Journal of Machine Learning Research 23 (2022) 1-75 Submitted 10/21; Revised 2/22; Published 3/22

Neural Estimation of Statistical Divergences

Sreejith Sreekumar SREEJITHSREEKUMAR@CORNELL.EDU
Electrical and Computer Engineering Department
Cornell University
Ithaca, NY 14850, USA

Ziv Goldfeld GOLDFELD@CORNELL.EDU
Electrical and Computer Engineering Department
Cornell University
Ithaca, NY 14850, USA

Editor: Jean-Philippe Vert

Abstract

Statistical divergences (SDs), which quantify the dissimilarity between probability distributions, are a basic constituent of statistical inference and machine learning. A modern method for estimating those divergences relies on parametrizing an empirical variational form by a neural network (NN) and optimizing over parameter space. Such neural estimators are abundantly used in practice, but corresponding performance guarantees are partial and call for further exploration. We establish non-asymptotic absolute error bounds for a neural estimator realized by a shallow NN, focusing on four popular f-divergences—Kullback-Leibler, chi-squared, squared Hellinger, and total variation. Our analysis relies on non-asymptotic function approximation theorems and tools from empirical process theory to bound the two sources of error involved: function approximation and empirical estimation. The bounds characterize the effective error in terms of NN size and the number of samples, and reveal scaling rates that ensure consistency. For compactly supported distributions, we further show that neural estimators of the first three divergences above with appropriate NN growth-rate are minimax rate-optimal, achieving the parametric convergence rate.

Keywords: Approximation theory, minimax estimation, empirical process theory, f-divergence, neural estimation, neural network, statistical divergence, variational form.

1. Introduction

Statistical divergences (SDs) measure the discrepancy between probability distributions. A variety of inference tasks, from generative modeling (Kingma and Welling, 2014; Nowozin et al., 2016; Arjovsky et al., 2017; Tolstikhin et al., 2018; Goldfeld et al., 2020a; Nietert et al., 2021) to homogeneity/goodness-of-fit/independence testing (Kac et al., 1955; Zhang et al., 2018b; Hallin et al., 2021) 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) (Zolotarev, 1983; Müller, 1997), and Wasserstein distances (Villani, 2008; Santambrogio, 2015). A common formulation that captures many of these is[1]

$$\mathsf{D}_{h,\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[h\circ f], \tag{1.1}$$

where $\mathcal{F}$ is a function class of 'discriminators' and h is sometimes called a 'measurement function' (cf., e.g., Arora et al., 2017). This variational form is at the core of various learning algorithms implemented based on SDs (Nowozin et al., 2016; Arjovsky et al., 2017), and has been recently leveraged for estimating SDs from samples—a technique termed neural estimation. While neural estimators (NEs) are popular in practice due to their computational scalability, a theoretic account of corresponding performance guarantees is missing. To address the deficit, this work provides a through study of consistency and non-asymptotic absolute error bounds for NEs realized by shallow neural networks (NNs).

1.1 Neural Estimation of Statistical Divergences

Typical applications to machine learning, e.g., generative adversarial networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017) or anomaly detection (Póczos et al., 2011; Zenati et al., 2018; Schlegl et al., 2019), favor estimators whose computation scales well with number of samples and is compatible with backpropagation and minibatch-based optimization. Neural estimation is a modern technique that adheres to these requirements (Arora et al., 2017; Zhang et al., 2018a; Belghazi et al., 2018; Mroueh et al., 2021). Neural estimators (NEs) parameterize the discriminator class $\mathcal{F}$ in (1.1) by a NN, approximate expectations by sample means, and then optimize the resulting empirical objective over parameter space. Denoting the samples from μ and ν by $X^n := (X_1,\ldots,X_n)$ and $Y^n := (Y_1,\ldots,Y_n)$, respectively, the said NE is

$$\hat{\mathsf{D}}_{h,\mathcal{G}}(X^n,Y^n) := \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\Big[g(X_i) - h\circ g(Y_i)\Big], \tag{1.2}$$

where $\mathcal{G}$ is the class of functions realized by a NN.

The performance of a NE is dictated by the quality of the NN approximation to the original function class $\mathcal{F}$ from (1.1), and the sample size needed to accurately estimate the parametrized form $\mathsf{D}_{h,\mathcal{G}}(\mu,\nu)$. The former is measured by the *approximation error*, $\left|\mathsf{D}_{h,\mathcal{F}}(\mu,\nu) - \mathsf{D}_{h,\mathcal{G}}(\mu,\nu)\right|$, whereas the latter by the *empirical estimation error*, $\left|\hat{\mathsf{D}}_{h,\mathcal{G}}(X^n,Y^n) - \mathsf{D}_{h,\mathcal{G}}(\mu,\nu)\right|$. While approximation needs $\mathcal{G}$ 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, 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 (1.1) is the NN class $\mathcal{G}$ to begin with, were derived in (Arora et al., 2017; Zhang et al., 2018a). These objects are known as NN distances and, by definition, overlook the approximation error. 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 (1.1). This assumption is often violated in

1. Specifically, (1.1) accounts for f-divergences, IPMs and the 1-Wasserstein distance.

practice, e.g., when using a NN class as done herein, or a reproducing kernel Hilbert space, as considered in (Nguyen et al., 2010).

Quantification of the approximation error, alongside the empirical estimation error, is pivotal for a complete account of neural estimation performance. This work thus studies non-asymptotic effective (approximation plus empirical estimation) error bounds for NEs realized by a k-neuron shallow NN and n samples from each distribution. Results are specialized to four popular f-divergences: Kullback-Leibler (KL), chi-squared (χ^2), squared Hellinger (H^2) distance, and total variation (TV) distance.

1.2 Contributions

This work extends our earlier conference paper (Sreekumar et al., 2021), where the first non-asymptotic effective error bounds for NEs of f-divergences was derived. Consistency results for appropriate scaling rates of the NN and the sample sizes were also provided. However, the analysis therein resulted in sub-optimal error rates, only considered compactly supported distributions, and was not applicable for TV distance estimation. These aspects are key for a complete account of the neural estimation performance, and serve to motivate the present work, which closes all the aforementioned gaps.

We first consider compactly supported distributions and show that the effective error of a NE based on k neurons and n samples for the KL divergence, χ^2 divergence, or the H^2 distance scales as

$$O\left(k^{-1/2}+n^{-1/2}\right). \tag{1.3}$$

Our bound is sharp in the sense that by scaling k proportional to n, NEs achieve minimax optimality, converging at the parametric $n^{-1/2}$ rate. The results assume a spectral norm bound on the optimal potential (i.e., maximizer of (1.1)) of the SD, which, in particular, is satisfied when the distributions have sufficiently smooth densities . Notably, this condition suffices to avoid the so-called curse of dimensionality (CoD) and attain parametric rates that do not degrade exponentially with dimension.[2]

The derivation of (1.3) relies on two key technical results that separately account for the approximation and estimation errors. The first is a sup-norm $O(k^{-1/2})$ universal approximation bound for shallow NNs (Klusowski and Barron, 2018), and the second is a $O(n^{-1/2})$ bound on the empirical estimation error of the parametrized form. Derivation of the latter result leverages tools from empirical process theory and bounds the entropy integral (Van Der Vaart and Wellner, 1996) associated with the NN class. To that end, we bound the covering number of the NN class by noting that it can be represented as (a subset of) the symmetric convex hull (Van Der Vaart and Wellner, 1996)) of a composition of a monotone function with a VC subgraph class.

Equipped with these results, we treat neural estimation of the KL and χ^2 divergences, and the H^2 and TV distances. We establish consistency and obtain (1.3) as a finite-sample absolute-error bound by combining the approximation and empirical estimation bounds and identifying the appropriate scaling of the NN width k and parameter norms with the sample size n for each f-divergence. Our analysis results in the parametric absolute-error convergence rate for the NEs of KL divergence, χ^2 divergence, and H^2 distance. We also show

2. A similar behavior was observed in (Kandasamy et al., 2015) for classic f-divergence estimators between densities with high (Hölder) smoothness.

an $\Omega(n^{-1/2})$ lower bound on the minimax absolute error risk for KL divergence estimation problem by reducing it to differential entropy estimation and using the lower bound from Goldfeld et al. (2020b) for the latter problem. This establishes minimax optimality of KL divergence NE, with similar claims holding for NEs of χ^2 divergence and H^2 distance. Our method also accounts for the mutual information neural estimator (MINE) (Belghazi et al., 2018), and provides the first non-asymptotic effective error bound and minimax optimality claim for it. Different from these, the TV distance NE requires a modified approach because the spectral norm of the optimal potential is infinite. To circumvent the issue, we apply Gaussian smoothing to this potential, and control the approximation error as the smoothing parameters shrinks with the NN size k. This results in an approximation-estimation error bound that depends on dimension, i.e., the CoD applies in this case.

We then extend our results to distributions with unbounded support. To that end, we exploit the fact that our approximation error bound depends on the support of the target function only via its spectral norm. Thus, bounds on the effective error in the unbounded case are obtained by quantifying the spectral norm of the optimal potential inside a ball and growing its radius appropriately with k. The resulting bound depends on the scaling of the radius and the tail decay of the underlying distributions (as quantified by the Orlicz norm of the densities). The results are specialized to the aforementioned divergences, focusing on Gaussian and sub-Gaussian distributions. We note that our analysis applies to distributions whose densities need not be bounded away from zero (see also, Berrett et al., 2019; Berrett and Samworth, 2019)—an assumption that is often imposed for f-divergence estimation.

1.3 Related Work

Many non-parametric estimators of SDs are available in the literature (Wang et al., 2005; Perez-Cruz, 2008; Sriperumbudur et al., 2012; Krishnamurthy et al., 2014; Singh and Póczos, 2014a,b; Kandasamy et al., 2015; Singh and Póczos, 2016; Noshad et al., 2017; Moon et al., 2018; Wisler et al., 2018; Berrett et al., 2019; Berrett and Samworth, 2019; Liang, 2019; Han et al., 2020). These estimators typically rely on classic methods (or their variants) such as plug-in, 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 (see Remark 21). However, kernel-based methods usually require at least one of the following to achieve optimal rates: (i) boundary bias correction mechanism such as the usage of a mirror image kernel which assumes knowledge of the boundaries of the support of the distributions (cf., e.g., Singh and Póczos, 2014a,b); (ii) assumptions ensuring smooth behaviour of densities at the boundaries (cf., e.g., Moon et al., 2018; Han et al., 2020). As will be evident later, smoothness assumptions imposed by the aforementioned spectral norm condition are sufficient for (1.3) to hold for the NE. Thus, knowledge of the support of distributions or boundary bias correction is not needed.

Focussing on NEs, the tradeoffs between approximation and estimation errors was previously studied for non-parametric regression using NNs (cf., e.g., Barron, 1994; Bach, 2017; Suzuki, 2019). The goal there is to fit the best NN proxy to an (unknown) target function based on data generated from it by minimizing a prescribed loss function. Assuming that the target function satisfies certain smoothness or spectral norm constraints, the approximation-estimation tradeoff in such problems has been analyzed for different loss functions. In particular, Barron (1994) derived upper bounds on the minimax mean squared error rate for shallow NN models under a spectral norm condition on the Fourier transform of the target function. Density estimation under general loss functions was considered in (Yang and Barron, 1999), where minimax rate bounds in terms of covering/packing entropy were established. In (Suzuki, 2019), the minimax rate for non-parametric regression using deep NNs (DNNs) when the target function is Besov was determined. More recently, (Uppal et al., 2019) established the minimax rate for density estimation under a so-called Besov IPM loss.

1.4 Organization

The paper is organized as follows. Section 2 provides background and preliminary definitions. Technical results characterizing the approximation error and empirical estimation error are stated in Section 3. In Section 4, we apply these results to obtain upper bounds on the neural estimation error of the aforementioned f-divergences. Corresponding error bounds for distributions with unbounded support are the topic of Section 5. Section 6 provides concluding remarks and discusses future research directions. Proofs are deferred to appendices.

2. Background and Definitions

2.1 Notation

Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$ and $x \cdot y$ designate the inner product. The ℓ^m ball of radius $r \geq 0$ in $\mathbb{R}^d$ centered at 0 is $B_d^m(r)$; in particular, the Euclidean ball is designated as $B_d(r)$. We use $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ for the extended reals. For $1 \leq r < \infty$, the L^r space over $\mathcal{X} \subseteq \mathbb{R}^d$ with respect to (w.r.t.) the measure μ is denoted by $L^r(\mathcal{X}, \mu)$, with $\|\cdot\|_{r,\mu}$ representing the norm. When μ is the Lebesgue measure λ, we use the shorthand $L^r(\mathcal{X})$ with norm $\|\cdot\|_{r,\mathcal{X}}$, or even L^r and $\|\cdot\|_r$ when $\mathcal{X}$ is clear from the context. For $r = \infty$, we use $\|\cdot\|_{\infty,\mu}$ and $\|\cdot\|_{\infty,\mathcal{X}}$ for the essential supremum norm and the standard sup-norm, respectively. Slightly abusing notation, for $\mathcal{X} \subseteq \mathbb{R}^d$, we set $\|\mathcal{X}\| := \sup_{x \in \mathcal{X}} \|x\|_\infty$.

The probability space on which all random variables are defined is denoted by $(\Omega, \mathcal{A}, \mathbb{P})$ (assumed to be sufficiently rich), with $\mathbb{E}$ designating 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 the expectation of f is taken w.r.t. $\mu \in \mathcal{P}(\mathcal{X})$, we write $\mathbb{E}_\mu[f] := \int f \mathrm{d}\mu$. For $\mu, \nu \in \mathcal{P}(\mathcal{X})$ with $\mu \ll \nu$, i.e., μ is absolutely continuous w.r.t. ν, we use $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ for the Radon-Nikodym derivative of μ w.r.t. ν. For $n \in \mathbb{N}$, $\mu^{\otimes n}$ denotes the n-fold product measure of μ.

We assume that all functions are Borel measurable. For a multi-index $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{Z}^d_{\geq 0}$, the partial derivative operator of order $\|\alpha\|_1 := \sum_{j=1}^d \alpha_j$ is designated by $D^\alpha := \frac{\partial^{\alpha_1}}{\partial^{\alpha_1} x_1} \cdots \frac{\partial^{\alpha_d}}{\partial^{\alpha_d} x_d}$. 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. For $b \geq 0$ and an integer $m \geq 0$, $\mathsf{C}_b^m(\mathcal{U}) := \big\{ f \in \mathsf{C}^m(\mathcal{U}) : \max_{\alpha : \|\alpha\|_1 \leq m} \|D^\alpha f\|_{\infty,\mathcal{U}} \leq b \big\}$ denotes the subclass of $\mathsf{C}^m(\mathcal{U})$ with partial derivatives of order up to m uniformly bounded by b. The restriction of $f : \mathbb{R}^d \to \mathbb{R}$ to a subset $\mathcal{X} \subseteq \mathbb{R}^d$ is denoted by $f|_\mathcal{X}$. The Fourier transform of $f \in L^1(\mathcal{X})$ is denoted by $\mathfrak{F}[f]$. For a function class $\mathcal{F}$ and a function g, $g \circ \mathcal{F} := \{g \circ f : f \in \mathcal{F}\}$ and $|g \circ \mathcal{F}| := \{|g \circ f| : f \in \mathcal{F}\}$, where $\circ$ denotes function composition (domains assumed to be compatible for composition).

We denote universal constants by c (or c_1, c_2, etc.) while constants that depend on a parameter x are denoted by c_x. *The values of c and c_x may change between different instances even within the same line of an equation.* We use the shorthand $a \lesssim_x b$ for $a \leq c_x b$ for some $c_x > 0$ ($a \lesssim b$ means $a \leq cb$ for a universal constant $c > 0$); similarly, $a \asymp_x b$ stands for $a = c_x b$. We also employ standard asymptotic notations such as O, Ω, $\tilde{O}$, etc., where the tilde designates hidden logarithmic factors. For $a, b \in \mathbb{R}$, $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. We proceed with preliminary definitions and technical background.

2.2 Statistical Divergences

Let $\mathcal{X} \subseteq \mathbb{R}^d$. A common variational formulation of a SD between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\mathsf{D}_{h,\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[h \circ f], \tag{2.1}$$

where $h : \mathbb{R} \to \bar{\mathbb{R}}$, and $\mathcal{F}$ is a class of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ for which the last expectation is finite. This formulation captures f-divergences, IPMs (for $h(x) = x$), as well as the 1-Wasserstein distance (which is an IPM w.r.t. the 1-Lipschitz function class). We next specialize the above variational form to the f-divergences for which we derive neural estimation error bounds.

<u>KL divergence:</u> The KL divergence between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) := \begin{cases} \mathbb{E}_\mu\left[\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right], & \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases}$$

A variational form for $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu)$ is obtained via Legendre-Fenchel duality, yielding:

$$\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) = \sup_{f : \mathcal{X} \to \mathbb{R}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu\big[e^f - 1\big], \tag{2.2}$$

where the supremum is over all measurable functions such that the last expectation in (2.2) is finite. This fits the framework of (2.1) with $h(x) = h_{\mathsf{KL}}(x) := e^x - 1$. When $\mu \ll \nu$, the supremum in (2.2) is achieved by $f_{\mathsf{KL}} := \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

<u>χ^2 divergence:</u> The χ^2 divergence between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\chi^2(\mu\|\nu) := \begin{cases} \mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right)^2\right], & \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases}$$

It admits the dual form:

$$\chi^2(\mu\|\nu) = \sup_{f : \mathcal{X} \to \mathbb{R}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu\left[f + f^2/4\right], \tag{2.3}$$

where the supremum is over all measurable functions such that the last expectation in (2.3) is finite. This dual form corresponds to (2.1) with $h(x) = h_{\chi^2}(x) := x + x^2/4$ and the supremum is achieved by $f_{\chi^2} := 2\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right)$, whenever $\mu \ll \nu$.

