Classification Logit Two-Sample Testing by Neural Networks for Differentiating Near Manifold Densities

Xiuyuan Cheng and Alexander Cloninger

***Abstract*— The recent success of generative adversarial networks and variational learning suggests that training a classification network may work well in addressing the classical two-sample problem, which asks to differentiate two densities given finite samples from each one. Network-based methods have the computational advantage that the algorithm scales to large datasets. This paper considers using the classification logit function, which is provided by a trained classification neural network and evaluated on the testing set split of the two datasets, to compute a two-sample statistic. To analyze the approximation and estimation error of the logit function to differentiate near-manifold densities, we introduce a new result of near-manifold integral approximation by neural networks. We then show that the logit function provably differentiates two sub-exponential densities given that the network is sufficiently parametrized, and for on or near manifold densities, the needed network complexity is reduced to only scale with the intrinsic dimensionality. In experiments, the network logit test demonstrates better performance than previous network-based tests using classification accuracy, and also compares favorably to certain kernel maximum mean discrepancy tests on synthetic datasets and hand-written digit datasets.**

***Index Terms*— Neural network two-sample test, neural network approximation theory, maximum mean discrepancy, manifold data analysis.**

I. Introduction

THE powerful expressiveness of neural networks and the recent progress in neural network optimization suggest the natural idea of using a network for the comparison of two unknown distributions p and q from finitely observed data samples, a problem known as *two-sample testing* in statistics [1]. As a central problem in statistics, two-sample testing is widely encountered in general data analysis of biomedical data, audio and imaging data, etc. [2]–[6], and particularly, in the machine learning application of training and evaluating generative models [7]–[16]. Many existing two-sample tests for multivariate data are based on certain estimators of a distance or divergence between p and q. Important examples include Maximum Mean Discrepancy (MMD), especially kernel-based MMD [17], [18] and distance of Reproducing Kernel Hilbert Space (RKHS) Mean Embedding [3], [4], and divergence based methods which may involve non-parametric estimation of density difference or density ratio [19]–[22]. While these methods have been intensively studied and theoretically well-understood, their application is often restricted to data with small dimensionality and/or small sample size due to model and computational limitations. More background information on two-sample tests can be found in Section I-B.

The idea of training a classifier to solve two-sample problem dates back to earlier statistical works, and was recently revisited in the machine learning literature [5]. Notably, the training of discriminator in generative adversarial networks (GAN) [9]–[11] is also to solve a two-sample problem: in the training of GAN, in each iteration a discriminator network (D-net) is trained to distinguish the density q produced by a generative network from the data density p which is only accessible via observed samples, that is, a two-sample problem. (Strictly speaking, the task of D-net is a goodness-of-fit problem as the model density q is analytically given [14]–[16]. The scenario of two-sample is also important since batch sampling is commonly used in training GAN and other generative networks.) The success of GAN and adversarial training in many applications suggests that training a neural network potentially provides a powerful tool to solve the two-sample problem, for various applications in machine learning and data analysis.

The current paper proposes classification logit two-sample test, a general method for two-sample problems based on training a classification neural network. The proposed test statistic is the log ratio of the class probabilities averaged over samples, which can be computed once a classifier network is trained. The theoretical results cover neural network approximation and estimation analysis, which we summarize in Section I-A. In our analysis, an important setting is when high dimensional data exhibit intrinsically low-dimensional structures, which is a scenario frequently encountered in generative models as well as other applications. Take GAN as an example: because

Manuscript received 24 November 2020; revised 20 January 2022; accepted 1 May 2022. Date of publication 17 May 2022; date of current version 15 September 2022. This work was supported by the NSF under Grant DMS-1818945. The work of Xiuyuan Cheng was supported in part by the NSF under Grant DMS-1820827, in part by the NIH under Grant R01GM131642, and in part by the Alfred P. Sloan Foundation. The work of Alexander Cloninger was supported in part by the NSF under Grant DMS-2012266, in part by the Russel Sage Foundation under Grant 2196, and in part by Intel Research. *(Corresponding author: Alexander Cloninger.)*

Xiuyuan Cheng is with the Department of Mathematics, Duke University, Durham, NC 27708 USA (e-mail: xiuyuan.cheng@duke.edu).

Alexander Cloninger is with the Department of Mathematics, Halicioğlu Data Science Institute, University of California at San Diego, San Diego, CA 92093 USA (e-mail: acloninger@ucsd.edu).

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3175691.

Digital Object Identifier 10.1109/TIT.2022.3175691

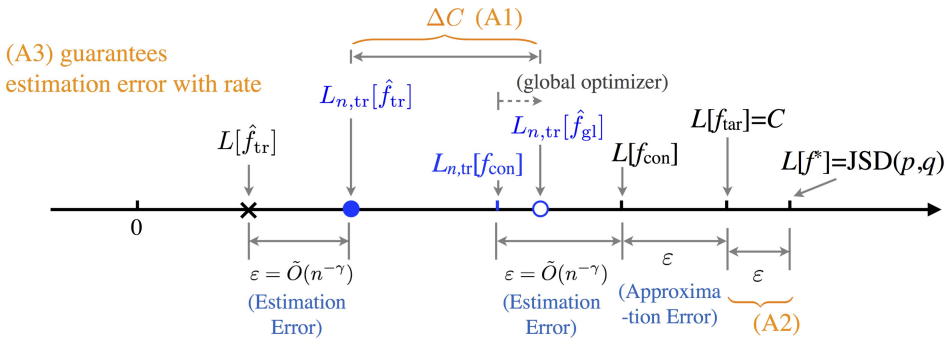


Fig. 1. Roadmap of analysis, illustrating the different functions f^*, f_{tar}, $\hat{f}_{\mathrm{gl}}$, $\hat{f}_{\mathrm{tr}}$ and f_{con}, (the last three lying in the network function family $\mathcal{F}_\Theta$). The goal of analysis is to bound $L[f_{\mathrm{tar}}]-L[\hat{f}_{\mathrm{tr}}]$ to be less than some L_{gap}, see (30) and (31), which is proved in Theorem IV.6(1).

a GAN traditionally maps from a low dimensional latent space to a higher dimensional data space where q is defined, the resulting model density q is concentrated on or near a low-dimensional manifold in the ambient space. This means it is critical that the analysis of the D-net focuses on this manifold or near manifold setting.

From an approximation theoretic perspective, recent work has shown that the D-net does not have to scale with the number of points in order to bound the generalization error [23], or for constructing GANs [24]. Other important directions in this line include studying the effects of depth of the network [25], [26], and showing that the space of fixed size networks is not closed under L^p norm [27]. While neural network approximation of functions on low-dimensional manifolds has been studied [28], [29], and the needed network complexity is reduced to be intrinsic, these works have not considered high dimensional data that are lying near low-dimensional manifolds. Finally, generative networks that explicitly model the low-dimensional manifold in the latent space have been studied for data lying on low-dimensional manifolds [30], [31].

A critical aspect of considering neural networks in a statistical context with finitely many observed samples is that the family of networks cannot be so large that overfitting becomes a concern. This means that the family networks constructed in any approximation theory argument must have bounded estimation error even as the approximation error goes to 0. This is violated if we allow for arbitrary growth of the network architecture without some additional assumption, and the traditional construction using local polynomial expansions [23]–[25] does not automatically satisfy this requirement. This motivates the use of a wavelet type argument [28], and one of the key results in this paper (Theorem III.2) that establishes an upper bound on the Lipschitz constants of the constructed networks independent of the architecture of the network. We review and discuss all these connections in more detail in Section I-C.

A. Main Results

The main results and contributions of the current paper are summarized as below:

- We introduce a neural network-based two-sample test statistic using classification logit function. The algorithm inherits the scalable computational efficiency of neural networks. Numerical experiments show that the proposed test compares favorably to kernel MMD tests and earlier neural network test based on classification accuracy, on both synthetic manifold data and hand-written digits datasets.
- Theoretical guarantee of testing power is proved (c.f. Theorem IV.6) for sub-exponential densities p and q in $\mathbb{R}^D$, and the needed network complexity is reduced to be intrinsic when p and q lie on or near to a low-dimensional manifold embedded in $\mathbb{R}^D$ even when D is high. The way of proving Theorem IV.6 consists of a combination of approximation and estimation error analysis, under an assumption on small optimization error, which bounds $L[\hat{f}_{\mathrm{tr}}]-L[f^*]$ to be sufficiently small and consequently obtains a strictly positive $L[\hat{f}_{\mathrm{tr}}]$. A roadmap of analysis is provided in Section II-D and Figure 1.
- We prove a result of near-manifold integral approximation (c.f. Theorem III.1), namely approximating the integral of some regular function f with respect to a density p, which concentrates near a low-dimensional manifold $\mathcal{M}$, by the integral of f_{con}, where f_{con} is constructed by a neural network to approximate f *on* the manifold $\mathcal{M}$ only. A key step in the analysis is to show that as we make the neural network approximation error $\|f-f_{\mathrm{con}}\|_{\infty,\mathcal{M}}$ approach zero by enlarging the network complexity, the Lipschitz constant of f_{con} in $\mathbb{R}^D$ can be made uniformly bounded, and the bound depends on $\mathcal{M}$ and f only. This result (c.f. Theorem III.2) extends the earlier result in [28], and can be of independent interest.

The needed theoretical assumptions for proving test power guarantee are given in Section II-D. The softmax classification loss corresponds to the Jensen-Shannon divergence (JSD) between two densities, which belongs to a general class of f-divergences. Thus our method and techniques may extend to a broad class of classification networks [11]. Since JSD is a prototypical case of f-divergence, we focus on softmax classifier network in this paper. We discuss future directions in the Section VI.

Notation: We provide a list of default notations in Table I. In particular, f_{con} is named as f_N in Theorem III.2 and

TABLE I
LIST OF DEFINITIONS AND NOTATIONS

p, q	two distributions in two-sample problem, also used to denote the two densities
X, Y	independent i.i.d. datasets drawn from p and q respectively
n_X, n_Y	number of samples in X and Y
$\mathcal{P}_\sigma$	class of near-manifold densities, c.f. (9), $\sigma > 0$ is the scale of sub-exponential decay
$\mathcal{M}$	d-dimensional manifold in $\mathbb{R}^D$
$\{(U_i, \phi_i)\}_{i=1}^K$	manifold atlas consisting of K patches, $U_i = B(x_i, \delta) \cap \mathcal{M}$
δ	radius of Euclidean ball in manifold atlas, $\delta > 0$ is a fixed $O(1)$ constant determined by $\mathcal{M}$
η_i	partition of unity function on $\mathcal{M}$, supp(η_i) is on U_i
f_θ	neural network *logit* function, c.f. 2
f^*	log ratio of densities $f^* = \log \frac{p}{q}$, c.f. (8)
f_{tar}	smooth and compactly supported surrogate function of f^*, also the target function for neural network approximation, c.f. Assumption 2
f_{con}	constructed neural network function
$\hat{f}_{\text{gl}}$	neural network solution that is global optimizer of the training loss $L_{n,\text{tr}}[f]$
$\hat{f}_{\text{tr}}$	neural network solution attained by actual optimization procedure on the training data
$T[f]$	population test statistic of logit function f, c.f. (4)
$L_{n,\text{tr}}[f]$	empirical training loss of n training samples of logit function f, c.f. (6)
$L[f]$	population loss of logit function f, c.f. (7)

its proof. Throughout the paper, for integrals we may use the abbreviated notation $\int f$ to represent $\int f(x)dx$. For a set A, $|A|$ stands for the cardinal number of A. For the asymptotic notations, big-O $O(\cdot)$: we write $f = O(g)$ if there is $C > 0$ such that $|f| \leq C|g|$ in the limit; Tilde $\sim$: we write $f \sim g$ for f, $g \geq 0$ if there are $C_1, C_2 > 0$ such that $C_1 g \leq f \leq C_2 g$ in the limit (note that $\sim$ also denotes "distributed as" for random variable, when there is no confusion); We use $\tilde{O}(\cdot)$ to stand for $O(\cdot)$ multiplied by another factor involving a log, to be defined specifically in the context.

B. Background: Two-Sample Problem and Test Statistic

We introduce certain terminologies from statistical literature about the two-sample problem. Formally, the two-sample problem asks to test the null hypothesis $\mathcal{H}_0 : p = q$ given datasets

$$X = \{x_i\}_{i=1}^{n_X}, \quad x_i \sim p \quad i.i.d., \quad Y = \{y_j\}_{j=1}^{n_Y}, \quad y_j \sim q \quad i.i.d.,$$

and X independent from Y. It is also of application interest to provide indication of where q differs from p. Similarly, K-sample problem can be considered where $K > 2$, namely to differentiate distributions among K data sets. We focus on the two-sample problem in the paper, and the neural network classifier method extends to the $K > 2$ situations. Throughout the paper, we assume that the distributions have density functions, and use p and q to stand for both the distribution and the density function.

Most two-sample testing method is based upon a test *statistic*

$$\hat{T} = \hat{T}(X, Y),$$

which is computed from the two datasets, and a test threshold τ. The null hypothesis $\mathcal{H}_0$ is rejected if $\hat{T} > \tau$. To control the false discovery under the null, the threshold τ is usually set to the smallest value s.t. $\Pr[\hat{T} > \tau | \mathcal{H}_0] \leq \alpha$, where $\alpha \in (0, 1)$ is a pre-specified number called the *significance level* of the test (typically $\alpha = 0.05$). Algorithm-wise, τ is given either by theory (the probabilistic distribution of $\hat{T}$ under $\mathcal{H}_0$) or computed from data. Specifically, *permutation test* is a standard procedure to determine τ by a bootstrap approach [32], [33], and was used in kernel MMD tests [18]. For a given test, the *test power* is measured by $\Pr[\hat{T} > \tau | \mathcal{H}_1]$, namely the probability of true discovery when p and q indeed differ. The test is called asymptotically consistent if the test power $\to 1$ as the number of samples $n_X, n_Y \to \infty$ and usually the ratio $n_X / n_Y \to$ a nonzero constant.

C. Related Works

1) Classification and Two-Sample Testing: The relations between two-sample testing, divergence estimation and binary classification has been pointed out earlier in [19], [34], [35]. [36] studied Fisher LDA classifier used for testing mean shift of Gaussian distributions. Discriminative approach has also been used to detect and correct covariant shifts [37], [38]. Training classifier provides an estimator of density ratio, as has been pointed out in [39] and in the formulation of learning generative models [40]. While distribution divergence estimation has been studied and used for two-sample problems [20]–[22], the use of neural network as a divergence estimator for two-sample testing was less investigated. In terms of theoretical guarantee of test power, the analysis in [5] assumes a non-zero population test statistic when $p \neq q$ but the expression is not specified, along with other approximations. Theoretical analysis of neural network two-sample testing power remains limited.

2) MMD and Kernel MMD Tests: MMD, also known as Integral Probability Metric (IPM), encloses a wide class of two-sample statistics such as Kolmogorov-Smirnov statistic, Wasserstein metric, etc. Particularly, kernel-based MMD [17], [18] has been widely applied due to its non-parametric form, and recently in training moment matching networks [7], [8] and evaluating generative models [13]. The population test statistic of MMD test takes the general form as

$$D(p, q) = \sup_{f \in \mathcal{F}} \int f(p - q), \tag{1}$$

where $\mathcal{F}$ is certain restricted family of functions. In kernel MMD, $\mathcal{F}$ is the L^2-unit ball in the RKHS. To optimize kernel choice, [18] considers selecting kernel bandwidth from data, [6] studies anisotropic kernels. Optimizations of kernel through convex combination of multiple kernels [41], adapting reference locations in the Mean Embedding test [4], and neural network parametrization [42] have been introduced which maximize estimated testing power. The combination of neural network feature learning and kernel MMD has been studied

in [8], where the training is typically more costly than that of a classifier network. Compared to kernel methods, neural networks are often algorithmically more scalable, and the current paper studies the theoretical guarantee of testing power, and compares performance in practice.

The proposed network logit test in this work is not an MMD test, because the training objective is an (empirical) f-divergence rather than an IPM. However, the test statistic resembles the form of MMD test statistic, which is evaluated at the supremum f in $\mathcal{F}$. We discuss the connection in Section II-B, and compare with kernel MMD tests in experiments.

3) Relation to Goodness-of-Fit Test and GAN: The goodness-of-fit problem differs from the two-sample problem in that one of the two densities is analytically accessible. Using the explicit formula of q, methods based on kernel Stein discrepancy have been developed in [14]–[16] and applied to generative model evaluation. However, the computation of the *score function* $\nabla \log q$ may be difficult for certain generative models, including many generative networks. Meanwhile, in many generative models including GAN the goodness-of-fit is evaluated by batch sampling, i.e. the two-sample setting: Kernel MMD is used in [7], [8]; GAN, Wasserstein GAN [10] and f-GAN [11] estimate density divergence by a trained network (the D-net). Since the success of GAN training relies on the discriminative power of the D-net, the efficiency of using a neural network for the two-sample test is important for the training and evaluating of such models.

4) Neural Network Approximation Theory: The expressiveness of neural networks and their ability to approximate functions to error ϵ considers the minimum architecture needed to approximate a family of functions, independent of whether an optimization scheme converges to the proposed constructed network. The number of parameters needed in a network to bound the point-wise error by ϵ is known to scale as $\epsilon^{-D/r}$ where D is dimension of the space and r is the smoothness of the target function [25], [43], [44]. It has also been established that if the function is in a Korobov-2 space (i.e., smooth mixed derivatives up to order $\partial_{x_1}^2 \dots \partial_{x_D}^2 f$), then the complexity can be reduced to $\epsilon^{-1/2}|\log(\epsilon)|^{3D/2}$ [45]. Similarly, when the data lies on a lower dimensional manifold of dimension $d < D$, the complexity of the networks scales as $\epsilon^{-d/2}$ for twice differentiable functions when the network depth is bounded [23], [28], [46], and regression of Hölder functions on such manifolds using deep ReLU networks was studied in [29] which proved estimation convergence rate depending on d. These methods are mostly applied to regression problems, and consider data distributions that are absolutely continuous in $\mathbb{R}^D$ or lying on a low-dimensional manifold, but have not considered the case where data is concentrated near low-dimensional structures.

II. Log-Ratio Test by Neural Network Classification

A. Classification Logit Test

The proposed test statistic is computed in the following way. Given two datasets X and Y as in Section I-B, without loss of generality suppose $n = n_X + n_Y$ is even integer. Same as in [5], we split the dataset $\mathcal{D} = \{(x_i, 0)\}_{i=1}^{n_X} \cup \{(y_j, 1)\}_{j=1}^{n_Y} = \{(z_i, l_i)\}_{i=1}^{n}$, $l_i \in \{0, 1\}$, into two halves, i.e. $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$, $|\mathcal{D}_{\text{tr}}| = |\mathcal{D}_{\text{te}}| = n/2$, $\mathcal{D}_{\text{te}} = X_{\text{te}} \cup Y_{\text{te}}$ and similarly for the training set. The method consists of two phases of training and testing:

- Training. A binary classification neural network is trained on $\mathcal{D}_{\text{tr}}$ using softmax loss (equivalent to applying logistic regression to the output of last layer before the loss layer), which gives estimated class probabilities as
$$\Pr[l = 0|x] = \frac{e^{u_\theta(x)}}{e^{u_\theta(x)} + e^{v_\theta(x)}},$$
$$\Pr[l = 1|x] = \frac{e^{v_\theta(x)}}{e^{u_\theta(x)} + e^{v_\theta(x)}},$$
where $u_\theta(x)$ and $v_\theta(x)$ are activations in the last hidden layer of the network, θ denoting the network parametrization. We define
$$f_\theta := u_\theta - v_\theta, \tag{2}$$
which is the *logit*. The mathematical formulation of network training is detailed in Section II-C.
- Testing. After f_θ is parametrized by a trained neural network, the test statistic is computed as
$$\hat{T} = \frac{1}{|X_{\text{te}}|} \sum_{x \in X_{\text{te}}} f_\theta(x) - \frac{1}{|Y_{\text{te}}|} \sum_{y \in Y_{\text{te}}} f_\theta(y), \tag{3}$$
which can be equivalently written as $\hat{T} = \int f_\theta(x)(\hat{p}_{\text{te}}(x) - \hat{q}_{\text{te}}(x))dx$ where $\hat{p}_{\text{te}}$ and $\hat{q}_{\text{te}}$ stand for the empirical measure of X_{te} and Y_{te} respectively.
1) Determination of τ: Once the logit function f_θ is evaluated on each testing sample, the test threshold τ can be computed by a bootstrap method [32], [33]: randomly permute the $|\mathcal{D}_{\text{te}}|$ many labels l_i on the test set, and recompute the test statistics for m_{perm} times, typically a few hundreds. Then τ is set to be the (1-α)-quantile of the empirical distribution so as to control the type-I error to be at most α. Note that this permutation test does not require retraining the network upon each permutation nor re-evaluation of the neural network on testing samples.

The above classifier logit test can be used with other classifiers than neural networks, e.g., logistic regression, which is equivalent to restricting f_θ to be a linear mapping or the network to have only one linear layer. A main advantage of neural network is the enlarged expressiveness of the class of functions f_θ that can be represented or approximated.

2) Computational Complexity: Given n data samples, evaluating the network output on each sample takes a fixed amount of flops, and thus computing the test statistic takes $O(n)$ operations. The permutation test to determine τ adds negligible cost since f_θ has been evaluated at each test sample, and permuting the class labels only reorders these computed values. The training phase is certainly more expensive, though theoretically the overall complexity is $O(n)$ assuming that training is terminated after a fixed number of epochs. Note that the computation can be conducted by batch sampling so the algorithm scales to large sample size and also to multiple sample problems.

B. Witness Function

Given the logit function f, the empirical test statistic is written as $\hat{T} = \int f(\hat{p}-\hat{q})$, the subscripts being omitted without causing confusion, and the population test statistic is

$$T[f] := \int f(p-q), \tag{4}$$

which is of the same form as the MMD discrepancy (1) evaluated at the sup-achieved f. In the literature of kernel MMD [18], such f is named the *witness function*, as it provides an indication of where q differs from p. The indicator of differential regions can be of more application interest than the hypothesis test itself. The witness function for kernel MMD is expressed via the reproducing kernel.

In our setting, the density differential indicator is provided by the logit function f_θ of the trained classifier network. We follow the MMD literature and call f_θ the (empirical) witness function of the proposed logit test. We will see in Section II-C that $f^* := \log \frac{p}{q}$ gives the unconstraint optimizer of the population training objective. We call f^* the population witness function of the logit test.

The witness function plays an important role in the ability of the test to distinguish two densities. When $p \neq q$, once a witness function f with $T[f] > 0$ is obtained (from the training set), the two-sample test (on the test set) using $\hat{T}$ will be asymptotically consistent as a direct result of Central Limit Theorem (CLT). The difference in test power thus depends on the quality of f, e.g., how large is the bias $T[f]$ compared to the variance of $\hat{T}$. Consequently, the efficiency of the network classification test lies in how well the neural network can express a good witness function and how it can be identified via the optimization, which is the central question of our analysis. Apart from theory, we also experimentally compare the witness function of different tests in Section V.

C. Identification of f_θ by Neural Network Training

Mathematically, the training of classification neural network optimizes the following objective

$$\sum_{x\in X_{\text{tr}}} \log D(x) + \sum_{y\in Y_{\text{tr}}} \log(1-D(y)), \quad D(x) := \frac{e^{f(x)}}{1+e^{f(x)}}, \tag{5}$$

called the (negative) empirical training loss.[1] After normalizing by number of samples, where we assume same of samples in X and Y, and then $|X_{\text{tr}}| = |Y_{\text{tr}}|$, for simplicity throughout the paper, the optimization of the empirical loss (up to a additive constant) can be written as

$$\begin{aligned}
&\max_{f\in\mathcal{F}_\Theta} L_{n,\text{tr}}[f] \\
&\quad = \frac{1}{2}\left(\frac{1}{|X_{\text{tr}}|}\sum_{x\in X_{\text{tr}}} \log D(x) \right. \\
&\qquad\qquad \left. + \frac{1}{|Y_{\text{tr}}|}\sum_{y\in Y_{\text{tr}}} \log(1-D(y)) + 2\log 2\right) \\
&\quad = \frac{1}{2}\left(\int \hat{p}_{\text{tr}}(x)\log D(x)dx \right. \\
&\qquad\qquad \left. + \int \hat{q}_{\text{tr}}(x)\ \log(1-D(x)) + 2\log 2\right),
\end{aligned} \tag{6}$$

where $\mathcal{F}_\Theta$ denotes the class of functions that can be expressed as the difference of the outputs in the last hidden layer of the classification network, and $\hat{p}_{\text{tr}}$ and $\hat{q}_{\text{tr}}$ stand for the empirical measure of X_{tr} and Y_{tr} respectively.

This training objective is the same as that of the the D-net in the standard GAN [9]. As number of samples $n \to \infty$, the corresponding population training loss can be expressed as

$$L[f] = \frac{1}{2}\left(\int p\log\frac{2e^f}{1+e^f} + \int q\log\frac{2}{1+e^f}\right), \tag{7}$$

and a direct verification shows that the maximizer is (see e.g. [9] where it is proved in terms of $D = \frac{e^f}{1+e^f}$)

$$\begin{aligned}
f^* &= \arg\max_f L[f] = \log\frac{p}{q}, \\
L[f^*] &= \frac{1}{2}\left(\int p\log\frac{2p}{p+q} + \int q\log\frac{2q}{p+q}\right) = \text{JSD}(p,q)
\end{aligned} \tag{8}$$

which characterizes the solution of (7) when the function class is arbitrarily large or large enough to contains f^*, JSD referring to the Jensen-Shannon divergence.

Note that the f_θ identified in practice, which we call $\hat{f}_{tr} \in \mathcal{F}_\Theta$ ("tr" for "trained"), differs from f^* for three reasons:

- (Approximation error) The neural network function class $\mathcal{F}_\Theta$ is of finite complexity and may not contain f^*.
- (Estimation error) Only finite training samples are used, which makes $L_{n,\text{tr}} \neq L$.
- (Optimization error) The optimization of $L_{n,\text{tr}}[f]$ may attain a local rather than global optimum. We call the global optimum $\hat{f}_{\text{gl}}$ ("gl" for "global").

The goal of analysis is thus to prove that the logit test (3) using $\hat{f}_{\text{tr}}$ obtained from training on the training set can distinguish different p and q with efficiency.

D. Assumptions and Roadmap of Analysis

We illustrate the differences among $\hat{f}_{tr}$, $\hat{f}_{gl}$, and f^* in the diagram in Figure 1, which also gives the roadmap of our analysis towards proving the test consistency guarantee (Theorem IV.6). To prove test consistency, we make the following assumptions: the first one is to handle the optimization error.

Assumption 1 (Optimization Error): The neural network training of optimizing $\max_{f\in\mathcal{F}_\Theta} L_{n,\text{tr}}[f]$ outputs $\hat{f}_{\text{tr}}$ which achieves a training loss that is ΔC-close to the global optimum, ΔC small enough, namely,

$$L_{n,\text{tr}}[\hat{f}_{\text{gl}}] - \Delta C \leq L_{n,\text{tr}}[\hat{f}_{\text{tr}}] \leq L_{n,\text{tr}}[\hat{f}_{\text{gl}}].$$

If the training algorithm is stochastic, then the above holds with high probability.

Some recent neural network optimization literature supports this assumption [47], [48]. In certain settings different than ours, it is proved that ΔC can be made small, especially

[1] "Loss" usually refers to minimization, in this paper we use L and L_n to denote the population and empirical negative softmax loss which is to maximize by the optimization.

when training with large samples and using over-parametrized networks [49]–[53]. Our experiments in Section II-E show that ΔC can be relatively small in practice. In the current paper we do not further analyze the optimization error, and more discussion can be found in Section VI.

