Non-asymptotic Performance Guarantees for Neural Estimation of f-Divergences

Sreejith Sreekumar
Cornell University

Zhengxin Zhang
Cornell University

Ziv Goldfeld
Cornell University

Abstract

Statistical distances (SDs), which quantify the dissimilarity between probability distributions, are central to machine learning and statistics. A modern method for estimating such distances from data relies on parametrizing a variational form by a neural network (NN) and optimizing it. These estimators are abundantly used in practice, but corresponding performance guarantees are partial and call for further exploration. In particular, there seems to be a fundamental tradeoff between the two sources of error involved: approximation and estimation. While the former needs the NN class to be rich and expressive, the latter relies on controlling complexity. This paper explores this tradeoff by means of non-asymptotic error bounds, focusing on three popular choices of SDs—Kullback-Leibler divergence, chi-squared divergence, and squared Hellinger distance. Our analysis relies on non-asymptotic function approximation theorems and tools from empirical process theory. Numerical results validating the theory are also provided.

1 INTRODUCTION

Statistical distances (SDs) measure the discrepancy between probability distributions. A variety of machine learning (ML) tasks, from generative modeling (Arjovsky et al., 2017; Goodfellow et al., 2014; Kingma and Welling, 2014; Nowozin et al., 2016; Tolstikhin et al., 2018) to those relying on barycenters (Dognin et al., 2019; Gramfort et al., 2015; Rabin et al., 2011), can be posed as measuring or optimizing a SD between the data distribution and the model. Popular SDs include f-divergences (Ali and Silvey, 1966; Csiszár, 1967), integral probability metrics (IPMs) (Müller, 1997; Zolotarev, 1983), and Wasserstein distances (Villani, 2008). A common formulation that captures many of these is[1]

$$\mathsf{H}_{\gamma,\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \mathbb{E}_P[f] - \mathbb{E}_Q[\gamma\circ f], \tag{1}$$

where $\mathcal{F}$ is a function class of 'discriminators' and γ is sometimes called a 'measurement function' (cf., e.g., Arora et al. (2017)). This variational form is at the core of many ML algorithms implemented based on SDs, and has been recently leveraged for estimating SDs from samples.

Various non-parametric estimators of SDs are available in the literature (Kandasamy et al., 2015; Krishnamurthy et al., 2014; Liang, 2019; Perez-Cruz, 2008; Wang et al., 2005). These classic methods typically rely on kernel density estimation (KDE) or k-nearest neighbors (kNN) techniques, and are known to achieve optimal estimation error rates for specific SDs, subject to smoothness and/or regularity conditions on the densities. To mention a few, Kandasamy et al. (2015) proposed a KDE-based estimator which achieves the parametric mean squared error (MSE) rate for KL divergence estimation, provided that the densities are bounded away from zero and belong to a Hölder class of sufficiently large smoothness. For the special case of entropy estimation in the high smoothness regime, Berrett et al. (2019) proposed an asymptotically efficient weighted kNN estimator that does not rely on the boundedness from below assumption. Recently, Han et al. (2020) proposed a minimax rate-optimal entropy estimator for densities of sufficient Lipschitz smoothness. While these classic estimators achieve optimal performance under appropriate assumptions, they are often hard to compute in high dimensions.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

[1] Specifically, (1) accounts for f-divergences, IPMs and the 1-Wasserstein distance.

1.1 Statistical Distances Neural Estimation

Typical applications to machine learning, e.g., generative adversarial networks (GANs) (Arjovsky et al., 2017; Nowozin et al., 2016) or anomaly detection (Póczos et al., 2011), favor estimators whose computation scales well with number of samples and is compatible with backpropagation and minibatch-based optimization. A modern estimation technique that adheres to these requirements is the so-called neural estimation method (Arora et al., 2017; Belghazi et al., 2018; Zhang et al., 2018). Neural estimators (NEs) parameterize the discriminator class $\mathcal{F}$ in (1) by a neural network (NN), approximate expectations by sample means, and then optimize the obtained empirical objective. Denoting the samples from P and Q by $X^n := (X_1, \cdots, X_n)$ and $Y^n := (Y_1, \cdots, Y_n)$, respectively, the resulting NE is

$$\hat{\mathsf{H}}_{\gamma,\mathcal{G}_k}(X^n, Y^n) := \sup_{g\in\mathcal{G}_k} \frac{1}{n}\sum_{i=1}^{n}\Big[g(X_i) - \gamma\circ g(Y_i)\Big], \quad (2)$$

where $\mathcal{G}_k$ is the class of functions realizable by a k-neuron NN. Despite the popularity of NEs in applications, their theoretical properties and corresponding performance guarantees remain largely obscure. Addressing this deficit is the objective of this work.

There is a fundamental tradeoff between the quality of approximation by NNs and the sample size needed for accurate estimation of the parametrized form. The former is measured by the *approximation error*, $\big|\mathsf{H}_{\gamma,\mathcal{F}}(P,Q) - \mathsf{H}_{\gamma,\mathcal{G}_k}(P,Q)\big|$, whereas the latter by the *estimation error*, $\big|\hat{\mathsf{H}}_{\gamma,\mathcal{G}_k}(X^n,Y^n) - \mathsf{H}_{\gamma,\mathcal{G}_k}(P,Q)\big|$. While approximation needs $\mathcal{G}_k$ to be rich and expressive, efficient estimation relies on controlling its complexity. Past works on NEs provide only a partial account of estimation performance. Belghazi et al. (2018) proved consistency of mutual information neural estimation (MINE), which boils down to estimating KL divergence, but do not quantify approximation errors. Non-asymptotic sample complexity bounds for the parameterized form, i.e., when $\mathcal{F}$ in (1) is the NN class $\mathcal{G}_k$ to begin with, were derived in Arora et al. (2017); Zhang et al. (2018). These objects are known as NN distances and, by definition, overlook the approximation error with respect to (w.r.t.) the original SD. Also related is Nguyen et al. (2010), where KL divergence estimation rates are provided under the assumption that the approximating class is large enough to contain an optimizer of (1). This assumption is often violated in practice, e.g., when using NNs as done herein, or a reproducing kernel Hilbert space, as considered in Nguyen et al. (2010). This makes the quantification of the approximation error pivotal for a complete account of estimation performance. In light of the above, our objective is to derive non-asymptotic neural estimation performance bounds that characterize the dependence of the error on k and n, and help understand tradeoffs between them.

1.2 Contributions

We show that the effective (approximation plus estimation) error of a NE realized by a k-neuron shallow NN with bounded parameters and n samples scales like

$$O\left(k^{-1/2} + h_\gamma(k)n^{-1/2}\right),$$

where $h_\gamma(k)$ grows with k at a rate that depends on the estimated SD. In order to bound the approximation error, we refine Theorem 1 in Barron (1992) to show that a k-neuron NN with *bounded parameters* can approximate any function in the Barron class (Barron, 1993) under the sup-norm within an $O(k^{-1/2})$ error. To control the empirical estimation error, we leverage tools from empirical process theory and bound the associated entropy integral (Van Der Vaart and Wellner, 1996) to achieve the $O\left(h_\gamma(k)n^{-1/2}\right)$ convergence rate.

