Journal of Machine Learning Research 27 (2026) 1-47 Submitted 2/25; Revised 12/25; Published 1/26

A Data-Augmented Contrastive Learning Approach to Nonparametric Density Estimation

Chenghao Li CHENGHAOLI@LINK.CUHK.EDU.HK
Department of Statistics and Data Science
The Chinese University of Hong Kong
Hong Kong SAR, China

Yuanyuan Lin YLIN@STA.CUHK.EDU.HK
Department of Statistics and Data Science
The Chinese University of Hong Kong
Hong Kong SAR, China

Editor: Aapo Hyvärinen

Abstract

In this paper, we introduce a data-augmented nonparametric noise contrastive estimation method to density estimation using deep neural networks. By leveraging the idea of contrastive learning, our density estimator exhibits efficiency with a one-step and simulation-free evaluation process, imposes no constraints on the neural network, and is shown to be consistent and asymptotically automatically normalized. A novel data augmentation procedure allows us to mitigate the influence of the choice of reference distribution on our method. Non-asymptotic upper bounds for the expected L_2-risk and the expected total variation distance have been established, which achieve minimax optimal rates. Moreover, our new method exhibits inherent adaptivity to low-dimensional structures of data with a faster convergence rate under a compositional structure assumption. Numerical experiments show the competitiveness of our new method compared with the state-of-the-art nonparametric density estimation methods.

Keywords: Contrastive learning, data augmentation, deep neural network, non-asymptotic error bound, nonparametric density estimation

1. Introduction

In a variety of modern statistical and machine learning problems, learning the underlying data distribution is one of the fundamental issues. Many problems can be formulated as learning some characteristics of the data distribution. Thus, a good distribution/density estimator enables a wide range of tasks to be performed, including classification (Schmah et al., 2008), denoising (Ballé et al., 2016), missing value imputation (Dinh et al., 2015), data synthesis (Theis and Bethge, 2015), and many others.

There are mainly two kinds of distribution learning methods: density estimation and generative modeling. Neural networks-based generative models include generative adversarial networks (Goodfellow et al., 2014), variational autoencoders (Kingma and Welling, 2022, 2019), normalizing flows (Rezende and Mohamed, 2015; Kobyzev et al., 2020), and diffusion models (Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021). There have been growing evidence and examples indicating the success of generative models in various applications, such as the utilization of diffusion models in conditional sample generation, black-box optimization, control, and reinforcement learning (Chen et al., 2024). As a class of implicit distribution learning approaches, generative models generally do not provide an explicit form of the estimated distribution. However, an accurate density estimator with explicit form could be essential in many statistical inference and prediction tasks, such as anomaly/outlier detection and universal likelihood ratio test (Wasserman et al., 2020).

In this paper, we consider the task of nonparametric density estimation based on observed data. It is well-known that estimating the density function of high-dimensional data is a challenging problem. Traditional density estimation methods such as histogram (Scott, 1979; Lugosi and Nobel, 1996) and kernel density estimation (Rosenblatt, 1956; Parzen, 1962) enjoy nice theoretical properties. But they are typically inapplicable in high dimension due to the deteriorating performance as the dimensionality increases. To overcome the difficulties associated with high dimensionality, extensive investigations have been conducted to harness the capacity of deep neural networks in approximating high-dimensional functions. Under autoregressive models, Germain et al. (2015), Uria et al. (2016) and Papamakarios et al. (2017) decompose the target density into a product of conditional densities based on the probability chain rule, where each conditional density can be parameterized by neural networks. Normalizing flows-based methods in Ballé et al. (2016), Dinh et al. (2017), and Grover et al. (2018) recover data with an invertible transformation of some latent variable with known density, where the transformation has a compositional structure and can be learnt by neural networks.

Recently, significant advancements have been made in nonparametric density estimation. Roundtrip (Liu et al., 2021) is a density estimation method based on two generative adversarial networks to learn the transformation from the target distribution to some known distribution as well as its inverse. Roundtrip is capable of both density estimation and sample generation in high-dimensional problems, a unique feature compared to other existing density estimation methods. Bos and Schmidt-Hieber (2024) introduces a novel two-stage density estimation method, which casts the unsupervised density estimation problem into a supervised regression problem. To be exact, they first compute a kernel density estimator based on a subsample for pseudo labeling. Then a density estimator is obtained by nonparametric regression technique.

Noise contrastive estimation (Gutmann and Hyvärinen, 2012), also known as contrastive learning, is an intriguing approach for learning unnormalized probabilistic models. The key idea is to transform the unsupervised learning problem of density estimation into a supervised classification between data and artificially generated noise (reference data). The connection between density estimation and classification has been discussed earlier in Hastie et al. (2009, Section 14.2.4), which offers valuable insights into extending this strategy to other unsupervised learning problems. A comprehensive review of the applications of contrastive learning can be found in Jaiswal et al. (2021).

In recent years, several stimulating contrastive learning-based density estimation methods have been proposed. Flow contrastive estimation (Gao et al., 2020) enhances noise contrastive estimation by employing an adaptive reference distribution defined by a trainable normalizing flow. Telescoping density-ratio estimation (Rhodes et al., 2020) estimates the density ratio between the data and noise by applying noise contrastive estimation to a sequence of intermediate distributions. Then, a density estimator can be obtained through multiplying the estimated density ratio by the known reference density.

In this paper, we propose a data-augmented contrastive learning approach to nonparametric density estimation. Compared with the state-of-the-art such as Roundtrip (Liu et al., 2021), kernel-based two-stage estimator (Bos and Schmidt-Hieber, 2024), flow contrastive estimation (Gao et al., 2020), and telescoping density-ratio estimation (Rhodes et al., 2020), our method has advantages in the following aspects.

First, our method is methodologically and computationally appealing. It inherits the advantages of noise contrastive estimation, such as automatic normalization and consistency. Instead of estimating the target density directly, we advocate a data augmentation procedure and estimate a mixture density, thus it is less sensitive to the choice of the reference distribution in comparison to the vanilla noise contrastive estimation (Gutmann and Hyvärinen, 2012). We minimize a well-defined objective function without any constraint, and the resulting density estimator is asymptotically automatically normalized. In this way, we avoid the estimation of the normalizing constant, which involves computationally expensive or even intractable integration of a high-dimensional function. Computationally, our method is easy to implement and numerically stable.

Second, our method is theoretical appealing. We derive non-asymptotic upper bounds for the expected L_2-risk and total variation distance under the assumption that the target density function belongs to some Hölder class. Our results allow for flexibility in network architectures and no constraints (e.g., sparsity and weight-norm constraints) on the network class. Under mild conditions on the reference distribution, it can be easily checked that the mixture density is always bounded away from zero and infinity, automatically satisfying the so-called strong density assumption (see, e.g., Definition 2.2 of Audibert and Tsybakov (2007)) commonly used in statistics. Moreover, when the target density function has a compositional structure, we show that our method can circumvent the curse of dimensionality and automatically adapt to the latent structure.

To summarize, our main contributions are as follows:

(i) We propose a data-augmented nonparametric noise contrastive estimation method to density estimation, which is applicable to a wide range of problems. By employing a data augmentation procedure, we can mitigate the influence of the choice of reference distribution on density estimation, and the mixture density to be estimated automatically satisfies the strong density assumption under mild conditions on the reference density.

(ii) We establish non-asymptotic error bounds under the expected L_2-risk and the expected total variation distance. The error bounds are explicitly determined by the architecture parameters of the neural network. With specific choices of network architectures, our error bounds can attain the nonparametric minimax optimal rate (Stone, 1982). In addition, the resulting density estimator is shown to be asymptotically automatically normalized, thus the problematic integration one constraint is avoided.

(iii) Under a compositional structure assumption of the target density, we show that our estimator can alleviate the curse of dimensionality with a faster rate of convergence depending on the intrinsic dimension, rather than the nominal dimension.

The remainder of the paper is organized as follows. In Section 2, we introduce our proposed data-augmented nonparametric noise constrastive estimation and the ReLU-activated feedforward neural network. Section 3 presents the error analysis and non-asymptotic error bounds for our estimator. In Section 4, we show that our method can circumvent the curse of dimensionality when the target density has a latent compositional structure. In Section 5, we compare our method with several state-of-the-art nonparametric density estimation methods using both simulated and real data. A few concluding remarks are given in Section 6. All proofs are deferred to Appendix A.

2. Data-augmented nonparametric noise contrastive estimation

In this section, we first describe the problem setup and the proposed data-augmented nonparametric noise constrastive estimation. After that, we briefly review the structure of deep neural networks, and introduce the network class used for density estimation.

2.1 Methodology

Let $X_1, \dots, X_n$ be n independent and identically distributed random vectors from an unknown distribution with probability density function f_0 on $\Omega \subseteq \mathbb{R}^d$. The task is to estimate f_0 nonparametrically based on the observed data $\{X_i\}_{i=1}^n$.

Our proposed data-augmented nonparametric noise contrastive density estimation is to first convert the unsupervised density estimation problem into a supervised task of classification, building on the ideas from Hastie et al. (2009) and Gutmann and Hyvärinen (2012). The key difference is that we estimate a mixture density $f_{\mathcal{M}}$ by classification, instead of estimating the target density f_0 directly. To achieve this, we need to introduce external data points sampled from some known density $f_{\mathcal{R}}$. Specifically, our proposed procedure consists of three main steps:

Step 1. Data augmentation.

Let $f_{\mathcal{R}}$ be a known density function on Ω, termed as the reference density. We generate a random sample $\{\tilde{Z}_1, \dots, \tilde{Z}_{\rho n}\}$ from the reference density $f_{\mathcal{R}}$, independent of the observed data $\{X_i\}_{i=1}^n$, where ρ is a positive constant such that $\rho n \in \mathbb{Z}_+$ and $\mathbb{Z}_+$ denotes the set of positive integers. Given the observed data $\{X_i\}_{i=1}^n \sim f_0$ and the reference data $\{\tilde{Z}_i\}_{i=1}^{\rho n} \sim f_{\mathcal{R}}$, we term $\{Y_i\}_{i=1}^{n+\rho n} :- \{X_i\}_{i=1}^n \cup \{\tilde{Z}_i\}_{i=1}^{\rho n}$ as the augmented data.

We assign a binary label $\tilde{C}_i$ to each Y_i according to whether it is the observed data. That is, we assign $\tilde{C}_i = 1$ if $Y_i \in \{X_j\}_{j=1}^n$ and $\tilde{C}_i = 0$ if $Y_i \in \{\tilde{Z}_j\}_{j=1}^{\rho n}$. Then, the conditional densities of Y are

$$f(y \mid \tilde{C} = 1) = f_0(y), \quad f(y \mid \tilde{C} = 0) = f_{\mathcal{R}}(y).$$

Note that the probability mass of label $\tilde{C}$ is $\Pr(\tilde{C} = 1) = 1/(1+\rho)$ and $\Pr(\tilde{C} = 0) = \rho/(1+\rho)$. Then, the marginal density of the augmented data Y is

$$\begin{aligned} f(y) &= f(y \mid \tilde{C} = 1)\Pr(\tilde{C} = 1) + f(y \mid \tilde{C} = 0)\Pr(\tilde{C} = 0) \\ &= \frac{1}{1+\rho} f_0(y) + \frac{\rho}{1+\rho} f_{\mathcal{R}}(y) :- f_{\mathcal{M}}(y). \end{aligned} \tag{1}$$

Here $f_{\mathcal{M}}$ is referred to as the mixture density. It can be easily derived from (1) that $f_0 = (1+\rho)f_{\mathcal{M}} - \rho f_{\mathcal{R}}$. We can obtain an estimator for f_0 once the estimator for the mixture density $f_{\mathcal{M}}$ is available.

Step 2. Estimating the mixture density $f_{\mathcal{M}}$ with a noise contrastive estimation.

We generate an additional set of reference data $\{Z_i\}_{i=1}^{\nu(n+\rho n)}$ from $f_{\mathcal{R}}$ independent of $\{\tilde{Z}_i\}_{i=1}^{\rho n}$, termed as the contrastive noise. Here ν is the ratio of sample sizes between the contrastive noise and the augmented data. We let $\nu \in \mathbb{Z}_+$ for convenience. Henceforth, we write the union of the augmented data and the contrastive noise as $S = \{Y_1, \ldots, Y_{n+\rho n}, Z_1, \ldots, Z_{\nu(n+\rho n)}\}$, the sample used for density estimation.

Similar to Step 1, we rewrite S as $\{U_t\}_{t=1}^{(1+\nu)(n+\rho n)}$ and assign a binary label C_t to each U_t according to whether it is the augmented data. That is, we assign $C_t = 1$ if $U_t \in \{Y_i\}_{i=1}^{n+\rho n}$, and $C_t = 0$ if $U_t \in \{Z_i\}_{i=1}^{\nu(n+\rho n)}$. Then, the conditional densities of U are:

$$f(u \mid C = 1) = f_{\mathcal{M}}(u), \quad f(u \mid C = 0) = f_{\mathcal{R}}(u).$$

The prior marginal probability mass of label C in the combined data S is $\Pr(C = 1) = 1/(1+\nu)$ and $\Pr(C = 0) = \nu/(1+\nu)$. Then, by the Bayes' formula, the posterior probabilities for the labels are

$$\Pr(C = 1 \mid u) = \frac{f_{\mathcal{M}}(u)}{f_{\mathcal{M}}(u) + \nu f_{\mathcal{R}}(u)}, \quad \Pr(C = 0 \mid u) = \frac{\nu f_{\mathcal{R}}(u)}{f_{\mathcal{M}}(u) + \nu f_{\mathcal{R}}(u)}.$$

Then, the log-likelihood of f for the sample S is

$$\begin{aligned}\ell(f; S) &= \sum_{t=1}^{(1+\nu)(n+\rho n)} \big\{C_t \log \Pr(C_t = 1 \mid U_t; f) + (1 - C_t) \log \Pr(C_t = 0 \mid U_t; f)\big\} \\ &= \sum_{i=1}^{n+\rho n} \log \Pr(C_i = 1 \mid Y_i; f) + \sum_{i=1}^{\nu(n+\rho n)} \log \Pr(C_i = 0 \mid Z_i; f),\end{aligned}$$

where $\Pr(C = 1 \mid u; f) :- f(u)/\{f(u) + \nu f_{\mathcal{R}}(u)\}$ and $\Pr(C = 0 \mid u; f) :- \nu f_{\mathcal{R}}(u)/\{f(u) + \nu f_{\mathcal{R}}(u)\}$. Intuitively, an estimator for $f_{\mathcal{M}}$ can be obtained by maximizing the log-likelihood function $\ell(f; S)$ over $f \in \mathcal{F}_n$, a prespecified function class. Equivalently, we minimize

$$\begin{aligned}\hat{R}_S(f) &:- -\frac{1}{(1+\rho)n}\Big\{\sum_{i=1}^{(1+\rho)n} \log \Pr(C_i = 1 \mid Y_i; f) + \sum_{i=1}^{\nu(1+\rho)n} \log \Pr(C_i = 0 \mid Z_i; f)\Big\} \\ &= -\frac{1}{(1+\rho)n}\Big\{\sum_{i=1}^{(1+\rho)n} \log \frac{f(Y_i)}{f(Y_i) + \nu f_{\mathcal{R}}(Y_i)} + \sum_{i=1}^{\nu(1+\rho)n} \log \frac{\nu f_{\mathcal{R}}(Z_i)}{f(Z_i) + \nu f_{\mathcal{R}}(Z_i)}\Big\},\end{aligned} \tag{2}$$

which equals to $-\ell(f; S)/\{(1+\rho)n\}$. We refer to $\hat{R}_S(f)$ as the empirical noise contrastive risk given sample S, which is also written as $\hat{R}(f)$ to suppress its dependence on sample S when there is no confusion.

As a result, we define the estimator for the mixture density as the empirical risk minimizer:

$$\hat{f}_{\mathcal{M}} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_n} \hat{R}(f). \tag{3}$$

Step 3. Deriving the estimator for the target density f_0.

In view of (1) and the estimator $\hat{f}_{\mathcal{M}}$ defined in (3), our proposed estimator for f_0 is given by

$$\tilde{f}_n = (1+\rho)\hat{f}_{\mathcal{M}} - \rho f_{\mathcal{R}}, \tag{4}$$

which is referred to as *data-augmented nonparametric noise contrastive estimation.*

Remark 1 *For any $f \in \mathcal{F}_n$, the population-level noise contrastive risk corresponding to (2) is*

$$\begin{aligned} R(f) &= -\big\{\mathbb{E}_{Y\sim f_{\mathcal{M}}} \log Pr(C=1 \mid Y; f) + \nu \mathbb{E}_{Z\sim f_{\mathcal{R}}} \log Pr(C=0 \mid Z; f)\big\} \\ &= -\Big\{\mathbb{E}_{Y\sim f_{\mathcal{M}}} \log \frac{f(Y)}{f(Y)+\nu f_{\mathcal{R}}(Y)} + \nu \mathbb{E}_{Z\sim f_{\mathcal{R}}} \log \frac{\nu f_{\mathcal{R}}(Z)}{f(Z)+\nu f_{\mathcal{R}}(Z)}\Big\}. \end{aligned}$$

The weak law of large numbers shows that $\hat{R}(f)$ converges in probability to $R(f)$ as $n \to \infty$. Moreover, following Theorem 1 in Gutmann and Hyvärinen (2012), it can be shown that $f_{\mathcal{M}}$ is the unique minimizer of $R(f)$ given any positive reference density $f_{\mathcal{R}}$. This provides theoretical support for our approach to density estimation.

Remark 2 *A critical issue in the vanilla contrastive learning (Gutmann and Hyvärinen, 2012) is the choice of the reference distribution $f_{\mathcal{R}}$. Intuitively, a good candidate for $f_{\mathcal{R}}$ is a distribution close to the target f_0. If $f_{\mathcal{R}}$ differs a lot from f_0, then the classification task might be too easy and would not require the network to learn much about the structure of data. Nonetheless, it could be hard to choose a proper reference distribution in practice. Our proposed data augmentation method in Step 1 offers an effective way to address the problem by providing a density estimator that is less sensitive to the choice of reference distribution. In comparison with f_0, the reference density $f_{\mathcal{R}}$ is closer to the mixture density $f_{\mathcal{M}}$. In such a way, our method can estimate $f_{\mathcal{M}}$ and thus f_0 satisfactorily in accordance with the insights in Gutmann and Hyvärinen (2012).*

2.2 Deep neural networks

We now introduce the class of neural networks used for density function approximation. In recent years, advancements on the approximation theory of the feedforward neural networks with Rectified Linear Unit (ReLU) activation function are reported by Yarotsky (2017), Chen et al. (2019), Shen et al. (2020), Farrell et al. (2021), and Lu et al. (2021). In this paper, we consider multi-layer perceptron, an important and widely-used subclass of feedforward neural networks. The following expression characterizes the architecture of multi-layer perceptron:

$$f_\theta(x) = \mathcal{L}_{\mathcal{D}} \circ \boldsymbol{\sigma} \circ \mathcal{L}_{\mathcal{D}-1} \circ \boldsymbol{\sigma} \circ \cdots \circ \mathcal{L}_1 \circ \boldsymbol{\sigma} \circ \mathcal{L}_0(x),\ x \in \mathbb{R}^{p_0},$$

where $\boldsymbol{\sigma}$ is the component-wise version of ReLU activation function $\sigma(x) = \max\{0, x\}$, and $\mathcal{L}_i(x) = W_i x + b_i, i = 0, \ldots, \mathcal{D}$ are linear transformations with weight matrices $W_i \in \mathbb{R}^{p_{i+1}\times p_i}$ and bias vectors $b_i \in \mathbb{R}^{p_{i+1}}$. Here, $\mathcal{D}$ is the depth of the neural network, that is, the number of hidden layers. The $(\mathcal{D}+2)$-tuple $(p_0, \ldots, p_{\mathcal{D}+1})$ denotes the number of neurons in each layer, and specifically, $p_0 = d$, $p_{\mathcal{D}+1} = 1$. The width of the network is defined as $\mathcal{W} = \max\{p_1, \ldots, p_{\mathcal{D}}\}$. The total number of parameters $\mathcal{S} = \sum_{i=0}^{\mathcal{D}} p_{i+1}(p_i + 1)$ is referred to as the size of the network. We denote the collection of networks f_θ described above as $\mathcal{NN}(\mathcal{D}, \mathcal{W}, \mathcal{S})$. Note that the architecture parameters of the network class satisfies $\max\{\mathcal{W}, \mathcal{D}\} \leq \mathcal{S} \leq \mathcal{W}(d+1) + (\mathcal{W}^2 + \mathcal{W})(\mathcal{D}-1) + \mathcal{W} + 1 = O(\mathcal{W}^2\mathcal{D})$.