The second assumption assumes a function $f_{\rm tar}$ that carries a sufficiently large $L[f_{\rm tar}]$, and will serve as the target function to be approximated by the constructed neural network function $f_{\rm con} \in \mathcal{F}_\Theta$.

Assumption 2 (Smooth Surrogate $f_{\rm tar}$): Suppose $p \neq q$ in $\mathcal{P}_{\rm exp}$ are given, then there exists a compactly-supported smooth function $f_{\rm tar}$ on $\mathbb{R}^D$ such that $L[f_{\rm tar}] := C > 0$.

An example of f^* and $f_{\rm tar}$ is shown in Figure 4. When f^* itself is sufficiently regular, one can set $f_{\rm tar}$ to be f^*. Otherwise, one constructs $f_{\rm tar}$ as a smooth surrogate such that $L[f_{\rm tar}] > L[f^*] - \varepsilon$, and then $C = L[f_{\rm tar}] > \text{JSD} - \varepsilon$ which is positive with small enough ε. Formally, Assumption 2 only requires C to be strictly positive, even not close to JSD, which suffices to prove Theorem IV.6. Evidently, in case of JSD being small, one would like C to be ε-close to JSD to guarantee the strict positivity of C. We further discuss the motivation and construction of $f_{\rm tar}$ in Section IV, where we explain the non-uniqueness of $f_{\rm tar}$ and the possible trade-off in the choice.

At last, we assume that the network function class has bounded Lipschitz constant in $\mathbb{R}^D$ uniformly as the network approximation error ε approaches zero. The motivation and validity of the assumption is provided in Section IV-C.1.

Assumption 3 (Lipschitz Regularization of Neural Network Function): To achieve decreasing approximation threshold ε, the network function family $\Theta = \Theta(\varepsilon)$ being used has increasing complexity, and regularization of $\mathcal{F}_\Theta$ can be applied such that for all ε,

$$\sup_{f \in \mathcal{F}_\Theta} \text{Lip}_{\mathbb{R}^D}(f) \leq L_\Theta.$$

Under the three assumptions, the building blocks of the analysis are as follows. We use ε to stand for a generic small number which may differ in different places. Starting from $L[f^*] = \text{JSD} > 0$, JSD standing for $\text{JSD}(p, q)$,

1. Use $f_{\rm tar}$ as the surrogate of f^* by Assumption 2, which gives that $L[f_{\rm tar}] > \text{JSD} - \varepsilon$. (As explained after Assumption 2, this ε may not need to be small as long as $L[f_{\rm tar}] > 0$. In this roadmap, we consider the case where $L[f_{\rm tar}]$ is ε-close to JSD for simplicity.)
2. A small $|L[f_{\rm con}] - L[f_{\rm tar}]|$ when network complexity is sufficiently large. This relies on the network approximation analysis, and we bound the needed network complexity to scale with the intrinsic dimension d when p and q lie on or near to low-dimensional manifolds. Then $L[f_{\rm con}] > \text{JSD} - 2\varepsilon$.
3. A small $|L[f_{\rm con}] - L_{n,\rm tr}[f_{\rm con}]|$ by concentration bound. Note that we will need to deal with the deviation of $L_{n,\rm tr}[\hat{f}_{\rm tr}]$ later, so we bound $\sup_{f \in \mathcal{F}_\Theta} |L[f] - L_{n,\rm tr}[f]|$ based on a bound of the covering number of the network function family $\mathcal{F}_\Theta$. In particular, for on-or-near manifold densities the convergence rate of the estimation error is improved to only involving the intrinsic dimensionality d. This step gives $L_{n,\rm tr}[f_{\rm con}] > \text{JSD} - 3\varepsilon$.
4. The optimization gives that $L_{n,\rm tr}[\hat{f}_{\rm gl}] \geq L_{n,\rm tr}[f_{\rm con}]$. Together with Assumption 1, we have that $L_{n,\rm tr}[\hat{f}_{\rm tr}] > \text{JSD} - \Delta C - 3\varepsilon$.
5. By the bound in Step 3., $|L[\hat{f}_{\rm tr}] - L_{n,\rm tr}[\hat{f}_{\rm tr}]|$ is small. Then $L[\hat{f}_{\rm tr}] > \text{JSD} - \Delta C - 4\varepsilon$.

The above steps derive a non-zero lower bound of $L[\hat{f}_{\rm tr}]$, which then leads to a non-zero population test statistic $T[\hat{f}_{\rm tr}]$. The testing consistency is then proved by the asymptotic normality of the empirical statistic $\hat{T}$ (3) (on testing set, rather than training set) by CLT. Throughout the steps, an important new analysis is for the near-manifold densities, which we detail in Section III in a more general form. The steps towards proving the test consistency are carried out in Section IV.

E. Examples of 1D Datasets

Before going to the theoretical analysis of the test consistency, we provide an illustrative example of 1D datasets. We conduct experiments on two sets of 1D data, in $\mathbb{R}$, and in $\mathbb{R}^2$ but lying near to a 1D curve, respectively. The experiments are set to verify the Assumption 1 and to observe the influence of the intrinsic dimensionality of data. More experiments of two-sample testing are reported in Section V.

1) Datasets, Model Training and Evaluation: The densities are Gaussian mixtures and have analytical formulas. Plots of the datasets are given in Figure 2 (left). Specifically,

- Example 1.
  Distribution of p is $\frac{1}{5}(\mathcal{N}(-2, \sigma_p^2) + \mathcal{N}(-1, \sigma_p^2) + \mathcal{N}(0, \sigma_p^2) + \mathcal{N}(1, \sigma_p^2) + \mathcal{N}(2, \sigma_p^2))$, where $\sigma_p = 0.8$; Distribution of q is $(1-\delta)\mathcal{N}(0, 1) + \frac{\delta}{2}\mathcal{N}(-3, \sigma_q^2) + \frac{\delta}{2}\mathcal{N}(4, \sigma_q^2)$, where σ_q = 0.5, $\delta = 0.1$.
- Example 2. The data vector (x, y) is defined by
  $$x = \frac{t}{2}, \quad y = \text{Sigmoid}(2t) + s, \quad \text{Sigmoid}(z) = \frac{1}{1 + exp(-z)},$$
  where the random variables $t \sim p$ or q as in Example 1, and $s \sim \mathcal{N}(0, \sigma_s^2)$ independently from t, $\sigma_s = 0.05$. The finite σ_s adds a small Gaussian noise to the y-coordinate such that the 2D data lie near to a 1D manifold, given by the curve when $\sigma_s = 0$.

For both examples, the network has one hidden layer of H neurons, $H = 4, 8, \cdots, 512$, and ReLU activation function. The training is by stochastic optimization and 40 replicas are conducted. For a given witness function f, either analytic or parametrized by a trained neural network, we compute the value of $L[f]$ by a numerical integration on the 1D or 2D domain.

2) Results: The values of $L[\hat{f}_{\rm tr}]$, averaged over replicas, are plotted against H for increasing number of training samples $n = |X_{\rm tr}| + |Y_{\rm tr}| = 250, 500, \cdots, 4000$, as shown in Figure 2 (right). The mean and standard deviation of $L[f_{\rm tr}]$ over replicas are shown in Table II. As shown in the plot and the table, the network training achieves $L[\hat{f}_{\rm tr}]$ more stably as n increases, and the curves indicate reliable pattern except for small values of n (n =250) and the first few small values of H ($H \leq 32$). The experimental results show that

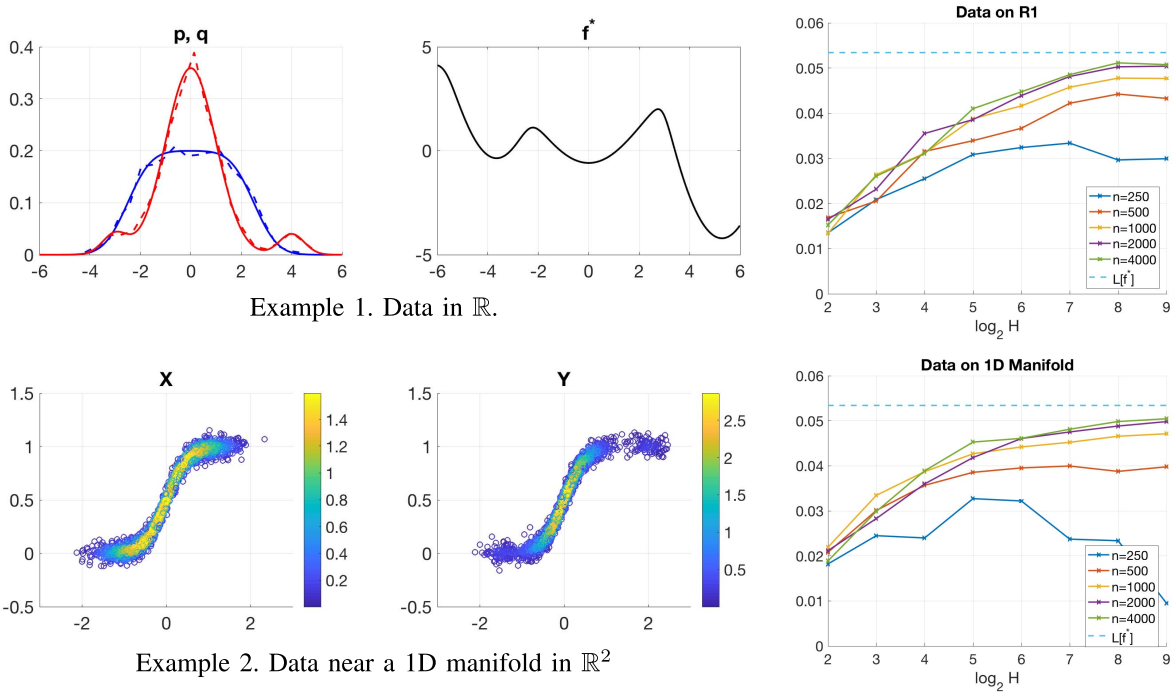


Fig. 2. (Left) Datasets $X \sim p$ and $Y \sim q$ of the two examples in Section II-E. In the plots of Example 1, the dashed lines show the histograms of samples in X and Y respectively, and the solid curves show the density functions p and q. In the plots of Example 2., samples in X and Y are shown as scatter plots in $\mathbb{R}^2$ color by the densities p and q respectively. (Right) The values of $L[\hat{f}_{\text{tr}}]$ (averaged over 40 replicas of experiments) plotted against $log_2(H)$, where H is the width of the hidden layer in the classification neural network, a proxy for the complexity of $\mathcal{F}_\Theta$. The dashed line shows $L[f^*] = \text{JSD}(p, q)$, which is the unconstraint maximizer of L as in (7).

- As the network complexity increases, the curve of $L[\hat{f}_{\text{tr}}]$ goes up and approach the unconstrained maximum $L[f^*]$. For larger n the difference $L[f^*] - L[\hat{f}_{\text{tr}}]$ is smaller, indicating that increasing training size n reduces the influence of finite-sample. Since $L[\hat{f}_{\text{gl}}]$ lies between $L[f^*]$ and $L[\hat{f}_{\text{tr}}]$, this means that the stochastic optimization achieves a loss which is ΔC-close to that of $\hat{f}_{\text{gl}}$, supporting Assumption 1, and one may furtherly expect ΔC to be small at least when H is large.
- The trend of the increase of $L[\hat{f}_{\text{tr}}]$ as the network complexity increases behaves similarly for Example 1 and Example 2, which indicates that it is the the intrinsic geometry of the datasets that affects the efficiency of the network to produce a witness function with differential power.

The second observation that for on or near-manifold datasets, only the intrinsic geometric complexity influences the performance of neural network methods has been reported in experimental literature [54], [55] and approximation literature [23], [28], [29]. This motivates our theoretical work to reduce the needed network complexity to only scale with the intrinsic complexity of the densities p and q. Our result provides an explanation from the viewpoint of approximation and estimation error analysis.

III. Approximation of Near-Manifold Integrals

This section establishes a result that, for a near-manifold density p (described by exponential decay of p away from the manifold $\mathcal{M}$), the uniform approximation of a $\mathbb{R}^D$-Lipschitz function in $L^\infty(\mathcal{M})$ implies approximation in $L^1(\mathbb{R}^D, p)$ up to an error proportional to the scale of the exponential tail. This serves to resolve Step 2 in Section II-D, where, since only on-manifold approximation suffices, the network complexity scales with the intrinsic manifold dimensionality and the target function restricted to $\mathcal{M}$. The approximation in $L^1(\mathbb{R}^D, p)$ implied by that in $L^\infty(\mathcal{M})$ is also used in the estimation error analysis in Steps 3-5 in Section IV.

We call this result (Theorem III.1) the near-manifold integral approximation result. It is given in a slightly more general form than that of $L[f]$ as in (7), and we think the theorem can be of independent interest. An important result which is used in proving Theorem III.1 is Theorem III.2, which constructs the neural network approximation of the on-manifold function. All proofs are in Section VII.

A. The Integral Approximation Result

Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact smooth manifold of dimension d. We define $\mathcal{P}_\sigma$ to be the class of densities in $\mathbb{R}^D$ which decay exponentially fast away from the manifold $\mathcal{M}$, that is, for some small $0 < \sigma < 1$, and absolute constant $c_1 > 0$,

$$\mathcal{P}_\sigma = \{p \text{ density on } \mathbb{R}^D \text{ s.t. } \Pr_{X\sim p}[d(X, \mathcal{M}) > t] \le c_1 e^{-\frac{t}{\sigma}}\}, \tag{9}$$

where $d(x, \mathcal{M}) := \inf_{y\in\mathcal{M}} \|x-y\|_2$ for any $x \in \mathbb{R}^D$. We treat c_1 as fixed throughout the analysis and σ as the parameter indicating the size of the exponential tail.

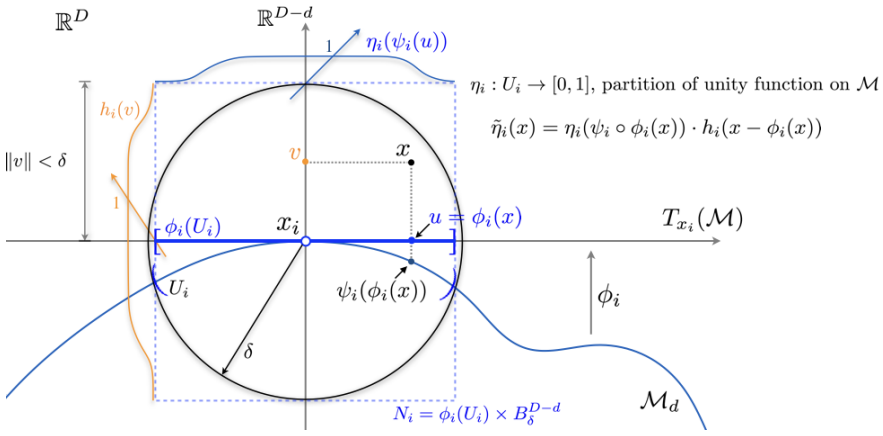


Fig. 3. A diagram showing a d-dimensional manifold $\mathcal{M}$ embedded in the ambient space $\mathbb{R}^D$ (only part of $\mathcal{M}$ is shown), an atlas $\{(U_i, \phi_i)\}_{i=1}^K$ which gives a δ-cover of $\mathcal{M}$, and the extended partition of unity function $\tilde{\eta}_i$ defined on the neighborhood N_i of U_i in $\mathbb{R}^D$.

The aim of the analysis is to approximately compute the integral

$$I[f] := \int_{\mathbb{R}^D} p(x) T(f(x)) dx, \tag{10}$$

by replacing f with a neural network function f_θ, where T is a 1D Lipschitz function so as to make the result more general, and $p \in \mathcal{P}_\sigma$ which concentrates near the manifold. Due to the exponential decay of the density p, we expect the integral $I[f]$ to be dominated by the contribution of the integral on the manifold. Indeed, restricting f to be on the manifold allows us to approximate $f|_\mathcal{M}$ by a neural network whose model complexity as approximation error $\to 0$ only depends on $f|_\mathcal{M}$ and the intrinsic manifold geometry. The bound of $|I[f] - I[f_\theta]|$, which involves integral in the ambient space, is then obtained by considering the corresponding on-manifold integrals of f and f_θ respectively, which gives the following Theorem.

The on-manifold function approximation by neural network is built upon the result in [28], which starts from an atlas on $\mathcal{M}$, and will be set up in Section III-B. The subscript in the notation 'Lip' denotes the domain on which the Lipschitz continuity is considered.

Theorem III.1: Suppose that $f : \mathbb{R}^D \to \mathbb{R}$ is Lipschitz on $\mathbb{R}^D$, $f|_\mathcal{M}$ is C^2 on the manifold, $T : \mathbb{R} \to \mathbb{R}$ is Lipschitz-1, and $p \in \mathcal{P}_\sigma$ with $\sigma < \frac{1}{2}$. Consider $I[f]$ as in (10), then for any $\varepsilon < 1$, there is a neural network architecture Θ with $O_{f,\mathcal{M}}(\varepsilon^{-d/2}) + N_0$ many trainable parameters, and a function $f_{con} \in \mathcal{F}_\Theta$ such that

$$|I[f_{\text{con}}] - I[f]| \le (1 + 2C_1(\mathcal{M})c_1\sigma)\varepsilon + C_3(f, \mathcal{M})c_1\sigma, \tag{11}$$

where

$$C_3(f, \mathcal{M}) := 2C_1(\mathcal{M})\|T \circ f\|_{L^\infty(\mathcal{M})} + C_2(\mathcal{M})(L_{\mathcal{M},f} + \text{Lip}_{\mathbb{R}^D}(f)), \tag{12}$$

$C_1(\mathcal{M})$, $C_2(\mathcal{M})$ are as in (19) and only depending on the manifold-atlas, the meaning of $O_{f,\mathcal{M}}(\cdot)$, N_0 and $L_{\mathcal{M},f}$ is the same as in Theorem III.2 (noting that f in Theorem III.2 is $f|_\mathcal{M}$ here). In particular, none of the three changes with ε, and N_0 is also independent of f but involves manifold-atlas and D.

Remark III.1: f in Theorem III.2 is $f|_\mathcal{M}$ here, and f_N in Theorem III.2 is f_{con} here. The existence of f_{con} is provided by Theorem III.2 with the stated properties therein.

The proof of Theorem III.1 combines the integral comparison for Lipschitz functions on $\mathbb{R}^D$ in Proposition III.5, and the on-manifold function approximation in Theorem III.2. In Theorem III.2, it is proved that $\text{Lip}_{\mathbb{R}^D}(f_{con}) \le L_{\mathcal{M},f}$ using the wavelet construction of f_{con}. In Section IV, we will introduce a Lipschitz regularization of the network family (Assumption 3), and further bound the ∞-norm of $T \circ f$ on $\mathcal{M}$ in (12) in the two-sample testing context.

B. Manifold Atlas and On-Manifold Function Approximation

We first establish some notations for the manifold and atlas cover. The manifold $\mathcal{M}$ can be covered with an atlas $\{(U_i, \phi_i)\}_{i=1}^K$, where $U_i = B(x_i, \delta) \cap \mathcal{M}$ is an open set on $\mathcal{M}$, $0 < \delta < 1$, the choice of which to be specified below. The orthogonal projection $\phi_i : U_i \to \mathbb{R}^d$ is the map that takes U_i to the tangent plane $T_{x_i}(\mathcal{M})$. We also define the map $\psi_i : \phi_i(U_i) \to U_i$, which is the inverse of ϕ_i due to the one-to-one correspondence between U_i and $\phi_i(U_i)$. Let $d_\mathcal{M}$ denote the manifold geodesic distance. In the construction, the Euclidean ball radius δ is chosen to be small enough such that $B(x, 2\delta) \cap \mathcal{M}$ is isomorphic to a ball in $\mathbb{R}^d$ and there exist positive α_i and β_i s.t. $\forall x, x' \in U_i$,

$$\alpha_i \|\phi_i(x) - \phi_i(x')\|_2 \le d_\mathcal{M}(x, x') \le \beta_i \|\phi_i(x) - \phi_i(x')\|_2, \tag{13}$$

and for all i,

$$\alpha_i \ge \alpha_\mathcal{M} \ge \frac{1}{2}, \quad 1 \le \beta_i \le \beta_\mathcal{M} \le 2. \tag{14}$$

To see why this is possible: for any point $x \in \mathcal{M}$, there is a $\delta_x > 0$ and constants $\frac{1}{2} < \alpha_x \le 1 \le \beta_x < 2$ such that whenever $\delta' < \delta_x$, the relation (13) with constants β_x, α_x holds on the neighborhood $U_x := B(x, \delta') \cap \mathcal{M}$, and at the same time $B(x, 2\delta') \cap \mathcal{M}$ is isomorphic to a ball in $\mathbb{R}^d$. This is because that the manifold is locally near Euclidean, which means that β_x, α_x can be made close to 1 if $\delta' \to 0+$.

The existence of δ_x and β_x, α_x is due to manifold smoothness and the finiteness of the local curvature near x. The $\inf_{x\in\mathcal{M}} \delta_x$ exists and is positive due to compactness and smoothness of $\mathcal{M}$. Setting that (and the minimum with 1) as δ for all x leads to a finite cover of $\mathcal{M}$ which is $\{U_i\}_{i=1}^K$ with constants α_i, β_i on each U_i, and then the global $\alpha_{\mathcal{M}}, \beta_{\mathcal{M}}$ exist. Note that while the atlas and particularly the radius δ depend on the embedding of $\mathcal{M}$ in the ambient space $\mathbb{R}^D$, the atlas remains valid if D increases to D', $\mathcal{M}$ is isometrically embedded into $\mathbb{R}^{D'}$, and the reach of the manifold is maintained [56], e.g., when $\mathbb{R}^D$ is isometrically embedded in $\mathbb{R}^{D'}$. Thus we view any quantity which only depends on the δ-atlas as intrinsic to the manifold geometry.

Given the covering atlas, there exists a partition of unity $\{\eta_i\}_{i=1}^K$ on $\mathcal{M}$ such that $\text{supp}(\eta_i) \subset U_i$, $\eta_i \in C^\infty(\mathcal{M})$, and $\sum_{i=1}^K \eta_i(x) = 1$ for all $x \in \mathcal{M}$. Under this setting, the following theorem constructs network function f_{con} (named as f_N) with the approximating property. The subscript N stands for the number of network parameters.

Theorem III.2: Notations being as above, suppose $f \in C^2(\mathcal{M})$, then for any $\varepsilon < 1$, there exists a four layer feed-forward network with rectified linear unit activations and $N = O_{f,\mathcal{M}}(\varepsilon^{-d/2}) + N_0$ parameters, such that the network function $f_N : \mathbb{R}^D \to \mathbb{R}$ satisfies that

$$\|f - f_N\|_{L^\infty(\mathcal{M})} \le \varepsilon,$$

where the constant in $O_{f,\mathcal{M}}(\cdot)$ only depends on f and manifold-atlas, specifically, it is $(d+K)\delta^d(KC_{f,\eta})^{d/2}$, K being the number of coverings in the δ-atlas, the constant $C_{f,\eta}$ depending on d and up to 2nd manifold-derivatives of f and the partition of unity functions of the δ-atlas. $N_0 = C(KdD + \frac{D}{\delta}\log\frac{D}{\delta})$, C being an absolute constant, and N_0 does not depend on f nor change with ε.

Furthermore, the constructed network function f_N is globally Lipschitz and $Lip_{\mathbb{R}^D}(f_N) \le L_{\mathcal{M},f}$, which is a constant depending on f and the manifold-atlas, but independent of ε and N.

The proof is by constructing a sub-network function $\bar{f}_{N,i}$ on a neighborhood $N_i \subset \mathbb{R}^D$ around each $U_i \subset \mathcal{M}$ to approximate $f\eta_i$ respectively, and then taking a sum over i using the partition of unity property. Specifically, for each U_i, the neighborhood N_i is defined as

$$N_i := \phi_i(U_i) \times B_\delta^{D-d}, \quad B_\delta^{D-d} := \{x \in \mathbb{R}^{D-d}, \|x\|_2 < \delta\}, \tag{15}$$

thus $\phi_i(N_i) = \phi_i(U_i)$, where we also denote ϕ_i as the orthogonal projection from $\mathbb{R}^D$ to the tangent space $T_{x_i}(\mathcal{M})$. The relation (13)(14) gives the following lemma,

Lemma III.3: For any $x \in U_i$, $\|x - \phi_i(x)\| < \delta\sqrt{1-\frac{1}{\beta_i^2}} \le \frac{\sqrt{3}}{2}\delta$.

The sub-network function $\bar{f}_{N,i}(x)$, $x = (u, v)$ in local coordinates, is constructed by first approximating $f\eta_i(\psi_i(u))$, viewed as a function on $\mathbb{R}^d$, by some $\hat{f}_{N,i}(u)$, $u \in \mathbb{R}^d$, and then extending constantly into N_i for a distance $< \delta$ by a tent-like function $g_i(v)$ on $\mathbb{R}^{D-d}$ which is 1 when $\|v\| < \frac{\sqrt{3}}{2}\delta$ and 0 when $\|v\| \ge 1$. Lemma III.3 guarantees that $\bar{f}_{N,i}$ restricted to $x \in U_i$ equals $\hat{f}_{N,i}(u)$ because $g_i(v) = 1$ on U_i. Thus f_N by taking a sum $\sum_{i=1}^K \bar{f}_{N,i}$ is uniformly approximating f on the manifold. The Lipschitz constant of f_N is proved by considering that of $\hat{f}_{N,i}$ and g_i respectively which bounds each $\text{Lip}_{\mathbb{R}^D}(\bar{f}_{N,i})$. The statement regarding the global Lipschitz continuity of f_N is a byproduct of the construction in [28], but was not explicitly stated therein. For completeness, we provide a proof of the Theorem in Section VII.

C. Comparison of on and Near-Manifold Integrals

As we would like to analyze the integral of near manifold density in the ambient space, we extend the partition of unity function η_i to the neighborhood N_i defined as in (15) as

$$\tilde{\eta}_i(x) \ = \eta_i(\psi_i \circ \phi_i(x)) \cdot h_i(x - \phi_i(x)), \tag{16}$$

where $h_i(x)$ is continuous on $\mathbb{R}^{D-d}$, vanishes outside B_δ^{D-d}, $0 \le h_i \le 1$ and

$$h_i(x) = 1 \text{ if } \|x\| \le \delta\sqrt{1-\frac{1}{\beta_i^2}}, \quad \text{Lip}(h_i) < \frac{2\beta_i^2}{\delta}. \tag{17}$$

This can be constructed, e.g., by $h_i(x) = g(\|x\|/\delta)$ where $g(r) = 1$ on $[0, \sqrt{1-\frac{1}{\beta_i^2}}]$, 0 outside $r > 1$, and linearly interpolating in between. This construction guarantees the following properties of the extended function $\tilde{\eta}_i$'s:

Lemma III.4: For $i = 1, \cdots, K$, $\tilde{\eta}_i$ as a function in $\mathbb{R}^D$ vanishes outside N_i, equals η_i on U_i, is Lipschitz continuous on $\mathbb{R}^D$, and, for all i,

$$\text{Lip}_{\mathbb{R}^D}(\tilde{\eta}_i) = \text{Lip}_{N_i}(\tilde{\eta}_i) \le L_{\mathcal{M}},$$

where $L_{\mathcal{M}}$ is an absolute constant depending on the manifold-atlas.