The effective error bound is then specialized to three predominant f-divergences: KL, chi-squared (χ^2 divergence, and squared Hellinger distance. We establish finite-sample absolute-error bounds of these NEs by identifying the appropriate scaling of the width k with the sample size n in the general bounds. This, in turn, implies consistency of the NEs. Our analysis is based on two key observations. First, to achieve a small approximation error, we would like $\mathcal{G}_k$ to universally approximate the original function class $\mathcal{F}$, which needs either width (Lu et al., 2017) or parameters (Stinchcombe and White, 1990) to be unbounded. On the other hand, to achieve the parametric estimation rate $n^{-1/2}$, the class $\mathcal{G}_k$ must not be too large. The effective error bound then relies on finding the appropriate scaling of k (and the uniform parameter norm) with n so that a small approximation error and fast estimation rates are both attained. Numerical results (on synthetic data) validating our theory are also provided.

Notation. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$, and $x \cdot y$ designate the inner product. The Euclidean ball of radius $r \geq 0$ centered at 0 is $B^d(r)$. We use $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ for the extended reals. For $1 \leq p < \infty$, the L^p space over $\mathcal{X} \subseteq \mathbb{R}^d$ w.r.t. the Lebesgue measure is denoted by $L^p(\mathcal{X})$, with $\|\cdot\|_p$ designating the norm. We let $(\Omega, \mathcal{A}, \mathbb{P})$ be the probability space on which all random variables are defined; $\mathbb{E}$ denotes the corresponding expectation. The class of Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $\mathcal{P}(\mathcal{X})$. To stress that an expectation of f is taken w.r.t. $P \in \mathcal{P}(\mathcal{X})$, we write $\mathbb{E}_P[f]$. We assume that all functions considered henceforth are Borel functions. The essential supremum of a function w.r.t. $P \in \mathcal{P}(\mathcal{X})$ is denoted by $\mathsf{ess\,sup}_P(f)$. For $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$, i.e., P is absolutely continuous w.r.t. Q, we use $\frac{\mathrm{d}P}{\mathrm{d}Q}$ for the Radon-Nikodym derivative of P w.r.t. Q. For $n \in \mathbb{N}$, $P^{\otimes n}$ denotes the n-fold product measure of P. For an open set $\mathcal{U} \subseteq \mathbb{R}^d$ and an integer $m \geq 0$, the class of functions such that all partial derivatives of order m exist and are continuous on $\mathcal{U}$ are denoted by $\mathsf{C}^m(\mathcal{U})$. In particular, $\mathsf{C}(\mathcal{U}) := \mathsf{C}^0(\mathcal{U})$ and $\mathsf{C}^\infty(\mathcal{U})$ denotes the class of continuous functions and infinitely differentiable functions on $\mathcal{U}$. The restriction of a function $f : \mathbb{R}^d \to \mathbb{R}$ to a subset $\mathcal{X} \subseteq \mathbb{R}^d$ is represented by $f|_\mathcal{X}$. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$, $D^{\boldsymbol{\alpha}} := \frac{\partial^{\alpha_1}}{\partial^{\alpha_1} x_1} \cdots \frac{\partial^{\alpha_d}}{\partial^{\alpha_d} x_d}$ denotes the partial derivative operator of order $|\boldsymbol{\alpha}| := \sum_{j=1}^d \alpha_j$.

2 Background and Preliminaries

Below, we provide a short background on the central technical ideas used in the paper.

Statistical distances. A common variational formulation of a SDs between $P, Q \in \mathcal{P}(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^d$, is

$$\mathsf{H}_{\gamma,\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P[f] - \mathbb{E}_Q[\gamma \circ f], \tag{3}$$

where $\gamma : \mathbb{R} \to \bar{\mathbb{R}}$, and $\mathcal{F}$ is a class of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ for which the expectations are finite. This formulation captures f-divergences (when γ is the convex conjugate of f), IPMs (for $\gamma(x) = x$) as well as the 1-Wasserstein distance (which is an IPM w.r.t. the 1-Lipschitz function class).

Approximated function class. Our approximation result requires the target function with domain $\mathcal{X}$ to have an extension on $\mathbb{R}^d$, which belongs to a certain class of functions introduced in Barron (1993).

Definition 1 (Barron class). *Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that has a Fourier representation $f(x) = \int_0^\infty e^{i\omega \cdot x} \tilde{F}(d\omega)$, where $\tilde{F}(d\omega)$ is a complex Borel measure over $\mathbb{R}^d$ with magnitude $F(d\omega)$ that satisfies*

$$B(f, \mathcal{X}) := \int_{\mathbb{R}^d} \sup_{x \in \mathcal{X}} |\omega \cdot x| \, F(d\omega) < \infty. \tag{4}$$

For $c \geq 0$, the Barron class is

$$\mathcal{B}_c(\mathcal{X}) := \left\{ f : \mathbb{R}^d \to \mathbb{R}, B(f, \mathcal{X}) \vee |f(0)| \leq c \right\}. \tag{5}$$

For $\tilde{f} : \mathcal{X} \to \mathbb{R}$, define

$$c_B^\star(\tilde{f}, \mathcal{X}) := \inf \left\{ c : \exists\, f \in \mathcal{B}_c(\mathcal{X}),\ \tilde{f} = f|_\mathcal{X} \right\}. \tag{6}$$

Stochastic processes. Our analysis of the estimation error requires the following definitions.

Definition 2 (Subgaussian process). *Let (Θ, d) be a metric space. A real-valued stochastic process $\{X_\theta\}_{\theta \in \Theta}$ with index set Θ is called subgaussian if it is centered and*

$$\mathbb{E}\left[e^{t(X_\theta - X_{\theta'})}\right] \leq e^{\frac{1}{2} t^2 d(\theta, \theta')^2},\ \forall\, \theta, \theta' \in \Theta,\ t \geq 0. \tag{7}$$

Definition 3 (Separable process). *A stochastic process $\{X_\theta\}_{\theta \in \Theta}$ on a metric space (Θ, d) is called separable if there exists a countable set $\Theta_0 \subseteq \Theta$, such that*

$$\lim_{\theta' \to \theta:\ \theta' \in \Theta_0} X_{\theta'} \to X_\theta,\ \forall\, \theta \in \Theta \qquad a.s. \tag{8}$$

Definition 4 (Covering number). *A set Θ' is an ϵ-covering for the metric space (Θ, d) if for every $\theta \in \Theta$, there exists a $\theta' \in \Theta'$ such that $d(\theta, \theta') \leq \epsilon$. The ϵ-covering number is*

$$N(\Theta, d, \epsilon) := \inf \left\{ |\Theta'| :\ \Theta' \textit{ is an } \epsilon\textit{-covering for } \Theta \right\}.$$

The next theorem gives a tail bound for the supremum of a subgaussian process in terms of the covering number. This result is key for our estimation error analysis.