We take $\mathcal{F}_n$ to be the following class of neural networks:

$$\mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma) :- \{f_\theta \in \mathcal{NN}(\mathcal{D}, \mathcal{W}, \mathcal{S}) : \gamma \leq f_\theta \leq \Gamma\},$$

for some positive constants γ, Γ. Note that the architecture parameters $\mathcal{D}, \mathcal{W}$, and $\mathcal{S}$ can depend on n, and we suppress the subscript n for simplicity. Furthermore, the output range constraint $\gamma \leq f_\theta \leq \Gamma$ can be satisfied by adding an additional layer $\phi(x) = \sigma(x-\gamma) - \sigma(x-\Gamma) + \gamma$ at the end of the network. Then, the network depth will increase by 1 accordingly. Such adjustment will only affect the constant prefactors in our non-asymptotic error bounds and will not affect the convergence rate, and thus the details are omitted for simplicity.

3. Theoretical analysis

In this section, we provide a theoretical analysis of the proposed data-augmented nonparametric noise contrastive estimator. We will study the convergence properties of the mixture density estimator $\hat{f}_{\mathcal{M}}$, which can imply the convergence properties of $\tilde{f}_n$, the estimator for the target density f_0.

3.1 Error decomposition

To evaluate the quality of an estimator $\hat{f}$ for a target function f_*, we use the expected L_2-risk $\mathbb{E}\|\hat{f} - f_*\|^2_{L_2(f_*)}$, where the expectation is taken with respect to training sample and

$$\|f - f_*\|^2_{L_2(f_*)} :- \mathbb{E}_{X\sim f_*}|f(X) - f_*(X)|^2 = \int_\Omega |f(x) - f_*(x)|^2 f_*(x)dx.$$

The main procedure for the error analysis is to decompose the expected L_2-risk of $\hat{f}_{\mathcal{M}}$ into stochastic error and approximation error, and bound them separately. Thereafter, an upper bound of $\mathbb{E}_S\|\tilde{f}_n - f_0\|^2_{L_2(f_0)}$ can be established based on that of $\mathbb{E}_S\|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}$.

We first present some conditions imposed on the target density f_0 and reference density $f_{\mathcal{R}}$. We take the domain Ω of f_0 and $f_{\mathcal{R}}$ to be the d-dimensional hypercube $[0,1]^d$ for simplicity. Extension to accommodate density with unbounded domain will be discussed in Section 3.5.

Assumption 3 *The reference density $f_{\mathcal{R}}$ is positive and continuous on $[0,1]^d$, and there exists a positive constant L_0 such that $f_0 \leq L_0$.*

Assumption 3 necessitates mild conditions on both the reference density and the target density. According to Remark 1, a reference distribution $f_{\mathcal{R}}$ that is positive on $[0,1]^d$ suffices to guarantee the uniqueness of the noise contrastive risk minimizer. For the target density f_0, we only require it to be upper bounded, thanks to the proposed data augmentation technique. Specifically, the estimation problem of f_0 is converted into the estimation of a mixture density, making the strong density assumption $l_* \leq f_* \leq L_*$ automatically satisfied under mild conditions on the reference density. The strong density assumption is a common assumption in the literature of density estimation; see Section 3.4 for more details.

The following lemma bounds the expected L_2-risk of $\tilde{f}_n$ for estimating f_0 by the expected L_2-risk of $\hat{f}_{\mathcal{M}}$ for estimating $f_{\mathcal{M}}$, with the proof provided in Appendix A.1. We will then focus on studying the convergence properties of $\hat{f}_{\mathcal{M}}$.

Lemma 4 *For any $\Omega \subseteq \mathbb{R}^d$, the data-augmented nonparametric noise contrastive estimator $\tilde{f}_n$ defined in (4) satisfies*

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \leq (1+\rho)^3 \mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})},$$

where S is the training sample.

With $f_{\mathcal{M}}$ being the unique minimizer of the population-level noise contrastive risk, the excess risk is $R(f) - R(f_{\mathcal{M}})$. We next present an inequality concerning the relation between the excess risk and the L_2 distance $\|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}$.

Lemma 5 (Calibration) *Under Assumption 3, there exist positive constants c_1 and c_2 such that, for all $f \in \mathcal{F}_n$,*

$$c_1 \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq R(f) - R(f_{\mathcal{M}}) \leq c_2 \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}. \tag{5}$$

Moreover, c_1 and c_2 only depend on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$.

The proof of Lemma 5 is given in Appendix A.2. The relationship in (5) is known as the calibration condition or the curvature condition around $f_{\mathcal{M}}$, satisfied by many loss functions in various statistical problems, including least squares regression, logistic regression, multinomial logistic regression, and Poisson regression, among others (Farrell et al., 2021). The calibration condition can lead to the following upper bound for the expected L_2-risk, with the proof provided in Appendix A.3.

Lemma 6 (Error decomposition) *Under Assumption 3, we have*

$$\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq \frac{1}{c_1} \mathbb{E}_S [R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] + \frac{c_2}{c_1} \inf_{f \in \mathcal{F}_n} \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}, \tag{6}$$

where c_1, c_2 are the same constants as in Lemma 5.

The first term of the right-hand side of (6) is the *stochastic error*, and the second term is the *approximation error*. The stochastic error quantifies the expected disparity between the risk and empirical risk evaluated at the estimator $\hat{f}_{\mathcal{M}}$, and can be upper bounded with respect to the complexity of the function space $\mathcal{F}_n$ using empirical process theory (Anthony and Bartlett, 1999; Bartlett et al., 2019; van der Vaart and Wellner, 2023). The approximation error measures how well the function $f_{\mathcal{M}}$ can be approximated using $\mathcal{F}_n$ with respect to L_2 distance, and can be properly bounded using the state-of-the-art approximation theory of deep neural networks.

3.2 Stochastic error and approximation error

The stochastic error of the density estimator $\hat{f}_{\mathcal{M}}$ will be controlled by the complexity measures of the neural network class $\mathcal{F}_n$. The definitions of uniform covering number and pseudo-dimension are given in Appendix B.

Lemma 7 (Stochastic error) *Let $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ be the class of feedforward neural networks activated by continuous piecewise linear activation function with finitely many inflection points. Let $\hat{f}_{\mathcal{M}} \in \arg\min_{f \in \mathcal{F}_n} \hat{R}(f)$ be the empirical risk minimizer over $\mathcal{F}_n$. Suppose that Assumption 3 holds. Then for $n \geq Pdim(\mathcal{F}_n)/\{(1+\nu)(1+\rho)\}$,*

$$\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] \leq c_3 \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n},$$

where c_3 is a constant depending only on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$.

The proof of Lemma 7 is given in Appendix A.4. To establish Lemma 7, the stochastic error $\mathbb{E}_S[R(\hat{f}_n) - \hat{R}(\hat{f}_n)]$ is initially bounded by a term determined by $\log(\mathcal{N}_\infty(\tilde{c}/n, \mathcal{F}_n, n + \nu n))$, the metric entropy of $\mathcal{F}_n$ with the radius of covering being $1/n$ multiplied by some constant. And the metric entropy of $\mathcal{F}_n$ can be further bounded by the pseudo dimension using Theorem 12.2 in Anthony and Bartlett (1999). Based on the VC-dimension bounds (Theorem 7) in Bartlett et al. (2019), the pseudo dimension of the class of neural networks activated by piecewise linear activation function with finitely many inflection points can be controlled by its architecture parameters $\mathcal{D}$ and $\mathcal{S}$, that is, $\text{Pdim}(\mathcal{F}_n) = O(\mathcal{S}\mathcal{D} \log \mathcal{S})$. While we use ReLU activated neural networks in this paper, the result in Bartlett et al. (2019) enables us to extend our stochastic error bound in Lemma 7 to piecewise linear function-activated neural networks. The proof of Lemma 7 provides the constant c_3 explicitly, which exhibits merely logarithmic growth in L_0, $1/\gamma$, and Γ.

To analyze the expected L_2-risk, Bos and Schmidt-Hieber (2024) showed an oracle inequality that bounds the risk by the sum of stochastic error, approximation error, and optimization error. Their stochastic error incorporates a similar metric entropy $\log(\mathcal{N}_{\mathcal{F}}(\delta))$, where $\mathcal{N}_{\mathcal{F}}(\delta)$ is the δ-covering number of function class $\mathcal{F}$ with respect to the supremum norm. The metric entropy $\log(\mathcal{N}_{\mathcal{F}}(\delta))$ is further bounded with respect to network architecture parameters, that is, $O(\mathcal{S}_* \mathcal{D} \log(\mathcal{S}_* \mathcal{D} d/\delta))$ where $\mathcal{S}_*$ is the sparsity defined by the number of nonzero network parameters. Their bound is similar to ours, yet network size is replaced by sparsity and there are additional logarithmic factors concerning network depth $\mathcal{D}$ and dimension d. While typically the sparsity $\mathcal{S}_*$ can be significantly smaller than the size $\mathcal{S}$, the constraints they impose on sparsity and weight norm may restrict the expressiveness of neural network. Balancing stochastic errors and approximation errors in respective settings through proper choices of network architectures, our network size $\mathcal{S}$ and the sparsity $\mathcal{S}_*$ in Bos and Schmidt-Hieber (2024) are of the same order. As a result, the stochastic errors bounds (as well as the overall error bounds) of our work and Bos and Schmidt-Hieber (2024) are comparable with distinct constant prefactors and logarithmic terms. Further details can be found in Section 4.

To bound the approximation error $\inf_{f \in \mathcal{F}_n} \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}$, we impose some smoothness conditions on the target density f_0 and the reference density $f_{\mathcal{R}}$. In this paper, we assume that both f_0 and $f_{\mathcal{R}}$ belong to some Hölder class (Definition 28, Appendix B).

Assumption 8 (Hölder smoothness) *The densities $f_0, f_{\mathcal{R}} \in \mathcal{H}^{\beta}([0,1]^d, B_0)$ for a given $\beta > 0$ and some finite constant $B_0 > 0$.*

The assumption on $f_{\mathcal{R}}$ in Assumption 8 is not restrictive, as many popular distributions satisfy this assumption, such as the uniform distribution on $[0,1]^d$. Notably, Assumption 8 implies that the mixture density $f_{\mathcal{M}} \in \mathcal{H}^{\beta}([0,1]^d, B_0)$.

Applying Theorem 3.3 in Jiao et al. (2023), we can derive an approximation error bound with respect to the L_2 distance; see Lemma 19 in Appendix A.5.

3.3 Non-asymptotic error bound

The following corollary establishes the consistency of our data-augmented nonparametric noise contrastive estimation.

Corollary 9 (Consistency) *Let $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ be the class of feedforward neural networks activated by continuous piecewise linear activation function with finitely many inflection points. Suppose that Assumption 3 holds, f_0 is continuous on $[0,1]^d$, and*

$$\mathcal{S} \to \infty \quad \textit{and} \quad \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} \to 0 \quad \textit{as } n \to \infty.$$

Then the data-augmented nonparametric noise contrastive estimation $\tilde{f}_n$ defined in (4) is consistent in the sense that

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \to 0 \quad \textit{as } n \to \infty.$$

Corollary 9 is a direct consequence of Lemmas 4, 6, and 7 and the approximation results of continuous function by piecewise linear neural networks in Yarotsky (2017) and Yarotsky (2018). In other words, the consistency result in Corollary 9 not only holds for ReLU-activated neural networks, but also for a broader neural network class with piecewise linear activation function.

Next, we present the non-asymptotic error bound of our proposed estimator, with the proof provided in Appendix A.6.

Theorem 10 (Non-asymptotic error bound) *Suppose that Assumptions 3 and 8 hold, and $\gamma \le \rho l/(1+\rho)$, $\Gamma \ge L_0 \vee L$ where l, L are positive constants such that $l \le f_{\mathcal{R}} \le L$. Then, for any $P, Q \in \mathbb{Z}_+$, the function class of ReLU-activated multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ with width $\mathcal{W} = 38(\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + 1} P \lceil \log_2(8P) \rceil$ and depth $\mathcal{D} = 21(\lfloor \beta \rfloor + 1)^2 Q \lceil \log_2(8Q) \rceil$, for $n \ge Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$, the data-augmented nonparametric noise contrastive estimation $\tilde{f}_n$ defined in (4) satisfies*

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \le c_4 \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_5 B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d},$$

where c_4 and c_5 are constants only depending on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$. Furthermore, if we set

$$\mathcal{W} = 114(\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + 1},$$
$$\mathcal{D} = 21(\lfloor \beta \rfloor + 1)^2 \lceil n^{d/\{2(d+2\beta)\}} \log_2(8 n^{d/\{2(d+2\beta)\}}) \rceil,$$
$$\mathcal{S} = O((\lfloor \beta \rfloor + 1)^6 d^{2\lfloor \beta \rfloor + 2} \lceil n^{d/\{2(d+2\beta)\}} \log_2 n \rceil),$$

then for $n \geq Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$, $\tilde{f}_n$ satisfies

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \leq c_6 (B_0 \vee 1)^2 (\lfloor \beta \rfloor + 1)^9 d^{2\lfloor \beta \rfloor + \beta \vee 3} n^{-\frac{2\beta}{d+2\beta}} (\log_2 n)^4, \tag{7}$$

where c_6 is a constant only depending on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$.

In Theorem 10, the expected L_2-risk is upper bounded by the sum of the stochastic error term $c_4 \mathcal{S}\mathcal{D} \log \mathcal{S} \log n/n$ and the approximation error term $c_5 B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1}$ $(PQ)^{-4\beta/d}$. This error bound is distinguished by two merits. On one hand, it is non-asymptotic and explicit in the sense that no obscurely-defined constant is involved. On the other hand, the error bound is a general result which holds for a variety of network architectures. The approximation rate $(PQ)^{-4\beta/d}$ corresponds to the width $\mathcal{W} = 38(\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + 1} P \lceil \log_2(8P) \rceil$ and depth $\mathcal{D} = 21(\lfloor \beta \rfloor + 1)^2 Q \lceil \log_2(8Q) \rceil$ of the neural network, instead of merely the network size $\mathcal{S}$. The approximation error deceases with a larger network size, while the stochastic error increases as the network complexity increases. To achieve the best error rate (7), we balance the stochastic error and the approximation error using a network architecture of fixed width and proper depth, though different network architectures can also be employed.

Our error bound (7) attains the minimax rate $n^{-2\beta/(d+2\beta)}$ in standard nonparametric regression (Stone, 1982) up to a logarithmic factor. It is worth noting that the prefactor in the error bound depends on the dimension of data d merely polynomially. Based on Bartlett et al. (2019), the pseudo dimension of network class $\mathcal{F}_n$ can be bounded explicitly as $\mathrm{Pdim}(\mathcal{F}_n) = O(\mathcal{S}\mathcal{D} \log \mathcal{S}) = O((s+1)^9 d^{2s+3} n^{d/(d+2\beta)} (\log_2 n)^3)$. So $n \geq n_0$ suffices for the prerequisite $n \geq \mathrm{Pdim}(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$, where n_0 is a constant only depending on d and β. As a result, our error bound with specific choice of network structure is non-asymptotic, in the sense that the result holds for all n greater than some constant depending on d and β.

Remark 11 *In estimating the mixture density $f_{\mathcal{M}}$, the constant ρ plays its part in sample size and is preferred to be large; but Lemma 4 implies that increasing ρ could deteriorate the upper bound of $\tilde{f}_n$. To select a proper ρ, we propose to minimize the constant prefactors of the error bound in Theorem 10 with respect to ρ. According to the proofs in Appendices A.2 - A.4, the prefactors of stochastic error and approximation error are respectively*

$$c_4 \propto \frac{(1+\rho)^4}{\rho}, \quad c_5 \propto \frac{(1+\rho)^6}{\rho^3},$$

where the dependencies on logarithmic terms of ρ and other quantities are ignored. Since $\arg\min_{\rho>0} (1+\rho)^4/\rho = 1/3$, $\arg\min_{\rho>0} (1+\rho)^6/\rho^3 = 1$, the minima of the overall error bound with respect to ρ approximately lies within the range of $[1/3, 1]$, which gives us a guideline for selecting ρ in practice. Based on the simulation results in Section 5.2, we would suggest setting a relatively large value for ρ (e.g., $\rho = 1$) in high-dimensional cases or when the reference distribution is likely to differ significantly from the target data distribution.

Remark 12 *Many existing works on noise contrastive estimation assume that the data domain is the entire Euclidean space $\mathbb{R}^d$ (see, for example, Gutmann and Hyvärinen (2012)). In this scenario, the error of noise contrastive estimation can increase exponentially with* *dimension, due to mismatches between the tails of the data and reference distributions (Lee et al., 2023; Chehab et al., 2023). An intuitive explanation is that noise contrastive estimation can be regarded as a variant of importance sampling (Pihlaja et al., 2010), which is known to be sensitive to the tails of distributions (Liu et al., 2015). More discussions can be found in Section 3.5.*

In this section, we assume that the data domain is the d-dimensional hypercube $[0,1]^d$*. In this case, there is no issue concerning tail behavior, allowing us to derive an error bound that depends polynomially on the dimension. In Section 3.5, we will extend our method to handle the unbounded case.*

3.4 Error bound in total variation distance

In this subsection, we establish non-asymptotic error bounds for our proposed density estimator under the total variation distance. Specifically, we consider the expected total variation distance $\mathbb{E}_S\mathrm{TV}(\tilde{f}_n, f_0)$ between our proposed estimator $\tilde{f}_n$ and the target density f_0, where

$$\mathrm{TV}(f, f_*) = \frac{1}{2}\int_\Omega |f(x) - f_*(x)|dx.$$

The following theorem establishes non-asymptotic error bounds in total variation distance.