We then establish a key result used in the proof of Theorem III.1, which compares the ambient space integral to a "projected" integral on the manifold.

Proposition III.5 (Integral Comparison): Notations of manifold-atlas and partition of unity functions as above. Suppose $g : \mathbb{R}^D \to \mathbb{R}$ is Lipschitz continuous on $\mathbb{R}^D$, $p \in \mathcal{P}_\sigma$ with $\sigma < \frac{1}{2}$, and define

$$\tilde{p}(x) = \sum_{i=1}^K \eta_i(x)\tilde{p}_i(x),$$

where $\tilde{p}_i$ is an atlas dependent "projection" of the density p to U_i, the explicit formula to be given below, then

$$\left|\int_{\mathbb{R}^D} g(x)p(x)dx - \int_{\mathcal{M}} g(x)\tilde{p}(x)d_{\mathcal{M}}(x)\right| \le \left(\|g\|_{L^\infty(\mathcal{M})}C_1(\mathcal{M}) + \text{Lip}_{\mathbb{R}^D}(g)C_2(\mathcal{M})\right)c_1\sigma, \tag{18}$$

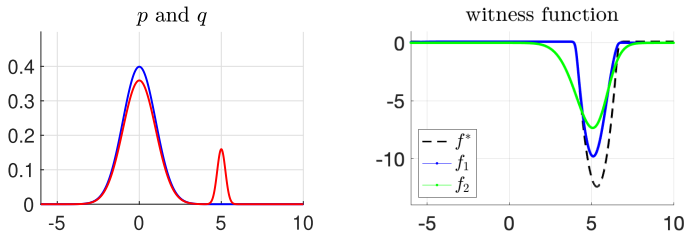
where c_1 as in the definition of $\mathcal{P}_\sigma$ (9),

$$C_1(\mathcal{M}) = 3KL_{\mathcal{M}}, \quad C_2(\mathcal{M}) = K(2L_{\mathcal{M}} + 1 + \beta_{\mathcal{M}}), \tag{19}$$

$L_{\mathcal{M}}$ as in Lemma III.4, and $\beta_{\mathcal{M}}$ as in (14). $C_1(\mathcal{M})$ and $C_2(\mathcal{M})$ are absolute constants determined by the manifold and atlas.

In particular, taking $g = 1$, the proposition gives that

$$\left|\int_{\mathcal{M}} \tilde{p}(x)d_{\mathcal{M}}(x) - 1\right| \le C_1(\mathcal{M})c_1\sigma, \tag{20}$$



	$L[f]$	$T[f]$
f^*	35.93 $\times 10^{-3}$	1.1221
f_1	35.92 $\times 10^{-3}$	0.9241
f_2	33.00 $\times 10^{-3}$	0.7067

Fig. 4. (Plots) Left plot shows 1D densities p and q. The middle plot shows the log density ratio f^* and two examples f_1 and f_2 of the smooth and compactly-supported surrogate function f_{tar}, which are of different regularity. (Table) The first column shows the values of $L[f]$ for f^* and two surrogates f_{tar} respectively. $L[f^*] = \text{JSD}(p, q)$. The second column shows the values of $T[f]$.

which means that the constructed $\tilde{p}$ is close to being integral 1 on $\mathcal{M}$ up to an error proportional to σ.

The proof of the main Theorem III.1 then follows by combining Proposition III.5 and Theorem III.2.

IV. Consistency of Network Logit Test

In this section, we fulfill the steps listed in Section II-D to prove the theoretical guarantee of the network logit two-sample testing. The manifold setting of densities reduces the dimensionality from D to d in both the approximation error (Step 2) and the estimation error analysis (Steps 3-5). All proofs are in Section VII.

A. Settings of Data Densities and Step 1. f_{tar}

We consider the following three settings of densities p and q,

- (General setting) Densities in $\mathbb{R}^D$ with sub-exponential tail. This includes compactly supported densities, which are mostly encountered in practice.
- (On-manifold setting) Densities constrained on a smooth compact manifold $\mathcal{M} \subset \mathbb{R}^D$.
- (Near-manifold setting) Densities which are exponentially decay away from the manifold $\mathcal{M}$, as defined in (9) with some positive c_1 and small σ.

The analysis is under the same framework, where bifurcations take place in the integral approximation and bounding of the estimation error. We consider the class of sub-exponential densities in $\mathbb{R}^D$ as

$$\mathcal{P}_{\text{exp}} = \{p \text{ density on } \mathbb{R}^D \text{ s.t. } \Pr_{X\sim p}[\|X\| > t] \leq Ce^{-t/c}\} \tag{21}$$

for some $C, c > 0$. One can always rescale the space to make p and q supported on a diameter-$O(1)$ domain, when compactly supported, or $c = 1$, when exponentially decay, even when the dimension D is large. This corresponds to normalizing the data vectors to be of $O(1)$ norm in practice. By this normalizing argument, we assume $\text{supp}(p + q)$ has diameter-$O(1)$ when compactly supported and generally exponentially decay with $c = 1$ in below.

We start from Step 1. in Section II-D, and provide motivation, construction and examples of f_{tar} which is assumed in Assumption 2. A sub-exponential density p may vanish at certain points in $\mathbb{R}^D$ or have discontinuity, e.g., when $\text{supp}(p)$ is compact. This makes $f^* = \log \frac{p}{q}$ possibly diverge at a point, e.g., when $\text{supp}(p)$ and $\text{supp}(q)$ partially do not overlap. Observe that unless $p = q$, one can always restrict to the interior of $\text{supp}(\frac{p+q}{2})$ and consider a bounded version of f^*, e.g., $f = \min\{\max\{f^*, M\}, -M\}$ for $M > 0$, such that $L[f]$ as in (7) is close to $\text{JSD}(p, q)$ and > 0. Furthermore, $L[f]$ can be written as

$$L[f] = \frac{1}{2}\left(\int_{\mathbb{R}^D} pT_p \circ f + \int_{\mathbb{R}^D} qT_q \circ f\right),$$
$$T_p(\xi) = \log\frac{2e^{\xi}}{1+e^{\xi}}, \quad T_q(\xi) = \log\frac{2}{1+e^{\xi}}, \tag{22}$$

where $T = T_p$ and T_q all satisfy that

$$T : \mathbb{R} \to \mathbb{R} \text{ is Lipschitz and } \text{Lip}(T) < 1, \text{ and } T(0) = 0. \tag{23}$$

As a result, approximating the bounded function f under $L^1(p)$ and $L^1(q)$ will approximate the integral $L[f]$. Thus we can choose the approximator of f to be smooth, and, by the sub-exponential decay of p and q, to be compactly supported. According to these procedures, one can construct a f_{tar} to be a smooth surrogate of f^* as in Assumption 2, satisfying that $L[f_{\text{tar}}]$ is sufficiently close to $L[f^*] = \text{JSD}(p, q) > 0$ and thus is also strictly positive. We thus call $L[f_{\text{tar}}]$ as $C > 0$ throughout the paper. In particular, if $(p + q)$ are compactly supported, then $\text{supp}(f_{\text{tar}})$ can be made within that domain. For general sub-exponential densities, the diameter of $\text{supp}(f_{\text{tar}})$ can be made proportional to the scale parameter c in $\mathcal{P}_{\text{exp}}$. By the normalizing argument above, the diameter of $\text{supp}(f_{\text{tar}})$ is $O(1)$. Since the above constructive procedures are standard, Assumption 2 is mild under our setting.

A example of f^* and f_{tar} is shown in Figure 4, where p is $\mathcal{N}(0, 1)$ and q is a Gaussian mixture in 1D. The f^* in this case is non-vanishing in $\mathbb{R}$, and has a peak around $x = 5$. The two examples of f_{tar} gives $L[f]$ significantly large compared to $L[f^*]$, as shown in the table. This example also shows that there are more than one choice of f_{tar} to fulfill Assumption 2. Generally, one may trade-off between the regularity of f_{tar} and making C arbitrarily close to $L[f^*]$. We further discuss this in Remark IV.1.

B. Step 2. Neural Network Approximation

The neural network approximation theory provides f_{con} which uniformly approximates f_{tar} on a compact domain Ω. We consider the three settings of densities respectively.

1) General Setting: When p and q are general sub-exponential densities on $\mathbb{R}^D$, we make use of the compact-supportedness of f_{tar} as in Assumption 2, and set it as the target function to approximate.

Proposition IV.1: Suppose that $p \neq q$, $\text{JSD}(p,q) > 0$, and p, q are in $\mathcal{P}_{\text{exp}}$ with $c = 1$. Under Assumption 2, there exists a smooth and compactly supported f_{tar} with $L[f_{\text{tar}}] = C > 0$, and diameter of $\text{supp}(f_{\text{tar}}) < C'$. Then for any $\varepsilon < 1$, there is a neural network architecture Θ with $O_{f_{tar},C'}(\varepsilon^{-D/r} \log \frac{1}{\varepsilon})$ many trainable parameters, and $f_{con} \in \mathcal{F}_\Theta$ such that

$$L[f_{\text{con}}] \geq C - \varepsilon > 0,$$

where the constant in $O_{f_{tar},C'}(\cdot)$ depends on up to the r-th derivative of f_{tar}, the diameter C', r and D, $r \geq 1$. When p and q are compactly supported, C' is the diameter of $\text{supp}(p+q)$.

As previously commented beneath Assumption 2, C' is an $O(1)$ constant even when dimensionality D can be large, thus the network complexity remains $O(\varepsilon^{-D/r})$. The proof of Proposition IV.1 is by a direct application of the standard network approximation theory, e.g. that in [25], where the sub-exponential tail of the distribution does not affect because the constructed network function can be made supported on the same domain of $\text{supp}(f_{\text{tar}})$. More recent approximation result which improves the approximation rate, such as [57], will improve the complexity needed accordingly. We comment in Remark IV.1 about the choice of f_{tar} in Assumption 2 and the r used in Proposition IV.1, These trade-offs can be made precise in specific settings of the densities, and we leave the choices abstract here.

Remark IV.1 (Trade-Off in the Choice of f_{tar} and Regularity Level r): There are two trade-offs in applying Proposition IV.1. First, generically (e.g., the two densities are absolutely continuous) one can make the constant C arbitrarily close to $\text{JSD}(p,q)$, and as a result the regularity of f_{tar} may worse, and then the needed network complexity will increase. Second, the regularity level r can be chosen to be large given that $f_{\text{tar}} \in C_c^\infty(\Omega)$, but the constant in the network complexity bound will grow with higher order r. We illustrate the first trade-off in Figure 4: in this example, q is significantly large on a region where p almost vanish, resulting in the log density ratio f^* having a peak in that region. Such a singularity may create a difficulty for network to approximate f^*. Replacing f^* by a regularized surrogate f_{tar} such as f_1, one can produce a significantly large $L[f_{\text{tar}}] = C > 0$ which allows Theorem IV.6 to apply, and at the same time have more efficient (theoretical) network approximation and possibly more efficient estimation in practice. By choosing f_{tar} of better regularity, such as f_2, the gap between C and $\text{JSD}(p,q)$ is larger and C is smaller, but the network approximation and estimation may be even better. In addition, using f_1 or f_2 as witness function may not necessarily sacrifice testing power because more regular f can reduce the variance of the test statistic (though reducing the bias $T[f]$ at the same time) and both bias and variance affect the test power, as suggested by the CLT result in Theorem IV.6.

2) On-Manifold Setting: When p and q are degenerate and constrained to a compact smooth manifold $\mathcal{M}$ in $\mathbb{R}^D$, the integral $L[f]$ is carried out on the manifold only. Replacing the Euclidean metric with the Riemannian geometry on $\mathcal{M}$, and the integral in $\mathbb{R}^D$ with integration on $\mathcal{M}$ with Riemannian volume element, the choice of a smooth f_{tar} with $L[f_{\text{tar}}] = C > 0$ as in Assumption 2 extends. This gives the on-manifold-version of Proposition IV.1, where the classical network approximation is replaced with Theorem III.2, which guarantees the existence of f_{con} such that

$$\|f_{\text{con}} - f^*\|_{L^\infty(\mathcal{M})} \leq \varepsilon,$$

where the needed network complexity of $O(\varepsilon^{-d/2})$, namely reducing the exponent factor from D to the intrinsic dimensionality d. The proof of Proposition IV.1 directly extends by replacing integration on Ω with that on $\mathcal{M}$ and the rest is the same.

3) Near-Manifold Setting: When the densities decays sub-exponentially away from the manifold, since $\mathcal{M}$ is compact, the densities belong to be sub-exponential family as in the general setting. While Proposition IV.1 still applies, the needed network complexity is not intrinsic. We apply the analysis in Section III to improve this.

Proposition IV.2: Suppose that $p \neq q$, $\text{JSD}(p,q) > 0$, and $p, q \in \mathcal{P}_\sigma$ as defined in (9) with $\sigma < \frac{1}{2}$. Under Assumption 2, there exists a smooth and compactly supported f_{tar} with $L[f_{\text{tar}}] = C > 0$, and there is one point $x_0 \in \mathcal{M}$ such that $f_{\text{tar}}(x_0) = 0$. Then for any $\varepsilon < 1$, there is a neural network architecture Θ with $O_{f_{tar},\mathcal{M}}(\varepsilon^{-d/2}) + N_0$ many trainable parameters, and $f_{con} \in \mathcal{F}_\Theta$ such that

$$L[f_{\text{con}}] \geq C - \Big(\varepsilon \cdot (1 + 2\ c_1 \sigma C_1(\mathcal{M})) + \sigma \cdot c_1 \left(C_4(\mathcal{M}) \text{Lip}_{\mathbb{R}^D}(f_{\text{tar}}) + C_2(\mathcal{M}) L_{\mathcal{M},f_{\text{tar}}} \right) \Big), \quad (24)$$

where

$$C_4(\mathcal{M}) := 2C_1(\mathcal{M}) \text{diam}(\mathcal{M}) + C_2(\mathcal{M}),$$

and C_1, C_2 as in (19), the constants C_1, C_2, C_4 only depend on the manifold-atlas. The meaning of $O_{f_{tar},\mathcal{M}}(\cdot)$, N_0, $L_{\mathcal{M},f_{\text{tar}}}$ is the same as in Theorem III.1 taking $f = f_{\text{tar}}$.

Remark IV.2: About the condition that f_{tar} vanishes at one point on $\mathcal{M}$: Since $p \neq q$, and by construction f_{tar} is the smooth surrogate of $f^* = \log \frac{p}{q}$, then f_{tar} vanishes at least at one point in $\mathbb{R}^D$. When both p and q are concentrating near the manifold $\mathcal{M}$, it generically holds that f_{tar} vanishes at least at one point on the manifold. Thus the condition is mild and does not pose a constraint.

The proposition bounds $C - L[f_{\text{con}}]$ to be $O(\varepsilon) + O(\sigma)$, where the integral of $L[f_{\text{tar}}]$ is approximated by constructing on-manifold approximation of the target function f_{tar} only, and the network complexity is reduced to be intrinsic. By (24), $L[f_{\text{con}}]$ is close to C when ε and σ are sufficiently small.

C. Concentration of L_n and Steps 3-5

Steps 3 and 5 are based on the concentration of $L_n[f_\theta]$ close to $L[f_\theta]$ when $f_\theta = f_{\text{con}}$ or trained on the training set. We omit subscript "tr" in L_n, which emphasizes that L_n is the empirical loss on the training samples. We will upper bound,

under proper Lipschitz regularization of network function class $\mathcal{F}_\Theta$, that

$$\sup_{f\in\mathcal{F}_\Theta} |L_n[f] - L[f]| \leq ?\text{w.h.p. for sufficiently large } n, \quad (25)$$

where w.h.p. stands for "with high probability" and will be made precise. For the three settings of densities, we prove that the bound in (25) is $\tilde{O}(n^{-\frac{1}{2+D}})$ for $\mathcal{P}_{\text{exp}}$ densities in $\mathbb{R}^D$, $\tilde{O}(n^{-\frac{1}{2+d}})$ for on-manifold densities, and $\tilde{O}(n^{-\frac{1}{2+d}}) + O(\sigma)$ for near-manifold densities in $\mathcal{P}_\sigma$, $\tilde{O}$ meaning that the constant may involve $\log n$ (Proposition IV.5). We first introduce the needed Lipschitz regularization of network functions, which leads to a bound of the covering number of $\mathcal{F}_\Theta$ that is used in the concentration analysis.

1) Lipschitz Regularization of Network Functions: When neural network is using ReLU or other continuous nonlinear activations, the network function is typically differentiable or piece-wise differentiable on $\mathbb{R}^D$, and is globally Lipschitz because all the weights are finite. However, $\text{Lip}_{\mathbb{R}^D}(f_\theta)$ for a member $f_\theta \in \mathcal{F}_\Theta$ may potentially be large. In the network approximation analysis, Theorem III.2 shows that as the approximation error ε gets small the constructed network function f_{con} to approximate f is globally Lipschitz with $\text{Lip}_{\mathbb{R}^D}(f_{\text{con}}) \leq L_{\mathcal{M},f}$, a constant only depending on manifold-atlas and $f|_{\mathcal{M}}$. This means that when the target function is smooth, constraining on bounded Lipschitz constant of f_θ does not prevent the network approximation of f to high accuracy. This approximation result however does not mean that the trained network function $\hat{f}_{n,tr}$ has certain bounded Lipschitz constant automatically. $\hat{f}_{n,tr}$ with poor Lipschitz regularity may incure larger variance of the statistic with finite samples. Applying regularization to the network function balances between bias and variance, and is commonly used in practice.

We thus consider network function families with Lipschitz regularization as assumed in Assumption 3. The universal Lipschitz constant bound L_Θ can be viewed as a network hyperparameter, and we call the restricted family $\mathcal{F}_{\Theta,\text{reg}}$.

2) Covering Number of Regularized $\mathcal{F}_\Theta$: The Lipschitz regularization enables bounding of the covering number of function space by the covering number of the domain when compact. For a compact set $K \subset \mathbb{R}^D$, define its covering number as

$$\mathcal{N}(K,r) := \inf\{\text{Card}(S),\, K \subset \bigcup_{s\in S} \bar{B}_r(x)\}, \quad r > 0.$$

where "Card" stands for cardinal number, $\bar{B}_r(x)$ is the Euclidean closed ball $\{y, \|y - x\| \leq r\}$, and the requirement of S is equivalent to that S is an r-net of K. The general covering number is denoted as $\mathcal{N}(X, r, \|\cdot\|)$, where X is the set to be covered, and r is the radius of the closed $\|\cdot\|$-ball.

Lemma IV.3: Let K be a compact set in $\mathbb{R}^D$ with covering number $\mathcal{N}(K,r)$, $r > 0$. Consider the function class

$$\mathcal{F} := \{f : \mathbb{R}^D \to \mathbb{R},\, \text{Lip}_{\mathbb{R}^D}(f) \leq L,\, \|f\|_{L^\infty(K)} \leq B\},$$

where $L > 0$, $B > 0$. Then for any subset $\mathcal{F}'$ of $\mathcal{F}$ (including $\mathcal{F}' = \mathcal{F}$), and any $0 < r < \frac{B}{L}$, there is a finite set $F \subset \mathcal{F}'$ which forms an $(4Lr)$-net of $\mathcal{F}'$, i.e., $\mathcal{F}' \subset \bigcup_{f\in F} \bar{B}(f, 4Lr, \|f\|_{L^\infty(K)})$ and

$$\text{Card}(F) \leq \left(\frac{2B}{Lr}\right)^{\mathcal{N}(K,r)}. \quad (26)$$

The lemma proves an upper bound for the covering number of the whole class $\mathcal{F}$, which will contain regularized network function class as a subset. When applying to analyzing the general and on-or-near manifold densities, the K will be a ball in $\mathbb{R}^D$ and the compact manifold respectively, the covering number $\mathcal{N}(K)$ then differs in a factor of r^{-D} v.s. that of r^{-d}.

3) Concentration of L_n Sup Over Network Function Family: The concentration of $L_n[f]$ for a single Lipschitz f is a direct result of Bernstein's inequality for sub-exponential random variables. Specifically, the following lemma, proved in Appendix A, verifies that the random variables $T \circ f(x_i)$ are sub-exponential due to that the density of x_i is in $\mathcal{P}_{\text{exp}}$.

Lemma IV.4: Suppose that $T : \mathbb{R} \to \mathbb{R}$, $\text{Lip}(T) \leq 1$, $f : \mathbb{R}^D \to \mathbb{R}$, $\text{Lip}_{\mathbb{R}^D}(f) \leq L$, $x_i \sim p$, $i = 1, \cdots, n$, i.i.d., and p is sub-exponential on $\mathbb{R}^D$, i.e. $p \in \mathcal{P}_{\text{exp}}$ with $c = 1$, then $\xi_i := T(f(x_i))$ are i.i.d 1D sub-exponential random variables, and specifically,

$$\Pr[|\xi_i - \mathbb{E}\xi_i| > t] \leq C' e^{-c'\frac{t}{L}}, \quad \forall t > 0, \quad (27)$$

where C' and c' are absolute positive constants.

Bernstein's inequality for sub-exponential random variables (Corollary 2.8.3 in [58]) then gives that

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n T(f(x_i)) - \mathbb{E}_{x\sim p} T(f(x))\right| \geq t\right] \leq 2\exp\left\{-c_0' n \frac{t^2}{L^2}\right\}, \quad \forall 0 < t < c_1' L, \quad (28)$$

where c_0', c_1' are absolute positive constant. We will control the sup over network function family by a covering argument. Define the regularized neural network function class for architecture Θ as

$$\mathcal{F}_{\Theta,\text{reg}}(\Omega) = \{f \in \mathcal{F}_\Theta,\, \text{Lip}_{\mathbb{R}^D}(f) \leq L,\, \exists x_0 \in \Omega, f(x_0) = 0\}, \quad (29)$$

where the dependence on Ω is only via the assumption on the location of a vanishing point of f. Under Assumption 3, L equals the universal constant L_Θ, the subscript Θ is omitted here for simplicity.

Proposition IV.5: Let B_R denote the ball in $\mathbb{R}^D$ centering at origin, $R \geq 1$. For the three density settings, suppose that

(1) General. $p, q \in \mathcal{P}_{\text{exp}}$ with $c = 1$. When p and q are compactly supported, $\text{supp}(p+q) \subset B_R$.
(2) On-manifold. Case (1) plus that $\text{supp}(p+q) \subset \mathcal{M} \subset B_R$ in $\mathbb{R}^D$.
(3) Near-manifold. $\mathcal{M}$ as in (2), case (1) plus that $p, q \in \mathcal{P}_\sigma$ as defined in (9) with $\sigma < \frac{1}{2}$.

Suppose that the network function family, denoted as $\mathcal{F}_{\Theta,\text{reg}}$ for all cases, is $\mathcal{F}_{\Theta,\text{reg}}(B_R)$ for case (1), and $\mathcal{F}_{\Theta,\text{reg}}(\mathcal{M})$

for cases (2) (3). Then, when n is sufficiently large, with probability $\to 1$ as $n \to \infty$,

$$\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}} |L_n[f] - L[f]| \leq \begin{cases} \tilde{C} \cdot L \log n (\log n/n)^{1/(2+D)}, & \text{case (1)} \\ \tilde{C}(\mathcal{M}) \cdot L (\log n/n)^{1/(2+d)}, & \text{case (2)} \\ \tilde{C}(\mathcal{M}) \cdot L(\sigma + (\log n/n)^{1/(2+d)}), & \text{case (3)} \end{cases}$$

where $\tilde{C}(\cdot)$ refers to a positive constant that may depend on $\cdot$, without $(\cdot)$ an absolute constant, and the notation stands for different constants in different cases. In case (1), if p, q are compactly supported, the bound can improve to $\tilde{C}(R) \cdot L(\log n/n)^{1/(2+D)}$ removing a $\log n$ factor. The $\to 1$ probability is exponentially high except for the non-compactly supported case in (1).

In case (1), though the densities may have sub-exponential tail, it is generic to assume that exists $x_0 \in B_R$ such that $p(x_0) = q(x_0)$. Since $\log \frac{p}{q}$ vanishes at least at one point inside B_R, we consider network functions that also have this property. The condition that $R \geq 1$ is only for convenience, and does not pose any constraint to our setting. Note that the estimation error in cases (2) improves from the generic case (1) by reducing the D to the intrinsic dimensionality d, and in the near-manifold case, the bound adds another term of $O(\sigma)$, similar to the result of the approximation error. The analysis of case (3) again uses the integral comparison technique in Section III.

As a remark, if p, q have compact support Ω which has a smaller volume than B_R^D, then the covering number $\mathcal{N}(\Omega, r)$ can be made $< \mathcal{N}(B_R^D, r)$, the latter leading to the $-\frac{1}{2+D}$ rate in case (1). This is the fundamental reason that the rate can be improved to $-\frac{1}{2+d}$ for on or near manifold densities in cases (2) and (3). Our proof technique bounds the estimation error by the covering complexity of the "essential" support of the densities, allowing certain sub-exponential tail, which can be viewed as capturing the essential complexity of the densities. The manifold setting is a special case that is natural in applications.

D. Testing Power and Consistency of Network-Logit Test

We have now finished the 5 steps in Section II-D, which allows to establish that (Theorem IV.6(1))

$$L[\hat{f}_{\text{tr}}] > L[f_{\text{tar}}] - L_{\text{gap}} > 0, \tag{30}$$

where $L_{\text{gap}} < L[f_{\text{tar}}] = C$, and

$$L_{\text{gap}} = \Delta C + O(\varepsilon) + O(\sigma) + \tilde{O}(n_{\text{tr}}^{-1/(2+d')}), \tag{31}$$

$$d' = \begin{cases} D, & \text{when } p, q \text{ are general sub-exponential densities in } \mathbb{R}^D, \\ d, & \text{when } p, q \text{ are on or near a } d\text{-dimensional manifold } \mathcal{M}. \end{cases}$$

Recall that f_{tar} is a smooth surrogate of $f^* = \log \frac{p}{q}$, and C can approach $L[f^*]$ =JSD(p, q) if f^* itself is regular. In (31), ε is the network approximation error, σ is the sub-exponential decay scale of near-manifold densities (and does not appear for general or on-manifold density settings), ΔC is the optimization error (Assumption 1), and the last term which decays with n_{tr} is the estimation error.

When $L_{\text{gap}} < C$, (30) guarantees that $L[\hat{f}_{\text{tr}}] > 0$. With an elementary lemma showing that $T[f] \geq 4L[f]$, which is Lemma VII.8 (the tightness of the relaxation is explained in the remark below and in Lemma VII.9), it gives that $T[\hat{f}_{\text{tr}}] > 0$. This sets a strictly positive expectation of the test statistic $\hat{T}$. The testing consistency then follows by standard CLT, after proving a bounded variance of the test statistic.

We are ready to prove the main theorem of this section:

Theorem IV.6: Notations being as above, under Assumptions 1, 2, 3, and the generic settings in Proposition IV.5. Let $\hat{f}_{\text{tr}}$ be the trained network function from n_{tr} many samples of X and Y, and T_n be the test statistic evaluated on the testing set where $|X_{\text{te}}| = |Y_{\text{te}}| = n$. If ΔC, ε, σ are sufficiently small and n_{tr} sufficiently large such that L_{gap} in (31) is less than C, then

(1) When $p \neq q$, $\mathbb{E}T_n = T[\hat{f}_{\text{tr}}] > 0$, and specifically $\mathbb{E}T_n > 4(C - L_{\text{gap}})$.
(2) When $p = q$, $\sqrt{n}T_n \to \mathcal{N}(0, \sigma^2_{\mathcal{H}_0})$ in distribution. When $p \neq$ q, $\sqrt{n}(T_n - \mathbb{E}T_n) \to \mathcal{N}(0, \sigma^2_{\mathcal{H}_1})$ in distribution. $\sigma_{\mathcal{H}_0}$ and $\sigma_{\mathcal{H}_1}$ are all bounded by L_Θ multiplied by an absolute constant.

