f(q||p) := \int_{\mathcal{Z}} p(z) f\left(q(z)/p(z)\right) dz$, if f has a bounded first-order derivative on its domain.*

4. Some theoretical extensions

In this section, we consider several extensions to relax some conditions of the main results in 3. These extensions include: (1) Extension to cases with unbounded support; (2) Extension to mitigate the curse of dimensionality; (3) Extension to cases with unbounded density ratio.

4.1 Extension to the case of unbounded support

Recall that the densities q^* and p^* are assumed to have the same support $[0,1]^d$. Such a hypercube assumption is made for technical convenience. In fact, our Theorem 6, Corollary 7 and Theorem 10 do not rely on the hypercube assumption. In practice, a density ratio with unbounded support can be often encountered, necessitating a relaxation of the hypercube assumption to accommodate scenarios with unbounded support. To this end, we only need to study the upper bound for the approximation error $\|D_{\mathrm{NN}} - D^*\|_{\max}$ more carefully. With unbounded support, we may bound $\|D_{\mathrm{NN}} - D^*\|_{\max}$ by the support truncation technique under some suitable additional assumptions, at a small price of a slightly larger logarithm term in the error bound.

Specifically, suppose that the p.d.f's are supported on $\mathbb{R}^d$, to bound the approximation error as in Theorem 10, in addition to Assumptions 4 and 5 and the Hölder class assumption, we need to impose the following assumption on the tail probabilities.

Assumption 12 *There exist two positive constants c and C such that*

$$\max(E_{p^*}\mathbb{I}\{\|Z\|_\infty \geq C\log n\},\ E_{q^*}\mathbb{I}\{\|Z\|_\infty \geq C\log n\}) \leq cn^{-\frac{2\beta}{d+2\beta}}.$$

Examples satisfying Assumption 12 include sub-Gaussian variables e.g. normal random variables and sub-exponential variables e.g. exponential random variables and χ^2-random variables etc. We provide a brief proof about this fact in Appendix C. The next theorem provides an extended result for target density ratio with an unbounded support.

Theorem 13 (Non-asymptotic error bound with an unbounded support) *Suppose that Assumptions 4, 5 and 12 are satisfied, $D^* \in \mathcal{H}^\beta(\mathbb{R}^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$, and $\mathcal{F}_{\mathrm{FNN}}$ is the function class of ReLU DNNs with width $\mathcal{W}$ and depth $\mathcal{D}$ specified as below,*

$$\begin{aligned}\mathcal{W} &= 114(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1},\\ \mathcal{D} &= 21(\lfloor\beta\rfloor + 1)^2 \left\lceil n^{\frac{d}{2(d+2\beta)}} \log_2\left(8n^{\frac{d}{2(d+2\beta)}}\right)\right\rceil.\end{aligned}$$

Then, for $M \geq 1$ and $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$,

$$E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{\max} \leq C_0(\lfloor\beta\rfloor + 1)^9 d^{2\lfloor\beta\rfloor+(\beta\vee 3)} n^{-\frac{2\beta}{d+2\beta}}(2\log^{2\lfloor\beta\rfloor\vee 3} n), \tag{16}$$

where the constant C_0 depends only on (μ, σ, M).

The proof strategy for Theorem 13 is based on truncating the input data to confine it within a bounded set $(\{z : \|z\|_\infty \leq C\log n\})$ and controlling the associated errors for data points lying outside this bounded set through the tail probability delineated in Assumption 12. Compared with the upper bound of the approximation error in Theorem 10, when the p.d.f's are supported on $\mathbb{R}^d$, the approximation error bound has an extra logarithmic factor $(2\log n)^{2\lfloor\beta\rfloor}$.

4.2 Extension to mitigate the curse of dimensionality

In many modern statistical and machine learning tasks, such as image processing and text analysis, the dimensionality d of the data can be high, which results in a very slow convergence rate even with a large sample size. This is known as the curse of dimensionality. Nonetheless, the data in various applications has been demonstrated to be supported or approximately supported in some subspaces or subsets with low intrinsic dimensionality (Nakada and Imaizumi, 2020). For regression problems, Nakada and Imaizumi (2020) have shown that DNNs can adaptively estimate the regression function through the low-dimensional structure of the data, and the resulting convergence rates no longer depend on the nominal high dimensionality d of the data, but on its low intrinsic dimension.

Motivated by these advancements, we assume that the data is concentrated on an approximate compact Riemannian submanifold $\mathcal{M}$ with the Riemannian dimension $d_{\mathcal{M}} \ll d$.

Assumption 14 *The target log-density ratio $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$, and the data from the densities p^*, q^* are concentrated on a set $\mathcal{M}_{\rho,\tau,V,\mathcal{R}} \subseteq [0,1]^d$ defined as*

$$\mathcal{M}_{\rho,\tau,V,\mathcal{R}} := \{x \in [0,1]^d : \text{ there exists } y \in \mathcal{M}_{\tau,V,\mathcal{R}},\ \|x-y\|_2 \leq \rho\},$$

where $\mathcal{M}_{\tau,V,\mathcal{R}}$ is a compact $d_{\mathcal{M}}$-dimensional Riemannian submanifold having condition number $1/\tau$, volume V, and geodesic covering regularity $\mathcal{R}$ and $\rho \in (0,1)$.

The next theorem provides a non-asymptotic error bound depending only on the intrinsic dimension.

Theorem 15 *Assume Assumptions 4, 5 and 14 hold. Suppose that $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$, $k \in \mathbb{N}^+$ and $a \in (0,1]$. If $\mathcal{F}_{\mathrm{FNN}}$ is the function class of ReLU FNNs with width and depth*

$$\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 {d_\delta}^{\lfloor\beta\rfloor+1} L\lceil\log_2(8L)\rceil$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 K\lceil\log_2(8K)\rceil,$$

where $K, L \in \mathbb{N}^+$ and $d_\delta = O\left(d_{\mathcal{M}}\log(dV\mathcal{R}\tau^{-1}/\delta)/\delta^2\right) \ll d$, then when $M \geq 1$, $n > \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$ and

$$\rho \leq (\lfloor\beta\rfloor + 1)^2 2^\beta d^{\beta-\frac{1}{2}} d_\delta^{\lfloor\beta\rfloor+(\beta-1/2)\vee(1/2)} (KL)^{-\frac{2\beta}{d_\delta}},$$

we have

$$E_{S_p,S_q}\|\widehat{D} - D^*\|_{\max}^2 \leq C(1-\delta)^{-2\beta}\Big[\xi_n^2 + C_2(KL)^{-\frac{4\beta}{d_\delta}}\Big],$$

where the constant C only depends on (μ, σ, M), $C_2 = (\lfloor\beta\rfloor + 1)^4(2d)^{2\beta} d_\delta^{3\beta+(\beta\vee 1)}$, and ξ_n is defined in (9).

Remark 3 *d_δ is the dimension of the Euclidean space, where the low-dimensional Riemannian manifold is embedded and the distance between two embedded points approximately equals to the distance between their original points on the manifold with a relative error δ; see (48) in the supplementary materials for more details. In Theorem 15, the dependence on ρ has been absorbed in the convergence rate in Theorem 15 as we require $\rho \leq (\lfloor\beta\rfloor+1)^2 2^\beta d^{\beta-1/2} d_\delta^{\lfloor\beta\rfloor+(\beta-1/2)\vee(1/2)}(KL)^{-2\beta/d_\delta}$.*

By Theorem 15, if we set $\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 {d_\delta}^{\lfloor\beta\rfloor+1}, \mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 \left\lceil n^{\zeta_\delta}\log_2\left(8n^{\zeta_\delta}\right)\right\rceil$, with $\zeta_\delta = d_\delta/(2(d_\delta+2\beta))$, then

$$E_{S_p,S_q}\|\widehat{D}-D^*\|_{\max}^2 \leq C_0 C_3 (1-\delta)^{-2\beta} n^{-\frac{2\beta}{d_\delta+2\beta}}\log^3 n, \tag{17}$$

where the constants C_0 only depends on (μ,σ,M) and

$$C_3 = (\lfloor\beta\rfloor+1)^9 \max\{d_\delta^{2\lfloor\beta\rfloor+3}, (2d)^{2\beta} d_\delta^{3\beta+(\beta\vee 1)}\}.$$

The convergence rate $n^{-2\beta/(d_\delta+2\beta)}$ up to some logarithmic factor in (17) only depends on $d_\delta \ll d$, instead of the ambient dimension d. Therefore, Theorem 15 shows that a low-dimensional Riemannian submanifold support assumption can alleviate the curse of dimensionality.

4.3 Extension to the case with an unbounded density ratio

Our previous results in Theorem 6 and Corollary 7 are derived under the boundedness condition in Assumption 5, which is commonly made in the relevant literature (Sugiyama et al., 2012a; Kato and Teshima, 2021; Kato et al., 2023). It is, however, somewhat restrictive in density-ratio estimation problems. For instance, it may not be satisfied in the scenario with density-chasm problem, i.e., the gap between two densities is large (Rhodes et al., 2020). To relax such an assumption, we consider the following partially-relaxed version.

Assumption 16 *There exists a constant $0 < M < \infty$ such that $D^*(z) \geq -M$ for every $z \in \mathcal{Z}$ and $\|D\|_\infty \leq M$ for every $D \in \mathcal{F}_{\mathrm{FNN}}$.*

Assumption 16 does not require the target log-density ratio D^* to be bounded above. Denote the truncated D^* and R^* by

$$D_M^*(z) = D^*(z)\mathbb{I}\{D^*(z) \leq M\} + M\mathbb{I}\{D^*(z) > M\},$$
$$R_M^*(z) = R^*(z)\mathbb{I}\{R^*(z) \leq e^M\} + e^M\mathbb{I}\{R^*(z) > e^M\},$$

where $0 < M < \infty$ and $\mathbb{I}\{\cdot\}$ is the indicator function. The next theorem establishes a non-asymptotic error bound involving the truncation error under Assumption 16.

Theorem 17 *Suppose Assumptions 4 and 16 hold. When $n > \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$, there exists two constants C depending only on (μ,σ,M) and C_0 depending only on (μ,σ), such that*

$$E_{S_p,S_q}\|\widehat{D}-D^*\|_p^2 \leq C_0 e^{2M}\|R^*-R_M^*\|_p^2 + C\left(\xi_n + \inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\|D-D_M^*\|_p^2\right),$$

where ξ_n is defined in (9).

The term $\|R^* - R^*_M\|_p^2$ is the truncation error for an unbounded R^* and the unboundedness also leads to the term $\xi_n = [\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})(\log n)/n]^{1/2}$ in the error bound, which is greater than ξ_n^2 in the bounded case. However, because no boundedness assumption is needed in this theorem, we can apply it to study the convergence properties of the telescoping density-ratio estimator of Rhodes et al. (2020) in subsection 4.4 below.

Remark 4 *To the best of our knowledge, the tight bound for density-ratio estimation in unbounded setting is an open question. Even in density estimation problems, when the density is unbounded above, it remains unsolved what the tight bound is. Our result in Theorem 17, albeit not tight, is an attempt to tackle the challenging unbounded density-ratio problem.*

4.4 An application to analysis of the telescoping density-ratio estimator

In this subsection, we apply the extended result in Theorem 17 for unbounded density ratios to investigate the convergence properties of the telescoping density-ratio estimator proposed by Rhodes et al. (2020). By leveraging the main results in Theorem 6 and Theorem 17, we theoretically demonstrate the advantages of the telescoping estimator for unbounded density ratios. This analysis provides insights into the performance of the telescoping approach under more general conditions.

When the difference or the ‘gap’ between two densities is large, a single-ratio estimation method may perform poorly. This is referred to as the *density-chasm problem* (Rhodes et al., 2020). To alleviate this problem, Rhodes et al. (2020) proposed an approach called Telescoping density-Ratio Estimation (TRE). This approach first gradually transports samples from q^* to samples from p^*, creating a chain of intermediate datasets, then estimates the density ratio between consecutive datasets along this chain. The chained ratios are combined via a telescoping product which yields an estimate of the original density ratio. The experiments conducted by Rhodes et al. (2020) demonstrate that TRE can yield substantial improvements over existing single-ratio methods for mutual information estimation, representation learning and energy-based modelling.

We now provide a theoretical analysis of TRE, which partially explains why TRE performs well. For notational simplicity, suppose $n_p = n_q \equiv n$ below.

For $k = 0, 1, \dots, K$, Rhodes et al. (2020) constructed a chain of intermediate samples connecting q^* and p^* by setting $Z_{k,i} = (1-\alpha_k^2)^{1/2} Z_{q,i} + \alpha_k Z_{p,i},\ i = 1, \dots, n$, where $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = 1$, and used these samples to build a TRE.

To simplify the analysis, we use a slightly different chain of intermediate samples as follows. For $k = 0, 1, \dots, K$, let

$$Z_{k,i} = (1-\delta_{k,i}) Z_{q,i} + \delta_{k,i} Z_{p,i},\ i = 1, \dots, n, \tag{18}$$

where $\delta_{k,i}, i = 1, \dots, n$, are i.i.d. Bernoulli random variables with success probability α_k. Let q_k be the density of the synthetic data $Z_{k,i}$ constructed this way. We have $q_k(z) = (1-\alpha_k) q^*(z) + \alpha_k p^*(z), k = 1, \dots, K-1$. Therefore, the distribution of the samples from q_k in the chain is a simple mixture of q^* and p^* with the mixing proportions $1-\alpha_k$ and α_k, instead of a more complex convolution of two densities using the construction of Rhodes et al. (2020). As α_k changes from $\alpha_0 = 0$ to $\alpha_K = 1$ over a grid $\{\alpha_0, \alpha_1, \dots, \alpha_K\} \subset [0, 1]$, the distributions of the samples in the chain move gradually from q^* to p^*. Let $q_0 = q^*$ and $q_K \equiv p^*$. Then,

$$R^*(z) = \frac{q^*(z)}{p^*(z)} = \prod_{i=0}^{K-1} R_k^*(z),\ z \in \mathcal{Z}, \tag{19}$$

where $R_k^*(z) = q_k(z)/q_{k+1}(z)$. For $k = 0, 1, \ldots, K-1$, applying the neural density-ratio estimator with $\{Z_{k,j}\}_{j=1}^{n_k}$ and $\{Z_{k+1,j}\}_{j=1}^{n_{k+1}}$ yields an estimator $\widehat{R}_k$ of R_k^*. Then the telescoping density ratio estimator of R^* is $\prod_{k=0}^{K-1} \widehat{R}_k$.

We consider the log-density ratio. Let $\widehat{D}_k$ be the neural estimator of $D_k^* \equiv \log(q_k/q_{k+1})$. Based on (19), the telescoping estimator of the log-density ratio $D^* \equiv \log R^*$ is

$$\widehat{D}_{\mathrm{TRE}} = \sum_{k=0}^{K-1} \widehat{D}_k. \tag{20}$$

In what follows, we show that under certain conditions, the telescoping estimator has an improved asymptotic error bound. The intuition is as follows: when q_k/q_{k+1} is bounded or $q_k(z)/q_{k+1}(z) \ll q^*(z)/p^*(z)$ for $z \in \mathcal{Z}$, where q_k and q_{k+1} are the densities of the synthetic data $\{Z_{k,j}\}_{j=1}^{n}$ and $\{Z_{k+1,j}\}_{j=1}^{n}$, respectively, the truncation error for q_k/q_{k+1} vanishes or is far less than that for q^*/p^*. This can help the telescoping density-ratio estimator perform better than a single-ratio estimator.

Assume that $q^* \geq c_1$ and $c_1 \leq p^* \leq c_2$, where $0 < c_1, c_2 < \infty$ are two constants. Thus, $D^* = \log(q^*/p^*) \geq \log(c_1/c_2)$. So Assumption 16 is satisfied. For any finite set $\mathcal{A} \subset \mathbb{R}$, $\max \mathcal{A}$ denotes the maximal value in $\mathcal{A}$. Let $M = \log(c_2/c_1)$ and M satisfy

$$M \geq \max \mathcal{A}_{M,\alpha}^{(2K)}, \tag{21}$$

where

$$\begin{aligned}\mathcal{A}_{M,\alpha}^{(2K)} = \{1\} &\cup \left\{\log \frac{1-\alpha_{k-1}}{1-\alpha_k}, k = 1, \ldots, K-1\right\} \\ &\cup \left\{\log \frac{(1-e^{-M})\alpha_k + e^{-M}}{(1-e^{-M})\alpha_{k-1} + e^{-M}}, k = 1, \ldots, K-1\right\}.\end{aligned}$$

Based on Theorem 17, we can establish an asymptotic error bound for the telescoping estimator $\widehat{D}_{\mathrm{TRE}}$ defined in (20), with

$$\widehat{D}_k \in \operatorname*{argmin}_{D \in \mathcal{F}_{\mathrm{FNN}}} \hat{\mathcal{B}}_\psi^k(e^D),$$

where

$$\hat{\mathcal{B}}_\psi^k(e^D) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_1(D(Z_{k+1,i})) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_2(D(Z_{k,i})),$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are defined in (6) and $\mathcal{F}_{\mathrm{FNN}}$ consists of DNNs D with $\|D\|_\infty \leq M$. To demonstrate the advantages of the telescoping estimator, we also consider the single-ratio estimator (SRE), $\widehat{D}_{\mathrm{SRE}} \in \operatorname{argmin}_{D \in \mathcal{F}_{\mathrm{FNN}}} \hat{\mathcal{B}}_\psi(e^D)$.

Proposition 18 *Assume that* $q^* \geq c_1$, $c_1 \leq p^* \leq c_2$, *where the constants* $0 < c_1 \leq c_2 < \infty$, *and the samples* $\{Z_{q,i}\}_{i=1}^n$ *from* q^* *and* $\{Z_{p,j}\}_{j=1}^n$ *from* p^* *are independent. Then, there exists a constant* $C_0(\mu, \sigma, c_1)$ *depending only on* (μ, σ, c_1) *such that for*

$$B_{\text{SRE}} = e^M C_0(\mu, \sigma, c_1) \|R^* - R_M^*\|_p,$$

we have

$$\limsup_{n\to\infty} E_{S_p,S_q} \|\widehat{D}_{\text{SRE}} - D^*\|_2 \leq B_{\text{SRE}},$$
$$\limsup_{n\to\infty} E_{S_p,S_q} \|\widehat{D}_{\text{TRE}} - D^*\|_2 \leq (1 - \alpha_{K-1}) B_{\text{SRE}},$$

where $\|f\|_2 = \left[\int_{\mathcal{Z}} f^2(z)dz\right]^{1/2}$ *for any square integrable function* f.

Proposition 18 shows that for a given sequence $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = 1$ satisfying (21), the upper bound for the asymptotic L_2-error of $\widehat{D}_{\text{TRE}}$ is reduced by a factor $(1 - \alpha_{K-1})$ with $0 < 1 - \alpha_{K-1} < 1$. This upper bound can be far less than that of $\widehat{D}_{\text{SRE}}$ when α_{K-1} is close to 1. Therefore, TRE can improve the asymptotic error bound over the bound for the single-ratio method.

To realize (21), it generally requires that for $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = 1$, $\max_{k=1,\ldots,K}(\alpha_k - \alpha_{k-1})$ is small enough or otherwise $\max \mathcal{A}_{M,\alpha}^{(2K)} \gg M$. Hence, it often needs a chain with relatively large K to achieve better performance. Nonetheless, one intermediate distribution ($K = 2$) may still give improved results when M is large enough.

Proposition 18 is generally not applicable to the original chain of TRE. The difficulty is due to the possibly intensive oscillation of density ratios caused by the convolution form for the density of the sum of two random variables. We illustrate this by a toy example: suppose Z_q, Z_p are i.i.d. uniform random variables on $[0, 1]$. For any $t \in (0, 1/2]$, $(1-t)Z_q + tZ_p$ has density

$$q_t^*(z) = \frac{z}{t(1-t)}\mathbb{I}\{0 \leq z \leq t\} + \frac{1}{1-t}\mathbb{I}\{t<z \leq 1-t\} + \frac{1-z}{t(1-t)}\mathbb{I}\{1-t<z \leq 1\}.$$

In this case, q^*/q_t^* is unbounded and oscillates sharply when z is close to 0 or 1. This makes it hard to estimate q^*/q_t^*. However, the chain we used does not have this problem, which may be a good choice in practice. In subsections 5.4 & 5.5, we conduct some numerical experiments to compare the performance of the telescoping ratio estimates based on our proposed mixing chain and the original convolution chain.

5. Numerical studies

In this section, we conduct simulation studies to validate our main theoretical results, evaluate the performance of our proposed KL estimator, test the effect of (1-α_{K-1}) in the proposed mixing telescoping density-ratio estimator (mTRE) and compare mTRE to the convolution-based telescoping density-ratio estimator (cTRE) in Rhodes et al. (2020).