Theorem 1. *(van Handel, 2016, Theorem 5.29) Let $\{X_\theta\}_{\theta \in \Theta}$ be a separable subgaussian process on the metric space (Θ, d). Then, for any $\theta_0 \in \Theta$ and $\delta \geq 0$, we have*

$$\mathbb{P}\left(\sup_{\theta \in \Theta} X_\theta - X_{\theta_0} \geq C \int_0^\infty \sqrt{\log N(\Theta, d, \epsilon)} d\epsilon + \delta \right) \\ \leq C e^{-\frac{\delta^2}{C \mathsf{diam}(\Theta)^2}}, \tag{9}$$

for $\mathsf{diam}(\Theta) := \sup_{\theta, \theta' \in \Theta} d(\theta, \theta')$ and a universal constant C.

3 Statistical Distances Neural Estimation

For simplicity of presentation, we henceforth fix $\mathcal{X} = [0, 1]^d$, although our results and analysis readily generalize to arbitrary compact supports $\mathcal{X} \subset \mathbb{R}^d$. Accordingly, $B(f, \mathcal{X})$, $\mathcal{B}_c(\mathcal{X})$ and $c_B^\star(\tilde{f}, \mathcal{X})$ are denoted by $B(f)$, $\mathcal{B}_c$ and $c_B^\star(\tilde{f})$, respectively. We first describe the neural estimation method, followed by two technical results that account for the approximation and the estimation errors. These results are later leveraged to derive effective error bounds for neural estimation of KL and χ^2 divergences, as well as the squared Hellinger distance. All proofs are deferred to the supplement.

$$\mathcal{G}_k(\mathbf{a}) := \left\{ g:\mathbb{R}^d \to \mathbb{R}: \begin{array}{c} g(x) = \sum_{i=1}^{k} \beta_i \phi\left(w_i \cdot x + b_i\right) + b_0,\ w_i \in \mathbb{R}^d,\ b_0, b_i, \beta_i \in \mathbb{R}, \\ \max_{\substack{i=1,\ldots,k \\ j=1,\ldots,d}} \left\{|w_{i,j}|, |b_i|\right\} \leq a_1,\ |\beta_i| \leq a_2,\ \ i = 1, \ldots, k,\ |b_0| \leq a_3 \end{array} \right\}. \tag{11}$$

$$\mathcal{H}^{l,\delta}_{b,c}(\mathcal{U}) := \left\{ f \in \mathsf{C}^l(\mathcal{U}): \begin{array}{l} \left|D^{\boldsymbol{\alpha}} f(x) - D^{\boldsymbol{\alpha}} f(x')\right| \leq c \left\|x - x'\right\|^{\delta},\ \forall x, x' \in \mathcal{U},\ |\boldsymbol{\alpha}| = l, \\ \max_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}| \leq l} \sup_{x \in \mathcal{U}} \left|D^{\boldsymbol{\alpha}} f(x)\right| \leq b \end{array} \right\}. \tag{14}$$

3.1 Neural Estimation

Let $P, Q \in \mathcal{P}(\mathcal{X})$. Consider a SD $\mathsf{H}_{\gamma,\mathcal{F}}(P,Q)$ between these distributions (see (3)), and assume that n independently and identically distributed (i.i.d.) samples $X^n := (X_1, \cdots, X_n)$ and $Y^n := (Y_1, \cdots, Y_n)$ from P and Q, respectively, are available. The NE of $\mathsf{H}_{\gamma,\mathcal{F}}(P,Q)$ based on a k-neuron shallow network (to parametrize the function class $\mathcal{F}$) and the samples X^n, Y^n (to approximate the expected values) is

$$\hat{\mathsf{H}}_{\gamma,\mathcal{G}_k(\mathbf{a})}(X^n, Y^n) := \sup_{g \in \mathcal{G}_k(\mathbf{a})} \frac{1}{n} \sum_{i=1}^{n} \Big[g(X_i) - \gamma \circ g(Y_i) \Big], \tag{10}$$

where $\mathcal{G}_k(\mathbf{a})$ is the NN class defined in (11) above, with parameter bounds specified by $\mathbf{a} = (a_1, a_2, a_3) \in \mathbb{R}^3_{\geq 0}$, and activation function $\phi : \mathbb{R} \to \mathbb{R}$, which is henceforth taken as the logistic sigmoid $\phi(x) = \frac{1}{1+e^{-x}}$. The results that follow extend to any measurable bounded variation sigmoidal (i.e., $\phi(z) \to 1$ as $z \to \infty$ and $\phi(z) \to 0$ as $z \to -\infty$) activation.

Our goal is to provide absolute-error performance guarantees for this NE, in terms of the approximation error and the statistical estimation error.[2]

3.2 Sup-norm Function Approximation

We start with a bound on the approximation error of a target function $\tilde{f}$ with domain $\mathcal{X}$ for which $c^\star_B(\tilde{f}) < \infty$.

Theorem 2 (Approximation). *Let $P, Q \in \mathcal{P}(\mathcal{X})$ and consider the NN class $\mathcal{G}^*_k(c) := \mathcal{G}_k\left(\sqrt{k}\log k, 2k^{-1}c, c\right)$ (see (11)) for some $c \geq 0$. Given $\tilde{f} : \mathcal{X} \to \mathbb{R}$ such that $c^\star_B(\tilde{f}) \leq c$, there exists $g \in \mathcal{G}^*_k(c)$ satisfying*

$$\left\|\tilde{f} - g\right\|_{\infty,P,Q} = O\left(k^{-\frac{1}{2}}\right), \tag{12}$$

where $\|f - g\|_{\infty,P,Q} := \operatorname{ess\,sup}_P |f-g| \vee \operatorname{ess\,sup}_Q |f-g|$. Moreover, for any $\mathbf{a} \in \mathbb{R}^3_{\geq 0}$, $c > 0$, and $\epsilon > 0$, there exists $\tilde{f} : \mathcal{X} \to \mathbb{R}$ with $c^\star_B(\tilde{f}) \leq c$ such that

$$\inf_{g \in \mathcal{G}_k(\mathbf{a})} \left\|\tilde{f} - g\right\|_2 = \Omega\left(k^{-\left(\frac{1}{2} + \frac{1}{d} + \epsilon\right)}\right). \tag{13}$$

The explicit dependence of d and c in the right hand side (R.H.S.) of (12) is given in (42).

The above theorem states that a k-neuron shallow NN can approximate a function $\tilde{f}$ on $\mathcal{X}$ within an $O(k^{-1/2})$ gap in the uniform norm, provided $\tilde{f}$ is the restriction of some f from the Barron class. The upper and lower bounds in (12) and (13) differ by $k^{-(1/d+\epsilon)}$, which becomes negligible for large d and small ϵ. Also observe that the lower bound is in terms of L^2 norm which implies a lower bound w.r.t. the L^∞ norm.