Theorem 13 (Non-asymptotic error bound in total variation distance) *Suppose that Assumptions 3 and 8 hold, and* $\gamma \le \rho l/(1+\rho)$*,* $\Gamma \ge L_0 \vee L$ *where* l*,* L *are positive constants such that* $l \le f_{\mathcal{R}} \le L$*. Then, for any* $P, Q \in \mathbb{Z}_+$*, the function class of ReLU-activated multi-layer perceptrons* $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ *with width* $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} P\lceil\log_2(8P)\rceil$ *and depth* $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 Q\lceil\log_2(8Q)\rceil$*, for* $n \ge Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$*, the data-augmented nonparametric noise contrastive estimation* $\tilde{f}_n$ *defined in (4) satisfies*

$$\mathbb{E}_S TV(\tilde{f}_n, f_0) \le c_7\Big(\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n}\Big)^{1/2} + c_8 B_0(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+(\beta\vee 1)/2}(PQ)^{-2\beta/d},$$

where c_7 *and* c_8 *are constants only depending on* $\nu, \rho, L_0, \gamma, \Gamma$*, and* $f_{\mathcal{R}}$*. Furthermore, if we set*

$$\mathcal{W} = 114(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1},$$
$$\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2\lceil n^{d/2(d+2\beta)}\log_2(8n^{d/\{2(d+2\beta)\}})\rceil,$$
$$\mathcal{S} = O((\lfloor\beta\rfloor+1)^6 d^{2\lfloor\beta\rfloor+2}\lceil n^{d/\{2(d+2\beta)\}}\log_2 n\rceil),$$

then for $n \ge Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$*,* $\tilde{f}_n$ *satisfies*

$$\mathbb{E}_S TV(\tilde{f}_n, f_0) \le c_9(B_0 \vee 1)(\lfloor\beta\rfloor+1)^{9/2} d^{\lfloor\beta\rfloor+(\beta\vee 3)/2} n^{-\frac{\beta}{d+2\beta}}(\log_2 n)^2, \tag{8}$$

where c_9 *is a constant only depending on* $\nu, \rho, L_0, \gamma, \Gamma$*, and* $f_{\mathcal{R}}$*.*

The proof of Theorem 13 is given in Appendix A.7. The rate of convergence in (8) is minimax optimal (Stone, 1982) up to a logarithmic factor.

Similar to Corollary 9, we can show that the data-augmented nonparametric noise contrastive estimation is consistent with respect to the total variation distance. This further implies that our density estimator $\tilde{f}_n$ is asymptotically automatically normalized as stated in the next corollary.

Corollary 14 (Asymptotic automatic normalization) *Let $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ be the class of feedforward neural networks activated by continuous piecewise linear activation function with finitely many inflection points. Suppose that Assumption 3 holds, f_0 is continuous on $[0,1]^d$, and*

$$\mathcal{S} \to \infty \quad \textit{and} \quad \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} \to 0 \quad \textit{as } n \to \infty.$$

Then the data-augmented nonparametric noise contrastive estimation $\tilde{f}_n$ defined in (4) is consistent with respect to total variation distance:

$$\mathbb{E}_S \, TV(\tilde{f}_n, f_0) \to 0 \quad \textit{as } n \to \infty,$$

and asymptotically normalized in the sense that

$$\mathbb{E}_S \Big[\int_{[0,1]^d} \tilde{f}_n(x) dx \Big] \to 1 \quad \textit{as } n \to \infty.$$

We conclude this subsection by underscoring a theoretical merit of our method. Under Assumption 3, it holds that $l \leq f_{\mathcal{R}} \leq L$ for some positive constants l and L. Then, the mixture density $f_{\mathcal{M}}$ satisfies that

$$f_{\mathcal{M}} = \frac{1}{1+\rho} f_0 + \frac{\rho}{1+\rho} f_{\mathcal{R}} \geq \frac{\rho}{1+\rho} f_{\mathcal{R}} \geq \frac{\rho l}{1+\rho},$$

That is, $f_{\mathcal{M}}$ is lower bounded even though the target density f_0 is not. The proofs of Theorem 13 and Lemmas 5 and 7 rely on the fact that $f_{\mathcal{M}}$ is lower bounded. By employing our proposed data augmentation method, we convert the estimation task of f_0 to the estimation of $f_{\mathcal{M}}$, and the latter satisfies the strong density assumption under mild conditions on the reference density.

3.5 Estimating densities with unbounded supports

In this subsection, we extend our method to handle densities with unbounded supports. Specifically, we let the target density f_0 and the reference density $f_{\mathcal{R}}$ be functions on $\Omega = \mathbb{R}^d$, the entire d-dimensional Euclidean space. To tackle the challenges associated with unbounded support, we consider a truncated version of noise contrastive risk.

We define the truncated domain Ω_t as a d-dimensional hypercube $[-t,t]^d$, where t is a positive constant (the truncation level) and can depend on the sample size n. Then, the truncated noise contrastive risk is defined as

$$R_t(f) = -\Big\{ \mathbb{E}_{Y \sim f_{\mathcal{M}}} \Big[\log \frac{f(Y)}{f(Y) + \nu f_{\mathcal{R}}(Y)} \mathbb{1}_{\Omega_t}(Y) \Big] + \nu \mathbb{E}_{Z \sim f_{\mathcal{R}}} \Big[\log \frac{\nu f_{\mathcal{R}}(Z)}{f(Z) + \nu f_{\mathcal{R}}(Z)} \mathbb{1}_{\Omega_t}(Z) \Big] \Big\},$$

where $\mathbb{1}_A$ denotes the indicator function on a set A, i.e $\mathbb{1}_A$ is equal to 1 on A and 0 outside A. Given the training sample $S = \{Y_1, \ldots, Y_{n+\rho n}, Z_1, \ldots, Z_{\nu(n+\rho n)}\}$, the truncated empirical risk is defined by replacing the expectations in R_t by sample averages:

$$\hat{R}_t(f) = -\frac{1}{(1+\rho)n}\left\{\sum_{i=1}^{(1+\rho)n}\log\frac{f(Y_i)}{f(Y_i)+\nu f_{\mathcal{R}}(Y_i)}\mathbb{1}_{\Omega_t}(Y_i) + \sum_{i=1}^{\nu(1+\rho)n}\log\frac{\nu f_{\mathcal{R}}(Z_i)}{f(Z_i)+\nu f_{\mathcal{R}}(Z_i)}\mathbb{1}_{\Omega_t}(Z_i)\right\}.$$

We define the estimator for the mixture density $f_{\mathcal{M}}$ as the minimizer of the truncated empirical risk, that is, $\check{f}_{\mathcal{M}} \in \arg\min_{f\in\mathcal{F}_n}\hat{R}_t(f)$. Then, similar to (4), the resulting estimator for the target density f_0 is given by

$$\bar{f}_n = (1+\rho)\check{f}_{\mathcal{M}} - \rho f_{\mathcal{R}}. \tag{9}$$

Remark 15 *The truncation level t can depend on the sample size n, making it possible to approximate the mixture density on the truncated domain Ω_t as n increases. Theoretically, Theorem 16 below indicates that optimal convergence rate of our estimator can be achieved by setting $t = \kappa\log n$, with the constant κ depending on the tail behavior of the target distribution. In practice, determining the constant κ is typically infeasible, so we suggest setting it as the maximum value of the data. In this case, the truncated empirical risk equals its untruncated counterpart defined in (2).*

In the error analysis for the refined estimator in (9), we assume that the target distribution is sub-exponential (Assumption 20 in Appendix A.8), which encompasses numerous distributions such as mixtures of Gaussian distributions, log-concave distributions (Bagnoli and Bergstrom, 2005; Lovász and Vempala, 2007), and distributions with bounded supports.

By Lemma 22 in Appendix A.8, the expected L_2-risk of $\bar{f}_n$ can be decomposed into the error on the truncated domain and the error due to truncation, and the former can be further decomposed into stochastic and approximation errors. We bound these errors in Lemmas 23, 25, and 26, respectively. With these results and a proper selection of the truncation level t, we can establish the following non-asymptotic error bound.

Theorem 16 (Non-asymptotic error bound) *Suppose that Assumptions 20 and 21 hold, $t = a^{-1}\log n$, $\gamma \le \rho l_{\mathcal{R}}/(1+\rho)$, and $\Gamma \ge L_0 \vee L_{\mathcal{R}}$. Then, for any $P, Q \in \mathbb{Z}_+$, the function class of ReLU-activated multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1}P\lceil\log_2(8P)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 Q\lceil\log_2(8Q)\rceil + 1$, for $n \ge [Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}] \vee \exp(a/2)$, the data-augmented nonparametric noise contrastive estimator $\bar{f}_n$ defined in (9) satisfies*

$$\begin{aligned}\mathbb{E}_S\|\bar{f}_n - f_0\|^2_{L_2(f_0)} \le\ & c_{16}\gamma^{-3}l_{\mathcal{R}}^{-3}\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n} \\ & + c_{17}\gamma^{-3}l_{\mathcal{R}}^{-2}(2/a)^{2\beta}(\lfloor\beta\rfloor+1)^4 d^{2\lfloor\beta\rfloor+\beta\vee 1}(PQ)^{-4\beta/d}(\log n)^{2\beta} \\ & + c_{18}dn^{-1},\end{aligned}$$

where c_{16}, c_{17}, and c_{18} are constants only depending on ν, ρ, C, L_0, Γ, $L_{\mathcal{R}}$, and B_0. Furthermore, if we set

$$\begin{aligned}\mathcal{W} &= 114(\lfloor\beta\rfloor+1)^2 d^{\lfloor\beta\rfloor+1}, \\ \mathcal{D} &= 21(\lfloor\beta\rfloor+1)^2\lceil n^{d/\{2(d+2\beta)\}}\log_2(8n^{d/\{2(d+2\beta)\}})\rceil + 1, \\ \mathcal{S} &= O((\lfloor\beta\rfloor+1)^6 d^{2\lfloor\beta\rfloor+2}\lceil n^{d/\{2(d+2\beta)\}}\log_2 n\rceil),\end{aligned}$$

then for $n \geq [Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}] \vee \exp(a/2)$, $\bar{f}_n$ *satisfies*

$$\mathbb{E}_S \|\bar{f}_n - f_0\|^2_{L_2(f_0)} \leq c_{19} \gamma^{-3} l_{\mathcal{R}}^{-3} \{(2/a) \vee 1\}^{2\beta} (\lfloor \beta \rfloor + 1)^9 d^{2\lfloor \beta \rfloor + 3(\beta \vee 1)} n^{-\frac{2\beta}{d+2\beta}} (\log_2 n)^{2(\beta \vee 2)}, \quad (10)$$

where c_{19} *is a constant only depending on* ν, ρ, C, L_0, Γ, $L_{\mathcal{R}}$, *and* B_0.

The proof of Theorem 16 can be found in Appendix A.8. For estimating densities with unbounded domain $\mathbb{R}^d$, the refined estimator in (9) attains the minimax rate of nonparametric regression, $n^{-2\beta/(d+2\beta)}$ (Stone, 1982), up to a logarithmic factor. Note that the term $\gamma^{-3} l_{\mathcal{R}}^{-3}$ in (10) will result in prefactors depending on the dimension d exponentially. More detailed discussions are given in Appendix A.8, subsequent to Assumption 20.

Under parametric distributional assumptions, Lee et al. (2023, Theorem 4) demonstrated that the mean square error of the noise contrastive estimator increases at least exponentially fast in the dimension. Chehab et al. (2023, Theorem 2) established a similar lower bound for noise contrastive estimation in estimating the normalizing constant (partition function) of an unnormalized density model. Our error bound in (10) involves prefactors that exhibit exponential growth in d, similar to those in Chehab et al. (2023), Lee et al. (2023) and many other works in the literature. The exponential growth of the prefactors may be attributed to the mismatches between the shapes, regions of high probability mass, or tail behaviors, of the data and reference distributions, as there could be many possible shapes for a distribution in high dimensions.

4. Circumventing curse of dimensionality

In many statistical and machine learning problems, the dimension d of data can be large, which leads to an extremely slow convergence rate even when the sample size is large. This phenomenon is commonly referred to as the curse of dimensionality. Promising ways to mitigate the curse of dimensionality are to impose structural assumptions on the target function (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Kohler et al., 2022), or low-dimensional support assumption on the data distribution (Schmidt-Hieber, 2019; Nakada and Imaizumi, 2020; Shen et al., 2020; Chen et al., 2022; Jiao et al., 2023).

Under these assumptions, it has been shown that the rate of convergence can be improved as it depends on the intrinsic dimension $d_* \ll d$ of the target function (e.g., under hierarchical and compositional structure assumptions), or the intrinsic dimension of the support of the data (e.g., under low-dimensional manifolds and low Minkowski dimension sets assumptions), rather than the nominal or ambient dimensionality d. In statistics, there are plenty of density models that possess a compositional structure, including graphical models, Bayesian networks, copulas, and mixture models, among others (Bos and Schmidt-Hieber, 2024). In this subsection, we will impose a compositional structure assumption on the target density function under which the rate of convergence can be improved, thus we can alleviate the curse of dimensionality.

For a given positive integer q and vectors $\mathbf{l} = (l_0, \ldots, l_q)$, $\mathbf{d} = (d_1, \ldots, d_q)$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$, let $\mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$ be the collection of functions f admitting the compositional structure

$$f = g^{(q)} \circ g^{(q-1)} \circ \cdots \circ g^{(1)},$$

where $g^{(i)} : \mathbb{R}^{l_{i-1}} \to \mathbb{R}^{l_i}$ $(i = 1, \ldots, q)$ is defined by $g^{(i)}(x) = (g_1^{(i)}(W_1^{(i)}x), \cdots, g_{l_i}^{(i)}(W_{l_i}^{(i)}x))^\top$ with $W_j^{(i)} \in \mathbb{R}^{d_i \times l_{i-1}}$ being a matrix and $g_j^{(i)} : \mathbb{R}^{d_i} \to \mathbb{R}$ being a function. Moreover, $g_j^{(i)} \in \mathcal{H}^{\beta_i}([a_i, b_i]^{d_i}, B_0)$ for some $\beta_i > 0$. Without loss of generality, each $W_j^{(i)}$ is assumed to have full row rank, and $[a_i, b_i]^{d_i}$ can be taken as $[0, 1]^{d_i}$.

Assumption 17 (Compositional structure) *The target density $f_0 \in \mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$ for some q, $\mathbf{l}$, $\mathbf{d}$, and $\boldsymbol{\beta}$.*

As f_0 is a density function on $\Omega = [0, 1]^d$, it follows that $l_0 = d$, $l_q = 1$. Let d_i denote the maximal number of variables (linear combinations of the inputs to $g^{(i)}$) on which each $g_j^{(i)}$ can depend. Thus, each component of $g^{(i)}$ is a d_i-variate function. It always holds that $d_i \le l_{i-1}$, and across various models, d_i can be significantly smaller than l_{i-1}. An interesting regime considered in Schmidt-Hieber (2020) is $d_i \le \min\{l_0, \ldots, l_{i-1}\}$ for all i, which means that no dimensions are added on deeper abstraction levels in the composition of functions. In particular, it implies that $d_i \le l_0 = d$.

The following theorem establishes non-asymptotic error bounds for data-augmented nonparametric noise contrastive estimation under Assumption 17. The proof is given in Appendix A.9.

Theorem 18 *Suppose that Assumptions 3, 8, and 17 hold, $f_{\mathcal{R}} \in \mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$, $\gamma \le \rho l/(1+\rho)$ and $\Gamma \ge L_0 \vee L$ where l, L are positive constants such that $l \le f_{\mathcal{R}} \le L$. Then for any $P, Q \in \mathbb{Z}_+$, the function class of ReLU-activated multi-layer perceptrons $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ with width $\mathcal{W} = 76(\lfloor\beta\rfloor + 1)^2 \max_i\{l_i 3^{d_i} d_i^{\lfloor\beta\rfloor+1}\} P\lceil\log_2(8P)\rceil$ and depth $\mathcal{D} = 21(q+1)(\lfloor\beta\rfloor+1)^2 Q\lceil\log_2(8Q)\rceil + 2\sum_{i=1}^q d_i + 3q + 2$, for $n \ge Pdim(\mathcal{F}_n)/\{(1+\rho)(1+\nu)\}$, the data-augmented nonparametric noise contrastive estimation $\tilde{f}_n$ defined in (4) satisfies*

$$\mathbb{E}_{S'}\|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \le c_{10}\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n} + c_{11}(\lfloor\beta\rfloor + 1)^4 d_*^{2\lfloor\beta\rfloor+\beta\vee 1} \max_i (PQ)^{-4\beta_i^*/d_i},$$

where $d_* = \max_i d_i$, $\beta_i^* = \beta_i \prod_{j=i+1}^q (\beta_j \wedge 1)$ $(1 \le i \le q-1)$, $\beta_q^* = \beta_q$, $a \wedge b :- \min\{a, b\}$, and c_{10}, c_{11} are constants only depending on ν, ρ, L_0, γ, Γ, q, B_0, $\{l_i\}_{i=1}^{q-1}$, and $f_{\mathcal{R}}$. Furthermore, if we set

$$\mathcal{W} = 228(\lfloor\beta\rfloor + 1)^2 \max_i\{l_i 3^{d_i} d_i^{\lfloor\beta\rfloor+1}\},$$

$$\mathcal{D} = 21(q+1)(\lfloor\beta\rfloor + 1)^2 \lceil n^{d_{i_0}/2(d_{i_0}+2\beta^*_{i_0})} \log_2(8n^{d_{i_0}/\{2(d_{i_0}+2\beta^*_{i_0})\}})\rceil + 2\sum_{i=1}^q d_i + 3q + 2,$$

where $i_0 \in \arg\min_i\{\beta_i^*/d_i\}$, then $\tilde{f}_n$ satisfies

$$\mathbb{E}_{S'}\|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \le c_{12}(\lfloor\beta\rfloor + 1)^9 d_*^{2\lfloor\beta\rfloor+\beta\vee 5} 9^{d_*} \phi_n (\log_2 n)^4, \qquad (11)$$

where $\phi_n = \max_i n^{-2\beta_i^*/(d_i+2\beta_i^*)}$, and c_{12} is a constant only depending on $\nu, \rho, L_0, \gamma, \Gamma, q, B_0$, $\{l_i\}_{i=1}^{q-1}$, and $f_{\mathcal{R}}$.

The β_i^*'s in Theorem 18 are referred to as the effective smoothness indices for compositional functions in $\mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, b_0)$. Observe that any function $f \in \mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$ has smoothness at least $\beta_* = \min_i \beta_i^*$. The condition $f_{\mathcal{R}} \in \mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$ in Theorem 18 implies that the reference distribution $f_{\mathcal{R}}$ also admits a compositional structure, whose complexity is constrained not to exceed that concerning f_0. In particular, it implies that the intrinsic dimension of $f_{\mathcal{R}}$ does not surpass that of f_0, and the effective smoothness indices of $f_{\mathcal{R}}$ are at least as large as those of f_0. Such a condition can be satisfied by setting $f_{\mathcal{R}}$ as the uniform distribution on $[0, 1]^d$.

It is possible that the target density function f_0 can be expressed as a composition of functions in various ways, which could lead to different convergence rates. Theorem 18 is applicable to any representation of compositional structure $f_0 = g^{(q)} \circ g^{(q-1)} \circ \cdots \circ g^{(1)}$, and as a result, our estimator can achieve the fastest convergence rate concerning all possible representations of target density. Since d_i should be consistently much smaller than d, the intrinsic dimension $d_* = \max_i d_i$ represents a significant improvement in mitigating the curse of dimensionality. To see this, note that $\phi_n \le n^{-2\beta_*/(d_*+2\beta_*)}$.

To derive the error bound in Theorem 18, we use an approximation result for compositional structure functions (Lemma C.2 in Wu et al. (2024) or Lemma 30 in Appendix B). The approximation result is with respect to the L_∞ norm on $[0, 1]^d$, at the price that the network width should incorporate a factor roughly 3^{d_*}. In the regime $d_i \le \min\{l_0, \ldots, l_{i-1}\}$ for all i, the convergence rate in (11) is, up to a logarithmic factor, the same as the minimax rate in the nonparametric regression model under the compositional structure assumption on regression function (Schmidt-Hieber, 2020). Bos and Schmidt-Hieber (2024) derived a similar error bound $C_d \phi_n (\log n)^5$ for their kernel-based two-stage density estimator, which is novel. But the prefactor C_d in their error bound depends on d exponentially; in contrast, our result depends exponentially on the intrinsic dimension d_*, which could be much smaller than d. Furthermore, Bos and Schmidt-Hieber (2024) requires the network parameters (weights and biases) to be bounded by 1 and satisfy a sparsity constraint, and we do not make such network assumptions for our method.