5.1 Simulation studies to validate the main theoretical results

In this subsection, we present numerical validations for Theorems 10 & 13. Note that Theorem 10 pertains to a special case of Theorem 6 and Corollary 7. The numerical verifications conducted for Theorem 10 also serve to demonstrate the validity of Theorem 6 and Corollary 7. Moreover, Theorem 13 is an extension of Theorem 6 for the unbounded support case.

In the first part, we conduct numerical validations for Theorem 10.

- Model (TN): Let $\mathrm{TN}_{d_0}(\Sigma, a)$ be a truncated d_0-dimensional Gaussian distribution with a mean $\mathbf{0} \in \mathbb{R}^{d_0}$, covariance matrix $\Sigma \in \mathbb{R}^{d_0 \times d_0}$ and a truncation level $a > 0$. That is, $V \sim \mathrm{TN}_d(\Sigma, a)$ implies that there exists a d_0-dimensional Gaussian random variable $U = \mathcal{N}(\mathbf{0}, \Sigma)$, such that $V = U\mathbb{I}(\|U\|_\infty \leq a)$. Let $Z = (Z_1, \ldots, Z_{2d_0})^\top \in \mathbb{R}^{2d_0}$ be the random vector of interest. For distribution p^*, $Z \sim \mathrm{TN}_{2d_0}(I_{2d_0}, a)$ and for distribution q^*, $Z \sim \mathrm{TN}_{2d_0}(\Sigma(\rho), a)$, where I_{2d_0} is the $2d_0 \times 2d_0$ identity matrix and $\Sigma(\rho) = (\sigma^\rho_{i,j}) \in \mathbb{R}^{2d_0 \times 2d_0}, \sigma^\rho_{i,j} = 1$ if $i = j$, $\sigma^\rho_{i,j} = \rho$ if $|i - j| = d_0$, and $\sigma^\rho_{i,j} = 0$ otherwise for $i, j = 1, 2, \ldots, 2d_0$.

  We discuss the Hölder properties of the truncated Gaussian distributions in details in Appendix B. In this example, the target log-density ratio has an infinite Hölder smoothness parameter ($\beta = \infty$) and has finite support. The parameters are specified as follows: $\rho = 0.1, d_0 = 2$ and $a = 1$.

- Model (NI): The second simulation below satisfies all the conditions outlined in Theorem 10. That is, the target log-density ratio has a finite Hölder smoothness parameter and finite support, different from the first model.

  Let X be a d_0-dimensional random vector. And

$$\begin{cases} q^* : Z = (Y, X), Y = Y_1\mathbb{I}\{a_1 \leq Y_1 \leq a_2\}, Y_1 = f_0(X) + \varepsilon, \\ p^* : Z = (Y, X), Y = Y_1\mathbb{I}\{a_1 \leq Y_1 \leq a_2\}, Y_1 = 1 + f_0(X) + \varepsilon, \end{cases} \tag{22}$$

  where $\varepsilon \sim \mathrm{TN}_1(1, 1)$ and a_1, a_2 are set to let q^* and p^* have the same support. We show at the end of Appendix A that

$$D^*(z) = \log \frac{q^*(y, x)}{p^*(y, x)} = f_0(x) - y + \frac{1}{2} + c \tag{23}$$

  for some constant c and any $z = (y, x)$, implying that D^* has a Hölder smoothness akin to that of $f_0(x)$.

  For this example, we set $d_0 = 3, a_1 = 1, a_2 = 2$, X follows a uniform distribution on $[0, 1]^{d_0}$ and

$$\begin{aligned} f_0(x) :&= (d_0 - 1)^{-1} \sum_{i=1}^{d_0 - 1} x_i x_{i+1} + d_0^{-1} \sum_{i=1}^{d_0} 2 \sin(2\pi x_i)\, \mathbb{I}\{x_i \leq 0.5\} \\ &+ d_0^{-1} \sum_{i=1}^{d_0} \left(4\pi(\sqrt{2} - 1)^{-1} \left(x_i - 2^{-1/2} \right)^2 - \pi(\sqrt{2} - 1) \right) \mathbb{I}\{x_i > 0.5\}. \end{aligned}$$

Notably, f_0 is a member of $\mathcal{H}^\beta([0,1]^{d_0}, M)$ with $\beta = 2$ and some constant M. Such a f_0 is also considered in the simulation studies in Nakada and Imaizumi (2020).

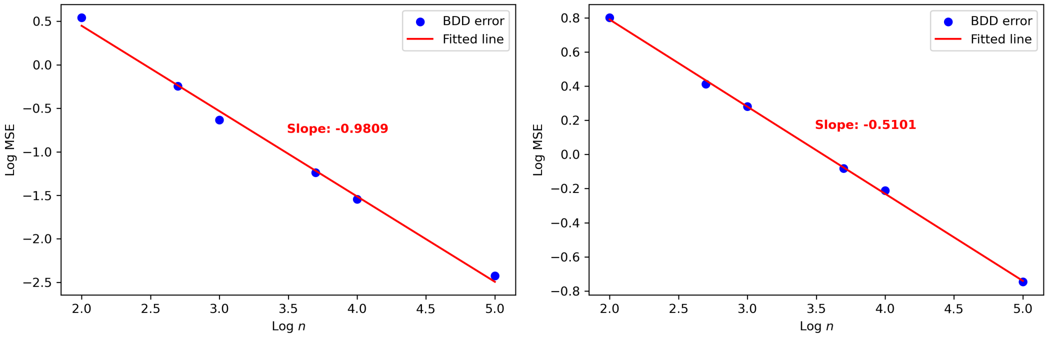
- Implementation details and results: For each model, we vary $n_p = n_q = n \in \{100, 500, 1000, 5000, 10000, 100000\}$ and each setup is repeated 10 times. The neural network architectures and additional implementation details can be found in Appendix G.1. The empirical mean squared error, denoted as $\widehat{\mathrm{MSE}}(\hat{D})$, is computed as detailed in Appendix G.1.

  The average MSEs and their SEs over 10 replications with different n under Models (TN) & (NI) are summarized in Table 2. From Table 2, it can be seen that the empirical MSEs decrease as the training sample size increases. According to Theorem 10, $\log \mathrm{MSE}(n) \approx (-2\beta/(2\beta+d)) \log n + C$ for some constant C. Denote the theoretical convergence factor as $\mathrm{CF} = -2\beta/(2\beta+d)$. We then fit a simple linear regression based on the data in Table 2 for Models (TN) and (NI) separately. In the linear regression, the response variable is the logarithm of the estimated MSE, denoted by $\log \widehat{\mathrm{MSE}}(\hat{D})$, and $\log n$ is the predictor. The log-transformed data and the fitted lines are displayed in Figure 1.

  For Model (TN) with $d = 2d_0 = 4$ and $\beta = \infty$, the theoretical convergence factor can be computed to be $\mathrm{CF} = -1$. And the empirical CF shown in Figure 1 is around -0.9809. For Model (NI) with $d = d_0 + 1 = 4$ and $\beta = 2$, the theoretical convergence factor is $\mathrm{CF} = -2\beta/(2\beta + d) = -1/2$. And the empirical CF in Figure 1 is around -0.5101. These empirical results provide supporting evidence for the validity of our Theorem 10.

Table 2: The average MSEs of the proposed deep log density ratio estimate (BDD) and the corresponding simulation standard errors (in parentheses) over 10 replications for verifying Theorem 10.

Model n	Model (TN)	Model (NI)
100	3.48502(0.69336)	6.32138(1.28762)
500	0.56943(0.12371)	2.58476(0.46108)
1000	0.23291(0.03584)	1.90831(0.53766)
5000	0.05760(0.00893)	0.82870(0.30309)
10000	0.02860(0.00540)	0.61421(0.13418)
100000	0.00376(0.00009)	0.17945(0.00595)



(a) Model (TN)

(b) Model (NI)

Figure 1: The log-transformed data based on Table 2, and the fitted lines based on the linear regression model, along with their slopes. The theoretical convergence factors for Models (TN) & (NI) are -1 and $-1/2$, respectively.

In the second part, we conduct numerical validations for Theorem 13, which underscore the importance of Assumption 12 in Theorem 13.

- Simulation model setup: We consider the data generating model in (22) with $d_0 = 1$ and $f_0(x) = \sin(|x|)$. In view of the non-differentiability of $|\cdot|$ and the Lipschitz continuity of f_0 with a constant of 1, we know $f_0 \in \mathcal{H}^\beta(\mathbb{R}, 1)$ with $\beta = 1$. Note that $Z = (Y, X)$. We consider two settings of X:

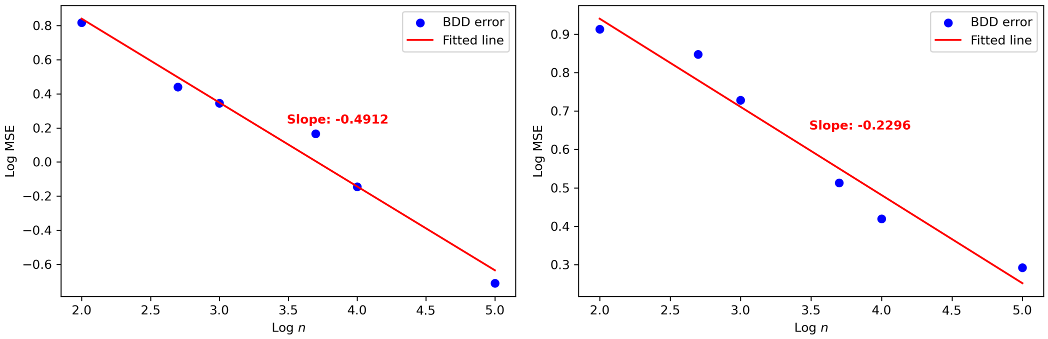
$$\begin{cases} X \text{ follows the standard normal distribution}, a_1 = 0.8, a_2 = 1.8 \\ X \text{ follows Student-t}(2), a_1 = 0.5, a_2 = 1.5 \end{cases},$$

  where Student-t(2) means the Student's-t distribution with degrees of freedom 2. When X follows normal distribution, $Z = (Y, X)$ (if Y is bounded) satisfies Assumption 12; when X follows Student-t(2), Assumption 12 is violated since random variables satisfy Assumption 12 at least have finite moments, while Student-t(2) random variable does not have finite variance. Other simulation implementation details are similar to those for Model (NI) in the first part; see Appendix G.1.

- Results: The average MSEs and their SEs (over 10 replications) with varying sample sizes n are summarized in Table 3. When Assumption 12 holds, the theoretical convergence factor CF $= -2\beta/(2\beta + d) = -1/2$ since $d = d_0 + 1 = 2$ and $\beta = 1$. The log-transformed empirical data in Table 3 and the fitted lines based on a linear regression model are shown in Figure 2. The empirical CF for normally distributed X in Figure 2 is around -0.4912, very close to the theoretical value of $-1/2$. Conversely, when X follows Student-t(2), the empirical CF is computed as -0.2296, indicating a decelerated convergence rate when Assumption 12 is violated.

Table 3: The average MSEs of the proposed deep log density ratio estimate (BDD) and the corresponding simulation standard errors (in parentheses) over 10 replications for verifying Theorem 13.

Model n	Normal	Student-t(2)
100	6.54752(3.75635)	8.19044(3.28793)
500	2.74838(1.54536)	7.04476(1.91902)
1000	2.21367(1.79875)	5.34623(2.25564)
5000	1.46320(0.37831)	3.25975(1.12520)
10000	0.71740(0.34525)	2.62748(0.92486)
100000	0.19501(0.00879)	1.95932(0.26622)



(a) N(0,1)

(b) Student-t(2)

Figure 2: The log-transformed data in Table 2, and the fitted lines based on a linear regression model, along with their slopes. The theoretical convergence factor is $-1/2$.

5.2 Simulation studies to examine KL divergence estimation

In this subsection, we use a simulation experiment to examine the asymptotic normality of KL divergence estimation in Theorem 11. Our simulation setting is as follows: Let $Z = (Z_1, Z_2, \ldots, Z_{d_0}, Z_{d_0+1}, Z_{d_0+2}, \ldots, Z_{2d_0})^\top \in \mathbb{R}^{2d_0}$ be some random vector of interest. For distribution p, $Z \sim N(0, I_{2d_0})$ and for distribution q, $Z \sim N(0, \Sigma(\rho))$, where I_{2d_0} and $\Sigma(\rho) = (\sigma^{\rho}_{i,j})$ are defined in the last subsection.

We calculate the bias magnitude (Bias), the standard errors of the KL divergence estimate in (15) across the replicated samples (SE), the estimated standard deviation (SD) based on the asymptotic normality in Theorem 11, and the actual confidence interval coverage under the nominal coverage of 0.95 (CP). The simulation results for $d = 1, 5$, $\rho = 0.4, 0.7$ and $m = 100, 200, 400$, where m is the testing sample size in Theorem 11, are reported in Table 4. These reported results are based on 200 simulation repetitions. From Table 4, we see that relatively small testing data sample size $m = 100$ produces the satisfactory actual coverage, and relatively large m, such as $m = 200$ or 400, may damage the actual coverage of the constructed confidence interval through Theorem 11. This coincides with the

Table 4: Simulation results for the KL divergence estimation

d	n	ρ	m	Bias	SE	SD	CP
1	5000	0.4	100	-0.0024	0.0414	0.0385	0.915
			200	-0.0064	0.0286	0.0274	0.945
			400	-0.0061	0.0206	0.0193	0.895
		0.7	100	-0.0022	0.0684	0.0655	0.930
			200	-0.0093	0.0446	0.0470	0.945
			400	-0.0090	0.0349	0.0337	0.930
	10000	0.4	100	-0.0073	0.0391	0.0383	0.960
			200	-0.0043	0.0303	0.0276	0.925
			400	-0.0025	0.0220	0.0195	0.910
		0.7	100	-0.0092	0.0707	0.0664	0.930
			200	-0.0077	0.0481	0.0473	0.950
			400	-0.0092	0.0372	0.0332	0.895
5	5000	0.4	100	-0.0215	0.0934	0.0882	0.935
			200	-0.0260	0.0665	0.0631	0.915
			400	-0.0320	0.0583	0.0444	0.805
		0.7	100	-0.0661	0.1633	0.1513	0.910
			200	-0.0325	0.1329	0.1064	0.865
			400	-0.0457	0.0948	0.0750	0.830
	10000	0.4	100	-0.0184	0.0887	0.0876	0.935
			200	-0.0179	0.0673	0.0620	0.920
			400	-0.0222	0.0550	0.0441	0.880
		0.7	100	-0.0421	0.1607	0.1467	0.900
			200	-0.0427	0.1155	0.1059	0.920
			400	-0.0423	0.0919	0.0744	0.850

requirement for m in Theorem 11, as we have seen that m cannot be too large in Theorem 11. All related implementation details can be found in Appendix G.2.

5.3 Simulation studies to check the effect of (1-α_{K-1}) in mTRE

In this subsection, we conduct simulations to examine the effect of $(1-\alpha_{K-1})$. The data generating model is as follows: Let $Z = (Z_1, Z_2, \ldots, Z_d)^\top \in \mathbb{R}^d$ be a random vector, where $Z_1, Z_2, \ldots, Z_d$ are i.i.d. random variables following distributions as below:

- For q^*, $Z_i = \sqrt{U_i}, U_i \sim \text{Uniform}([0,1]), i = 1, \ldots, d$. It can be easily checked that $q^*(z) = (2^d \prod_{i=1}^d \sqrt{z_i})^{-1} \geq 2^{-d}$ for any $z = (z_1, z_2, \ldots, z_d)^\top \in [0,1]^d$ and $\lim_{\|z\|_2 \to 0} q^*(z) = \infty$;

- For p^*, $Z_i = (U_i + U_i^2)/2, U_i \sim \text{Uniform}([0,1]), i = 1, \ldots, d$. Similarly, one can easily show that $p^*(z) = \prod_{i=1}^d (z_i + 1/2)$ and thus, $2^{-d} \leq p^*(z) \leq (3/2)^d$.

Such a simulation model meets the conditions in Proposition 18. In each experimental iteration, we sample 2000 data points from q^* and p^* respectively to facilitate the training of the neural network and the subsequent density ratio estimation. For other implementation details, we refer readers to Appendix G.3 of the paper. The experiments are conducted for $d = 1$ and $d = 5$, respectively, with each configuration repeated 10 times. The average MSEs (over 10 replications) and their respective SEs for varying values of α_{K-1} are summarized in Table 5.

Table 5: The average MSEs of the estimated TRE and the corresponding simulation standard errors (in parentheses) over 10 replications.

d α_{K-1}	1	5
0.9	$1.473\times10^{1}(1.994\times10^{-1})$	$5.296\times10^{13}(1.498\times10^{14})$
0.99	$2.255\times10^{0}(3.931\times10^{-2})$	$1.016\times10^{5}(4.460\times10^{1})$
0.999	$5.170\times10^{-2}(7.967\times10^{-4})$	$6.820\times10^{4}(2.553\times10^{1})$
0.9999	$5.765\times10^{-4}(1.440\times10^{-5})$	$1.601\times10^{4}(4.727\times10^{0})$
0.99999	$5.741\times10^{-6}(1.467\times10^{-6})$	$5.069\times10^{2}(1.281\times10^{0})$
0.999999	$1.873\times10^{-7}(2.425\times10^{-7})$	$6.039\times10^{0}(1.489\times10^{-1})$
0.9999999	$1.542\times10^{-7}(1.943\times10^{-7})$	$6.689\times10^{-2}(1.513\times10^{-2})$

Table 5 tells that the estimation errors decrease as α_{K-1} approaches 1, which aligns with the theoretical results in Proposition 18. Meanwhile, the density ratio estimation is quite challenging in this simulation model, as the density ratio has a non-vanishing probability at extremely large values, especially when the dimension d is higher; see the simulation results for $d = 5$ and relatively small α_{K-1} in Table 5. Nevertheless, the proposed telescoping idea proves instrumental in mitigating this challenge and enhancing the density ratio estimation.

5.4 Simulation on comparison of mTRE and cTRE

In this subsection, we conduct simulation studies to evaluate the performance of the telescoping ratio estimates (TREs) based on our proposed mixing chain (denoted by mTRE) and the original convolution chain (denoted by cTRE). Our simulation settings are as follows.

- Beta setting: Let $Z = (Z_1, Z_2, \ldots, Z_d)^\top \in \mathbb{R}^d$ be a random vector of interest, where $Z_1, Z_2, \ldots, Z_d$ are i.i.d. random variables following Beta distribution, denoted by Beta(α, β). Set p^* as the p.d.f of Z with Beta$(2.2, 1.5)$ and q^* as the p.d.f of Z with Beta$(2, 2)$. In this setting, we set $d = 5$.

- Normal setting: Let $Z \in \mathbb{R}^{2d_0}$ be some random vector of interest generated as in subsection 5.2. But in this setting, $d_0 = 5$ and $\rho = 0.9$.

After obtaining a log density-ratio estimate $\hat{D}$, suppose we have n_t testing samples $\{Z_{t,i}\}_{i=1}^{n_t}$ from q^*, we calculate the MSE between the log density-ratio estimate $\hat{D}$ and its true value

Table 6: The MSEs averaged over 10 replications and the corresponding standard errors in parentheses for mTRE and cTRE under different settings of n, K. The bold one is the best among the two estimates in a specific setting.

Setting	Method	(n,K)			
		(5000,5)	(5000,10)	(10000,5)	(10000,10)
Beta	mTRE	**0.9850(0.0269)**	**0.8840(0.0180)**	**1.0109(0.0171)**	**0.9299(0.0194)**
	cTRE	1.4670(0.0606)	1.2935(0.0274)	1.3674(0.0625)	1.2850(0.0293)
Normal	mTRE	**2.7426(0.0370)**	2.8330(0.0450)	**2.7483(0.0367)**	2.7813(0.0265)
	cTRE	2.7987(0.0586)	**2.7076(0.0293)**	2.8184(0.0347)	**2.7503(0.0297)**

D_0 as below,

$$\widehat{\mathrm{MSE}}(\hat{D}) = \left[\frac{1}{n_t}\sum_{i=1}^{n_t}\{\hat{D}(Z_{t,i}) - D_0(Z_{t,i})\}^2\right]^{1/2}.$$

We calculate the averaged MSEs over 10 replications and the corresponding standard errors for mTRE and cTRE under different settings of n, K, where n is the training data sample size and K is the length of the chain. In addition, we use $n_t = 5000$ in this simulation. The implementation details are given in Appendix G.4 of the paper. The results are summarized in Table 6. Clearly, Table 6 shows that, for the models considered in the simulation studies, the proposed mixing chain performs comparably or better compared with the original convolution chain.