Remark 1 (Relation to previous results). *A result reminiscent to Theorem 2 appears in Barron (1992), but some technical details had to be adapted to apply the bound to neural estimation of SDs. Theorem 2 generalizes Barron (1992, Theorem 2) and Yukich et al. (1995, Theorem 2.2) from unbounded NN weights and bias parameters to bounded ones. We note that while Yukich et al. (1995, Theorem 2.2) allows unbounded parameters (input weight and bias), a more general problem of approximating a function and its derivatives is treated therein. Here we only consider the approximation of the function itself.*

We next show that a sufficiently smooth Hölder function on $\mathcal{X}$ is the restriction of some function in the Barron class. To that end we first define the Hölder function class.

Definition 5 (Holder class). *For $b, c, \delta \geq 0$, an integer $l \geq 0$, and an open set $\mathcal{U} \subseteq \mathbb{R}^d$, the bounded holder class $\mathcal{H}^{l,\delta}_{b,c}(\mathcal{U})$ is defined in (14) above.*

We have the following universal approximation property for Hölder functions.

Corollary 1 (Approximation of Hölder functions). *Given a function $\tilde{f} : \mathcal{X} \to \mathbb{R}$, suppose there exists an open set $\mathcal{U} \supset \mathcal{X}$, $b, c, \delta \geq 0$, and $f \in \mathcal{H}^{s,\delta}_{b,c}(\mathcal{U})$, $s := \lfloor \frac{d}{2} \rfloor + 2$, such that $\tilde{f} = f|_{\mathcal{X}}$. Then, there exists $g \in \mathcal{G}^*_k(\bar{c}_{b,c,d})$ such that*

$$\left\|\tilde{f} - g\right\|_{\infty,P,Q} = O\left(k^{-\frac{1}{2}}\right), \tag{15}$$

where $\bar{c}_{b,c,d}$ is given in (54) in Appendix A.2.

[2] In practice, an optimization error is also present, but its exploration is left for future work.

3.3 Estimation of Parameterized Distances

We next bound the error of estimating the parametrized SD $\mathsf{H}_{\gamma,\mathcal{G}_k(\mathbf{a})}(P,Q)$ (i.e., (3) for a NN function class) by its empirical version from (10). Throughout this section we assume that X^n and Y^n are, respectively, i.i.d. samples from P and Q.

Theorem 3 (Empirical estimation error tail bound)**.** *Let $P,Q \in \mathcal{P}(\mathcal{X})$. Assume that $\mathbf{a} \in \mathbb{R}^3_{\geq 0}$, $\mathcal{G}_k(\mathbf{a})$, and $\gamma : \mathbb{R} \to \bar{\mathbb{R}}$ are such that $\mathsf{H}_{\gamma,\mathcal{G}_k(\mathbf{a})}(P,Q) < \infty$ and*

$$\bar{\gamma}'_{\mathcal{G}_k(\mathbf{a})} := \sup_{\substack{x\in\mathcal{X},\\ g\in\mathcal{G}_k(\mathbf{a})}} \gamma' \circ g(x) < \infty, \tag{16}$$

where γ' is the derivative of γ. Then, for the universal constant C from Theorem 1 and any $\delta > 0$, we have

$$\mathbb{P}\Big(\Big|\hat{\mathsf{H}}_{\gamma,\mathcal{G}_k(\mathbf{a})}(X^n,Y^n) - \mathsf{H}_{\gamma,\mathcal{G}_k(\mathbf{a})}(P,Q)\Big| \geq \delta + CE_{k,\mathbf{a},n,\gamma}\Big) \leq 2Ce^{-\frac{n\delta^2}{V_{k,\mathbf{a},\gamma}}}, \tag{17}$$

where $E_{k,\mathbf{a},n,\gamma} = O\left(n^{-1/2}\right)$ and explicit expressions for $V_{k,\mathbf{a},\gamma}$ and $E_{k,\mathbf{a},n,\gamma}$ are given in (58)-(59) *in Appendix A.3.*

The proof of Theorem 3 (see Appendix A.3) involves upper bounding the estimation error by a separable subgaussian process and invoking Theorem 1.

Remark 2 (NN distances)**.** *The SD $\mathsf{H}_{\gamma,\mathcal{G}_k(\mathbf{a})}(P,Q)$ is the so-called NN distance, studied in Arora et al. (2017); Zhang et al. (2018) in the context of GANs. Theorem 3 can be understood as a sample complexity bound for NN distance estimation from data. Taking δ of order $n^{-1/2}$, the estimation error attains the parametric rate with high probability.*

3.4 f-Divergence Neural Estimation

Having Theorems 2-3 and Corollary 1, we analyze neural estimation of three important SDs: KL divergence, χ^2 divergence and squared Hellinger distance.

3.4.1 KL Divergence

The KL divergence between $P,Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$ is $\mathsf{D}_{\mathsf{KL}}(P\|Q) := \mathbb{E}_P\left[\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]$ (and infinite when P is not absolutely continuous w.r.t. Q). A variational form for $\mathsf{D}_{\mathsf{KL}}(P\|Q)$ is obtained via Legendre-Fenchel duality, yielding:

$$\mathsf{D}_{\mathsf{KL}}(P\|Q) = \sup_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[f] - \mathbb{E}_Q\left[e^f - 1\right], \tag{18}$$

where the supremum is over all measurable functions such that expectations are finite. This fits the framework of (3) with $\gamma(x) = \gamma_{\mathsf{KL}}(x) := e^x - 1$. The supremum in (18) is achieved by $f_{\mathsf{KL}} := \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)$.

Let $\hat{D}_{\mathcal{G}_k(\mathbf{a}_k)}(X^n,Y^n) := \hat{\mathsf{H}}_{\gamma_{\mathsf{KL}},\mathcal{G}_k(\mathbf{a}_k)}(X^n,Y^n)$ be a NE of $\mathsf{D}_{\mathsf{KL}}(P\|Q)$, where $\mathbf{a}_k \in \mathbb{R}^3_{\geq 0}$ for all $k \in \mathbb{N}$. The effective error achieved by the estimator can be bounded as the sum of the approximation and estimation errors.

To present error bounds, we require a few definitions. Let $\mathcal{P}_{\mathsf{KL}}(\mathcal{X})$ be the set of all pairs $(P,Q) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $P \ll Q$ and $\mathsf{D}_{\mathsf{KL}}(P\|Q) < \infty$, and set

$$\mathcal{I}(m) := \{f : \mathcal{X} \to \mathbb{R},\ c^\star_B(f) \vee \|f\|_\infty \leq m\}. \tag{19}$$

As a consequence of proof of Corollary 1, $\mathcal{I}(m)$ is nonempty since it contains any $f \in \mathcal{H}^{l,\delta}_{b,c}(\mathcal{U})$ for some $\mathcal{U} \supseteq \mathcal{X}$ and appropriately chosen parameters l,b,c,δ. For any m, the aforementioned condition is satisfied, e.g., by Gaussian densities with suitable parameters.

The following theorem establishes the consistency of $\hat{D}_{\mathcal{G}_k(\mathbf{a}_k)}(X^n,Y^n)$ and bounds the effective (approximation and estimation) error in terms of the NN and sample sizes, which reveals the tradeoff between them.