5. Numerical studies

In the numerical studies, we compare the performance of our method with the state-of-the-art density estimation methods, namely Roundtrip (Liu et al., 2021), kernel-based two-stage estimation (Bos and Schmidt-Hieber, 2024), flow contrastive estimation (Gao et al., 2020), and telescoping density-ratio estimation (Rhodes et al., 2020), using both simulated and real data. We present the models and results of the simulation studies in Sections 5.1 and 5.2, respectively. The detailed simulation setups are deferred to Appendix C. The real data studies are presented in Section 5.3. The source code is available at `https://github.com/Li-Chenqhao/DANNCE`.

5.1 Simulation models

In the simulations, we consider the following three models of 2-D distributions studied in Liu et al. (2021). Let $N_d(\mathbf{m}, \boldsymbol{\Sigma})$ be the d-dimensional Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$; specifically, we denote it as $N(m, \sigma^2)$ when $d = 1$.

1. Independent Gaussian mixture model: $\mathbf{x} = (x_1, x_2)$ with $x_i \sim (1/3)(N(-1, 0.1^2) + N(0, 0.1^2) + N(1, 0.1^2))$, $i = 1, 2$.
2. Eight-octagon Gaussian mixture model: $\mathbf{x} \sim (1/8)\sum_{i=1}^{8} N_2(\mathbf{m}_i, \mathbf{\Sigma}_i)$, where $\mathbf{m}_i = (3\cos(\pi i/4), 3\sin(\pi i/4))$ and
$$\mathbf{\Sigma}_i = \begin{bmatrix} \cos^2(\frac{\pi i}{4}) + 0.16^2 \sin^2(\frac{\pi i}{4}) & (1 - 0.16^2)\sin(\frac{\pi i}{4})\cos(\frac{\pi i}{4}) \\ (1 - 0.16^2)\sin(\frac{\pi i}{4})\cos(\frac{\pi i}{4}) & \sin^2(\frac{\pi i}{4}) + 0.16^2\cos^2(\frac{\pi i}{4}) \end{bmatrix}$$
for $i = 1, 2, \ldots, 8$.
3. Involute model: $\mathbf{x} = (x_1, x_2)$ with $x_1 \sim N(r\sin(2r), 0.4^2)$ and $x_2 \sim -N(r\cos(2r), 0.4^2)$, where $r \sim \text{Uniform}(0, 2\pi)$.

Model 1 and Model 2 are Gaussian mixture models with multiple components. Model 1 involves two independent dimensions, each of which comprises a mixture of three Gaussian distributions with equal weights. In contrast, the two dimensions of Model 2 are conjuncted by covariance matrices that vary across components, resulting in a complicated covariance structure. Model 3 comprises an infinite mixture of Gaussian distributions that concentrate around an involute of a circle, exhibiting a highly nonlinear structure. While inherently unimodal, this distribution would manifest multimodality when conditioned on either of its dimensions. The intricate characteristics of these models typically cannot be effectively captured by traditional methods. Note that the distributions in all three models are supported on $\mathbb{R}^2$, aligned with our theoretical results in Section 3.5. The simulation studies empirically validate the effectiveness of our method for density estimation with unbounded support.

To evaluate the performance of data-augmented nonparametric noise contrastive estimation in higher-dimensional scenarios, we also consider Model 1 with different dimension d, that is, $\mathbf{x} = (x_1, \ldots, x_d)$ and $x_i \sim (1/3)(N(-1, 0.1^2) + N(0, 0.1^2) + N(1, 0.1^2))$ for $i = 1, \ldots, d$, which is a Gaussian mixture model with 3^d components.

5.2 Simulation results

To compare the performance in 2-D density estimation of different methods, we calculate numerically some discrepancy measures of distributions, between the true density f_0 and the estimated density $\hat{f}$, including the L_2 distance, Kullback-Leibler (KL) divergence, and Jensen-Shannon divergence. Since the KL divergence is asymmetric, we calculate both $\text{KL}(f_0 \| \hat{f})$ and $\text{KL}(\hat{f} \| f_0)$. In addition, we compare the first and second moments of the true and estimated densities. A good density estimation shall give similar moments to the true ones.

The configurations for our data-augmented nonparametric noise contrastive estimation method and neural network training are detailed in Appendix C. For each model, a sample of size $20,000$ is generated. After their corresponding network training processes, the density estimators of Roundtrip, kernel-based two-stage estimation, flow contrastive estimation, telescoping density-ratio estimation, and our data-augmented nonparametric noise contrastive estimation are assessed via discrepancy measures and moments. We visualize the true and estimated densities of the three models in Figure 1, and present the means and standard deviations of assessment quantities based on 20 repetitions of experiments in Tables 1 and 2.

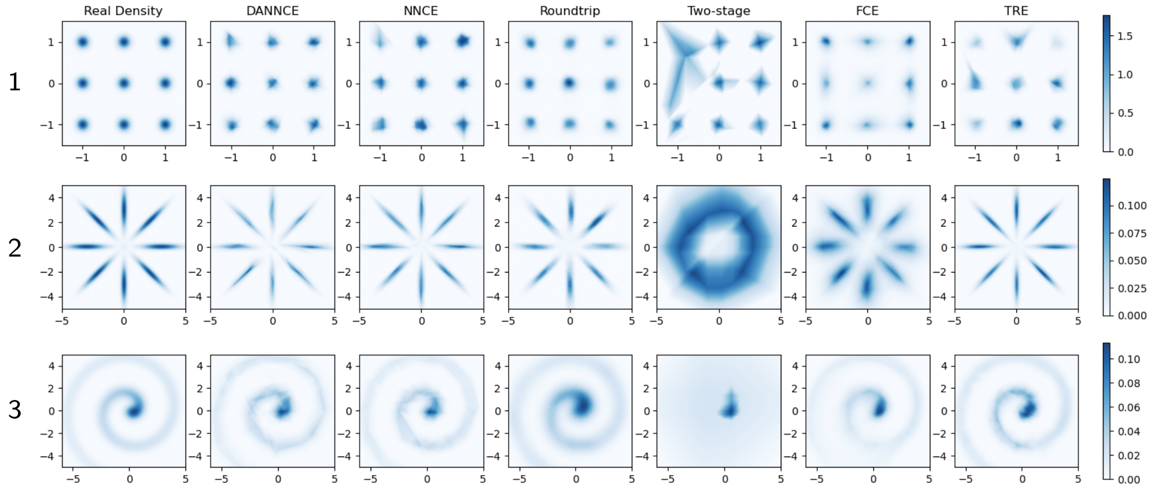

Figure 1: True density and density estimations visualized on 2-D bounded regions. Each row gives results of a model, and the first column shows corresponding true densities. DANNCE, NNCE, Two-stage, FCE, and TRE refer to our data-augmented nonparametric noise contrastive estimation, the nonparametric noise contrastive estimation (our method without data augmentation ($\rho = 0$)), kernel-based two-stage estimation, flow contrastive estimation, and telescoping density-ratio estimation, respectively. Each color bar indicates the scaling of true and estimated densities in a model.

As displayed by Fig. 1, our method, Roundtrip, and telescoping density-ratio estimation effectively learn the data distributions of all three models. Moreover, our method demonstrates higher accuracy in Model 3, especially in a neighborhood of the mode. In comparison, kernel-based two-stage estimator struggles to accurately identify the components of Model 1. For Models 2 and 3, the two-stage estimator fails to capture the underlying structure of data distribution, leading to nearly uniform estimations between components. These results can be attributed to the fact that the performance of the two-stage estimator is highly dependent on that of kernel density estimation.

In terms of moments and discrepancy measures, our method generally outperforms Roundtrip, kernel-based two-stage estimator, and flow contrastive estimation across the three models under consideration. The performance of our method and telescoping density-ratio estimation is comparable: our method gives better results in terms of discrepancy measures, while telescoping density-ratio estimation is more effective in moment estimation. In this 2-D case, the nonparametric noise contrastive estimator (our method without data augmentation) performs comparably to our data-augmented estimator; but for higher-dimensional experiments presented later, our data-augmented estimator significantly outperforms the estimator without data-augmentation. We also observe that the KL divergences $KL(p_d \| p_m)$ for Roundtrip are negative across the three models, while the inverse

KL divergences $KL(p_d \| p_m)$ are very high. This suggests that Roundtrip is possibly unnormalized and over-estimate density at certain low-density region.

Table 1: Simulation Results of Moments

Model	Method	1st moment		2nd moment		
	True	0	0	.6767	0	.6767
	Roundtrip	6.450 ± 47.57	6.244 ± 30.59	.9263 ± .0339	−.0167 ± .0342	.9121 ± .0368
	Two-stage	125.3 ± 399.4	−57.00 ± 296.2	1.126 ± .5761	−.0287 ± .2069	1.091 ± .4819
1	FCE	13.80 ± 79.71	2.048 ± 77.60	.8204 ± .0793	.0080 ± .0315	.8168 ± .0734
	TRE	**−6.613 ± 23.30**	**3.716 ± 28.83**	.4860 ± .0504	.0069 ± .0195	.5025 ± .0573
	DANNCE	9.071 ± 44.58	−.5126 ± 38.58	.7070 ± .0323	**.0149 ± .0171**	.7045 ± .0382
	NNCE	16.79 ± 55.95	16.59 ± 30.60	**.6774 ± .0388**	−.0232 ± .0178	**.6923 ± .0489**
	True	0	0	4.849	0	4.849
	Roundtrip	−3.563 ± 15.60	−2.758 ± 10.99	7.043 ± .4121	.0131 ± .2510	7.175 ± .4197
	Two-stage	15.49 ± 55.48	−8.401 ± 61.57	17.22 ± 7.679	.4285 ± 2.251	16.07 ± 7.710
2	FCE	5.617 ± 21.98	−3.554 ± 22.79	6.458 ± .8541	.0273 ± .3570	6.222 ± .6119
	TRE	−2.436 ± 11.79	**.3589 ± 10.24**	5.013 ± .2393	.0713 ± .1388	5.143 ± .2510
	DANNCE	1.247 ± 10.57	1.039 ± 10.51	5.115 ± .1904	−.0012 ± .1161	**5.015 ± .2012**
	NNCE	**5.630 ± 7.808**	1.884 ± 11.02	**4.926 ± .2150**	**−.0096 ± .0745**	5.054 ± .2037
	True	−.3894	−.2094	6.313	−.3479	4.918
	Roundtrip	−.5348 ± .0488	−.1978 ± .0630	8.292 ± .2407	−.5693 ± .1711	6.679 ± .1637
	Two-stage	−.2651 ± .4579	−.0880 ± .1649	11.44 ± 5.042	−.1040 ± .4861	9.544 ± 4.086
3	FCE	−.4529 ± .1297	−.1083 ± .1843	6.355 ± .4189	**−.3458 ± .2354**	5.312 ± .3437
	TRE	**−.3769 ± .0409**	−.1859 ± .0396	**6.335 ± .2681**	−.3027 ± .0778	**4.917 ± .0891**
	DANNCE	−.4046 ± .0970	−.1966 ± .0658	6.440 ± .2312	−.2563 ± .0824	4.956 ± .1414
	NNCE	−.4145 ± .0745	**−.2089 ± .0624**	6.497 ± .2219	−.2612 ± .0725	4.956 ± .1539

Notes: True moments and moment estimations calculated from estimated densities. DANNCE, NNCE, Two-stage, FCE, and TRE refer to our data-augmented nonparametric noise contrastive estimation, the nonparametric noise contrastive estimation (our method without data augmentation ($\rho = 0$)), kernel-based two-stage estimation, flow contrastive estimation, and telescoping density-ratio estimation, respectively. Each entry is the mean ± standard deviation from 20 replicas. The first moments for Model 1 and Model 2 are in 10^{-3} and 10^{-2}, respectively. The moments with true value 0 are compared in terms of standard deviation. The best performance among all methods are shown in bold.

Table 2: Simulation Results of Discrepancy Measures

Model	Method	L_2 Distance	$KL(p_d\|p_m)$	$KL(p_m\|p_d)$	$JS(p_d\|p_m)$
1	Roundtrip	$.3865 \pm .0255$	$-2.131 \times 10^{-4} \pm 1.032 \times 10^{-5}$	812.9 ± 37.93	406.4 ± 18.97
	Two-stage	$.6781 \pm .2707$	4.167 ± 7.964	4.445 ± 4.275	4.306 ± 3.789
	FCE	$.3954 \pm .0801$	$.1385 \pm .0962$	$1.950 \pm .7429$	$1.044 \pm .3952$
	TRE	$.4182 \pm .1110$	$.4076 \pm .1455$	$.2128 \pm .3781$	$.3102 \pm .2478$
	DANNCE	$\mathbf{.2129} \pm \mathbf{.0210}$	$.0373 \pm .0556$	$.1468 \pm .0557$	$.0920 \pm .0231$
	NNCE	$.2228 \pm .0404$	$.0387 \pm .0708$	$\mathbf{.0933} \pm \mathbf{.0788}$	$\mathbf{.0661} \pm \mathbf{.0625}$
2	Roundtrip	$.1353 \pm .0110$	$-2.896 \times 10^{-3} \pm 1.634 \times 10^{-4}$	81.63 ± 5.508	40.81 ± 2.754
	Two-stage	$.3323 \pm .0908$	2.486 ± 7.424	27.35 ± 16.32	14.92 ± 7.074
	FCE	$.1537 \pm .0061$	$.4621 \pm .0780$	$4.722 \pm .7917$	$2.592 \pm .3774$
	TRE	$.0469 \pm .0080$	$-.0317 \pm .0364$	$.0774 \pm .0384$	$.0228 \pm .0052$
	DANNCE	$.0471 \pm .0045$	$.0492 \pm .0493$	$.1251 \pm .0516$	$.0871 \pm .0359$
	NNCE	$\mathbf{.0415} \pm \mathbf{.0081}$	$.0032 \pm .0327$	$\mathbf{.0419} \pm \mathbf{.0814}$	$\mathbf{.0219} \pm \mathbf{.0069}$
3	Roundtrip	$5.312 \pm .2948$	$-1.936 \times 10^{-3} \pm 6.437 \times 10^{-5}$	53.12 ± 2.259	26.56 ± 1.129
	Two-stage	10.64 ± 1.285	2.769 ± 6.090	3.094 ± 1.394	2.932 ± 2.446
	FCE	$4.182 \pm .4854$	$.0774 \pm .0486$	$.2678 \pm .0872$	$.1736 \pm .0306$
	TRE	$2.651 \pm .4691$	$.0193 \pm .0274$	$.0326 \pm .0282$	$.0259 \pm .0231$
	DANNCE	$2.463 \pm .1960$	$.0679 \pm .0274$	$.0467 \pm .0260$	$.0573 \pm .0063$
	NNCE	$\mathbf{2.442} \pm \mathbf{.2687}$	$.0193 \pm .0282$	$\mathbf{.0218} \pm \mathbf{.0247}$	$\mathbf{.0206} \pm \mathbf{.0047}$

Notes: Discrepancy measures between the target density and the estimated densities. DANNCE, NNCE, Two-stage, FCE, and TRE refer to our data-augmented nonparametric noise contrastive estimation, the nonparametric noise contrastive estimation (our method without data augmentation ($\rho = 0$)), kernel-based two-stage estimation, flow contrastive estimation, and telescoping density-ratio estimation, respectively. Each entry is the mean ± standard deviation from 20 replicas. The L_2 distances for Model 3 are in 10^{-2}. The best performance among all methods are shown in bold, except for $KL(p_d\|p_m)$.

For higher-dimensional cases, we evaluate the performance of a density estimator by calculating the Spearman rank correlation between the true density and the estimated density on a test set. A density estimation with a Spearman rank correlation close to 1 would be preferred. For each dimension $d = 3, 4, \ldots, 10$, we sample a training set of size $10000 \cdot d^3$ and a test set of size $100,000$ from the data distribution. Following model training, the Spearman rank correlations are computed and displayed in Table 3. For our method and the telescoping density-ratio estimation, we present the mean and maximum of Spearman rank correlations based on 5 repetitions of experiments. For Roundtrip, kernel-based two-stage estimator, and flow contrastive estimation, we report their best results from 5 replicas, due to the instability exhibited by these methods in this simulation study.

Table 3 demonstrates that our method outperforms Roundtrip and nonparametric noise contrastive estimation almost consistently across this experiment. Moreover, our method outperforms telescoping density-ratio estimation when $d \leq 7$, while the latter exhibits superior accuracy when $d \geq 8$.

The Spearman rank correlations calculated from kernel-based two-stage estimation are $0.991, 0.979$, and 0.897 for $d = 3, 4$, and 5, respectively. The performance of the two-stage estimator is impressive at $d = 3$ and 4, but shows a sharp decline as d increases to 5. When the dimension $d \geq 6$, the two-stage estimator cannot produce valid results for the data distribution model being considered. Flow contrastive estimation yields Spearman rank correlations of 0.831 and 0.762 for $d = 3$ and 4 respectively, while it performs poorly when $d \geq 5$ for the model under consideration.

Table 3: Simulation Results of Higher Dimensions

Dimension d		3	4	5	6	7	8	9	10
max	Roundtrip	0.959	0.957	0.950	0.935	0.923	0.813	0.791	0.789
	TRE	**0.977**	0.963	0.949	0.933	**0.927**	**0.918**	**0.934**	**0.942**
	DANNCE	0.976	0.972	**0.964**	**0.952**	0.920	0.914	0.898	0.882
	NNCE	0.977	**0.974**	0.958	0.939	0.916	0.891	0.872	0.841
mean	TRE	0.954	0.936	0.910	0.924	0.912	**0.913**	**0.924**	**0.929**
	DANNCE	**0.973**	**0.969**	**0.959**	**0.941**	**0.918**	0.908	0.892	0.875
	NNCE	0.969	0.968	0.955	0.933	0.911	0.884	0.865	0.821

Notes: Spearman rank correlations between true density and density estimations of different methods in higher dimensions, with the best performance shown in bold. DANNCE, NNCE, and TRE refer to our data-augmented nonparametric noise contrastive estimation, the nonparametric noise contrastive estimation (our method without data augmentation ($\rho = 0$)), and telescoping density-ratio estimation, respectively. In DANNCE, ρ is set to 0.3 for $d \leq 8$ and to 1 for $d = 9$ and 10. Max and mean denote the maximum and average values based on 5 replicas.

We also conduct experiments to illustrate the influence of the hyperparameter ρ on the performance of our method. Figure 2 visualizes the L_2 distances (2-D case) and Spearman rank correlations (higher-dimensional cases) between the true and estimated densities with different values of ρ. The means and 95% confidence intervals of these assessment quantities are presented, based on 50 repeated trials for the 2-D case and 20 trials for the higher-dimensional cases. In the 2-D case, increasing ρ cannot lead to a better estimation. However, the performance exhibits a significant increase as ρ varies from 0 to 5, in the higher-dimensional cases with $d \geq 5$. For dimensions of 3 and 4, similar results are obtained across different values of ρ. Relatively high Spearman rank correlations can be achieved with $\rho = 1$, which supports our theoretical optima range of $[1/3, 1]$ in Section 3.3. Balancing estimation accuracy and computational efficiency, such a range remains effective. We can employ values of ρ around 1/3 in low-dimensional cases and values around 1 in higher-dimensional cases.