5.5 Comparison of mTRE and cTRE on the MNIST dataset

We now apply the proposed mixing chain (18) for density ratio estimation to the MNIST dataset (http://yann.lecun.com/exdb/mnist), which consists of 60,000 training and 10,000 testing grayscale images with 28 × 28 pixels. In the implementation, to accelerate the computation, we use the subsampling method with a training subsample size of 20,000 and a relatively small DenseNet network structure (Huang et al., 2017); see Table 10 of the Appendix for the specification of the network architectures and its subsection G.5 for the implementation details.

Similar to the results in Table 1 of Rhodes et al. (2020), we calculate the average negative log-likelihood (ANLL) in bits per dimension (bpd, smaller is better) on the 10,000 testing grayscale images; see its detailed definition in (64) of the supplementary materials. We denote the estimate based on the proposed mixing chain (18) with the chain length B by "mTRE-B". We obtain the averaged ANLLs and their empirical standard errors for mTRE-5 and mTRE-10 over 5 random training subsamples. For fairer comparison, we have run the original convolution chain (cTRE) in Rhodes et al. (2020) with a chain length 5 and 10 based on 20000 samples, mirroring the setups applied to the proposed mTRE. The simulation results are presented in Table 7. From Table 7, one can see that mTRE performs better than the corresponding cTRE, and cTRE-10 failed to yield results due to recurring occurrences of a "NAN" output in five random re-samplings.

Table 7: The averaged ANLL's and their empirical standard errors in parentheses for mTRE-5, mTRE-10, cTRE-5 and cTRE-10 over 5 random training subsamples.

Estimator	mTRE-5	mTRE-10	cTRE-5	cTRE-10
ANLL	1.40 (0.0045)	1.39 (0.0077)	1.41 (0.0002)	–

6. Related work: comparison with the NN-BD estimator

There has been much work on the error analysis of nonparametric density-ratio estimation (Nguyen et al., 2010; Sugiyama et al., 2008; Kanamori et al., 2012a; Yamada et al., 2013). These results show that when the targeted density-ratio belongs to certain function space $\mathcal{H}$, such as RKHS, and thus no approximation error is incurred, their estimators achieve certain nonparametric convergence rate decided by the complexity of $\mathcal{H}$. Compared to these works, our results consider the approximation error using neural network functions and still achieves the minimax optimality up to some logarithmic factor under some mild conditions.

Our work is most related to the paper by Kato and Teshima (2021), who proposed a non-negative Bregman divergence (NN-BD) method to tackle the possible over-fitting problem due to the unboundedness of certain Bregman divergences. We compare our theoretical results with those for the NN-BD estimator of Kato and Teshima (2021). Using the notation in this paper, we restate two conditions required in Kato and Teshima (2021):

(a) Let $\mathcal{F}^R_{\mathrm{FNN}}$ be a class of FNNs with output taking values in $[e^{-M}, e^M]$ for some finite $M > 0$. The target density-ratio $R^* \in \mathcal{F}^R_{\mathrm{FNN}}$. Moreover, for any function in $\mathcal{F}^R_{\mathrm{FNN}}$, its Frobenius norm of the parameter matrix W_j in the jth layer is bounded by $\mathcal{B}_j \geq 0$ and the activation functions are 1-Lipschitz positive-homogeneous.

(b) The function $\psi(\cdot)$ is σ-strongly convex. Let $\ell_1(t) = \psi^*(t)t - \psi(t) + A, \ell_2(t) = -\tilde{\psi}(t), t \in [e^{-M}, e^M]$, where $\psi^*(t) = C_{nn}\{\psi'(t)t - \psi(t)\} + \tilde{\psi}(t)$. Here $\tilde{\psi}(t)$ is a function bounded above, C_{nn} and A are user-selected constants. Suppose $\ell_1(\cdot)$ and $\ell_2(\cdot)$ are Lipschitz functions on $[e^{-M}, e^M]$.

Under these two conditions, Kato and Teshima (2021) rewrote the BD in (4) as

$$\mathcal{B}_\psi(R) = E_p\ell_1(R(Z)) - C_{nn}E_q\ell_1(R(Z)) + E_q\ell_2(R(Z)) + (1 - C_{nn})A,$$

and proposed the density-ratio estimator $\widehat{R}_{\mathrm{KT}}$ defined as

$$\widehat{R}_{\mathrm{KT}} = \operatorname*{argmin}_{R\in\mathcal{F}_{\mathrm{FNN}}} \Big\{\frac{1}{n_q}\sum_{i=1}^{n_q}\ell_2(R(Z_{q,i})) + \Big[\frac{1}{n_p}\sum_{j=1}^{n_p}\ell_1(R(Z_{p,j})) - \frac{C_{nn}}{n_q}\sum_{j=1}^{n_q}\ell_1(R(Z_{q,j}))\Big]_+\Big\},$$

where $[a]_+ = \max(0, a)$ for any $a \in \mathbb{R}$. They showed that

$$\|\widehat{R}_{\mathrm{KT}} - R^*\|_p = O_p\left(n^{-\frac{1}{2+a}}\right) \tag{24}$$

for any $0<a<2$.

As a comparison, we have the following error bound for our density-ratio estimator $\widehat{R} = \exp(\widehat{D})$ based on Theorem 6.

Corollary 19 *Under Assumption* 4*, when* $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}^{R})$*, there exists a constant* C *depending only on* (μ, σ, M) *such that, for any* $\gamma \geq 0$*, with probability at least* $1 - \exp(-\gamma)$*,*

$$\|\widehat{R} - R^*\|_p \leq C \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}^{R}) \log n}{n}} + \sqrt{\frac{\gamma}{n}} \right).$$

Corollary 19 implies that $\|\widehat{R} - R^*\|_p = O_p\big(\sqrt{\log n/n}\big)$, when the true density-ratio $R^* \in \mathcal{F}_{\mathrm{FNN}}$. This convergence rate is faster than the rate for $\widehat{R}_{\mathrm{KT}}$ given in (24) obtained in Kato and Teshima (2021). Moreover, the boundedness assumption for the weights of the neural network functions, as imposed by Kato and Teshima (2021), is not needed in our result. Corollary 19 also shows that, if the target ratio is assumed to belong to the optimization space (or hypothesis space), i.e., $R^* \in \mathcal{F}_{\mathrm{FNN}}^{R}$ without approximation error, then the convergence rate does not depend on the dimension of the data. In other words, the estimation of R^* does not suffer from the curse of dimensionality. However, the well-specified model assumption is not realistic. Therefore, it is important to consider the approximation error since most likely $R^* \notin \mathcal{F}_{\mathrm{FNN}}^{R}$ in applications. Note that there is essential difference between our proof of Theorem 6 and those of Theorem 2 in Kato and Teshima (2021): the proof of Theorem 2 in Kato and Teshima (2021) operates in an asymptotic framework, whereas our proof of Theorem 6 is in a non-asymptotic sense. A detailed comparison is provided in Appendix F.

7. Conclusions

In this work, we have established non-asymptotic error bounds for the BDD density-ratio estimator. Under reasonable conditions, we have shown that this estimator achieves the optimal minimax convergence rate up to some logarithmic factor. When the data distribution is supported on an approximate low-dimensional manifold, we have shown that the BDD estimator can mitigate the curse of dimensionality. We have extended the analysis to the cases that the target density ratio has unbounded support or is unbounded itself. As an application of our theoretical results, we proposed an estimator of the KL divergence that is asymptotically normal based on our convergence results for density ratio estimation and a data-splitting procedure. We have also analyzed the convergence properties of the telescoping density ratio estimator (Rhodes et al., 2020) and provided sufficient conditions under which it has a lower error bound than a single-ratio estimator.

A limitation of our work is that certain boundedness assumption on the target density ratio such as Assumption 5 or 16 is needed. Also, when the boundedness assumption is partially relaxed as in Assumption 16, the error bound in Theorem 17 is not as sharp as that with the boundedness assumption in Theorem 6. It would be interesting to further relax or remove such assumptions. In addition, as we have stated in Remark 4, to the best of our knowledge, what a tight bound is for density-ratio estimation in an unbounded setting is still an open problem. Our result in Theorem 17, albeit not tight, is an attempt to tackle the difficult unbounded density-ratio estimation problem. These are interesting and challenging problems that deserve further study in the future.

Acknowledgments

The authors wish to thank the action editor Professor Kenji Fukumizu and two reviewers for their insightful comments and constructive suggestions that significantly helped improve the article. This work was conducted when Siming Zheng was a postdoctoral fellow in Department of Statistics and Data Science, The Chinese University of Hong Kong. Guohao Shen's research was partially supported by the National Key R&D Program of China (No. 2024YFA1015700), the National Natural Science Foundation of China (No. 12501713), the Hong Kong Research Grants Council (No. 15305523), and a research grant from The Hong Kong Polytechnic University (No. P0048718). Yuanyuan Lin's research was partially supported by the Hong Kong Research Grants Council (No. 14304523) and The Hong Kong Jockey Club Equine Welfare Research Foundation, Direct Grants for Research, The Chinese University of Hong Kong. Jian Huang's research was supported by the National Natural Science Foundation of China grant (No. 72331005) and the research grants from The Hong Kong Polytechnic University (Nos. P0046811, P0042888, P0045417, P0045931).

Appendix

In the appendix, we provide the technical details and proofs of the results stated in the paper, and include the implementation details for the experiments reported in Section 5.

Appendix A. Proofs and technical details

Proof [Proof of Theorem 6] For notational convenience, denote $\epsilon_n = \|D_{\mathrm{NN}} - D^*\|_{\max}$ and use E_I to denote E_{I^*}, $I = p, q$. Recall that E_{n_I} denotes the expectation with respect to (w.r.t) the empirical distribution of $\{Z_{I,t}\}_{t=1}^{n_I}$ for $I = p, q$. As $\widehat{D} \in \operatorname{argmin}_{D\in\mathcal{F}_{\mathrm{FNN}}} \mathcal{L}_{n_p,n_q}(D)$, where $\mathcal{L}_{n_p,n_q}(D) = 1/n_p \sum_{j=1}^{n_p} \mathcal{L}_1(D(Z_{p,j})) + 1/n_q \sum_{i=1}^{n_q} \mathcal{L}_2(D(Z_{q,i}))$, we have

$$\begin{aligned}
& c_0\|\widehat{D} - D^*\|_{\max}^2 \\
\leq\ & \mathcal{B}_\psi\left(e^{\widehat{D}}\right) - \mathcal{B}_\psi\left(e^{D^*}\right) \\
\leq\ & \mathcal{B}_\psi\left(e^{\widehat{D}}\right) - \mathcal{B}_\psi\left(e^{D^*}\right) - \mathcal{L}_{n_p,n_q}(\widehat{D}) + \mathcal{L}_{n_p,n_q}(D_{\mathrm{NN}}) \\
=\ & \mathcal{B}_\psi\left(e^{\widehat{D}}\right) - \mathcal{L}_{n_p,n_q}(\widehat{D}) - \left\{\mathcal{B}_\psi\left(e^{D^*}\right) - \mathcal{L}_{n_p,n_q}(D^*)\right\} \\
+\ & \left\{\mathcal{L}_{n_p,n_q}(D_{\mathrm{NN}}) - \mathcal{L}_{n_p,n_q}(D^*)\right\} \\
=\ & (E_{p^*} - E_{n_p})\{\mathcal{L}_1(\widehat{D}) - \mathcal{L}_1(D^*)\} + (E_q - E_{n_q})\{\mathcal{L}_2(\widehat{D}) - \mathcal{L}_2(D^*)\} \\
+\ & E_{n_p}\{\mathcal{L}_1(D_{\mathrm{NN}}) - \mathcal{L}_1(D^*)\} + E_{n_q}\{\mathcal{L}_2(D_{\mathrm{NN}}) - \mathcal{L}_2(D^*)\}.
\end{aligned} \tag{25}$$

By Lemmas 25 and 30, with probability at least $1 - \exp(-\gamma_1)$,

$$E_{n_p}\{\mathcal{L}_1(D_{\mathrm{NN}}) - \mathcal{L}_1(D^*)\} \leq E_p\{\mathcal{L}_1(D_{\mathrm{NN}}) - \mathcal{L}_1(D^*)\} + \sqrt{2}C_1\|D_{\mathrm{NN}} - D^*\|_{\max}\sqrt{\frac{\gamma_1}{n}} + \frac{16C_1M\gamma_1}{3n}. \tag{26}$$

Also, with probability at least $1-\exp(-\gamma_1)$,

$$E_{n_q}\{\mathcal{L}_2(D_{\rm NN})-\mathcal{L}_2(D^*)\} \le E_q\{\mathcal{L}_2(D_{\rm NN})-\mathcal{L}_2(D^*)\}+\sqrt{2}C_2\|D_{\rm NN}-D^*\|_{\max}\sqrt{\frac{\gamma_1}{n}}+\frac{16C_2M\gamma_1}{3n}, \tag{27}$$

where $C_1=2e^{2M}\mu, C_2=e^M\mu$. In what follows, we provide the detailed derivation for (26) and omit the one for (27) as it is structurally the same as those for (26). By Lemma 25, $\mathcal{L}_1(t)$ is a function with a Lipschitz constant $C_1=2e^{2M}\mu$.

First, we will show that

$$\mathrm{Var}_{Z\sim p}(\mathcal{L}_1(D_{\rm NN}(Z))-\mathcal{L}_1(D^*(Z))) \le C_1^2\|D_{\rm NN}-D^*\|_{\max}^2. \tag{28}$$

To this end, using the basic inequality $\mathrm{Var}(X)=EX^2-E^2X\le EX^2$ for any random variable X, we have

$$\begin{aligned}\mathrm{Var}_{Z\sim p}(\mathcal{L}_1(D_{\rm NN}(Z))-\mathcal{L}_1(D^*(Z))) &\le E_{Z\sim p}(\mathcal{L}_1(D_{\rm NN}(Z))-\mathcal{L}_1(D^*(Z)))^2\\ &\le E_{Z\sim p}\{C_1(D_{\rm NN}(Z)-D^*(Z))\}^2\\ &= C_1^2E_{Z\sim p}(D_{\rm NN}(Z)-D^*(Z))^2\\ &= C_1^2\|D_{\rm NN}-D^*\|_p^2\\ &\le C_1^2\|D_{\rm NN}-D^*\|_{\max}^2,\end{aligned}$$

where the second inequality uses the fact that $\mathcal{L}_1(t)$ has a Lipschitz constant C_1 and the last inequality is because $\|f\|_{\max}=\max\{\|f\|_p,\|f\|_q\}$ by its definition. This proves (28).

For any $z\in[0,1]^d$, by the Lipschitz property of $\mathcal{L}_1$ and $\|D_{\rm NN}\|_\infty,\|D^*\|_\infty\le M$, we have

$$|\mathcal{L}_1(D_{\rm NN}(z))-\mathcal{L}_1(D^*(z))|\le C_1|D_{\rm NN}(z)-D^*(z)|\le 2MC_1.$$

By Bernstein's inequality in Lemma 30, it holds that with probability at least $1-\exp(-\gamma_1)$,

$$\begin{aligned}&E_{n_p}\{\mathcal{L}_1(D_{\rm NN})-\mathcal{L}_1(D^*)\}\\ =\;&\frac{1}{n_p}\sum_{j=1}^{n_p}\{\mathcal{L}_1(D_{\rm NN}(Z_{p,j}))-\mathcal{L}_1(D^*(Z_{p,j}))\}\\ \le\;&E_p\{\mathcal{L}_1(D_{\rm NN})-\mathcal{L}_1(D^*)\}+\frac{2MC_1\gamma_1}{3n_p}+\sqrt{2}C_1\|D_{\rm NN}-D^*\|_{\max}\sqrt{\frac{\gamma_1}{n_p}}\\ \le\;&E_p\{\mathcal{L}_1(D_{\rm NN})-\mathcal{L}_1(D^*)\}+\sqrt{2}C_1\|D_{\rm NN}-D^*\|_{\max}\sqrt{\frac{\gamma_1}{n}}+\frac{16C_1M\gamma_1}{3n}.\end{aligned}$$

In the last inequality we have used the fact that $n\le n_p$. This completes the proof of (26).

The inequalities (26) and (27) together imply that with probability at least $1-2\exp(-\gamma_1)$,

$$\begin{aligned}&E_{n_p}\{\mathcal{L}_1(D_{\rm NN})-\mathcal{L}_1(D^*)\}+E_{n_q}\{\mathcal{L}_2(D_{\rm NN})-\mathcal{L}_2(D^*)\}\\ \le\;&E_p\{\mathcal{L}_1(D_{\rm NN})-\mathcal{L}_1(D^*)\}+E_q\{\mathcal{L}_2(D_{\rm NN})-\mathcal{L}_2(D^*)\}\\ +\;&\sqrt{2}(C_1+C_2)\|D_{\rm NN}-D^*\|_{\max}\sqrt{\frac{\gamma_1}{n}}+\frac{16(C_1+C_2)M\gamma_1}{3n}\\ =\;&\mathcal{B}_\psi\left(e^{D_{\rm NN}}\right)-\mathcal{B}_\psi\left(e^{D^*}\right)+\sqrt{\frac{2\gamma_1}{n}}(C_1+C_2)\|D_{\rm NN}-D^*\|_{\max}+\frac{16(C_1+C_2)M\gamma_1}{3n}\\ \le\;&C_0\|D_{\rm NN}-D^*\|_{\max}^2+\sqrt{\frac{2\gamma_1}{n}}(C_1+C_2)\|D_{\rm NN}-D^*\|_{\max}+\frac{16(C_1+C_2)M\gamma_1}{3n}.\end{aligned} \tag{29}$$

***Step* 1**. Let $g = (D - D^*)^2$, then $g \leq 4M^2$ by Assumption 5. If $\|D - D^*\|_{\max} \leq r$, then

$$\mathrm{var}_p(g) \leq E_p(g^2) = E_p(D - D^*)^4 \leq 4M^2 E_p(D - D^*)^2 \leq 4M^2r^2.$$

Regarding g as a function of $D - D^*$, we have

$$\begin{aligned} |g(D_1 - D^*) - g(D_2 - D^*)| &= |D_1^2 - 2D_1D^* - (D_2^2 - 2D_2D^*)| \\ &= |(D_1 + D_2 - 2D^*)(D_1 - D_2)| \\ &= |(D_1 + D_2 - 2D^*)\{(D_1 - D^*) - (D_2 - D^*)\}| \\ &\leq 4M|(D_1 - D^*) - (D_2 - D^*)|. \end{aligned}$$

Thus g can be viewed as the function of $D - D^*$ with a Lipschitz constant $4M$. Denote $\mathcal{F}_{\mathrm{FNN}}^{D^*,r} = \{D \in \mathcal{F}_{\mathrm{FNN}}, \|D - D^*\|_{\max} \leq r\}$, and

$$R_{n_I}\mathcal{F} = \sup_{f \in \mathcal{F}} \frac{1}{n_I} \sum_{i=1}^{n_I} \eta_i^I f(Z_{I,i}),\ I = p, q,$$

where $\eta_i^I, i = 1, 2, \ldots, n_I$ are i.i.d. Rademacher variables. For the rest of the proof of Theorem 6, we use $E_\eta R_{n_I}\mathcal{F}$ to denote the conditional expectation of $R_{n_I}\mathcal{F}$ w.r.t $\eta_i^I, i = 1, 2, \ldots, n_I$, given $Z_{I,i}, i = 1, 2, \ldots, n_I$ and $E_{I,\eta}R_{n_I}\mathcal{F}$ to denote the expectation of $R_{n_I}\mathcal{F}$ jointly w.r.t $\eta_i^I, Z_{I,i}, i = 1, 2, \ldots, n_I$. By Lemma 29, with probability at least $1 - \exp(-\gamma_1)$,

$$\begin{aligned} &\|D - D^*\|_{p,n_p}^2 - \|D - D^*\|_p^2 \\ \leq\ & 3E_{p,\eta}R_{n_p}\left\{(D - D^*)^2 : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,r}\right\} + 2\sqrt{\frac{2\gamma_1}{n}}M + \frac{16M^2}{3}\frac{\gamma_1}{n} \\ \leq\ & 24ME_{p,\eta}R_{n_p}\left\{(D - D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,r}\right\} + 2\sqrt{\frac{2\gamma_1}{n}}Mr + \frac{16M^2}{3}\frac{\gamma_1}{n}, \end{aligned} \tag{30}$$

where the last inequality follows from Talagland's contraction theorem. Similarly, with probability at least $1 - \exp(-\gamma_1)$,

$$\|D - D^*\|_{q,n_q}^2 - \|D - D^*\|_q^2 \leq 24ME_{q,\eta}R_{n_q}\left\{(D - D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,r}\right\} + 2\sqrt{\frac{2\gamma_1}{n}}Mr + \frac{16M^2}{3}\frac{\gamma_1}{n}. \tag{31}$$