Theorem 4 (KL neural estimation)**.** *Let $(P,Q) \in \mathcal{P}_{\mathsf{KL}}(\mathcal{X})$. For any $\alpha > 0$:*

(i) If $f_{\mathsf{KL}} \in \mathsf{C}(\mathcal{X})$, then for $\{k_n\}_{n\in\mathbb{N}}$, n such that $k_n \to \infty$ and $k_n \leq (\frac{1}{2} - \alpha)\log n$,

$$\hat{D}_{\mathcal{G}_{k_n}(\mathbf{1})}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\mathsf{KL}}(P\|Q), \quad \mathbb{P}-a.s. \tag{20}$$

(ii) Suppose there exists an M such that $f_{\mathsf{KL}} \in \mathcal{I}(M)$. Then, for k and n such that $k^3 = O\left(n^{1-\alpha}\right)$,

$$\mathbb{E}\left[\left|\hat{D}_{\mathcal{G}^*_k(0.5\log k)}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}(P\|Q)\right|\right] = O\left(k^{-\frac{1}{2}} + k^{\frac{3}{2}}n^{-\frac{1}{2}}\right). \tag{21}$$

The consistency result (Part (i)) in the above theorem uses the fact that $\mathcal{G}_{k_n}(\mathbf{1})$ is a universal approximator for the class of continuous functions on compact sets as $k_n \to \infty$. The error bound in (21) utilizes Theorems 2-3 to bound the effective error as the sum of the approximation and estimation errors. From (12), the former error is $O(k^{-1/2})$ if $c^\star_B(f_{\mathsf{KL}}) \leq M$ and $\mathbf{a}$ is such that $\mathcal{G}^*_k(M) \subseteq \mathcal{G}_k(\mathbf{a})$. As M is often unavailable (due to P and Q being unknown), in order to achieve the above error, we take $\mathcal{G}_k(\mathbf{a}_k) = \mathcal{G}^*_k(m_k)$ for some increasing positive sequence $\{m_k\}_{k\in\mathbb{N}}$ ($m_k = 0.5\log k$ in Theorem 4 above) such that $m_k \to \infty$. This ensures that $m_k \geq M$ for sufficiently large k.

Remark 3 (KL approximation-estimation tradeoff)**.** *In Appendix B.1, we state the KL neural estimation error bound for an arbitrary increasing sequence $\{m_k\}_{k\in\mathbb{N}}$ (see* (92)*). If M (such that $f_{\mathsf{KL}} \in \mathcal{I}(M)$) is known when picking the NN parameters, then for a network of size $k = O\left(n^{(1-\alpha)}\right)$ with $m_k = M$, we have (see Remark 10 in Appendix B.1)*

$$\mathbb{E}\left[\left|\hat{D}_{\mathcal{G}^*_k(M)}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}(P\|Q)\right|\right]$$

$$\mathcal{L}_{\mathsf{KL}}(b,c) := \left\{ (P,Q) \in \mathcal{P}_{\mathsf{KL}}(\mathcal{X}) : \begin{array}{l} \exists\, f, \bar{f} \in \mathcal{H}^{s,\delta}_{b,c}(\mathcal{U}) \text{ for some } \delta \geq 0 \text{ and open set } \mathcal{U} \supset \mathcal{X} \ \text{ s.t.} \log p = f|_{\mathcal{X}}, \\ \log q = \bar{f}|_{\mathcal{X}} \end{array} \right\}. \tag{23}$$

$$= O\left(k^{-\frac{1}{2}} + \sqrt{k}\; n^{-\frac{1}{2}}\right). \tag{22}$$

Remark 4 (KL effective sample complexity)**.** *The optimal choice of k for* (22) *is $k = \sqrt{n}$ (for $\alpha < 0.5$). Inserting this into* (22)*, we obtain the effective error bound $O\left(n^{-1/4}\right)$. Although this rate is polynomial in n, it is slower than the parametric $n^{-1/2}$ rate that can be achieved for KL divergence estimation via KDE techniques in the very smooth density regime (Kandasamy et al., 2015).*[3]

Remark 5 (L^2 neural estimation of a function)**.** *A reminiscent analysis for the sample complexity of learning a NN approximation of a bounded range function from samples was employed in Barron (1994). This differs from our setup since SDs are given as a supremum over a function class as opposed to a single function. As such, our results require stronger sup-norm approximation results, as opposed to the L^2 bound used in Barron (1994).*

Theorem 4 provides conditions on f_{KL} under which bounds on the effective error of neural estimation can be obtained (namely, that $f_{\mathsf{KL}} \in \mathcal{I}(M)$ for some M). A primitive condition in terms of the densities of P and Q is given next. Let μ be a measure that dominates both P and Q, i.e., $P, Q \ll \mu$, and denote the corresponding densities by $p := \frac{\mathsf{d}P}{\mathsf{d}\mu}$ and $q := \frac{\mathsf{d}Q}{\mathsf{d}\mu}$.

Proposition 1 (KL sufficient condition)**.** *For $b, c \geq 0$, consider the class $\mathcal{L}_{\mathsf{KL}}(b,c)$ of pairs of distributions defined in* (23) *above. Suppose $(P,Q) \in \mathcal{L}_{\mathsf{KL}}(b,c)$. Then, Part (ii) of Theorem 4 and* (22) *hold with $M = 2\bar{c}_{b,c,d}$, with $\bar{c}_{b,c,d}$ as defined in Corollary 1.*

Remark 6. *[Feasible distributions] For appropriately chosen $b, c \geq 0$, the class $\mathcal{L}_{\mathsf{KL}}(b,c)$ contains distribution pairs $(P,Q) \in \mathcal{P}_{\mathsf{KL}}(\mathcal{X})$ whose densities w.r.t. a common dominating measure (e.g., $(P{+}Q)/2$) are bounded (from above and below) on $\mathcal{X}$ with a smooth extension on an open set covering $\mathcal{X}$. In particular, this includes uniform distributions, truncated Gaussians, truncated Cauchy distributions, etc.*

3.4.2 χ^2 Divergence

The χ^2 (chi-squared) divergence between $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$ is $\chi^2\left(P\|Q\right) = \mathbb{E}_Q\left[\left(\frac{\mathsf{d}P}{\mathsf{d}Q} - 1\right)^2\right]$. It admits the dual form:

$$\chi^2\left(P\|Q\right) = \sup_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[f] - \mathbb{E}_Q\left[f + f^2/4\right], \tag{24}$$

where the supremum is over all f such that expectations are finite. This dual form corresponds to (3) with $\gamma(x) = \gamma_{\chi^2}(x) := x + \frac{x^2}{4}$. The supremum in (24) is achieved by $f_{\chi^2} = 2\left(\frac{\mathsf{d}P}{\mathsf{d}Q} - 1\right)$.