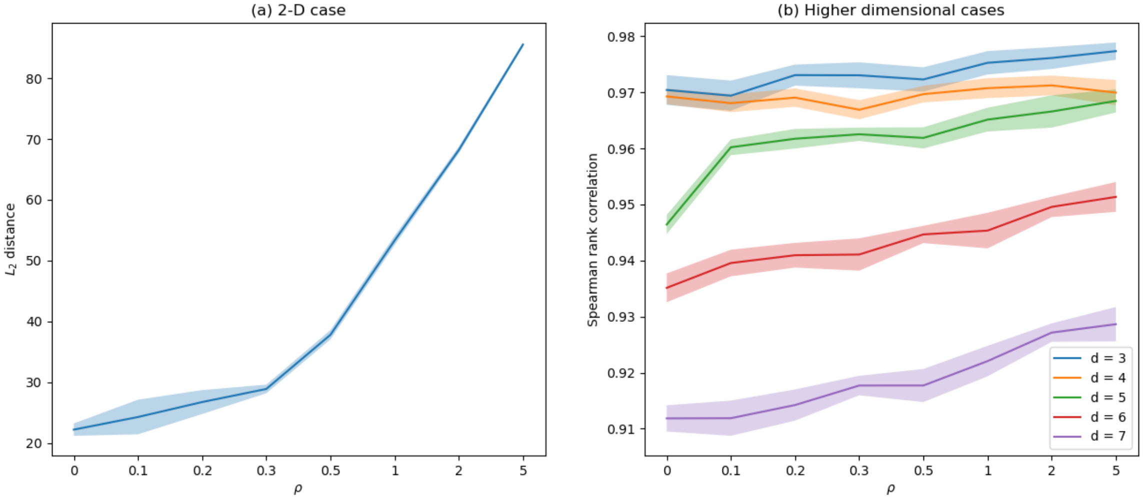

Figure 2: The performance of our method as ρ changes. We use a sequence of ρ values, 0, 0.1, 0.2, 0.3, 0.5, 1, 2, and 5. The lines denote the means of the assessment quantities, and the shaded bands indicate the corresponding 95% confidence intervals from repeated trials. (a) The L_2 distance between true and estimated densities in the 2-D case. All values have been multiplied by 100. Results are based on 50 replicas. (b) The Spearman rank correlation between true and estimated densities, for d = 3, 4, 5, 6, and 7. A higher Spearman rank correlation indicates better performance. Results are based on 20 replicas.

5.3 Real data studies: Shuttle and Mammography datasets

We apply our method to two real-world datasets, Shuttle (Feng et al., 1993) and Mammography (Woods et al., 1993), both of which are available from the ODDS database (Rayana, 2016). We conduct anomaly/outlier detection task on these datasets, a practical application of density estimation. Specifically, a data point with a significantly low density value is considered likely to be an anomaly.

The Shuttle dataset contains $49,097$ samples from seven classes. Each sample consists of an 8-dimensional feature vector and a class label. For the outlier detection task, the largest class (Class 1) is regarded as the inlier set, the second-largest class (Class 4) is discarded, and all other classes are combined to form the outlier set. Consequently, approximately 7% of the samples ($3,511$ instances) are considered outliers. The Mammography dataset contains $11,183$ samples collected for detecting microcalcifications in mammogram images. Each sample consists of 6 features extracted from mammograms, and a binary label indicating whether the instance is normal (inlier) or abnormal (outlier). Among these samples, 260 instances (approximately 2.32%) are outliers. For each dataset, we first combine the inlier and outlier samples and then randomly split the data, with 90% of the samples used for training and 10% for testing.

For our data-augmented nonparametric noise contrastive estimation method, we set the hyperparameters $\nu = 5$ and $\rho = 1$. The Gaussian distribution $N_d(\mathbf{m}, \sigma^2\mathbf{I})$ is used as the reference distribution, where $\mathbf{m}$ is the sample mean and σ^2 is the maximum sample variance across all data dimensions, calculated from the training set. We compare our data-augmented nonparametric noise contrastive estimation with Roundtrip and telescoping density-ratio estimation. The methods are evaluated using precision at k, which measures the proportion of true outliers among the top-k points identified as outliers by the detector; in our case, they correspond to the top-k points with the lowest estimated density. We set k as the number of true outliers in the test set, and a precision closer to 1 indicates better performance.

The analysis results are reported in Table 4, which presents the average precision based on three repeated trials. From Table 4, one can see that our method exhibits superior performance on the Shuttle dataset compared with Roundtrip and telescoping density-ratio estimation. For the Mammography dataset, the precision of our method is comparable to Roundtrip and is higher than that of telescoping density-ratio estimation. The anomaly percentage of the Mammography dataset is relatively low (2.32%), and our method performs competitively and stably. The nonparametric noise contrastive estimation (our method without data augmentation) does not perform well on the two datasets, and thus its results are not reported.

Table 4: Anomaly detection results on the Shuttle and Mammography datasets

	Roundtrip	TRE	DANNCE
Shuttle	0.959	0.914	**0.970**
Mammography	**0.482**	0.432	0.481

Notes: The precision at k of different methods on two ODDS datasets, with the best performance shown in bold. DANNCE and TRE refer to our data-augmented nonparametric noise contrastive estimation and telescoping density-ratio estimation, respectively. Each entry is the average value from three replicas. The results of Roundtrip are from Liu et al. (2021).

6. Conclusions

In this paper, we propose a data-augmented nonparametric contrastive learning approach to density estimation, which leverages the power of deep neural networks in approximating high-dimensional functions. Our density estimator is simple and easy-to-implement. The proposed data augmentation technique addresses the selection of reference distribution to some extent and improves performance in some practical applications. It also enjoys some theoretical merits, for example, the lower-bounded density assumption can be avoided. Under relative mild conditions, we establish non-asymptotic error bounds for our estimator and provide theoretical guarantees for our method. With certain choice of network parameters, we show that our error bound can achieve the minimax optimal rate $n^{-2\beta/(d+2\beta)}$ of standard nonparametric regression. Under a compositional structural assumption on the target density, we have shown that our estimator can circumvent the curse of dimensionality with a faster convergence rate $\max_i n^{-2\beta_i^*/(d_i+2\beta_i^*)}$, which depends merely on the intrinsic dimension of the target density, and attains the minimax optimal rate under compositional structure assumption.

Acknowledgments

The authors thank the Action Editor Professor Aapo Hyvärinen and four anonymous reviewers for their insightful comments and constructive suggestions that helped improve our paper. Yuanyuan Lin's research was partially supported by the Hong Kong Research Grants Council (Grant No: 14304523) and The Hong Kong Jockey Club Equine Welfare Research Foundation, Direct Grants for Research, The Chinese University of Hong Kong.

Appendix A. Proofs of lemmas and theorems

In this section of the Appendix, we include the proofs for the lemmas and theorems stated in Sections 3 - 4.

A.1 Proof of Lemma 4

Proof Notice that $f_0 = (1+\rho)f_{\mathcal{M}} - \rho f_{\mathcal{R}} \leq (1+\rho)f_{\mathcal{M}}$ and $\tilde{f}_n = (1+\rho)\hat{f}_{\mathcal{M}} - \rho f_{\mathcal{R}}$, we have

$$\begin{aligned}\|\tilde{f}_n - f_0\|^2_{L_2(f_0)} &= \int_\Omega |\tilde{f}_n(x) - f_0(x)|^2 f_0(x)dx \\ &= (1+\rho)^2 \int_\Omega |\hat{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x)dx \\ &\leq (1+\rho)^3 \int_\Omega |\hat{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_{\mathcal{M}}(x)dx \\ &= \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}\end{aligned}$$

given any training sample S. Taking expectation with respect to S gives the desired inequality. ■

A.2 Proof of Lemma 5

Proof We first prove a somewhat more general result than Lemma 5. Suppose that there exist positive constants l, L, l_*, L_*, γ and Γ such that $l \leq f_{\mathcal{R}} \leq L$, $l_* \leq f_* \leq L_*$, and $\gamma \leq f \leq \Gamma$ for all $f \in \mathcal{F}$ on $[0,1]^d$, we show that $c_1\|f - f_*\|^2_{L_2(f_*)} \leq R(f) - R(f_*) \leq c_2\|f - f_*\|^2_{L_2(f_*)}$ for some positive constants c_1 and c_2.

In this case, there exist positive constants m and M such that

$$\begin{gathered} mf_{\mathcal{R}} \leq f_* \leq Mf_{\mathcal{R}}, \\ mf_{\mathcal{R}} \leq f \leq Mf_{\mathcal{R}},\ \forall f \in \mathcal{F}. \end{gathered} \tag{12}$$

Specifically, we can choose $m = (l_* \wedge \gamma)/L$ and $M = (L_* \vee \Gamma)/l$.

According to the definition, we can rewrite the noise contrastive risk as

$$R(f) = -\Big\{ \int f_*(u) \log \frac{f(u)}{f(u) + \nu f_{\mathcal{R}}(u)} du + \nu \int f_{\mathcal{R}}(u) \log \frac{\nu f_{\mathcal{R}}(u)}{f(u) + \nu f_{\mathcal{R}}(u)} du \Big\}.$$

Therefore,

$$R(f) - R(f_*) = \int \Big[\{f_*(u) + \nu f_{\mathcal{R}}(u)\} \log \frac{f(u) + \nu f_{\mathcal{R}}(u)}{f_*(u) + \nu f_{\mathcal{R}}(u)} + f_*(u) \log \frac{f_*(u)}{f(u)} \Big] du. \tag{13}$$

We denote the integrand in (13) as $W(f, f_*, f_{\mathcal{R}})(u)$. Consider the function

$$g_a(b) = \frac{e^a + \nu c}{e^a} \log \frac{e^b + \nu c}{e^a + \nu c} + a - b.$$

We have

$$g_a'(b) = \frac{e^a + \nu c}{e^a} \frac{e^b}{e^b + \nu c} - 1, \quad g_a''(b) = \frac{e^a + \nu c}{e^a} \frac{\nu c}{(e^b + \nu c)^2} e^b,$$

and that $g_a(a) = 0$, $g_a'(a) = 0$. For any a, b such that $mc \le e^a \le Mc, mc \le e^b \le Mc$, we have

$$g_a''(b) \ge \frac{m\nu}{M(M + \nu)}.$$

Hence,

$$g_a(b) \ge \frac{m\nu}{2M(M + \nu)}(b - a)^2.$$

While $a, b \le \log(L_* \vee \Gamma)$, it holds that $|e^b - e^a| \le (L_* \vee \Gamma)|b - a|$ and

$$g_a(b) \ge \frac{m\nu}{2M(M + \nu)(L_* \vee \Gamma)^2}(e^b - e^a)^2.$$

Taking $a = \log f_*(u), b = \log f(u)$ and $c = f_{\mathcal{R}}(u)$, we obtain

$$\frac{W(f, f_*, f_{\mathcal{R}})(u)}{f_*(u)} \ge \frac{m\nu}{2M(M + \nu)(L_* \vee \Gamma)^2}\{f(u) - f_*(u)\}^2. \tag{14}$$

Now consider the function

$$h_a(b) = \frac{a + \nu c}{a} \log \frac{b + \nu c}{a + \nu c} + \log \frac{a}{b}.$$

Likewise, when $b \ge \gamma$, it holds that $h_a(b) \le (b - a)^2/(2\gamma^2)$. Taking $a = f_*(u)$, $b = f(u)$ and $c = f_{\mathcal{R}}(u)$, we have

$$\frac{W(f, f_*, f_{\mathcal{R}})(u)}{f_*(u)} \le \frac{1}{2\gamma^2}\{f(u) - f_*(u)\}^2. \tag{15}$$

Multiplying by $f_*(u)$ and taking integrals in (14) and (15), we have $c_1\|f - f_*\|^2_{L_2(f_*)} \le R(f) - R(f_*) \le c_2\|f - f_*\|^2_{L_2(f_*)}$, where $c_1 = m\nu/\{2M(M + \nu)(L_* \vee \Gamma)^2\}$ and $c_2 = 1/(2\gamma^2)$.

Under Assumption 3, there exist positive constants l, L, and L_0 such that $l \le f_{\mathcal{R}} \le L$, $f_0 \le L_0$. Thus, $\rho l/(1 + \rho) \le f_{\mathcal{M}} \le L \vee L_0$. By definition of $\mathcal{F}_n$, $\gamma \le f \le \Gamma$ holds for all $f \in \mathcal{F}_n$. Therefore, $c_1\|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \le R(f) - R(f_{\mathcal{M}}) \le c_2\|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}$ holds for any $f \in \mathcal{F}_n$, where c_1 and c_2 are constants only depending on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$. This concludes the proof of Lemma 5. ■

A.3 Proof of Lemma 6

Proof By Lemma 5, there exists positive constants c_1 and c_2 such that

$$c_1\|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \le R(\hat{f}_{\mathcal{M}}) - R(f_{\mathcal{M}}) \le c_2\|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}.$$

Taking expectation with respect to sample S, we have

$$\mathbb{E}_S\|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \le \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - R(f_{\mathcal{M}})].$$

For any $\epsilon > 0$, there exists $f_\epsilon \in \mathcal{F}_n$ such that $\|f_\epsilon - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq \inf_{f\in\mathcal{F}_n} \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} + \epsilon$. Since $\hat{f}_{\mathcal{M}}$ minimizes $\hat{R}(f)$, we have $\hat{R}(\hat{f}_{\mathcal{M}}) \leq \hat{R}(f_\epsilon)$, and that

$$\mathbb{E}_S \hat{R}(\hat{f}_{\mathcal{M}}) \leq \mathbb{E}_S \hat{R}(f_\epsilon) = R(f_\epsilon).$$

Therefore,

$$\begin{aligned}
\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_0)} &\leq \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - R(f_{\mathcal{M}})] \\
&= \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - R(f_\epsilon) + R(f_\epsilon) - R(f_{\mathcal{M}})] \\
&\leq \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] + \frac{1}{c_1}[R(f_\epsilon) - R(f_{\mathcal{M}})] \\
&\leq \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] + \frac{c_2}{c_1}\|f_\epsilon - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \\
&\leq \frac{1}{c_1}\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] + \frac{c_2}{c_1}\inf_{f\in\mathcal{F}}\|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} + \frac{c_2}{c_1}\epsilon.
\end{aligned}$$

Letting $\epsilon \to 0$, we get the desired result. ■

A.4 Proof of Lemma 7

Proof By definition, we have

$$\begin{aligned}
&R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}}) \\
=& -\int \Big\{ f_{\mathcal{M}}(u)\log\frac{\hat{f}_{\mathcal{M}}(u)}{\hat{f}_{\mathcal{M}}(u) + \nu f_{\mathcal{R}}(u)} + \nu f_{\mathcal{R}}(u)\log\frac{\nu f_{\mathcal{R}}(u)}{\hat{f}_{\mathcal{M}}(u) + \nu f_{\mathcal{R}}(u)}\Big\}du \\
&+ \frac{1}{(1+\rho)n}\sum_{i=1}^{(1+\rho)n}\log\frac{\hat{f}_{\mathcal{M}}(Y_i)}{\hat{f}_{\mathcal{M}}(Y_i) + \nu f_{\mathcal{R}}(Y_i)} + \frac{1}{(1+\rho)n}\sum_{i=1}^{\nu(1+\rho)n}\log\frac{\nu f_{\mathcal{R}}(Z_i)}{\hat{f}_{\mathcal{M}}(Z_i) + \nu f_{\mathcal{R}}(Z_i)}.
\end{aligned}$$

Partition the sample $S = \{Y_1, \ldots, Y_{n+\rho n}, Z_1, \ldots, Z_{\nu(n+\rho n)}\}$ into contrastive sequences $\bigcup_{i=1}^n S_i$. Each contrastive sequence S_i is composed of an augmented data point and ν contrastive noise data points, i.e. $S_i = (Y_i, Z_{\nu(i-1)+1}, \ldots, Z_{\nu i})$ for $i = 1, \ldots, (1+\rho)n$. Let S' be an independent copy of S. Define

$$g(f, (y, z_1, \ldots, z_\nu)) :- \log\frac{f(y)}{f(y) + \nu f_{\mathcal{R}}(y)} + \sum_{j=1}^{\nu}\log\frac{\nu f_{\mathcal{R}}(z_j)}{f(z_j) + \nu f_{\mathcal{R}}(z_j)},$$

for any $f \in \mathcal{F}_n$ and sequence $(y, z_1, \ldots, z_\nu)$. With these notations, we can rewrite $R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})$ as

$$R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}}) = \frac{1}{(1+\rho)n}\sum_{i=1}^{(1+\rho)n}\{g(\hat{f}_{\mathcal{M}}, S_i) - \mathbb{E}_{S'}g(\hat{f}_{\mathcal{M}}, S'_i)\}. \tag{16}$$

To provide an upper bound for $\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})]$, we leverage expression (16). By Assumption 3 and (12), for any $f \in \mathcal{F}_n$, we have

$$\begin{aligned}-g(f,(y,z_1,\cdots,z_\nu)) &= \log\frac{f(y)+\nu f_{\mathcal{R}}(y)}{f(y)} + \sum_{j=1}^{\nu}\log\frac{f(z_j)+\nu f_{\mathcal{R}}(z_j)}{\nu f_{\mathcal{R}}(z_j)}\\ &\le \log\Big(1+\frac{\nu}{m}\Big) + \nu\log\Big(1+\frac{M}{\nu}\Big).\end{aligned}$$

And it is straightforward to see that $g(f,\cdot)\le 0$. Therefore, for all $f \in \mathcal{F}_n, g(f,\cdot)\in[-K,0]$ where $K = \log(1+\nu/m)+\nu\log(1+M/\nu)$. It follows from Lemma 29 in Appendix B that (set $\alpha = t$ and $\epsilon = 1/2$ therein), for any $t > 0$,

$$\begin{aligned}&\Pr\Big(\frac{1}{(1+\rho)n}\sum_{i=1}^{(1+\rho)n}\{g(\hat{f}_{\mathcal{M}},S_i) - \mathbb{E}_{S'}g(\hat{f}_{\mathcal{M}},S_i')\} > t\Big)\\ &\le \Pr\Big(\exists f\in\mathcal{F}_n, \frac{1}{(1+\rho)n}\sum_{i=1}^{(1+\rho)n}\{g(f,S_i)-\mathbb{E}_{S'}g(f,S_i')\} > t\Big)\\ &\le 4\mathbb{E}\mathcal{N}_1\Big(\frac{t(1+\rho)n}{10},\mathcal{G},S_{1:(1+\rho)n}\Big)\exp\Big(-\frac{3t(1+\rho)n}{160K}\Big),\end{aligned}$$

where $\mathcal{G} = \{g(f,\cdot) : f \in \mathcal{F}_n\}$. Since

$$\mathbb{E}\mathcal{N}_1\Big(\frac{t(1+\rho)n}{10},\mathcal{G},S_{1:(1+\rho)n}\Big) \le \mathcal{N}_1\Big(\frac{t(1+\rho)n}{10},\mathcal{G},(1+\rho)n\Big) \le \mathcal{N}_\infty\Big(\frac{t}{10},\mathcal{G},(1+\rho)n\Big),$$

our next step is to bound $\mathcal{N}_\infty(t/10,\mathcal{G},(1+\rho)n)$ by $\mathcal{N}_\infty(\epsilon,\mathcal{F}_n,(1+\nu)(n+\rho n))$ for some $\epsilon > 0$.

By the definition of uniform covering number, for any $u_{1:(1+\nu)(n+\rho n)}$, there exists $\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}} \subset \mathbb{R}^{(1+\nu)(n+\rho n)}$ with cardinality $\big|\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}}\big| \le \mathcal{N}_\infty(\epsilon,\mathcal{F}_n,(1+\nu)(n+\rho n))$, such that for all $f\in\mathcal{F}_n$, there exists vector $\bar{f}\in\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}}$ satisfying $\max_{1\le i\le(1+\nu)(n+\rho n)}|f(u_i)-\bar{f}_i| < \epsilon$. Note that for any u, $\gamma \le f(u) \le \Gamma$, we can assume without loss of generality that $\gamma \le \bar{f}_i \le \Gamma$ for all $1 \le i \le (1+\nu)(n+\rho n)$.