Squared Hellinger distance: Let $\eta \in \mathcal{P}(\mathcal{X})$ be a probability measure that dominates both $\mu, \nu \in \mathcal{P}(\mathcal{X})$, i.e., $\mu, \nu \ll \eta$ (e.g., $\eta = (\mu+\nu)/2$), and denote the corresponding densities by $p = \frac{\mathrm{d}\mu}{\mathrm{d}\eta}$ and $q = \frac{\mathrm{d}\nu}{\mathrm{d}\eta}$. The squared Hellinger distance between μ, ν is[3]

$$\mathsf{H}^2(\mu,\nu) := \mathbb{E}_\eta\left[\left(\sqrt{p}-\sqrt{q}\right)^2\right]. \tag{2.4}$$

When $\mu \ll \nu$, the above expression can be written as

$$\mathsf{H}^2(\mu,\nu) = \mathbb{E}_\nu\left[\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}}-1\right)^2\right],$$

with the corresponding dual form

$$\mathsf{H}^2(\mu,\nu) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\\ f(x)<1, \forall x\in\mathcal{X}}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu\left[\frac{f}{1-f}\right], \tag{2.5}$$

where the supremum is over all functions such that the expectations are finite. This form corresponds to (2.1) with $h(x) = h_{\mathsf{H}^2}(x) := x/(1-x)$, and the supremum in (2.5) is achieved by $f_{\mathsf{H}^2} := 1 - \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^{-1/2}$. Also note that $\sqrt{\mathsf{H}^2}$ defines a metric on $\mathcal{P}(\mathcal{X})$ and that $0 \leq \mathsf{H}^2(\mu,\nu) \leq 2$, for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

Total variation distance: The TV distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is

$$\delta_{\mathsf{TV}}(\mu,\nu) := \sup_{\mathcal{C}} 2\left|\mu(\mathcal{C}) - \nu(\mathcal{C})\right|, \tag{2.6}$$

where the supremum is over all Borel subsets of $\mathcal{X}$. The corresponding variational form is

$$\delta_{\mathsf{TV}}(\mu,\nu) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R},\\ \|f\|_\infty\leq 1}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f], \tag{2.7}$$

which pertains to (2.1) with $h(x) = h_{\mathsf{TV}}(x) := x$. Denoting the densities of μ and ν w.r.t. a common dominating measure $\eta \in \mathcal{P}(\mathcal{X})$ by p and q, respectively, the supremum in (2.7) is achieved by $f_{\mathsf{TV}} := \mathbb{1}_{\mathcal{C}^*} - \mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*}$, where

$$\mathcal{C}^* := \left\{x \in \mathcal{X} : p(x) \geq q(x)\right\}. \tag{2.8}$$

Furthermore, δ_{TV} is a metric on $\mathcal{P}(\mathcal{X})$ with $0 \leq \delta_{\mathsf{TV}}(\mu,\nu) \leq 2$.

2.3 Stochastic Processes

Our analysis of the estimation error needs the following definitions.

Definition 1 (Sub-Gaussian process) *A real-valued stochastic process $(X_\theta)_{\theta\in\Theta}$ on a metric space (Θ, d) is sub-Gaussian if it is centered, i.e., $\mathbb{E}[X_\theta] = 0$ for all $\theta \in \Theta$, and*

$$\mathbb{E}\left[e^{t(X_\theta - X_{\tilde{\theta}})}\right] \leq e^{\frac{1}{2}t^2\mathsf{d}(\theta,\tilde{\theta})^2}, \quad \forall\, \theta, \tilde{\theta} \in \Theta,\ t \geq 0.$$

3. The standard definition of the squared Hellinger distance has an extra factor of 0.5. We use the current definition as it simplifies the statements of some results and proofs, while clearly having no effect on the qualitative conclusions. The same applies for the TV distance given in (2.6).

Definition 2 (Separable process) *A stochastic process $(X_\theta)_{\theta\in\Theta}$ on a metric space (Θ, d) is said to be separable if there exists a null set N and a countable subset $\Theta_0 \subseteq \Theta$, such that for every $\omega \notin N$ and $\theta \in \Theta$, there is a sequence $(\theta_m)_{m\in\mathbb{N}}$ in Θ_0 with $\mathsf{d}(\theta_m, \theta) \to 0$ and $X_{\theta_m}(\omega) \to X_\theta(\omega)$.*

Definition 3 (Covering and packing numbers) *Let (Θ, d) be a metric space.*

(i) A set $\Theta' \subseteq \Theta$ is an ϵ-covering of (Θ, d) if for every $\theta \in \Theta$, there exists $\tilde{\theta} \in \Theta'$ such that $\mathsf{d}(\theta, \tilde{\theta}) \le \epsilon$; the ϵ-covering number is $N(\epsilon, \Theta, \mathsf{d}) := \inf\{|\Theta'| : \Theta' \text{ is an } \epsilon\text{-covering of } \Theta\}$.

(ii) A set $\Theta' \subseteq \Theta$ is an ϵ-packing of (Θ, d) if $\mathsf{d}(\theta, \tilde{\theta}) > \epsilon$ for every $\theta, \tilde{\theta} \in \Theta'$ such that $\theta \neq \tilde{\theta}$; the ϵ-packing number is $T(\epsilon, \Theta, \mathsf{d}) := \sup\{|\Theta'| : \Theta' \text{ is an } \epsilon\text{-packing of } \Theta\}$.

2.4 Function Classes

Our approximation result requires the target function on $\mathcal{X}$ to have an extension to $\mathbb{R}^d$, whose spectral norm (as introduced in (Barron, 1993) and (Klusowski and Barron, 2018)) is finite. The class of functions with such bounded spectral norm is defined next.

Definition 4 (Approximation class) *Let $m \in \mathbb{N}$. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that has a Fourier representation $f(x) = \int_0^\infty e^{i\omega\cdot x} F(\mathrm{d}\omega)$, where $i = \sqrt{-1}$ is the imaginary unit and $F(\mathrm{d}\omega)$ is a complex Borel measure over $\mathbb{R}^d$ with magnitude $|F|(\mathrm{d}\omega)$ that satisfies*

$$S_m(f) := \int_{\mathbb{R}^d} \|\omega\|_1^m \, |F|(\mathrm{d}\omega) < \infty. \tag{2.9}$$

For $c \ge 0$, $m = 1, 2$, and $\mathcal{X} \subseteq \mathbb{R}^d$, define

$$\mathcal{B}_{c,m,\mathcal{X}}(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} : \|\mathcal{X}\| \, S_m(f) \vee |f(0)| \vee \|\nabla f(0)\|_1 \, \mathbb{1}_{\{m=2\}} \le c \right\},$$

and for $f : \mathcal{X} \to \mathbb{R}$, set

$$c^\star(f, m, \mathcal{X}) := \inf\left\{ c : \exists\ \tilde{f} \in \mathcal{B}_{c,m,\mathcal{X}}(\mathbb{R}^d),\ f = \tilde{f}|_{\mathcal{X}} \right\}.$$

We refer to $\mathcal{B}_{c,1,\mathcal{X}}(\mathbb{R}^d)$, $\mathcal{B}_{c,2,\mathcal{X}}(\mathbb{R}^d)$, $c^\star_{\mathsf{B}}(f, \mathcal{X}) := c^\star(f, 1, \mathcal{X})$ and $c^\star_{\mathsf{KB}}(f, \mathcal{X}) := c^\star(f, 2, \mathcal{X})$ as the Barron class, Klusowski-Barron class, Barron coefficient, and Klusowski-Barron coefficient, respectively.

For TV distance neural estimation, analysis of the NN approximation error for step functions is required. Such functions naturally belong to the Lipschitz function class defined below.

Definition 5 (Lipschitz class) *For $r \in (0, \infty]$, $m \in \mathbb{N}$, and $f \in L^r(\mathbb{R}^d)$, the m^{th} modulus of smoothness of f is*

$$\xi_{m,r}(f, t) := \sup_{u\in\mathbb{R}^d, \|u\|\le t} \|\Delta_u^m f\|_{r,\mathbb{R}^d}, \tag{2.10}$$

where $\Delta_u^m f(x) = \sum_{j=0}^m (-1)^{m-j} f(x + ju)$. For $\mathcal{X} \subseteq \mathbb{R}^d$ and $0 < s \le 1$, the Lipschitz class with smoothness parameter s is

$$\mathsf{Lip}_{s,r,b}(\mathcal{X}) := \left\{ f \in L^r(\mathbb{R}^d) : \|f\|_{\mathsf{Lip}(s,r)} \le b, \mathsf{supp}(f) = \mathcal{X} \right\},$$

where $\|f\|_{\mathsf{Lip}(s,r)} := \|f\|_r + \sup_{t>0} t^{-s} \xi_{1,r}(f, t)$ is the Lipschitz seminorm.

Note that norm in (2.10) is taken over $\mathbb{R}^d$ (despite the assumption that f nullifies outside of $\mathcal{X}$). For $d = 1$, the class of functions of bounded variation over $\mathcal{X} \subset \mathbb{R}$ is contained in $\cup_{b\in\mathbb{R}}\mathsf{Lip}_{1,1,b}(\mathcal{X})$.

The Vapnik-Chervonenkis (VC) type class of functions will play a prominent role in our empirical estimation error analysis.

Definition 6 (VC-type class) *Let $\mathcal{F}$ be a class of Borel measurable functions with domain $\mathcal{X}$ and a finite measurable envelope F, i.e., $\sup_{f\in\mathcal{F}}|f(x)| \leq F(x) < \infty,\ \forall x \in \mathcal{X}$. Then, $\mathcal{F}$ is a VC-type class with envelope F if there exists finite constants $l_{\mathsf{vc}}(\mathcal{F}) = l_{\mathsf{vc}}(\mathcal{F}, F)$ and $u_{\mathsf{vc}}(\mathcal{F}) = u_{\mathsf{vc}}(\mathcal{F}, F)$ such that*

$$\sup_{\gamma\in\mathcal{P}(\mathcal{X})} N\left(\epsilon \left\|F\right\|_{2,\gamma}, \mathcal{F}, \|\cdot\|_{2,\gamma}\right) \leq \left(l_{\mathsf{vc}}(\mathcal{F})\epsilon^{-1}\right)^{u_{\mathsf{vc}}(\mathcal{F})}, \quad \forall\ 0 < \epsilon \leq 1. \tag{2.11}$$

Finally, we introduce the function class of shallow NNs.

Definition 7 (NN class) *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a (non-linear) measurable activation function. The class of shallow NNs (i.e., with a single hidden layer) with k neurons and bounds on its parameters specified by $\mathbf{a} = (a_1, a_2, a_3, a_4) \in \mathbb{R}^4_{\geq 0}$ is*

$$\mathcal{G}_k(\mathbf{a}, \phi) := \left\{ g : \mathbb{R}^d \to \mathbb{R} : \begin{array}{c} g(x) = \displaystyle\sum_{i=1}^{k} \beta_i \phi\left(w_i \cdot x + b_i\right) + w_0 \cdot x + b_0, \\ \displaystyle\max_{1\leq i\leq k} \|w_i\|_1 \vee |b_i| \leq a_1,\ \ \max_{1\leq i\leq k} |\beta_i| \leq a_2,\ \ |b_0| \leq a_3, \|w_0\|_1 \leq a_4 \end{array} \right\}.$$

Let $\phi_{\mathsf{S}}(z) = (1 + e^{-z})^{-1}$ and $\phi_{\mathsf{R}}(z) = z \vee 0$ denote the logistic sigmoid[4] and the rectified linear unit (ReLU) activation functions, respectively. Further, for $a \geq 0$, define the shorthands $\mathcal{G}^{\mathsf{S}}_k(a) := \mathcal{G}_k\big(k^{1/2}\log k, 2k^{-1}a, a, 0, \phi_{\mathsf{S}}\big)$, $\mathcal{G}^{\mathsf{R}}_k(a) := \mathcal{G}_k\big(1, 2k^{-1}a, a, a, \phi_{\mathsf{R}}\big)$, and $\mathcal{G}^{*}_k(\phi) := \mathcal{G}_k\big(\mathbf{a}^*, \phi\big)$ with $\mathbf{a}^* = (1, 1, 1, 0)$. Throughout, we will assume $\phi \in \{\phi_{\mathsf{S}}, \phi_{\mathsf{R}}\}$.

2.5 Minimax Estimation Risk

To investigate the decision-theoretic fundamental limit of estimating a SD $\mathsf{D}_{h,\mathcal{F}}$ as defined in (2.1), we now define the minimax risk. Let $\mathcal{P}^2_{\mathcal{X}} \subseteq \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ be a class of pairs of distributions between which $\mathsf{D}_{h,\mathcal{F}}$ is finite and fix $(\mu, \nu) \in \mathcal{P}^2_{\mathcal{X}}$. Let $X^n := (X_1, \ldots, X_n)$ and $Y^n := (Y_1, \ldots, Y_n)$ be n independently and identically distributed (i.i.d.) samples from μ and ν, respectively.[5] An estimator of $\mathsf{D}_{h,\mathcal{F}}$ based on these samples is denoted by $\hat{\mathsf{D}}_{h,\mathcal{F}}(X^n, Y^n)$. The minimax absolute-error risk is

$$\mathcal{R}^{\star}_{h,\mathcal{F}}(n, \mathcal{P}^2_{\mathcal{X}}) := \inf_{\hat{\mathsf{D}}_{h,\mathcal{F}}} \sup_{(\mu,\nu)\in\mathcal{P}^2_{\mathcal{X}}} \mathbb{E}\left[\left|\mathsf{D}_{h,\mathcal{F}}(\mu, \nu) - \hat{\mathsf{D}}_{h,\mathcal{F}}(X^n, Y^n)\right|\right]. \tag{2.12}$$

4. The results that follow with ϕ_{S} as activation straightforwardly applies to any continuous monotone bounded activation, e.g., any sigmoidal activation with $\phi(z) \to 1$ as $z \to \infty$ and $\phi(z) \to 0$ as $z \to -\infty$.
5. For simplicity, we restrict attention to the case where an equal number of samples is available from both μ and ν, but our analysis readily extends to the mismatched scenario with the corresponding bounds obtained by replacing $n^{-1/2}$ by $(m^{-1} + n^{-1})^{1/2}$, where m denotes the number of samples from μ (say).

We explore the performance of the NE

$$\hat{\mathsf{D}}_{h,\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n,Y^n) := \sup_{g\in\mathcal{G}_k(\mathbf{a}_k,\phi)} \frac{1}{n}\sum_{i=1}^{n}\Big[g(X_i) - h\circ g(Y_i)\Big], \tag{2.13}$$

under the above framework. By appropriately scaling the NN size k (and parameter norm) with the sample size n, we show that NEs of KL and χ^2 divergences as well as H^2 distance converges at the parametric $n^{-\frac{1}{2}}$ rate uniformly over certain classes of distribution pairs satisfying regularity conditions. We further show (see, e.g., Corollary 18) that the minimax risk is at least $\Omega(n^{-1/2})$ over this class, thus establishing the minimax optimality of NEs.

3. Preliminary Technical Results

We next present two technical results that account for the NN approximation error and the empirical estimation error of the parametrized SD. These results are later leveraged to derive effective error bounds for neural estimation of KL and χ^2 divergences, squared Hellinger distance and TV distance.

3.1 Sup-norm Function Approximation

We start with a bound on the approximation error of a target function f with a compact domain $\mathcal{X}$ for which $c^\star(f,m,\mathcal{X})<\infty$, $m=1,2$. A reminiscent result for the case $m=1$ was given in (Barron, 1992), albeit without explicitly quantifying the dependence on dimension or addressing how the NN parameters scale with k. The bounds for $m=2$ are taken from (Klusowski and Barron, 2018).

Theorem 8 (Approximation error bound) *Let $\mathcal{X}$ be compact. Given $f:\mathcal{X}\to\mathbb{R}$ with $c^\star_{\mathsf{KB}}(f,\mathcal{X})\le a$, there exists $g\in\mathcal{G}_k^{\mathsf{R}}(a)$ such that*

$$\|f-g\|_\infty \lesssim a d^{\frac{1}{2}} k^{-\frac{1}{2}}. \tag{3.1}$$

Similarly, given $f:\mathcal{X}\to\mathbb{R}$ such that $c^\star_{\mathsf{B}}(f,\mathcal{X})\le a$, there exists $g\in\mathcal{G}_k^{\mathsf{S}}(a)$ satisfying (3.1).

The above theorem states that a k-neuron shallow NN can approximate a function f on $\mathcal{X}$ within an $O(k^{-1/2})$ gap in the sup-norm, provided f is the restriction of some $\tilde{f}$ from the Barron class or Klusowski-Barron class. The bound in (3.1) follows from (Klusowski and Barron, 2018, Theorem 2), up to rescaling the domain therein. The proof of the second claim pertaining to approximation by NN class $\mathcal{G}_k^{\mathsf{S}}(a)$ is provided in Appendix A.1.1, and is based on ideas from (Barron, 1992, 1993; Yukich et al., 1995). The error bounds stated in Theorem 8 are representative of the approximation capabilities of shallow NNs with ReLU (unbounded) and sigmoid (bounded) activations, respectively. Note that $\mathcal{G}_k^{\mathsf{R}}(a)$ has bounded parameters independent of k, albeit with an extra affine term (see Definition 7) compared to functions in $\mathcal{G}_k^{\mathsf{S}}(a)$. On the other hand, achieving $O(k^{-1/2})$ approximation error using the latter class requires the bounds on the hidden layer weights and biases to scale as $k^{1/2}\log k$.

Remark 9 (Related approximation results) *Several related approximation bounds to Theorem 8 are available in the literature, which can also be leveraged to analyze the approximation error of NEs. In particular, Yukich et al. (1995, Theorem 2.2) provides sup-norm* *error bounds for approximating a target function and its derivatives by a sigmoidal NN with unbounded input weights and biases. A further improvement over (Barron, 1992, Theorem 2) by a $k^{-1/2d}$ factor is reported in (Makovoz, 1998) for NNs with step activation functions, under a different regularity condition on the Fourier transform of target function. A sup-norm approximation result for squared ReLU activation is given in (Klusowski and Barron, 2018, Theorem 3) for functions f with bounded $S_3(f)$ (see (2.9)). Also related are NN approximation bounds derived in (Domingo-Enrich and Mroueh, 2021) for a function with bounded $\mathcal{R},\mathcal{U}$-norm, where the latter is based on $\mathcal{R}$-norm introduced in (Ongie et al., 2020).*

The next proposition shows that a sufficiently smooth function over a compact domain can be approximated to within $O(k^{-1/2})$ error by a shallow NN.