Let $R_n(r)/(24M) = \max_{I \in \{p,q\}} \left\{E_{I,\eta}R_{n_I}\left\{(D - D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,r}\right\}\right\}$. When

$$r^2 \geq R_n(r), r^2 \geq \frac{16M^2\gamma}{3n}, \tag{32}$$

(30) and (31) indicate that with probability at least $1 - 2\exp(-\gamma_1)$,

$$\begin{aligned} \|D - D^*\|_{n_p,n_q}^2 &= \max\{\|D - D^*\|_{p,n_p}^2, \|D - D^*\|_{q,n_q}^2\} \\ &\leq \max\{\|D - D^*\|_p^2, \|D - D^*\|_q^2\} + R_n(r) + 2\sqrt{\frac{2\gamma_1}{n}}Mr + \frac{16M^2}{3}\frac{\gamma_1}{n} \\ &= \|D - D^*\|_{\max}^2 + R_n(r) + 2\sqrt{\frac{2\gamma_1}{n}}Mr + \frac{16M^2}{3}\frac{\gamma_1}{n} \\ &\leq (2r)^2. \end{aligned}$$

Thus, when (32) holds, with probability at least $1-2\exp(-\gamma_1)$,

$$\|D-D^*\|_{\max} \le r \Rightarrow \|D-D^*\|_{n_p,n_q} \le 2r. \tag{33}$$

***Step* 2**. Suppose $\|\widehat{D}-D^*\|_{\max} \le r_0$ and let

$$\mathcal{G}_i = \left\{\mathcal{L}_i(D)-\mathcal{L}_i(D^*) : D \in \mathcal{F}_{\text{FNN}}^{D^*,r_0}\right\}, i=1,2.$$

For each $(I,i) \in \{(p,1),(q,2)\}$, with probability at least $1-2\exp(-\gamma_1)$,

$$(E_I - E_{n_I})\{\mathcal{L}_i(\widehat{D})-\mathcal{L}_i(D^*)\} \le 6E_\eta R_{n_I}\mathcal{G}_i + \sqrt{2}C_i r_0\sqrt{\frac{\gamma_1}{n}} + \frac{46C_iM\gamma_1}{3n}. \tag{34}$$

Denote $\hat{\mathcal{F}}_{\text{FNN}}^{D^*,r} = \{D \in \mathcal{F}_{\text{FNN}}, \|D-D^*\|_{n_p,n_q} \le r\}$. By (33) in *Step* 1, when $r_0^2 \ge R_n(r_0)$ and $r_0^2 \ge 16M^2\gamma_1/(3n)$, with probability at least $1-2\exp(-\gamma_1)$, for each $(I,i) \in \{(p,1),(q,2)\}$,

$$E_\eta R_{n_I}\mathcal{G}_i \le 2C_iE_\eta R_{n_I}\left\{(D-D^*) : D \in \mathcal{F}_{\text{FNN}}^{D^*,r_0}\right\} \le 2C_iE_\eta R_{n_I}\left\{(D-D^*) : D \in \hat{\mathcal{F}}_{\text{FNN}}^{D^*,2r_0}\right\}.$$

Denote $\hat{\mathcal{F}}_I^{D^*,r} = \{D \in \mathcal{F}_{\text{FNN}}, \|D-D^*\|_{I,n_I} \le r\}$. When $n \ge \text{Pdim}(\mathcal{F}_{\text{FNN}}), r_0 \ge 1/n$ and $n \ge (2eM)^2$, we have

$$E_\eta R_{n_I}\{(D-D^*) : D \in \hat{\mathcal{F}}_I^{D^*,2r_0}\} \le 64r_0\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n}{n}}, \tag{35}$$

and thus

$$E_\eta R_{n_I}\mathcal{G}_i \le 128C_ir_0\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n}{n}}. \tag{36}$$

Combining (25) (29) (34) and (36), with probability at least $1-8\exp(-\gamma_1)$, we have

$$\begin{aligned} c_0\|\widehat{D}-D^*\|_{\max}^2 &\le 768(C_1+C_2)r_0\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n}{n}} \\ &\quad + \sqrt{\frac{2\gamma_1}{n}}(C_1+C_2)r_0 + \frac{46(C_1+C_2)M\gamma_1}{3n} + C_0\epsilon_n^2 \\ &\quad + \sqrt{\frac{2\gamma_1}{n}}(C_1+C_2)\epsilon_n + \frac{16(C_1+C_2)M\gamma_1}{3n} \\ &= (C_1+C_2)r_0\left(768\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n}{n}} + \sqrt{\frac{2\gamma_1}{n}}\right) \\ &\quad + C_0\epsilon_n^2 + \sqrt{\frac{2\gamma_1}{n}}(C_1+C_2)\epsilon_n + \frac{62(C_1+C_2)M\gamma_1}{3n}. \end{aligned}$$

Therefore, when $\max\left\{\sqrt{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n/n}, \epsilon_n\right\} \ll r_0$, there exists $r_1 \ll r_0$ such that $\|\widehat{D}-D^*\|_{\max} \ll r_1$.

***Step* 3**. Let $r_* = \inf\{r \ge 0 : R_n(s) \le s^2, \text{ for } s \ge r\}$ and E be the event on which $\|D-D^*\|_{n_p,n_q} \le 4r_*$ for all $D \in \mathcal{F}_{\text{FNN}}^{D^*,2r_*}$ We intend to prove

$$r_* \le \kappa M\sqrt{\frac{\text{Pdim}(\mathcal{F}_{\text{FNN}})\log n}{n}}, \ \kappa = 24 \times 130. \tag{37}$$

When $r_* \leq 2\sqrt{3}M\sqrt{\log n/n}/3$, the inequality is trivial. When $r_* \geq 2\sqrt{3}M\sqrt{\log n/n}/3$, by the result in Step 1, $P(E) \geq 1 - 2/n$. As a result,

$$r_*^2 \leq R_n(r_*) \leq R_n(2r_*) = 24M \max_{I\in\{p,q\}} \left\{E_{I,\eta}R_{n_I}\left\{(D-D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,2r_*}\right\}\right\}.$$

For each $I \in \{p, q\}$,

$$\begin{aligned} E_{I,\eta}R_{n_I}\left\{(D-D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,2r_*}\right\} &= E_I E_\eta R_{n_I}\left\{(D-D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,2r_*}\right\} \\ &= E_I E_\eta R_{n_I}\left\{(D-D^*) : D \in \mathcal{F}_{\mathrm{FNN}}^{D^*,2r_*}\right\}(I_E + I_{E^c}) \\ &\leq E_I E_\eta R_{n_I}\left\{(D-D^*) : D \in \hat{\mathcal{F}}_{\mathrm{FNN}}^{D^*,4r_*}\right\} + \frac{4M}{n}. \end{aligned}$$

It follows from (35) that

$$\begin{aligned} r_*^2 &\leq 24M\left(128r_*\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}} + \frac{4M}{n}\right) \\ &= 24M\left(128r_*\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}} + r_* \cdot \frac{4M}{n} \cdot \frac{1}{r_*}\right) \\ &\leq 24Mr_*\left(128\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}} + \sqrt{\frac{3}{n\log n}}\right) \\ &\leq \kappa\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}Mr_*, \end{aligned}$$

where $\kappa = 24 \times 130$. Thus, $r_* \leq \kappa M\sqrt{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n/n}$ and (37) is proved.

***Step* 4**. Let $B_{\max}(D^*, r) = \{D \in \mathcal{F}_{\mathrm{FNN}}, \|D - D^*\|_{\max} \leq r\}$, $\bar{r} \geq \max\left(\sqrt{\log n/n}, r_*\right)$ and $l = \left\lfloor \log_2(2M/\sqrt{\log n/n})\right\rfloor$. Then, the neural network function space $\mathcal{F}_{\mathrm{FNN}}$ can be divided into

$$B_{\max}(D^*, \bar{r}), B_{\max}(D^*, 2\bar{r})\backslash B_{\max}(D^*, \bar{r}), \ldots, B_{\max}(D^*, 2^l\bar{r})\backslash B_{\max}(D^*, 2^{l-1}\bar{r}).$$

As $\bar{r} \geq r_*$, it then follows from the definition of r_* that $\bar{r}^2 \geq R_n(\bar{r})$. Further, if $\bar{r}^2 \geq 16M^2\gamma_1/(3n)$, according to (33) in *Step* 1, with probability at least $1 - 2l\exp(-\gamma_1)$, for any $j = 1, 2, \ldots, l$,

$$\|D - D^*\|_{\max} \leq 2^j\bar{r} \Rightarrow \|D - D^*\|_{n_p,n_q} \leq 2^{j+1}\bar{r}.$$

Suppose that for some $j \leq l$, $\widehat{D} \in B_{\max}(D^*, 2^j\bar{r})\backslash B_{\max}(D^*, 2^{j-1}\bar{r})$, then by the results in Step 2, with probability at least $1 - 8\exp(-\gamma_1)$,

$$\begin{aligned} c_0\|\widehat{D} - D^*\|_{\max}^2 \leq{}& (C_1 + C_2)2^j\bar{r}\left(768\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}} + \sqrt{\frac{2\gamma_1}{n}}\right) \\ &+ C_0\epsilon_n^2 + \sqrt{\frac{2\gamma_1}{n}}(C_1 + C_2)\epsilon_n + \frac{62(C_1 + C_2)M\gamma_1}{3n}. \end{aligned}$$

If

$$\frac{1}{c_0}(C_1+C_2)\left(768\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{2}\sqrt{\frac{\gamma_1}{n}}\right)\leq\frac{1}{8}2^j\bar{r}, \tag{38}$$

and

$$\frac{1}{c_0}\left[C_0\epsilon_N^2+\sqrt{2}(C_1+C_2)\epsilon_N\sqrt{\frac{\gamma_1}{n}}+\frac{62(C_1+C_2)M\gamma_1}{3n}\right]\leq\frac{1}{8}2^{2j}\bar{r}^2, \tag{39}$$

then

$$\|\widehat{D}-D^*\|_{\max}^2\leq 2^{2j-2}\bar{r}^2\Leftrightarrow\|\widehat{D}-D^*\|_{\max}\leq 2^{j-1}\bar{r}.$$

In short, to obtain this inequality, we need $\bar{r}$ satisfying (38), (39) and

$$\bar{r}\geq\max\left(\sqrt{\frac{\log n}{n}},M\sqrt{\frac{16\gamma_1}{3n}},r_*\right).$$

As $r_*\leq\kappa M\sqrt{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n/n}$, let

$$C_*=\max\left\{\frac{3072(C_1+C_2)}{c_0},3120M,\sqrt{\frac{\max\{6C_0,124(C_1+C_2)M,5(C_1+C_2)\}}{c_0}}\right\},$$

then

$$\bar{r}=C_*\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{\frac{\gamma_1}{n}}+\epsilon_n\right)$$

satisfies all the requirements. As a result, with probability at least $1-10l\exp(-\gamma_1)$,

$$\|\widehat{D}-D^*\|_{\max}\leq\bar{r}\text{ and }\|\widehat{D}-D^*\|_{n_p,n_q}\leq 2\bar{r}.$$

Let $\gamma_1=\log 10l+\gamma, l=\lfloor\log_2(2M/\sqrt{\log n/n})\rfloor$. By the basic inequality $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for $a,b\geq 0$, with probability at least $1-\exp(-\gamma)$,

$$\begin{aligned}\|\widehat{D}-D^*\|_{\max}\leq\bar{r} &= C_*\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{\frac{\log 10l+\gamma}{n}}+\epsilon_n\right)\\ &\leq C_*\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{\frac{\gamma}{n}}+\epsilon_n+\sqrt{\frac{\log 10l}{n}}\right)\\ &\leq (2C_*+\log_2 M+1)\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{\frac{\gamma}{n}}+\epsilon_n\right).\end{aligned}$$

Similarly, we have

$$\|\widehat{D}-D^*\|_{n_p,n_q}\leq 2(2C_*+\log_2 M+1)\left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+\sqrt{\frac{\gamma}{n}}+\epsilon_n\right). \tag{40}$$

Since $C_1 = 2e^{2M}\mu, C_2 = e^M\mu$, $c_0 = \sigma e^{-3M}/2$ and $C_0 = \mu e^{3M}/2$, $(2C_* + \log_2 M + 1)$ can be significantly simplified through bounding each (M, μ, σ)-related quantity in $(2C_* + \log_2 M + 1)$ by $(Me^{5M}\mu)/\sigma$ up to some constant prefactors, when $\sigma \leq \mu$ (It is satisfied by the ψ in Table 1) and $M \geq 1$. Here we illustrate how to bound $3072(C_1 + C_2)/c_0$ and $3120M$. The other quantities can be bounded in some similar ways. Since $M \geq 1$, obviously we have

$$\frac{3072(C_1 + C_2)}{c_0} = \frac{12288e^{5M}(1 + e^{-M}/2)\mu}{\sigma} \leq \frac{24576e^{5M}\mu}{\sigma} \leq \frac{24576Me^{5M}\mu}{\sigma}.$$

Because $M \geq 1$ and $\sigma \leq \mu$, we know $(e^{5M}\mu)/\sigma \geq 1$ and then clearly

$$3120M \leq 3120M \times \frac{e^{5M}\mu}{\sigma} \leq \frac{3120Me^{5M}\mu}{\sigma}.$$

Given these discussions, there exists some large universal positive constant $\tilde{C}$ such that

$$2C_* + \log_2 M + 1 \leq \frac{\tilde{C}Me^{5M}\mu}{\sigma}.$$

The proof of Theorem 6 is completed. ■

Proof [Proof of Theorem 10] Since $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0, 1]$, by Lemma 9, for the $\mathcal{F}_{\text{FNN}}$, a function class consists of ReLU FNN with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} L\lceil\log_2(8L)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 K\lceil\log_2(8K)\rceil$, where $K, L \in \mathbb{N}^+$, there exists a function $\phi_0 \in \mathcal{F}_{\text{FNN}}$ such that

$$\sup_{x\in[0,1]^d\setminus H_{B,\delta}} |D^* - \phi_0| \leq 18M(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+(\beta\vee1)/2}(KL)^{-\frac{2\beta}{d}}, \tag{41}$$

where $H_{B,\delta} = \cup_{i=1}^d \{x = [x_1, \ldots, x_d] : x_i \in \cup_{b=1}^{B-1} (b/B - \delta, b/B)\}, B = \lceil(KL)^{2/d}\rceil, \delta \in (0, 1/(3B)]$. As $D_{\text{NN}} \in \text{argmin}_{D\in\mathcal{F}_{\text{FNN}}} \|D - D^*\|_{\max}$, then $\|D_{\text{NN}} - D^*\|^2_{\max} \leq \|\phi_0 - D^*\|^2_{\max}$. By the result in (41), for $I = p$ or q, we have

$$\begin{aligned} \|\phi_0 - D^*\|_I^2 &= \int_{[0,1]^d\setminus H_{B,\delta}} |D^* - \phi_0|^2 I^*(x)dx + \int_{H_{B,\delta}} |D^* - \phi_0|^2 I^*(x)dx \\ &\leq 324M^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee1)}(KL)^{-\frac{4\beta}{d}} + 4M^2 \int_{H_{B,\delta}} I^*(x)dx. \end{aligned}$$

As $p^*(\cdot), q^*(\cdot)$ are the density functions of some measures on $[0, 1]^d$ which are absolutely continuous with respect to the Lebesgue measure and δ can be arbitrarily small, $\int_{H_{B,\delta}} I_0(x)dx$ is also arbitrarily small. Thus we have

$$\|\phi_0 - D^*\|_I^2 \leq 324M^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee1)}(KL)^{-\frac{4\beta}{d}}$$

and

$$\begin{aligned} \|D_{\text{NN}} - D^*\|^2_{\max} &\leq \|\phi_0 - D^*\|^2_{\max} \\ &\leq 324M^2(\lfloor\beta\rfloor + 1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee1)}(KL)^{-\frac{4\beta}{d}} \\ &= 324M^2 C_1(\beta, d)(KL)^{-\frac{4\beta}{d}}. \end{aligned}$$

By Corollary 7, there exists a constant C_1 only depending on (μ, σ, M) such that

$$\begin{aligned} E_{S_p,S_q}\|\widehat{D}-D^*\|^2_{\max} &\le C_1\left(\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}+\|D_{\mathrm{NN}}-D^*\|^2_{\max}\right) \\ &\le C_1\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}+324M^2C_1(\beta,d)(KL)^{-\frac{4\beta}{d}}\right\} \\ &\le 324M^2C_1\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}+C_1(\beta,d)(KL)^{-\frac{4\beta}{d}}\right\}. \qquad (42) \end{aligned}$$

This completes the proof of the first part of Theorem 10.

As for the second part of this theorem, based on Lemma 31, for a specific ReLU network f_ϕ, where ϕ contains the parameters in the network, there exists a universal constant C_2 such that

$$\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \le C_2\mathcal{S}\mathcal{D}\log\mathcal{S},$$

where $\mathcal{S}$ is the total number of parameters in the network f_ϕ. For a ReLU FNN with width $\mathcal{W}$ and depth $\mathcal{D}$, it can be easily checked that $\mathcal{S} = O(\mathcal{W}^2\mathcal{D})$. Now for $\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+1}$, $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\left\lceil n^{d/\{2(d+2\beta)\}}\log_2\left(8n^{d/\{2(d+2\beta)\}}\right)\right\rceil$, and $\mathcal{W},\mathcal{D}$ satisfy

$$O(\mathcal{W}^2\mathcal{D}) = O\left((\lfloor\beta\rfloor+1)^6d^{2\lfloor\beta\rfloor+2}\left\lceil n^{\frac{d}{2(d+2\beta)}}\log n\right\rceil\right),$$

which means $L=1, K=\left\lceil n^{d/\{2(d+2\beta)\}}\right\rceil$, and there exist three universal constants C_3, C_4, C_5 such that

$$\begin{aligned} &\frac{\mathcal{S}\mathcal{D}\log\mathcal{S}\log n}{n} \\ \le\ & C_3\left\{(\lfloor\beta\rfloor+1)^6d^{2\lfloor\beta\rfloor+2}\left\lceil n^{\frac{d}{2(d+2\beta)}}\log n\right\rceil\right\}\times\left(\log\left[C_3\left\{(\lfloor\beta\rfloor+1)^6d^{2\lfloor\beta\rfloor+2}\left\lceil n^{\frac{d}{2(d+2\beta)}}\log n\right\rceil\right\}\right]\right) \\ & \times\left\{21(\lfloor\beta\rfloor+1)^2\left\lceil n^{\frac{d}{2(d+2\beta)}}\log_2\left(8n^{\frac{d}{2(d+2\beta)}}\right)\right\rceil\log n/n\right\} \\ \le\ & \frac{C_4}{n}\left\{(\lfloor\beta\rfloor+1)^8d^{2\lfloor\beta\rfloor+2}n^{\frac{2d}{2(d+2\beta)}}\log^2 n\right\} \\ & \times\left\{6\log(\lfloor\beta\rfloor+1)+2(\lfloor\beta\rfloor+1)\log d+\frac{d}{2(d+2\beta)}\log n\right\} \\ \le\ & \frac{C_4}{n}\left\{(\lfloor\beta\rfloor+1)^8d^{2\lfloor\beta\rfloor+2}n^{\frac{2d}{2(d+2\beta)}}\log^2 n\right\}\{6(\lfloor\beta\rfloor+1)+2(\lfloor\beta\rfloor+1)d+\log n\} \\ \le\ & C_5(\lfloor\beta\rfloor+1)^9d^{2\lfloor\beta\rfloor+3}n^{-\frac{2\beta}{d+2\beta}}\log^3 n. \qquad (43) \end{aligned}$$

It follows from (42) that

$$\begin{aligned} & E_{S_p,S_q}\|\widehat{D}-D^*\|^2_{\max} \\ \le\ & 324M^2C_1\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}+C_1(\beta,d)(KL)^{-\frac{4\beta}{d}}\right\} \\ \le\ & 324M^2C_1\left\{\frac{C_2\mathcal{S}\mathcal{D}\log\mathcal{S}\log n}{n}+C_1(\beta,d)(KL)^{-\frac{4\beta}{d}}\right\} \\ \le\ & 324M^2C_1\left\{C_2C_5(\lfloor\beta\rfloor+1)^9d^{2\lfloor\beta\rfloor+3}n^{-\frac{2\beta}{d+2\beta}}\log^3 n+(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+(\beta\vee 1)}n^{-\frac{2\beta}{d+2\beta}}\right\} \\ \le\ & 324M^2C_1(C_2C_5+1)(\lfloor\beta\rfloor+1)^9d^{2\lfloor\beta\rfloor+(\beta\vee 3)}n^{-\frac{2\beta}{d+2\beta}}\log^3 n. \end{aligned}$$