Furthermore, the needed network complexity is bounded by

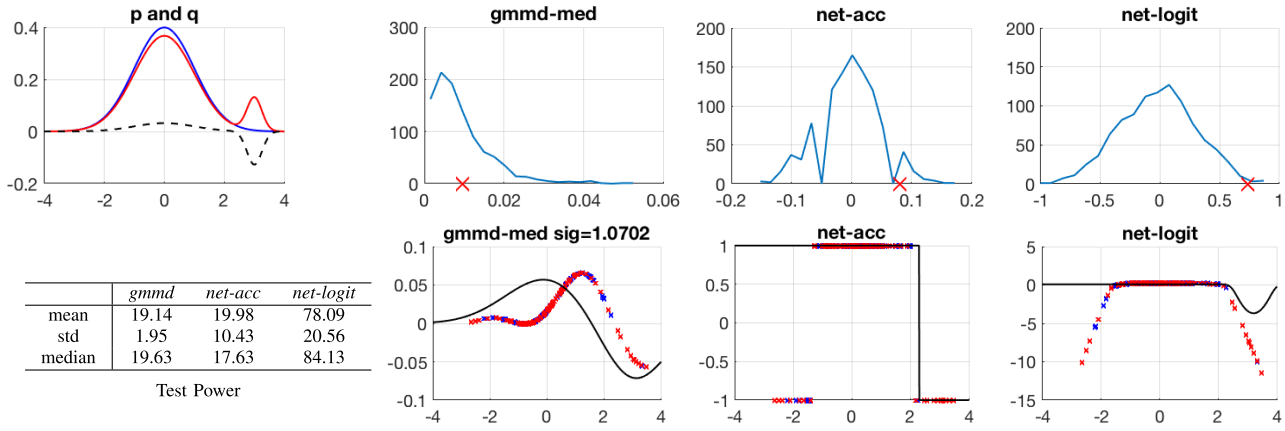
- $O(\varepsilon^{-D/r})$ for r-regular f_{tar}, when p, q are general sub-exponential densities in $\mathbb{R}^D$,
- $O(\varepsilon^{-d/2})$ for C^2 $f_{\text{tar}}|_{\mathcal{M}}$, when p, q are on or near a d-dimensional compact smooth manifold $\mathcal{M}$.

Remark IV.3: The constants in all the big O's in (31) may depend on the network Lipchitz upper bound L_Θ, the function f_{tar}, the diameter of the domain where the densities lie on (in case of exponential decay, the scale c in $\mathcal{P}_{\text{exp}}$), and the manifold-atlas if the densities are on-or-near manifold.

The above theorem directly gives the asymptotic test consistency in the next corollary:

Corollary IV.7: Suppose Theorem IV.6 holds and notations are as therein, and when $p \neq q$, $T := \mathbb{E}T_n > 4(L[f_{\text{tar}}] - L_{\text{gap}}) > 0$. Let $0 < \alpha < 1$ be the two-sample test level, typically $\alpha = 0.05$, and the test threshold be $\tau_n = \frac{\sigma_{\mathcal{H}_0}}{\sqrt{n}} \Psi^{-1}(\alpha)$, where $\Psi(x) := \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$. Then, when $p = q$, as $n = n_{te} \to \infty$, $\Pr[T_n > \tau_n] \to \alpha$; When $p \neq q$, as $n \to \infty$, $\Pr[T_n > \tau_n] \to 1$.

To go beyond the asymptotic consistency as $n \to \infty$ proved in the corollary, one may derive finite-sample testing power lower bound using a control of speed of convergence in CLT, e.g., the Berry-Esseen theorem, together with the moment bound of the random variable $f(x_i)$ by the universal Lipschitz constant bound L_Θ similarly as in the proof of Theorem IV.6. Specifically, the standard Berry-Esseen bound implies that for large but finite n, a test power of $1 - \epsilon - O(n^{-1/2})$ can be guaranteed when $\sqrt{n}T$ is greater than $\Psi^{-1}(\epsilon) + O(1)$ up to multiplying an $O(1)$ constant. This shows that the test power approaches 1 as long as the value of T (which is lower-bounded by $4(L[f_{\text{tar}}] - L_{\text{gap}})$ by our analysis) exceeds a threshold at the order of $n^{-1/2}$. Further details omitted here.



	gmmd	*net-acc*	*net-logit*
mean	19.14	19.98	78.09
std	1.95	10.43	20.56
median	19.63	17.63	84.13

Test Power

Fig. 5. (Plots) Top-left: Two densities p and q in Eg.3, $\delta = 0.08$. Right three columns: The test statistic $\hat{T}$ on $|X_{\text{te}}| = |Y_{\text{te}}| = 100$ samples i.e. under $\mathcal{H}_1$ (red cross) and the histogram of $\hat{T}$ under 1000 permutation tests i.e. $\mathcal{H}_0$ (blue curve). The population witness function (black curve) and the empirical one evaluated on test data (red cross for X_{te}, blue crosses for Y_{te}) of the three methods are shown in the bottom row. (Table) The mean, standard deviation ("std") and median of the test power computed over $n_{\text{rep}} = 20$ replicas. The randomness of test power estimation is explained in Appendix C.

V. Experiments

This section conducts numerical experiments of the proposed two-sample test and compares with alternatives, on synthetic 1D and manifold densities and evaluating handwritten digits generating models. Codes are available at the public repository https://github.com/xycheng/net_logit_test.

A. Synthetic 1D Normal Density Departure

The following three examples all have $x_i \sim \mathcal{N}(0, 1)$ and

- Eg.1. Mean shift, $y_i \sim \mathcal{N}(\delta, 1)$;
- Eg.2. Dilation of variance, $y_i \sim \mathcal{N}(0, (1+\delta)^2)$;
- Eg.3. Mixture with bump at tail, $y_j \sim (1-\delta)\mathcal{N}(0, 1) + \delta\mathcal{N}(3, \frac{1}{16})$.

We examine the tests: (1) *net-acc*, (2) *net-logit*, (3) *gmmd* (setting kernel bandwidth σ to be median distance), (4) *gmmd-ad* (selecting σ adaptively using the training set), and (5) *gmmd+* (using all training and testing samples, median σ), (6) *gmmd++* (using all data and post-selecting σ over a range). Tests (1)-(4) only use the test set when measuring the power, while (5)-(6) access both the training and test sets. More details about experimental set-ups are in Appendix C. The test powers of all the methods are plotted in Figure 6 for the three examples. For Eg.1 and Eg.2, *gmmd+*, *gmmd++* are performing consistently better than the other four which only access the test data set, particularly in Eg.1. In Eg.3, *net-logit* gives stronger power than *gmmd+*, *gmmd++* when $n_{\text{all}} > 200$, where *net-acc* remains inferior to the two. Among the four methods (1)-(4), the performances on Eg.1 are comparable, and *net-logit* gives better power on Eg.2 and Eg. 3. This is especially the case of Eg. 3, where *net-logit* shows the most significant advantage. The test powers, empirical and population witness functions of tests (1)(2)(3) on E.g. 3 are shown in Figure 5, and more details in Appendix C-F.

B. Comparison of Test-Power by Witness Function

Generally, there is no best test methods to use for all data sets, and the test power depends on both method and data. Here we study Eg. 3. in more detail, comparing the witness functions of the network-based and kernel-based methods, so as to explain the empirical advantage of *net-logit* in this case.

The analysis and experiments in Section II-E show that, with large samples and sufficiently large neural network, the trained witness function $\hat{f}_{tr}$ approaches the population witness function f^* in terms of the population divergence $L[f]$. Similar theories hold for kernel MMD tests. Thus comparing test power based on population witness functions sheds light on the behavior of the tests.

1) Witness Function: The population witness functions for *gmmd* is $w_\sigma(x) := \int g_\sigma(x-y)(p(y)-q(y))dy$, $g_\sigma(z) := e^{-|z|^2/(2\sigma^2)}$, and its empirical counterpart is by replacing p and q with the empirical densities of X_{te} and Y_{te} respectively. Recall that the population and empirical witness function for *net-logit* test are $f^*(x) = \log\frac{p(x)}{q(x)}$ and f_θ respectively. For *net-acc*, when $|X_{\text{te}}| = |Y_{\text{te}}|$, it is equivalent to using the sign (taking value of ± 1) of f_θ instead of f_θ in computing the test statistic in (3), as shown in (A.1). Thus we call $\text{Sign}(f_\theta(x))$ the empirical witness function for the *net-acc* test, and $\text{Sign}(f^*(x))$ the population one. The population and empirical witness functions (in one test run) are plotted in Figure 5. Comparing to *gmmd*, the witness function of *net-logit*, i.e., the log density ratio, weighs larger at the differential region which is at the tail of the density p. Taking the sign of f_θ as done in *net-acc* test introduces discontinuity of at the decision boundary neat $x = 2$, which leads to comparatively larger variance of $\hat{T}$. This intuitively explains why the *net-logit* test performs better.

2) Quantitative Comparison: To further explain the testing power difference, we give a quantitative comparison. Let w be the population witness function of the three methods respectively, and define

$$\textbf{Mean} := \mathbb{E}_{x\sim p, Y\sim q}(w(X) - w(Y)),$$
$$\textbf{Std} := \sqrt{\text{Var}_{x\sim p}(w(X)) + \text{Var}_{Y\sim q}(w(Y))}.$$

The larger the **Mean**, and the smaller the **Std**, the more powerful the test is going to be. To remove the scaling

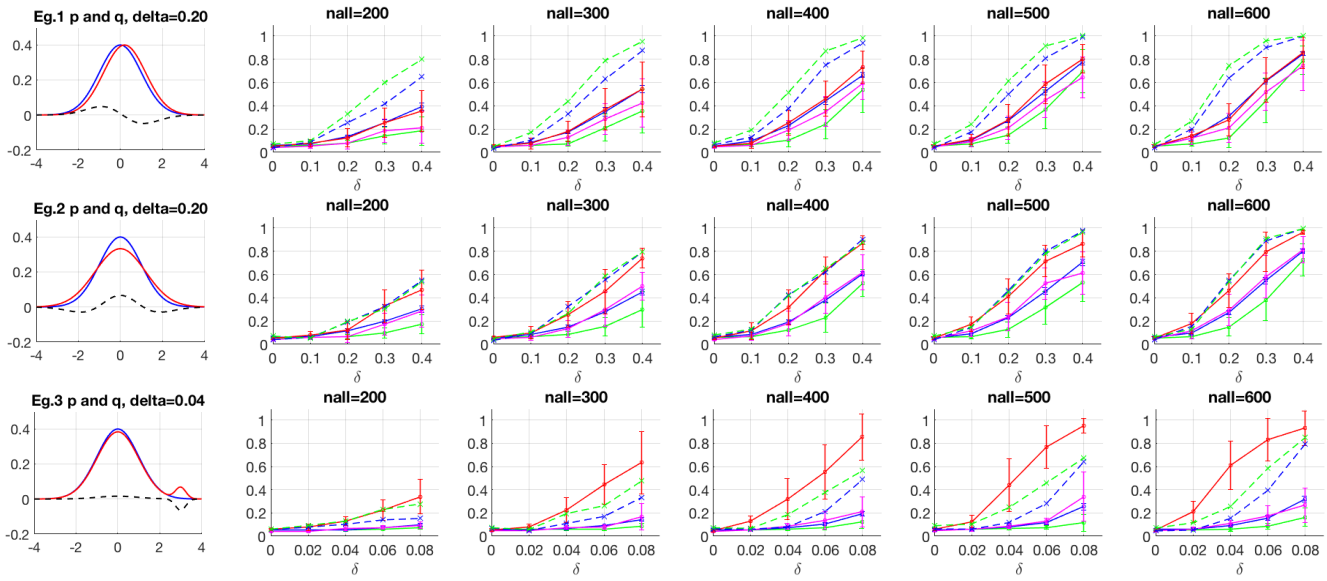


Fig. 6. Three examples of 1D data in Section V-A. Test power of: *gmmd* (blue), *gmmd-ad* (green), *net-acc* (pink), *net-logit* (red), error bar standing for the standard deviation of the estimated power over 20 replicas, and *gmmd*+ (blue dash), *gmmd*++ (green dash). $n_{all} = |X| + |Y|$ including half-half training-testing split.

	Mean	Std	Mean/Std
gmmd	0.0087	0.0421	0.2075
net-acc	0.1579	0.6087	0.2594
net-logit	0.2445	0.9011	0.2714

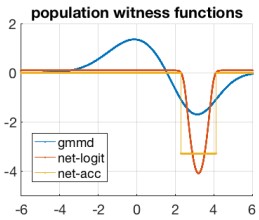


Fig. 7. (Table) The values of **Mean**, **Std**, and their ratio of the three tests in Figure 5. (Right plot) The population witness functions of the three tests, normalized to have **Std** $= 1$.

equivalence (a test statistic multiplied by a positive constant gives the same test power), we will use the ratio of **Mean** and **Std** as an indicator of test power. A similar approach has been taken to study kernel MMD [6], [59]. Using the explicit formula of p and q in Eg. 3, the values of **Mean** and **Std** can be analytically computed, shown in the table in Figure 7, where *net-logit* gives the largest ratio. The normalized witness functions are plotted in Figure 7, where *net-logit* witness function gives the largest weights to the differential region of p and q in this example.

C. Synthetic 2D Manifold Density

The example consists of p and q which lie on the sphere S^2, a 2-dimensional manifold embedded in $\mathbb{R}^3$. A realization of samples X and Y is shown in the left of Figure 8, and the formula of densities are given in Appendix D. Figure 8 plots the test power of the 5 methods over increasing density departure δ and sample size. It can been seen that *net-logit* gives the fastest growth of power as δ increases and the strongest average power for all n_{all}, but the variation can be large if the power is not close to 1. The *gmmd-ad* improves the power of *gmmd*, but does not do as well as *net-acc*, which again performs inferior to *net-logit*. *net-logit* performs better than *gmmd*++ (post-selecting σ, green dash) and the advantage is more evident when $n_{\text{all}} > 200$. This indicates that larger sample size can be particularly in favor of network-based tests, which rely on the search in the network parameter space optimized on a separated training set.

D. Generated vs Authentic MNIST Data

As a real-world data example, we study the task of distinguishing "faked" MNIST samples produced by a pre-trained generative network from authentic ones. The MNIST dataset consists of gray-scale hand-written digits of size 28×28 falling into 10 classes, which is relatively simple and thus is viewed to lie near to low-dimensional manifolds in the ambient space of $\mathbb{R}^{784}$. More details about the generative and classification networks in Appendix E. We compare (1) *net-acc* (2) *net-logit* (3) *gmmd* (4) *gmmd-ad* on two samples X and Y, half of $\mathcal{D} = X \cup Y$ used for training. X consists of authentic MNIST samples, and Y of a mixture of authentic and faked ones, i.e. $p = p_{\text{data}}$ and $q = (1-\delta)p_{\text{data}} + \delta p_{\text{model}}$, $\delta \in [0, 1]$. The test power for increasing δ and sample size $n_{all} = |\mathcal{D}|$ up to 500 is shown in Figure 10, where *net-logit* gives the strongest power throughout all cases, and the two network-based tests significantly outperforms the other two when $n_{all} \geq 300$. The adaptive choice of kernel bandwidth also improves the power over the median-distance choice. The standard deviation of the *net-acc* and *net-logit* power is less than that of *gmmd-ad* power when $n_{\text{all}} = 300$ and $\delta \geq 0.4$, when the former two give near to 1 power. We also observe that the training of the CNN classifier in this experiment is more stable than that of the previous fully-connected network on low-dimensional synthetic data, as revealed in the training error evolution plots, c.f. Figure 12 Figure 15. With another pre-trained model which generates faked images that are closer to authentic ones,

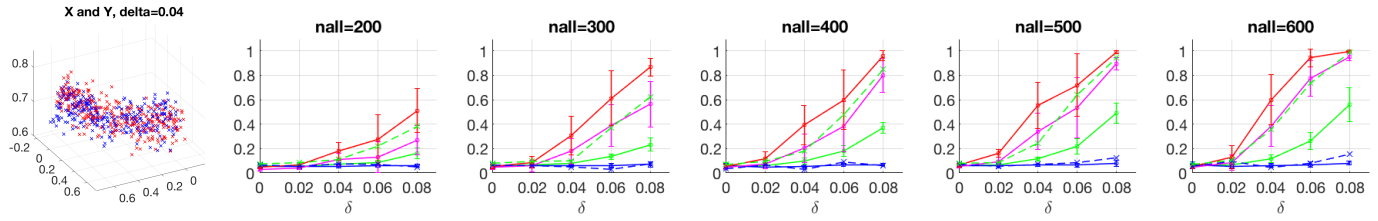


Fig. 8. Test power of the different tests on data on sphere in $\mathbb{R}^3$ in Section V-C. Markers same as in Figure 6.

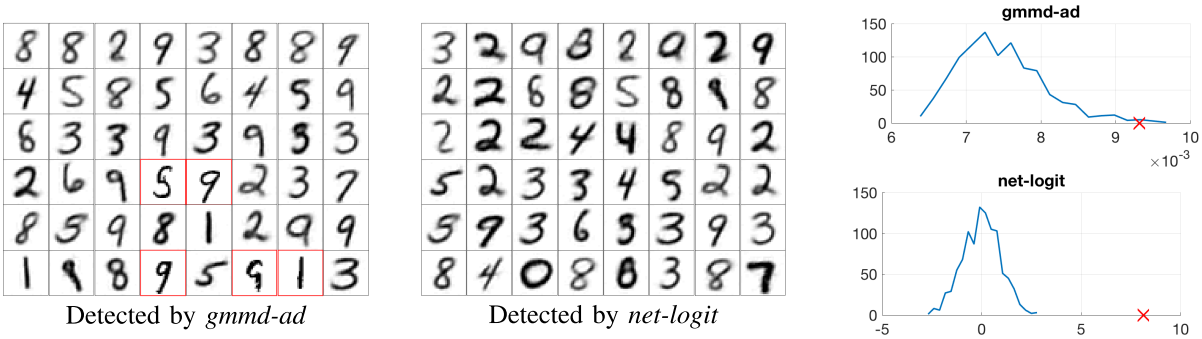


Fig. 9. Two-sample problem of differentiating p, the density of authentic MNIST digits, and q which contains a $\delta = 0.4$ fraction of digits "faked" by a generative model. $|X| = |Y| = 500$. The *gmmd-ad* and *net-logit* tests use half as training set, and test on the other $|\mathcal{D}_{\text{te}}| = 500$ samples. Left and middle: the most likely fake digits identified by the empirical witness functions of the two tests, red box indicates authentic digits incorrectly identified. Right: The test statistic $\hat{T}$ ($\mathcal{H}_1$) and the histogram of its value under 1000 permutation tests ($\mathcal{H}_0$).

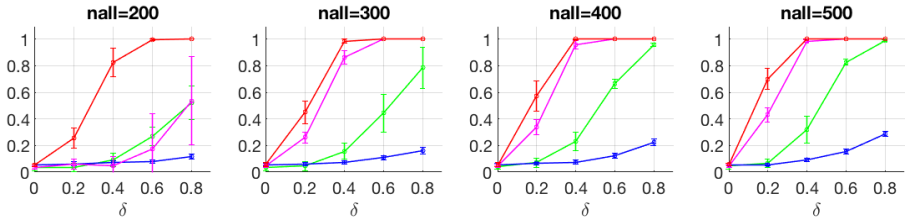


Fig. 10. Test power of *gmmd* (blue) *gmmd-ad* (green) *net-acc* (pink) *net-logit* (red) on differentiating authentic vs synthesized MNIST digits produced by a generative model, where sample X has all authentic ones, and δ stands for the fraction of synthesized ones in Y, $n_{all} = |X| + |Y|$ including half-half training-testing split.

net-logit again shows the best discriminative power, *net-acc* gives comparable performance starting $n_{\text{all}} = 300$, while *gmmd* and *gmmd-ad* gives trivial power up to $n_{all} = 500$, c.f. Figure 14.

Setting $n_{\text{all}} = 1000$, $\delta = 0.4$, the results of *gmmd-ad* and *net-logit* in one test run is shown in Figure 9. Based on the $n_{\text{all}} = 500$ plot in Figure 10, both tests shall have non-trivial power, and that of *net-logit* shall be close to 1. In this test, both methods correctly rejects $\mathcal{H}_0$, yet the *net-logit* statistic deviates from the distribution of $\hat{T}|\mathcal{H}_0$ more significantly, indicating stronger power (shown in the histogram plots). To compare the detecting ability of the empirical witness function $\hat{w}$ of *gmmd-ad* and *net-logit*, for each method, we sort the 250 samples in Y_{te} (among which 100 are faked ones) in ascending order of the value of $\hat{w}$ and select the first 100 samples. These are samples which the model views as most likely to be faked ones. The success rate of identifying faked samples is about 60 by *gmmd-ad* $\hat{w}$, and about 90 by *net-logit* $\hat{w}$. The first 48 most likely faked digits identified with both witness functions are plotted in Figure 9, where *gmmd-ad* $\hat{w}$ incorrectly includes 5 authentic samples, and none by *net-logit* $\hat{w}$.

VI. Discussion

The neural network approximation analysis is under the framework of wavelet construction which are realized by a shallow network. It can be worthwhile to go beyond the framework of [28] and consider deeper network architecture, as well as more than feed-forward network architectures, e.g., residual networks. Method-wise, for the objective of logistic (softmax) loss, as has been pointed out in [39], the training of the classifier can be interpreted as minimizing a Bregman divergence between the estimated logit f_θ and the true log density ratio $f^* = \log(p/q)$. If one views the trained f_θ as an estimator of f^*, the approximation $f_\theta \approx f^*$ suggests that $T[f_\theta] \approx T[f^*] = \int (p-q) \log \frac{p}{q} = \text{KL}(p||q) + \text{KL}(q||p) = \text{SKL}(p,q)$ which is the symmetric KL divergence (SKL). As a result, the proposed statistic $\hat{T}$ can be viewed as estimating $\text{SKL}(p,q)$. The analysis in the current paper

does not directly gives estimation guarantee of $\mathrm{SKL}(p,q)$, however, the approximation analysis constructs neural network approximator of f^* and our approximation theory may be used in future analysis of estimating SKL. Meanwhile, one can try to extend to other training objectives than softmax loss, such as f-divergences [11], [22] and testing power estimators [13], [41], [42].

For the two-sample problem, the analysis in this paper covers approximation and estimation error, and more understanding of network optimization is needed so as to better study network classification two-sample test methods. Empirically, we have not systematically explored the influence of different network architectures on optimization. Theoretically, a future direction is to extend our work to optimization guarantees, currently contained within Assumption 1. Among recent results on controlling optimization error of neural networks is the analysis of the neural tangent kernel [52], [53], [60], which also implies control of generalization error [61] and approximation error [62]. For two-sample testing, NTK analysis of training dynamics has been recently considered for a trained kernel MMD [63]. However, most NTK analysis in literature particularly applies to L2 cost functions and linear output networks, and certain assumptions of the NTK break down when these are no longer the case [64]. As the current paper focuses on the logistic (cross-entropy) loss, the connection between NTK, as well as other neural network optimization analysis, can be further explored in future work.

VII. Proofs

A. Proofs in Section III

1) Proofs in Section III-B:

Proof of Lemma III.3: For any $x \in U_i = B(x_i,\delta)\cap\mathcal{M}$, we have that (note that $\phi_i(x_i)=x_i$)

$$\|x-x_i\| \le d_{\mathcal{M}}(x,x_i) \le \beta_i\|\phi_i(x)-x_i\|,$$

where the first inequality is by that geodesic distance is always larger than the Euclidean distance, and the second inequality is by (13). This gives that

$$\|\phi_i(x)-x_i\| \ge \frac{1}{\beta_i}\|x-x_i\|. \tag{32}$$

Meanwhile, as ϕ_i is an orthogonal projection, we have that $\|x-x_i\|^2 = \|\phi_i(x)-x_i\|^2 + \|x-\phi_i(x)\|^2$, and then with (32) it gives that

$$\|x-\phi_i(x)\| \le \|x-x_i\|\sqrt{1-1/\beta_i^2} < \delta\sqrt{1-1/\beta_i^2},$$

which is further bounded by $\frac{\sqrt{3}}{2}\delta$ by $\beta_i \le 2$ as in (14). □

Proof of Theorem III.2: The construction of f_N was originally posed in [28], but added here for completeness. The bound on the Lipschitz constant uses additional new analysis.

Under the setup of manifold atlas $\{(U_i,\phi_i)\}_{i=1}^K$, the network function is

$$f_N(x) = \sum_{i=1}^K \bar{f}_{N,i}(x), \quad x\in\mathbb{R}^D, \tag{33}$$

where each $\bar{f}_{N,i}$ is constructed near U_i in the following way. For each i, we work with the local coordinates $x=(u,v)$, $u := \phi_i(x)\in\mathbb{R}^d$, and $v\in\mathbb{R}^{D-d}$, and we use the bar to denote that the function is defined in $\mathbb{R}^D$. We utilize a wavelet-type construction through combinations of Relu activation units in order to create a multiscale approximation scheme.

This begins with the construction of a trapezoid based scaling function

$$\begin{aligned} t(x) :=& \ \mathrm{Relu}(x+3)-\mathrm{Relu}(x+1)\\ &-\mathrm{Relu}(x-1)+\mathrm{Relu}(x-3) \end{aligned} \tag{34}$$

$$\varphi_{k,b}(u) := c_d \mathrm{Relu}\left(\sum_{j=1}^d t(2^{\frac{k}{d}}(u_j-b_j)) - 2(d-1)\right),$$
$$k=0,1,2,\cdots, \quad b\in 2^{-\frac{k}{d}}\mathbb{Z}_d, \tag{35}$$

where c_d is a constant that normalizes the scaling function to be unit norm. Similarly, one can construct a wavelet function

$$\psi_{k,b}(u) := 2^{\frac{k}{2}}\left(\varphi_{k,b}(u) - 2^{-1}\varphi_{k-1,b}(u)\right). \tag{36}$$

The wavelet terms are summed to give a local approximation,

$$\hat{f}_i(u) := \sum_{(k,b)} c_{k,b}\psi_{k,b}(u), \quad k=0,1,2,\cdots, \quad b\in 2^{-\frac{k}{d}}\mathbb{Z}_d. \tag{37}$$

The finite-term summation $\hat{f}_{N,i}$ will be a truncation $k\le k_{max}$ of $\hat{f}_i$, and the network function

$$\begin{aligned} \bar{f}_{N,i}(x) :=& \ \mathrm{Relu}(\mathrm{Relu}(\hat{f}_i(u)) + F_0 g_\delta(v) \ - F_0)\\ &- \mathrm{Relu}(\mathrm{Relu}(-\hat{f}_i(u)) + F_0 g_\delta(v) \ - F_0),\\ F_0 :=& \ \|f\|_{L^\infty(\mathcal{M})}+1, \end{aligned} \tag{38}$$

where $g_\delta:\mathbb{R}^D\to\mathbb{R}$ will be constructed such that, C_1 and C_2 being absolute constants,

(1) g_δ is continuous on $\mathbb{R}^D$, $\mathrm{supp}(g_\delta)\subset B_\delta^D$, $0\le g_\delta(v)\le 1$ and $g_\delta(v)=1$ if $\|v\|\le\frac{\sqrt{3}}{2}\delta$.
(2) g_δ is piece-wise differentiable on $\mathbb{R}^D$, and $\|\nabla g_\delta\|\le\frac{C_1}{\delta}$ when differentiable.
(3) $g_\delta(v)$ can be represented by a network with $\le C_2 D\frac{\log D+\log 1/\delta}{\delta}$ many parameters.