Proposition 10 (Approximation of smooth functions) *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact and $f : \mathcal{X} \to \mathbb{R}$. Suppose that there exists an open set $\mathcal{U} \supset \mathcal{X}$, $b \geq 0$, and $\tilde{f} \in \mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$, $s_{\mathsf{KB}} := \lfloor d/2 \rfloor + 3$, such that $f = \tilde{f}|_{\mathcal{X}}$. Then, there exists $g \in \mathcal{G}_k^{\mathsf{R}}\left(\bar{c}_{b,d,\|\mathcal{X}\|}\right)$, where $\bar{c}_{b,d,\|\mathcal{X}\|}$ is given in* (A.15)*, such that $\|f-g\|_\infty \lesssim c_{b,d,\|\mathcal{X}\|} d^{1/2} k^{-1/2}$. The same holds with s_{KB} and $\mathcal{G}_k^{\mathsf{R}}$ replaced with $s_{\mathsf{B}} := \lfloor d/2 \rfloor + 2$ and $\mathcal{G}_k^{\mathsf{S}}$, respectively.*

The proof of Proposition 10 (see Appendix A.1.2) shows that any sufficiently smooth function on $\mathcal{X}$ can be extended to a function in the Barron or the Klusowksi-Barron class with domain $\mathbb{R}^d$. This is done by nullifying the partial derivatives of order s_{KB} (or s_{B}) outside $\mathcal{X}$ and multiplying by a smooth bump function that equals 1 on $\mathcal{X}$ and smoothly decays outside. Note that for an integer $s \geq 0$ and a real number $\tilde{s} \geq s$, $\mathsf{C}_b^s(\mathcal{U})$ contains the Hölder class with smoothness $\tilde{s}$ and radius b.

3.2 Estimation of Parameterized Divergences

For $\mu, \nu \in \mathcal{P}(\mathcal{X})$, consider the SD $\mathsf{D}_{h,\mathcal{F}}(\mu,\nu)$ defined in (2.1). Let X^n and Y^n be n i.i.d. samples from μ and ν, respectively. Consider a NE for $\mathsf{D}_{h,\mathcal{F}}(\mu,\nu)$ realized by a shallow NN, i.e., $\hat{\mathsf{D}}_{h,\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n,Y^n)$ (see (2.13)). Our next result provides a tail inequality for the error in estimating the parametrized divergence $\mathsf{D}_{h,\mathcal{G}_k^*(\phi)}(\mu,\nu)$ by $\hat{\mathsf{D}}_{h,\mathcal{G}_k^*(\phi)}(X^n,Y^n)$, which will be used to prove consistency of the NE. To state it, given a class of functions $\mathcal{F}$ with domain $\mathcal{X}$, define $\underline{C}(\mathcal{F},\mathcal{X}) := \inf_{x\in\mathcal{X}, f\in\mathcal{F}} f(x)$ and $\bar{C}(\mathcal{F},\mathcal{X}) := \sup_{x\in\mathcal{X}, f\in\mathcal{F}} f(x)$.

Theorem 11 (Empirical estimation error tail bound) *Let $\mu,\nu \in \mathcal{P}(\mathcal{X})$ and consider the NN class $\mathcal{G}_k^*(\phi)$ given in Definition 7. Assume $\mathcal{X}$ and ϕ are such that $\bar{C}\big(|\mathcal{G}_k^*(\phi)|,\mathcal{X}\big) < \infty$, h is differentiable in $\big[\underline{C}\big(\mathcal{G}_k^*(\phi),\mathcal{X}\big), C\big(\mathcal{G}_k^*(\phi),\mathcal{X}\big)\big]$ with derivative h', $\mathsf{D}_{h,\mathcal{G}_k^*(\phi)}(\mu,\nu) < \infty$, and*

$$\bar{C}\left(\left|h' \circ \mathcal{G}_k^*(\phi)\right|, \mathcal{X}\right) < \infty. \tag{3.2}$$

Then there exists a constant $c > 0$ such that for any $\delta \geq 0$, we have

$$\sup_{\substack{\mu,\nu\in\mathcal{P}(\mathcal{X}):\\ \mathsf{D}_{h,\mathcal{G}_k^*(\phi)}(\mu,\nu)<\infty}} \mathbb{P}\Big(\Big|\hat{\mathsf{D}}_{h,\mathcal{G}_k^*(\phi)}(X^n,Y^n) - \mathsf{D}_{h,\mathcal{G}_k^*(\phi)}(\mu,\nu)\Big| \geq \delta + E_{k,h,\phi,\mathcal{X}} n^{-\frac{1}{2}}\Big) \leq c e^{-\frac{n\delta^2}{V_{k,h,\phi,\mathcal{X}}}}, \tag{3.3}$$

with upper bounds for $V_{k,h,\phi,\mathcal{X}}$ and $E_{k,h,\phi,\mathcal{X}}$ available in (A.20) *and* (A.21)*, respectively.*

The proof of Theorem 11 (see Appendix A.1.3) relies on upper bounding the estimation error by a separable sub-Gaussian process and invoking the chaining tail inequality (see Theorem 56 in Appendix A.1.3).

The next theorem provides an upper bound on the expected empirical estimation error. It will be used to obtain effective error bounds for the NE in the forthcoming sections.

Theorem 12 (Empirical estimation error bound) *Let $a > 0$ and $\mathcal{G}_k \in \{\mathcal{G}_k^{\mathsf{R}}(a), \mathcal{G}_k^{\mathsf{S}}(a)\}$. Suppose h is differentiable in $\big[\underline{C}(\mathcal{G}_k, \mathcal{X}), \bar{C}(\mathcal{G}_k, \mathcal{X})\big]$ with derivative h', $\bar{C}\left(|h' \circ \mathcal{G}_k|, \mathcal{X}\right) \vee \bar{C}(|h \circ \mathcal{G}_k|, \mathcal{X}) \vee \bar{C}\big(|\mathcal{G}_k|, \mathcal{X}\big) \lesssim_{a,h,\|\mathcal{X}\|} 1$ for all $k \in \mathbb{N}$. Then, for all $k, n \in \mathbb{N}$,*

$$\sup_{\mu,\nu \in \mathcal{P}(\mathcal{X})} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{h,\mathcal{G}_k}(X^n, Y^n) - \mathsf{D}_{h,\mathcal{G}_k}(\mu,\nu)\right|\right] \lesssim_{h,a,\|\mathcal{X}\|} d^{\frac{3}{2}} n^{-\frac{1}{2}}. \tag{3.4}$$

Theorem 12 follows from a more general result that we establish in Appendix A.1.4 (namely, Theorem 58), where $\mathcal{G}_k$ as above is replaced by an arbitrary VC-type class satisfying certain technical conditions. The proof of the latter relies on standard maximal inequalities from empirical process theory. To prove (3.4), we also require a bound on the entropy integral of the NN class. This is obtained by noting that $\mathcal{G}_k$ is a subset of the symmetric convex hull of the composition of a monotone function with a VC subgraph class, and upper bounding the covering numbers of such convex hulls.

Remark 13 (NN distances) *The SD $\mathsf{D}_{h,\mathcal{G}_k(\mathbf{a},\phi)}(\mu,\nu)$ is the so-called NN distance, studied in (Arora et al., 2017; Zhang et al., 2018a) in the context of GANs. Theorem 11 and 12 can thus be understood, respectively, as a tail bound and as an error bound for NN distance estimation from data, and implies that the estimation error rate is parametric in n.*

For $\mathcal{G}_k \in \{\mathcal{G}_k^{\mathsf{R}}(a), \mathcal{G}_k^{\mathsf{S}}(a)\}$, we have $\bar{C}\left(|\mathcal{G}_k|\right) \leq 3a(\|\mathcal{X}\|+1)$, $\bar{C}\left(|h \circ \mathcal{G}_k|\right) \leq \sup\{|h(z)|, z \in [-3a(\|\mathcal{X}\|+1), 3a(\|\mathcal{X}\|+1)]\} < \infty$ and $\mathsf{D}_{h,\mathcal{G}_k} < \infty$ for all k and $h \in \{h_{\mathsf{KL}}, h_{\chi^2}\}$. Similarly, $\bar{C}\left(|h' \circ \mathcal{G}_k|\right)$ is finite and bounded by a quantity independent of k for these h (see (B.3) and (B.5)). Hence, h_{KL} and h_{χ^2} satisfies the assumptions in Theorem 12, and consequently, (3.4) applies for KL and χ^2 divergences. These bounds also hold for H^2 and TV distances for appropriate NN classes (see Theorems 27 and 32 below). In the next section, we use the above results to analyze the effective error for neural estimation of SDs.

4. Neural Estimation of f-Divergences

We now turn to analyze neural estimation performance of several important f-divergences, encompassing KL, χ^2, H^2, and TV. Throughout this section, we assume for simplicity that $\mathcal{X} = [0,1]^d$, but the results and proof techniques readily extend to arbitrary compact domains. Further, we present results for ReLU NNs, although all statements also hold for sigmoid nets with a slightly modified spectral norm condition defining the class of distributions. We comment about this once in Remark 15 below, but omit further mention to avoid repetition.

4.1 KL Divergence

Let $\hat{\mathsf{D}}_{\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n, Y^n) := \hat{\mathsf{D}}_{h_{\mathsf{KL}},\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n, Y^n)$ be a NE of $\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)$, where $\mathbf{a}_k \in \mathbb{R}^4_{\geq 0}$ for all $k \in \mathbb{N}$. To state performance guarantees for this NE, some definitions are needed. Let

$\mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$ be the set of all pairs $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $\mu \ll \nu$ and $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) < \infty$, and for any $M \geq 0$ define

$$\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X}) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X}) : c^\star_{\mathsf{KB}}(f_{\mathsf{KL}},\mathcal{X}) \vee \mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \leq M\right\}. \tag{4.1}$$

For appropriately chosen $M, b \geq 0$, $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ contains $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$ for which $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \leq M$ and $f_{\mathsf{KL}} = \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \in \mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$ for some $\mathcal{U} \supseteq \mathcal{X}$. To see this, note that a smoothness order of s_{KB} for f_{KL} ensures that $c^\star_{\mathsf{KB}}(f_{\mathsf{KL}},\mathcal{X}) \leq \bar{c}_{b,d,\|\mathcal{X}\|}$ (see Proposition 10). Hence, for any $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$ and $M \geq \bar{c}_{b,d,\|\mathcal{X}\|} \vee \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)$, $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$. In particular, $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$, for sufficiently large M, contains Gaussian densities, truncated and normalized to be supported on $\mathcal{X}$.

Since the class $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ becomes larger as M increases, it is to be expected that a larger NN class would be required for accurate neural estimation of KL divergence between distributions in this class. This means that the range of the NN parameters has to be selected depending on M. However, often it is hard to ascertain such an M for the distributions of interest. To account for this, we do not assume that M is known in advance. Instead, we take a NN class $\mathcal{G}^{\mathsf{R}}_k(m_k)$ for some non-decreasing positive sequence $(m_k)_{k\in\mathbb{N}}$ with $m_k \to \infty$, for obtaining neural estimation error bounds.

The following theorem establishes the consistency of KL divergence NE and uniformly bounds the effective error in terms of the NN and sample sizes.

Theorem 14 (KL divergence neural estimation) *The following hold:*

(i) Let $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$ be such that $f_{\mathsf{KL}} \in \mathsf{C}(\mathcal{X})$. Then, for any $0 < \rho < 1$, $(k_n)_{n\in\mathbb{N}}$ with $k_n \to \infty$, $k_n \leq \frac{1}{4}(1-\rho)\log n$ and $\mathcal{G}_n = \mathcal{G}^*_{k_n}(\phi)$,

$$\hat{\mathsf{D}}_{\mathcal{G}_n}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\mathsf{KL}}(\mu\|\nu), \quad \mathbb{P}-a.s. \tag{4.2}$$

(ii) For any $M \geq 0$, $m_k = \log\log k \vee 1$, $\mathcal{G}_k = \mathcal{G}^{\mathsf{R}}_k(m_k)$,

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\right|\right] \lesssim_M d^{\frac{1}{2}} k^{-\frac{1}{2}} + d^{\frac{3}{2}}(\log k)^7 n^{-\frac{1}{2}}. \tag{4.3}$$

The proof of Theorem 14 is presented in Appendix A.2.1. The consistency result in Part (i) relies on $\mathcal{G}^*_k(\phi)$ being a universal approximator for the class of continuous functions on compact sets as $k \to \infty$ and Theorem 11. For Part (ii), we derive (4.3) by utilizing Theorems 8 and 12 to bound the sum of the approximation and estimation errors. From Theorem 8, the former is $O(k^{-1/2})$ if $c^\star_{\mathsf{KB}}(f_{\mathsf{KL}},\mathcal{X}) \leq M$ and k is such that $M \leq \log\log k \vee 1$. On the other hand, for k violating this condition, the effective error is bounded by $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \leq M$. The growing NN parameters contribute an extra polylog(k) factor to the empirical estimation error bound.

Remark 15 (Effective error bound for sigmoid NN class) *It can be seen from the proof of* (4.3) *that the same bound applies to sigmoid NN class $\mathcal{G}^{\mathsf{S}}_k(m_k)$ when $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ is replaced by $\mathcal{P}^2_{\mathsf{KL},\mathsf{B}}(M,\mathcal{X}) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X}) : c^\star_{\mathsf{B}}(f_{\mathsf{KL}},\mathcal{X}) \vee \mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \leq M\right\}$. Similar remarks apply for all the effective error bounds henceforth, which we omit to avoid repetition.*

Remark 16 (Effective error bound based on M) *If M in the definition of the class $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ is known when picking the NN parameters (i.e., they can depend on M), then with $m_k = M$ and $\mathcal{G}_k = \mathcal{G}^{\mathsf{R}}_k(M)$, we have (see* (A.44) *and the last statement in the proof of Theorem 14 in Appendix A.2.1)*

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right] \lesssim_M \; d^{\frac{1}{2}}k^{-\frac{1}{2}} + d^{\frac{3}{2}}n^{-\frac{1}{2}}, \tag{4.4}$$

which removes the polylog factor in the empirical estimation bound (2nd term in (4.3)*).*

Remark 17 (L^2 neural estimation of a function) *In (Barron, 1994), a reminiscent approximation-estimation error analysis for learning a NN approximation of a bounded range function is presented. 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).*

The error bounds in (4.4) and (4.3) imply that the KL divergence NE achieves the parametric and near parametric error rates, respectively.

Corollary 18 (Minimax optimality) *The KL divergence NE $\hat{\mathsf{D}}_{\mathcal{G}_n}(X^n,Y^n)$ is minimax rate-optimal over $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ and $\mathcal{P}^2_{\mathsf{KL,B}}(M,\mathcal{X})$ with $\mathcal{G}_n = \mathcal{G}^{\mathsf{R}}_n(M)$ and $\mathcal{G}_n = \mathcal{G}^{\mathsf{S}}_n(M)$, respectively, achieving the $O\big(n^{-1/2}\big)$ minimax risk. If M is unknown, then this NE with M replaced by $m_n = \log\log n \vee 1$, is near minimax optimal achieving $\tilde{O}\big(n^{-1/2}\big)$ minimax risk.*

The corollary is proven in Appendix A.2.2, where the upper bound follows directly from Theorem 16, Remark 15 and (4.4) by setting $k = n$. For the lower bound, we present a reduction of the KL divergence estimation problem to differential entropy estimation, and invoke the $\Omega(n^{-1/2})$ lower bound from Goldfeld et al. (2020b) for the latter problem.

Theorem 14 and Corollary 18 impose conditions on f_{KL} to bound the effective neural estimation error (namely, assuming that $c^\star_{\mathsf{KB}}(f_{\mathsf{KL}},\mathcal{X}) \leq M$, for some M). A primitive sufficient condition in terms of the densities p and q of μ and ν, respectively, w.r.t. an arbitrary common dominating measure η is given next.

Proposition 19 (Sufficient condition for Theorem 14) *For $b \geq 0$ and $s_{\mathsf{KB}} = \lfloor d/2\rfloor + 3$, consider the class $\widetilde{\mathcal{P}}^2_{\mathsf{KL}}(b,\mathcal{X})$ of pairs of distributions given by*

$$\widetilde{\mathcal{P}}^2_{\mathsf{KL}}(b,\mathcal{X}) := \left\{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(\mathcal{X}) : \begin{array}{l}\exists\ \tilde{p},\tilde{q}\in \mathsf{C}^{s_{\mathsf{KB}}}_b(\mathcal{U}) \text{ for some open set } \mathcal{U}\supset\mathcal{X} \\ s.t.\ \ \log p = \tilde{p}|_{\mathcal{X}},\ \ \log q = \tilde{q}|_{\mathcal{X}}\end{array}\right\}.$$

Then, (4.3) *and* (4.4) *hold with $M = 2\bar{c}_{b,d,\|\mathcal{X}\|} \vee 2b$, where $\bar{c}_{b,d,\|\mathcal{X}\|}$ is given in* (A.15)*,*[6] *and $\widetilde{\mathcal{P}}^2_{\mathsf{KL}}(b,\mathcal{X})$ in place of $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$.*

Remark 20 (Feasible distributions) *$\widetilde{\mathcal{P}}^2_{\mathsf{KL}}(\cdot,\mathcal{X})$ contains distributions $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$ whose densities (p,q) 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.*

6. Although $\mathcal{X}$ is taken to be $[0,1]^d$, we will retain the dependence of $\mathcal{X}$ in the error bounds, which will be used later for extending the results to the unbounded support case.