This completes the proof of the second part of Theorem 10. ■

Proof [Proof of (14)] Write

$$\begin{aligned}
& \left|\widehat{\mathrm{KL}}(q^*||p^*) - \mathrm{KL}(q^*||p^*)\right| \\
= & \left|\frac{1}{n_q}\sum_{i=1}^{n_q}\widehat{D}(Z_{q,i}) - \mathrm{KL}(q^*||p^*)\right| \\
= & \left|\frac{1}{n_q}\sum_{i=1}^{n_q}\widehat{D}(Z_{q,i}) - \frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) + \frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) - E_{Z\sim q^*}D^*(Z)\right| \\
\leq & \left|\frac{1}{n_q}\sum_{i=1}^{n_q}\widehat{D}(Z_{q,i}) - \frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i})\right| + \left|\frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) - E_{Z\sim q^*}D^*(Z)\right| \\
\leq & \|\widehat{D} - D^*\|_{q,n_q} + \left|\frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) - E_{Z\sim q^*}D^*(Z)\right| \\
\leq & \|\widehat{D} - D^*\|_{n_p,n_q} + \left|\frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) - E_{Z\sim q^*}D^*(Z)\right|. \qquad (44)
\end{aligned}$$

Applying the result for $\|D_{\mathrm{NN}} - D^*\|_{n_p,n_q}$ in Theorem 6 and following the same proof of Theorem 10 but with $E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{\max}$ replaced by $E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{n_p,n_q}$, under the conditions that $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$, and $\mathcal{F}_{\mathrm{FNN}}$ is the function class of ReLU DNNs with width $\mathcal{W}$ and depth $\mathcal{D}$ satisfying

$$\mathcal{W} = 114(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1}, \quad \mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 \left\lceil n^{\frac{d}{2(d+2\beta)}} \log_2\left(8n^{\frac{d}{2(d+2\beta)}}\right)\right\rceil,$$

we have

$$E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{n_p,n_q} \leq C_0(\lfloor\beta\rfloor + 1)^9 d^{2\lfloor\beta\rfloor+(\beta\vee 3)} n^{-\frac{2\beta}{d+2\beta}} \log^3 n.$$

This implies that

$$\|\widehat{D} - D^*\|_{n_p,n_q} = O_p\left(n^{-\frac{\beta}{d+2\beta}} \log^{\frac{3}{2}} n\right). \qquad (45)$$

By the classic central limit theorem, we have

$$\left|\frac{1}{n_q}\sum_{i=1}^{n_q}D^*(Z_{q,i}) - E_{Z\sim q^*}D^*(Z)\right| = O_p\left(n_q^{-\frac{1}{2}}\right) = O_p\left(n^{-\frac{1}{2}}\right), \qquad (46)$$

as $n \leq n_q$. As a result, by (44), (45), and (46), we have

$$\left|\widehat{\mathrm{KL}}(q^*||p^*) - \mathrm{KL}(q^*||p^*)\right| = O_p\left(n^{-\frac{\beta}{d+2\beta}} \log^{\frac{3}{2}} n\right).$$

This completes the proof of (14). ■

Proof [Proof of Theorem 11] For convenience, recall that

$$S_q = \{Z_{q,1}, Z_{q,2}, \ldots, Z_{q,n_q}\},$$

$$S_p = \{Z_{p,1}, Z_{p,2}, \ldots, Z_{p,n_p}\},$$

and we write

$$S_q^{\text{test}} = \{\tilde{Z}_{q,1}, \tilde{Z}_{q,2}, \ldots, \tilde{Z}_{q,m}\}.$$

Obviously, we have

$$\begin{aligned} & \sqrt{m}\left(\widetilde{\mathrm{KL}}_m(q^*||p^*) - \mathrm{KL}(q^*||p^*)\right) \\ = \ & \sqrt{m}\left(\widetilde{\mathrm{KL}}_m(q^*||p^*) - E_{q^*}D^*(Z)\right) \\ = \ & \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\}\right] + \sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m}\{D^*(\tilde{Z}_{q,i}) - E_{q^*}D^*(Z)\}\right]. \end{aligned}$$

Since $m = o(n^{2\beta/(d+2\beta)}\log^3 n)$, it suffices to prove that

$$\frac{1}{m}\sum_{i=1}^{m}\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\} = O_p\left(n^{-\frac{\beta}{d+2\beta}}\log^{\frac{3}{2}} n\right). \tag{47}$$

By the Cauchy-Schwarz inequality, we have

$$\left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\}\right]^2 \le \frac{1}{m}\sum_{i=1}^{m}\left\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\right\}^2.$$

Therefore,

$$\begin{aligned} & E_{S_q,S_p,S_q^{\text{test}}}\left[\frac{1}{m}\sum_{i=1}^{m}\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\}\right]^2 \\ \le \ & E_{S_q,S_p,S_q^{\text{test}}}\left[\frac{1}{m}\sum_{i=1}^{m}\left\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\right\}^2\right] \\ = \ & \frac{1}{m}E_{S_q,S_p,S_q^{\text{test}}}\sum_{i=1}^{m}\{\widehat{D}(\tilde{Z}_{q,i}) - D^*(\tilde{Z}_{q,i})\}^2 \\ = \ & E_{S_q,S_p,Z\sim q^*}\{\widehat{D}(Z) - D^*(Z)\}^2 \\ = \ & E_{S_q,S_p}\|\widehat{D} - D^*\|_q^2 \\ \le \ & E_{S_q,S_p}\|\widehat{D} - D^*\|_{\max}^2 \\ = \ & E_{S_p,S_q}\|\widehat{D} - D^*\|_{\max}^2 \\ = \ & O\left(n^{-\frac{2\beta}{d+2\beta}}\log^3 n\right). \end{aligned}$$

By Markov's inequality, (47) holds and the asymptotic normality follows from the classic central limit theorem. The proof of the theorem is completed. ■

Proof [Proof of Theorem 13] Without loss of generality, we assume $C = c = 1$. For any $D \in \mathcal{F}_{\text{FNN}}$ specified in the theorem, we have

$$\begin{aligned}
&E_{p^*}[D(Z) - D^*(Z)]^2 \\
&\leq E_{p^*}[\{D(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \geq \log n)] + E_{p^*}[\{D(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \leq \log n)] \\
&\leq 4M^2 E_{p^*}\mathbb{I}(\|Z\|_\infty \geq \log n) + E_{p^*}[\{D(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \leq \log n)],
\end{aligned}$$

where the second inequality follows from the facts that $\|D^*\|_\infty \leq M, \|D\|_\infty \leq M$ under Assumption 5. Since $D^* \in \mathcal{H}^\beta(\mathbb{R}^d, M)$, $D^*(2t\log n - \log n\mathbf{1}_d) \in \mathcal{H}^\beta([0,1]^d, (2\log n)^{\lfloor\beta\rfloor}M)$ as a function of t, where $\mathbf{1}_d$ is the d-dimensional all-one vector. By Lemma 9, there exists a function $\phi_0 \in \mathcal{F}_{\text{FNN}}$ such that

$$\sup_{t\in[0,1]^d\setminus H_{B,\delta}} |D^*(2t\log n - \log n\mathbf{1}_d) - \phi_0| \leq 18(2\log n)^{\lfloor\beta\rfloor}M(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2} n^{-\frac{\beta}{d+2\beta}},$$

where $H_{B,\delta} = \cup_{i=1}^d \big\{t = [t_1,\ldots,t_d] : t_i \in \cup_{b=1}^{B-1}(b/B - \delta, b/B)\big\}, B = \lceil n^{\frac{1}{d+2\beta}}\rceil, \delta \in (0, 1/(3B)]$. Thus

$$\sup_{z\in[-\log n,\log n]^d\setminus\tilde{H}^d_{B,\epsilon}} \left|D^*(z) - \phi_0\left(\frac{z+\log n\mathbf{1}_d}{2\log n}\right)\right| \leq 18(2\log n)^{\lfloor\beta\rfloor}M(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2} n^{-\frac{\beta}{d+2\beta}},$$

where $\tilde{H}^d_{B,\delta} = \left\{2t\log n - \log n : t \in H^d_{B,\delta}\right\}$. Let $\tilde{\phi}_0(z) = \phi_0\left(\frac{z+\log n\mathbf{1}_d}{2\log n}\right) \in \mathcal{F}_{\text{FNN}}$. As δ can be arbitrarily small, it then follows from similar lines as in the proof of Theorem 10 that

$$E_{p^*}[\{\tilde{\phi}_0(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \leq \log n)] \leq 324M^2(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee 1)}(2\log n)^{2\lfloor\beta\rfloor} n^{-\frac{2\beta}{d+2\beta}}.$$

Similarly, we can obtain

$$E_{q^*}[\{\tilde{\phi}_0(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \leq \log n)] \leq 324M^2(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee 1)}(2\log n)^{2\lfloor\beta\rfloor} n^{-\frac{2\beta}{d+2\beta}}.$$

Since $D_{\text{NN}} \in \operatorname{argmin}_{D\in\mathcal{F}_{\text{FNN}}} \|D - D^*\|_{\max}$, we have

$$\begin{aligned}
&\|D_{\text{NN}} - D^*\|^2_{\max} \\
&\leq \|\tilde{\phi}_0 - D^*\|^2_{\max} \\
&\leq \max_{h=p,q}\{4M^2 E_{h^*}\mathbb{I}(\|Z\|_\infty \geq \log n) + E_{h^*}[\{\tilde{\phi}_0(Z) - D^*(Z)\}^2\mathbb{I}(\|Z\|_\infty \leq \log n)]\} \\
&\leq 328M^2(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+(\beta\vee 1)}(2\log n)^{2\lfloor\beta\rfloor} n^{-\frac{2\beta}{d+2\beta}}.
\end{aligned}$$

Then using Theorem 6 and

$$\xi_n^2 \leq C(\lfloor\beta\rfloor+1)^9 d^{2\lfloor\beta\rfloor+3} n^{-\frac{2\beta}{d+2\beta}} \log^3 n$$

from the proof of Theorem 10, where C is a constant depending only on (μ, σ, M), we can obtain (16). This proves (16). ■

Proof [Proof of Theorem 15] Based on Theorem 3.1 in Baraniuk and Wakin (2009), fix $\delta \in (0,1)$ and $\gamma \in (0,1)$. Let $A = \sqrt{d/d_\delta}\Phi$, where $\Phi \in \mathbb{R}^{d_\delta \times d}$ is a random orthogonal projection with

$$d_{\delta,\gamma} = O\left(\frac{d_{\mathcal{M}} \ln\left(dV\mathcal{R}\tau^{-1}\delta^{-1}\right)\ln(1/\gamma)}{\delta^2}\right).$$

If $d_{\delta,\gamma} \le d$, then with probability at least $1-\gamma$: For every $x, y \in \mathcal{M}_{\tau,V,\mathcal{R}}$

$$(1-\delta)\|x-y\|_2 \le \|Ax - Ay\|_2 \le (1+\delta)\|x-y\|_2.$$

This result shows that if for $\delta \in (0,1)$,

$$d_\delta = O\left(\frac{d_{\mathcal{M}} \ln\left(dV\mathcal{R}\tau^{-1}\delta^{-1}\right)}{\delta^2}\right) \ll d,$$

then there exists a linear projection $A \in \mathbb{R}^{d_\delta \times d}$ such that $AA^T = dI_{d_\delta}/d_\delta$, where $I_{d_\delta} \in \mathbb{R}^{d_\delta \times d_\delta}$ is an identity matrix, and for any $x, y \in \mathcal{M}_{\tau,V,\mathcal{R}}$,

$$(1-\delta)\|x-y\|_2 \le \|Ax - Ay\|_2 \le (1+\delta)\|x-y\|_2. \tag{48}$$

Then we have

$$A(\mathcal{M}_{\rho,\tau,V,\mathcal{R}}) \subseteq A\left([0,1]^d\right) \subseteq \left[-\frac{d}{\sqrt{d_\delta}}, \frac{d}{\sqrt{d_\delta}}\right]^{d_\delta}.$$

Note that for any $z \in A(\mathcal{M}_{\tau,V,\mathcal{R}})$, there exits a unique $x \in \mathcal{M}_{\tau,V,\mathcal{R}}$ such that $z = Ax$. Otherwise, suppose we can find $x, x' \in \mathcal{M}_{\tau,V,\mathcal{R}}, x \neq x'$ such that $z = Ax = Ax'$, then by (48), we know $\|x - x'\|_2 = 0$ and thus $x = x'$, which contradicts the assumption that $x \neq x'$. This uniqueness allows us to define a linear operator $\mathcal{SL} : A(\mathcal{M}_{\tau,V,\mathcal{R}}) \to \mathcal{M}_{\tau,V,\mathcal{R}}$ such that $A[\mathcal{SL}(z)] = z$. By (48), we have

$$(1-\delta)\|\mathcal{SL}(z_1) - \mathcal{SL}(z_2)\|_2 \le \|z_1 - z_2\|_2 \le (1+\delta)\|\mathcal{SL}(z_1) - \mathcal{SL}(z_2)\|_2.$$

This implies that the norm of $\mathcal{SL}$ belongs to $[1/(1+\delta), 1/(1-\delta)]$. For the high-dimensional function $D^* : [0,1]^d \to \mathbb{R}$ whose support is $\mathcal{M}_{\rho,\tau,V,\mathcal{R}}$, it has a approximate low-dimensional representation $\tilde{D}^*$ as follows:

$$\tilde{D}^*(z) = D^*(\mathcal{SL}(z)), \ \forall \ z \in A(\mathcal{M}_{\tau,V,\mathcal{R}}).$$

As $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$, we have $\tilde{D}^* \in \mathcal{H}^\beta\left(A(\mathcal{M}_{\tau,V,\mathcal{R}}), M/(1-\delta)^\beta\right)$. By the extended version of Whitney's extension theorem in Fefferman (2006), since $A(\mathcal{M}_{\tau,V,\mathcal{R}}) \subseteq A\left([0,1]^d\right) \subseteq \left[-d/\sqrt{d_\delta}, d/\sqrt{d_\delta}\right]^{d_\delta}$, there exists $\tilde{D}^*_E \in \mathcal{H}^\beta\left(\left[-d/\sqrt{d_\delta}, d/\sqrt{d_\delta}\right]^{d_\delta}, M/(1-\delta)^\beta\right)$ such that $\tilde{D}^*_E \equiv \tilde{D}^*$ on $A(\mathcal{M}_{\tau,V,\mathcal{R}})$. If $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 {d_\delta}^{\lfloor\beta\rfloor+1} L\lceil\log_2(8L)\rceil$ and $\mathcal{D} = 21(\lfloor\beta\rfloor + 1)^2 K\lceil\log_2(8K)\rceil$, by the first result of Lemma 9, there exists a function ϕ_0 implemented by a ReLU network with width $\mathcal{W}$ and depth $\mathcal{D}$ such that

$$\sup_{z \in [0,1]^{d_\delta} \setminus H^{d_\delta}_{B,\epsilon}} \left|\tilde{D}^*_E\left(\frac{2dz - d\mathbf{1}_{d_\delta}}{\sqrt{d_\delta}}\right) - \phi_0(z)\right| \le \frac{18M}{(1-\delta)^\beta}(\lfloor\beta\rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor\beta\rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}}.$$

where $H_{B,\epsilon}^{d_\delta} = \cup_{i=1}^{d_\delta} \big\{ x = [x_1, x_2, \ldots, x_{d_\delta}] : x_i \in \cup_{b=1}^{B-1} \left(b/B - \epsilon, b/B\right) \big\}$ and $B = \lceil (KL)^{2/d} \rceil, \epsilon \in (0, 1/(3B)]$. Thus

$$\begin{aligned} & \sup_{z \in \left[-\frac{d}{\sqrt{d_\delta}}, \frac{d}{\sqrt{d_\delta}}\right]^{d_\delta} \backslash \tilde{H}_{B,\epsilon}^{d_\delta}} \left| \tilde{D}_E^*(z) - \phi_0\left(\frac{\sqrt{d_\delta} z + d\mathbf{1}_{d_\delta}}{2d} \right) \right| \\ \leq \; & \frac{18M}{(1-\delta)^\beta} (\lfloor \beta \rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor \beta \rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}}, \end{aligned}$$

where $\tilde{H}_{B,\epsilon}^{d_\delta} = \left\{ (2dt - d\mathbf{1}_{d_\delta})/\sqrt{d_\delta} : t \in H_{B,\epsilon}^{d_\delta} \right\}$.

Let

$$\tilde{\phi}_0(x) = \phi_0\left((\sqrt{d_\delta} A x + d\mathbf{1}_{d_\delta})/(2d) \right)$$

and

$$\bar{H}_{*B,\epsilon}^{d} = \left\{ x \in [0,1]^{d \times d} : (\sqrt{d_\delta} A x + d\mathbf{1}_{d_\delta})/(2d) \in H_{B,\epsilon}^{d_\delta} \right\}.$$

It can be easily checked that $\tilde{\phi}_0$ is also a function implemented by a ReLU network with the same structure as ϕ_0, except that the input layer of $\tilde{\phi}_0$ has d units, instead of d_δ units. For any $x \in \mathcal{M}_{\rho,\tau,V,\mathcal{R}} \backslash \bar{H}_{*B,\epsilon}^{d}$, $Ax \in \left[-d/\sqrt{d_\delta}, d/\sqrt{d_\delta}\right]^{d_\delta} \backslash \tilde{H}_{B,\epsilon}^{d_\delta}$ and there exists a $x' \in \mathcal{M}_{\tau,V,\mathcal{R}}$ satisfying $\|x - x'\|_2 \leq \rho$. Since $\tilde{D}_E^* \in \mathcal{H}^\beta\left(\left[-d/\sqrt{d_\delta}, d/\sqrt{d_\delta}\right]^{d_\delta}, M/(1-\delta)^\beta \right)$ and $D^* \in \mathcal{H}^\beta([0,1]^d, M)$,

$$\begin{aligned} & |\tilde{\phi}_0(x) - D^*(x)| \\ \leq \; & |\tilde{\phi}_0(x) - \tilde{D}_E^*(Ax)| + |\tilde{D}_E^*(Ax) - \tilde{D}_E^*(Ax')| + |\tilde{D}_E^*(Ax') - D^*(x)| \\ \leq \; & \frac{18M}{(1-\delta)^\beta} (\lfloor \beta \rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor \beta \rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}} + \frac{M}{(1-\delta)^\beta} \|Ax' - Ax\|_2 + \rho M \\ \leq \; & \frac{18M}{(1-\delta)^\beta} (\lfloor \beta \rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor \beta \rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}} + \frac{M\sqrt{d}}{(1-\delta)^\beta \sqrt{d_\delta}} \rho + \rho M \\ \leq \; & \frac{18M}{(1-\delta)^\beta} (\lfloor \beta \rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor \beta \rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}} + \frac{2M\sqrt{d}}{(1-\delta)^\beta \sqrt{d_\delta}} \rho \\ \leq \; & \frac{20M}{(1-\delta)^\beta} (\lfloor \beta \rfloor + 1)^2 (2d)^\beta d_\delta^{\lfloor \beta \rfloor + (\beta \vee 1 + \beta)/2} (KL)^{-\frac{2\beta}{d_\delta}}, \end{aligned} \tag{49}$$

where the last inequality holds when $\rho \leq (\lfloor \beta \rfloor + 1)^2 2^\beta d^{\beta - \frac{1}{2}} d_\delta^{\lfloor \beta \rfloor + (\beta - 1/2) \vee (1/2)} (KL)^{-\frac{2\beta}{d_\delta}}$. As $D_{\mathrm{NN}} \in \operatorname{argmin}_{D \in \mathcal{F}_{\mathrm{FNN}}} \|D - D^*\|_{\max}$,

$$\|D_{\mathrm{NN}} - D^*\|_{\max}^2 \leq \|\tilde{\phi}_0 - D^*\|_{\max}^2.$$

By the result in (49), for $I = p$ or q, it holds

$$\begin{aligned} \|\tilde{\phi}_0 - D^*\|_I^2 &= \int_{[0,1]^d \backslash H_{B,\delta}} |D^* - \tilde{\phi}_0|^2 I^*(x) dx + \int_{H_{B,\delta}} |D^* - \tilde{\phi}_0|^2 I^*(x) dx \\ &\leq \frac{400M^2}{(1-\delta)^{2\beta}} (\lfloor \beta \rfloor + 1)^4 (2d)^{2\beta} d_\delta^{\beta \vee 1 + 3\beta} (KL)^{-\frac{4\beta}{d_\delta}} + \frac{4M^2}{(1-\delta)^{2\beta}} \int_{\bar{H}_{*B,\epsilon}^{d}} I^*(x) dx. \end{aligned}$$