Note that while $v=x-\phi_i(x)$ lies in $(T_{x_i}(\mathcal{M}))^\perp$ which is $(D-d)$-dimensional space, in practice our network g_δ takes the coordinates of v in $\mathbb{R}^D$ (to avoid D-by-D dense connection in the change of coordinate layer), thus g_δ is constructed to be a mapping from $\mathbb{R}^D$ to $\mathbb{R}$ with the properties (1)-(3).

Lemma VII.1: Given $0<\delta\le 1$, a function $g_\delta:\mathbb{R}^D\to\mathbb{R}$ that satisfies (1)(2)(3) can be constructed.

The approximation construction proceeds by setting the coefficients $c_{k,b}=\langle f\eta_i,\tilde{\psi}_{k,b}\rangle$ in (37), where η_i is the partition of unity function on U_i, the function $f\eta_i$ is viewed as a function on $\mathbb{R}^d$ (due to one-to-one correspondence between U_i and $\phi_i(U_i)$) which is supported on $\phi_i(U_i)$, and the inner product is taking on $\mathbb{R}^d$. The function $\tilde{\psi}_{k,b}$ is the dual of basis $\psi_{k,b}$, which is compactly supported on $\mathbb{R}^d$.

The following two technical lemmas provide two properties (P1) and (P2) concerning the convergence of the infinite summation in (37). First observe that while b is on an infinite grid in $\mathbb{R}^d$ as in (37), since the functions $\varphi_{k,b}(u)$ are compactly supported, only finitely many b are involved in the summation

for each k. Specifically, the $2^{-\frac{k}{d}}$ spacing of b_j in each of the d direction matches the wavelet basis spacial rescaling $2^{\frac{k}{d}}$ in (35), which gives the following lemma.

Lemma VII.2 (P1): In (37), at each $u \in \mathbb{R}^d$, for each k, at most 12^d many b's are involved in the summation.

Another important property is the decaying of the wavelet coefficients $c_{k,b}$ as k increases due to the C^2 regularity of the function $f\eta_i$ on $\mathbb{R}^d$.

Lemma VII.3 (P2): For any b and $k = 0, 1, 2, \cdots$, $|c_{k,b}| \leq C_{d,f,\eta_i} 2^{-(\frac{2k}{d}+\frac{k}{2})}$, the constant depending on the 2nd derivative of $f\eta_i$ as a function on $\mathbb{R}^d$ and the dimension d.

For any $\epsilon < 1$, let $\hat{f}_{N,i}(u)$ be the $k \leq k_{max}$ truncation of (37), and k_{max} large enough such that

$$\sum_{k>k_{max}} 12^d \cdot C_{d,f,\eta_i} 2^{-(\frac{2k}{d}+\frac{k}{2})} 4\ c_d 2^{k/2} < \epsilon, \tag{39}$$

which is possible due to the summability of $2^{-2k/d}$. By Lemmas VII.2 and VII.3, this proves that

$$|\hat{f}_{N,i}(u) - f\eta_i(u)| < \epsilon, \quad \forall u \in \mathbb{R}^d \text{ in the local coordinate on } T_{x_i}(\mathcal{M}). \tag{40}$$

We now claim that, with the definition (38),

$$|\bar{f}_{N,i}(x) - f\eta_i(x)| < \epsilon, \quad \forall x \in \mathcal{M} \text{ even not in } U_i, \tag{41}$$

and to verify this, consider three cases respectively,

(i) $x = (u, v) \in U_i$. By Property (1) of g_δ and Lemma III.3, $g_\delta(v) = 1$. Then (38) becomes

$$\begin{aligned}\bar{f}_{N,i}(x) &= \text{Relu}(\text{Relu}(\hat{f}_{N,i}(u))) - \text{Relu}(\text{Relu}(-\hat{f}_{N,i}(u))) \\ &= \hat{f}_{N,i}(u),\end{aligned}$$

and then (41) follows by (40).

(ii) $x \notin U_i$, but $\phi_i(x) \in \phi_i(U_i)$. This only happens if $\|x - \phi_i(x)\| > \delta$. (Otherwise, since $\|\phi_i(x) - x_i\| < \delta$, then $\|x - x_i\| < 2\delta$. This means that $x \in B^D_{2\delta}(x_i)$, while $B^D_{2\delta}(x_i) \cap \mathcal{M}$ is isomorphic to ball in $\mathbb{R}^d$ by construction, c.f. beginning of Section III-B, thus $\phi_i(x) \in \phi_i(U_i)$ only when $x \in U_i$, drawing a contradiction.) Thus in the local coordinates $x = (u, v)$, $u = \phi_i(x)$, $\|v\| > \delta$, and then by Property (1) of g_δ, $g_\delta(v) = 0$. Note that while $f\eta_i(x) = 0$, $f\eta_i(u)$ may not be zero due to that $u = \phi_i(x) \in \phi_i(U_i)$. However, $|f\eta_i(u)| \leq |f(u)| \leq \|f\|_{L^\infty(\mathcal{M})}$, and then (40) gives that $|\hat{f}_{N,i}(u)| < \epsilon + |f\eta_i(u)| < 1 + \|f\|_{L^\infty(\mathcal{M})} = F_0$. Inserting into (38) gives that

$$\begin{aligned}\bar{f}_{N,i}(x) = &\text{Relu}(\text{Relu}(\hat{f}_{N,i}(u)) - F_0) \\ &- \text{Relu}(\text{Relu}(-\hat{f}_{N,i}(u)) - F_0) = 0,\end{aligned}$$

thus (41) holds.

(iii) $x \notin U_i$, and $\phi_i(x) \notin \phi_i(U_i)$. Again $f\eta_i(x) = 0$. Since $f\eta_i(u)$ vanishes outside $\phi_i(U_i)$, $f\eta_i(u) = 0$. By (40), $|\hat{f}_{N,i}(u)| < \epsilon$. Note that $g_\delta(v)$ may not be zero, but remains between 0 and 1. Note that $\text{Relu}(\hat{f}_{N,i}(u)) = (\hat{f}_{N,i}(u))^+ \in [0, \varepsilon)$, and then

$$\begin{aligned}\bar{f}_{N,i}(x)^+ &:= \text{Relu}((\hat{f}_{N,i}(u))^+ + F_0 g_\delta(v) - F_0) \\ &\leq (\hat{f}_{N,i}(u))^+ < \epsilon.\end{aligned} \tag{42}$$

Similarly,

$$\begin{aligned}\bar{f}_{N,i}(x)^- &:= \text{Relu}((\hat{f}_{N,i}(u))^- + F_0 g_\delta(v) - F_0) \\ &\leq (\hat{f}_{N,i}(u))^- < \epsilon.\end{aligned} \tag{43}$$

For each u, only one of $(\hat{f}_{N,i}(u))^\pm$ is non zero, and when $(\hat{f}_{N,i}(u))^+ = 0$ then so is $\bar{f}_{N,i}(x)^+$, same for minus. Thus $\bar{f}_{N,i}(x) = \bar{f}_{N,i}(x)^+ - \bar{f}_{N,i}(x)^-$ satisfies that $|\bar{f}_{N,i}(x)| < \epsilon$, which proves (41).

Given (41), back to (33), and by that $\sum_{i=1}^K \eta_i = 1$ on $\mathcal{M}$, for any $x \in \mathcal{M}$,

$$|f_N(x) - f(x)| \leq \sum_{i=1}^K |\bar{f}_{N,i}(x) - f\eta_i(x)| < \epsilon K.$$

Setting $\epsilon := \frac{\varepsilon}{K}$ to begin with, which determines k_{max} in (39), finishes the approximation part of the Theorem, and to verify the claimed number of neural network parameters, we use big O to denote multiplying an absolute constant here:

- The first layer which conducts change of coordinate of $x \in \mathbb{R}^D$ to local coordinates around U_i, for each i, takes $O(dD)$ parameters. Because $u \in \mathbb{R}^d$ is determined by the projection ϕ_i, which is a D-to-d linear transform, of $x_c := (x - x_i)$, and $v = x_c - \phi_i(x_c)$ can be computed in another layer which has $O(dD)$ weights.
- On the branch sub-network for each i, the layer which produces $\bar{f}_{N,i}$ takes the input of $\hat{f}_{N,i}(u)$ and $g_\delta(v)$ and uses $O(1)$ weights. In $\hat{f}_{N,i}(u)$, the number of basis $\#\{(k,b)\} = \sum_{k=0}^{k_{max}} (2\delta)^d 2^k = O(\delta^d 2^{k_{max}})$, because the diameter of $\phi_i(U_i) \leq 2\delta$ due to $U_i \subset B_\delta(x_i)^D$, thus a grid of b on $(-\delta, \delta)^d$ suffices. k_{max} is set in (39) which satisfies $2^{k_{max}} \leq (\frac{\varepsilon}{KC'_{d,f,\eta_i}})^{-d/2}$, where C'_{d,f,η_i} is a constant depending on $f\eta_i$, d and atlas. Let $C' := \max_{i=1}^K C'_{d,f,\eta_i}$, and then $2^{k_{max}} \leq \varepsilon^{-d/2}(KC')^{d/2}$. We use this k_{max} for all atlas i subnet, thus $\#\{(k,b)\} \leq O(\delta^d \varepsilon^{-d/2}(KC')^{d/2})$. The bottom layer which constructs $\varphi_{k,b}(u)$ takes $O(d\#\{(k,b)\})$ many weights, and can be shared across i. The upper layer which linearly combines $\psi_{k,b}$ from $\varphi_{k,b}$'s to form $\hat{f}_{N,i}$ involves i-atlas specific coefficients $c_{k,b}$, and uses $O(\#\{(k,b)\})$ many weights. In $g_\delta(v)$, the 2 layer network uses $O(\frac{D}{\delta}\log\frac{D}{\delta})$ parameters by Property (3) of g_δ, and are shared across all i.

Summing over the K atlas sub-networks, the total number of parameters is, omitting absolute constant,

$$\begin{aligned}&KdD + (d+K)\delta^d \varepsilon^{-d/2}(KC')^{d/2} + \frac{D}{\delta}\log\frac{D}{\delta} \\ &\quad = C_{f,\mathcal{M}} \varepsilon^{-d/2} + N_0, \\ &C_{f,\mathcal{M}} := (d+K)\delta^d (KC')^{d/2}, \quad N_0 := KdD + \frac{D}{\delta}\log\frac{D}{\delta}.\end{aligned}$$

We now prove the Lipschitz constant part of the Theorem. For fixed i, we first bound $\text{Lip}_{\mathbb{R}^D}(\bar{f}_{N,i})$. Working with local coordinates $x = (u, v)$, for any $x, x' \in \mathbb{R}^D$,

$$\begin{aligned}&|\bar{f}_{N,i}(x) - \bar{f}_{N,i}(x')| \\ &\leq |\bar{f}_{N,i}(x)^+ - \bar{f}_{N,i}(x')^+| + |\bar{f}_{N,i}(x)^- - \bar{f}_{N,i}(x')^-| \\ &\leq 2(|\hat{f}_{N,i}(u) - \hat{f}_{N,i}(u')| + F_0 |g_\delta(v) - g_\delta(v')|).\end{aligned} \tag{44}$$

To bound $\text{Lip}_{\mathbb{R}^d}(\hat{f}_{N,i})$, recall that

$$\hat{f}_{N,i}(u) = \sum_{k=0}^{k_{max}} \sum_b c_{k,b} \psi_{k,b}(u),$$

and in (36), (35), by that t is piece-wise differentiable on $\mathbb{R}$, the function $\varphi_{k,b}$ is piece-wise differentiable on $\mathbb{R}^d$, and then so is $\psi_{k,b}$. We have that

$$\begin{aligned}\|\nabla\varphi_{k,b}(u)\| &\le |c_d|\|2^{k/d}\mathbf{1}_d\| = |c_d|(d\cdot 2^{2k/d})^{1/2}\\ &=: c'_d\cdot 2^{k/d},\quad \forall k=0,1,2,\cdots,\end{aligned}\tag{45}$$

where c'_d is a constant depending on d. Then,

$$\begin{aligned}\|\nabla\psi_{k,b}(u)\| &\le 2^{\frac{k}{2}}\left(\|\nabla\varphi_{k,b}(u)\| + 2^{-1}\|\nabla\varphi_{k-1,b}(u)\|\right)\\ &\le 2^{\frac{k}{2}}c'_d(2^{\frac{k}{d}} + 2^{-1}2^{\frac{k-1}{d}}) \le \frac{3}{2}c'_d 2^{\frac{k}{2}}2^{\frac{k}{d}},\end{aligned}\tag{46}$$

and then, for any $u,u'\in\mathbb{R}^d$,

$$\begin{aligned}&|\hat{f}_{N,i}(u) - \hat{f}_{N,i}(u')|\\ &= \left|\sum_{k=0}^{k_{max}}\left(\sum_{b\in\mathcal{B}_u} c_{k,b}\psi_{k,b}(u) - \sum_{b\in\mathcal{B}_{u'}} c_{k,b}\psi_{k,b}(u')\right)\right|\\ &\quad (B_u \text{ denoting the set of } b\text{'s involved in the summation})\\ &\le \sum_{k=0}^{k_{max}}\sum_{b\in\mathcal{B}_u\cup\mathcal{B}_{u'}} |c_{k,b}||\psi_{k,b}(u)-\psi_{k,b}(u')|\\ &\le \sum_{k=0}^{k_{max}}(2\cdot 12^d)C_{d,f,\eta_i}2^{-(\frac{2k}{d}+\frac{k}{2})}\cdot\frac{3}{2}c'_d 2^{\frac{k}{2}}2^{\frac{k}{d}}\|u-u'\| \quad (47)\\ &=: \|u-u'\|C_{f,\mathcal{M}}\sum_{k=0}^{k_{max}}2^{-\frac{k}{d}}\\ &\le \|u-u'\|C_{f,\mathcal{M}}(1-2^{-1/d})^{-1} =: \|u-u'\|C'_{f,\mathcal{M}},\end{aligned}$$

where $C_{f,\mathcal{M}}$ is a constant depending on function f and manifold-atlas, including d and η_i, and so is $C'_{f,\mathcal{M}}$. In the derivation above, the inequality (47) follows by (P1) Lemma VII.2, (P2) Lemma VII.3, Eqn. (46), and the piece-wise continuous differentiability of $\psi_{k,b}$ in $\mathbb{R}^d$. This proves that $\text{Lip}_{\mathbb{R}^d}(\hat{f}_{N,i}) \le C'_{f,\mathcal{M}}$.

Meanwhile, Property (2) of g_δ gives that $\text{Lip}_{\mathbb{R}^{D-d}}(g_\delta) \le \frac{C_1}{\delta}$, C_1 being absolute constant. Then (44) continues as below, $x=(u,v)$ being orthogonal decomposition,

$$\begin{aligned}(44) &\le 2(C'_{f,\mathcal{M}}\|u-u'\| + F_0\frac{C_1}{\delta}\|v-v'\|)\\ &\le 2(C'_{f,\mathcal{M}} + F_0\frac{C_1}{\delta})\sqrt{2}\|x-x'\| =: C''_{f,\mathcal{M}}\|x-x'\|,\end{aligned}$$

and this proves that $\text{Lip}_{\mathbb{R}^D}(\bar{f}_{N,i}) \le C''_{f,\mathcal{M}}$. By (33),

$$\text{Lip}_{\mathbb{R}^D}(f_N) \le KC''_{f,\mathcal{M}} =: L_{\mathcal{M},f}$$

which is a constant depending on f and manifold-atlas only. □

Proof of Lemma VII.1: Given $0<\delta\le 1$ fixed, we prove the construction of $g = g_\delta : \mathbb{R}^m \to \mathbb{R}$ for any dimension m, and take $m = D$. Let g be in the form of

$$\begin{aligned}&g(v) := t_\delta\left(\sum_{j=1}^m y(v_j)\right), \text{ where for } x\in\mathbb{R},\\ &t_\delta(x) := 1 - \text{Relu}\left(\frac{x-0.8\delta^2}{0.2\delta^2}\right) + \text{Relu}\left(\frac{x-\delta^2}{0.2\delta^2}\right),\end{aligned}\tag{48}$$

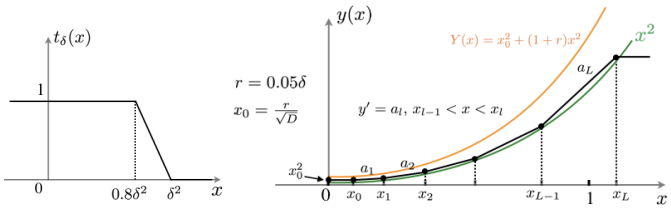


Fig. 11. The construction of $g_\delta : \mathbb{R}^D \to \mathbb{R}$. (Left) t_δ as in (48). (Right) $y(x)$ by $2(L+1)$ Relu's as in (49) (50), only the $x>0$ part is shown and $y(x) = y(-x)$.

and $y:\mathbb{R}\to\mathbb{R}$ will be an approximation of $y(x)\approx x^2$ for $|x|\le 1$. $t_\delta(x) = 1$ when $x < 0.8\delta^2$, 0 when $x > \delta^2$, and linearly interpolating in between, see Figure 11. To construct y, define $r := \frac{1}{20}\delta$, $0<r\le\frac{1}{20}$, define a sequence of points

$$\begin{aligned}&x_0 := \frac{r}{\sqrt{m}},\quad x_l := \rho x_{l-1},\quad \rho = 1+2r > 1,\\ &l = 1,\cdots,L,\ L \text{ is the smallest integer s.t. } x_L = x_0\rho^L > 1.\end{aligned}$$

and let $y(x)$ be a piece-linear function, $y(-x) = y(x)$, and $y(x) = x_0^2$ when $|x|\le x_0$, $y(x_l) = x_l^2$, $l = 1,\cdots,L$, and $y(x) = x_L^2$ for $x > x_L$, see Figure 11. Such $y(x)$ can be represented by $2(L+1)$ many Relu functions, specifically,

$$\begin{aligned}y(x) &:= y_+(x) + y_+(-x) - x_0^2,\\ a_l &:= x_{l-1} + x_l,\ l = 1,\cdots,L,\ a_{l+1} > a_l,\end{aligned}\tag{49}$$

$$\begin{aligned}y_+(x) = x_0^2 &+ a_1\text{Relu}(x-x_0) + (a_2-a_1)\text{Relu}(x-x_1) +\\ &\cdots + (a_L - a_{L-1})\text{Relu}(x-x_{L-1}) - a_L\text{Relu}(x-x_L).\end{aligned}\tag{50}$$

Since $y(x_l) = x_l^2$ for all $0\le l\le L$, and by convexity of x^2, we have that

(p1) $y(x)\ge x^2$ whenever $|x|\le x_L$, and $y(x) = x_L^2 > 1$ when $|x| > x_L$.

We also claim that

(p2) $0\le y(x)\le x_0^2 + (1+r)x^2 := Y(x)$, for all x.

(p3) y is piece-wise linear on $\mathbb{R}$, when $x\ne x_l$, $y'(x)$ exists, and $y' = 0$ if $|x| < x_0$ or $|x| > x_L$, $|y'(x)| < 2.1|x|$ if $x_0 < |x| < x_L$.

To verify (p2) and (p3): By symmetry of y, only consider when $x\ge 0$. When $x\le x_0$, $y(x) = x_0^2 \le Y(x)$. For $x\in[x_{l-1},x_l]$, on the left end point $y(x_{l-1}) = x_{l-1}^2 \le (1+r)x_{l-1}^2 < Y(x_{l-1})$, and $y'(x) = a_l$ on the interval, also $x_l = (1+2r)x_{l-1}$, then

$$\begin{aligned}(Y-y)'(x) &= 2(1+r)x - a_l = 2(1+r)x - (x_{l-1}+x_l)\\ &= 2(1+r)(x-x_{l-1}) \ge 0,\quad x_{l-1}\le x\le x_l,\end{aligned}$$

Thus $Y\ge y$ on $[x_{l-1},x_l]$. When $|x| > x_L$, $Y(x)\ge(1+r)x^2 \ge x_L^2 = y(x)$. Thus (p2) holds.

The differentiability of y is by construction, and when $x_{l-1} < x < x_l$, $l = 1,\cdots,L$, $y'(x) = a_l = x_{l-1} + x_l = x_{l-1}2(1+r) < x2(1+r) \le 2.1\ x$, by that $r\le 0.05$. Thus (p3) holds.

We are ready to prove the properties (1)-(3) of g defined as in (48):

(1) $y:\mathbb{R}\to\mathbb{R}$ is continuous on $\mathbb{R}$, and so is t_δ, then g is continuous on $\mathbb{R}^m$. t_δ takes value between 0 and 1, and so is g. For any $\|v\|_{\mathbb{R}^m}\geq\delta$, $\sum_{j=1}^m v_j^2\geq\delta^2$, by (p1) we have that $\sum_{j=1}^m y(v_j)\geq\delta^2$. (If all $|v_j|\leq x_L$, $y(v_j)\geq v_j^2$, then the sum $\geq\sum_j v_j^2\geq\delta^2$. If any one $|v_j|>x_L$, $y(v_j)>1\geq\delta^2$, then so is the sum.) Then by that $t_\delta(x)=0$ if $x\geq\delta^2$, $g(v)=0$. This proves that g vanishes outside B_δ^m.

For any $\|v\|_{\mathbb{R}^m}\leq\frac{\sqrt{3}}{2}\delta$, we have $\sum_{j=1}^m v_j^2\leq 0.75\delta^2$, then (p2) gives that

$$\begin{aligned}\sum_{j=1}^m y(v_j)&\leq\sum_j(x_0^2+(1+r)v_j^2)\\&=x_0^2m+(1+r)\|v\|^2=r^2+(1+r)\|v\|^2\\&\leq(0.05\delta)^2+1.05\cdot 0.75\delta^2\\&<0.8\delta^2,\quad\text{(by that } r=0.05\delta\leq 0.05)\end{aligned}$$

and then, by that $t_\delta(x)=1$ when $x<0.8\delta^2$, $g(v)=1$.

(2) $y(v_j)$ applies to each coordinate v_j and y is piece-wise linear on $\mathbb{R}$, and t_δ is also piece-wise linear on $\mathbb{R}$, thus g is piece-wise linear on $\mathbb{R}^m$ and thus piece-wise differentiable. By chain rule, $\partial_{v_j}g(v)=t_\delta'(\sum_j y(v_j))y'(v_j)$ and t_δ' is non-zero only if $\sum_j y(v_j)\in(0.8\delta^2,\delta^2)$, in which case $|t_\delta'(\sum_j y(v_j))|=\frac{1}{0.2\delta^2}$ and $|y'(v_j)|\leq 2.1|v_j|$ by (p3). (Because $y(v_j)\geq 0$ and $\sum_j y(v_j)\leq\delta^2$, each $y(v_j)\leq\delta^2\leq 1$, thus $|v_j|<x_L$, and $|y'(x)|\leq 2.1|x|$ as long as $|x|<x_L$). Also note that as proved in (1), one needs $\|v\|<\delta$ to make $\sum_j y(v_j)<\delta^2$. This means that $\nabla g(v)$, when existing, is non-zero only when $\|v\|<\delta$ and $\sum_j y(v_j)\in(0.8\delta^2,\delta^2)$, and then

$$|\partial_{v_j}g(v)|\leq\frac{1}{0.2\delta^2}2.1|v_j|,\quad j=1,\cdots,m,$$

which gives that

$$\|\nabla g(v)\|^2\leq(\frac{10.5}{\delta^2})^2\|v\|^2\leq(\frac{10.5}{\delta^2})^2\delta^2=\frac{(10.5)^2}{\delta^2}.$$

This proves (2) with $C_1=10.5$.

(3) Representing g as a two-layer neural network, the top layer t_δ has 2 neurons and each takes m inputs, thus it has $O(m)$ weights, here we use big O to denote multiplying an absolute constant and same below. The bottom layer branches for the m coordinates, each branch is a sub-network $y(v_j)$ with one hidden layer of width $2(L+1)$ and a scalar input v_j, and thus it has $O(L)$ weights, and all the m branches has $O(mL)$ weights. Thus the total number of parameters is $O(m(1+L))$. By definition, $x_0(1+2r)^{L-1}\leq 1$, thus, by that $\log(1+2r)>r$ $(0<r\leq 0.05)$,

$$L\leq 1+\frac{\log\sqrt{m}+\log\frac{1}{r}}{r}=1+\frac{\frac{1}{2}\log m+\log\frac{20}{\delta}}{0.05\delta},$$

this proves that total number of network parameters $\leq Cm(2+\frac{\frac{1}{2}\log m+\log\frac{20}{\delta}}{0.05\delta})$ for an absolute constant C, which is (3) with C_2 being an absolute constant. □

Proof of Lemma VII.2: Note that $\mathrm{supp}(t(x))\subset[-3,3]$. Because of this, for a fixed k, $\mathrm{supp}(\phi_{k,0})\subset 2^{-k/d}[-3,3]^d$ and $\mathrm{supp}(\psi_{k,0})\subset 2^{-(k-1)/d}[-3,3]^d\subset 2^{-k/d}[-6,6]^d$. Recall that the grid of shifts satisfies $b\in 2^{-k/d}\mathbb{Z}_d$. Because the support and the shifts scale identically with k, we only need count the overlap for $k=0$. For a given u_j for $j=1,\ldots,d$, there are at most 12 b_j values where u_j is contained inside the support. This is true for each j, meaning there are 12^d wavelets $\psi_{0,b}$ such that $\psi_{0,b}(u)\neq 0$. □

Proof of Lemma VII.3: For ease of notation, let $g=f\eta_i$. First, note that the coefficients satisfy

$$\begin{aligned}c_{k,b}&=\langle\tilde\psi_{k,b},g\rangle\\&=2^{k/2}\int\tilde\psi(2^{k/d}(x-b))g(x)dx\\&=2^{k/2}\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)g(2^{-k/d}y+b)dy.\end{aligned}$$

By the assumption that g is twice differentiable, we can take a Taylor expansion of g around b and arrive at

$$\begin{aligned}&\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)g(2^{-k/d}y+b)dy\\&=\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)\Big(g(b)+2^{-k/d}\langle y,\nabla g(b)\rangle\\&\qquad+\frac{1}{2}\|\nabla^2g(b)\|(2^{-k/d}\|y\|_2)^2+O(\|y\|^3)\Big)dy.\end{aligned}$$

By construction of $\psi_{k,b}$, a simple calculation shows that $\psi_{k,b}$ has two vanishing moments (see [28] Proposition C.1 for a full calculation). Because the dual wavelet $\tilde\psi_{k,b}$ can be expressed in terms of a convolution with $\psi_{k,0}$ (see [65]), $\tilde\psi_{k,b}$ inherits two vanishing moments as well. This means

$$|\langle\tilde\psi_{k,b},g\rangle|\leq C2^{-k/2}2^{-2k/d}\|\nabla^2g(b)\|\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)\|y\|^2\,dy.$$

The only thing left to show is that $\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)\|y\|^2\,dy<\infty$. This can be shown as follows: Because ψ is compactly supported, it trivially satisfies

$$|\psi(x)|\leq\frac{C_\alpha}{(1+\|x\|^d)^{1+\alpha}},$$

for any $\alpha>0$. Since $\psi_{k,b}$ is constructed to satisfy the necessary properties to be a wavelet frame (i.e., the decay assumptions of Theorem 3.25 in [65]), $\tilde\psi(y)$ is also a wavelet frame and satisfies

$$|\tilde\psi(x)|\leq\frac{C_{\alpha'}}{(1+\|x\|^d)^{1+\alpha'}},$$

for any $\alpha'<\alpha$. Because the choice of α was arbitrary, we can choose α' large enough that $\int_{\mathrm{supp}(\tilde\psi)}\tilde\psi(y)\|y\|^2dy<\infty$. Combining results, this gives

$$|\langle\tilde\psi_{k,b},f\eta_i\rangle|\leq C_{d,f,\eta_i}2^{-(\frac{2k}{d}+\frac{k}{2})}.$$

□

B. Other Proofs in Section III

Proof of Theorem III.1: Since f is Lipschitz on $\mathbb{R}^D$, T is Lip-1 on $\mathbb{R}$, the composed function $T\circ f$ is Lipschitz on $\mathbb{R}^D$, and

$$\mathrm{Lip}_{\mathbb{R}^D}(T\circ f)\leq\mathrm{Lip}_{\mathbb{R}^D}(f).$$

By that $I[f] = \int_{\mathbb{R}^D} p \cdot T \circ f$, applying Proposition III.5 gives that

$$I[f] = \int_{\mathcal{M}} T \circ f(x)\ \tilde{p}(x) d_{\mathcal{M}}(x) + r_1,$$

and

$$|r_1| \le \left(\|T \circ f\|_{L^\infty(\mathcal{M})} C_1 + \text{Lip}_{\mathbb{R}^D}(f) C_2 \right) c_1 \sigma, \tag{51}$$

where $C_1 := 3KL_{\mathcal{M}}$, $C_2 := K(2L_{\mathcal{M}} + 1 + \beta_{\mathcal{M}})$ as in Proposition III.5.