Remark 21 (Relation to other works) *Corollary 18 and Proposition 19 together imply that the KL divergence NE achieves the parametric minimax error rate over the class of densities with at least* $s_{\mathsf{KB}} = \lfloor d/2 \rfloor + 3$ *derivatives (or* $s_{\mathsf{B}} = \lfloor d/2 \rfloor + 2$ *derivatives in the case of sigmoid activation). To compare with existing results, it is known that variants of classic kernel-based estimators (Kandasamy et al., 2015; Singh and Póczos, 2014a; Moon et al., 2018; Berrett et al., 2019) achieve the optimal minimax risk of* $O(n^{-1/2})$ *when the densities are Hölder smooth with at least* $d/2$ *or* d *derivatives. We also note that the parametric rate achieved by NE is an improvement over the* $n^{-1/4}$ *rate shown in Sreekumar et al. (2021). Furthermore, we observe that* (4.3), (4.4) *and minimax rate optimality holds for the class of distributions obtained by replacing* $c^\star_{\mathsf{KB}}(f_{\mathsf{KL}}, \mathcal{X}) \leq M$ *in* (4.1) *with* $\|\mathcal{X}\| \, \|f_{\mathsf{KL}}\|_{\mathcal{R},\mathcal{U}} \vee |f_{\mathsf{KL}}(0)| \vee \|\nabla f_{\mathsf{KL}}(0)\|_1 \leq M$*, where* $\mathcal{R},\mathcal{U}$*-norm is defined in (Domingo-Enrich and Mroueh, 2021, Equation 6). This follows by using (Domingo-Enrich and Mroueh, 2021, Theorem 2) in place of Theorem 8 to analyze the approximation error. Similar conclusions hold for NEs of other SDs considered below.*

4.1.1 Neural Estimation via Donsker-Varadhan Formula

Another well known variational representation for KL divergence is the Donsker-Varadhan (DV) formula:

$$\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) = \sup_{f \in \mathcal{F}} \mathbb{E}_\mu[f] - \log \mathbb{E}_\nu\big[e^f\big],$$

where the supremum is over all measurable f such that the last expectation is finite. Parametrizing $\mathcal{F}$ by a NN and replacing expectation with sample means leads to the DV-NE for KL, given by

$$\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}}(X^n, Y^n) := \sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n g(X_i) - \log \frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}.$$

In (Belghazi et al., 2018), the authors studied the special case of DV-NE pertaining to estimation of mutual information, termed MINE. They established consistency along with sample complexity bounds (without accounting for the approximation error). In Appendix C, we show that consistency of the DV-NE holds under similar conditions as in Theorem 14 (see (C.1)). We also prove that the effective error bound given in (4.3) applies to DV-NE, albeit with different constants (see (C.2)). In particular, the latter establishes the minimax optimality of DV-NE with the scaling $k = n$. Instantiating these results for $\mu = P_{AB}$ and $\nu = P_A \otimes P_B$ (i.e., a joint probability law versus the product of its marginals), translates these performance guarantees to MINE, now accounting for finite-size NNs, the associated approximation error, and minimax convergence rates.

4.2 χ^2 Divergence

Let $\hat{\chi}^2_{\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n, Y^n) := \hat{\mathsf{D}}_{h_{\chi^2},\mathcal{G}_k(\mathbf{a}_k,\phi)}(X^n, Y^n)$ denote the NE of $\chi^2(\mu\|\nu)$. Set $\mathcal{P}^2_{\chi^2}(\mathcal{X})$ as the collection of all $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $\mu \ll \nu$ and $\chi^2(\mu\|\nu) < \infty$, and let

$$\mathcal{P}^2_{\chi^2}(M, \mathcal{X}) := \Big\{(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathcal{X}) : c^\star_{\mathsf{KB}}\big(f_{\chi^2}, \mathcal{X}\big) \vee \chi^2(\mu\|\nu) \leq M\Big\}.$$

The next theorem establishes consistency of the NE and bounds its effective absolute-error.

Theorem 22 (χ^2 divergence neural estimation) *The following hold:*

(i) Let $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathcal{X})$ be such that $f_{\chi^2} \in \mathsf{C}(\mathcal{X})$. Then, for any $0<\rho<1$, $(k_n)_{n\in\mathbb{N}}$ with $k_n \to \infty$, $k_n = O\left(n^{(1-\rho)/5}\right)$ and $\mathcal{G}_n = \mathcal{G}^*_{k_n}(\phi)$, we have

$$\hat{\chi}^2_{\mathcal{G}_n}(X^n, Y^n) \xrightarrow[n\to\infty]{} \chi^2(\mu\|\nu), \quad \mathbb{P}-a.s. \tag{4.5}$$

(ii) For any $M \geq 0$, $m_k = \log k$, and $\mathcal{G}_k = \mathcal{G}^{\mathsf{R}}_k(m_k)$, we have

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\chi^2}(M,\mathcal{X})} \mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n,Y^n) - \chi^2(\mu\|\nu)\right|\right] \lesssim_M dk^{-\frac{1}{2}} + d^{\frac{3}{2}}(\log k)^2 n^{-\frac{1}{2}}. \tag{4.6}$$

The proof strategy for Theorem 22 is similar to that of Theorem 14, with appropriate adaptations to account for the difference between f_{χ^2} and f_{KL} (see Appendix A.2.4). Comparing (4.5)-(4.6) to (4.2)-(4.3), we see that consistency for χ^2 divergence estimation holds under milder conditions and that the effective error bound is better in terms of dependence on k than for KL divergence.

Remark 23 (Effective error based on M) *If the NN parameters can depend on M, then setting $m_k = M$ in (4.6) yields*

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\chi^2}(M,\mathcal{X})} \mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n,Y^n) - \chi^2(\mu\|\nu)\right|\right] \lesssim_M dk^{-\frac{1}{2}} + d^{\frac{3}{2}} n^{-\frac{1}{2}}. \tag{4.7}$$

Choosing $k=n$ in (4.7) (resp. (4.6)), we have that the χ^2 NE achieves the parametric (resp. near parametric) error rate over the class $\mathcal{P}^2_{\chi^2}(M,\mathcal{X})$. The proof is similar to that of Corollary 18, and is omitted for brevity. Let $\mathcal{P}^2_{\chi^2,\mathsf{B}}(M,\mathcal{X}) := \big\{(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathcal{X}) : c^\star_{\mathsf{B}}\big(f_{\chi^2},\mathcal{X}\big) \vee \chi^2(\mu\|\nu) \leq M\big\}$.

Corollary 24 (Minimax optimality) *The χ^2 NE $\hat{\chi}^2_{\mathcal{G}_n}(X^n,Y^n)$ is minimax rate-optimal over $\mathcal{P}^2_{\chi^2}(M,\mathcal{X})$ and $\mathcal{P}^2_{\chi^2,\mathsf{B}}(M,\mathcal{X})$ with $\mathcal{G}_n = \mathcal{G}^{\mathsf{R}}_n(M)$ and $\mathcal{G}_n = \mathcal{G}^{\mathsf{S}}_n(M)$, respectively, achieving the $O(n^{-1/2})$ risk. This NE achieves the near parametric $\tilde{O}\big(n^{-1/2}\big)$ minimax risk when M is replaced by $\log n$, which is applicable to the scenario of unknown M.*

Given next is the counterpart of Proposition 19 for χ^2 divergence (proven in Appendix A.2.5), which provides primitive conditions in terms of densities under which the effective error bounds in Theorem 22 and Corollary 24 hold.

Proposition 25 (Sufficient condition for Theorem 22) *For $b \geq 0$ and $s_{\mathsf{KB}} = \lfloor d/2 \rfloor + 3$, let*

$$\widetilde{\mathcal{P}}^2_{\chi^2}(b,\mathcal{X}) := \left\{(\mu,\nu)\in\mathcal{P}^2_{\chi^2}(\mathcal{X}) : \begin{array}{l} \exists\ \tilde{p},\tilde{q} \in \mathsf{C}^{s_{\mathsf{KB}}}_b(\mathcal{U}) \text{ for some open set } \mathcal{U}\supset\mathcal{X} \\ s.t.\ \ p = \tilde{p}|_{\mathcal{X}},\ \ q^{-1} = \tilde{q}|_{\mathcal{X}} \end{array}\right\}.$$

Then, (4.6) and (4.7) hold with $M = \left(\kappa_d d^{3/2}\|\mathcal{X}\| \vee 1\right)\big(2 + 2^{s_{\mathsf{KB}}+1}\bar{c}^2_{b,d,\|\mathcal{X}\|}\big) \vee (b^2+1)$, where κ_d and $\bar{c}_{b,d,\|\mathcal{X}\|}$ are given in (A.3) and (A.15), respectively, and $\widetilde{\mathcal{P}}^2_{\chi^2}(b,\mathcal{X})$ in place of $\mathcal{P}^2_{\chi^2}(M,\mathcal{X})$,

Remark 26 (Feasible distributions) *The class $\widetilde{\mathcal{P}}^2_{\chi^2}(\cdot,\mathcal{X})$ contains $(\mu,\nu)\in\mathcal{P}^2_{\chi^2}(\mathcal{X})$, whose densities p,q, 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 20.*

4.3 Squared Hellinger Distance

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

$$\tilde{\mathcal{G}}_{k,t}\left(\mathbf{a},\phi\right):=\Big\{g:\mathbb{R}^d\to\mathbb{R}:g(x)=(1-t)\wedge\tilde{g}(x),\ \tilde{g}\in\mathcal{G}_k(\mathbf{a},\phi)\Big\}.$$

Set $\mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$ as the collection of all $(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})$ such that $\mu\ll\nu$, and

$$\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X}):=\left\{(\mu,\nu)\in\mathcal{P}_{\mathsf{H}^2}(\mathcal{X}):c^\star_{\mathsf{KB}}(f_{\mathsf{H}^2},\mathcal{X})\vee\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty,\eta}\leq M\right\}. \tag{4.8}$$

Also, let $\tilde{\mathcal{G}}^*_{k,t}(\phi):=\tilde{\mathcal{G}}_{k,t}\big(1,1,1,0,\phi\big)$, and for $a\geq 0$, define

$$\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t}(a):=\tilde{\mathcal{G}}_{k,t}\big(1,2k^{-1}a,a,a,\phi_{\mathsf{R}}\big). \tag{4.9}$$

The next theorem establishes consistency of the NE and bounds its effective absolute-error.

Theorem 27 (Squared Hellinger distance neural estimation) *The following hold:*

(i) If $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$ is such that $f_{\mathsf{H}^2}\in\mathsf{C}(\mathcal{X})$ and there exists $M>0$ such that $\|\mathrm{d}\mu/\mathrm{d}\nu\|_{\infty,\eta}\leq M$, then, for any $0<\rho<1$, $(k_n)_{n\in\mathbb{N}}$ with $k_n\to\infty$, $k_n=O\left(n^{(1-\rho)/3}\right)$ and $\mathcal{G}_n=\tilde{\mathcal{G}}^*_{k_n,M^{-1/2}}(\phi)$, we have

$$\hat{\mathsf{H}}^2_{\mathcal{G}_n}(X^n,Y^n)\xrightarrow[n\to\infty]{}\mathsf{H}^2(\mu,\nu),\quad\mathbb{P}-a.s. \tag{4.10}$$

(ii) For any $M\geq 0$, $m_k=\log k$, $t_k=(\log k)^{-1}$, and $\mathcal{G}_k=\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)$, we have

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})}\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n)-\mathsf{H}^2(\mu,\nu)\right|\right]\lesssim_M d^{\frac{1}{2}}k^{-\frac{1}{2}}\log k+d^{\frac{3}{2}}(\log k)^3n^{-\frac{1}{2}}. \tag{4.11}$$

The proof of Theorem 27 is presented in Appendix A.2.6. To prove consistency and establish effective error bounds for H^2 NE, we use a truncated NN class $\tilde{\mathcal{G}}_{k,t}\left(\mathbf{a},\phi\right)$ that saturates the NN output to $1-t$ for some $t>0$. This is done since $h_{\mathsf{H}^2}(x)$ has a singularity at $x=1$ and the NN outputs must be truncated below 1 so as to satisfy (3.2) for bounding the empirical estimation error. To get the effective error bounds under this constraint, we take $t=t_k$ for some non-increasing positive sequence $t_k\to 0$. The bound in (4.11) uses $t_k=(\log k)^{-1}$.

Remark 28 (Effective error based on M) *If M is known when selecting the NN parameters, then for $\mathcal{G}_k=\tilde{\mathcal{G}}^{\mathsf{R}}_{k,M^{-1/2}}(M)$, we have (see* (A.58)*)*

$$\sup_{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})}\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n)-\mathsf{H}^2(\mu,\nu)\right|\right]\lesssim_M\ d^{\frac{1}{2}}k^{-\frac{1}{2}}+d^{\frac{3}{2}}n^{-\frac{1}{2}}.$$

Addressing minimax optimality, we set $k = n$ in the above equation (resp.(4.11)) to attain the parametric (resp. near parametric) rate for the H^2 NE over the class $\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$. Let $\tilde{\mathcal{G}}^{\mathsf{S}}_{k,t}(a) := \tilde{\mathcal{G}}_{k,t}\big(k^{1/2}\log k, 2k^{-1}a, a, 0, \phi_{\mathsf{S}}\big)$, and $\mathcal{P}^2_{\mathsf{H}^2,\mathsf{B}}(M,\mathcal{X})$ denote the class of distribution pairs with $c^\star_{\mathsf{KB}}(f_{\mathsf{H}^2},\mathcal{X})$ in (4.8) replaced with $c^\star_{\mathsf{B}}(f_{\mathsf{H}^2},\mathcal{X})$.

Corollary 29 (Minimax optimality) *The H^2 NE $\hat{\mathsf{H}}^2_{\mathcal{G}_n}(X^n,Y^n)$ is minimax rate-optimal over the class $\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$ and $\mathcal{P}^2_{\mathsf{H}^2,\mathsf{B}}(M,\mathcal{X})$ with $\mathcal{G}_n = \tilde{\mathcal{G}}^{\mathsf{R}}_{n,M^{-1/2}}(M)$ and $\mathcal{G}_n = \tilde{\mathcal{G}}^{\mathsf{S}}_{n,M^{-1/2}}(M)$, respectively, achieving $O\big(n^{-1/2}\big)$ minimax risk. Further, relevant to the case when M is unknown, the same NE achieves $\tilde{O}\big(n^{-1/2}\big)$ risk, when M is replaced with $\log n$ and $M^{-1/2}$ with $(\log n)^{-1}$.*

Below, we provide a sufficient condition in terms of densities under which the effective error bounds in Theorem 27 as well as Corollary 29 applies, similar in spirit to Proposition 19 (see Appendix A.2.7 for the proof).

Proposition 30 (Sufficient condition for Theorem 27) *For $b \geq 0$ and $s_{\mathsf{KB}} = \lfloor d/2 \rfloor + 3$, consider the class $\widetilde{\mathcal{P}}^2_{\mathsf{H}^2}(b,\mathcal{X})$ of pairs of distributions given by*

$$\widetilde{\mathcal{P}}^2_{\mathsf{H}^2}(b,\mathcal{X}) := \left\{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X}) : \begin{array}{l}\exists\ \tilde{p},\tilde{q}\in\mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U}) \textit{ for some open set } \mathcal{U}\supset\mathcal{X} \\ \textit{s.t.}\quad p^{-\frac{1}{2}} = \tilde{p}|_{\mathcal{X}},\ q^{\frac{1}{2}} = \tilde{q}|_{\mathcal{X}},\ \textit{and}\ \left\|p\vee q^{-1}\right\|_{\infty,\eta}\leq b\end{array}\right\}.$$

Then, (4.11) and Remark 28 hold with $M = \big(\kappa_d d^{\frac{3}{2}}\left\|\mathcal{X}\right\|\vee 1\big)\big(1+2^{s_{\mathsf{KB}}}\bar{c}^2_{b,d,\|\mathcal{X}\|}\big)\vee b^2$, where $\bar{c}_{b,d,\|\mathcal{X}\|}$ and κ_d are given in (A.15) and (A.3), respectively, and $\widetilde{\mathcal{P}}^2_{\mathsf{H}^2}(b,\mathcal{X})$ in place of $\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$.

Remark 31 (Feasible distributions) *$\widetilde{\mathcal{P}}^2_{\mathsf{H}^2}(\cdot,\mathcal{X})$ includes $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$, whose densities p,q, are bounded (from above and away from zero) on $\mathcal{X}$ with an extension that is sufficiently smooth on an open set covering $\mathcal{X}$. This contains the distributions mentioned in Remark 20.*

4.4 Total Variation Distance

Consider the NN class obtained by truncating the functions in $\mathcal{G}_k(\mathbf{a},\phi)$ to $[-1,1]$, i.e.,

$$\bar{\mathcal{G}}_k(\mathbf{a},\phi) := \left\{g:\ g(x) = \mathbb{1}_{\{|\tilde{g}(x)|\leq 1\}}\tilde{g}(x) + \mathbb{1}_{\{\tilde{g}(x)>1\}} - \mathbb{1}_{\{\tilde{g}(x)<-1\}} \text{ for some } \tilde{g}\in\mathcal{G}_k(\mathbf{a},\phi)\right\}. \tag{4.12}$$

Also, let $\hat{\delta}_{\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(X^n,Y^n) := \hat{\mathsf{D}}_{h_{\mathsf{TV}},\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(X^n,Y^n)$, and set $\bar{\mathcal{G}}^{\mathsf{R}}_k(a) := \bar{\mathcal{G}}_k\big(1,2k^{-1}a,a,a,\phi_{\mathsf{R}}\big)$, and $\bar{\mathcal{G}}^*_k(\phi) := \bar{\mathcal{G}}_k\big(1,1,1,0,\phi\big)$. Denote the densities of μ and ν w.r.t. λ by p and q, respectively, and for $M\geq 0$ define

$$\mathcal{P}^2_{\mathsf{TV}}(M,\mathcal{X}) := \Big\{(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X}) : \mu,\nu\ll\lambda,\ \ \|p\vee q\|_{\infty,\mathcal{X}}\leq M\Big\}. \tag{4.13}$$

The following theorem bounds the effective error for TV distance neural estimation.