As $p^*(\cdot), q^*(\cdot)$ are the density functions of some measures on $[0,1]^d$ which are absolutely continuous w.r.t the Lebesgue measure and ϵ can be arbitrarily small for the given δ, $\int_{\bar{H}^d_{*B,\epsilon}} I_0(x)dx$ is also arbitrarily small for the given δ. Thus we have

$$\|\tilde{\phi}_0 - D^*\|_I^2 \;\leq\; \frac{400M^2}{(1-\delta)^{2\beta}}(\lfloor\beta\rfloor+1)^4(2d)^{2\beta}d_\delta^{\beta\vee 1+3\beta}(KL)^{-\frac{4\beta}{d_\delta}}$$

and

$$\begin{aligned}\|D_{\mathrm{NN}} - D^*\|_{\max}^2 &\leq \|\phi_0 - D^*\|_{\max}^2\\ &\leq \frac{400M^2}{(1-\delta)^{2\beta}}(\lfloor\beta\rfloor+1)^4(2d)^{2\beta}d_\delta^{\beta\vee 1+3\beta}(KL)^{-\frac{4\beta}{d_\delta}}\\ &= \frac{400M^2}{(1-\delta)^{2\beta}}C_2(\beta,d,d_\delta)(KL)^{-\frac{4\beta}{d_\delta}}.\end{aligned}$$

By Corollary 7, there exists a constant C_1 only depending on (μ,σ,M), such that

$$\begin{aligned}&E_{S_p,S_q}\|\widehat{D} - D^*\|_{\max}^2\\ \leq\; & C_1\left(\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n} + \|D_{\mathrm{NN}} - D^*\|_{\max}^2\right)\\ \leq\; & C_1\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n} + \frac{400M^2}{(1-\delta)^{2\beta}}C_2(\beta,d,d_\delta)(KL)^{-\frac{4\beta}{d_\delta}}\right\}\\ \leq\; & \frac{400M^2C_1}{(1-\delta)^{2\beta}}\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n} + C_2(\beta,d,d_\delta)(KL)^{-\frac{4\beta}{d_\delta}}\right\}.\end{aligned} \tag{50}$$

For

$$\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 d_\delta{}^{\lfloor\beta\rfloor+1},$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\left\lceil n^{\frac{d_\delta}{2(d_\delta+2\beta)}}\log_2\left(8n^{\frac{d_\delta}{2(d_\delta+2\beta)}}\right)\right\rceil,$$

and $\mathcal{W}, \mathcal{D}$ satisfy

$$\mathcal{O}(\mathcal{W}^2\mathcal{D}) = \mathcal{O}\left((\lfloor\beta\rfloor+1)^6 d_\delta^{2\lfloor\beta\rfloor+2}\left\lceil n^{\frac{d_\delta}{2(d_\delta+2\beta)}}\log n\right\rceil\right),$$

along the derivation of (43), there exists a universal constants C^* such that

$$\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n} \leq C^*(\lfloor\beta\rfloor+1)^9 d_\delta^{2\lfloor\beta\rfloor+3} n^{-\frac{2\beta}{d_\delta+2\beta}}\log^3 n.$$

Based on the result of (50),

$$\begin{aligned}&E_{S_p,S_q}\|\widehat{D} - D^*\|_{\max}^2\\ \leq\; & \frac{400M^2C_1}{(1-\delta)^{2\beta}}\left\{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n} + C_2(\beta,d,d_\delta)(KL)^{-\frac{4\beta}{d_\delta}}\right\}\\ \leq\; & \frac{400M^2C_1}{(1-\delta)^{2\beta}}\left\{C^*(\lfloor\beta\rfloor+1)^9 d_\delta^{2\lfloor\beta\rfloor+3}n^{-\frac{2\beta}{d_\delta+2\beta}}\log^3 n + C_2(\beta,d,d_\delta)n^{-\frac{2\beta}{d_\delta+2\beta}}\right\}\\ \leq\; & \frac{800M^2C_1C^*}{(1-\delta)^{2\beta}}(\lfloor\beta\rfloor+1)^9\max\left\{d_\delta^{2\lfloor\beta\rfloor+3}, (2d)^{2\beta}d_\delta^{\beta\vee 1+3\beta}\right\}n^{-\frac{2\beta}{d_\delta+2\beta}}\log^3 n\\ =\; & \frac{800M^2C_1C^*C_3(\beta,d,d_\delta)}{(1-\delta)^{2\beta}}n^{-\frac{2\beta}{d_\delta+2\beta}}\log^3 n.\end{aligned}$$

This completes the proof of the theorem and (17). ■

Lemma 20 *The following excess risk decomposition always holds:*

$$\mathcal{B}_\psi\left(e^{\widehat{D}}\right)-\mathcal{B}_\psi\left(e^{D^*}\right)=\left\{\mathcal{B}_\psi\left(e^{\widehat{D}}\right)-\inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\mathcal{B}_\psi\left(e^{D}\right)\right\}+\left\{\inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\mathcal{B}_\psi\left(e^{D}\right)-\mathcal{B}_\psi\left(e^{D^*}\right)\right\}.$$

Under Assumptions 4 *and* 16*, when* $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$*, there exist three constants* C, C_0, C_**, with* C, C_0 *depending only on* (μ,σ,M) *and* C_* *depending only on* (μ,σ)*, such that*

$$E_{S_p,S_q}\left\{\mathcal{B}_\psi\left(e^{\widehat{D}}\right)-\inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\mathcal{B}_\psi\left(e^{D}\right)\right\}\leq C\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}, \tag{51}$$

and

$$E_{S_p,S_q}\|\widehat{D}-D^*\|_p^2\leq C_0\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})\log n}{n}}+C_*e^{2M}\inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\|e^D-e^{D^*}\|_p^2.$$

Proof [Proof of Lemma 20] To show (51) is the key step in the proof of this theorem, thus we focus on the proof of (51). Let

$$D_0\in\underset{D\in\mathcal{F}_{\mathrm{FNN}}}{\mathrm{argmin}}\,\mathcal{B}_\psi\left(e^D\right).$$

Then,

$$\begin{aligned}
&E_{S_p,S_q}\left\{\mathcal{B}_\psi\left(e^{\hat{D}}\right)-\inf_{D\in\mathcal{F}_{\mathrm{FNN}}}\mathcal{B}_\psi\left(e^{D}\right)\right\}\\
&=E_{S_p,S_q}\left\{\mathcal{B}_\psi\left(e^{\hat{D}}\right)-\mathcal{B}_\psi\left(e^{D_0}\right)\right\}\\
&\leq E_{S_p,S_q}\left\{\mathcal{B}_\psi\left(e^{\hat{D}}\right)-\hat{\mathcal{B}}_\psi\left(e^{\hat{D}}\right)+\hat{\mathcal{B}}_\psi\left(e^{\hat{D}}\right)-\hat{\mathcal{B}}_\psi\left(e^{D_0}\right)\right\}\\
&\quad+E_{S_p,S_q}\left\{\hat{\mathcal{B}}_\psi\left(e^{D_0}\right)-\mathcal{B}_\psi\left(e^{D_0}\right)\right\}\\
&\leq 2E_{S_p,S_q}\left\{\sup_{D\in\mathcal{F}_{\mathrm{FNN}}}|\hat{\mathcal{B}}_\psi\left(e^{D}\right)-\mathcal{B}_\psi\left(e^{D}\right)|\right\}.
\end{aligned}\tag{52}$$

By the symmetrization technique, Talagrand's lemma, (35) and the fact that $\|D\|_\infty \leq M$ for any $D\in\mathcal{F}_{\mathrm{FNN}}$, we can show inequality (51) based on (52). ■

Proof [Proof of Theorem 17] Theorem 17 is a direct corollary of Lemma 20. We omit the details here. ■

Proof [Proof of Proposition 18] For $k=0,\ldots,K-2$, the densities $q_k(), q_{k+1}()$ of the synthetic data $\{Z_{k,j}\}_{j=1}^n$ and $\{Z_{k+1,j}\}_{j=1}^n$ satisfy

$$\frac{q_k(t)}{q_{k+1}(t)}=\frac{(1-\alpha_k)q^*(z)+\alpha_kp^*(z)}{(1-\alpha_{k+1})q^*(z)+\alpha_{k+1}p^*(z)}\in\left[\frac{(1-e^{-M})\alpha_k+e^{-M}}{(1-e^{-M})\alpha_{k+1}+e^{-M}},\frac{1-\alpha_k}{1-\alpha_{k+1}}\right].$$

As $\|f\|_2 = (\int_{\mathcal{Z}} f^2(x)dx)^{\frac{1}{2}}$, then for any density g satisfying $g \geq c$, $\|f\|_2 = \left(\int_{\mathcal{Z}} f^2(x)dx\right)^{\frac{1}{2}} \leq \left(\int_{\mathcal{Z}} f^2(x)g(x)/cdx\right)^{\frac{1}{2}} = \|f\|_g/\sqrt{c}$. Using an appropriate $\mathcal{F}_{\text{FNN}}$ whose element D satisfies $\|D\|_\infty \leq M$, for the direct estimate $\widehat{D}_{\text{SRE}}$, as $\log(q^*/p^*)$ is only bounded from below by $-M_0$, by Theorem 17, we have

$$\limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_{\text{SRE}} - D^*\|_2 \leq e^M C_*(\mu,\sigma,c_1)\|R^* - R^*_M\|_p.$$

For $k = 0, 1, \ldots, K-2$, as $|\log\{q_k(t)/q_{k+1}(t)\}|$ is bounded by M, by Corollary 7, we have

$$\limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_k - D^*_k\|_2 = 0.$$

Let $R^*_{K-1,M} = (1-\alpha_{K-1})R^*_M + \alpha_{K-1}$. As the logarithm of $R^*_{K-1} = (1-\alpha_{K-1})q^*/p^* + \alpha_{K-1}$ is also only bounded from below by $-M$, again, by Theorem 17,

$$\begin{aligned}\limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_{K-1} - D^*_{K-1}\|_2 &\leq e^M C_*(\mu,\sigma,c_1)\|R^*_{K-1} - R^*_{K-1,M}\|_p \\ &= (1-\alpha_{K-1})e^M C_*(\mu,\sigma,c_1)\|R^* - R^*_M\|_p.\end{aligned}$$

Thus

$$\begin{aligned}\limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_{\text{TRE}} - D^*\|_2 &\leq \sum_{k=0}^{K-1}\limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_k - D^*_k\|_2 \\ &= \limsup_{n\to\infty} E_{S_p,S_q}\|\widehat{D}_{K-1} - D^*_{K-1}\|_2 \\ &\leq (1-\alpha_{K-1})e^M C_*(\mu,\sigma,c_1)\|R^* - R^*_M\|_p.\end{aligned}$$

This completes the proof of Proposition 18. ■

Proof [Proof of (23)] Recall that $\varepsilon \sim \text{TN}_1(1,1)$ has a density function:

$$\phi_{\text{TN},d}(v) = \frac{1}{c_{1,1}}\exp\left(-\frac{v^2}{2}\right), \text{ for } |v| \leq 1,$$

where $c_{1,1} = \int_{-1}^{1}\exp\left(-v^2/2\right)dv$. Denote the density function of X by ψ_X.

For q^*, the density for the latent variable pair (Y_1, X) is

$$f_{q^*}(y,x) = \frac{1}{c_{1,1}}\exp\left[-\frac{\{y - f_0(x)\}^2}{2}\right]\psi_X(x).$$

Hence, q^* is expressed by

$$q^*(y,x) = \frac{1}{c_{q^*,d}}\exp\left[-\frac{\{y - f_0(x)\}^2}{2}\right]\psi_X(x),$$

where $c_{q^*,d} = \int_{[a_1,a_2]\times\mathcal{X}}\exp\left[-\{y - f_0(x)\}^2/2\right]\psi_X(x)dydx$. Similarly,

$$p^*(y,x) = \frac{1}{c_{p^*,d}}\exp\left[-\frac{\{y - f_0(x) - 1\}^2}{2}\right]\psi_X(x),$$

where $c_{p^*,d} = \int_{[a_1,a_2]\times\mathcal{X}} \exp\left[-\{y - f_0(x) - 1\}^2/2\right] \psi_X(x) dydx$. Then,

$$\begin{aligned} D^*(y,x) = \log \frac{q^*(y,x)}{p^*(y,x)} &= \log \left(\frac{\frac{1}{c_{q^*,d}} \exp\left[-\frac{\{y-f_0(x)\}^2}{2}\right] \psi_X(x)}{\frac{1}{c_{p^*,d}} \exp\left[-\frac{\{y-f_0(x)-1\}^2}{2}\right] \psi_X(x)} \right) \\ &= \log \left(\frac{\exp\left[-\frac{\{y-f_0(x)\}^2}{2}\right]}{\exp\left[-\frac{\{y-f_0(x)-1\}^2}{2}\right]} \right) + \log\left(\frac{c_{p^*,d}}{c_{q^*,d}}\right) \\ &= f_0(x) - y + \frac{1}{2} + \log\left(\frac{c_{p^*,d}}{c_{q^*,d}}\right). \end{aligned}$$

Hence, (23) is proved. ■

Appendix B. Examples of Hölder function class

Let p^* be the density function of a truncated d-dimensional multivariate Gaussian with mean zero and covariance matrix $\Sigma_p \in \mathbb{R}^{d\times d}$ in $[0,1]^d$. That means

$$p^*(z) = \frac{\exp\left(-\frac{z'\Sigma_p^{-1}z}{2}\right)}{c(\Sigma_p)}, \quad c(\Sigma_p) = \int_{[0,1]^d} \exp\left(-\frac{t'\Sigma_p^{-1}t}{2}\right) dt, \ z \in [0,1]^d.$$

Similarly, let

$$q^*(z) = \frac{\exp\left(-\frac{z'\Sigma_q^{-1}z}{2}\right)}{c(\Sigma_q)}$$

for some positive definite matrix Σ_q. For any matrix $A \in \mathbb{R}^{d\times d}$, $A_{i,\cdot}$ is the ith row of A for $i = 1, 2, \ldots, d$ and

$$\|A\|_{2,\infty} := \sup_{\|z\|_\infty \le 1} \|Az\|_2.$$

Then,

$$D^*(z) = \log \frac{q^*(z)}{p^*(z)} = \frac{1}{2} z'(\Sigma_p^{-1} - \Sigma_q^{-1})z + \log(c(\Sigma_p) - c(\Sigma_q)), \ z \in [0,1]^d.$$

Let

$$M = \max\left\{ \frac{1}{2}(\|\Sigma_p^{-1/2}\|_{2,\infty}^2 + \|\Sigma_q^{-1/2}\|_{2,\infty}^2) + |\log[c(\Sigma_p) - c(\Sigma_q)]|, \|(\Sigma_p^{-1} - \Sigma_q^{-1})_{i,\cdot}\|_2, i = 1, 2, \ldots, d \right\}.$$

It is straightforward to check that

$$D^* \in \mathcal{H}^2([0,1]^d, M).$$

It implies the Hölder smoothness parameter β is 2 for this example.

Moreover, the truncated multivariate Gaussian distributions considered above are special cases of the exponential distribution class defined below. Define the density function class

$$\mathrm{Exp}(\beta, B) := \left\{ p(z) = \exp(g(z))/c_g : z \in [0,1]^d, c_g = \int_{[0,1]^d} \exp(g(t))dt, g \in \mathcal{H}^\beta([0,1]^d, B) \right\}.$$

Suppose that $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite and let $g(z) = z'\Sigma z/2$. Then, $g \in \mathcal{H}^2([0,1]^d, M_\Sigma)$, where $M_\Sigma = \max\left\{\frac{1}{2}(\|\Sigma^{1/2}\|_{2,\infty}^2, \|\Sigma_{i,\cdot}\|_2, i = 1, 2, \dots, d\right\}$. If $p^*, q^* \in \mathrm{Exp}(\beta, B)$, we have $D^*(z) = \log[q^*(z)/p^*(z)] \in \mathcal{H}^\beta([0,1]^d, 4B)$.

Appendix C. Examples of random variables satisfying Assumption 12

Let $Z = (Z_1, Z_2, \dots, Z_d)$. If Z_i is a sub-Gaussian variable with mean 0, there exists $\sigma_i > 0$ such that

$$\mathbb{P}(|Z_i| \geq t) \leq 2e^{-\frac{t^2}{2\sigma_i^2}} \quad \text{for all } t \geq 0.$$

Then, by Bonferroni's inequality, for any fixed d and $\beta < \infty$,

$$E\mathbb{I}\{\|Z\|_\infty \geq \sqrt{2} \max_{i=1,\dots,d} \sigma_i \log n\} \leq E\mathbb{I}\{\|Z\|_\infty \geq \sigma\sqrt{2\log n}\} \leq 2dn^{-1} \leq 2dn^{-\frac{2\beta}{d+2\beta}}.$$

As a result, Assumption 12 holds for the multivariate sub-Gaussian distribution. The famous multivariate sub-Gaussian distribution is the multivariate normal distribution.

In addition, it can be proved that sub-exponential variables also satisfy Assumption 12. When Z_i is a sub-exponential variable with mean 0, by Theorem 2.13 of Wainwright (2019), there exist c_{i1} and c_{i2} such that

$$\mathbb{P}(|Z_i| \geq t) \leq c_{i1}e^{-c_{i2}t} \quad \text{for all } t \geq 0.$$

Then for $Z = (Z_1, Z_2, \dots, Z_d)$, by Bonferroni's inequality again,

$$E\mathbb{I}\left\{\|Z\|_\infty \geq \frac{\log n}{\min_{i=1,\dots,d} c_{i2}}\right\} \leq \left(\sum_{i=1}^d c_{i1}\right) n^{-1} \leq \left(\sum_{i=1}^d c_{i1}\right) n^{-\frac{2\beta}{d+2\beta}},$$

for any fixed d and $\beta < \infty$. Examples of sub-exponential variables include the exponential random variable and χ^2-random variable etc.

Appendix D. Extension to sparse neural networks

In this section, we extend our main results: Theorems 6 & 10, to the cases when the neural networks are sparse as considered by Schmidt-Hieber (2020). We first define the neural network: For $\mathcal{D} \in \mathbb{N}$ and $\mathbf{d} = (d_0, \dots, d_{\mathcal{D}+1}) \in \mathbb{N}^{\mathcal{D}+2}$ with $d_0 = d$,

$$\mathcal{F}_{\mathcal{D},\mathbf{d}}^S := \{f : x \mapsto A_{\mathcal{D}}\sigma_{v_{\mathcal{D}}}A_{\mathcal{D}-1}\sigma_{v_{\mathcal{D}-1}} \cdots A_1\sigma_{v_1}A_0x : A_i \in \mathbb{R}^{d_{i+1}\times d_i}, v_i \in \mathbb{R}^{d_i}, i = 0, \dots, \mathcal{D}\},$$

where $\sigma_v(t) := \sigma(t-v)$ and $\sigma(t) = \max\{t, 0\}$ is applied componentwisely. Then, for $s \in \mathbb{N}, M \geq 0, \mathcal{D} \in \mathbb{N}$, and $\mathbf{d} \in \mathbb{N}^{\mathcal{D}+2}$, define

$$\mathcal{H}_{\mathcal{D},\mathbf{d},s,M} := \left\{ f \in \mathcal{F}^S_{\mathcal{D},\mathbf{d}} : \sum_{j=0}^{\mathcal{D}} \|A_j\|_0 + \|v_j\|_0 \leq s, \max_{j=0,\ldots,L} \|A_j\|_\infty \vee \|v_j\|_\infty \leq 1, \|f\|_\infty \leq M \right\},$$

where $\|A_j\|_0, \|v_j\|_0$ denote the numbers of non-zero entries of A_j and v_j, respectively, and $\|A_j\|_\infty, \|v_j\|_\infty$ represent the maximums of the absolute entries of A_j and v_j, respectively.

The condition $\sum_{j=0}^{\mathcal{D}} \|A_j\|_0 + \|v_j\|_0 \leq s$ implies that the total number of nonzero parameters in the neural network is bounded by s. It implies that the majority of the weights in the network are zero, corresponding to a sparse structure. In other words, the sparsity of a neural network is enforced by regulating the number of nonzero parameters to be less than the specified value s.

Denote

$$\xi_{S,n} = \sqrt{\frac{(s+1)\log(n(\mathcal{D}+1)V)}{n}} \text{ and } D^S_{\mathrm{NN}} \in \underset{D \in \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}}{\operatorname{argmin}} \|D - D^*\|_{\max}, \tag{53}$$

where $V := \prod_{j=0}^{\mathcal{D}+1} (d_j + 1)$.