Let f_{con} be given by Theorem III.2 to uniformly approximate f on $\mathcal{M}$ up to ε, $\text{Lip}_{\mathbb{R}^D}(f_{\text{con}}) \le L_{\mathcal{M},f}$. Repeat the above argument on f_{con} in place of f, we have that

$$I[f_{\text{con}}] = \int_{\mathcal{M}} T \circ f_{\text{con}}(x)\ \tilde{p}(x) d_{\mathcal{M}}(x) + r_2,$$

where since $\|f_{con} - f\|_{L^\infty(\mathcal{M})} \le \varepsilon$ and $\text{Lip}(T) \le 1$, $\|T \circ f_{\text{con}}\|_{L^\infty(\mathcal{M})} \le \|T \circ f\|_{L^\infty(\mathcal{M})} + \varepsilon$, then

$$|r_2| \le \left((\|T \circ f\|_{L^\infty(\mathcal{M})} + \varepsilon) C_1 + L_{\mathcal{M},f} C_2 \right) c_1 \sigma. \tag{52}$$

Comparing the integrals on the manifold,

$$\begin{aligned}
&\left| \int_{\mathcal{M}} T \circ f(x)\ \tilde{p}(x) d_{\mathcal{M}}(x) - \int_{\mathcal{M}} T \circ f_{\text{con}}(x)\ \tilde{p}(x) d_{\mathcal{M}}(x) \right| \\
&\le \int_{\mathcal{M}} |T \circ f(x)\ - T \circ f_{\text{con}}(x)| \tilde{p}(x) d_{\mathcal{M}}(x) \\
&\le \int_{\mathcal{M}} |f(x)\ - f_{\text{con}}(x)| \tilde{p}(x) d_{\mathcal{M}}(x) \quad \text{(by that } T \text{ is Lip-1)} \\
&\le \varepsilon \int_{\mathcal{M}} \tilde{p}(x) d_{\mathcal{M}}(x) \quad \text{(by uniform approximation on } \mathcal{M}) \\
&\le \varepsilon (1 + 3KL_{\mathcal{M}} c_1 \sigma). \quad \text{(by (20))}
\end{aligned} \tag{53}$$

Collecting (51), (52), (53), $|I[f] - I[f_{\text{con}}]|$ is bounded by the sum of the three, which is

$$\begin{aligned}
&(1 + C_1 c_1 \sigma)\varepsilon + \Big\{ C_1 (\varepsilon + 2\|T \circ f\|_{L^\infty(\mathcal{M})}) \\
&\qquad + C_2 (L_{\mathcal{M},f} + \text{Lip}_{\mathbb{R}^D}(f)) \Big\} c_1 \sigma,
\end{aligned}$$

as stated in the Theorem. □

Proof of Lemma III.4: For a fixed i, by the definition (16), and that (i) η_i vanishes outside U_i, and $\phi(N_i) = \phi(U_i)$, and (ii) h_i vanishes outside B_δ^{D-d}, we have that $\text{supp}(\tilde{\eta}_i) \subset N_i$. To see that $\tilde{\eta}_i|_{U_i} = \eta_i$, it suffices to show that $h_i(x - \phi_i(x)) = 1$ on U_i, which follows by the definition of h_i and Lemma III.3.

Finally, we prove the Lipschitz continuity of $\tilde{\eta}_i$ on $\mathbb{R}^D$. First, $\tilde{\eta}_i$ is continuous on $\mathbb{R}^D$. This is because $\tilde{\eta}_i$ has the factorized definition as in (16), and that $\eta_i(\psi_i(u))$ as a function on $\mathbb{R}^d$ is smooth, plus that $h_i(v)$ as a function on $\mathbb{R}^{D-d}$ is continuous, thus the product function $\tilde{\eta}_i$ is continuous on $\mathbb{R}^D$. Next we prove the Lipschitz constant. By the global continuity and that $\text{supp}(\tilde{\eta}_i) \subset N_i$, $\text{Lip}_{\mathbb{R}^D}(\tilde{\eta}_i) = \text{Lip}_{N_i}(\tilde{\eta}_i)$. For the latter, consider $x, x' \in N_i$, and let $y := \psi_i \circ \phi_i(x) \in U_i$,

$$\tilde{\eta}_i(x) = \eta_i(y) h_i(x - \phi_i(x)),$$

similarly for x', y'. Since η_i is smooth on $\mathcal{M}$ and compactly supported on U_i, we assume that

$$|\eta_i(y_1) - \eta_i(y_2)| \le c_i d_{\mathcal{M}}(y_1, y_2), \quad \forall y_1, y_2 \in U_i,$$

i.e., $c_i = \text{Lip}(\eta_i)$ w.r.t. manifold geometry, and c_i is determined once the manifold-atlas is fixed. Then

$$\begin{aligned}
&|\tilde{\eta}_i(x) - \tilde{\eta}_i(x')| \le |\eta_i(y) - \eta_i(y')||h_i(x - \phi_i(x))| \\
&\qquad + |\eta_i(y')||h_i(x - \phi_i(x)) - h_i(x' - \phi_i(x'))| \\
&\le c_i d_{\mathcal{M}}(y, y') + |h_i(x - \phi_i(x)) - h_i(x' - \phi_i(x'))| \\
&\qquad \text{(by that } \eta_i \text{ and } h_i \text{ are bounded by 1)} \\
&\le c_i \beta_i \|\phi_i(y) - \phi_i(y')\|_2 \\
&\qquad + \frac{2\beta_i^2}{\delta} \|(x - \phi_i(x)) - (x' - \phi_i(x'))\| \quad \text{(by (13),(17))} \\
&\le c_i \beta_i \|x - x'\| + \frac{2\beta_i^2}{\delta} \|x - x'\|,
\end{aligned}$$

where the last row is by that $\phi_i(y) = \phi_i(x)$, similarly for y', and ϕ_i is orthogonal projection. This proves that $\text{Lip}(\tilde{\eta}_i) \le c_i \beta_i + \frac{2\beta_i^2}{\delta}$ which can be upper bounded by $2c_i + \frac{8}{\delta}$ by (14). Taking maximum over i gives that $\sup_i \text{Lip}(\tilde{\eta}_i) \le L_{\mathcal{M}}$, which is an absolute content determined by the manifold and atlas. □

Proof of Proposition III.5: We need two technical lemmas, the proofs are elementary and in Appendix A.

Lemma VII.4: For any $p \in \mathcal{P}_\sigma$ defined as in (9), if $\sigma < \frac{1}{2}$, then

$$\int_{\mathbb{R}^D} d(x, \mathcal{M}) p(x) dx, \int_{\mathbb{R}^D} d(x, \mathcal{M})^2 p(x) dx < c_1 \sigma.$$

The requirement of $\sigma < 1/2$ is only used to simply the bound to be $c_1 \sigma$, and the integral of $d(x, \mathcal{M})^2$ actually gives a abound of $2c_1 \sigma^2$.

Lemma VII.5: If $g : \mathbb{R}^D \to \mathbb{R}$ is globally Lipschitz continuous, then

$$|g(x)| \le \|g\|_{L^\infty(\mathcal{M})} + \text{Lip}(g) \cdot d(x, \mathcal{M}), \quad \forall x \in \mathbb{R}^D,$$

where $\|g\|_{L^\infty(\mathcal{M})}$ is finite due to that g is continuous and $\mathcal{M}$ is compact.

For each $i = 1, \cdots, K$, let $H_i := \phi_i(U_i) = \phi_i(N_i)$, $H_i \subset T_{x_i}(\mathcal{M})$. We will show that

$$\begin{aligned}
&\int_{\mathbb{R}^D} p(x) g(x) dx \approx \int_{\mathbb{R}^D} p(x) g(x) \sum_{i=1}^K \tilde{\eta}_i(x) \quad \text{(error 1)} \\
&= \sum_{i=1}^K \int_{N_i} g(x) \tilde{\eta}_i(x) p(x) dx \quad (\text{supp}(\tilde{\eta}_i) \subset N_i, \text{ Lemma III.4}) \\
&\approx \sum_i \int_{N_i} g(\psi_i \circ \phi_i(x)) \tilde{\eta}_i(x) p(x) dx \quad \text{(error 2)},
\end{aligned} \tag{54}$$

and then, on each $N_i = H_i \times B_\delta^{D-d}$, we change to local coordinate $x = (u, v)$, $u = \phi_i(x) \in H_i$, $v = x - \phi_i(x) \in B_\delta^{D-d}$, and use (16) namely

$$\tilde{\eta}_i(x) = \eta_i(\psi_i(u)) h_i(v)$$

to integrate w.r.t v first, which continues (54) as

$$\begin{aligned}
&= \sum_i \int_{H_i} (g \cdot \eta_i)(\psi_i(u)) \left(\int_{B_\delta^{D-d}} h_i(v) p(u, v) dv \right) du \\
&=: \sum_i \int_{U_i} g(z) \eta_i(z) \tilde{p}_i(z) d_{\mathcal{M}}(z) \quad \text{(definition of } \tilde{p}_i) \\
&= \int_{\mathcal{M}} g(z) \left(\sum_i \eta_i(z) \tilde{p}_i(z) \right) d_{\mathcal{M}}(z) = \int_{\mathcal{M}} g(z) \tilde{p}(z) d_{\mathcal{M}}(z),
\end{aligned}$$

where $d_{\mathcal{M}}(z)$ stands for the Reimannian volume measure on $\mathcal{M}$. To prove the proposition, it suffices to bound (error 1) and (error 2) and show that the sum $\leq$ the right hand side of (18).

Bound of (error 1):

$$\begin{aligned}&\left|\int_{\mathbb{R}^D} p(x)g(x)dx - \int_{\mathbb{R}^D} p(x)g(x)\sum_{i=1}^K \tilde{\eta}_i(x)\right| \\ &\leq \int_{\mathbb{R}^D} p(x)\left|g(x)(1-\sum_{i=1}^K \tilde{\eta}_i(x))\right| dx \\ &=: \int_{\mathbb{R}^D} p(x)|g(x)\xi(x)|dx, \end{aligned} \tag{55}$$

where $\xi := (1-\sum_{i=1}^K \tilde{\eta}_i)$, $\xi|_{\mathcal{M}} = 0$, and ξ is Lipschitz on $\mathbb{R}^D$ with

$$\mathrm{Lip}(\xi) \leq \sum_{i=1}^K \mathrm{Lip}(\tilde{\eta}_i) \leq KL_{\mathcal{M}}$$

by Lemma III.4. By Lemma VII.5, $\forall x \in \mathbb{R}^D$,

$$\begin{aligned}|g(x)\xi(x)| &\leq (\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g)\cdot d(x,\mathcal{M})) \\ &\quad \cdot (0 + \mathrm{Lip}(\xi)\cdot d(x,\mathcal{M})) \\ &= \mathrm{Lip}(\xi)(\|g\|_{L^\infty(\mathcal{M})}\cdot d(x,\mathcal{M}) + \mathrm{Lip}(g)\cdot d(x,\mathcal{M})^2),\end{aligned}$$

and then (55) continues as

$$\begin{aligned}(55) &\leq \int_{\mathbb{R}^D} p(x)\mathrm{Lip}(\xi)\left(\|g\|_{L^\infty(\mathcal{M})}\cdot d(x,\mathcal{M})\right. \\ &\qquad \left. +\mathrm{Lip}(g)\cdot d(x,\mathcal{M})^2\right)dx \\ &= \mathrm{Lip}(\xi)\left(\|g\|_{L^\infty(\mathcal{M})}\int_{\mathbb{R}^D} p(x)d(x,\mathcal{M})\right. \\ &\qquad \left. +\mathrm{Lip}(g)\int_{\mathbb{R}^D} p(x)d(x,\mathcal{M})^2\right) \\ &< \mathrm{Lip}(\xi)\left(\|g\|_{L^\infty(\mathcal{M})}c_1\sigma + \mathrm{Lip}(g)c_1\sigma\right),\end{aligned}$$

where the last line is by Lemma VII.4. This proves that

$$(\text{error 1}) < KL_{\mathcal{M}}(\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g))c_1\sigma. \tag{56}$$

Bound of (error 2):

$$\begin{aligned}&\left|\sum_{i=1}^K \int_{N_i} (g(x) - g(\psi_i\circ\phi_i(x)))\tilde{\eta}_i(x)p(x)dx\right| \\ &\leq \sum_{i=1}^K \int_{N_i} |g(x) - g(\psi_i\circ\phi_i(x))|\tilde{\eta}_i(x)p(x)dx.\end{aligned} \tag{57}$$

Define $\xi_i(x) := g(x) - g(\psi_i\circ\phi_i(x))$ for $x \in N_i$, and we derive bound for $|\xi_i(x)|\tilde{\eta}_i(x)$ on N_i. For any $x \in N_i$, exists $x^*_{\mathcal{M}} \in \mathcal{M}$ such that $\|x - x^*_{\mathcal{M}}\| = d(x,\mathcal{M})$, but the point $x^*_{\mathcal{M}}$ may not lie in U_i. Consider the two cases respectively:

(i) If $x^*_{\mathcal{M}} \in U_i$, then $\xi_i(x^*_{\mathcal{M}}) = 0$, and then

$$\begin{aligned}|\xi_i(x)| &= |\xi_i(x) - \xi_i(x^*_{\mathcal{M}})| \\ &\leq \mathrm{Lip}_{N_i}(\xi_i)\|x - x^*_{\mathcal{M}}\| = \mathrm{Lip}_{N_i}(\xi_i)d(x,\mathcal{M}).\end{aligned} \tag{58}$$

For each i, one can verify that $g(\psi_i\circ\phi_i(x))$ has Lipschitz constant $\mathrm{Lip}(g)\beta_i$ on N_i: For any $x, x' \in N_i$, let $y = \psi_i\circ\phi_i(x)$, $y' = \psi_i\circ\phi_i(x')$, $y, y' \in U_i$, then

$$\begin{aligned}&|g(\psi_i\circ\phi_i(x)) - g(\psi_i\circ\phi_i(x'))| \\ &= |g(y) - g(y')| \leq \mathrm{Lip}(g)\|y - y'\| \\ &\leq \mathrm{Lip}(g)d_{\mathcal{M}}(y,y') \quad \text{(geodesic larger than Euclidean)} \\ &\leq \mathrm{Lip}(g)\beta_i\|\phi_i(y) - \phi_i(y')\| \quad \text{(by (13))} \\ &= \mathrm{Lip}(g)\beta_i\|\phi_i(x) - \phi_i(x')\| \leq \mathrm{Lip}(g)\beta_i\|x - x'\|.\end{aligned}$$

As a result, we have that

$$\mathrm{Lip}_{N_i}(\xi_i) \leq \mathrm{Lip}(g)(1+\beta_i). \tag{59}$$

Back to (58), by that $|\tilde{\eta}_i(x)| \leq 1$, and (14), we then have that

$$|\xi_i\cdot\tilde{\eta}_i(x)\ | \leq \mathrm{Lip}(g)(1+\beta_{\mathcal{M}})\cdot d(x,\mathcal{M}). \tag{60}$$

(ii) If $x^*_{\mathcal{M}} \notin U_i$, then $x^*_{\mathcal{M}}$ must be outside N_i. (Otherwise, suppose $x^* := x^*_{\mathcal{M}}$ is in N_i, $\|x^* - \phi_i(x^*)\| \leq \delta$ and $\phi_i(x^*) \in \phi_i(U_i)$. By construction in the beginning of Section III-B, $B^D_{2\delta}(x_i)\cap\mathcal{M}$ is isomorphic to Euclidean ball, thus there is one-to-one correspondence between points in $B^D_{2\delta}(x_i)\cap\mathcal{M}$ and their projected image under ϕ_i. Since $\psi_i(\phi_i(x^*)) \in U_i$, x^* cannot be $\psi_i(\phi_i(x^*))$, thus x_i cannot $\in B^D_{2\delta}(x_i)$. This draws a contradiction, because $\|\phi_i(x^*) - x_i\| < \delta$, then $\|x^* - x_i\| < 2\delta$ by triangle inequality.) Then the line from x to $x^*_{\mathcal{M}}$ intersects with the boundary of N_i at a point x', and

$$\|x - x'\| \leq \|x - x^*_{\mathcal{M}}\| = d(x,\mathcal{M}).$$

Then by that $\tilde{\eta}_i(x') = 0$,

$$\begin{aligned}|\tilde{\eta}_i(x)| &= |\tilde{\eta}_i(x) - \tilde{\eta}_i(x')| \\ &\leq \mathrm{Lip}(\tilde{\eta}_i)\|x - x'\| \leq \mathrm{Lip}(\tilde{\eta}_i)d(x,\mathcal{M}).\end{aligned}$$

Meanwhile, Lemma VII.5 gives that

$$\begin{aligned}|\xi_i(x)| &\leq |g(x)| + |g(\psi_i\circ\phi_i(x))| \\ &\leq 2\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g)d(x,\mathcal{M}).\end{aligned}$$

Together, and by the bound of $\mathrm{Lip}(\tilde{\eta}_i)$ in Lemma III.4,

$$|\xi_i(x)\tilde{\eta}_i(x)| \leq \left(2\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g)d(x,\mathcal{M})\right)L_{\mathcal{M}}d(x,\mathcal{M}). \tag{61}$$

Combining the two cases, we simply bound $|\xi_i(x)|\tilde{\eta}_i(x)$ on N_i by the sum of (60) and (61), and then (57) continues as

$$\begin{aligned}(\text{error 2}) &\leq \sum_{i=1}^K \int_{N_i} \Big\{\mathrm{Lip}(g)(1+\beta_{\mathcal{M}})\cdot d(x,\mathcal{M}) \\ &+ \left(2\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g)d(x,\mathcal{M})\right)L_{\mathcal{M}}d(x,\mathcal{M})\Big\}p(x)dx \\ &< Kc_1\sigma\left\{\mathrm{Lip}(g)(1+\beta_{\mathcal{M}}) + L_{\mathcal{M}}(2\|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(g))\right\}\end{aligned} \tag{62}$$

where Lemma VII.4 is used.

Finally, combining the two bounds of (error 1) and (error 2) (56) and (62) proves the claim. □

C. Proofs in Section IV

1) Approximation Error Analysis of $L[f]$:

Proof of Proposition IV.1: We first consider when $\Omega :=\text{supp}(p+q)$ is compact. Then $\text{supp}(f_{tar})$ is inside Ω and let C' be the diameter of Ω. Result in [25] guarantees the existence of f_{con} such that

$$\|f_{\text{tar}} - f_{\text{con}}\|_{L^\infty(\Omega)} \le \varepsilon \tag{63}$$

with the needed neural network complexity stated in the proposition. Then, by (22),

$$\begin{aligned}
|L[f_{\text{tar}}] - L[f_{\text{con}}]| &\le \frac{1}{2}\Big(\int p|T_p \circ f_{\text{tar}} - T_p \circ f_{\text{con}}| \\
&\quad + \int q|T_q \circ f_{\text{tar}} - T_q \circ f_{\text{con}}|\Big) \\
&\le \frac{1}{2}\left(\int p|f_{\text{tar}} - f_{\text{con}}| + \int q|f_{\text{tar}} - f_{\text{con}}|\right) \quad \text{(by (23))} \\
&\le \varepsilon \frac{1}{2}(\int_\Omega p + \int_\Omega q) = \varepsilon, \quad \text{(by } \text{supp}(p+q) \subset \Omega \text{ and (63))}
\end{aligned}$$

which proves the claim.

When the two densities are merely sub-exponential, by a re-centering of the origin we assume that $\text{supp}(f_{\text{tar}})$ lies inside $\Omega := (-\frac{C'}{2}, \frac{C'}{2})^D$, and all derivatives of $f = f_{\text{tar}}$ vanishes at $\partial\Omega$. The constructive proof in [25] utilizes a partition of unity of the box Ω by evenly divided sub-boxes, and approximate f by a Taylor expansion on each sub-box. As a result, the f_{con} which fulfills (63) also vanishes outside Ω. This is because the only sub-boxes whose support are not in Ω are those that are centered on the boundary of Ω, and then the coefficients in Taylor expansion vanish, thus those sub-boxes can be removed from the formula of f_{con}. Note that since $T(0) = 0$ (23), whenever f vanishes, so does $T \circ f$, thus $\text{supp}(T_p \circ f_{\text{tar}})$, $\text{supp}(T_p \circ f_{\text{con}}) \subset \Omega$. Thus we have

$$\begin{aligned}
&|\int pT_p \circ f_{\text{tar}} - \int pT_p \circ f_{\text{con}}| \\
&\le \int_\Omega p|T_p \circ f_{\text{tar}} - T_p \circ f_{\text{con}}| \\
&\le \int_\Omega p|f_{\text{tar}} - f_{\text{con}}| \le \varepsilon \int_\Omega p \le \varepsilon,
\end{aligned}$$

and similarly for the integral w.r.t. q. Putting together, it gives that $|L[f_{\text{tar}}] - L[f_\theta]| \le \varepsilon$. □

Proof of Proposition IV.2: Under Assumption 2, f_{tar} is smooth and compactly supported in $\mathbb{R}^D$ and then globally Lipschitz. Since $\mathcal{M}$ is compact smooth manifold, $f_{\text{tar}}|_{\mathcal{M}}$ is smooth on the manifold. Since $T = T_p$ and T_q are Lipschitz-1, Theorem III.1 applies to $f = f_{\text{tar}}$ and guarantees the existence of a f_{con} in the network function family with the claimed complexity to bound $|L[f_{\text{tar}}] - L[f_{\text{con}}]|$. (Strictly speaking, Theorem III.1 only applies to bound $|I[f_{\text{tar}}] - I[f_{\text{con}}]|$ where $I[f] = \int pT_p \circ f$, or $\int qT_q \circ f$, and $L[f]$ equals the average of the two. Nevertheless, the proof of Theorem III.1 uses the integral comparison lemma Proposition III.5 and the uniform approximation of f_{tar} on the manifold, which directly extends to prove the same result for $I[f] = L[f]$.)

To prove the proposition, it suffices to bound the quantities

$$\|T \circ f_{\text{tar}}\|_{L^\infty(\mathcal{M})}, \quad T = T_p, T_q.$$

Let $f = f_{\text{tar}}$, since $f(x_0) = 0$ for $x_0 \in \mathcal{M}$, and then $T \circ f(x_0) = 0$, and $\forall x \in \mathcal{M}$,

$$\|T(f(x))\| \le \text{Lip}_{\mathbb{R}^D}(f)\|x - x_0\| \le \text{Lip}_{\mathbb{R}^D}(f)\text{diam}(\mathcal{M}).$$

Thus

$$\|T \circ f_{\text{tar}}\|_{L^\infty(\mathcal{M})} \le \text{Lip}_{\mathbb{R}^D}(f_{\text{tar}})\text{diam}(\mathcal{M}), \quad T = T_p, T_q.$$

Combining with (11) and (12) leads to (24). □

2) Estimation Error Analysis of $L_n[f]$:

Proof of Lemma IV.3: For fixed $0 < r < \frac{B}{L}$, let $X := \{x_i\}_{i=1}^N$ be a r-net of K such that $N = \mathcal{N}(K, r)$. Let $t := 4Lr > 0$, and $F := \{f_j\}_{j=1}^M$ be a maximal t-separated set in $\mathcal{F}'$, meaning that for any $j \ne j'$, $\|f_j - f_{j'}\|_{L^\infty(K)} > t$, and no more member in $\mathcal{F}'$ can be added to preserve this property. Such F always exists because it can be generated by adding points from an arbitrary point while preserving the t-separation property. By construction, F is a t-net of $\mathcal{F}'$. We will show that $M \le$ (26).

Partition the 1D interval $[-B, B]$ into N_1 many disjoint sub-intervals, each of length $\le 2Lr$, and N_1 can be made $< \frac{B}{Lr} + 1 \le \frac{2B}{Lr}$. For any $j \ne j'$, there must be one $x_i \in X$ such that $f_j(x_i)$ and $f_{j'}(x_i)$ lie in distinct subintervals. (Otherwise, $|f_j(x_i) - f_{j'}(x_i)| \le 2Lr$, and by that both f_j and $f_{j'}$ are L-Lipschitz, $|f_j(x) - f_{j'}(x)| < 4Lr$ for any $x \in B_r(x_i)$. Since $\cup_{i=1}^N B_r(x_i)$ cover K, this means that $\|f_j - f_{j'}\|_{L^\infty(K)} \le 4Lr = t$, contradicting with that f_j and $f_{j'}$ are t-separated.) Then each function f_j corresponds to a string of interval indices $(I_1, \cdots, I_N)$, where $I_i \in \{1, \cdots, N_1\}$, and these strings are distinct for the M many f_j's. There are at most N_1^N distinct strings, which means that $M \le N_1^N$. □

Proof of Proposition IV.5: Recall that ($|X_{tr}| = |Y_{tr}| = n$)

$$\begin{aligned}
L_n[f] &= \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^n T_p \circ f(x_i) + \frac{1}{n}\sum_{i=1}^n T_q \circ f(y_i)\right), \\
L[f] &= \mathbb{E}L_n[f],
\end{aligned}$$

where $x_i \sim p$ i.i.d., $y_i \sim q$ i.i.d., and x_i's and y_i's are independent, and T_p, T_q as in (23). We prove the three cases respectively, where for case (1), we prove the compactly supported case first, and the exponential-tail case as (1').