Theorem 32 (TV distance neural estimation) *The following hold:*

(i) For any $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $0 < \rho < 1$, $(k_n)_{n\in\mathbb{N}}$ *with* $k_n \to \infty$, $k_n = O\big(n^{(1-\rho)/2}\big)$ *and* $\mathcal{G}_n = \bar{\mathcal{G}}^*_{k_n}(\phi)$, *we have*

$$\hat{\delta}_{\mathcal{G}_n}(X^n, Y^n) \xrightarrow[n\to\infty]{} \delta_{\mathsf{TV}}(\mu,\nu)\,, \quad \mathbb{P}-a.s. \tag{4.14}$$

(ii) For any $0 < s \leq 1$, $M \geq 0$, $\tilde{c}_{k,d,s,M,\|\mathcal{X}\|} = O_{d,M,\|\mathcal{X}\|}\big(k^{(d+2)/2(s+d+2)}\big)$ *as defined in* (A.75) *and* $\mathcal{G}_k = \bar{\mathcal{G}}^{\mathsf{R}}_k\big(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\big)$, *we have*

$$\sup_{\substack{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{TV}}(M,\mathcal{X}):\\ f_{\mathsf{TV}}\in\mathsf{Lip}_{s,1,M}(\mathcal{X})}} \mathbb{E}\left[\left|\hat{\delta}_{\mathcal{G}_k}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right] \lesssim_{d,M,s} k^{-\frac{s}{2(s+d+2)}} + k^{\frac{d+2}{2(s+d+2)}} n^{-\frac{1}{2}}. \tag{4.15}$$

The proof of Theorem 32 is provided in Appendix A.2.8. A key technical challenge arises from the fact that $f_{\mathsf{TV}} = \mathbb{1}_{\mathcal{C}^*} - \mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*}$ (see (2.8)) contains step discontinuities in its domain, and hence, it does not belong to the Klusowski-Barron class. Consequently, Theorem 8 is not directly applicable for bounding the approximation error as was done for the SDs considered until now. To overcome this issue, we apply a Gaussian smoothing kernel to f_{TV} so that the smoothed version belongs to the Klusowski-Barron class. The width of the kernel is then adjusted as a function of k such that L^1 norm of the difference between f_{TV} and its smoothed version decreases as k increases. The need for the smoothing operation results in a slower approximation and empirical estimation error rate that depends on d.

Remark 33 (Curse of dimensionality) *Setting* $k = n$ *in* (4.15), *we achieve the effective error rate* $O\big(n^{-s/2(s+d+2)}\big)$. *Note that this rate suffers from CoD, different from NEs of other SDs considered above where the parametric rate is achieved.*

In practice, the condition $f_{\mathsf{TV}} \in \mathsf{Lip}_{s,1,M}(\mathcal{X})$ required for (4.15) may be hard to verify. A simple sufficient condition in terms of the densities of μ and ν is given below. To state it, we need the following definition.

Definition 34 (Critical zero) *Given* $f : \mathcal{X} \to \mathbb{R}$, *a point* $x_0 \in \mathcal{X}$ *is called a critical zero of* f *if* $f(x_0) = 0$ *and every neighbourhood* $\mathcal{U}_{x_0}$ *of* x_0 *contains an* $x \in \mathcal{U}_{x_0} \cap \mathcal{X}$ *such that* $f(x) \neq 0$. *In particular, if* $f(x_0) = 0$ *and* f *is differentiable at* x_0 *with derivative* $f'(x_0) > 0$, *then* x_0 *is a critical zero. Let* $\mathcal{Z}(f)$ *denote the set of critical zeros of* f.

Based on the above, for $N \in \mathbb{N}$ and $b \geq 0$, define

$$\mathcal{T}_{b,N}(\mathcal{X}) := \big\{f : \mathcal{X} \to \mathbb{R} : \|x - x'\| \geq b, \forall\, x, x' \in \mathcal{Z}(f), |\mathcal{Z}(f)| \leq N\big\}, \tag{4.16}$$

as the class of functions on $\mathcal{X}$ with at most N critical zeros at pairwise (Euclidean) distance of at least b from each other. We are now ready to state the sufficient condition for TV distance estimation; see Appendix A.2.9 for proof.

Proposition 35 (Sufficient condition for Theorem 32) *For* $N \in \mathbb{N}$ *and* $b \geq 0$, *consider the class*

$$\widetilde{\mathcal{P}}^2_{\mathsf{TV}}(b,N,\mathcal{X}) := \big\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{TV}}(b,\mathcal{X}) : \exists\ f \in \mathcal{T}_{b,N}(\mathcal{X})\ s.t.\ \ p - q = f\big\}\,.$$

Then, for any $0 < s \leq 1$, (4.15) *holds with* $M = \lambda(\mathcal{X}) + \big(2b^{-s}\lambda(\mathcal{X}) \vee 2N\pi^{d/2}b^{d-s}\Gamma(d/2 + 1)^{-1}\big)$ *and supremum over* $(\mu,\nu) \in \widetilde{\mathcal{P}}^2_{\mathsf{TV}}(b,N,\mathcal{X})$ *in place of that over* $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{TV}}(M,\mathcal{X})$, *where* $\lambda(\mathcal{X})$ *is the Lebesgue measure of* $\mathcal{X}$ *and* Γ *is the gamma function.*

Remark 36 (Feasible distributions) *The set $\widetilde{\mathcal{P}}^2_{\mathsf{TV}}(\cdot,\cdot,\mathcal{X})$ includes generalized Gaussian distributions, Gaussian mixtures, exponential families, Cauchy distributions, etc., truncated and normalized to be supported on $\mathcal{X}$. It also includes distributions whose densities are analytic functions, e.g., non-negative polynomials on $\mathcal{X}$. These inclusions are easy to verify since $p-q$ has finitely many separated critical zeros for such distributions (cf., e.g., (Smale, 1986; Kalantari, 2004) for the case of analytic functions).*

5. Neural Estimation for Distributions with Unbounded Support

Thus far, we considered compactly supported μ and ν. In this section, we consider neural estimation of KL, χ^2, H^2 and TV with $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$. Throughout, unless stated otherwise, we will assume that $\mu,\nu \ll \lambda$ with p,q denoting the respective Lebesgue densities. For each SD, we first prove consistency of the NE under certain regularity conditions on the densities. Then, we present effective error bounds under an Orlicz norm constraint on the densities, which are subsequently specialized to multivariate Gaussian distributions. We next introduce the required definitions below.

Definition 37 (Orlicz space) *An increasing convex function $\psi : [0,\infty) \to [0,\infty)$ with $\psi(0) = 0$ and $\lim_{x\to\infty} \psi(x) = \infty$ is called an Orlicz function. For a given ψ and $M \geq 0$, the bounded Orlicz space[7] is*

$$L_\psi(M) = \left\{ f : \mathbb{R}^d \to \mathbb{R} : \|f\|_\psi \leq M \right\},$$

where $\|f\|_\psi := \inf \left\{ c > 0 : \int_{\mathbb{R}^d} \psi\left(\|x\|/c\right) f(x)\mathrm{d}x \leq 1 \right\}$.

Examples of Orlicz functions include $\hat{\psi}_r(z) = z^r$ and $\psi_r(z) = e^{z^r} - 1$, $z \in \mathbb{R}$, for $r \geq 1$; in particular, ψ_r with $r = 2$ correspond to the sub-Gaussian class defined next.

Definition 38 (Sub-Gaussian distribution) *A distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$ is σ^2-sub-Gaussian for $\sigma > 0$ if $X \sim \mu$ satisfies*

$$\mathbb{E}\left[e^{u\cdot(X-\mathbb{E}[X])} \right] \leq e^{\frac{\sigma^2 \|u\|^2}{2}}, \quad \forall\ u \in \mathbb{R}^d.$$

For $M \geq 0$, let $\mathcal{SG}(M)$ be the set of all σ^2-sub-Gaussian distributions with $\sigma^2 \vee \|\mathbb{E}[X]\| \leq M$.

With some abuse of notation, we henceforth use boldface letters to denote infinite sequences, e.g., $\mathbf{v} = (v_k)_{k\in\mathbb{N}}$; this will simplify some of the subsequent notation. In particular, we use $\mathbf{r} = (r_k)_{k\in\mathbb{N}}$ for an increasing positive divergent sequence (i.e., $r_k \to \infty$) with $r_k \geq 1$, and $\mathbf{m} = (m_k)_{k\in\mathbb{N}}$ for a non-decreasing positive sequence with $m_k \geq 1$. Let $\hat{\mathcal{G}}_k(\mathbf{a},\phi,r) := \left\{ g\mathbb{1}_{B_d(r)} : g \in \mathcal{G}_k(\mathbf{a},\phi) \right\}$, $\hat{\mathcal{G}}^{\mathsf{R}}_k(a,r) := \left\{ g\mathbb{1}_{B_d(r)} : g \in \mathcal{G}^{\mathsf{R}}_k(a) \right\}$, and $\hat{\mathcal{G}}^{*}_k(\phi,r) := \left\{ g\mathbb{1}_{B_d(r)} : g \in \mathcal{G}^{*}_k(\phi) \right\}$ denote the NN classes $\mathcal{G}_k(\mathbf{a},\phi)$, $\mathcal{G}^{\mathsf{R}}_k(a)$, and $\mathcal{G}^{*}_k(\phi)$, respectively, after nullifying the functions outside of $B_d(r)$.

7. It is possible to generalize the results in this section to $\mu,\nu \ll \gamma$, where γ is an arbitrary positive σ-finite Borel measure. Accordingly, the Orlicz norm in Definition 37 is replaced with $\|f\|_{\psi,\gamma} := \inf \left\{ c \in [0,\infty] : \int_{\mathbb{R}^d} \psi\big(\|x\|/c\big) f(x) d\gamma(x) \leq 1 \right\}$. We adopt the current definition for simplicity.

5.1 KL Divergence

For $M \geq 0$, $\ell \in \mathbb{N}$, $\mathbf{r}$ and $\mathbf{m}$ as above, consider the following class of distributions:

$$\bar{\mathcal{P}}^2_{\mathsf{KL},\psi}(M,\ell,\mathbf{r},\mathbf{m}) := \left\{ (\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathbb{R}^d) : \begin{array}{l} \mu,\nu \ll \lambda,\ p,q \in L_\psi(M),\ \|f_{\mathsf{KL}}\|_{\ell,\mu} \leq M, \\ c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k,\ k \in \mathbb{N} \end{array} \right\}.$$

In words, the class above contains pairs of distributions whose (i) densities have a ψ-Orlicz norm bounded by M, (ii) f_{KL} has $L^\ell(\mu)$ norm at most M, and (iii) the restriction of f_{KL} to $B_d(r_k)$ has a Klusowski-Barron coefficient that is at most m_k.

The following is the counterpart of Theorem 14 for distributions supported on $\mathbb{R}^d$; the proof is provided in Appendix A.3.1.

Theorem 39 (KL divergence neural estimation) *For any $0 < \rho < 1$, the following hold:*

(i) Let $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathbb{R}^d)$ be such that $f_{\mathsf{KL}} \in \mathsf{C}(\mathbb{R}^d)$ and $\|f_{\mathsf{KL}}\|_{1,\mu} < \infty$. Then, for k_n, r_n, n satisfying $k_n \to \infty$, $r_n \to \infty$, $k_n^{3/2} r_n e^{k_n(r_n+1)} = O\left(n^{(1-\rho)/2}\right)$ and $\mathcal{G}_n = \hat{\mathcal{G}}^*_{k_n}(\phi, r_n)$,

$$\hat{\mathsf{D}}_{\mathcal{G}_n}(X^n, Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\mathsf{KL}}(\mu\|\nu), \quad \mathbb{P} - a.s.$$

(ii) Let $\ell > 1$, $M \geq 0$, $\ell^* = \ell/(\ell-1)$, and $\mathbf{m}$ be such that $1 \leq m_k \lesssim k^{(1-\rho)/2}$. Then, for $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$, we have

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{KL},\psi}(M,\ell,\mathbf{r},\mathbf{m})} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\right|\right] \\ \lesssim_{d,M,\rho,\psi,\ell} m_k k^{-\frac{1}{2}} + m_k r_k e^{3m_k(r_k+1)}\ n^{-\frac{1}{2}} + \left(\psi\left(r_k M^{-1}\right)\right)^{\frac{-1}{\ell^*}}. \tag{5.1}$$

The proof of the consistency claim in Part (i) follows similar to (4.2) by using the universal approximation property of $\hat{\mathcal{G}}^*_{k_n}(\phi, r_n)$ on Euclidean balls, controlling the residual approximation error via integrability assumption on f_{KL}, and using Theorem 11 to bound the empirical estimation error. The proof of (5.1) is based on the following observations. First, we note that if $c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)}, B_d(r_k)\right)$ can be bounded for every k, then Theorem 8 implies that the NN class $\hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$ with $m_k, r_k \to \infty$ at an appropriate rate can approximate f_{KL} to within an error of $\lesssim d^{1/2} m_k k^{-1/2}$ inside the Euclidean ball $B_d(r_k)$. An upper bound on $c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)}, B_d(r_k)\right)$ is guaranteed, for instance, by Proposition 10 when f_{KL} is sufficiently smooth on $B_d(r_k)$. Moreover, since every Borel probability measure on $\mathbb{R}^d$ is tight, $\mu\left(B^c_d(r_k)\right) \vee \nu\left(B^c_d(r_k)\right) \to 0$ for every $r_k \to \infty$. The proof then follows by an analysis of the approximation error outside $B_d(r_k)$ under the Orlicz norm constraint on the densities of μ and ν, along with an account of the empirical estimation error. The Orlicz norm constraint controls the rate of tail decay of the densities.

Remark 40 (Feasible distributions) *Based on Proposition 10,* (5.1) *holds for distributions $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathbb{R}^d)$, $\mu,\nu \ll \lambda$, such that their densities are sufficiently smooth and bounded (from above and away from zero) on Euclidean balls $B_d(r)$ for any $r > 0$, and $\|f_{\mathsf{KL}}\|_{\ell,\mu}$ is finite for some $\ell > 1$. This includes multivariate Gaussians, Gaussian mixtures, Cauchy distributions, etc., to name a few.*

As an instance of an explicit effective error bound, we now specialize Theorem 39 to the important case of Gaussian distributions. Define the class

$$\mathcal{P}^2_{\mathsf{N}}(M) := \left\{ \big(\mathcal{N}(\mathsf{m}_p, \Sigma_p), \mathcal{N}(\mathsf{m}_q, \Sigma_q)\big) : \begin{array}{l} \|\mathsf{m}_p\|, \|\mathsf{m}_q\| \leq M \\ \|\Sigma_p\|_{\mathsf{op}}, \|\Sigma_p^{-1}\|_{\mathsf{op}}, \|\Sigma_q\|_{\mathsf{op}}, \|\Sigma_q^{-1}\|_{\mathsf{op}} < M \end{array} \right\},$$

of pairs of non-singular multivariate Gaussian distributions with appropriate bounded operator norm (denoted by $\|\cdot\|_{\mathsf{op}}$). The following corollary quantifies the effective error for pairs of Gaussian distributions. However, as the proof (see Appendix A.3.2) requires a tedious evaluation of a bound on the Klusowski-Barron coefficient, we restrict attention to isotropic Gaussians, i.e., whose covariance matrix is $\Sigma = \sigma^2 \mathrm{I}_d$, for some $\sigma > 0$. The (sub)class of isotropic Gaussian measures is denoted by $\bar{\mathcal{P}}^2_{\mathsf{N}}(M)$. Nevertheless, we stress that the argument can be generalized to account for the entire $\mathcal{P}^2_{\mathsf{N}}(M)$ class above.

Corollary 41 (Gaussian effective error) *For any $1 < M < \infty$, there exists $c_{d,M} > 0$ such that for $m_k \asymp_{d,M} (\log k)^{0.5(d+3)}$, $r_k := 1 \vee M + \tilde{r}_k$, $\tilde{r}_k \asymp_{d,M} \sqrt{\log k}$ and $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$,*

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{N}}(M)} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n, Y^n) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\right|\right] \lesssim_{d,M} (\log k)^{\frac{d+4}{2}} \left(k^{-\frac{1}{2}} + k^{c_{d,M}(\log k)^{\frac{d+2}{2}}} n^{-\frac{1}{2}}\right).$$

Remark 42 (Gaussian error rate) *Optimizing over k in the above equation yields an effective error rate of $n^{-(\log n)^{c_{d,M}}} \log n$ for some $c_{d,M} > -1$. Despite the dependence of this rate on d, in Appendix D.1 we show that for certain classes of sub-Gaussian distributions, a NE effective error rate of $n^{-1/3}$ can be achieved independent of dimension. This is to stress that the NE can produce dimension-free convergence rates even when supports are unbounded.*

5.2 χ^2 Divergence

We next consider χ^2 divergence. Consider the following class of distributions:

$$\bar{\mathcal{P}}^2_{\chi^2,\psi}(M, \ell, \mathbf{r}, \mathbf{m}) := \left\{ (\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathbb{R}^d) : \begin{array}{l} \mu,\nu \ll \lambda,\ p,q \in L_\psi(M),\ \left\|f_{\chi^2}\right\|_{\ell,\mu} \leq M, \\ c^\star_{\mathsf{KB}}\left(f_{\chi^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k,\ k \in \mathbb{N} \end{array} \right\}.$$

The following theorem states consistency of the χ^2 NE and bounds the effective error.

Theorem 43 (χ^2 neural estimation) *The following hold:*

(i) Let $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathbb{R}^d)$ satisfy $f_{\chi^2} \in \mathsf{C}(\mathbb{R}^d)$ and $\left\|f_{\chi^2}\right\|_{1,\mu} \vee \left\|h_{\chi^2} \circ f_{\chi^2}\right\|_{1,\nu} < \infty$. Then, for $k_n \to \infty$, $r_n \to \infty$, n satisfying $k_n^{5/2} r_n^2 = O\left(n^{(1-\rho)/2}\right)$ for some $0 < \rho < 1$ and $\mathcal{G}_n = \hat{\mathcal{G}}^*_{k_n}(\phi, r_n)$, we have

$$\hat{\chi}^2_{\mathcal{G}_n}(X^n, Y^n) \xrightarrow[n\to\infty]{} \chi^2(\mu\|\nu), \quad \mathbb{P}-a.s.$$

(ii) For any $M \geq 0$, $\ell > 1$, $\ell^* = \ell/(\ell-1)$ and $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$, we have

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\chi^2,\psi}(M,\ell,\mathbf{r},\mathbf{m})} \mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n, Y^n) - \chi^2(\mu\|\nu)\right|\right] \lesssim_{M,\psi,\ell} m_k^2 d k^{-\frac{1}{2}} + d^{\frac{3}{2}} m_k^2 r_k^2 n^{-\frac{1}{2}} + \left(\psi\left(r_k M^{-1}\right)\right)^{-\frac{1}{\ell^*}}.$$

The proof of Theorem 43 is similar to that of Theorem 39 and is given in Appendix A.3.3.