For $1 \le i \le (1+\rho)n$, let $I_i = \{(1+\rho)n+\nu(i-1)+1,\ldots,(1+\rho)n+\nu i\}$ and $s_i = (u_i, u_{I_i}) :- (u_i, u_{(1+\rho)n+\nu(i-1)+1},\ldots,u_{(1+\rho)n+\nu i})$. For any $g(f,\cdot)\in\mathcal{G}$ with corresponding $f\in\mathcal{F}_n$ and $\bar{f}\in\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}}$, define $\bar{g} = (\bar{g}_1,\ldots,\bar{g}_n)\in\mathbb{R}^{(1+\rho)n}$ by

$$\bar{g}_i = \log\frac{\bar{f}_i}{\bar{f}_i+\nu f_{\mathcal{R}}(u_i)} + \sum_{j\in I_i}\log\frac{\nu f_{\mathcal{R}}(u_j)}{\bar{f}_j+\nu f_{\mathcal{R}}(u_j)}, \quad 1\le i\le(1+\rho)n.$$

Let $\mathbf{G} : \mathbb{R}^{(1+\nu)(n+\rho n)} \to \mathbb{R}^{(1+\rho)n}$ be the function mapping $\bar{f}$ to $\bar{g}$ as above. The collection of vectors $\bar{g}$ can be written as

$$\mathcal{G}|^{(\epsilon)}_{s_{1:(1+\rho)n}} = \{\bar{g} = \mathbf{G}(\bar{f}) : \bar{f}\in\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}}\}.$$

We have $\big|\mathcal{G}|^{(\epsilon)}_{s_{1:(1+\rho)n}}\big| \le \big|\mathcal{F}_n|^{(\epsilon)}_{u_{1:(1+\nu)(n+\rho n)}}\big|$. Moreover, for all $1 \le i \le (1+\rho)n$,

$$\begin{aligned}
|g(f, s_i) - \bar{g}_i| &\le \Big| \log \frac{f(u_i)}{f(u_i) + \nu f_{\mathcal{R}}(u_i)} - \log \frac{\bar{f}_i}{\bar{f}_i + \nu f_{\mathcal{R}}(u_i)} \Big| + \nu \max_{j \in I_i} \Big| \log \frac{\bar{f}_j + \nu f_{\mathcal{R}}(u_j)}{f(u_j) + \nu f_{\mathcal{R}}(u_j)} \Big| \\
&\le \frac{\nu f_{\mathcal{R}}(u_i)}{f(u_i)\bar{f}_i} |f(u_i) - \bar{f}_i| + \nu \max_{j \in I_i} \frac{1}{\nu f_{\mathcal{R}}(u_j)} |f(u_j) - \bar{f}_j| \\
&\le \Big(\frac{\nu}{m\gamma} + \frac{1}{l} \Big) \epsilon.
\end{aligned}$$

So $\max_{1 \le i \le (1+\rho)n} |g(f, s_i) - \bar{g}_i| \le \kappa\epsilon$, where $\kappa = \nu/(m\gamma) + 1/l$. That is, the set $\mathcal{G}|^{(\epsilon)}_{s_{1:(1+\rho)n}} \subset \mathbb{R}^n$ is a $\kappa\epsilon$-cover of $\mathcal{G}|_{s_{1:(1+\rho)n}}$ with respect to the $\|\cdot\|_\infty$ norm. This implies $\mathcal{N}_\infty(\kappa\epsilon, \mathcal{G}, s_{1:(1+\rho)n}) \le |\mathcal{G}|^{(\epsilon)}_{s_{1:(1+\rho)n}}| \le \mathcal{N}_\infty(\epsilon, \mathcal{F}_n, (1+\nu)(n+\rho n))$. By taking all $s_{1:(1+\rho)n}$ (i.e. all $u_{1:(1+\nu)(n+\rho n)}$) into account, we have

$$\mathcal{N}_\infty(\kappa\epsilon, \mathcal{G}, (1+\rho)n) \le \mathcal{N}_\infty(\epsilon, \mathcal{F}_n, (1+\nu)(n+\rho n)),$$

and $\mathcal{N}_\infty(t/10, \mathcal{G}, (1+\rho)n) \le \mathcal{N}_\infty(t/(10\kappa), \mathcal{F}_n, (1+\nu)(n+\rho n))$ follows directly.

Given the tail probability bound

$$\begin{aligned}
&\Pr\Big(\frac{1}{(1+\rho)n} \sum_{i=1}^{(1+\rho)n} \{ g(\hat{f}_{\mathcal{M}}, S_i) - \mathbb{E}_{S'} g(\hat{f}_{\mathcal{M}}, S'_i) \} > t \Big) \\
&\le 4\mathbb{E}\mathcal{N}_1\Big(\frac{t(1+\rho)n}{10}, \mathcal{G}, S_{1:(1+\rho)n} \Big) \exp\Big(-\frac{3t(1+\rho)n}{160K} \Big) \\
&\le 4\mathcal{N}_\infty\Big(\frac{t}{10\kappa}, \mathcal{F}_n, (1+\nu)(n+\rho n) \Big) \exp\Big(-\frac{3t(1+\rho)n}{160K} \Big),
\end{aligned}$$

we can bound the expectation as below:

$$\begin{aligned}
&\mathbb{E}_S\Big[\frac{1}{(1+\rho)n} \sum_{i=1}^{(1+\rho)n} \{ g(\hat{f}_{\mathcal{M}}, S_i) - \mathbb{E}_{S'} g(\hat{f}_{\mathcal{M}}, S'_i) \} \Big] \\
&\le a_n + \int_{a_n}^{\infty} \Pr\Big(\frac{1}{(1+\rho)n} \sum_{i=1}^{(1+\rho)n} \{ g(\hat{f}_{\mathcal{M}}, S_i) - \mathbb{E}_{S'} g(\hat{f}_{\mathcal{M}}, S'_i) \} > t \Big) dt \\
&\le a_n + \int_{a_n}^{\infty} 4\mathcal{N}_\infty\Big(\frac{t}{10\kappa}, \mathcal{F}_n, (1+\nu)(n+\rho n) \Big) \exp\Big(-\frac{3t(1+\rho)n}{160K} \Big) dt \\
&\le a_n + \int_{a_n}^{\infty} 4\mathcal{N}_\infty\Big(\frac{a_n}{10\kappa}, \mathcal{F}_n, (1+\nu)(n+\rho n) \Big) \exp\Big(-\frac{3t(1+\rho)n}{160K} \Big) dt \\
&\le a_n + 4\mathcal{N}_\infty\Big(\frac{a_n}{10\kappa}, \mathcal{F}_n, (1+\nu)(n+\rho n) \Big) \exp\Big(-\frac{3a_n(1+\rho)n}{160K} \Big) \frac{160K}{3(1+\rho)n}.
\end{aligned}$$

Take $a_n = \log(4\mathcal{N}_\infty(K/\{\kappa(1+\rho)n\}, \mathcal{F}_n, (1+\nu)(n+\rho n))) \cdot 160K/\{3(1+\rho)n\}$. Note that $a_n/(10\kappa) \ge K/\{\kappa(1+\rho)n\}$, and $\mathcal{N}_\infty(K/\{\kappa(1+\rho)n\}, \mathcal{F}_n, (1+\nu)(n+\rho n)) \ge \mathcal{N}_\infty(a_n/(10\kappa), \mathcal{F}_n,$

$(1+\nu)(n+\rho n))$. Then,

$$\mathbb{E}_S\Big[\frac{1}{(1+\rho)n}\sum_{i=1}^{(1+\rho)n}\{g(\hat{f}_{\mathcal{M}},S_i)-\mathbb{E}_{S'}g(\hat{f}_{\mathcal{M}},S_i')\}\Big]$$

$$\leq \frac{160K}{3(1+\rho)}\cdot\frac{\log\big(4\mathcal{N}_\infty\big(\frac{K}{\kappa(1+\rho)n},\mathcal{F}_n,(1+\nu)(n+\rho n)\big)\big)+1}{n}.$$

In view of (16), we have

$$\mathbb{E}_S[R(\hat{f}_{\mathcal{M}})-\hat{R}(\hat{f}_{\mathcal{M}})]\leq C_1K\frac{\log\big(4\mathcal{N}_\infty\big(\frac{K}{\kappa(1+\rho)n},\mathcal{F}_n,(1+\nu)(n+\rho n)\big)\big)}{(1+\rho)n}, \tag{17}$$

for some universal constant C_1.

Our final step is to bound the uniform covering number. It follows from Theorem 12.2 in Anthony and Bartlett (1999) that, for $n\geq \mathrm{Pdim}(\mathcal{F}_n)/\{(1+\nu)(1+\rho)\}$,

$$\mathcal{N}_\infty\Big(\frac{K}{\kappa(1+\rho)n},\mathcal{F}_n,(1+\nu)(n+\rho n)\Big)\leq\Big(\frac{e(1+\nu)(1+\rho)^2n^2\kappa\Gamma}{K\mathrm{Pdim}(\mathcal{F}_n)}\Big)^{\mathrm{Pdim}(\mathcal{F}_n)}.$$

Furthermore, by Theorem 7 in Bartlett et al. (2019), the pseudo-dimension of the class of networks is bounded by its architecture parameters:

$$\mathrm{Pdim}(\mathcal{F}_n)\leq C'\mathcal{S}\mathcal{D}\log\mathcal{S},$$

where C' is a universal constant. Combining the upper bounds of uniform covering number and pseudo-dimension in (17), we have

$$\mathbb{E}_S[R(\hat{f}_{\mathcal{M}})-\hat{R}(\hat{f}_{\mathcal{M}})]\leq c_3\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n},$$

where $c_3=C_2K(1+\rho)^{-1}\log((1+\nu)(1+\rho)^2\kappa\Gamma/K)$, for some universal constant C_2. This concludes the proof of Lemma 7. ■

A.5 Approximation error bound

Lemma 19 (Approximation error) *Suppose that Assumptions 3 and 8 hold, and $\gamma\leq\rho l/(1+\rho)$, $\Gamma\geq L_0\vee L$ where l, L are positive constants such that $l\leq f_{\mathcal{R}}\leq L$. Then for any $P,Q\in\mathbb{Z}_+$, there exists a function $f_*\in\mathcal{F}_n=\mathcal{F}(\mathcal{D},\mathcal{W},\mathcal{S},\gamma,\Gamma)$ with width $\mathcal{W}=38(\lfloor\beta\rfloor+1)^2d^{\lfloor\beta\rfloor+1}P\lceil\log_2(8P)\rceil$ and depth $\mathcal{D}=21(\lfloor\beta\rfloor+1)^2Q\lceil\log_2(8Q)\rceil$ such that*

$$\|f_*-f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}\leq 324B_0^2(\lfloor\beta\rfloor+1)^4d^{2\lfloor\beta\rfloor+\beta\vee1}(PQ)^{-4\beta/d}.$$

Proof For $J\in\mathbb{Z}_+$ and $\delta\in(0,1/J)$, define a region $\Omega([0,1]^d,J,\delta)\subset[0,1]^d$ as

$$\Omega([0,1]^d,J,\delta)=\bigcup_{i=1}^{d}\Big\{x=[x_1,\dots,x_d]^\top:x_i\in\bigcup_{j=1}^{J-1}\Big(\frac{j}{J}-\delta,\frac{j}{J}\Big)\Big\}.$$

By Theorem 3.3 in Jiao et al. (2023), there exists a function $f_* \in \mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ with width $\mathcal{W} = 38(s+1)^2 d^{s+1} P \lceil \log_2(8P) \rceil$ and depth $\mathcal{D} = 21(s+1)^2 Q \lceil \log_2(8Q) \rceil$, such that

$$|f_*(x) - f_{\mathcal{M}}(x)| \leq 18 B_0 (s+1)^2 d^{s+(\beta \vee 1)/2} (PQ)^{-2\beta/d},$$

for any $x \in [0,1]^d \setminus \Omega([0,1]^d, J, \delta)$ where $J = \lceil (PQ)^{2/d} \rceil$ and δ is an arbitrary number in $(0, 1/(3J)]$. Note that the Lebesgue measure of $\Omega([0,1]^d, J, \delta)$ is no more than $dJ\delta$, which can be arbitrarily small if δ is arbitrarily small. By absolute continuity of the distribution corresponding to $f_{\mathcal{M}}$ with respect to Lebesgue measure, we have

$$\|f_* - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq 18^2 B_0^2 (s+1)^4 d^{2s+\beta \vee 1} (PQ)^{-4\beta/d}.$$

Thus we complete the proof of Lemma 19. ■

A.6 Proof of Theorem 10

Proof Due to Lemma 6, the expected L_2-risk for estimating $f_{\mathcal{M}}$ can be decomposed into the sum of stochastic error and approximation error:

$$\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq \frac{1}{c_1} \mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] + \frac{c_2}{c_1} \inf_{f \in \mathcal{F}_n} \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}. \tag{18}$$

By Lemma 7, we bound the stochastic error by

$$\mathbb{E}_S[R(\hat{f}_{\mathcal{M}}) - \hat{R}(\hat{f}_{\mathcal{M}})] \leq c_3 \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n}. \tag{19}$$

By Lemma 19 in Appendix A.5, we bound the approximation error by

$$\inf_{f \in \mathcal{F}_n} \|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq 324 B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d}. \tag{20}$$

Combining (18), (19), and (20), we have

$$\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq c_4' \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_5' B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d}. \tag{21}$$

Since c_1, c_2 and c_3 only depend on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$, it is clear that c_4', c_5' are constants only depending on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$. Finally, Lemma 4 implies that

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|^2_{L_2(f_0)} \leq c_4 \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_5 B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d},$$

where $c_4 = (1+\rho)^3 c_4'$ and $c_5 = (1+\rho)^3 c_5'$. Setting $P = 1$ and $Q = \lceil n^{d/\{2(d+2\beta)\}} \rceil$, we obtain the error bound (7) in Theorem 10. This concludes the proof of Theorem 10. ■

A.7 Proof of Theorem 13

Proof Suppose that $f_{\mathcal{R}} \geq l$, then $f_{\mathcal{M}} \geq \rho l/(1+\rho)$. It follows from the definition of total variation distance that

$$\begin{aligned}
\mathbb{E}_S \mathrm{TV}(\hat{f}_{\mathcal{M}}, f_{\mathcal{M}}) &= \frac{1}{2}\mathbb{E}_S \int |\hat{f}_{\mathcal{M}}(y) - f_{\mathcal{M}}(y)| dy \\
&\leq \frac{1}{2}\mathbb{E}_S \int |\hat{f}_{\mathcal{M}}(y) - f_{\mathcal{M}}(y)| f_{\mathcal{M}}(x) \frac{1+\rho}{\rho l} dy \\
&= \frac{1+\rho}{2\rho l}\mathbb{E}_S[\mathbb{E}_{Y\sim f_{\mathcal{M}}} |\hat{f}_{\mathcal{M}}(Y) - f_{\mathcal{M}}(Y)|] \\
&\leq \frac{1+\rho}{2\rho l}\big\{\mathbb{E}_S[\mathbb{E}_{Y\sim f_{\mathcal{M}}} |\hat{f}_{\mathcal{M}}(Y) - f_{\mathcal{M}}(Y)|^2]\big\}^{1/2} \\
&= \frac{1+\rho}{2\rho l}\big\{\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})}\big\}^{1/2},
\end{aligned}$$

where the penultimate line follows from the Jensen's inequality. By inequality (21) in Appendix A.6,

$$\mathbb{E}_S \|\hat{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}})} \leq c_4' \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_5' B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d},$$

where c_4' and c_5' are constants only depending on $\nu, \rho, L_0, \gamma, \Gamma$ and, $f_{\mathcal{R}}$. Therefore,

$$\begin{aligned}
\mathbb{E}_S \mathrm{TV}(\hat{f}_{\mathcal{M}}, f_{\mathcal{M}}) &\leq \frac{1+\rho}{2\rho l}\big\{\mathbb{E}_S \|\hat{f}_n - f_0\|^2_{L_2(f_0)}\big\}^{1/2} \\
&\leq \frac{1+\rho}{2\rho l}\Big\{c_4 \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_5 B_0^2 (\lfloor \beta \rfloor + 1)^4 d^{2\lfloor \beta \rfloor + \beta \vee 1} (PQ)^{-4\beta/d}\Big\}^{1/2} \\
&\leq \frac{(1+\rho)\sqrt{c_4'}}{2\rho l}\Big(\mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n}\Big)^{1/2} \\
&\qquad + \frac{(1+\rho)\sqrt{c_5'}}{2\rho l} B_0 (\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + (\beta \vee 1)/2} (PQ)^{-2\beta/d}.
\end{aligned}$$

Note that $\mathrm{TV}(\tilde{f}_n, f_0) = (1+\rho)\mathrm{TV}(\hat{f}_{\mathcal{M}}, f_{\mathcal{M}})$. Letting $c_7 = (1+\rho)^2\sqrt{c_4'}/(2\rho l)$, $c_8 = (1+\rho)^2\sqrt{c_5'}/(2\rho l)$, we have

$$\mathbb{E}_S \mathrm{TV}(\tilde{f}_n, f_0) \leq c_7 \Big(\mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n}\Big)^{1/2} + c_8 B_0 (\lfloor \beta \rfloor + 1)^2 d^{\lfloor \beta \rfloor + (\beta \vee 1)/2} (PQ)^{-2\beta/d},$$

and c_7, c_8 only depend on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$. Setting $P = 1$ and $Q = \lceil n^{d/\{2(d+2\beta)\}} \rceil$, we obtain the error bound (8) in Theorem 13. This concludes the proof of Theorem 13. ■

A.8 Error analysis for the unbounded-support case

In this subsection, we present an error analysis of our estimator for estimating densities with unbounded support, with a focus on the differences between the theoretical analysis in this setting and the bounded case. First, a positive and continuous reference density supported on a compact domain is lower bounded, while this property does not hold for unbounded domains. Second, conventional neural network approximation theory for approximating function with compact domain and is not directly applicable to the unbounded support case. To address the complications associated with unbounded support, we will leverage the truncation technique and assume that the target distribution has a sub-exponentially decaying tail.

In the following, we denote the target distribution (probability measure) as μ_0; that is, $\mu_0(dx) = f_0(x)dx$. We make the following assumptions on the target density f_0 and the reference density $f_{\mathcal{R}}$:

Assumption 20 *The target distribution μ_0 is sub-exponential in the sense that*

$$\mu_0(\mathbb{R}^d \setminus [-x,x]^d) \leq Cd\exp(-ax),\ \forall\ x > 0, \tag{22}$$

for some positive constants a and C. Furthermore, there exist positive constants L_0 and $L_{\mathcal{R}}$ such that $f_0 \leq L_0$ and $0 < f_{\mathcal{R}} \leq L_{\mathcal{R}}$ on $\mathbb{R}^d$. In addition, there exists a positive number $l_{\mathcal{R}}$ that depends on t and d, such that $f_{\mathcal{R}} \geq l_{\mathcal{R}}$ on Ω_t.

Assumption 21 (Hölder smoothness) *The densities $f_0, f_{\mathcal{R}} \in \mathcal{H}^{\beta}(\mathbb{R}^d, B_0)$ for a given $\beta > 0$ and some finite constant $B_0 > 0$.*

In Assumption 20, we require the reference distribution to satisfy $f_{\mathcal{R}} \geq l_{\mathcal{R}}$ on Ω_t. Since the integral of $f_{\mathcal{R}}$ over Ω_t is upper bounded by 1, it follows that $l_{\mathcal{R}} \leq (2t)^{-d}$. Therefore, $l_{\mathcal{R}}$ depends on both t and d. We also select the lower bound of the neural network, γ, to be a function of t and d. In our theoretical results, we establish the dependence of the error bounds on $l_{\mathcal{R}}$ and γ explicitly.

When the truncation level t is a function of the sample size n, it follows that the reference distribution also depends on n. Assumption 20 is satisfied when the reference distribution is $N_d(\mathbf{0}, \sigma^2\mathbf{I})$, Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$, where σ is a function of n. Further details will be provided subsequently, following the proof of Theorem 16.