Let $\hat{\chi}^2_{\mathcal{G}_k(\mathbf{a}_k)}(X^n, Y^n) := \hat{\mathsf{H}}_{\gamma_{\chi^2}, \mathcal{G}_k(\mathbf{a}_k)}(X^n, Y^n)$ denote the NE of $\chi^2\left(P\|Q\right)$. Set $\mathcal{P}_{\chi^2}(\mathcal{X})$ as the collection of all $(P,Q) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $P \ll Q$ and $\chi^2\left(P\|Q\right) < \infty$. The next theorem establishes consistency of the NE and bounds its effective absolute-error.

Theorem 5 (χ^2 neural estimation)**.** *Let $(P,Q) \in \mathcal{P}_{\chi^2}(\mathcal{X})$. For any $\alpha > 0$:*

(i) If $f_{\chi^2} \in \mathsf{C}\left(\mathcal{X}\right)$, then for $\{k_n\}_{n\in\mathbb{N}}$, n such that $k_n \to \infty$ and $k_n = O\left(n^{(1-\alpha)/5}\right)$,

$$\hat{\chi}^2_{\mathcal{G}_{k_n}(\mathbf{1})}(X^n, Y^n) \xrightarrow[n\to\infty]{} \chi^2\left(P\|Q\right), \quad \mathbb{P} - a.s. \tag{25}$$

(ii) Suppose there exists an M such that $f_{\chi^2} \in \mathcal{I}(M)$ (see (19)*). Then, for k and n such that $\sqrt{k}\log^2 k = O\left(n^{(1-\alpha)/2}\right)$, we have*

$$\mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}^*_k(0.5\log k)}(X^n, Y^n) - \chi^2\left(P\|Q\right)\right|\right] = O\left(k^{-\frac{1}{2}} + \sqrt{k}\log^2 k\; n^{-\frac{1}{2}}\right). \tag{26}$$

The proof of Theorem 5 (see Appendix C.1) is similar to that of Theorem 4.

Remark 7 (χ^2 effective sample complexity)**.** *In Appendix C.1, we obtain general error bounds (see* (105)*) assuming an arbitrary increasing sequence $\{m_k\}_{k\in\mathbb{N}}$, as mentioned in Remark 3. Given M with $f_{\chi^2} \in \mathcal{I}(M)$, for $m_k = M$ and $k = O\left(n^{(1-\alpha)}\right)$, we have*

$$\mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}^*_k(M)}(X^n, Y^n) - \chi^2\left(P\|Q\right)\right|\right] = O\Big(k^{-\frac{1}{2}} + \sqrt{k}\; n^{-\frac{1}{2}}\Big). \tag{27}$$

Comparing (25)*-*(27) *to* (20)*-*(22)*, we see that consistency holds under milder conditions and that the effective error bound is slightly better for χ^2 divergence than for KL divergence. As in Remark 4, the optimal choice of k in* (27) *is $k = \sqrt{n}$ (for $\alpha < 0.5$). This results in an effective error bound of $O(n^{-1/4})$.*

The next result is the counterpart of Proposition 1 to χ^2 divergence (see Appendix C.2 for proof).

[3] The latter relies on a different technical assumption in terms of Hölder-smoothness of underlying densities.

$$\mathcal{L}_{\chi^2}(b,c) := \left\{ (P,Q) \in \mathcal{P}_{\chi^2}(\mathcal{X}) : \begin{matrix} \exists\, f, \bar{f} \in \mathcal{H}^{s,\delta}_{b,c}(\mathcal{U}) \text{ for some } \delta \geq 0 \text{ and open set } \mathcal{U} \supset \mathcal{X} \text{ s.t. } p = f|_{\mathcal{X}}, \\ q^{-1} = \bar{f}|_{\mathcal{X}} \end{matrix} \right\}. \tag{28}$$

Proposition 2 (χ^2 sufficient condition)**.** *For $b, c \geq 0$, consider the class $\mathcal{L}_{\chi^2}(b,c)$ of pairs of distributions defined in* (28) *above, and suppose that $(P,Q) \in \mathcal{L}_{\chi^2}(b,c)$. Then, Part (ii) of Theorem 5 and* (27) *hold with $M = (2 + \bar{c}^2_{b,c,d} 2^{\lfloor d/2 \rfloor + 3})(\kappa_d \sqrt{d} \vee 1)$, where κ_d and $\bar{c}_{b,c,d}$ are given in* (36b) *and* (54)*, respectively.*

Remark 8 (Feasible distributions)**.** *The class $\mathcal{L}_{\chi^2}(b,c)$, for appropriately chosen $b, c \geq 0$, contains all $(P,Q) \in \mathcal{P}_{\chi^2}(\mathcal{X})$, whose densities p, q, w.r.t. a common dominating measure are bounded (upper bounded for p and bounded away from zero for q) on $\mathcal{X}$ with an extension that is sufficiently smooth on an open set covering $\mathcal{X}$. This includes the distributions mentioned in Remark 6.*

3.4.3 Squared Hellinger distance

The squared Hellinger distance between $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$ is $H^2(P,Q) := \mathbb{E}_Q\left[\left(\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}} - 1\right)^2\right]$, and

$$H^2(P,Q) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R}, \\ f(x)<1, \forall x \in \mathcal{X}}} \mathbb{E}_P[f] - \mathbb{E}_Q\big[f/(1-f)\big], \tag{29}$$

is its dual form, where the supremum is over all functions such that the expectations are finite ((29) corresponds to (3) with $\gamma(x) = \gamma_{H^2}(x) := \frac{x}{1-x}\Big)$. The supremum in (29) is achieved by $f_{H^2} = 1 - \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^{-1/2}$.

Let $\hat{H}^2_{\tilde{\mathcal{G}}_k(\mathbf{a}_k,t)}(X^n,Y^n) := \hat{\mathsf{H}}_{\gamma_{H^2}, \tilde{\mathcal{G}}_k(\mathbf{a}_k,t)}(X^n,Y^n)$, where $t > 0$ and $\tilde{\mathcal{G}}_k(\mathbf{a},t)$ is the NN class

$$\tilde{\mathcal{G}}_k(\mathbf{a},t) := \left\{ g : \mathbb{R}^d \to \mathbb{R} : \begin{matrix} g(x) = (1-t) \wedge \tilde{g}(x), \\ \tilde{g} \in \mathcal{G}_k(\mathbf{a}) \end{matrix} \right\}. \tag{30}$$

Set

$$\mathcal{I}_{H^2}(m) := \left\{ f : \mathcal{X} \to \mathbb{R}, \begin{matrix} c^\star_B(f) \vee \big\|(1-f)^{-1}\big\|_\infty \\ \vee \|f\|_\infty \leq m \end{matrix} \right\},$$

and $\mathcal{P}_{H^2}(\mathcal{X})$ as the collection of all $(P,Q) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $P \ll Q$ (note that $0 \leq H^2(P,Q) \leq 2$). Define the shorthands $\tilde{\mathcal{G}}^{(1)}_{k,t} = \tilde{\mathcal{G}}_k(\mathbf{1},t)$,

$$\tilde{\mathcal{G}}^{(2)}_{k,m,t} := \tilde{\mathcal{G}}_k\left(\sqrt{k}\log k, 2k^{-1}m, m, t\right).$$

The next theorem establishes consistency of the NE and bounds its effective absolute-error (see Appendix D.1 for proof).