Theorem 21 *Suppose Assumptions 4 and 5 hold. If $\mathcal{F}_n$ in (5) is replaced by $\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$, there exists a constant C depending on (μ, σ, M) such that for any $\gamma > 0$, with probability at least* $1 - \exp(-\gamma)$*,*

$$\|\widehat{D} - D^*\|_{\max} \leq C \left(\xi_{S,n} + \|D^S_{\mathrm{NN}} - D^*\|_{\max} + \sqrt{\frac{\gamma}{n}} \right),$$

and

$$\|\widehat{D} - D^*\|_{n_p,n_q} \leq 2C \left(\xi_{S,n} + \|D^S_{\mathrm{NN}} - D^*\|_{\max} + \sqrt{\frac{\gamma}{n}} \right).$$

Proof With sparse neural networks, the proof of is structurally similar to those for Theorem 6 except for a main difference in (35). Specifically, (35) will be replaced by the fact that, when $r_0 \geq 1/n$,

$$E_\eta R_{n_I}\{(D - D^*) : D \in \hat{\mathcal{F}}_I^{D^*,2r_0}\} \leq 64 r_0 \sqrt{\frac{(s+1)\log(n(\mathcal{D}+1)V)}{n}}, \quad I = p, q. \tag{54}$$

In the following, we will focus on proving (54). Recall $\hat{\mathcal{F}}_I^{D^*,r} = \{D \in \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}, \|D - D^*\|_{I,n_I} \leq r\}$. Using Dudley's Chaining in Lemma 26 and the covering entropy bound in

Lemma 27, we have

$$
\begin{aligned}
& E_\eta R_{n_I}\{(D-D^*): D \in \hat{\mathcal{F}}_I^{D^*,2r_0}\} \\
\leq \;& \inf_{0<\alpha<2r_0}\left\{4\alpha + \frac{12}{\sqrt{n_I}}\int_\alpha^{2r_0}\sqrt{\log \mathcal{N}\left(\delta, \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}, \|\cdot\|_{I,n_I}\right)}d\delta\right\} \\
\leq \;& \inf_{0<\alpha<2r_0}\left\{4\alpha + \frac{12}{\sqrt{n_I}}\int_\alpha^{2r_0}\sqrt{\log \mathcal{N}\left(\delta, \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}|_{Z_{I,1},\ldots,Z_{I,n_I}}, \infty\right)}d\delta\right\} \\
\leq \;& \inf_{0<\alpha<2r_0}\left\{4\alpha + \frac{12}{\sqrt{n_I}}\int_\alpha^{2r_0}\sqrt{\log \mathcal{N}\left(\delta, \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}, \|\cdot\|_{\infty}\right)}d\delta\right\} \\
\leq \;& \inf_{0<\alpha<2r_0}\left\{4\alpha + \frac{12}{\sqrt{n_I}}\int_\alpha^{2r_0}\sqrt{(s+1)\log\left(2\delta^{-1}(L+1)V^2\right)}d\delta\right\} \\
\leq \;& \frac{32r_0\sqrt{(s+1)\log\left(2n(\mathcal{D}+1)V^2/r_0\right)}}{\sqrt{n}}
\end{aligned}
$$

where the last inequality follows from a specific choice $\alpha = 2r_0/n$ and $n \leq n_I$. When $r_0 \geq 1/n$, we have

$$
\begin{aligned}
\frac{32r_0\sqrt{(s+1)\log\left(2n(\mathcal{D}+1)V^2/r_0\right)}}{\sqrt{n}} \;&\leq\; \frac{32r_0\sqrt{(s+1)\log\left(2n^2(\mathcal{D}+1)V^2\right)}}{\sqrt{n}} \\
&\leq\; \frac{64r_0\sqrt{(s+1)\log\left(n(\mathcal{D}+1)V\right)}}{\sqrt{n}}.
\end{aligned}
$$

This finishes the proof of (54). The rest of the proof is the same as those for Theorem 6 and we omit them to avoid repeated discussions. ■

Theorem 22 (Non-asymptotic error bound for $\widehat{D}$ using sparse neural networks) *Suppose that Assumptions 4 and 5 hold, $D^* \in \mathcal{H}^\beta([0,1]^d, M)$ with $\beta = k + a$ where $k \in \mathbb{N}^+$ and $a \in (0,1]$, and $\mathcal{F}_n$ in (5) is replaced by $\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ which satisfies*

$$\mathcal{D} = 8 + \left(\lceil \log n\rceil + 5\right)\left(1 + \lceil \log_2(d \vee \beta)\rceil\right),$$

$\mathbf{d} = (d_0, \ldots, d_{\mathcal{D}+1}) \in \mathbb{N}^{\mathcal{D}+2},\ d_0 = d,\ d_i = 6(d+\lceil\beta\rceil)\lceil n\phi(n,\beta,d)\rceil,\ i = 1,2,\ldots,\mathcal{D}, d_{\mathcal{D}+1} = 1,$

and

$$s = 141(d+\beta+1)^{3+d}\lceil n\phi(n,\beta,d)\rceil(\lceil\log n\rceil + 6),$$

where $\phi(n,\beta,d) = n^{-2\beta/(2\beta+d)}$. Then,

$$E_{S_p,S_q}\|\widehat{D} - D^*\|^2_{\max} \leq C_0 C_1(\beta,d,M) n^{-\frac{2\beta}{d+2\beta}}\log^2 n,$$

where the constant C_0 depends only on (μ,σ,M) and

$$C_1(\beta,d,M) = \max\{(d+\beta+1)^{6+d}, 16(2M+1)^2\left(1+d^2+\beta^2\right)^2 6^{2d}, 4M^2 3^{2\beta} 2^{-\frac{2\beta}{d}}\}.$$

Proof [Proof of Theorem 22] Similar to Corollary 7, under the conditions of Theorem 21, there exists a constant C_0 depending only on (μ, σ, M), such that,

$$E_{S_p,S_q}\|\widehat{D}-D^*\|_{\max}^2 \leq C_0\left(\xi_{S,n}^2+\|D_{\mathrm{NN}}^S-D^*\|_{\max}^2\right). \tag{55}$$

An upper bound for $\xi_{S,n}^2$: By the expression of $\xi_{S,n}$ and the network structure in the theorem, we have

$$\begin{aligned}
\xi_{S,n}^2 &= \frac{(s+1)\log\left(n(\mathcal{D}+1)V\right)}{n}\\
&= \frac{\{141(d+\beta+1)^{3+d}\lceil n\phi(n,\beta,d)\rceil(\lceil\log n\rceil+6)+1\}\log\left(n(\mathcal{D}+1)V\right)}{n}\\
&\leq 4512(d+\beta+1)^{3+d}\phi(n,\beta,d)\log\left(n(\mathcal{D}+1)V\right)\\
&= 4512(d+\beta+1)^{3+d}\phi(n,\beta,d)\log\left(2n(d+1)(\mathcal{D}+1)(6(d+\lceil\beta\rceil)\lceil n\phi(n,\beta,d)\rceil+1)^{\mathcal{D}}\right)\\
&\leq C(d+\beta+1)^{6+d}\phi(n,\beta,d)\log^2 n\\
&= C(d+\beta+1)^{6+d}n^{-2\beta/(2\beta+d)}\log^2 n
\end{aligned}$$

where C is an absolute constant.

An upper bound for $\|D_{\mathrm{NN}}^S-D^*\|_{\max}$: By Lemma 28, there exists $\tilde{D}\in\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ such that

$$\begin{aligned}
\|\tilde{D}-D^*\|_\infty &\leq (2M+1)\left(1+d^2+\beta^2\right)6^d\lceil n\phi(n,\beta,d)\rceil/n+M3^\beta(\lceil n\phi(n,\beta,d)\rceil)^{-\frac{\beta}{r}}\\
&\leq 2\max\{2(2M+1)\left(1+d^2+\beta^2\right)6^d, M3^\beta 2^{-\frac{\beta}{r}}\}n^{-\beta/(2\beta+d)}.
\end{aligned}$$

By the definition of D_{NN}^S in (53), we know that

$$\begin{aligned}
\|D_{\mathrm{NN}}^S-D^*\|_{\max} &\leq \|\tilde{D}-D^*\|_{\max}\\
&\leq \|\tilde{D}-D^*\|_\infty\\
&\leq 2\max\{2(2M+1)\left(1+d^2+\beta^2\right)6^d, M3^\beta 2^{-\frac{\beta}{r}}\}n^{-\beta/(2\beta+d)}.
\end{aligned}$$

Combining the upper bounds for $\xi_{S,n}^2$ and $\|D_{\mathrm{NN}}^S-D^*\|_{\max}$ and using (55), we have that there exists an absolute constant C such that

$$\begin{aligned}
E_{S_p,S_q}\|\widehat{D}-D^*\|_{\max}^2 &\leq C_0\left(\xi_{S,n}^2+\|D_{\mathrm{NN}}^S-D^*\|_{\max}^2\right)\\
&\leq C_0CC_1(\beta,d,M)n^{-2\beta/(2\beta+d)}\log^2 n,
\end{aligned}$$

where

$$C_1(\beta,d,M)=\max\{(d+\beta+1)^{6+d}, 16(2M+1)^2\left(1+d^2+\beta^2\right)^2 6^{2d}, 4M^2 3^{2\beta}2^{-\frac{2\beta}{r}}\}.$$

This completes the proof of Theorem 22. ■

We apply Theorem 21 to the setups of Kato and Teshima (2021) and Kato et al. (2023), that is $D^*\in\mathcal{F}_n=\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ and $\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ is fixed. Similar to Corollary 19, we can show that $\|\widehat{R}-R^*\|_p=O_p\left(\sqrt{\log n/n}\right)=O_p\left(n^{-1/(2+a)}\right)$ for any $a>0$, which does not require the assumption that $a<2$ as in Theorem 2 in Kato and Teshima (2021) and Theorem 3.7 in Kato et al. (2023). Based on this result, similar to Theorem 11, we can establish the following corollary.

Corollary 23 *Suppose that $D^* \in \mathcal{F}_n = \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ and $\mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ is fixed. Let $m = n^b$ for any $0 < b < 1$. Then, for the independent data-based KL divergence estimator $\widetilde{\mathrm{KL}}_m(q^*||p^*)$ defined in (15), it holds that*

$$\sqrt{m}\left(\widetilde{\mathrm{KL}}_m(q^*\|p^*) - \mathrm{KL}(q^*\|p^*)\right) \to N(0,\sigma^2)$$

in distribution, where $\sigma^2 = E_{q^*}D^{*2}(Z) - E^2_{q^*}D^*(Z)$.

Since $\sqrt{m} = (\sqrt{n})^b$ and b can be arbitrarily close to 1, Corollary 23 provides a *nearly* $\sqrt{n}$-asymptotic normality which does not depend on the unknown parameters.

Appendix E. Supporting lemmas

We first show the following lemmas.

Lemma 24 1. *If the convex function $f : \mathbb{R} \to \mathbb{R}$ is μ-smooth over $\mathbb{R}$, then for any $x, y \in \mathbb{R}$, the following inequality holds*

$$f(y) \le f(x) + f'(x)(y-x) + \frac{\mu}{2}(y-x)^2.$$

2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a first-order differentiable and convex function. If f is σ-strongly convex, then for any $x, y \in \mathbb{R}$, the following inequality holds*

$$f(y) \ge f(x) + f'(x)(y-x) + \frac{\sigma}{2}(y-x)^2.$$

Proof [Proof of Lemma 24] The proof of Lemma 24 is standard and can be found in Beck (2017). ■

Lemma 25 *Under Assumptions* 4-5, *we have*

(a). There exist two constants $c_0 = \sigma e^{-3M}/2, C_0 = \mu e^{3M}/2$, such that

$$c_0\|D - D^*\|^2_{\max} \le \mathcal{B}_\psi\left(e^D\right) - \mathcal{B}_\psi\left(e^{D^*}\right),$$

and

$$\mathcal{B}_\psi\left(e^D\right) - \mathcal{B}_\psi\left(e^{D^*}\right) \le C_0\|D - D^*\|^2_{\max}.$$

(b). For $t_1, t_2 \in [-M, M]$, there exist two constants C_1, C_2, such that

$$|\mathcal{L}_1(t_1) - \mathcal{L}_1(t_2)| \le C_1|t_1 - t_2|,$$

and

$$|\mathcal{L}_2(t_1) - \mathcal{L}_2(t_2)| \le C_2|t_1 - t_2|.$$

Actually, we can take $C_1 = 2e^{2M}\mu, C_2 = e^M\mu$.

Proof [Proof of Lemma 25] (a) Recall from (1) that $\Delta_\psi(x,y) = \psi(x) - \psi(y) - \psi'(x)(x-y)$. Since $E_{p^*}\Delta_\psi(e^{D(Z)}, e^{D^*(Z)}) = \mathcal{B}_\psi(e^D) - \mathcal{B}_\psi(e^{D^*})$ and ψ is μ-smooth and σ-strongly convex, by Lemma 24,

$$\frac{\sigma}{2}E_{p^*}\{e^{D(Z)} - e^{D^*(Z)}\}^2 \le E_{p^*}\Delta_\psi(e^{D(Z)}, e^{D^*(Z)}) \le \frac{\mu}{2}E_{p^*}\{e^{D(Z)} - e^{D^*(Z)}\}^2,$$

and then by Assumption 5,

$$\frac{\sigma e^{-2M}}{2}E_{p^*}\{D(Z) - D^*(Z)\}^2 \le E_{p^*}\Delta_\psi(e^{D(Z)}, e^{D^*(Z)}) \le \frac{\mu e^{2M}}{2}E_{p^*}\{D(Z) - D^*(Z)\}^2. \quad (56)$$

As $E_{p^*}\{D(Z) - D^*(Z)\}^2 = E_{q^*}e^{-D^*(Z)}\{D(Z) - D^*(Z)\}^2$ and $\|D^*\|_\infty \le M$, we have

$$e^{-M}E_{q^*}\{D(Z) - D^*(Z)\}^2 \le E_{p^*}\{D(Z) - D^*(Z)\}^2 \le e^M E_{q^*}\{D(Z) - D^*(Z)\}^2. \quad (57)$$

Let $c_0 = \sigma e^{-3M}/2, C_0 = \mu e^{3M}/2$, then (56) and (57) imply that

$$c_0\|D - D^*\|^2_{\max} \le \mathcal{B}_\psi\left(e^D\right) - \mathcal{B}_\psi\left(e^{D^*}\right) \le C_0\|D - D^*\|^2_{\max}.$$

(b) Obviously, for $t_1, t_2 \in [-M, M]$,

$$\begin{aligned} |\mathcal{L}_1(t_1) - \mathcal{L}_1(t_2)| &= |\psi'(e^{t_1})e^{t_1} - \psi(e^{t_1}) - (\psi'(e^{t_2})e^{t_2} - \psi(e^{t_2}))| \\ &\le e^{t_1}|\psi'(e^{t_1}) - \psi'(e^{t_2})| + |\psi(e^{t_1}) - \psi(e^{t_2}) - \psi'(e^{t_2})(e^{t_1} - e^{t_2})| \\ &\le e^M\mu|e^{t_1} - e^{t_2}| + \frac{\mu}{2}|e^{t_1} - e^{t_2}|^2 \\ &\le 2e^M\mu|e^{t_1} - e^{t_2}| \ \ (As \ \ |e^{t_1} - e^{t_2}| \le 2e^M) \\ &\le 2e^{2M}\mu|t_1 - t_2|, \end{aligned}$$

and

$$\begin{aligned} |\mathcal{L}_2(t_1) - \mathcal{L}_2(t_2)| &= |\psi'(e^{t_1}) - \psi'(e^{t_2})| \\ &\le \mu|e^{t_1} - e^{t_2}| \\ &\le e^M\mu|t_1 - t_2|. \end{aligned}$$

The proof of the lemma is completed. ■

For a given sequence $z = (z_1, \ldots, z_m) \in \mathcal{Z}^m$ and a function class $\mathcal{F}$ on $\mathcal{Z}$, let $\mathcal{F}|_z = \{(f(z_1), \ldots, f(z_m) : f \in \mathcal{F}\}$ be the subset of $\mathbb{R}^m$. For any $\delta > 0$, $\mathcal{N}(\delta, \mathcal{F}|_z, \infty)$ denotes the covering number of $\mathcal{F}|_z$ under the norm $\|\cdot\|_\infty$ with radius δ. Recall that the definition of Rademacher given a function class $\mathcal{F}$ and $z = (z_1, \ldots, z_m)$ is defined by $\mathbb{E}_\eta R_m(\mathcal{F}; z)$, where $\eta = (\eta_1, \ldots, \eta_m)$ is a sequence of i.i.d. Rademacher random variables and

$$R_m(\mathcal{F}; z) = \sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m \eta_i f(z_i).$$

Lemma 26 (Dudley's Chaining, Lemma 3 of Farrell et al. (2021)) *Let* $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_m)$ *denote the covering number for class* $\mathcal{F}$ *with radius* δ *and metric* $\|\cdot\|_m$, *where* $\|f\|_m = \{(1/m)\sum_{i=1}^m f^2(z_i)\}^{1/2}$ *for any* $f \in \mathcal{F}$, *then*

$$\mathbb{E}_\eta R_m(\{f : f \in \mathcal{F}, \|f\|_m \leq r\}\,; z) \leq \inf_{0<\alpha<r}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^r \sqrt{\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_m)}d\delta\right\}.$$

Furthermore, since $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_m) \leq \mathcal{N}(\delta, \mathcal{F}|_z, \infty)$, *the upper bound also holds with* $\mathcal{N}(\delta, \mathcal{F}|_z, \infty)$.

Lemma 27 (Lemma 5 in Schmidt-Hieber (2020)) *For* $\mathcal{D} \in \mathbb{N}$ *and* $\mathbf{d} = (d_0, \ldots, d_{\mathcal{D}+1}) \in \mathbb{N}^{\mathcal{D}+2}$, *let* $V = \prod_{j=0}^{\mathcal{D}+1}(d_j + 1)$. *Then for any* $\delta > 0$,

$$\log \mathcal{N}(\delta, \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}, \|\cdot\|_\infty) \leq (s+1)\log\left(2\delta^{-1}(\mathcal{D}+1)V^2\right),$$

where $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ *denotes the covering number for class* $\mathcal{F}$ *with radius* δ *and metric* $\|\cdot\|_\infty$.

Lemma 28 (Theorem 5 of Schmidt-Hieber (2020)) *For any function* $f \in \mathcal{H}^\beta([0,1]^d, M)$ *and any integers* $m \geq 1$ *and* $N \geq (\beta+1)^d \vee (M+1)e^d$, *there exists a network* $\tilde{f} \in \mathcal{H}_{\mathcal{D},\mathbf{d},s,M}$ *with depth*

$$\mathcal{D} = 8 + (m+5)\left(1 + \lceil \log_2(d \vee \beta)\rceil\right),$$

widths

$$\mathbf{d} = (d_0, \ldots, d_{\mathcal{D}+1}) \in \mathbb{N}^{\mathcal{D}+2},\ d_0 = d,\ d_i = 6(d + \lceil \beta \rceil)N,\ i = 1, 2, \ldots, \mathcal{D}, d_{\mathcal{D}+1} = 1,$$

and number of parameters

$$s \leq 141(d+\beta+1)^{3+d}N(m+6),$$

such that

$$\|\tilde{f} - f\|_\infty \leq (2M+1)\left(1 + d^2 + \beta^2\right)6^d N 2^{-m} + K3^\beta N^{-\frac{\beta}{d}}.$$

The following lemma is from Theorem 2.1 of Bartlett et al. (2005) and also can be found in Lemma 5 of Farrell et al. (2021).