Next, we decompose the expected L_2-risk of $\bar{f}_n$ into the error on the truncated domain and the error due to truncation. Before proceeding, similar to the definition of $\|f - f_*\|^2_{L_2(f_*)}$, we let

$$\|g\|^2_{L_2(f_*;A)} := \mathbb{E}_{X\sim f_*}\left[|g(X)|^2 \mathbb{1}_A(X)\right] = \int_A |g(x)|^2 f_*(x)dx$$

for any probability density f_* on $\mathbb{R}^d$, function $g : \mathbb{R}^d \to \mathbb{R}$, and subset $A \subseteq \mathbb{R}^d$.

Lemma 22 (Error decomposition) *Given any $t \geq 1/2$, suppose that Assumption 20 holds, and $\gamma \leq 1$. Then, the expected L_2-risk of the data-augmented nonparametric noise contrastive estimator $\bar{f}_n$ defined in (9) satisfies*

$$\begin{aligned}\mathbb{E}_S\|\bar{f}_n - f_0\|^2_{L_2(f_0)} \leq c_{13}\gamma^{-1}l_{\mathcal{R}}^{-2}\mathbb{E}_S[R_t(\check{f}_{\mathcal{M}}) - \hat{R}_t(\check{f}_{\mathcal{M}})] + c_{14}\gamma^{-3}l_{\mathcal{R}}^{-2}\inf_{f\in\mathcal{F}_n}\|f - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)} \\ + c_{15}\mathbb{E}_S\|\check{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^{\complement})},\end{aligned} \tag{23}$$

where c_{13} and c_{14} are constants only depending on ν, ρ, L_0, Γ, and $L_{\mathcal{R}}$, and c_{15} is a constant only depending on ρ.

Proof According to the proof of Lemma 4, we have

$$\|\bar{f}_n - f_0\|^2_{L_2(f_0)} = (1+\rho)^2 \int_{\mathbb{R}^d} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x) dx \tag{24}$$

given any training sample S. Furthermore, it follows from $f_0 = (1+\rho) f_{\mathcal{M}} - \rho f_{\mathcal{R}} \leq (1+\rho) f_{\mathcal{M}}$ that

$$\begin{aligned} &\int_{\mathbb{R}^d} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x) dx \\ &= \int_{\Omega_t} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x) dx \\ &\qquad + \int_{\Omega_t^{\complement}} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x) dx \\ &\leq (1+\rho) \int_{\Omega_t} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_{\mathcal{M}}(x) dx \\ &\qquad + \int_{\Omega_t^{\complement}} |\check{f}_{\mathcal{M}}(x) - f_{\mathcal{M}}(x)|^2 f_0(x) dx \\ &= (1+\rho) \|\check{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)} + \|\check{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^{\complement})}. \end{aligned} \tag{25}$$

In the following, we intend to show a calibration condition in terms of $\|f - f_*\|^2_{L_2(f_*,\Omega_t)}$ and $R_t(f) - R_t(f_*)$, an inequality similar to (5) in Lemma 5. Suppose that $f_* \leq L_*$ and $\gamma \leq f \leq \Gamma$ for all $f \in \mathcal{F}$ on $\mathbb{R}^d$, and there exist positive constants m and M that may depend on t such that

$$\begin{gathered} m\gamma f_{\mathcal{R}} \leq f_* \leq \frac{M}{l_{\mathcal{R}}} f_{\mathcal{R}}, \\ m\gamma f_{\mathcal{R}} \leq f \leq \frac{M}{l_{\mathcal{R}}} f_{\mathcal{R}},\ \forall f \in \mathcal{F} \end{gathered} \tag{26}$$

on Ω_t. Then, according to the proof of Lemma 5, we have

$$c_0 \gamma l^2_{\mathcal{R}} \|f - f_*\|^2_{L_2(f_*;\Omega_t)} \leq R_t(f) - R_t(f_*) \leq \frac{1}{2\gamma^2} \|f - f_*\|^2_{L_2(f_*;\Omega_t)}, \tag{27}$$

where $c_0 = m\nu / \{2M(M+\nu)(L_* \vee \Gamma)^2\}$. To show this, we only need to replace the domain with Ω_t; specifically, we only consider the values of functions f, f_*, and $f_{\mathcal{R}}$ and the integrals on Ω_t.

Under Assumption 20, there exist positive constants L_0 and $L_{\mathcal{R}}$, as well as a positive number $l_{\mathcal{R}}$, such that $f_0 \leq L_0$ and $l_{\mathcal{R}} \leq f_{\mathcal{R}} \leq L_{\mathcal{R}}$ on Ω_t. Thus, $f_{\mathcal{M}} \leq L_0 \vee L_{\mathcal{R}}$. By the definition of $\mathcal{F}_n$, $\gamma \leq f \leq \Gamma$ for all $f \in \mathcal{F}_n$. Therefore, (26) holds for $f_* = f_{\mathcal{M}}$ and $\mathcal{F} = \mathcal{F}_n$ by setting $m = (1/L_{\mathcal{R}}) \wedge \{\rho/(1+\rho)\}$ and $M = \max\{L_0, L_{\mathcal{R}}, \Gamma\}$. It follows that (27) holds for $f_* = f_{\mathcal{M}}$ and $f \in \mathcal{F}_n$, with constant c_0 depending solely on ν, ρ, L_0, Γ, and $L_{\mathcal{R}}$.

It can be shown along similar lines as in the proof of Lemma 6 that

$$\mathbb{E}_S \|\check{f}_{\mathcal{M}} - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)} \leq \frac{1}{c_0 \gamma l^2_{\mathcal{R}}} \mathbb{E}_S[R_t(\check{f}_{\mathcal{M}}) - \hat{R}_t(\check{f}_{\mathcal{M}})] + \frac{1}{2c_0 \gamma^3 l^2_{\mathcal{R}}} \inf_{f \in \mathcal{F}_n} \|f - f_{\mathcal{M}}\|_{L_2(f_{\mathcal{M}};\Omega_t)}. \tag{28}$$

Taking expectations on both sides of (24) and (25), and combining with (28), we obtain the desired inequality. ■

The three terms on the right-hand side of (23) are the stochastic error, the approximation error and the truncation error, respectively. In the following, we establish upper bounds for these error terms in Lemmas 23, 25, and 26 respectively.

Lemma 23 (Stochastic error) *Let $\mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ be the class of feedforward neural networks activated by continuous piecewise linear activation function with finitely many inflection points, and let $\gamma \leq 1$. Given any $t \geq 1/2$, let $\check{f}_{\mathcal{M}} \in \arg\min_{f \in \mathcal{F}_n} \hat{R}_t(f)$ be a minimizer of the truncated empirical risk over $\mathcal{F}_n$, and suppose that Assumption 20 holds. Then, for $n \geq Pdim(\mathcal{F}_n)/\{(1+\nu)(1+\rho)\}$,*

$$\mathbb{E}_S[R_t(\check{f}_{\mathcal{M}}) - \hat{R}_t(\check{f}_{\mathcal{M}})] \leq c_1' \gamma^{-2} l_{\mathcal{R}}^{-1} \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n},$$

where c_1' is a constant depending only on ν, ρ, L_0, Γ, and $L_{\mathcal{R}}$.

Proof According to the proof of Lemma 22, we have

$$m\gamma f_{\mathcal{R}} \leq f_{\mathcal{M}} \leq \frac{M}{l_{\mathcal{R}}} f_{\mathcal{R}},$$
$$m\gamma f_{\mathcal{R}} \leq f \leq \frac{M}{l_{\mathcal{R}}} f_{\mathcal{R}}, \ \forall f \in \mathcal{F}_n$$

on Ω_t, where $m = (1/L_{\mathcal{R}}) \wedge \{\rho/(1+\rho)\}$ and $M = \max\{L_0, L_{\mathcal{R}}, \Gamma\}$. We can show along similar lines as in the proof of Lemma 7 that

$$\mathbb{E}_S[R_t(\check{f}_{\mathcal{M}}) - \hat{R}_t(\check{f}_{\mathcal{M}})] \leq c_1' \gamma^{-2} l_{\mathcal{R}}^{-1} \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n},$$

where $c_1' = CK(1+\rho)^{-1} \log(e(1+\nu)(1+\rho)^2 \kappa\Gamma/K)$, $K = \log(1+\nu/m) + \nu \log(1+M/\nu)$, $\kappa = 1 + \nu/m$, and C is a universal constant. This completes the proof of Lemma 23. ■

Lemma 24 (Clipping functions) *Given $K > 0$, we define $\eta_K : \mathbb{R} \to [-K, K]$ by*

$$\eta_K(z) = \begin{cases} -K, & z \in (-\infty, -K), \\ z, & z \in [-K, K], \\ K, & z \in (K, +\infty). \end{cases}$$

There exists a clipping function $\mathcal{C}_K : \mathbb{R}^d \to [-K, K]^d$ at level K implemented by a ReLU neural network with depth 1 and width $2d$ such that for any $x = [x_1, x_2, \ldots, x_d]^\top \in \mathbb{R}^d$,

$$\mathcal{C}_K(x) = [\eta_K(x_1), \eta_K(x_2), \ldots, \eta_K(x_d)]^\top.$$

Proof It is clear that $\mathcal{C}_K(x) = \boldsymbol{\sigma}(x + K\mathbf{1}_d) - \boldsymbol{\sigma}(x - K\mathbf{1}_d) - K\mathbf{1}_d$, where $\mathbf{1}_d$ denotes the d-dimensional vector with all elements equal to 1. This expression implies that the clipping function $\mathcal{C}_K$ can be implemented by a ReLU neural network with depth 1 and width $2d$. Thus, we conclude the proof of Lemma 24. ■

Lemma 25 (Approximation error) *Given any $t > 0$, suppose that Assumptions 20 and 21 hold, $\gamma \leq \rho l_{\mathcal{R}}/(1+\rho)$, and $\Gamma \geq L_0 \vee L_{\mathcal{R}}$. Then for any $P, Q \in \mathbb{Z}_+$, there exists a function $f_* \in \mathcal{F}_n = \mathcal{F}(\mathcal{D}, \mathcal{W}, \mathcal{S}, \gamma, \Gamma)$ with width $\mathcal{W} = 38(\lfloor\beta\rfloor + 1)^2 d^{\lfloor\beta\rfloor+1} P\lceil\log_2(8P)\rceil$ and depth $\mathcal{D} = 21(\lfloor\beta\rfloor+1)^2 Q\lceil\log_2(8Q)\rceil + 1$ such that*

$$\|f_* - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)} \leq 324(2t)^{2\beta} B_0^2 (s+1)^4 d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}.$$

Proof We construct an approximation f_* of the mixture density $f_{\mathcal{M}}$ on the truncated domain Ω_t, and bound the $L_2(f_{\mathcal{M}};\Omega_t)$ approximation error.

First of all, we use the clipping function $\mathcal{C}_K$ defined in Lemma 24 (set $K = t$ therein) to clip the unbounded domain. That is, for any $x \in \mathbb{R}^d$, we have $\mathcal{C}_t(x) \in \Omega_t$. We only need to consider the approximation of $f_{\mathcal{M}}$ on the truncated domain Ω_t.

Then, we employ the mapping $\tilde{x} = \mathcal{T}(x) = (x + t\mathbf{1}_d)/(2t)$ to transform the domain Ω_t into the domain $[0,1]^d$. Notice that the mapping $\mathcal{T}$ is linear and invertible. We denote its inverse as $x = \mathcal{T}^{-1}(\tilde{x})$. We further define a new target function $f^\diamond$ by $f^\diamond(\tilde{x}) = f_{\mathcal{M}}(\mathcal{T}^{-1}(\tilde{x}))$ for any $\tilde{x} \in \mathbb{R}^d$. Clearly, $f_{\mathcal{M}}(x) = f^\diamond(\mathcal{T}(x))$ for any $x \in \mathbb{R}^d$. Under Assumption 21, the new function $f^\diamond$ belongs to the Hölder class $\mathcal{H}^\beta(\mathbb{R}^d, (2t)^\beta B_0)$.

For $J \in \mathbb{Z}_+$ and $\delta \in (0, 1/J)$, we define $\Omega([0,1]^d, J, \delta) \subset [0,1]^d$ as the same region presented in the proof of Lemma 19. By Theorem 3.3 in Jiao et al. (2023), there exists a function $f^\dagger$ implemented by a deep ReLU neural network with width $38(s+1)^2 d^{s+1} P\lceil\log_2(8P)\rceil$ and depth $21(s+1)^2 Q\lceil\log_2(8Q)\rceil$ such that

$$|f^\dagger(\tilde{x}) - f^\diamond(\tilde{x})| \leq 18(2t)^\beta B_0 (s+1)^2 d^{s+(\beta\vee 1)/2}(PQ)^{-2\beta/d},$$

for any $\tilde{x} \in [0,1]^d \setminus \Omega([0,1]^d, J, \delta)$ where $J = \lceil (PQ)^{2/d}\rceil$ and δ is an arbitrary number in $(0, 1/(3J)]$. Note that the Lebesgue measure of $\Omega([0,1]^d, J, \delta)$ is no more than $dJ\delta$, which can be arbitrarily small if δ is arbitrarily small. Moreover, $(2t)^d f^\diamond$ is a probability density function on $\mathbb{R}^d$. By the absolute continuity of the distribution corresponding to $(2t)^d f^\diamond$ with respect to Lebesgue measure, we have

$$\|f^\dagger - f^\diamond\|^2_{L_2((2t)^d f^\diamond;[0,1]^d)} \leq 18^2 (2t)^{2\beta} B_0^2 (s+1)^4 d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}. \tag{29}$$

Let $f_*(x) = f^\dagger(\mathcal{T} \circ \mathcal{C}_t(x))$. We claim that f_* is a good approximation of the original target function $f_{\mathcal{M}}$ on the truncated domain Ω_t, and can be implemented by a deep ReLU neural network. According to the error bound (29), the approximation error of f_* on Ω_t is

$$\begin{aligned}
\|f_* - f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)} &= \int_{\Omega_t} |f_*(x) - f_{\mathcal{M}}(x)|^2 f_{\mathcal{M}}(x)dx \\
&= \int_{\Omega_t} |f^\dagger(\mathcal{T} \circ \mathcal{C}_t(x)) - f^\diamond(\mathcal{T}(x))|^2 f^\diamond(\mathcal{T}(x))dx \\
&= \int_{\Omega_t} |f^\dagger(\mathcal{T}(x)) - f^\diamond(\mathcal{T}(x))|^2 f^\diamond(\mathcal{T}(x))dx \\
&= \int_{[0,1]^d} |f^\dagger(\tilde{x}) - f^\diamond(\tilde{x})|^2 (2t)^d f^\diamond(\tilde{x})d\tilde{x} \\
&= \|f^\dagger - f^\diamond\|^2_{L_2((2t)^d f^\diamond;[0,1]^d)} \\
&\leq 324(2t)^{2\beta} B_0^2 (s+1)^4 d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}.
\end{aligned}$$

Finally, it remains to calculate the complexity of the ReLU neural network implementing f_*. By Lemma 24, $\mathcal{C}_t$ can be implemented by a ReLU neural network with depth 1 and width $2d$. Notice that $\mathcal{T}$ is linear and the ReLU neural network implementing $f^\dagger$ has width $38(s+1)^2 d^{s+1} P\lceil\log_2(8P)\rceil$ and depth $21(s+1)^2 Q\lceil\log_2(8Q)\rceil$. It follows that f_* can be implemented by a ReLU neural network with width $38(s+1)^2 d^{s+1} P\lceil\log_2(8P)\rceil$ and depth $21(s+1)^2 Q\lceil\log_2(8Q)\rceil+1$. This concludes the proof of Lemma 25. ■

Lemma 26 (Truncation error) *Given any $t>0$, suppose that Assumption 20 holds. Then, the truncation error satisfies*

$$\mathbb{E}_S\|\check{f}_{\mathcal{M}}-f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^{\complement})}\le c_2' C d\exp(-at),$$

where c_2' is a constant only depending on L_0, Γ, and $L_{\mathcal{R}}$.

Proof Given any training sample S,

$$\begin{aligned}\|\check{f}_{\mathcal{M}}-f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^{\complement})}&=\int_{\Omega_t^{\complement}}|\check{f}_{\mathcal{M}}(x)-f_{\mathcal{M}}(x)|^2 f_0(x)dx\\&\le(\Gamma+L_0\vee L_{\mathcal{R}})^2\int_{\Omega_t^{\complement}}f_0(x)dx\\&=(\Gamma+L_0\vee L_{\mathcal{R}})^2\mu_0\big(\Omega_t^{\complement}\big)\\&\le(\Gamma+L_0\vee L_{\mathcal{R}})^2 C d\exp(-at),\end{aligned}$$

where the last inequality follows from (22) in Assumption 20. This concludes the proof of Lemma 26. ■

According to Lemmas 23, 25, and 26, there exists a trade-off between these error terms. As the truncation level t rises, the approximation error increases, whereas the truncation error decreases. Moreover, the stochastic error bound depends on t only through $l_{\mathcal{R}}$ and γ. With a proper selection of t, we can establish the non-asymptotic error bounds in Theorem 16.

Proof of Theorem 16 By Lemma 22, the expected L_2-risk of $\bar{f}_n$ can be decomposed into the sum of stochastic error, approximation error, and truncation error:

$$\begin{aligned}\mathbb{E}_S\|\bar{f}_n-f_0\|^2_{L_2(f_0)}\le{}& c_{13}\gamma^{-1}l_{\mathcal{R}}^{-2}\mathbb{E}_S[R_t(\check{f}_{\mathcal{M}})-\hat{R}_t(\check{f}_{\mathcal{M}})]+c_{14}\gamma^{-3}l_{\mathcal{R}}^{-2}\inf_{f\in\mathcal{F}_n}\|f-f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)}\\&+c_{15}\mathbb{E}_S\|\check{f}_{\mathcal{M}}-f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^{\complement})}.\end{aligned}\tag{30}$$

By Lemma 23, we bound the stochastic error by

$$\mathbb{E}_S[R_t(\check{f}_{\mathcal{M}})-\hat{R}_t(\check{f}_{\mathcal{M}})]\le c_1'\gamma^{-2}l_{\mathcal{R}}^{-1}\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n}.\tag{31}$$

By Lemma 25, we bound the approximation error by

$$\inf_{f\in\mathcal{F}_n}\|f-f_{\mathcal{M}}\|^2_{L_2(f_{\mathcal{M}};\Omega_t)}\le 324(2t)^{2\beta}B_0^2(s+1)^4d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}. \tag{32}$$

By Lemma 26, we bound the truncation error by

$$\mathbb{E}_S\|\check{f}_{\mathcal{M}}-f_{\mathcal{M}}\|^2_{L_2(f_0;\Omega_t^\complement)}\le c_2'Cd\exp(-at)=c_2'Cdn^{-1}. \tag{33}$$

Combining (30), (31), (32) and (33), we have

$$\begin{aligned}\mathbb{E}_S\|\bar{f}_n-f_0\|^2_{L_2(f_0)}&\le c_{16}\gamma^{-3}l_{\mathcal{R}}^{-3}\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n}+c_{17}\gamma^{-3}l_{\mathcal{R}}^{-2}(2t)^{2\beta}(s+1)^4d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}\\&\quad+c_{18}dn^{-1}\\&=c_{16}\gamma^{-3}l_{\mathcal{R}}^{-3}\mathcal{S}\mathcal{D}\log\mathcal{S}\frac{\log n}{n}\\&\quad+c_{17}\gamma^{-3}l_{\mathcal{R}}^{-2}(2/a)^{2\beta}(s+1)^4d^{2s+\beta\vee 1}(PQ)^{-4\beta/d}(\log n)^{2\beta}\\&\quad+c_{18}dn^{-1}.\end{aligned}$$

Notice that c_{13}, c_{14}, and c_1' are constants only depending on ν, ρ, L_0, Γ, and $L_{\mathcal{R}}$, c_{15} only depends on ρ, and c_2' only depends on L_0, Γ, and $L_{\mathcal{R}}$. Then, it is clear that

- c_{16} is a constant only depending on ν, ρ, L_0, Γ, and $L_{\mathcal{R}}$;
- c_{17} is a constant only depending on ν, ρ, L_0, Γ, $L_{\mathcal{R}}$, and B_0;
- c_{18} is a constant only depending on ρ, C, L_0, Γ, $L_{\mathcal{R}}$.