Remark 44 (Feasible distributions) *Theorem 43 (ii) holds for any distributions $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathbb{R}^d)$, $\mu,\nu \ll \lambda$, such that their densities are sufficiently smooth and bounded (from above for p and away from zero for q) on Euclidean balls, and $\left\|f_{\chi^2}\right\|_{\ell,\mu}$ is finite for some $\ell > 1$. This encompasses the distributions mentioned in Remark 40 for certain parameter ranges.*

The corollary below (see Appendix A.3.4 for proof) provides effective error bounds for the following class of Gaussian distributions:

$$\bar{\mathcal{P}}^2_{\chi^2,\mathsf{N}}(M) := \left\{ \left(\mathcal{N}(\mathsf{m}_p, \sigma_p^2 \mathrm{I}_d), \mathcal{N}(\mathsf{m}_q, \sigma_q^2 \mathrm{I}_d)\right) : \begin{array}{l} 1/M < \sigma_p^2 < 2\sigma_q^2 < M, \\ 2\sigma_q^2 - \sigma_p^2 > 1/M, \ \ \|\mathsf{m}_p\| \vee \|\mathsf{m}_q\| \leq M \end{array} \right\},$$

where the constraint $\sigma_p^2 < 2\sigma_q^2$ is required for $\chi^2\left(\mu\|\nu\right)$ to be finite.

Corollary 45 (Gaussian effective error) *For $1 < M < \infty$, we have with $m_k \asymp_{d,M} k^{2M^5/(4M^5+1)}(\log k)^{0.5(s_{\mathsf{KB}}+d+1)}$, $r_k = 1 \vee M + \tilde{r}_k$, $\tilde{r}_k \asymp_M \sqrt{\log k}$ and $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k\left(m_k, r_k\right)$ that*

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\chi^2,\mathsf{N}}(M)} \mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n,Y^n) - \chi^2\left(\mu\|\nu\right)\right|\right] \lesssim_{d,M} (\log k)^{2(s_{\mathsf{KB}}+d+1)}\left(k^{-\frac{1}{2+8M^5}} + k^{\frac{4M^5}{1+4M^5}} n^{-\frac{1}{2}}\right).$$

Remark 46 (Gaussian error rate) *The optimum in the right hand side (RHS) of the equation above over (k,n) is attained at $k = n^{(1+4M^5)/(1+8M^5)}$, and results in an effective error rate of $n^{-1/(2+16M^5)}\,(\log n)^{2(s_{\mathsf{KB}}+d+1)}$. Note that this rate degrades with increasing d or M. Nevertheless, in Proposition 70 in Appendix D.2, we show that a dimension-free improvement of $n^{-1/2}$ (up to logarithmic factors) can be achieved for a certain class of sub-Gaussian distributions with unbounded support.*

5.3 Squared Hellinger Distance

Next, we consider the squared Hellinger distance. For $M, \mathbf{r}, \mathbf{m}$ as above, let

$$\bar{\mathcal{P}}^2_{\mathsf{H}^2,\psi}(M,\mathbf{r},\mathbf{m}) := \left\{ (\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}(\mathbb{R}^d) : \begin{array}{c} \mu,\nu \ll \lambda, \ p,q \in L_\psi(M), \\ c^\star_{\mathsf{KB}}\left(f_{\mathsf{H}^2}|_{B_d(r_k)}, B_d(r_k)\right) \vee \left\|\dfrac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty, B_d(r_k)} \leq m_k, \ \forall\, k \in \mathbb{N} \end{array} \right\}.$$

Also, consider the following NN class obtained from $\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t}(\cdot)$ (see (4.9)) by nullifying the functions outside of $B_d(r)$:

$$\check{\mathcal{G}}^{\mathsf{R}}_{k,t}(a,r) := \left\{ g\mathbb{1}_{B_d(r)} : g \in \tilde{\mathcal{G}}^{\mathsf{R}}_{k,t}(a)\right\}. \tag{5.2}$$

The next theorem provides conditions under which consistency holds for H^2 neural estimation and bounds the effective error; see Appendix A.3.5 for the proof.

Theorem 47 (Squared Hellinger distance neural estimation) *Let $\mathbf{m}$ satisfy $m_k = o(k^{1/4})$. The following hold:*

(i) For $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\mathsf{H}^2,\psi}(M,\mathbf{r},\mathbf{m})$*, and* $\mathbf{k}$*,* $\mathbf{r}$*,* $\mathbf{m}$*,* n *such that* $k_n \to \infty$*,* $r_{k_n} \to \infty$*,* $m_{k_n} \to \infty$*,* $k_n^{1/2} m_{k_n}^2 r_{k_n} = O\left(n^{(1-\rho)/2}\right)$ *for some* $0<\rho<1$ *and* $\mathcal{G}_n = \check{\mathcal{G}}^{\mathsf{R}}_{k_n, m_{k_n}^{-1/2}}(m_{k_n}, r_{k_n})$*, we have*

$$\hat{\mathsf{H}}^2_{\mathcal{G}_n}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{H}^2(\mu,\nu), \quad \mathbb{P}-a.s.$$

(ii) For any $M \geq 0$ *and* $\mathcal{G}_k = \check{\mathcal{G}}^{\mathsf{R}}_{k, m_k^{-1/2}}(m_k, r_k)$*, we obtain*

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{H}^2,\psi}(M,\mathbf{r},\mathbf{m})} \mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{H}^2(\mu,\nu)\right|\right]$$
$$\lesssim_{M,\psi} m_k^2 d^{\frac{1}{2}} k^{-\frac{1}{2}} + d^{\frac{3}{2}} m_k^2 r_k n^{-\frac{1}{2}} + \Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{2}}.$$

The proof of Theorem 47 follows along similar lines to Theorem 39. Notice that the NN class $\check{\mathcal{G}}^{\mathsf{R}}_{k,t}$ is used to overcome the issue of singularity of f_{H^2} as in Theorem 27.

Remark 48 (Feasible distributions) *Theorem 47 applies for any distributions* $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}(\mathbb{R}^d)$*,* $\mu,\nu \ll \lambda$*, such that their densities* p,q *are sufficiently smooth and bounded (from above and below) on Euclidean balls. To list a few, this includes multivariate Gaussians, mixture Gaussians, Cauchy distributions, etc.*

The next corollary provides effective error bounds for the class of isotropic Gaussian distributions with bounded parameters, $\bar{\mathcal{P}}^2_{\mathsf{N}}(M)$, considered in Section 5.1; see Appendix A.3.6 for the proof.

Corollary 49 (Gaussian effective error) *For* $1 < M < \infty$*,* $r_k = 1 \vee M + (M + 8M^2)^{-1/2}(\log k)^{1/2}$*,* $m_k \asymp_{d,M} k^{2M/(1+8M)}(\log k)^{0.5(s_{\mathsf{KB}}+d+1)}$*, and* $\mathcal{G}_k = \check{\mathcal{G}}^{\mathsf{R}}_{k, m_k^{-1/2}}(m_k, r_k)$*,*

$$\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{N}}(M)} \mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{H}^2(\mu,\nu)\right|\right] \lesssim_{d,M} (\log k)^{s_{\mathsf{KB}}+d+2} k^{-\frac{1}{2+16M}} \left(1 + k^{\frac{1}{2}} n^{-\frac{1}{2}}\right).$$

Remark 50 (Gaussian error rate) *Setting* $k=n$ *in the equation above yields an effective error rate of* $n^{-1/(2+16M)}(\log n)^{s_{\mathsf{KB}}+d+2}$*. While this rate deteriorates with* M *and* d*, in Proposition 72 in Appendix D.3, we show that a rate of* $n^{-1/2}$ *(up to logarithmic factors) is possible independent of dimension for a certain class of sub-Gaussian distributions with unbounded support.*

5.4 TV Distance

Finally, we consider neural estimation of TV distance for distributions with unbounded support. For $M \geq 0, s \geq 0, b \geq 0$, $N \in \mathbb{N}$, sequences $\mathbf{r}$ and $\mathbf{m}$ as above, let

$$\bar{\mathcal{P}}^2_{\mathsf{TV},\psi}(M,s,\mathbf{r},\mathbf{m}) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{TV}}(M,\mathbb{R}^d) : \begin{array}{l} \mu,\nu \ll \lambda, p,q \in L_\psi(M), \\ f_{\mathsf{TV}} \mathbb{1}_{B_d(r_k)} \in \mathsf{Lip}_{s,1,m_k}(B_d(r_k)) \end{array}\right\},$$
$$\hat{\mathcal{P}}^2_{\mathsf{TV}}(b,M,N) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{TV}}(M,\mathbb{R}^d) : \mu,\nu \in \mathcal{SG}(M),\ \exists\ f \in \mathcal{T}_{b,N}\big(\mathbb{R}^d\big) \text{ s.t. } p-q=f\right\},$$

where $\mathcal{P}^2_{\mathsf{TV}}(M,\mathbb{R}^d)$ and $\mathcal{T}_{b,N}(\mathbb{R}^d)$ are defined in (4.13) and (4.16), respectively. Also, define the NN classes $\vec{\mathcal{G}}^{\mathsf{R}}_k(a,r) := \{g\mathbb{1}_{B_d(r)} : g \in \bar{\mathcal{G}}^{\mathsf{R}}_k(a)\}$, and $\vec{\mathcal{G}}^{*}_k(\phi,r) := \{g\mathbb{1}_{B_d(r)} : g \in \bar{\mathcal{G}}^{*}_k(\phi)\}$, where $\bar{\mathcal{G}}_k$ is given in (4.12).

The next theorem is the analogue of Theorem 39 for TV distance neural estimation. Its proof is presented in Appendix A.3.7.

Theorem 51 (TV distance neural estimation) *The following hold:*

(i) For $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$, any $0<\rho<1$, and $\mathbf{k},\mathbf{r},n$ such that $k_n\to\infty$, $r_n\to\infty$, $k_n r_n^{1/2} = O\big(n^{(1-\rho)/2}\big)$ and $\mathcal{G}_n = \vec{\mathcal{G}}^{*}_{k_n}(\phi,r_n)$, we have

$$\hat{\delta}_{\mathcal{G}_n}(X^n,Y^n) \xrightarrow[n\to\infty]{} \delta_{\mathsf{TV}}(\mu,\nu)\,, \quad \mathbb{P}-a.s.$$

(ii) For any $M\geq 0$ and $0<s\leq 1$, $\mathcal{G}_k = \vec{\mathcal{G}}^{\mathsf{R}}_k\big(\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}}, r_k\big)$, where $\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}}$ is given in (A.98)*, we obtain*

$$\begin{aligned}\sup_{(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{TV},\psi}(M,s,\mathbf{r},\mathbf{m})} &\mathbb{E}\left[\left|\hat{\delta}_{\mathcal{G}_k}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right]\\ &\lesssim_{d,M,s,\rho} \left(m_k^{d+2} r_k^{s(d+1)} k^{-\frac{s}{2}}\right)^{\frac{1}{s+d+2}} + n^{-\frac{1}{2}}\left(m_k r_k^{s+1} k^{\frac{1}{2}}\right)^{\frac{d+2}{s+d+2}} + \psi\big((r_k M^{-1})\big)^{-1}.\end{aligned}$$

The following corollary (see Appendix A.3.8 for proof) provides effective error bounds for sub-Gaussian distributions such that $p-q$ has finite number of critical zeros pairwise separated by Euclidean distance bounded away from zero.

Corollary 52 (Sub-Gaussian effective error) *For any $0<s\leq 1$, $b\geq 0$, $M\geq 0$, $N\in\mathbb{N}$, $r_k = M\vee 1 + 4\sqrt{dM\log k}$, $m_k = c_{d,s,b,N,r_k}$ (see* (A.101)*) and $\mathcal{G}_k = \vec{\mathcal{G}}^{\mathsf{R}}_k\big(\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}}, r_k\big)$, we have*

$$\begin{aligned}\sup_{(\mu,\nu)\in\hat{\mathcal{P}}^2_{\mathsf{TV}}(b,M,N)} &\mathbb{E}\left[\left|\hat{\delta}_{\mathcal{G}_k}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right]\\ &\lesssim_{d,s,b,N} (\log k)^{\frac{(s+d)(d+2)}{2(s+d+2)}} k^{\frac{-s}{2(s+d+2)}} + (\log k)^{\frac{d+2}{2}} k^{\frac{d+2}{2(s+d+2)}} n^{-\frac{1}{2}}.\end{aligned}$$

Remark 53 (Sub-Gaussian error rate) *Setting $k=n$ in the bound above, the effective error rate is $n^{-s/2(s+d+2)}(\log n)^{(d+2)/2}$.*

Remark 54 (Feasible distributions) *$\hat{\mathcal{P}}^2_{\mathsf{TV}}(\cdot,\cdot,\cdot)$ includes generalized Gaussian distributions, mixture Gaussians, and in general, distributions pairs with smooth bounded densities having finite number of modes and sub-Gaussian tails.*

6. Concluding Remarks

This paper studied neural estimation of SDs, aiming to characterize the performance of NEs via an approximation-estimation error analysis. We showed that NEs of f-divergences, such as the KL and χ^2 divergences, squared Hellinger distance, and TV distance are consistent, provided the appropriate scaling of the NN size k with the sample size n. We further derived non-asymptotic absolute-error upper bounds that quantify the dependence on k and n. In the compactly supported case, the derived bounds enabled to establish the minimax optimality of NEs for KL divergence, χ^2 divergence, and H^2 distance. The key results leading to these bounds are Theorems 8 and 12, which, respectively, bound the sup-norm approximation error by NNs and the empirical estimation error of the parametrized SD. Our theory covers distributions whose densities belong to an appropriate Orlicz class (e.g., sub-Gaussian distributions), but faster (optimal) parametric rates are attained when supports are compact.

Going forward, we aim to extend our results to additional SDs such as Wasserstein distances and IPMs. While our analysis strategy extends to these examples, new approximation bounds for the appropriate function classes (e.g., 1-Lipschitz) are needed. Generalizing our results to NEs based on deep nets is another natural direction. Recent results on the approximation capabilities of DNNs (e.g., Yarotsky, 2017) appears useful for this purpose. While our analysis does not account for the optimization error, this is another important component of the overall error and we plan to examine it in the future. Through the results herein and the said future directions, we hope to couple neural estimators with the theory to guarantee their performance and/or elucidate their limitations.

Acknowledgments

The work of S. Sreekumar is supported by the TRIPODS Center for Data Science National Science Foundation Grant CCF-1740822. Z. Goldfeld is supported by the NSF CRII Grant CCF-1947801, the NSF CAREER Award under Grant CCF-2046018, and the 2020 IBM Academic Award. We thank Prof. Kengo Kato for many useful suggestions that helped us improve the manuscript. We also thank Zhengxin Zhang for his involvement in an earlier version of this work.

Appendix A. Proofs

This section contains proofs of the results presented in Section 3-5, each given in a different subsection. For fluidity, derivations of lemmas used in those proofs are relegated to Appendix B.

We first state an auxiliary result which will be useful in several proofs that follow. For $b \geq 0$ and an integer $s \geq 0$, define the function classes:

$$\mathcal{L}^{\mathsf{KB}}_{s,b}(\mathbb{R}^d) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \begin{array}{l} |f(0)| \vee \|\nabla f(0)\|_1 \leq b, \left\|D^{\tilde{\alpha}} f\right\|_1 < \infty, \forall\ \|\tilde{\alpha}\|_1 \leq s \\ \|D^{\alpha} f\|_2 \leq b, \forall\ \|\alpha\|_1 \in \{2, s\} \end{array} \right\}, \tag{A.1}$$

$$\mathcal{L}^{\mathsf{B}}_{s,b}(\mathbb{R}^d) := \left\{ f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \begin{array}{l} |f(0)| \leq b, \left\|D^{\tilde{\alpha}} f\right\|_1 < \infty, \forall\ \|\tilde{\alpha}\|_1 \leq s \\ \|D^{\alpha} f\|_2 \leq b, \forall\ \|\alpha\|_1 \in \{1, s\} \end{array} \right\}. \tag{A.2}$$

The next lemma states that functions in $\mathcal{L}^{\mathsf{KB}}_{s,b}(\mathbb{R}^d)$ (resp. $\mathcal{L}^{\mathsf{B}}_{s,b}(\mathbb{R}^d)$) with sufficient smoothness order s belong to the Klusowski-Barron (resp. Barron) class. Its proof is given in Appendix B.1 and borrows arguments from (Barron, 1993).

Lemma 55 (Smoothness and Klusowksi-Barron class) *Recall* $s_{\mathsf{KB}} = \lfloor 0.5d \rfloor + 3$ *and* $s_{\mathsf{B}} := \lfloor 0.5d \rfloor + 2$. *If* $f \in \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$, *then we have* $S_2(f) \leq bd^{3/2}\kappa_d$, *while if* $f \in \mathcal{L}_{s_{\mathsf{B}},b}(\mathbb{R}^d)$, *then* $S_1(f) \leq bd^{1/2}\kappa_d$, *where*

$$\kappa_d^2 := \left(d + d^{s_{\mathsf{B}}}\right) \int_{\mathbb{R}^d} \left(1 + \|\omega\|^{2(s_{\mathsf{B}}-1)}\right)^{-1} \mathrm{d}\omega < \infty. \tag{A.3}$$

Consequently, for $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{L}_{s_{\mathsf{KB}},b}(\mathbb{R}^d) \subseteq \mathcal{B}_{c,2,\mathcal{X}}(\mathbb{R}^d)$ *and* $\mathcal{L}_{s_{\mathsf{B}},b}(\mathbb{R}^d) \subseteq \mathcal{B}_{c,1,\mathcal{X}}(\mathbb{R}^d)$ *with* $c = b \vee bd^{3/2}\kappa_d \|\mathcal{X}\|$ *and* $c = b \vee bd^{1/2}\kappa_d \|\mathcal{X}\|$, *respectively.*

A.1 Proofs for Section 3

A.1.1 Proof of Theorem 8

For $\mathbf{a} = (a_1, a_2, a_3, a_4)$, we denote the set of feasible parameters of $\mathcal{G}_k(\mathbf{a}, \phi)$ by $\Theta_k(\mathbf{a})$, i.e.,

$$\Theta_k(\mathbf{a}) := \left\{ \left(\{\beta_i, w_i, b_i\}_{i=1}^k, w_0, b_0\right) : \begin{array}{c} w_i \in \mathbb{R}^d,\ b_i, \beta_i \in \mathbb{R}, \max\limits_{1 \leq i \leq k} \|w_i\|_1 \vee |b_i| \leq a_1, \\ \max\limits_{1 \leq i \leq k} |\beta_i| \leq a_2,\ \ |b_0| \leq a_3, \|w_0\|_1 \leq a_4 \end{array} \right\}. \tag{A.4}$$

Also, throughout this section, we write $g_\theta(x)$ to denote $g(x) = \sum_{i=1}^k \beta_i \phi\left(w_i \cdot x + b_i\right) + w_0 \cdot x + b_0$ with $\theta = \left(\{\beta_i, w_i, b_i\}_{i=1}^k, w_0, b_0\right)$, whenever the underlying θ is to be emphasized.