Lemma 29 (Theorem 2.1 of Bartlett et al. (2005)) *Let* $\mathcal{F}$ *be a class of functions on* $\mathcal{Z}$ *and* $\{Z_i\}_{i=1}^m$ *be independent random variables distributed according to a probability measure* P *on* $\mathcal{Z}$. *Assume that there are some* $B, C > 0$ *such that for every* $f \in \mathcal{F}, \|f\|_\infty \leq B, \operatorname{Var}[f(Z_i)] \leq C$. *For every* $t > 0$, *with probability at least* $1 - e^{-t}$,

$$\sup_{f\in\mathcal{F}}\left\{E_{Z\sim P}f(Z) - \frac{1}{m}\sum_{i=1}^m f(Z_i)\right\} \leq 3ER_m(\mathcal{F}; \{Z_i\}_{i=1}^m) + \sqrt{\frac{2Ct}{n}} + \frac{4Bt}{3n}.$$

Lemma 30 (Bernstein's inequality) *Let* $Z_1, \ldots, Z_m \sim Z$ *be i.i.d. random variables bounded in the interval* $[-c, c]$ *for* $c > 0$. *Then, for any* $\delta \in [0,1]$, *with probability at least* $1 - \delta$,

$$\frac{1}{m}\sum_{i=1}^m Z_i - EZ \leq \frac{c}{3m}\log(1/\delta) + \sqrt{\frac{2(\operatorname{Var} Z)\log(1/\delta)}{m}}.$$

Lemma 31 (Theorem 6 of Bartlett et al. (2019)) *For a ReLU network* $\mathcal{F}$ *with a total number of parameters* $\mathcal{S}$ *and depth* $\mathcal{D}$, *there exists two universal constants* $c, C > 0$ *such that*

$$c\mathcal{S}\mathcal{D}\log(\mathcal{S}/\mathcal{D}) \leq Pdim(\mathcal{F}) \leq C\mathcal{S}\mathcal{D}\log\mathcal{S}.$$

Appendix F. Discussions on the proof comparison with those of Theorem 2 of Kato and Teshima (2021)

To employ the error decomposition for the proof of Theorem 2 on page 16 of the supplementary material in Kato and Teshima (2021), we set the correction factor C in Kato and Teshima (2021), which is denoted by C_{nn} in our manuscript, as 0. Following the error decomposition result in Kato and Teshima (2021), we have

$$\begin{aligned} & \mathcal{B}_\psi\left(e^{\widehat{D}}\right)-\mathcal{B}_\psi\left(e^{D^*}\right) \\ = \quad & (E_{p^*}-E_{n_p})\{\mathcal{L}_1(\widehat{D})-\mathcal{L}_1(D^*)\}+(E_q-E_{n_q})\{\mathcal{L}_2(\widehat{D})-\mathcal{L}_2(D^*)\} \\ + \quad & E_{n_p}\{\mathcal{L}_1(D_{\text{NN}})-\mathcal{L}_1(D^*)\}+E_{n_q}\{\mathcal{L}_2(D_{\text{NN}})-\mathcal{L}_2(D^*)\}. \end{aligned} \tag{58}$$

Note that the error decomposition result for the proof of Theorem 2 with $C = 0$ in Kato and Teshima (2021) is a special case of (58) with $D^* \in \mathcal{F}_n$. If $D^* \in \mathcal{F}_n$, then $D^* = D_{\text{NN}}$ and the approximation error term $E_{n_p}\{\mathcal{L}_1(D_{\text{NN}})-\mathcal{L}_1(D^*)\}+E_{n_q}\{\mathcal{L}_2(D_{\text{NN}})-\mathcal{L}_2(D^*)\} = 0$. This tells that there is no need to consider the approximation error when $D^* \in \mathcal{F}_n$. However, in the setup of Theorem 6 in this paper, D^* is a general function and $D^* \notin \mathcal{F}_n$ in general. This will lead to $D^* \neq D_{\text{NN}}$ and $E_{n_p}\{\mathcal{L}_1(D_{\text{NN}})-\mathcal{L}_1(D^*)\}+E_{n_q}\{\mathcal{L}_2(D_{\text{NN}})-\mathcal{L}_2(D^*)\} \neq 0$.

Note that the error decomposition inequality (58) is the same as the one for Theorem 6, which is (25) in the paper. Hence, except for the approximation error terms, another main difference between the proof of Theorem 2 in Kato and Teshima (2021) and those for Theorem 6 is how to deal with the stochastic error

$$A = (E_{p^*}-E_{n_p})\{\mathcal{L}_1(\widehat{D})-\mathcal{L}_1(D^*)\}+(E_q-E_{n_q})\{\mathcal{L}_2(\widehat{D})-\mathcal{L}_2(D^*)\}.$$

Following the proof of Theorem 2 in Kato and Teshima (2021), we know that A is bounded by

$$A \le A_{p^*,n_p} + A_{q^*,n_q}, \tag{59}$$

where

$$A_{p^*,n_p} = |(E_{p^*}-E_{n_p})\{\mathcal{L}_1(\widehat{D})-\mathcal{L}_1(D^*)\}|,$$

and

$$A_{q^*,n_q} = |(E_q-E_{n_q})\{\mathcal{L}_2(\widehat{D})-\mathcal{L}_2(D^*)\}|.$$

In what follows, it suffices to bound A_{p^*,n_p} and A_{q^*,n_q}, separately. However, we will explain why we cannot follow the steps in the proof of Theorem 2 on page 17 of the supplementary material in Kato and Teshima (2021), under the setup of Theorem 6.

The steps to bound A_{p^*,n_p} and A_{q^*,n_q} in Kato and Teshima (2021) involve the applications of their Lemmas 9, 10, 11 (Lemma 5.14 in van de Geer (2000)). Among them, their Lemmas 10 and 11 are for asymptotic analysis, while the proof for Theorem 6 should be in a non-asymptotic sense. And their Lemmas 9 and 11 requires the hypothesis function class, which is $\mathcal{F}_{\text{FNN}}$ in this paper, to be fixed and satisfy a condition on the bracketing entropy $H_B\left(\delta, \mathcal{F}_{\text{FNN}}, \|\cdot\|_{L^2(P)}\right)$ under L_2 norm with respect to the underlying probability measure. That is, for any $0 < \gamma < 2$,

$$H_B\left(\delta, \mathcal{F}_{\text{FNN}}, \|\cdot\|_{L^2(P)}\right) = \log N_B\left(\delta, \mathcal{F}_{\text{FNN}}, \|\cdot\|_{L^2(P)}\right) = O\left(\frac{1}{\delta}\right)^\gamma, \quad \text{for all} \quad \delta > 0. \tag{60}$$

Regarding the definition of the bracketing number $N_B\left(\delta, \mathcal{F}_{\mathrm{FNN}}, \|\cdot\|_{L^2(P)}\right)$, we refer to Definition 2.1.6 of van der Vaart and Wellner (1996) or Definition 2.2 of van de Geer (2000). When $\mathcal{F}_{\mathrm{FNN}}$ is a fixed ReLU network satisfying the structure proposed by Schmidt-Hieber (2020) (that is $\mathcal{F}_{\mathrm{FNN}}$ does not change as the training sample size n increases and its weights are bounded by a constant 1 and sparse), Kato and Teshima (2021) showed that the bracketing entropy condition (60) holds by Lemma 5 in Schmidt-Hieber (2020) and they proved their Theorem 2 under such a network structure.

In contrast, Theorem 6 in this work only assumes that $\mathcal{F}_{\mathrm{FNN}}$ is a feedforward and fully connected ReLU network with a pseudo dimension $\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$. So $\mathcal{F}_{\mathrm{FNN}}$ in our Theorem 6 can depend on the training sample size n and no boundedness condition on its weights is imposed, thus Lemma 5 of Schmidt-Hieber (2020) is not applicable. It is known that the pseudo dimension is equal to the VC dimension for feedforward ReLU network (Theorem 14.1 of Anthony and Bartlett (1999)). Although some results, such as Corollary 1 of Adams and Nobel (2010), demonstrate that under certain conditions, the finite VC dimension of $\mathcal{F}_{\mathrm{FNN}}$ implies that the bracketing number $N_B\left(\delta, \mathcal{F}_{\mathrm{FNN}}, \|\cdot\|_{L^2(P)}\right) < \infty$ for all $\delta > 0$, the stringent bracketing entropy condition (60) cannot hold for general $\mathcal{F}_{\mathrm{FNN}}$.

In view of this, we summarize that the proof of Theorem 2 in Kato and Teshima (2021) can not be directly adapted to prove Theorem 6 in this paper. However, we still can bound A_{p^*,n_p} and A_{q^*,n_q} using the symmetrization technique, Talagrand's lemma, the inequality (35) in the paper and the fact that $\|D\|_\infty \leq M$ for any $D \in \mathcal{F}_{\mathrm{FNN}}$, which result in

$$A_{p^*,n_p} \leq C_0 \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n}{n}}, \quad A_{p^*,n_p} \leq C_0 \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n}{n}}, \tag{61}$$

for some positive constant C_0, when $n \geq \mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}})$. Combining (59) and (61), it holds

$$A \leq 2C_0 \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n}{n}}.$$

Following the proof of Theorem 17 in the paper, we have that

$$E_{S_p,S_q} \|\widehat{D} - D^*\|_{\max}^2 \leq C \left(\sqrt{\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n}{n}} + \inf_{D \in \mathcal{F}_{\mathrm{FNN}}} \|D - D^*\|_{\max}^2 \right). \tag{62}$$

However, in our proof of Theorem 6, we handle two interactive empirical processes w.r.t the two involved samples carefully through a novel localization technique, and we do not bound the approximation error and stochastic error separately. We use the localization technique to directly bound $\|\widehat{D} - D^*\|_{\max}$ and Theorem 6 and its corollary (Corollary 7) in the paper show that

$$E_{S_p,S_q} \|\widehat{D} - D^*\|_{\max}^2 \leq C \left(\frac{\mathrm{Pdim}(\mathcal{F}_{\mathrm{FNN}}) \log n}{n} + \inf_{D \in \mathcal{F}_{\mathrm{FNN}}} \|D - D^*\|_{\max}^2 \right),$$

which is faster than the rate in (62).

Appendix G. Implementation details

In all experiments, we apply the Adam optimizer (Kingma and Ba, 2014) in Pytorch with a learning rate *lr* and a weight decay parameter *wd*. All experiments are conducted on a laptop with an *Intel(R) Core(TM) i7-8750H @ 2.20GHz* CPU having 6 cores.

G.1 Implementation details for subsection 5.1

For Model (TN), we use a FNN with widths $(128, 128, 128)$ and the testing MSE is computed in the following way: For the estimated log density-ratio $\hat{D}$, we generate 5000 data points $\{Z_{p,i}^t\}_{i=1}^{5000}$ from p^* and calculate

$$\widehat{\mathrm{MSE}}(\hat{D}) = \frac{1}{5000}\sum_{i=1}^{5000}\left\{\hat{D}(Z_{p,i}^t) - \log\frac{q^*(Z_{p,i}^t)}{p^*(Z_{p,i}^t)}\right\}^2 .$$

In this simulation, $lr = 0.001$ and $wd = 0$. The maximum number of epochs is 50 and the batch size is $n/5$.

As for Model (NI), the testing MSE is computed as below: For the estimated log density-ratio $\hat{D}$, we generate $\{Z_{p,i}^t\}_{i=1}^{5000}$ from p^* and $\{Z_{q,i}^t\}_{i=1}^{5000}$ from q^*, respectively, and calculate

$$\widehat{\mathrm{MSE}}(\hat{D}) = \frac{1}{5000}\sum_{i=1}^{5000}\left\{(\hat{D}(Z_{p,i}^t) - \hat{D}(Z_{q,i}^t)) - (\tilde{D}^*(Z_{p,i}^t) - \tilde{D}^*(Z_{q,i}^t))\right\}^2, \tag{63}$$

where $\tilde{D}^*(z) = f_0(x) - y$. The rationale to employ such an evaluation metric lies in the non-analytical nature of c in (23), which introduces additional errors that are challenging to control when numerical integration is used to approximate c. Using (63), there is no need to estimate the constant c. Since $\tilde{D}^*(z_1) - \tilde{D}^*(z_2) = D^*(z_1) - D^*(z_2)$ for any $z_1 = (y_1, x_1)$ and $z_2 = (y_2, x_2)$ by (23), $\widehat{\mathrm{MSE}}(\hat{D})$ in (63) is an estimate of

$$\mathrm{MSE}_{p,q}(\hat{D}) = E_{Z_1\sim p^*, Z_2\sim q^*}\|(\hat{D}(Z_1) - \hat{D}(Z_2)) - (D^*(Z_1) - D^*(Z_2))\|^2,$$

and upholds the property $\mathrm{MSE}_{p,q}(\hat{D}) \leq 4\|\widehat{D} - D^*\|_{\max}^2$. As a result, the rate in Theorem 10 also holds for $\mathrm{MSE}_{p,q}(\hat{D})$ with a different prefactor constant. We use a FNN with larger width corresponding to the increasing training sample size n, aligning with the network structure conditions outlined in Theorem 10. Specifically, we use a FNN with widths $(w(n), w(n), w(n))$ specified in Table 8.

Table 8: The $w(n)$ for Model (NI).

n	100	500	1000	5000	10000	100000
w(n)	16	24	32	48	64	128

In this simulation, lr decreases as n increases, with a starting value 0.001, and $wd = 0$. The maximum number of epochs is 150 and the batch size is $n/5$.

As for the simulations to verify Theorem 13, the maximum number of epochs is 150, the batch size is $n/5$ and $wd = 0$. The learning rate lr decreases as n increases, with a starting value 0.0003 when X follows normal distribution, and 0.0005 when X is Student-t(2). We use a FNN with widths $(w(n), w(n), w(n))$ with $w(n)$ specified in Table 9.

Table 9: The $w(n)$ for simulations to verify Theorem 13.

n	100	500	1000	5000	10000	100000
w(n)	8	12	16	24	32	64

G.2 Implementation details for subsection 5.2

We apply the Adam optimizer (Kingma and Ba, 2014) in Pytorch with a learning rate $lr = 0.001$ and a weight decay parameter $wd = 0.0001$. A neural network with 2 hidden layers with widths $(32, 32)$ and the ReLU as its activation function, is used in the experiment. The maximum number of epochs is 400. In this experiment, the training data sample size n is 5000 or 10000. A validation data with sample size 1000 is used. The batch size is 500, and an early-stopping technique is applied with *patience* = 30, where *patience* is the number of epochs until termination if no improvement is made on the validation dataset.

G.3 Implementation details for subsection 5.3

We wish to highlight that Proposition 18 is an asymptotic result, valid under the condition that the training sample size approaches infinity. In such a case, all intermediate density ratios will be estimated accurately, except the last density ratio $R^*_{K-1} = q_{K-1}/p^*$, where $q_{K-1} = (1 - \alpha_{K-1})q^* + \alpha_{K-1}p^*$. To isolate the impact of $(1 - \alpha_{K-1})$ and ensure accurate simulation, we presume the knowledge of q^*/q_{K-1}, which is equivalent to assuming that all intermediate density ratios $R^*_k, k = 0, 1, \ldots, K-2$ are known, where $q_k = (1 - \alpha_k)q^* + \alpha_k p^*, k = 0, 1, \ldots, K-2$ and $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = 1$. Subsequently, we employ the proposed methodology to estimate R^*_{K-1} and denoted the estimate by $\hat{R}_{K-1}$. To evaluate the estimation error, we generate 2000 data points $\{Z_{u,i}\}_{i=1}^{2000}$ from Uniform$([0,1]^d)$ and calculate the mean squared error (MSE) for the ideal TRE using α_{K-1} as:

$$\widehat{\mathrm{MSE}}(\alpha_{K-1}) = \frac{1}{2000} \sum_{i=1}^{2000} \left[\left\{ \prod_{k=0}^{K-2} R^*_k(Z_{u,i}) \right\} \hat{R}_{K-1}(Z_{u,i}) - \frac{q^*(Z_{u,i})}{p^*(Z_{u,i})} \right]^2 .$$

An FNN with widths $(256, 256, 256)$, $lr = 0.01$ and $wd = 0$ are used in this experiment. The maximum number of epochs is 500 and the batch size is $n/5$.

G.4 Implementation details for subsection 5.4

In this simulation study, we apply a learning rate $lr = 0.0001$ and a weight decay parameter $wd = 0.0001$. A neural network with 2 hidden layers with widths $(64, 64)$ and ReLU activation function, is used in the experiment. The maximum number of epochs is 20000. In this experiment, the training data size n is 5000 (10000). A validation data is used. The batch size is 500 (1000), and an early-stopping technique is applied with *patience* = 100 for Beta setting and *patience* = 1000 for Normal setting, where *patience* is the number of epochs until termination if no improvement is made on the validation dataset. We use the LR-Bregman divergence in this example. For the sequence $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{K-1} < \alpha_K = 1$, we use the linearly spaced α_k's, that is $\alpha_k = k/K$, $k = 0, 1, 2, \ldots, K$.

G.5 Analysis on the MNIST dataset

In this study, the batch size is 512, $lr = 0.001$ and $wd = 0.0001$. The maximum number of epochs is 1000. Table 10 contains the specification of the network architectures we adopt for the proposed chain. In this part, the comparison criteria is the average negative log-likelihood (ANLL) in bits per dimension, which is calculated as below. Suppose we have i.i.d training data $\{Z_{q,j}\}_{j=1}^{n_q}$ and testing data $\{Z_{t,j}\}_{j=1}^{n_t}$ which have an unknown density q^* on $\mathcal{Z} \subseteq \mathbb{R}^d$. For an reference distribution, it has a known density p^* on $\mathcal{Z}$, e.g. $\mathcal{Z} = \mathbb{R}^d$ and the standard Gaussian distribution, and can conveniently generate i.i.d. samples $\{Z_{p,j}\}_{j=1}^{n_p}$. Suppose we have an estimate $\hat{D}$ of log density ratio $\log(q^*/p^*)$ based on $\{Z_{q,j}\}_{j=1}^{n_q}$ and $\{Z_{p,j}\}_{j=1}^{n_p}$. The ANLL on $\{Z_{t,j}\}_{j=1}^{n_t}$ is calculated by

$$\text{ANLL} = -\left[\frac{1}{n_t}\sum_{i=1}^{n_t}\{\hat{D}(Z_{t,i}) + \log p^*(Z_{t,i})\}\right]/d, \tag{64}$$

which essentially is an estimate of $-E_{q^*}\log q^*(Z)/d$.

The reference distribution for our mTRE is taken to be the standard Gaussian distribution. Here, the reference distribution is the same as the noise distribution in the MNIST experiments of (Rhodes et al., 2020).

Table 10: Architecture for mTRE

Layers	Details	Output size
Convolution	3 × 3 Conv	12 × 28 × 28
Transition Layer 1	BN, ReLU, 2 × 2 Average Pool,1 × 1 Conv	12 × 14 × 14
Dense Block 1	BN, 1 × 1 Conv, BN, 3 × 3 Conv	24 × 14 × 14
Transition Layer 1	BN, ReLU, 2 × 2 Average Pool,1 × 1 Conv	12 × 7 × 7
Dense Block 1	BN, 1 × 1 Conv, BN, 3 × 3 Conv	24 × 7 × 7
Pooling	BN, ReLU, 7 × 7 Average Pool, Reshape	24
Fully connected	Linear	1

References

Terrence M. Adams and Andrew B. Nobel. Uniform approximation and bracketing properties of vc classes. *arXiv preprint arXiv:1007.4037*, 2010.

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 641–665, PMLR, 2024.

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

Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2009.

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

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudo-dimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Amir Beck. *First-Order Methods in Optimization.* Society for Industrial and Applied Mathematics, 2017.

Peter J. Bickle, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York, 1998.

Lev M. Bregman. 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(3):200–217, 1967.

Ciwan Ceylan and Michael U. Gutmann. Conditional noise-contrastive estimation of unnormalised models. In *Proceedings of the 35th International Conference on Machine Learning*, pages 725–733, PMLR, 2018.

Trevor F. Cox and Gillian Ferry. Robust logistic discrimination. *Biometrika*, 78(4):841–849, 1991.

Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimates. *The Annals of Statistics*, 24(6):2499–2512, 1996.

Luc Devroye and Gábor Lugosi. *Combinatorial Methods in Density Estimation.* Springer, New York, 2001.

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

Charles Fefferman. Whitney's extension problem for c^m. *Annals of Mathematics*, 164(1): 313–359, 2006.

Yuan Gao, Jian Huang, Yuling Jiao, Jin Liu, Xiliang Lu, and Zhijian Yang. Deep generative learning via euler particle transport. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 336–368, PMLR, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Alex J. Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M. Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, Cambridge, 2009.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546): 1136–1154, 2024.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

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

Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and gans. *Applied and Computational Harmonic Analysis*, 65:249–278, 2023b.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012a.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58:708–720, 2012b.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, and James M. Robins. Nonparametric von mises estimators for entropies, divergences and mutual informations. *Advances in Neural Information Processing Systems*, 2015.

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

Masahiro Kato, Masaaki Imaizumi, and Kentaro Minami. Unified perspective on probability divergence via the density-ratio likelihood: Bridging KL-divergence and integral probability metrics. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 5271–5298, PMLR, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

George V. Moustakides and Kalliopi Basioti. Training neural networks for likelihood/density ratio estimation. *arXiv:1911.00405*, 2019.

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

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, 2016.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Hongxiang Qiu, Edgar Dobriban, and Eric Tchetgen Tchetgen. Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1680–1705, 2023.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In *Advances in Neural Information Processing Systems*, 2020.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1916–1921, 2020.

Bernard W. Silverman. Density ratios, empirical likelihood and cot death. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 27(1):26–33, 1978.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Von Bunau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning.* Cambridge University Press, Cambridge, 2012a.

Masashi Sugiyama, Teruyuki Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: A unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012b.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 2019.

Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17(2):138–155, 2009.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2008.

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

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics.* Springer, New York, 1996.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, Cambridge, 2019.

Larry Wasserman. *All of Nonparametric Statistics.* Springer, New York, 2006.

Makoto Yamada and Masashi Sugiyama. Direct importance estimation with gaussian mixture models. *IEICE Transactions on Information and Systems*, E92.D(10):2159–2162, 2009.

Makoto Yamada, Masashi Sugiyama, Gordon Wichern, and Jaak Simm. Direct importance estimation with a mixture of probabilistic principal component analyzers. *IEICE Transactions on Information and Systems*, E93.D(10):2846–2849, 2010.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 2024.