Setting $P=1$ and $Q=\lceil n^{d/\{2(d+2\beta)\}}\rceil$, we obtain the error bound (10) in Theorem 16. This concludes the proof of Theorem 16. ■

Lastly, we demonstrate that Assumption 20 is satisfied when the reference distribution is $N_d(\mathbf{0},\sigma^2\mathbf{I})$, based on which we can derive a refined error bound. Specifically, the choice $\sigma=a^{-1}\log n$ leads to a lower bound $f_{\mathcal{R}}\ge l_{\mathcal{R}}:-(2\pi e)^{-d/2}(a^{-1}\log n)^{-d}$ on the truncated domain Ω_t. Then, by setting $\gamma=\rho l_{\mathcal{R}}/(1+\rho)$, we can establish the following error bound:

$$\begin{aligned}&\mathbb{E}_S\|\bar{f}_n-f_0\|^2_{L_2(f_0)}\\&\le c_{20}(\pi e/2)^{3d}\{(2/a)\vee 1\}^{2\beta+6d}(\lfloor\beta\rfloor+1)^9d^{2\lfloor\beta\rfloor+3(\beta\vee 1)}n^{-\frac{2\beta}{d+2\beta}}(\log_2 n)^{2(\beta\vee 2)+6d},\end{aligned}$$

where c_{20} is a constant only depending on ν, ρ, C, L_0, Γ, $L_{\mathcal{R}}$, and B_0.

A.9 Proof of Theorem 18

Proof By Assumption 17 and $f_{\mathcal{R}}\in\mathcal{CS}(q,\mathbf{l},\mathbf{d},\boldsymbol{\beta},B_0)$,

$$\begin{aligned}f_0&=g^{(q)}\circ g^{(q-1)}\circ\cdots\circ g^{(1)},\\f_{\mathcal{R}}&=h^{(q)}\circ h^{(q-1)}\circ\cdots\circ h^{(1)},\end{aligned}$$

with $g^{(i)}$ and $h^{(i)}$ $(i = 1, \dots, q)$ defined by

$$g^{(i)}(x) = (g_1^{(i)}(W_1^{(i)}x), \cdots, g_{l_i}^{(i)}(W_{l_i}^{(i)}x))^\top,$$
$$h^{(i)}(x) = (h_1^{(i)}(V_1^{(i)}x), \cdots, h_{l_i}^{(i)}(V_{l_i}^{(i)}x))^\top.$$

We can show that $f_{\mathcal{M}}$ is also a composition of functions by placing the compositional structures $g^{(q)} \circ g^{(q-1)} \circ \cdots \circ g^{(1)}$ and $h^{(q)} \circ h^{(q-1)} \circ \cdots \circ h^{(1)}$ in parallel. Specifically, the functions $g^{(i)}$ and $h^{(i)}$ at each level of composition are concatenated as follows:

$$\bar{g}^{(i)}(x) = (g_1^{(i)}(\bar{W}_1^{(i)}x), \cdots, g_{l_i}^{(i)}(\bar{W}_{l_i}^{(i)}x), h_1^{(i)}(\bar{V}_1^{(i)}x), \cdots, h_{l_i}^{(i)}(\bar{V}_{l_i}^{(i)}x))^\top, \tag{34}$$

where $\bar{g}^{(1)} : \mathbb{R}^d \to \mathbb{R}^{2l_1}$, $\bar{W}_j^{(1)} = W_j^{(1)}$, $\bar{V}_j^{(1)} = V_j^{(1)}$ $(1 \le j \le l_1)$, $\bar{g}^{(i)} : \mathbb{R}^{2l_{i-1}} \to \mathbb{R}^{2l_i}$ $(1 < i \le q)$, and

$$\bar{W}_j^{(i)} = \begin{bmatrix} W_j^{(i)} & \mathbf{0}_{d_i \times l_{i-1}} \end{bmatrix}, \quad \bar{V}_j^{(i)} = \begin{bmatrix} \mathbf{0}_{d_i \times l_{i-1}} & V_j^{(i)} \end{bmatrix}, \ 1 < i \le q, \ 1 \le j \le l_i.$$

Then, we have

$$\bar{g}^{(q)} \circ \bar{g}^{(q-1)} \circ \cdots \circ \bar{g}^{(1)} = (f_0, f_{\mathcal{R}})^\top.$$

In addition, we define $\bar{g}^{(q+1)}(x) = g_1^{(q+1)}(\bar{W}_1^{(q+1)}x)$, with $g_1^{(q+1)}(x) = (L_0 \vee L)x$ and $\bar{W}_1^{(q+1)} = (L_0 \vee L)^{-1}(1/(1+\rho), \rho/(1+\rho))$. As a result,

$$\bar{g}^{(q+1)} \circ \bar{g}^{(q)} \circ \cdots \circ \bar{g}^{(1)} = \frac{1}{1+\rho} f_0 + \frac{\rho}{1+\rho} f_{\mathcal{R}} = f_{\mathcal{M}}.$$

By (34), $\bar{g}^{(i)}(x) = (\bar{g}_1^{(i)}(\bar{W}_1^{(i)}x), \cdots, \bar{g}_{l_i}^{(i)}(\bar{W}_{l_i}^{(i)}x))^\top$ with $\bar{g}_j^{(i)} = g_j^{(i)}$ $(1 \le j \le l_i)$ and $\bar{g}_j^{(i)} = h_{j-l_i}^{(i)}$ $(l_i + 1 \le j \le 2l_i)$. Hence, $\bar{g}_j^{(i)} \in \mathcal{H}^{\beta_i}([0,1]^{d_i}, B_0)$ for all $1 \le i \le q+1$, $1 \le j \le l_i$. Note that $d_{q+1} = 1$ and β_{q+1} can be arbitrarily large.

Therefore, $f_{\mathcal{M}} \in \mathcal{CS}(q+1, (d, 2l_1, \dots, 2l_{q-1}, 2, 1), (\mathbf{d}, 1), (\boldsymbol{\beta}, \beta_{q+1}), B_0)$. By Lemma 30, for any $P, Q \in \mathbb{Z}_+$, there exists a function $f_{\mathcal{M}*} : \mathbb{R}^d \to \mathbb{R}$ implemented by a ReLU network with width $\mathcal{W} = 76(\lfloor\beta\rfloor + 1)^2 \max_i\{l_i 3^{d_i} d_i^{\lfloor\beta\rfloor+1}\} P \lceil \log_2(8P) \rceil$ and depth $\mathcal{D} = 21(q+1)(\lfloor\beta\rfloor + 1)^2 Q \lceil \log_2(8Q) \rceil + 2\sum_{i=1}^q d_i + 3q + 2$ such that

$$|f_*(x) - f_{\mathcal{M}}(x)| \le c' q (B_0 \vee 1)^{q+1} (\lfloor\beta\rfloor + 1)^2 d_*^{\lfloor\beta\rfloor + (\beta \vee 1)/2} \max_i (PQ)^{-2\beta_i^*/d_i},$$

for all $x \in [0,1]^d$ and some c' depending on $\{l_i\}_{i=1}^{q-1}$. Consequently,

$$\|f_{\mathcal{M}*}(x) - f_{\mathcal{M}}(x)\|_{L_2(f_{\mathcal{M}})}^2 \le (c'q)^2 (B_0 \vee 1)^{2q+2} (\lfloor\beta\rfloor + 1)^4 d_*^{2\lfloor\beta\rfloor + \beta \vee 1} \max_i (PQ)^{-4\beta_i^*/d_i}.$$

By Lemmas 4, 6, and 7, we have

$$\mathbb{E}_S \|\tilde{f}_n - f_0\|_{L_2(f_0)}^2 \le c_{10} \mathcal{S}\mathcal{D} \log \mathcal{S} \frac{\log n}{n} + c_{11} (\lfloor\beta\rfloor + 1)^4 d_*^{2\lfloor\beta\rfloor + \beta \vee 1} \max_i (PQ)^{-4\beta_i^*/d_i},$$

where c_{10} is a constant only depending on $\nu, \rho, L_0, \gamma, \Gamma$, and $f_{\mathcal{R}}$, and $c_{11} = (1+\rho)^3 c_2 c_1^{-1} (c'q)^2 (B_0 \vee 1)^{2q+2}$ with c_1 and c_2 being the same constants as in Lemma 5. Lastly, we obtain the error bound (11) in Theorem 18 by setting $P = 1$ and $Q = \lceil n^{d_{i_0}/\{2(d_{i_0} + 2\beta_{i_0}^*)\}} \rceil$, where $i_0 \in \arg\min_i \beta_i^*/d_i$. This concludes the proof of Theorem 18. ■

Appendix B. Supporting definitions and lemmas

This section contains the supporting definitions and lemmas needed in the proofs of theoretical results.

Definition 27 *Suppose that $\mathcal{G}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$, given a sequence $x_{1:k} :- (x_1, \dots, x_k) \in \mathcal{X}^k$, let $\mathcal{N}_p(\epsilon, \mathcal{G}, x_{1:k})$ be the ϵ-covering number of $\mathcal{G}|_{x_{1:k}} :- \{(g(x_1), \dots, g(x_k)) : g \in \mathcal{G}\} \subset \mathbb{R}^k$ with respect to the $\| \cdot \|_p$ $(1 \le p \le \infty)$ norm. The uniform covering number $\mathcal{N}_p(\epsilon, \mathcal{G}, k)$ is defined as the maximum of $\mathcal{N}_p(\epsilon, \mathcal{G}, x_{1:k})$ over all $x_{1:k} \in \mathcal{X}^k$, i.e.*

$$\mathcal{N}_p(\epsilon, \mathcal{G}, k) = \max\{\mathcal{N}_p(\epsilon, \mathcal{G}, x_{1:k}) : x_{1:k} \in \mathcal{X}^k\}.$$

The pseudo-dimension Pdim($\mathcal{G}$) is the largest interger k for which there exists $(x_1, \dots, x_k, y_1, \dots, y_k) \in \mathcal{X}^k \times \mathbb{R}^k$ such that, for any $(a_1, \dots, a_k) \in \{0, 1\}^k$ there exists $g \in \mathcal{G}$ such that $g(x_i) > y_i \iff a_i = 1$ for all i.

The pseudo-dimension is closely related with VC-dimension in the sense that $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathcal{F})$ if $\mathcal{F}$ is a class of functions generated by a neural network with fixed architecture and fixed activation function (Theorem 14.1 in Anthony and Bartlett (1999)).

Definition 28 (Hölder class) *Let $\beta = s + r > 0$ with $s = \lfloor \beta \rfloor \in \mathbb{N}$ and $r \in (0, 1]$, where $\lfloor \beta \rfloor$ denotes the largest integer strictly less than β and $\mathbb{N}$ is the set of nonnegative integers. For a finite $B_0 > 0$, the Hölder class $\mathcal{H}^\beta(\mathbb{R}^d, B_0)$ is defined as*

$$\begin{aligned}&\mathcal{H}^\beta(\mathbb{R}^d, B_0)\\ &= \{f : \mathbb{R}^d \to \mathbb{R}, \max_{\|\alpha\|_1 \le s} \|\partial^\alpha f\|_\infty \le B_0, \max_{\|\alpha\|_1 = s} \sup_{x \ne y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x - y\|_2^r} \le B_0\},\end{aligned}$$

where $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, $\partial^\alpha = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$ and $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$. For any subset $\mathcal{X} \subseteq \mathbb{R}^d$, we denote $\mathcal{H}^\beta(\mathcal{X}, B_0) = \{f : \mathcal{X} \to \mathbb{R}, f \in \mathcal{H}^\beta(\mathbb{R}^d, B_0)\}$.

Lemma 29 (Theorem 11.6 in Györfi et al. (2002)) *Let $B \ge 1$ and let $\mathcal{G}$ be a set of functions $g : \mathbb{R}^d \to [0, B]$. Let $Z, Z_1, \dots, Z_n$ be i.i.d. $\mathbb{R}^d$-valued random variables. Assume $\alpha > 0, 0 < \epsilon < 1$, and $n \ge 1$. Then,*

$$\begin{aligned}&Pr\Big(\sup_{g \in \mathcal{G}} \frac{\frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)}{\alpha + \frac{1}{n}\sum_{i=1}^n g(Z_i) - \mathbb{E}g(Z)} > \epsilon\Big)\\ &\le 4\mathbb{E}\mathcal{N}_1\Big(\frac{\alpha\epsilon n}{5}, \mathcal{G}, Z_{1:n}\Big) \exp\Big(-\frac{3\epsilon^2 \alpha n}{40B}\Big).\end{aligned}$$

Lemma 29 follows directly from Theorem 11.6 in Györfi et al. (2002), wherein the definition of complexity is slightly different from that in Definition 27.

Lemma 30 (Lemma C.2 in Wu et al. (2024)) *Assume that $f \in \mathcal{H}^\beta([0, 1]^d, B_0) \cap \mathcal{CS}(q, \mathbf{l}, \mathbf{d}, \boldsymbol{\beta}, B_0)$. For any $P, Q \in \mathbb{Z}_+$, there exists a function φ_0 implemented by a ReLU network with width $\mathcal{W} = 38(\lfloor \beta \rfloor + 1)^2 \max_i\{l_i 3^{d_i} d_i^{\lfloor \beta \rfloor + 1}\} P \lceil \log_2(8P) \rceil$ and depth $\mathcal{D} = 21q(\lfloor \beta \rfloor + 1)^2 Q \lceil \log_2(8Q) \rceil + 2\sum_{i=1}^q d_i + 3(q - 1)$ such that*

$$|f(x) - \varphi_0(x)| \le c' q(B_0 \vee 1)^q (\lfloor \beta \rfloor + 1)^2 d_*^{\lfloor \beta \rfloor + (\beta \vee 1)/2} \max_i (PQ)^{-2\beta_i^*/d_i},$$

for all $x \in [0, 1]^d$, where $d_* = \max_i d_i$, $\beta_i^* = \beta_i \prod_{j=i+1}^q (\beta_j \wedge 1)$ $(1 \le i \le q - 1)$, $\beta_q^* = \beta_q$, and $c' = 19 \prod_{i=1}^{q-1} l_i$.

Appendix C. Simulation details

This section provides the detailed density estimation and neural network training setups in our simulation studies.

In our data-augmented nonparametric noise contrastive estimation method, we set the hyperparameters $\nu = 5$ and $\rho = 0.3$, unless otherwise specified. We use the isotropic Gaussian distribution $N_d(\mathbf{0}, \sigma^2\mathbf{I})$ as the reference distribution, where $\mathbf{0}$ and $\mathbf{I}$ denote the zero vector and identity matrix of proper dimension, respectively. For Models 1 - 3, the variance σ^2 in the reference distribution is set to be $0.5^2, 2^2, 2.5^2$, respectively. For the sake of practical considerations and computational tractability, the small values of density estimations by our method are truncated at level 10^{-16}. We also consider a special case of our method, $\rho = 0$. That is, we don't perform data augmentation on the observed data and the target density is estimated directly. Under this setting, there is no need for truncating the density estimation.

The neural networks utilized in our method are implemented with PyTorch (Paszke et al., 2019). In the 2-D experiments, the networks are configured with a depth of 4, a width of 16, and the leaky ReLU activation function $\bar{\sigma}(x) = \max\{0.1x, x\}$. Additionally, a scaled sigmoid function is incorporated after the final layer to truncate the output of network to a bounded interval. An Adam optimizer (Kingma and Ba, 2017) with a learning rate of 0.0002 is used for backpropagation and updating network parameters. The network training is stopped after $2,500$ epochs, as beyond this point, there is no significant improvement in loss function.

References

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

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *4th International Conference on Learning Representations*, 2016.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Thijs Bos and Johannes Schmidt-Hieber. A supervised deep learning method for nonparametric density estimation. *Electronic Journal of Statistics*, 18(2):5601 – 5658, 2024.

Omar Chehab, Aapo Hyvarinen, and Andrej Risteski. Provable benefits of annealing for estimating normalizing constants: Importance sampling, noise-contrastive estimation, and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pages 45945–45970, 2023.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*, volume 32, pages 8174–8184, 2019.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. Opportunities and challenges of diffusion models for generative AI. *National Science Review*, 11(12):nwae348, 2024.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. 2015. *arXiv*: 1410.8516v6.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *5th International Conference on Learning Representations*, 2017.

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

Cao Feng, Alistair Sutherland, Ross D. King, Stephen H. Muggleton, and Robert J. Henery. Statlog Project. UCI Machine Learning Repository, 1993. DOI: `https://doi.org/10.24432/C5XS3B`.

Ruiqi Gao, Erik Nijkamp, Diederik P. Kingma, Zhen Xu, Andrew M. Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2020.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 881–889, 2015.

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*, volume 27, pages 2672–2680, 2014.

Aditya Grover, Manik Dhar, and Stefano Ermon. Flow-GAN: Combining maximum likelihood and adversarial learning in generative models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, New York, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017. *arXiv*: 1412.6980v9.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2022. *arXiv*: 1312.6114v11.

Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE Transactions on Information Theory*, 68 (6):4032–4042, 2022.

Holden Lee, Chirag Pabbaraju, Anish Prasad Sevekari, and Andrej Risteski. Pitfalls of Gaussians as a noise distribution in NCE. In *11th International Conference on Learning Representations*, 2023.

Qiang Liu, Jian Peng, Alexander Ihler, and John Fisher. Estimating the partition function by discriminance sampling. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, page 514–522, 2015.

Qiao Liu, Jiaze Xu, Rui Jiang, and Wing Hung Wong. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15): e2101344118, 2021.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

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

Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21 (1):7018–7055, 2020.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, pages 2338–2347, 2017.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037, 2019.

Miika Pihlaja, Michael U. Gutmann, and Aapo Hyvärinen. A family of computationally efficient and simple estimators for unnormalized statistical models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, page 442–449, 2010.

Shebuti Rayana. ODDS Library. Stony Brook University, Department of Computer Science, 2016. URL `https://shebuti.com/outlier-detection-datasets-odds/`. Accessed: September 18, 2025.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538, 2015.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4905–4916, 2020.

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

Tanya Schmah, Geoffrey E. Hinton, Steven Small, Stephen Strother, and Richard Zemel. Generative versus discriminative training of RBMs for classification of fMRI images. In *Advances in Neural Information Processing Systems*, volume 21, pages 1409–1416, 2008.

Johannes Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. 2019. *arXiv*: 1908.00695v1.

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

David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11918–11930, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations*, 2021.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Lucas Theis and Matthias Bethge. Generative image modeling using spatial LSTMs. In *Advances in Neural Information Processing Systems*, volume 28, pages 1927–1935, 2015.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17 (1):7184–7220, 2016.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, Cham, 2023.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Kevin S. Woods, Christopher C. Doss, Kevin W. Bowyer, Jeffrey L. Solka, Carey E. Priebe, and W. Philip Kegelmeyer Jr. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6):1417–1436, 1993.

Di Wu, Yuling Jiao, Li Shen, Haizhao Yang, and Xiliang Lu. Neural network approximation for pessimistic offline reinforcement learning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024.

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

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 639–649, 2018.