(1) For any $f \in \mathcal{F}_{\Theta,\text{reg}}(B_R)$, there is $x_0 \in B_R$, $f(x_0) = 0$. Then $\forall x \in B_R(0)$, $|f(x)| \le \text{Lip}(f)\|x - x_0\| \le 2LR$, i.e., $\|f\|_{L^\infty(B_R(0))} \le 2LR$. Thus $\mathcal{F}_{\Theta,\text{reg}}(B_R)$ is contained in the function space of

$$\begin{aligned}
\mathcal{F} := \{&f, \text{Lip}_{\mathbb{R}^D}(f) \le L, \|f\|_{L^\infty(B_R(0))} \le 2LR\}, \\
&\text{equipped with } \|\cdot\|_{L^\infty(B_R(0))}.
\end{aligned}$$

By Lemma IV.3, for any $r < 1 < 2R$, $t := 4Lr$, there exists a finite set F in $\mathcal{F}_{\Theta,\text{reg}}(B_R)$ which form an t-net that covers $\mathcal{F}_{\Theta,\text{reg}}(B_R)$ under the metric of $\|\cdot\|_{L^\infty(B_R(0))}$, where

$$\text{Card}(F) \le (\frac{4R}{r})^{\mathcal{N}(B_R(0),r)}.$$

The covering number of a Euclidean ball can be bounded by $\mathcal{N}(B_R(0), r) \le \left(\frac{2R}{r} + 1\right)^D < \left(\frac{3R}{r}\right)^D$, see e.g.

Section 4.2 of [58], using $r < 1$. Thus,

$$\text{Card}(F) \le \exp\left\{(3R)^D r^{-D} \log \frac{4R}{r}\right\}.$$

Given $r < \min\{c_1'/4, 1\}$, for each $f \in F$, $\text{Lip}_{\mathbb{R}^D}(f) \le L$, then Lemma IV.4 and (28) apply to give concentration of the independent sums over x_i's and y_i's respectively. By a union bound,

$$\Pr[\exists f \in F,\ |L_n[f] - L[f]| \ge t] \le \text{Card}(F) \cdot 4\, e^{-c_0' n \frac{t^2}{L^2}}, \quad (64)$$

where $\frac{t}{L} = 4r$ is chosen to $< c_1'$ to begin with. The r.h.s of (64) is upper bounded by

$$\exp\left\{\log 4 + (3R)^D r^{-D} \log \frac{4R}{r} - c_0' n (4r)^2\right\}, \quad \text{for } 0 < r < \min\{c_1'/4, 1\},$$

which is further upper bounded by

$$\exp\left\{-\frac{2}{3} n^{\frac{D}{D+2}} (\log n)^{\frac{2}{2+D}} + \log 4\right\}, \quad (65)$$

if $\gamma := \max\{3R, \frac{1}{4\sqrt{c_0'}}\}$, $r = \gamma(\frac{\log n}{n})^{\frac{1}{2+D}}$, and $(\log n)^{\frac{1}{2+D}} > 4/3$. Note that as $n \to \infty$, $r \to 0$, thus the constraint of $r < \min\{c_1'/4, 1\}$ is satisfied for sufficiently large n.

We consider the good event where $|L_n[f] - L[f]| < t$ for all $f \in F$. Since F is a t-net that covers $\mathcal{F}_{\Theta,\text{reg}}(B_R)$ with unions of closed $\|\cdot\|_{L^\infty(B_R)}$-balls around points in F, for any $f \in \mathcal{F}_{\Theta,\text{reg}}(B_R)$, there is an $f_0 \in F$ such that $\|f - f_0\|_{L^\infty(B_R)} \le t$. Since $\text{supp}(p+q) \subset B_R$, this implies that $|L[f] - L[f_0]| \le t$ and $|L_n[f] - L_n[f_0]| \le t$, then

$$\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}(B_R)} |L_n[f] - L[f]| \le 3t.$$

This proves that with sufficiently large n, with probability $\ge 1 -$ (65),

$$\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}(B_R)} |L_n[f] - L[f]| \le 3 \cdot 4L \max\{3R, \frac{1}{4\sqrt{c_0'}}\} \left(\frac{\log n}{n}\right)^{\frac{1}{2+D}}.$$

(1') The proof extends that in (1) by a truncation argument for the exponential tail of the densities. We rename the R in the statement as R_0 in below.

The following lemma gives the decay of the integration with the tail of sub-exponential densities for any $f \in \mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$, proved in Appendix A.

Lemma VII.6: Suppose p is in $\mathcal{P}_{\text{exp}}$ with $c = 1$, function $f : \mathbb{R}^D \to \mathbb{R}$ has $\text{Lip}_{\mathbb{R}^D}(f) \le L$ and vanishes at some point x_0, $\|x_0\| < R_0$, then, C as in the definition of $\mathcal{P}_{\text{exp}}$,

$$\int_{\|x\| > R} |f| p < L(R_0 + 1 + R) C e^{-R}, \quad \forall R > R_0.$$

We introduce a sequence of domains $\Omega_n := B_{R_n}(0)$ in $\mathbb{R}^D$, where $R_n = \alpha \log n$, $\alpha > 0$ is a constant to be determined. For n samples of x_i's and y_i's, since p, q are in $\mathcal{P}_{\text{exp}}$ with $c = 1$,

$$\Pr[\exists i,\ \|x_i\| > R \text{ or } \exists i',\ \|y_{i'}\| > R] < 2nCe^{-R}, \quad \forall R > 0.$$

We call

$$(\text{good event 1})_n = \{\|y_{i'}\|,\ \|x_i\| \le R_n,\ \forall i, i'\}.$$

For any large enough n such that $R_n > \max\{R_0, 1\}$, the functional space $\mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$ lies inside

$$\mathcal{F}_n := \{f,\ \text{Lip}_{\mathbb{R}^D}(f) \le L,\ \|f\|_{L^\infty(B_{R_n})} \le 2LR_n\} \quad \text{equipped with } \|\cdot\|_{L^\infty(B_{R_n})}.$$

Similarly as in (1), for any $r < 1$ thus $< 2R_n$, $t := 4Lr$, $\mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$ has a subset F_n which is an t-net that covers $\mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$ under $\|\cdot\|_{L^\infty(B_{R_n})}$, and

$$\text{Card}(F_n) \le \exp\left\{(3R_n)^D r^{-D} \log \frac{4R_n}{r}\right\}.$$

We choose

$$r = \gamma_n(\frac{\log n}{n})^{\frac{1}{2+D}},$$
$$\gamma_n = 3R_n\ (\sim \log n > \frac{1}{4\sqrt{c_0'}} \text{ for large } n),$$

then if $(\log n)^{\frac{1}{2+D}} > 4/3$,

$$(\text{good event 2})_n := \{\forall f \in F_n,\ |L_n[f] - L[f]| < t\}$$

fails with probability $\le$ (65).

Restricting to $(\text{good event 1})_n$ and $(\text{good event 2})_n$. For any $f \in \mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$, exists $f_0 \in F_n$ such that $\|f - f_0\|_{L^\infty(B_{R_n})} \le t$, then $|L_n[f] - L_n[f_0]| \le t$. Meanwhile, since both f and f_0 are in $\mathcal{F}_{\Theta,\text{reg}}(B_{R_0}(0))$, Lemma VII.6 applies to both (since $R_n > R_0$), then

$$\begin{aligned}
|L[f] - L[f_0]| &\le \frac{1}{2}\Big(\int p|T_p \circ f - T_p \circ f_0| + \int q|T_q \circ f - T_q \circ f_0|\Big) \\
&\le \frac{1}{2}\left(\int p|f - f_0| + \int q|f - f_0|\right) \quad \text{(by (23))} \\
&\le \frac{1}{2}\left(t\int_{B_{R_n}} (p+q) + \int_{\mathbb{R}^D \setminus B_{R_n}} (p+q)(|f| + |f_0|)\right) \\
&\le t + 2CL(R_0 + 1 + R_n)e^{-R_n},
\end{aligned}$$

which means that

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}(B_{R_0})} |L_n[f] - L[f]| \\
&\le 3\ t + 2CL(R_0 + 1 + R_n)e^{-R_n} \\
&= 3 \cdot 4L \cdot 3R_n \left(\frac{\log n}{n}\right)^{\frac{1}{2+D}} + 2CL(R_0 + 1 + R_n)e^{-R_n} \\
&< \tilde{C}L \cdot \alpha \log n \cdot ((\log n/n)^{\frac{1}{2+D}} + n^{-\alpha}), \quad (66) \\
&\quad \text{(setting } R_n = \alpha \log n > R_0)
\end{aligned}$$

where $\tilde{C}$ is an absolute constant. This happens with probability $\ge 1 - p_{\text{fail}}$, and

$$\begin{aligned}
p_{\text{fail}} &\le \Pr[\text{ fail of } (\text{good event 1})_n] \\
&\quad + \Pr[\text{ fail of } (\text{good event 2})_n\ ] \\
&\le 2nCe^{-R_n} + (65) = (2C)n^{-\alpha+1} + (65).
\end{aligned}$$

To make $p_{\text{fail}} \to 0$, one can set $\alpha = 1 + \epsilon$ for some $\epsilon > 0$, then in (66) the $n^{-\alpha}$ term is dominated by the term of $(\log n/n)^{\frac{1}{2+D}} \gg O(n^{-1/3})$.

Putting together, we have that when n is sufficiently large, specifically, $(\log n)^{\frac{1}{2+D}} \geq 4/3$ and $R_n = (1+\epsilon)\log n > \max\{1, R_0, 1/(12\sqrt{c_0'})\}$, then with probability $\geq 1 - (2C)n^{-\epsilon} -$ (65), the bound (66) holds, which, for large n, is

$$\sim L \log n (\log n/n)^{\frac{1}{2+D}},$$

omitting the absolute constant in front.

(2) Since $\mathcal{M} \subset B_R$, similarly as in (1), enlarge the network function space to be

$$\mathcal{F} := \{f, \text{Lip}_{\mathbb{R}^D}(f) \leq L, \|f\|_{L^\infty(\mathcal{M})} \leq 2LR\} \text{ equipped with } \|\cdot\|_{L^\infty(\mathcal{M})}.$$

When applying Lemma IV.3, the t-net F can be chosen such that

$$\text{Card}(F) \leq \exp\left\{c(\mathcal{M}) r^{-d} \log \frac{4R}{r}\right\},$$

because the covering number of the manifold $\mathcal{N}(\mathcal{M}, r) \leq \frac{c(\mathcal{M})}{r^d}$, where $c(\mathcal{M})$ involves the intrinsic volume of $\mathcal{M}$ integrated over its Riemannian volume element, and it scales with the diameter of $\mathcal{M}$. The rest of the proof is similar, which gives that, with sufficiently large n, with probability $\geq 1 - \exp\left\{-\frac{2}{3} n^{\frac{d}{d+2}} (\log n)^{\frac{2}{2+d}} + \log 4\right\}$,

$$\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}(\mathcal{M})} |L_n[f] - L[f]| \leq 3 \cdot 4L \max\left\{c(\mathcal{M})^{1/d}, \frac{1}{4\sqrt{c_0'}}\right\} \left(\frac{\log n}{n}\right)^{\frac{1}{2+d}}.$$

(3) We consider the same enlarged $\mathcal{F}$ as in (3) equipped with $\|\cdot\|_{L^\infty(\mathcal{M})}$, which can be covered by the same t-net F as there. The large deviation of $|L_n[f] - L[f]|$ for any $f \in F$ is the same, because though the densities p, q are no longer supported on the manifold they still belong to the $\mathcal{P}_{\text{exp}}$ class, and so is the union bound. The difference is in the control of $|L_n[f] - L_n[f_0]|$ and $|L[f] - L[f_0]|$ where f is an arbitrary member in $\mathcal{F}_{\Theta,\text{reg}}(\mathcal{M})$ and $f_0 \in F$ such that $\|f - f_0\|_{L^\infty(\mathcal{M})} \leq t$.

For $|L[f] - L[f_0]|$, we can use the integral comparison Proposition III.5 and the uniform approximation of f by f_0 on the manifold. Specifically, similarly as in the proof of Proposition IV.2, we have that

$$|L[f] - L[f_0]| \leq 2\Big(2LRC_1(\mathcal{M}) + LC_2(\mathcal{M})\Big)c_1\sigma + \Big(1 + C_1(\mathcal{M})c_1\sigma\Big)t, \tag{67}$$

where $C_1(\mathcal{M})$, $C_2(\mathcal{M})$ are manifold-atlas-dependent only constants defined in (19).

For the empirical L_n, we need the concentration of the independent sum of the random variables $d(x_i, \mathcal{M})$ (and similarly $d(y_i, \mathcal{M})$), which is the following lemma proved in Appendix A.

Lemma VII.7: Suppose x_i, $i = 1, \cdots, n \sim p$ i.i.d., $p \in \mathcal{P}_\sigma$ as defined in (9) with $\sigma < \frac{1}{2}$ and also in $\mathcal{P}_{\text{exp}}$ with $c = 1$. Then $\forall 0 < \tau < 1$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n d(x_i, \mathcal{M}) > (c_1 + \tau)\sigma\right] \leq e^{-c_0'' n \tau^2},$$

where c_0'' is an absolute constant.

To proceed, observe that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (T_p \circ f(x_i) - T_p \circ f_0(x_i))\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n |f(x_i) - f_0(x_i)| \quad \text{(by (23))} \\
&\leq \frac{1}{n}\sum_{i=1}^n |f(x_i) - f((x_i)^*_{\mathcal{M}})| + |f((x_i)^*_{\mathcal{M}}) - f_0((x_i)^*_{\mathcal{M}})| \\
&\quad + |f_0((x_i)^*_{\mathcal{M}}) - f_0(x_i)| \quad \text{(for each } x_i, \exists (x_i)^*_{\mathcal{M}} \in \mathcal{M}, \\
&\qquad\qquad \text{and } \|x_i - (x_i)^*_{\mathcal{M}}\| = d(x_i, \mathcal{M})) \\
&\leq \frac{1}{n}\sum_{i=1}^n (2Ld(x_i, M) + t) \quad \text{(by that } \|f - f_0\|_{L^\infty(\mathcal{M})} \leq t) \\
&= t + 2L\left(\frac{1}{n}\sum_{i=1}^n d(x_i, \mathcal{M})\right).
\end{aligned}$$

Similarly for the independent sums with $T_q \circ f(y_i)$, this gives that

$$|L_n[f] - L_n[f_0]| \leq t + L\left(\frac{1}{n}\sum_{i=1}^n d(x_i, \mathcal{M}) + \frac{1}{n}\sum_{i=1}^n d(y_i, \mathcal{M})\right). \tag{68}$$

Putting together (67), (68), Lemma VII.7, and the union bound over the t-net such that $\sup_{f \in F} |L_n[f] - L[f]| < t$, choosing $\tau = c_1$, $t = 4Lr$, $r \sim (\log n/n)^{1/(2+d)}$ as in (3), we have that, when n is sufficiently large, with probability $\geq 1 - \exp\left\{-\frac{2}{3} n^{\frac{d}{d+2}} (\log n)^{\frac{2}{2+d}} + \log 4\right\} - \exp\{-c_0'' c_1^2 n\}$,

$$\sup_{f \in \mathcal{F}_{\Theta,\text{reg}}(\mathcal{M})} |L_n[f] - L[f]| \leq \tilde{C}(\mathcal{M}) \cdot L(\sigma + (\log n/n)^{\frac{1}{2+d}}),$$

where $\tilde{C}(\mathcal{M})$ is a constant that depends on manifold-atlas only. □

3) Testing Power Analysis on the Testing Set:

Proof of Theorem IV.6: To prove (1): The bounding of L_{gap} is by triangle inequality as detailed in Section II-D and collecting the following bounds:

- ΔC is the optimization error as in Assumption 1.
- The $O(\varepsilon)$ term is the network approximation error of $L[f_{\text{tar}}]$, which is proved by Proposition IV.1 for general and on-manifold densities p and q.
  For near-manifold densities, the bound of approximation error by $O(\varepsilon)+O(\sigma)$ is proved by Proposition IV.2, where σ is the sub-exponential decay scale as defined in (9).
- The $\tilde{O}(n_{\text{tr}}^{-1/(2+d')})$ term is the estimation error proved by Proposition IV.5.

The needed neural network complexity is as stated in the theorem, and the constant dependence in big-O is as in Remark IV.3.

Note that under the setting of Proposition IV.5, which only imposes the extra assumption that network functions need to vanish at least at one point in a bounded domain beyond Assumption 3, the approximation results hold. Particularly, in Proposition IV.2, the $\mathrm{Lip}(f_{\mathrm{con}})$ can be bounded by L_Θ which is a universal constant, and all other constants depends on manifold-atlas, the target function f_{tar}.

To prove (2): Because T_n is the sum of two independent sums of a fixed Lipschitz function $\hat{f}_{\mathrm{tr}}$ averaged on x_i's and y_i's respectively, CLT applies to give the asymptotically normality, and it suffices to bound the variance to prove the claim. Let $f = \hat{f}_{\mathrm{tr}}$, under Assumption 3, $\mathrm{Lip}_{\mathbb{R}^D}(f) \le L_\Theta$. Note that $\mathrm{Var}(\sqrt{n}T_n) = \mathrm{Var}_{x\sim p}(f(x)) + \mathrm{Var}_{y\sim q}(f(y))$. Since in all three settings in Proposition IV.5, p and q are in $\mathcal{P}_{\exp}$ with $c = 1$, then similar to Lemma IV.4, we know that $(f(x_i) - \mathbb{E}_{x\sim p} f(x))$ is 1D sub-exponential random variables satisfying (27) where $L = L_\Theta$. This proves that $\mathrm{Var}(f(x_i)) \le c'' L_\Theta^2$, where c'' is an absolute constant. Same argument applies to $f(y_i)$, and thus $\mathrm{Var}(\sqrt{n}T_n) \le L_\Theta^2$ multiplied by an absolute constant. □

Proof of Corollary IV.7: When $p = q$, the claim is equivalent to the asymptotic normality of $\sqrt{n}T_n/\sigma_{\mathcal{H}_0}$. When $p \neq q$, we show that for any small $\epsilon > 0$, there exists n_ϵ s.t. when $n > n_\epsilon$ then $\Pr[T_n > \tau_n] \ge 1 - 2\epsilon$. For fixed $\epsilon < 1$, define $t := \Psi^{-1}(\epsilon)$ which is a positive constant. There exists n_1 s.t. when $n > n_1$, then $\sqrt{n}T > \sigma_{\mathcal{H}_1} t + \sigma_{\mathcal{H}_0}\Psi^{-1}(\alpha)$, which implies that

$$\Pr[T_n > \tau_n] \ge \Pr\left[\frac{\sqrt{n}(T_n - T)}{\sigma_{\mathcal{H}_1}} > -t\right] := b_n.$$

By the asymptotic normality of $\sqrt{n}(T_n - T)/\sigma_{\mathcal{H}_1}$, the sequence b_n as $n \to \infty$ converges to $1 - \Psi(t) = 1 - \epsilon$. This means that, there is another n_2 such that $b_n > 1 - 2\epsilon$ whenever $n > n_2$. Taking n_ϵ as $\max\{n_1, n_2\}$ proves the claim. □

Lemma VII.8: For any f so that the integrals are defined, $T[f] \ge 4L[f]$.

The relaxation in Lemma VII.8 may not be sharp, particularly, when p and q nearly non-overlap on their supports, $T[f^*] = 2\mathrm{SKL}(p, q)$ diverges to infinity, while $L[f^*]$ remains bounded (by $2\log 2$). When f is close to zero, which as discussed above is the more relevant scenario for two-sample test, the following lemma quantifies the tightness of the relaxation. Proofs of both lemmas are in Appendix A.

Lemma VII.9: For any f s.t. f^2 is integrable w.r.t p and q,

$$0 \le \frac{1}{2}T[f] - 2L[f] \le \int (p+q)\frac{f^2}{2}.$$

Appendix A
Proofs of Technical Lemmas

Proof of Lemma IV.4: $\xi_i = g(x_i)$, where $g := T \circ f$, $\mathrm{Lip}_{\mathbb{R}^D}(g) \le L$.

$$|g(x_i) - \mathbb{E}_{x\sim p} g(x)| \le \int_{\mathbb{R}^D} |g(x_i) - g(x)| p(x)dx$$
$$\le L \int \|x_i - x\| p(x)dx \le L(\|x_i\| + \mathbb{E}_{x\sim p}\|x\|)$$

and by (21), $\mathbb{E}_{x\sim p}\|x\| = \int_0^\infty \Pr_{x\sim p}[\|x\| > t]dt < C$ which is an absolute constant. This means that the random variable $(\xi_i - \mathbb{E}\xi_i)/L$ in absolute value is upper bounded by $\|x_i\| + C$, where $y := \|x_i\|$ as a 1D random variable satisfies that $\Pr[|y| > t] < Ce^{-t}$, thus $(\xi_i - \mathbb{E}\xi_i)/L$ satisfies the sub-exponential tail claimed in the lemma with some other absolute constants C' and c'. □

Proof of Lemma VII.4: Let $X \sim p$, then $d(X, \mathcal{M})$ is a non-negative random variable, and

$$\int_{\mathbb{R}^D} d(x,\mathcal{M})p(x)dx = \int_{\mathbb{R}^D}\left(\int_0^{d(x,\mathcal{M})} dt\right)p(x)dx$$
$$= \int_0^\infty \left(\int_{\mathbb{R}^D} \mathbf{1}_{\{0<t<d(x,\mathcal{M})\}} p(x)dx\right)dt$$
$$= \int_0^\infty \Pr[d(X,\mathcal{M}) > t]dt \le \int_0^\infty c_1 e^{-\frac{t}{\sigma}}dt = c_1\sigma.$$

Similarly,

$$\int_{\mathbb{R}^D} d(x,\mathcal{M})^2 p(x)dx = \int_{\mathbb{R}^D}\left(\int_0^{d(x,\mathcal{M})} 2tdt\right)p(x)dx$$
$$= \int_0^\infty \left(\int_{\mathbb{R}^D} \mathbf{1}_{\{0<t<d(x,\mathcal{M})\}} p(x)dx\right)2tdt$$
$$= \int_0^\infty \Pr[d(X,\mathcal{M}) > t]2tdt \le \int_0^\infty c_1 e^{-\frac{t}{\sigma}}2tdt = 2c_1\sigma^2.$$

□

Proof of Lemma VII.5: By compactness and smoothness of $\mathcal{M}$, for any $x \in \mathbb{R}^D$, there exists $x^*_{\mathcal{M}} \in \mathcal{M}$ s.t. $\|x^*_{\mathcal{M}} - x\| = d(x,\mathcal{M})$. Thus,

$$|g(x)| \le |g(x^*_{\mathcal{M}})| + |g(x) - g(x^*_{\mathcal{M}})|$$
$$\le \sup_{x'\in\mathcal{M}} |g(x')| + \mathrm{Lip}(\xi)\|x - x^*_{\mathcal{M}}\|$$
$$= \|g\|_{L^\infty(\mathcal{M})} + \mathrm{Lip}(\xi)\cdot d(x,\mathcal{M}).$$

□

Proof of Lemma VII.6: Let $\mathrm{Lip}(f) \le L$, $f(x_0) = 0$, $\|x_0\| < R_0$, then

$$|f(x)| \le |f(x_0)| + L\|x - x_0\| \le L(R_0 + \|x\|), \quad \forall x,\ \|x\| > R,$$

Thus,

$$\int_{\|x\|>R} |f(x)|p(x)dx \le L\int_{\|x\|>R}(R_0 + \|x\|)p(x)dx$$
$$= L\left((R_0 + R)\int_{\|x\|>R} p + \int_{\|x\|>R}(\|x\| - R)p\right),$$

and

$$\int_{\|x\|>R}(\|x\| - R)p = \int_{\|x\|>R}\int_R^{\|x\|} dtp(x)dx$$
$$= \int_R^\infty \int_{\|x\|>t} p(x)dxdt$$
$$= \int_R^\infty \Pr[\|X\| > t]dt < \int_R^\infty Ce^{-t}dt = Ce^{-R},$$

then

$$\int_{\|x\|>R} |f|p < L(R_0 + R + 1)Ce^{-R}.$$

□

Proof of Lemma VII.7: By definition of $\mathcal{P}_\sigma$, $d(x_i, \mathcal{M})$'s are i.i.d. non-negative sub-exponential random variables, and

TABLE II

MEAN AND STANDARD DEVIATION (IN BRACKETS) OF $L[\hat{f}_{\text{TR}}]$, OVER 40 REPLICAS OF TRAINING OF THE NEURAL NETWORK, IN THE EXPERIMENT IN SECTION II-E AND FIGURE 2

n_{tr}	250	500	1000	2000	4000
H=4	0.0135 (0.0095)	0.0169 (0.0105)	0.0134 (0.0097)	0.0165 (0.0128)	0.0152 (0.0086)
H=8	0.0209 (0.0127)	0.0206 (0.0110)	0.0264 (0.0141)	0.0232 (0.0139)	0.0261 (0.0143)
H=16	0.0255 (0.0122	0.0315 (0.0129)	0.0311 (0.0139)	0.0355 (0.0124)	0.0310 (0.0149)
H=32	0.0309 (0.0105)	0.0339 (0.0107)	0.0388 (0.0096)	0.0386 (0.0123)	0.0410 (0.0094)
H=64	0.0324 (0.0189)	0.0367 (0.0086)	0.0416 (0.0086)	0.0439 (0.0081)	0.0447 (0.0058)
H=128	0.0334 (0.0100)	0.0422 (0.0046)	0.0458 (0.0032)	0.0481 (0.0031)	0.0485 (0.0030)
H=256	0.0297 (0.0177)	0.0442 (0.0053)	0.0478 (0.0022)	0.0503 (0.0017)	0.0511 (0.0008)
H=512	0.0299 (0.0123)	0.0433 (0.0068)	0.0477 (0.0040)	0.0504 (0.0011)	0.0508 (0.0011)

Example 1. Data in $\mathbb{R}$.

n_{tr}	250	500	1000	2000	4000
H=4	0.0182 (0.0116)	0.0208 (0.0111)	0.0219 (0.0115)	0.0212 (0.0103)	0.0189 (0.0115)
H=8	0.0245 (0.0115)	0.0301 (0.0117)	0.0335 (0.0101)	0.0283 (0.0100)	0.0299 (0.0116)
H=16	0.0240 (0.0159)	0.0357 (0.0085)	0.0387 (0.0083)	0.0360 (0.0102)	0.0389 (0.0097)
H=32	0.0327 (0.0086)	0.0386 (0.0061)	0.0427 (0.0059)	0.0418 (0.0069)	0.0453 (0.0034)
H=64	0.0322 (0.0085)	0.0395 (0.0051)	0.0442 (0.0021)	0.0461 (0.0027)	0.0461 (0.0051)
H=128	0.0237 (0.0202)	0.0400 (0.0066)	0.0452 (0.0029)	0.0475 (0.0024)	0.0481 (0.0023)
H=256	0.0234 (0.0173)	0.0388 (0.0081)	0.0466 (0.0025)	0.0488 (0.0020)	0.0498 (0.0020)
H=512	0.0095 (0.0239)	0.0398 (0.0049)	0.0471 (0.0026)	0.0498 (0.0016)	0.0505 (0.0009)

Example 2. Data near a 1D manifold in $\mathbb{R}^2$

$\mathbb{E}_{x\sim p} d(x, \mathcal{M}) < c_1\sigma$ by Lemma VII.4. The claims follows by the Bernstein's control of the positive tail for independent sum of i.i.d. sub-exponential random variables. □

Proof of Lemma VII.8: By definition,

$$\begin{aligned} 2L[f] &= \int p \log \frac{2}{1+e^{-f}} + \int q \log \frac{2}{1+e^{f}} \\ &= -\int p \log \frac{1+e^{-f}}{2} - \int q \log \frac{1+e^{f}}{2} \\ &\le -\int p \log e^{-\frac{f}{2}} - \int q \log e^{\frac{f}{2}} \\ &\qquad \text{(for any real number } \xi,\ \frac{1+e^{\xi}}{2} \ge e^{\frac{\xi}{2}}) \\ &= \int \frac{f}{2}(p-q) = \frac{1}{2}T[f]. \end{aligned}$$

□

Proof of Lemma VII.9:

$$\begin{aligned} \frac{1}{2}T[f] - 2L[f] &= \int p\frac{f}{2} - \int q\frac{f}{2} - \int p \log \frac{2e^{f}}{1+e^{f}} \\ &\qquad - \int q \log \frac{2}{1+e^{f}} \\ &= \int p \log \frac{1+e^{f}}{2e^{f/2}} + \int q \log \frac{1+e^{f}}{2e^{f/2}} \\ &= \int (p+q) \log \frac{e^{-f/2}+e^{f/2}}{2}, \end{aligned}$$

and by that $\frac{e^{x}+e^{-x}}{2} \le e^{x^2/2}$,

$$\frac{1}{2}T[f] - 2L[f] \le \int (p+q) \log e^{f^2/2} = \int \frac{f^2}{2}(p+q).$$

□

APPENDIX B
OPTIMIZATION EXPERIMENTS IN SECTION II-E

The experiment is merely to optimize the loss $L_{n,tr}$ using a dataset, and numerically compute the values of the obtained $L[\hat{f}_{tr}]$ using the analytical formula of the densities.