Theorem 6 (H^2 neural estimation)**.** *Let $(P,Q) \in \mathcal{P}_{H^2}(\mathcal{X})$. For any $\alpha > 0$:*

(i) If $f_{H^2} \in \mathsf{C}(\mathcal{X})$, then, for $\{k_n, t_{k_n}\}_{n\in\mathbb{N}}$, such that $k_n \to \infty$, $t_{k_n} > 0$, $t_{k_n} \to 0$, and $k_n^{\frac{3}{2}} t_{k_n}^{-2} = O\left(n^{(1-\alpha)/2}\right)$,

$$\hat{H}^2_{\tilde{\mathcal{G}}^{(1)}_{k_n,t_{k_n}}}(X^n,Y^n) \xrightarrow[n\to\infty]{} H^2(P,Q), \quad \mathbb{P} - a.s. \tag{31}$$

(ii) Suppose there exists M such that $f_{H^2} \in \mathcal{I}_{H^2}(M)$. Then, for k, n with $\log^3 k\sqrt{k} = O\left(n^{(1-\alpha)/2}\right)$, $m_k = 0.5\log k$ and $t_k = \log^{-1} k$, we have

$$\mathbb{E}\left[\left|\hat{H}^2_{\tilde{\mathcal{G}}^{(2)}_{k,m_k,t_k}}(X^n,Y^n) - H^2(P,Q)\right|\right] = \left(\log k\ k^{-\frac{1}{2}}\right) + O\left(\log^3 k\sqrt{k}\ n^{-\frac{1}{2}}\right). \tag{32}$$

To establish effective error bounds for squared Hellinger distance, we used a truncated NN class $\tilde{\mathcal{G}}_k(\mathbf{a},t)$ given in (30), which is the function class obtained by saturating the shallow NN output to $1-t$ for some $t > 0$. This is done since $\gamma_{H^2}(x)$ has a singularity at $x = 1$ and the NN outputs must be truncated below 1 so as to satisfy (16) for bounding the empirical estimation error. For obtaining effective error bounds under this constraint, we scale the parameter t with k as $\{t_k\}_{k\in\mathbb{N}}$ for some decreasing positive sequence $t_k \to 0$. The bound in (32) uses $t_k = \log^{-1} k$.

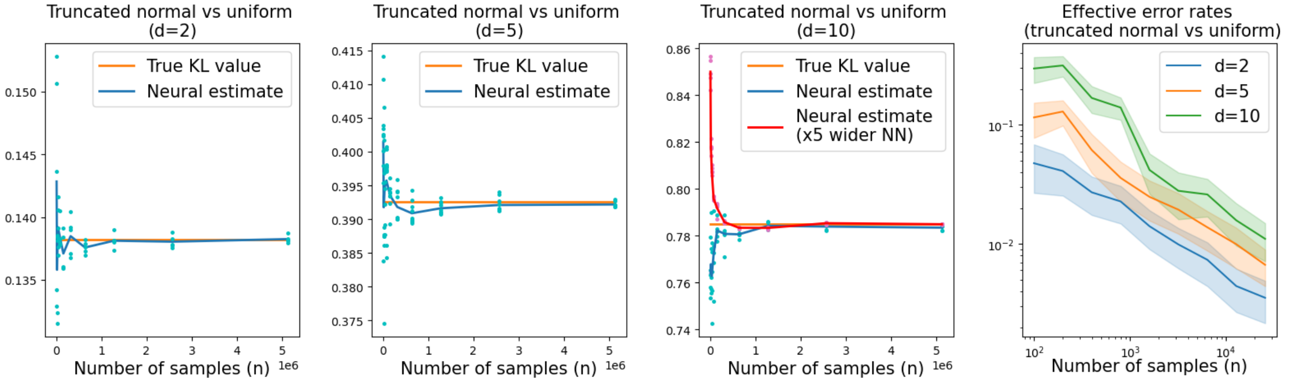
Remark 9 (Effective sample complexity)**.** *In Appendix D.1, we obtain effective error bounds (see* (119)*) for an arbitrary decreasing positive sequence $\{t_k\}_{k\in\mathbb{N}}$, with $t_k \to 0$, and an increasing positive divergent sequence $\{m_k\}_{k\in\mathbb{N}}$. If $f_{H^2} \in \mathcal{I}_{H^2}(M)$ and the NN parameters can depend on M, then, for k, $t_k = \log^{-1} k$ and n such that $\sqrt{k}\log^2 k = O\left(n^{(1-\alpha)/2}\right)$, setting $m_k = M$ in* (119) *yields*

$$\mathbb{E}\left[\left|\hat{H}^2_{\tilde{\mathcal{G}}^{(2)}_{k,M,t_k}}(X^n,Y^n) - H^2(P,Q)\right|\right] = O\Big(k^{-\frac{1}{2}}\log k\Big) + O\Big(\sqrt{k}\ \log^2 k n^{-\frac{1}{2}}\Big). \tag{33}$$

The optimal choice of k in (33) *is $k = n^{\frac{1}{2(1+\eta)}}$, $\big($for $\alpha < 0.5(1-2\eta)(1+\eta)^{-1}\big)$, where $\eta > 0$ is an arbitrarily small. The resulting effective error bound is $O(n^{-1/4})$.*

4 Empirical Results

We illustrate the performance of KL divergence neural estimation via some simple simulations. The considered NN class is $\mathcal{G}^*_k(M)$ (see Section 3.4) with M appropriately chosen. The number of samples n varies from $n = 10^5$ to $n = 6.4 \times 10^6$, and we scale the NN



(a) (b) (c) (d)

Figure 1: Neural estimate of KL divergence: (a) estimate versus n convergence in dimension $d = 2$, for P given by $\mathcal{N}(\mathbf{0}, \mathrm{I}_2)$ truncated to be supported inside $\mathcal{X} = [0.1, 2] \times [-1, 0]$ and $Q = \mathsf{Unif}(\mathcal{X})$; (b) estimate versus n convergence in dimension $d = 5$, for P given by $\mathcal{N}(\mathbf{0}, \mathrm{I}_5)$ truncated to $\mathcal{X} = [0.1, 2] \times [-1, 0] \times [2, 3] \times [-2, -1.5] \times [-1, 1]$ and $Q = \mathsf{Unif}(\mathcal{X})$; (c) similar to (b) but in dimension $d = 10$ and with compact support $\mathcal{X} \times \mathcal{X}$; (d) effective error rates versus number of samples.

size as $k = n^{1/5}$ (in accordance with $k = O\left(n^{1-\alpha}\right)$ for $\alpha > 0$ sufficiently small, see (22)). The NN is trained using Adam optimizer (Kingma and Ba, 2017) for 200 epochs. The initial learning rate of 10^{-2} is reduced to 10^{-3} after the first 100 epochs. We use batch size $n \times 10^{-3}$, and present plots averaged over 10 different runs (shown as dots).