We prove the second claim in Theorem 8. The proof relies on arguments from (Barron, 1992) and (Barron, 1993), along with the uniform central limit theorem (CLT) for uniformly bounded VC-type classes. Fix an arbitrary (small) $\delta > 0$, and let $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ be such that $f = \tilde{f}|_{\mathcal{X}}$ and $\|\mathcal{X}\|\, S_1(\tilde{f}) \vee \tilde{f}(0) \leq a + \delta$. Such an $\tilde{f}$ exists since $c^\star_{\mathsf{B}}(f, \mathcal{X}) \leq a$. Then, since $\mathcal{X}$ is compact, it follows from the proof of (Barron, 1993, Theorem 2) that

$$\tilde{f}_0(x) := \tilde{f}(x) - \tilde{f}(0) = \int_{\omega \in \mathbb{R}^d \setminus \{0\}} \varrho(x, \omega) \gamma(\mathrm{d}\omega),$$

where

$$\varrho(x, \omega) := \frac{L\left(\tilde{f}, \mathcal{X}\right)}{\sup_{x \in \mathcal{X}} |\omega \cdot x|} \left(\cos(\omega \cdot x + \zeta(\omega)) - \cos(\zeta(\omega)) \right),$$
$$\gamma(\mathrm{d}\omega) := \frac{\sup_{x \in \mathcal{X}} |\omega \cdot x| \left|\tilde{F}\right|(\mathrm{d}\omega)}{L\left(\tilde{f}, \mathcal{X}\right)},$$

with $L\left(\tilde{f}, \mathcal{X}\right) := \int_{\mathbb{R}^d} \sup_{x \in \mathcal{X}} |\omega \cdot x| \left|\tilde{F}\right|(\mathrm{d}\omega)$. Here $\left|\tilde{F}\right|(\mathrm{d}\omega)$ and $\zeta(\omega)$ are the magnitude and phase of the complex Borel measure in the Fourier representation of $\tilde{f}$, respectively. Note that γ defined above is a probability measure on $\mathbb{R}^d$.

Let $\tilde{\Theta} := \tilde{\Theta}\left(k, L\left(\tilde{f}, \mathcal{X}\right)\right) := \Theta_1\left(k^{1/2} \log k, 2L\left(\tilde{f}, \mathcal{X}\right), 0, 0\right)$ (see (A.4)). Then, it further follows from the proofs of (Barron, 1993, Lemma 2-Lemma 4,Theorem 3) that there exists a probability measure $\gamma_k \in \tilde{\mathcal{P}}_k := \mathcal{P}\left(\tilde{\Theta}\right)$ (see Barron, 1993, Eqns. (28)-(32)) such that

$$\left\| \tilde{f}_0 - \int_{\tilde{\theta} \in \tilde{\Theta}} g_{\tilde{\theta}}(\cdot)\ \gamma_k\left(\mathrm{d}\tilde{\theta}\right) \right\|_{\infty, \mathcal{X}} \lesssim L\left(\tilde{f}, \mathcal{X}\right) k^{-\frac{1}{2}}, \tag{A.5}$$

where $g_{\tilde{\theta}}(x) = \tilde{\beta}\phi_{\mathsf{S}}\big(\tilde{w}\cdot x + \tilde{b}\big)$ for $\tilde{\theta} = (\tilde{\beta}, \tilde{w}, \tilde{b}, 0, 0)$ and ϕ_{S} is the logistic sigmoid. The previous step needs further elaboration. The claims in (Barron, 1993, Lemma 2- Lemma 4, Theorem 3) are stated for L^2 norm, but it is not hard to see from the proof therein that the same also holds for sup-norm, apart from the following subtlety. In the proof of Lemma 3, it is shown that $\varrho(x,\omega)$, $\omega \in \mathbb{R}^d$, lies in the convex closure of a certain class of step functions, whose discontinuity points are adjusted to coincide with the continuity points of the underlying measure η. While this can be shown to account for universal approximation under the essential supremum w.r.t. η, to obtain a sup-norm bound one additional step is needed. Specifically, by using modified step functions whose value at 0 is 0.5 (instead of 1), using their linear combinations for approximation of the target function in Lemma 3, and subsequently replacing each such step function by sigmoids with coinciding values at zero, it can be seen that $\varrho(x,\omega)$ lies in the point-wise closure of convex hull of the desired sigmoid function class.

Next, for each fixed x, let $\upsilon_x : \tilde{\Theta} \to \mathbb{R}$ be given by $\upsilon_x(\tilde{\theta}) := \tilde{\beta}\phi_{\mathsf{S}}\big(\tilde{w}\cdot x + \tilde{b}\big)$ for $\tilde{\theta} = (\tilde{\beta}, \tilde{w}, \tilde{b}, 0, 0)$, and consider the function class $\tilde{\mathcal{F}}_k := \big\{\upsilon_x,\ x \in \mathbb{R}^d\big\}$. Note that every $\upsilon_x \in \tilde{\mathcal{F}}_k$ is a composition of an affine function in $(\tilde{w}, \tilde{b})$ with the bounded monotonic function $\tilde{\beta}\phi_{\mathsf{S}}(\cdot)$. Hence, (Van Der Vaart and Wellner, 1996, Lemma 2.6.15, Lemma 2.6.18) yields that $\tilde{\mathcal{F}}_k$ is a VC type class with index at most $d+3$ for each $k \in \mathbb{N}$. Hence, it follows from (Van Der Vaart and Wellner, 1996, Theorem 2.6.7) that for every $0 < \epsilon \leq 1$,

$$\sup_{\gamma \in \tilde{\mathcal{P}}_k} N\left(2\epsilon L\big(\tilde{f}, \mathcal{X}\big), \tilde{\mathcal{F}}_k, \|\cdot\|_{2,\gamma}\right) \leq \sup_{\gamma \in \tilde{\mathcal{P}}_\infty} N\left(2\epsilon L\big(\tilde{f}, \mathcal{X}\big), \tilde{\mathcal{F}}_\infty, \|\cdot\|_{2,\gamma}\right) \lesssim (d+3)(16e)^{d+3}\epsilon^{-2(d+2)}.$$

Moreover, by (Van Der Vaart and Wellner, 1996, Theorem 2.8.3), $\tilde{\mathcal{F}}_k$ is a uniform Donsker class (in particular, γ_k-Donsker) for all probability measures $\gamma \in \tilde{\mathcal{P}}_k$. Consequently, the uniform CLT (Dudley, 1999) applied to a VC-type class uniformly bounded by $2L\big(\tilde{f}, \mathcal{X}\big)$ yields that there exists k parameter vectors, $\tilde{\theta}_i := (\tilde{\beta}_i, \tilde{w}_i, \tilde{b}_i, 0, 0) \in \tilde{\Theta},\ 1 \leq i \leq k$, such that (see also Yukich et al., 1995, Theorem 2.1)

$$\left\| \int_{\tilde{\theta}\in\tilde{\Theta}} g_{\tilde{\theta}}(\cdot)\ \gamma_k(\mathrm{d}\tilde{\theta}) - \frac{1}{k}\sum_{i=1}^{k} g_{\tilde{\theta}_i}(\cdot) \right\|_{\infty,\mathbb{R}^d} \lesssim d^{\frac{1}{2}} L\big(\tilde{f}, \mathcal{X}\big) k^{-\frac{1}{2}}. \tag{A.6}$$

The RHS above is independent of γ_k and depends on $\tilde{f}$ and $\mathcal{X}$ only through $L\big(\tilde{f}, \mathcal{X}\big)$.

From (A.5)-(A.6) and triangle inequality, we obtain

$$\left\| \tilde{f}_0 - \frac{1}{k}\sum_{i=1}^{k} g_{\tilde{\theta}_i} \right\|_{\infty,\mathcal{X}} \lesssim d^{\frac{1}{2}} L\big(\tilde{f}, \mathcal{X}\big) k^{-\frac{1}{2}}.$$

Setting $\theta = \big(\big\{(\tilde{\beta}_i/k, \tilde{w}_i, \tilde{b}_i)\big\}_{i=1}^{k}, 0, \tilde{f}(0)\big)$ and $g_\theta(x) = k^{-1}\sum_{i=1}^{k}\tilde{\beta}_i\phi_{\mathsf{S}}(\tilde{w}_i\cdot x + \tilde{b}_i) + \tilde{f}(0)$ and noting that $L\big(\tilde{f}, \mathcal{X}\big) \leq \|\mathcal{X}\|\, S_1(\tilde{f})$ by Cauchy-Schwartz, we have

$$\left\| \tilde{f} - g_\theta \right\|_{\infty,\mathcal{X}} \lesssim d^{\frac{1}{2}}\, \|\mathcal{X}\|\, S_1(\tilde{f}) k^{-\frac{1}{2}} \leq d^{\frac{1}{2}}(a+\delta)k^{-\frac{1}{2}}.$$

Next, note that $\big\|\tilde{f} - g_\theta\big\|_{\infty,\mathcal{X}} = \|f - g_\theta\|_{\infty,\mathcal{X}}$ and $g_\theta \in \mathcal{G}_k^{\mathsf{S}}\big(\, \|\mathcal{X}\|\, S_1(\tilde{f}) \vee \tilde{f}(0)\big) \subseteq \mathcal{G}_k^{\mathsf{S}}\,(a+\delta)$. Since $\delta > 0$ is arbitrary and ϕ_{S} is continuous, we obtain that there exists $g_\theta \in \mathcal{G}_k^{\mathsf{S}}\,(a)$ with

$$\|f - g_\theta\|_{\infty,\mathcal{X}} \lesssim a d^{\frac{1}{2}} k^{-\frac{1}{2}}. \tag{A.7}$$

A.1.2 Proof of Proposition 10

To prove the first claim, consider $\tilde{f} \in \mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$ such that $f = \tilde{f}|_{\mathcal{X}}$. By Theorem 8, it suffices to show that there exists an extension f_{ext} of $\tilde{f}$ from $\mathcal{U}$ to $\mathbb{R}^d$ such that $\|\mathcal{X}\| \, S_2(f_{\mathsf{ext}}) \vee |f_{\mathsf{ext}}(0)| \vee \|\nabla f_{\mathsf{ext}}(0)\|_1 \leq \bar{c}_{b,d,\|\mathcal{X}\|}$. Let $\alpha_{|j}$ denote a multi-index of order j. Consider an extension of $D^{\alpha_{|s_{\mathsf{KB}}}} \tilde{f}$ from $\mathcal{U}$ to $\mathbb{R}^d$, which is zero outside $\mathcal{U}$. Fixing $D^{\alpha_{|s_{\mathsf{KB}}}} \tilde{f}$ on $\mathbb{R}^d$ induces an extension of all lower order derivatives $D^{\alpha_{|j}} f$, $0 \leq j < s_{\mathsf{KB}}$ to $\mathbb{R}^d$, which can be defined recursively as $D^{\alpha_{|1}} D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}(x) = D^{\alpha_{|1}+\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}(x)$, $x \in \mathbb{R}^d$, for all $\alpha_{|1}$, $\alpha_{|s_{\mathsf{KB}}-j}$ and $1 \leq j \leq s_{\mathsf{KB}}$.

Let $\mathcal{U}' := \left\{x' \in \mathbb{R}^d : \exists x \in \mathcal{X},\, \|x' - x\| < 1\right\}$ and first assume the strict inclusion $\mathcal{U} \subsetneq \mathcal{U}'$. In that case, the mean value theorem yields that for any $x, x' \in \mathcal{U}'$ and $1 \leq j \leq s_{\mathsf{KB}}$, we have

$$\left|D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}(x')\right| \leq \left|D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}(x)\right| + \sqrt{d} \max_{\tilde{x} \in \mathcal{U}',\, \alpha_{|1}} \left|D^{\alpha_{|s_{\mathsf{KB}}-j}+\alpha_{|1}} \tilde{f}(\tilde{x})\right| \left\|x - x'\right\|, \tag{A.8}$$

where we also used the fact that $\|x - x'\|_1 \leq \sqrt{d}\, \|x - x'\|$. Further, note that $\left\|D^{\alpha_{|s_{\mathsf{KB}}}} \tilde{f}\right\|_{\infty,\mathcal{U}'} \leq b$ ($D^{\alpha_{|s_{\mathsf{KB}}}} \tilde{f}$ equals zero outside $\mathcal{U}$), and since $\tilde{f} \in \mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$, we have $\left\|D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}(x)\right\|_{\infty,\mathcal{U}} \leq b$. Then, for any $x' \in \mathcal{U}'$, taking $x \in \mathcal{X}$ with $\|x - x'\| \leq 1$ (such an x exists by definition of $\mathcal{U}'$) in (A.8) yields $\left|D^{\alpha_{|s_{\mathsf{KB}}-1}} \tilde{f}(x')\right| \leq b + b\sqrt{d}$. Having this, we recursively apply (A.8) to obtain for $1 \leq j \leq s_{\mathsf{KB}}$ that

$$\left\|D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}\right\|_{\infty,\mathcal{U}'} \leq b \sum_{i=1}^{j} d^{\frac{i-1}{2}} + b d^{\frac{j}{2}} \leq b \frac{1 - d^{\frac{s_{\mathsf{KB}}}{2}}}{1 - \sqrt{d}} + b d^{\frac{s_{\mathsf{KB}}}{2}} =: \tilde{b}. \tag{A.9}$$

If $\mathcal{U}' \subseteq \mathcal{U}$, then $\left\|D^{\alpha_{|s_{\mathsf{KB}}-j}} \tilde{f}\right\|_{\infty,\mathcal{U}'} \leq b$ by definition since $\tilde{f} \in \mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$. Hence, (A.9) holds in both cases as $\tilde{b} \geq b$.

The desired final extension is $f_{\mathsf{ext}} := \tilde{f} \cdot f_{\mathsf{c}}$, where f_{c} is the smooth cut-off function

$$f_{\mathsf{c}}(x) := \mathbb{1}_{\mathcal{X}'} * \Psi_{\frac{1}{2}}(x) := \int_{\mathbb{R}^d} \mathbb{1}_{\mathcal{X}'}(y) \Psi_{\frac{1}{2}}(x - y) \mathrm{d}y, \ x \in \mathbb{R}^d, \tag{A.10}$$

with $\mathcal{X}' := \left\{x' \in \mathbb{R}^d : \exists x \in \mathcal{X},\, \left\|x' - x\right\| \leq 0.5\right\}$ and $\Psi(x) \propto \exp\left(-\frac{1}{0.5 - \|x\|^2}\right) \mathbb{1}_{\{\|x\| < 0.5\}}$ as the canonical mollifier normalized to have unit mass. Since $\Psi \in \mathsf{C}^\infty\left(\mathbb{R}^d\right)$, we have $f_{\mathsf{c}} \in \mathsf{C}^\infty\left(\mathbb{R}^d\right)$. Also, observe that $f_{\mathsf{c}}(x) = 1$ for $x \in \mathcal{X}$, $f_{\mathsf{c}}(x) = 0$ for $x \in \mathbb{R}^d \setminus \mathcal{U}'$ and $f_{\mathsf{c}}(x) \in (0,1)$ for $x \in \mathcal{U}' \setminus \mathcal{X}$. Hence, $f_{\mathsf{ext}}(x) = \tilde{f}(x)$ for $x \in \mathcal{X}$, $f_{\mathsf{ext}}(x) = 0$ for $x \in \mathbb{R}^d \setminus \mathcal{U}'$ and $|f_{\mathsf{ext}}(x)| \leq \left|\tilde{f}(x)\right|$ for $x \in \mathcal{U}' \setminus \mathcal{X}$, thus satisfying $f_{\mathsf{ext}}|_{\mathcal{X}} = \tilde{f}|_{\mathcal{X}} = f$ as required. Moreover, for all $0 \leq j \leq s_{\mathsf{KB}}$, we have $D^{\alpha_{|j}} f_{\mathsf{ext}}(x) = 0$, for $x \notin \mathcal{U}'$, and

$$\left\|D^{\alpha_{|j}} f_{\mathsf{ext}}\right\|_{\infty,\mathcal{U}'} \leq 2^j \tilde{b} \max_{\alpha : \|\alpha\|_1 \leq j} \left\|D^{\alpha} f_{\mathsf{c}}\right\|_{\infty,\mathcal{U}'} \leq 2^{s_{\mathsf{KB}}} \tilde{b} \max_{\alpha : \|\alpha\|_1 \leq s_{\mathsf{KB}}} \left\|D^{\alpha} \Psi\right\|_{\infty,\, B_d(0.5)} =: \hat{b}, \tag{A.11}$$

where the first inequality follows using product rule for differentiation and (A.9), while the second is due to (A.10).