The training conducts Adam for $100 \cdot \frac{8000}{n_{tr}}$ epochs to ensure that same number of samples are processed in the experiment when n_{tr} changes. The batch size $= 100$, and learning rate 1e-3.

The numerical values of the mean and standard deviation of $L[\hat{f}_{tr}]$ for various H and n_{tr} is shown in Table II, where the values of the mean are plotted in Figure 2.

APPENDIX C
TWO-SAMPLE TESTS ON 1D DATA IN SECTION V-A

A. The Different Test Methods

We consider two types of alternative two-sample tests, which are

- (*net-acc*) The test based on classification accuracy [5]. The equivalent form as an IPM test is explained in C-B.
- (*gmmd*) Gaussian kernel MMD. The kernel bandwidth σ in *gmmd* is set to be the median of the pairwise distances among all samples [18].

Where the three tests, *net-logit* (the proposed), *net-acc* and *gmmd* all use the test set for two-sample problem, the first two network-based methods are trained on the stand-along training set. One may observe that this comparison to kernel MMD is not fair: First, kernel MMD with median-distance σ does not use the training set, thus it would be a more fair comparison if *gmmd* can use all the data samples without training-test splitting. Second, the median setting of σ may not be optimal and can be improved by existing methods

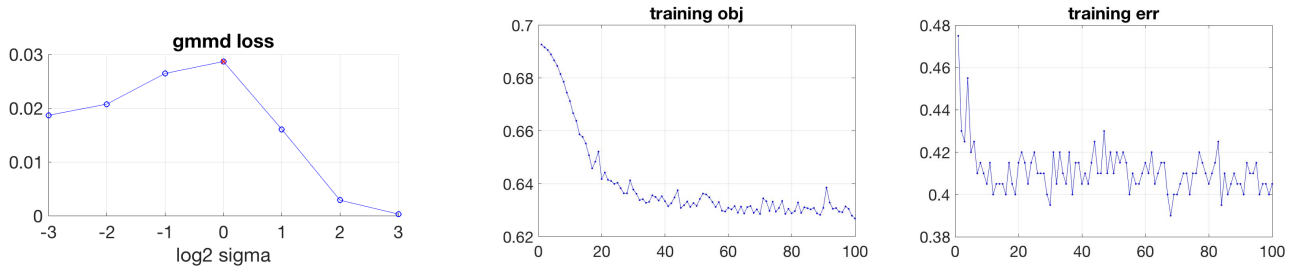


Fig. 12. Left: MMD discrepancy on trained set used by *gmmd-ad* to select kernel bandwidth σ. Middle and right: training of the classification network used in net-based tests. On the synthetic 1D dataset.

in literature. We thus consider three more variants of the Gaussian kernel MMD (GMMD) tests

- (*gmmd+*) GMMD using all samples. *gmmd* with median-distance σ which uses all the samples without training-test splitting.
- (*gmmd-ad*) GMMD using the test set only, but with adaptively selected bandwidth σ on the training set. The selection procedure is explained in C-C.
- (*gmmd++*) GMMD on the whole data sets with post-selected σ. The tests are conducted over a range of values of σ, which are $\{2^{-3}, 2^{-2}, \cdots 2^3\}$, and the best test power is post-selected. Note that this value of kernel bandwidth choice is not available in an algorithm, and the results are for theoretical comparison only.

These three tests are included for a more complete comparison between network-based tests and kernel tests. Training and testing split is half-and-half, and samples in X and Y are of the same number, in all cases.

B. Equivalent Form of Net-Acc Test

Here we show that the *net-acc* test studied in [5] is equivalent to using $\text{Sign}(f_\theta)$ instead of f_θ in (3) when $n_X = n_Y$, up to multiplying and adding constants. Specifically, by the definition of test statistic in [5], and recall that $|X_{te}| = |Y_{te}| = \frac{1}{2}|\mathcal{D}_{te}|$, $\text{Sign}(z) = 1$ if $z \geq 0$ and -1 if $z < 0$,

$$\begin{aligned}
\hat{T}_{\text{net-acc}} &= \frac{1}{2}\left(\frac{1}{|X_{te}|}\sum_{x\in X_{te}} \mathbf{1}_{\{f_\theta(x)\geq 0\}} \right.\\
&\qquad \left. + \frac{1}{|Y_{te}|}\sum_{y\in Y_{te}} \mathbf{1}_{\{f_\theta(y)<0\}}\right)\\
&= \frac{1}{2}\left(\frac{1}{|X_{te}|}\sum_{x\in X_{te}} \frac{1}{2}(1+\text{Sign}(f_\theta(x))) \right.\\
&\qquad \left. + \frac{1}{|Y_{te}|}\sum_{y\in Y_{te}} \frac{1}{2}(1-\text{Sign}(f_\theta(x)))\right)\\
&= \frac{1}{2} + \frac{1}{4}\left(\frac{1}{|X_{te}|}\sum_{x\in X_{te}} \text{Sign}(f_\theta(x)) \right.\\
&\qquad \left. - \frac{1}{|Y_{te}|}\sum_{y\in Y_{te}} \text{Sign}(f_\theta(x))\right). \qquad \text{(A.1)}
\end{aligned}$$

C. Adaptive Choice of σ in Gmmd-Ad

In the training phase, the algorithm computes the Gaussian kernel MMD discrepancy

$$\begin{aligned}
\hat{T}_{\text{MMD}}(X,Y) = &\frac{1}{|X|^2}\sum_{x,x'\in X} k_\sigma(x,x') + \frac{1}{|Y|^2}\sum_{y,y'\in Y} k_\sigma(y,y')\\
&- \frac{2}{|X||Y|}\sum_{x\in X,\, y\in Y} k_\sigma(x,y)
\end{aligned}$$

on the training set $X = X_{\text{tr}}$, $Y = Y_{\text{tr}}$, for a range of values of the kernel bandwidth σ, i.e. $\sigma = \{2^{-3}, \cdots, 2^3\}$. $k_\sigma(x,y) = \exp\{-\frac{|x-y|^2}{2\sigma^2}\}$ is the isotropic Gaussian kernel. A plot of MMD discrepancy as a function of varying σ is given in Figure 12. The σ which maximizes the MMD discrepancy on the training set is then chosen to compute the test statistic on the test set. The MMD test statistic also takes the form as $\hat{T}_{\text{MMD}}$ in [8], [18].

Note that the MMD loss may not be the optimal objective to select σ, and previous works have proposed to use the estimate of testing power as the optimization objective [13], [41], [42]. We use the MMD loss to select σ for simplicity. This method is also equivalent to the training process in [8] with only one trainable parameter which is the kernel bandwidth σ. In experiments, *gmmd-ad* improves the two-sample test performance over median choice on the 2D manifold dataset in Section V-C and the MNIST dataset in Section V-D.

D. Training of Neural Networks

In all the experiments, the classifier network used by *net-acc* and *net-logit* is a two-layer fully-connected neural network with 32 hidden nodes in each hidden layer, and the bottom layer has the same dimension as the input data. The training of the network is conducted via Adam [66]. Specifically, 100 epochs of Adam with learning rate 10^{-3}, and batch size 100 when the size of training set >100. A typical plot of evolution of training loss and training error is given in Figure 12. Training via SGD with momentum 0.9 produces similar result. The result is qualitatively the same when the number of hidden units varies from 16 to 1024. We have not investigated the optimal choice of network architecture hyperparameters for the two-sample problem.

We use fixed learning rate over a fixed number of epochs, and it is entirely possible that our training procedure is over simplified and better usage of stochastic gradient descent

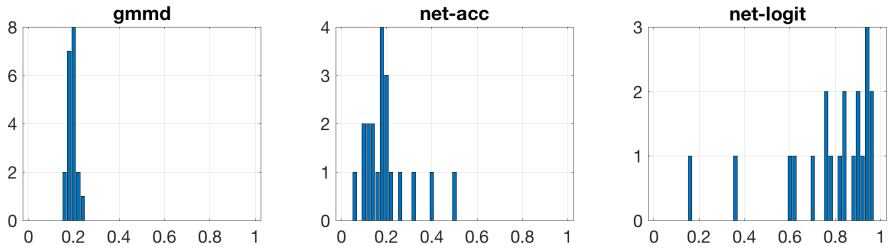


Fig. 13. Histogram of estimated test power from 400 test runs of *gmmd*, *net-acc* and *net-logit* over 20 replicas of training (no training for *gmmd*), on the example in Figure 5.

method as studied in [67]–[71] may lead to improved performance.

E. Test Power Estimation

The experiments with tests (1)-(4) use $n_{\text{run}} = 400$ test runs to estimate the power as the frequency of rejecting $\mathcal{H}_0$, and the whole experiments are repeated for $n_{\text{rep}} = 20$ replicas to compute mean and standard deviation of the estimated power. The test with (5) and (6) uses 200 test runs to estimate the power, since these gmmd methods demonstrate less variation in estimated power, explained as below.

The way of computing the test power is empirical and has randomness: for kernel mmd the variation is due to the finite number of runs (n_{run} times), and for network based tests there is extra variation due to the stochastic optimization of the network. Thus we use experiment replicas to recored the variations of the test power. The empirical distribution of the test power of the three methods over training replicas is given in 13, corresponding to the experiment in Figure 5. The plots of the two net-based methods indicate large variation of the power given by each trained network, that is, the "quality" of the trained net to discriminate the two densities varies. This instability is due to limited training samples as well as the randomness in the optimization algorithm.

We observe decreased power variation with larger training set, and the trained network gives better two-sample test power. This is consistent with the observation in Section II-E, and indicates that larger training set benefits the network training. However, as a price to pay, the testing set will be smaller given finitely many samples in total.

F. Detailed Experimental Results on Eg. 3

1) Set-Up: In Eg.3, the number $\delta \in [0, 1]$ controls the difference between the densities p and q. A plot for p and q with $\delta = 0.08$ is illustrated in Figure 5. 200 training and 200 testing samples are used, with half of the samples coming from X and half from Y in both the training and testing sets.

G. Test Power

The table in Figure 5 lists the power for the three methods (1)(2)(3), where *net-logit* gives significantly better average power about 80%, and the power of *net-acc* and *gmmd* are similar, both are about 20%. Table III gives the full table of test power including that of the methods *gmmd*+ and *gmmd*++.

TABLE III

THE MEAN, STANDARD DEVIATION ("STD") AND MEDIAN OF THE TEST POWER OF THE VARIOUS METHODS COMPUTED FROM $n_{run} = 400$ TEST RUNS OVER $n_{\text{REP}} = 20$ REPLICAS ON EG. 3 IN SECTION V-A. THE *gmmd*, *Net-Acc*, *Net-Logit* TESTS ARE COMPUTED ON $|X_{\text{TE}}| = |Y_{\text{TE}}| = 100$ SAMPLES, WHERE *Net-Acc* AND *Net-Logit* TRAIN A CLASSIFICATION NETWORK ON ANOTHER TRAINING SET OF SIZE $|X_{\text{TR}}| = |Y_{\text{TR}}| = 100$. *gmmd* ONLY USES THE TEST SET AND SETS THE KERNEL BANDWIDTH σ TO BE THE MEDIAN DISTANCE. *gmmd*+ AND *gmmd*++ ACCESSES BOTH THE TRAINING AND TEST SETS, WHERE *gmmd*+ USES THE MEDIAN DISTANCE AS σ, AND *gmmd*++ REPORTS THE BEST POWER OVER VARYING RANGE OF CHOICES OF σ, AS DESCRIBED IN C-A. THE RESULTS OF *gmmd*, *Net-Acc*, *Net-Logit* ARE ALSO REPORTED IN FIGURE 5

	gmmd	*gmmd+*	*gmmd++*	*net-acc*	*net-logit*
mean	19.14	46.63	57.29	19.98	78.09
std	1.95	2.49	1.578	10.43	20.56
median	19.63	47.13	57.38	17.63	84.13

where *gmmd*+ achieves a test power of 47% and *gmmd*++ a power of 57%, remaining inferior to *net-logit*, while both with small variation (std $\lesssim 2$) and thus are more stable than net-based tests. Results with other values of δ and sample sizes are reported in Figure 6.

The variation of the power is much larger for the two net-based tests, as explained in C-E and Figure 13. We note that such large variation is due to the instability of network training at small training size, and is likely to be a limitation of the current net-based methods.

APPENDIX D
TWO-SAMPLE TESTS ON 2D DATA IN SECTION V-C

We introduce the construction of $x_i \sim p$ and $y_j \sim q$. Let $x_i = T(u_i)$, $y_j = T(v_j)$ where $T : \mathbb{R}^2 \to \mathbb{R}^3$ is a smooth mapping from unit square to the spherical surface given by

$$T(x_1, x_2) = \frac{1}{R}\left(x_1, x_2, \sqrt{R^2 - x_1^2 - x_2^2}\right), \quad R = 1.5,$$

and u_i, v_j are i.i.d. copies of random variables u and v in $\mathbb{R}^2$ distributed as the following: $\epsilon = 0.05$,

$$u = t_u + \eta_u, \quad v = t_v + \eta_v, \quad \eta_u, \eta_v \sim \mathcal{N}(0, \epsilon^2 I_2),$$

$t_u \sim$ uniformly on a quarter circle in $[0, 1] \times [0, 1]$, and $t_v \sim$ the distribution of t_u rotated around $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ by angle δ, where the 4 random variables are all independent. The other experimental set-ups are the same as in Section V-A.

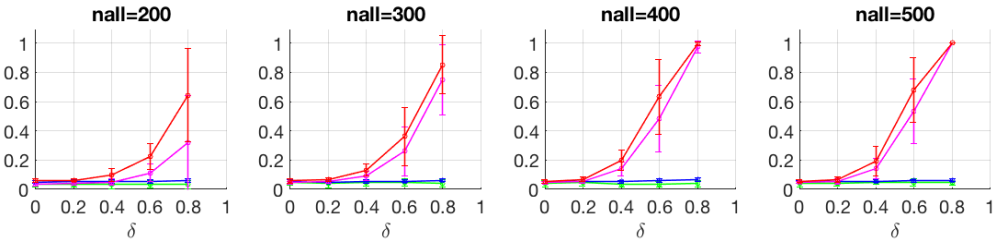


Fig. 14. Same plot as Figure 10 with another pre-trained generative model which produces faked images that are closer to the authentic ones.

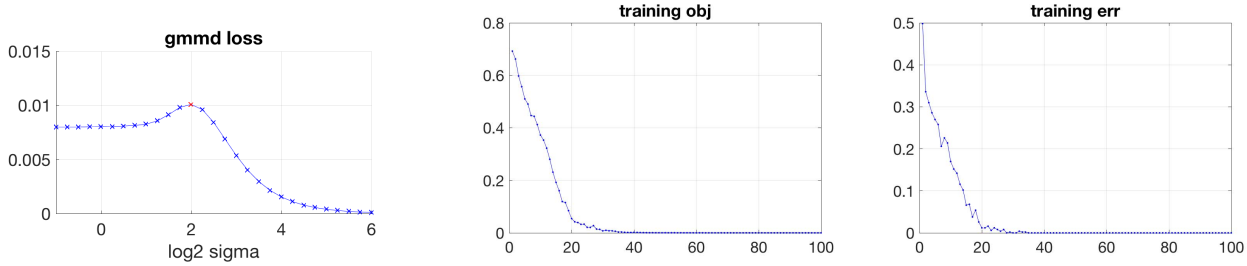


Fig. 15. Same plot as Figure 12 on MNIST data.

APPENDIX E
TWO-SAMPLE TESTS ON MNIST DATA IN SECTION V-D

The classifier network used in the experiment is a convolutional neural network (CNN) with 2 convolutional layers.

The pre-trained generated model is based on a convolutional auto-encoder:

c5x5x1x16 - re - ap 2x2 - c5x5x16x32 - re - ap 2x2 - fc128 - re
- fc10 - re $\leftarrow$ **code space** $\mathbb{R}^{10}$
- fc128 - re - ct 5x5x128x32 - re
- ct5x5x32x16 (upsample 2x2) - re - ct5x5x16x1 (upsample 2x2) - Euclidean loss

where "c" stands for convolutional layer, "ct" for transposed convolutional layers, "re" for Relu activation, and "ap" for average pooling. The auto-encoder is trained on 50000 MNIST dataset for 20 epochs using Adam with learning rate decreasing from 10^{-3} to 10^{-6} and batch size 100.

The sampling of generative model is conducted by adding a small isotropic Gaussian noise ("giggering") to the 10-dimensional codes of authentic MNIST digits computed by an encoder, and then mapping through the decoder to $\mathbb{R}^{784}$.

We also prepare another generative model by removing the bottleneck layer in the above auto-encoder architecture and retrain the model, which gives smaller reconstruction error and a higher-dimensional code space of $\mathbb{R}^{128}$. The generative model is conducted in the same way by sampling in the code space using Gaussian noise of smaller variance per coordinate. This produces faked images that are closer to the authentic ones in Euclidean distance in $\mathbb{R}^{784}$, however less explore the "manifold" of p_{data}. The test power of the four methods is shown in Figure 14.

The classification network used in *net-logit* is the following CNN

c5x5x1x16 - re - ap 2x2
- c5x5x16x32 - re - ap 2x2
- fc128 - re - fc2 - softmax loss

where dropout is used between the last 2 fully-connected layers. The classification CNN is trained for 100 epochs using Adam with learning rate 10^{-3} and batch size 100. A typical plot of evolution of training loss and training error is given in Figure 15.

The procedure of adaptive selection of σ by *gmmd-ad* is same as in Section V-A, where the bandwidth search range is $\sigma = \{2^{-1}, \cdots, 2^{6}\}$. The test power is evaluated on 400 test runs and the training is repeated for 20 replicas.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for helpful comments.

REFERENCES

[1] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing Statistical Hypotheses*, vol. 3. New York, NY, USA: Springer, 2005.
[2] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Scholkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, Jul. 2006.
[3] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton, "Fast two-sample testing with analytic representations of probability measures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1981–1989.
[4] W. Jitkrittum, Z. Szabó, K. P. Chwialkowski, and A. Gretton, "Interpretable distribution features with maximum testing power," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 181–189.
[5] D. Lopez-Paz and M. Oquab, "Revisiting classifier two-sample tests," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–15.
[6] X. Cheng, A. Cloninger, and R. R. Coifman, "Two-sample statistics based on anisotropic kernels," *Inf. Inference, A J. IMA*, vol. 9, no. 3, pp. 677–719, Sep. 2020.
[7] Y. Li, K. Swersky, and R. S. Zemel, "Generative moment matching networks," in *Proc. ICML*, 2015, pp. 1718–1727.
[8] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "MMD GAN: Towards deeper understanding of moment matching network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2203–2213.
[9] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[11] S. Nowozin, B. Cseke, and R. Tomioka, "F-GAN: Training generative neural samplers using variational divergence minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 271–279.

[12] J. R. Lloyd and Z. Ghahramani, "Statistical model criticism using kernel two sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 829–837.

[13] D. J. Sutherland *et al.*, "Generative models and model criticism via optimized maximum mean discrepancy," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–11.

[14] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *Proc. Workshop Conf.*, 2016, pp. 2606–2615.

[15] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 276–284.

[16] W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton, "A linear-time kernel goodness-of-fit test," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 262–271.

[17] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *J. Multivariate Anal.*, vol. 50, no. 1, pp. 41–54, 1994.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Jan. 2012.

[19] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet, "On integral probability metrics, Φ-divergences and binary classification," 2009, *arXiv:0901.2698*.

[20] M. Wornowizki and R. Fried, "Two-sample homogeneity tests based on divergence measures," *Comput. Statist.*, vol. 31, no. 1, pp. 291–313, Mar. 2016.

[21] M. Sugiyama, T. Kanamori, T. Suzuki, M. C. Du Plessis, S. Liu, and I. Takeuchi, "Density-difference estimation," *Neural Comput.*, vol. 25, no. 10, pp. 2734–2775, Oct. 2013.

[22] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 708–720, Sep. 2012.

[23] J. Schmidt-Hieber, "Deep ReLU network approximation of functions on a manifold," 2019, *arXiv:1908.00695*.

[24] M. Chen, W. Liao, H. Zha, and T. Zhao, "Statistical guarantees of generative adversarial networks for distribution estimation," 2020, *arXiv:2002.03938*.

[25] D. Yarotsky, "Error bounds for approximations with deep ReLU networks," *Neural Netw.*, vol. 94, pp. 103–114, Oct. 2017.

[26] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, "Optimal approximation with sparsely connected deep neural networks," *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 8–45, Jan. 2019.

[27] P. Petersen, M. Raslan, and F. Voigtlaender, "Topological properties of the set of functions generated by neural networks of fixed size," *Found. Comput. Math.*, vol. 21, no. 2, pp. 375–444, Apr. 2021.

[28] U. Shaham, A. Cloninger, and R. R. Coifman, "Provable approximation properties for deep neural networks," *Appl. Comput. Harmon. Anal.*, vol. 44, pp. 537–557, May 2018.

[29] M. Chen, H. Jiang, W. Liao, and T. Zhao, "Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery," 2019, *arXiv:1908.01842*.

[30] G. Mishne, U. Shaham, A. Cloninger, and I. Cohen, "Diffusion nets," *Appl. Comput. Harmon. Anal.*, vol. 47, no. 2, pp. 259–285, 2019.

[31] H. Li, O. Lindenbaum, X. Cheng, and A. Cloninger, "Variational diffusion autoencoders with random walk sampling," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 362–378.

[32] M. A. Arcones and E. Gine, "On the bootstrap of U and V statistics," *Ann. Statist.*, pp. 655–674, Jun. 1992.

[33] J. J. Higgins, *An Introduction to Modern Nonparametric Statistics*. Pacific Grove, CA, USA: Brooks/Cole, 2004.

[34] J. Friedman, "On multivariate goodness-of-fit and two-sample testing," Stanford Linear Accelerator Center, Menlo Park, CA, USA, Tech. Rep., 2004.

[35] M. D. Reid and R. C. Williamson, "Information, divergence and risk for binary experiments," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 731–817, Mar. 2011.

[36] I. Kim, A. Ramdas, A. Singh, and L. Wasserman, "Classification accuracy as a proxy for two-sample testing," *Ann. Statist.*, vol. 49, no. 1, pp. 411–434, Feb. 2021.

[37] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning for differing training and test distributions," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 81–88.

[38] S. Bickel, and M. Brückner, and T. Scheffer, "Discriminative learning under covariate shift," *J. Mach. Learn. Res.*, vol. 10, pp. 2137–2155, Sep. 2009.

[39] A. Menon and C. S. Ong, "Linking losses for density ratio and class-probability estimation," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 304–313.

[40] S. Mohamed and B. Lakshminarayanan, "Learning in implicit generative models," 2016, *arXiv:1610.03483*.

[41] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.

[42] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, "Learning deep kernels for non-parametric two-sample tests," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6316–6326.

[43] H. N. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.*, vol. 8, no. 1, pp. 164–177, Jan. 1996.

[44] Z. Shen, H. Yang, and S. Zhang, "Nonlinear approximation via compositions," *Neural Netw.*, vol. 119, pp. 74–84, Nov. 2019.

[45] H. Montanelli and Q. Du, "New error bounds for deep ReLU networks using sparse grids," *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 78–92, Jan. 2019.

[46] Z. Shen, "Deep network approximation characterized by number of neurons," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1768–1811, Jun. 2020.

[47] H. Lu and K. Kawaguchi, "Depth creates no bad local minima," 2017, *arXiv:1702.08580*.

[48] K. Kawaguchi and L. Kaelbling, "Elimination of all bad local minima in deep learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 853–863.

[49] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 33, pp. E7665–E7671, Aug. 2018.

[50] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, "On exact computation with an infinitely wide neural net," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8139–8148.

[51] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 322–332.

[52] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 242–252.

[53] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1675–1685.

[54] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, "Intrinsic dimension of data representations in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6109–6119.

[55] S. Rifai, Y. N. Dauphin, P. Vincent, Y. Bengio, and X. Müller, "The manifold tangent classifier," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2294–2302.

[56] E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo, and L. Wasserman, "Estimating the reach of a manifold," *Electron. J. Statist.*, vol. 13, no. 1, pp. 1359–1399, Jan. 2019.

[57] D. Yarotsky, "Optimal approximation of continuous functions by very deep ReLU networks," in *Proc. Conf. Learn. Theory*, 2018, pp. 639–649.

[58] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*. Cambridge, U.K.: Cambridge Univ. Press, 2018, vol. 47.

[59] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, vol. 162. Hoboken, NJ, USA: Wiley, 2009.

[60] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep ReLU networks," *Mach. Learn.*, vol. 109, no. 3, pp. 467–492, Mar. 2020.

[61] Y. Cao and Q. Gu, "Generalization bounds of stochastic gradient descent for wide and deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10836–10846.

[62] Z. Ji, M. Telgarsky, and R. Xian, "Neural tangent kernels, transportation mappings, and universal approximation," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–33.

[63] X. Cheng and Y. Xie, "Neural tangent kernel maximum mean discrepancy," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–13.

[64] C. Liu, L. Zhu, and M. Belkin, "On the linearity of large non-linear models: When and why the tangent kernel is constant," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–11.

[65] D. Deng and Y. Han, *Harmonic Analysis on Spaces of Homogeneous Type*. Springer, 2008.
[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[67] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Springer, 2010, pp. 177–186.
[68] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
[69] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.
[70] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 71–79.
[71] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

Xiuyuan Cheng received the Ph.D. degree from Princeton University in 2013. She held a post-doctoral position at the École Normale Supérieure in Paris and a Gibbs Assistant Professorship at Yale University. She joined the Duke Department of Mathematics in 2017. She is currently an Assistant Professor of mathematics at Duke University. She works on theoretical and computational techniques to solve problems in high-dimensional statistics, signal processing, and machine learning.

Alexander Cloninger received the Ph.D. degree in applied mathematics and scientific computation from the University of Maryland, College Park, in 2014. He was then an NSF Post-Doctoral Researcher and a Gibbs Assistant Professor of mathematics at Yale University. He joined UC San Diego (UCSD) in 2017. He is currently an Associate Professor of mathematics and the Halicioğlu Data Science Institute, UC San Diego. He researches problems in the areas of geometric data analysis, applied harmonic analysis, and machine learning.