Figure 1a shows convergence of the NE of $\mathsf{D}_{\mathsf{KL}}\left(P \| Q\right)$ versus number of samples, when P is a 2-dimensional truncated Gaussians (adhering to the compact support assumption) and Q is uniform distribution on the same support. For Figure 1a, we start from $\tilde{P} = \mathcal{N}(\mathbf{0}, \mathrm{I}_2)$, where $\mathrm{I}_d \in \mathbb{R}^{d \times d}$ is the identity, and truncate (and normalize) it to $\mathcal{X} = [0.1, 2] \times [-1, 0]$ to obtain P, and set $Q = \mathsf{Unif}(\mathcal{X})$. Figure 1b repeats the experiment but with P as a 5-dimensional Gaussian $\mathcal{N}(\mathbf{0}, \mathrm{I}_5)$ truncated to $\mathcal{X} := [0.1, 2] \times [-1, 0] \times [2, 3] \times [-2, -1.5] \times [-1, 1]$ and $Q = \mathsf{Unif}(\mathcal{X})$. The same setup but in dimension $d = 10$ and with $\mathcal{X} \times \mathcal{X}$ (instead of $\mathcal{X}$) is presented in Figure 1c (blue curve). Corresponding error rates versus number of samples (on a log-log scale) are shown in Figure 1d. It can be seen therein that the convergence rate is parametric for large enough values of n.

While convergence is evident in all dimensions, the trajectories are different: convergence happens from above when d is small and from below for large d (with $d = 5$ sitting in between and presenting a mixed trend). This happens because the same NN size $k = n^{1/5}$ were used in all three experiments, without factoring in the dimension (generally, higher-dimensional distribution need a larger NN). This results in the NN being relatively large when $d = 2$, which causes overfitting and, in turn, overestimation of the KL divergence for small n values. For $d = 10$, that same NN is relatively small, resulting in underestimation for small n. In accordance with the above, the $d = 5$ case exhibits a mixed trend. To verify this effect, we increased the NN size by a factor of 5 in the $d = 10$ experiment—the obtained neural estimator is shown by the red curve in Figure 1c. As expected, the larger networks results in convergence from above, similarly to the original $d = 2$ example.

5 Concluding Remarks

This paper studied neural estimation of SDs, aiming to characterize tradeoffs between approximation and empirical estimation errors. We showed that NEs of f-divergences, such as the KL and χ^2 divergences and the squared Hellinger distance, are consistent, provided the appropriate scaling of the NN size k with the sample size n. We then derived non-asymptotic absolute-error upper bounds that quantify the desired tradeoff between k and n. The key technical results leading to these bounds are Theorems 2-3, which, respectively, bound the sup-norm approximation error by NNs and the empirical estimation error of the parametrized SD.

Going forward, we aim to extend our results to additional SDs such as the total variation distance, the 1-Wasserstein distance, etc. While the high level analysis extends to these examples, new approximation bounds for the appropriate function classes (bounded or 1-Lipschitz) are needed. Another extension of interest is to P and Q that are not compactly supported. This is possible within our framework under proper tail decay, but we leave the details for future work. While we have neglected the optimization error from our current analysis, this is an important component of the overall estimation error and we plan to examine it in the future. Lastly, generalizing our analysis to NEs based on deep neural networks is another important extension. Through the results herein and the said future directions, we hope to provide useful performance guarantees for NEs that would facilitate a principled usage thereof in ML applications and beyond.

Acknowledgement

We thank the anonymous reviewers and meta-reviewer for their feedback which helped us improve the paper. The work of S. Sreekumar is supported by the TRIPODS Center for Data Science National Science Foundation Grant CCF-1740822. The work of Z. Goldfeld is supported in part by the NSF CRII Award under Grant CCF-1947801, in part by the 2020 IBM Academic Award, and in part by the NSF CAREER Award under Grant CCF-2046018.

References

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, Jan. 1966.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML-2017)*, pages 214–223, Sydney, Australia, Jul. 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the International Conference on Machine Learning (ICML-2017)*, pages 224–232, Sydney, Australia, Jul. 2017.

A. R. Barron. Neural net approximation. Proceedings of Seventh Yale Workshop on Adaptive and Learning Systems, CT, USA, 20–22 May 1992.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1): 115–133, Jan. 1994.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540, Stockholm Sweden, 10–15 Jul 2018.

T. B. Berrett, R. J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47 (1):288–318, 2019.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Ccientiarum Mathematicarum Hungarica*, 2: 229–318, 1967.

P. Dognin, I. Melnyk, Y. Mroueh, J. Ross, C. D. Santos, and T. Sercu. Wasserstein barycenter model ensembling. In *Proceedings of the International Conference on Learning Representations (ICLR-2019)*, New Orleans, Louisiana, US, May 2019.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS-2014)*, pages 2672–2680, 2014.

A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *Proceedings of the International Conference on Information Processing in Medical Imaging*, pages 261–272, Hong Kong, China, Jun. 2015.

Y. Han, J. Jiao, T. Weissman, and Y. Wu. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228 – 3250, 2020.

K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS-2015)*, pages 397–405, Montréal, Canada, 2015.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR-2014)*, Banff, Canada, Apr. 2014.

A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman. Nonparametric estimation of Rényi divergence and friends. In *Proceedings of the International Conference on Machine Learning (ICML-2014)*, pages 919–927, Beijing, China, Jun. 2014.

T. Liang. Estimating certain Integral Probability Metric (IPM) is as hard as estimating under the IPM. *arXiv preprint arXiv:1911.00730*, Nov. 2019.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS-2017)*, pages 6231–6239, Long Beach, CA, US, Dec. 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

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

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NeurIPS-2016)*, pages 271–279, Barcelona, Spain, Dec. 2016.

F. Perez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, pages 1666–1670, 2008.

B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 599–608. AUAI Press, 2011.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Proceedings of the International Conference on Scale Space and Variational Methods in Computer Vision (SSVM-2011)*, pages 435–446, Gedi, Israel, May 2011.

M. Stinchcombe and H. White. Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-1990)*, pages 7–16, San Diego, CA, US, Jun. 1990.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR-2018)*, Vancouver, Canada, Apr.-May 2018.

A. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, New York, 1996.

R. van Handel. *Probability in High Dimension: Lecture Notes-Princeton University.* [Online]. Available: `https://web.math.princeton.edu/~rvan/APC550.pdf`, 2016.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Q. Wang, S. R. Kulkarni, and S. Verdu. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

J. E. Yukich, M. B. Stinchcombe, and H. White. Sup-norm approximation bounds for networks through probabilistic methods. *IEEE Transactions on Information Theory*, 41(4):1021–1027, 1995.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discrimination-generalization tradeoff in GANs. In *Proceedings of the International Conference on Learning Representations (ICLR-2018)*, Vancouver, Canada, Apr.-May 2018.

V. M. Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2):264–287, 1983.