Consequently, for $0 \leq j \leq s_{\mathsf{KB}}$ and $i = 1, 2$, we have

$$\|D^{\alpha_{|j}} f_{\mathsf{ext}}\|_i^i = \int_{\mathcal{U}'} (D^{\alpha_{|j}} f_{\mathsf{ext}})^i(x) \mathrm{d}x \leq \hat{b}^i \ \lambda\big(B_d(\mathsf{rad}(\mathcal{X}) + 1)\big) = \hat{b}^i \frac{\pi^{\frac{d}{2}}}{\Gamma(0.5d + 1)} \big(\mathsf{rad}(\mathcal{X}) + 1\big)^d, \tag{A.12}$$

where λ denotes the Lebesgue measure, $\mathsf{rad}(\mathcal{X}) = 0.5 \sup_{x,x' \in \mathcal{X}} \|x - x'\|$, and Γ denotes the gamma function. Defining $b' := \hat{b} d \pi^{d/2} \Gamma(d/2+1)^{-1} \big(\mathsf{rad}(\mathcal{X})+1\big)^d$ and noting that $b' \geq \hat{b}$, we have from (A.11)-(A.12) that $f_{\mathsf{ext}} \in \tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big)$, where

$$\tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big) := \left\{ f \in L^1\big(\mathbb{R}^d\big) \cap L^2\big(\mathbb{R}^d\big) : \begin{array}{l} |f(0)| \leq b', \|D^\alpha f\|_2 \leq b' \text{ for } 1 \leq \|\alpha\|_1 \leq s_{\mathsf{KB}} \\ \|\nabla f(0)\|_1 \leq b', \big\|D^{\tilde{\alpha}} f\big\|_1 < \infty \text{ for } \|\tilde{\alpha}\|_1 \leq s_{\mathsf{KB}} \end{array} \right\}. \tag{A.13}$$

Since $\tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big) \subseteq \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big)$ (see (A.1)), Lemma 55 yields $S_2(f_{\mathsf{ext}}) \leq \kappa_d d^{3/2} b'$ and

$$f_{\mathsf{ext}} \in \mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\big(\mathbb{R}^d\big) \cap \tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big) \subseteq \mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\big(\mathbb{R}^d\big) \cap \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big), \tag{A.14}$$

where

$$\begin{aligned} \bar{c}_{b,d,\|\mathcal{X}\|} &:= \big(\kappa_d d^{\frac{3}{2}} \|\mathcal{X}\| \vee 1\big) \\ &\quad \times \underbrace{\pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2}+1\right)^{-1} \big(\mathsf{rad}(\mathcal{X})+1\big)^d 2^{s_{\mathsf{KB}}} b d \left(\frac{1 - d^{\frac{s_{\mathsf{KB}}}{2}}}{1-\sqrt{d}} + d^{\frac{s_{\mathsf{KB}}}{2}} \right) \max_{\|\alpha\|_1 \leq s_{\mathsf{KB}}} \big\|D^\alpha \Psi\big\|_{\infty, B_d(0.5)}}_{=:b'}, \end{aligned} \tag{A.15}$$

and $\kappa_d^2 := \big(d + d^{s_{\mathsf{B}}}\big) \int_{\mathbb{R}^d} \Big(1 + \|\omega\|^{2(s_{\mathsf{B}}-1)}\Big)^{-1} \mathrm{d}\omega$. It then follows from Theorem 8 that there exists $g \in \mathcal{G}_k^{\mathsf{R}}\big(\bar{c}_{b,d,\|\mathcal{X}\|}\big)$ such that $\|f-g\|_{\infty,\mathcal{X}} \lesssim \bar{c}_{b,d,\|\mathcal{X}\|} d^{1/2} k^{-1/2}$. This proves the first claim of the proposition. Repeating the same arguments starting with $\tilde{f} \in \mathsf{C}_b^{s_{\mathsf{B}}}(\mathcal{U})$, the second claim follows again from Theorem 8, thus completing the proof.

A.1.3 Proof of Theorem 11

We require the following theorem which gives a tail probability bound for the deviation of supremum of a sub-Gaussian process from its associated entropy integral.

Theorem 56 *(Van Handel, 2016, Theorem 5.29) Let $(X_\theta)_{\theta \in \Theta}$ be a separable sub-Gaussian process on the metric space (Θ, d). Then, there exists $c > 0$ such that for any $\theta_0 \in \Theta$ and $\delta \geq 0$, we have*

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

where $\mathsf{diam}(\Theta, \mathsf{d}) := \sup_{\theta, \tilde{\theta} \in \Theta} \mathsf{d}(\theta, \tilde{\theta})$.

We will also use the following lemma which bounds the covering number of $\mathcal{G}_k(\mathbf{a}_k, \phi)$ w.r.t. to metric induced by $\|\cdot\|_{\infty,\mathcal{X}}$.

Lemma 57 *Let ϕ be a continuous monotone activation whose Lipschitz constant is bounded by L, and $U_{a,\mathcal{X}}(\phi) := \phi\big(a(\|\mathcal{X}\|+1)\big) \vee \phi\big(-a(\|\mathcal{X}\|+1)\big)$. Then*

$$N\big(\epsilon, \mathcal{G}_k(\mathbf{a}_k, \phi), \|\cdot\|_{\infty,\mathcal{X}}\big) \leq \big(1 + 10 k a_{2,k} U_{a_{1,k},\mathcal{X}}(\phi) \epsilon^{-1}\big)^k \big(1 + 10 a_{4,k} \|\mathcal{X}\| \epsilon^{-1}\big)^d \big(1 + 10 a_{3,k} \epsilon^{-1}\big)$$

$$\times \left(1+10Lka_{1,k}a_{2,k}\left\|\mathcal{X}\right\|\epsilon^{-1}\right)^{dk}\left(1+10Lka_{1,k}a_{2,k}\left\|\mathcal{X}\right\|\epsilon^{-1}\right)^{k}.$$

In particular, for $\phi \in \{\phi_{\mathsf{R}}, \phi_{\mathsf{S}}\}$*, we have*

$$N\big(\epsilon, \mathcal{G}_k^{\mathsf{R}}(a), \|\cdot\|_{\infty,\mathcal{X}}\big) \leq \left(1+20a(\|\mathcal{X}\|+1)\epsilon^{-1}\right)^{(d+2)k+d+1}, \tag{A.16}$$

$$N\big(\epsilon, \mathcal{G}_k^{\mathsf{S}}(a), \|\cdot\|_{\infty,\mathcal{X}}\big) \leq \left(1+20a(\|\mathcal{X}\|+1)k^{\frac{1}{2}}(\log k+1)\epsilon^{-1}\right)^{(d+2)k+1}, \tag{A.17}$$

$$N\big(\epsilon, \mathcal{G}_k^{*}(\phi), \|\cdot\|_{\infty,\mathcal{X}}\big) \leq \left(1+10k(\|\mathcal{X}\|+1)\epsilon^{-1}\right)^{(d+2)k+1}. \tag{A.18}$$

The proof of Lemma 57 (see Appendix B.2) is based on the fact that the covering number of $B_d^m(r)$ w.r.t. $\|\cdot\|_m$ norm, $m \geq 1$, satisfies

$$N\big(\epsilon, B_d^m(r), \|\cdot\|_m\big) \leq \left(2r\epsilon^{-1}+1\right)^d. \tag{A.19}$$

Continuing with the proof of Theorem 11, we will show that the claim holds with

$$V_{k,h,\phi,\mathcal{X}} \lesssim \bar{C}\big(|\mathcal{G}_k^*(\phi)|, \mathcal{X}\big)^2 \left(\bar{C}\left(\left|h'\circ\mathcal{G}_k^*(\phi)\right|, \mathcal{X}\right)+1\right)^2, \tag{A.20}$$

$$E_{k,h,\phi,\mathcal{X}} \lesssim k\sqrt{d(\|\mathcal{X}\|+1)}\big(\bar{C}\big(\left|h'\circ\mathcal{G}_k^*(\phi)\right|, \mathcal{X}\big)+1\big)\sqrt{\bar{C}\big(|\mathcal{G}_k^*(\phi)|, \mathcal{X}\big)}, \tag{A.21}$$

where we recall that $\bar{C}\left(|\mathcal{F}|, \mathcal{X}\right) := \sup_{x\in\mathcal{X}, f\in\mathcal{F}}|f(x)|$. In the following, we will suppress the dependence of ϕ, h, and $\mathcal{X}$ for simplicity (unless explicitly needed), e.g., $\mathcal{G}_k(\mathbf{a}_k)$ instead of $\mathcal{G}_k(\mathbf{a}_k, \phi)$.

Fix $\mu, \nu \in \mathcal{P}(\mathcal{X})$ such that $\mathsf{D}_{h,\mathcal{G}_k(\mathbf{a}_k)}(\mu,\nu) < \infty$. We have

$$\begin{aligned}
&\hat{\mathsf{D}}_{h,\mathcal{G}_k(\mathbf{a}_k)}(x^n,y^n) - \mathsf{D}_{h,\mathcal{G}_k(\mathbf{a}_k)}(\mu,\nu)\\
&= \sup_{g_\theta\in\mathcal{G}_k(\mathbf{a}_k)} \frac{1}{n}\sum_{i=1}^n g_\theta(x_i) - \frac{1}{n}\sum_{i=1}^n h\circ g_\theta(y_i) - \left(\sup_{g_\theta\in\mathcal{G}_k(\mathbf{a}_k)} \mathbb{E}_\mu\big[g_\theta\big] - \mathbb{E}_\nu\big[h\circ g_\theta\big]\right)\\
&\leq \sup_{g_\theta\in\mathcal{G}_k(\mathbf{a}_k)} \frac{1}{n}\sum_{i=1}^n g_\theta(x_i) - \frac{1}{n}\sum_{i=1}^n h\circ g_\theta(y_i) - \mathbb{E}_\mu\big[g_\theta\big] + \mathbb{E}_\nu\big[h\circ g_\theta\big].
\end{aligned} \tag{A.22}$$

Consider the stochastic process $(Z_{g_\theta})_{g_\theta\in\mathcal{G}_k(\mathbf{a}_k)}$ defined by

$$Z_{g_\theta} := \frac{1}{n}\sum_{i=1}^n g_\theta(X_i) - \frac{1}{n}\sum_{i=1}^n h\circ g_\theta(Y_i) - \mathbb{E}_\mu\big[g_\theta\big] + \mathbb{E}_\nu\big[h\circ g_\theta\big]. \tag{A.23}$$

To apply Theorem 56, we now show that $(Z_{g_\theta})_{g_\theta\in\mathcal{G}_k(\mathbf{a}_k)}$ is a separable sub-Gaussian process on $\big(\mathcal{G}_k(\mathbf{a}_k), \mathsf{d}_{k,\mathbf{a}_k,n}\big)$, where $\mathsf{d}_{k,\mathbf{a}_k,n}$ will be defined below. Note that $\mathbb{E}\left[Z_{g_\theta}\right] = 0$ for all $g_\theta \in \mathcal{G}_k(\mathbf{a}_k)$, and

$$\begin{aligned}
\left|Z_{g_\theta} - Z_{g_{\tilde{\theta}}}\right| \leq \sum_{i=1}^n \frac{1}{n}&\left|g_\theta(X_i) - g_{\tilde{\theta}}(X_i) - \mathbb{E}_\mu\big[g_\theta - g_{\tilde{\theta}}\big]\right|\\
&+ \frac{1}{n}\left|h\circ g_\theta(Y_i) - h\circ g_{\tilde{\theta}}(Y_i) - \mathbb{E}_\nu\left[h\circ g_\theta - h\circ g_{\tilde{\theta}}\right]\right|.
\end{aligned} \tag{A.24}$$

By an application of the mean value theorem, we have for all $g_\theta, g_{\tilde{\theta}} \in \mathcal{G}_k(\mathbf{a}_k)$,

$$\left|h \circ g_\theta(x) - h \circ g_{\tilde{\theta}}(\tilde{x})\right| \leq \bar{C}\left(\left|h' \circ \mathcal{G}_k(\mathbf{a}_k)\right|\right)\left|g_\theta(x) - g_{\tilde{\theta}}(\tilde{x})\right|. \tag{A.25}$$

Hence, we have that almost surely

$$\begin{aligned}
&\frac{1}{n}\left|g_\theta(X_i) - g_{\tilde{\theta}}(X_i) - \mathbb{E}_\mu\left[g_\theta - g_{\tilde{\theta}}\right]\right| + \frac{1}{n}\left|h \circ g_\theta(Y_i) - h \circ g_{\tilde{\theta}}(Y_i) - \mathbb{E}_\nu\left[h \circ g_\theta - h \circ g_{\tilde{\theta}}\right]\right| \\
&\leq \frac{1}{n}\Big[\left|g_\theta(X_i) - g_{\tilde{\theta}}(X_i)\right| + \left|\mathbb{E}_\mu\left[g_\theta - g_{\tilde{\theta}}\right]\right| + \left|h \circ g_\theta(Y_i) - h \circ g_{\tilde{\theta}}(Y_i)\right| + \left|\mathbb{E}_\nu\left[h \circ g_\theta - h \circ g_{\tilde{\theta}}\right]\right|\Big] \\
&\leq 2n^{-1}\big(\bar{C}\left(\left|h' \circ \mathcal{G}_k(\mathbf{a}_k)\right|\right) + 1\big)\left\|g_\theta - g_{\tilde{\theta}}\right\|_{\infty,\mathcal{X}}.
\end{aligned} \tag{A.26}$$

Let $\mathsf{d}_{k,\mathbf{a}_k,n}\big(g_\theta, g_{\tilde{\theta}}\big) := R_{k,\mathbf{a}_k}\|g_\theta - g_{\tilde{\theta}}\|_{\infty,\mathcal{X}} n^{-\frac{1}{2}}$, where $R_{k,\mathbf{a}_k} := 2\left(\bar{C}\left(|h' \circ \mathcal{G}_k(\mathbf{a}_k)|\right) + 1\right)$. Then, it follows from (A.24) and (A.26) via Hoeffding's lemma that

$$\mathbb{E}\left[e^{t\left(Z_{g_\theta} - Z_{g_{\tilde{\theta}}}\right)}\right] \leq e^{\frac{1}{2}t^2 \mathsf{d}_{k,\mathbf{a}_k,n}\left(g_\theta, g_{\tilde{\theta}}\right)^2}.$$

Thus, $(Z_{g_\theta})_{g_\theta \in \mathcal{G}_k(\mathbf{a}_k)}$ is a separable sub-Gaussian process on the metric space $\big(\mathcal{G}_k(\mathbf{a}_k), \mathsf{d}_{k,\mathbf{a}_k,n}\big)$, where the separability follows from (A.26) by the denseness of the countable subset of $\mathcal{G}_k(\mathbf{a}_k)$ obtained by quantizing each of the finite number of bounded NN parameters to rational numbers (recall that a finite union of countable sets is countable and the activation ϕ is assumed continuous).

Specializing to the NN class $\mathcal{G}_k^*(\phi) := \mathcal{G}_k(\mathbf{a}^*, \phi)$, we next bound its covering number w.r.t. $\mathsf{d}_{k,\mathbf{a}^*,n}$, where $\mathbf{a}^* = (1, 1, 1, 0)$. We have

$$\begin{aligned}
N\big(\epsilon, \mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big) &:= N\big(\epsilon, \mathcal{G}_k^*, R_{k,\mathbf{a}^*} n^{-\frac{1}{2}}\|\cdot\|_{\infty,\mathcal{X}}\big) \\
&= N\big(\epsilon/(R_{k,\mathbf{a}^*} n^{-\frac{1}{2}}), \mathcal{G}_k^*, \|\cdot\|_{\infty,\mathcal{X}}\big) \\
&\leq \big(1 + 10k(\|\mathcal{X}\| + 1)R_{k,\mathbf{a}^*} n^{-\frac{1}{2}}\epsilon^{-1}\big)^{(d+2)k+1},
\end{aligned}$$

where the last inequality uses (A.18). Also, we have that $N\big(\epsilon, \mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big) = 1$ for $\epsilon \geq \mathsf{diam}\big(\mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big) := \max_{g_\theta, g_{\tilde{\theta}} \in \mathcal{G}_k^*} \mathsf{d}_{k,\mathbf{a}^*,n}\big(g_\theta, g_{\tilde{\theta}}\big)$. Then,

$$\begin{aligned}
&\int_0^\infty \sqrt{\log N\big(\epsilon, \mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big)}\,\mathrm{d}\epsilon \\
&\qquad = \int_0^{\mathsf{diam}\left(\mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\right)} \sqrt{\log N\big(\epsilon, \mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big)}\,\mathrm{d}\epsilon \\
&\qquad \lesssim \sqrt{kd}\int_0^{\mathsf{diam}\left(\mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\right)} \sqrt{\log\left(1 + 10k(\|\mathcal{X}\| + 1)R_{k,\mathbf{a}^*} n^{-\frac{1}{2}}\epsilon^{-1}\right)}\,\mathrm{d}\epsilon \\
&\qquad \lesssim k\sqrt{d(\|\mathcal{X}\| + 1)}R_{k,\mathbf{a}^*}\sqrt{\bar{C}\big(|\mathcal{G}_k^*|\big)}n^{-\frac{1}{2}},
\end{aligned}$$

where the last step uses $\log(1+x) \leq x,\ x \geq -1$, and $\mathsf{diam}\big(\mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\big) \leq 2R_{k,\mathbf{a}^*}\bar{C}\big(|\mathcal{G}_k^*|\big)n^{-1/2}$.

It follows from Theorem 56 with $Z_0 = 0$ and the definitions of V_k and E_k (see (A.20) and (A.21)) that there exists a constant $c > 0$ such that

$$\mathbb{P}\left(\sup_{g_\theta \in \mathcal{G}_k^*} Z_{g_\theta} \geq cE_k n^{-\frac{1}{2}} + \delta\right) \leq c\, e^{-\frac{\delta^2}{c\,\mathsf{diam}\left(\mathcal{G}_k^*, \mathsf{d}_{k,\mathbf{a}^*,n}\right)^2}} = c\, e^{-\frac{n\delta^2}{V_k}}, \quad \forall\, \delta \geq 0.$$

Noting that this also holds with $-Z_{g_\theta}$ in place of Z_{g_θ}, the union bound gives

$$\mathbb{P}\left(\sup_{g_\theta \in \mathcal{G}_k^*} |Z_{g_\theta}| \geq \delta + cE_k n^{-\frac{1}{2}}\right) \leq 2ce^{-\frac{n\delta^2}{V_k}}.$$

From (A.22)-(A.23) and the above equation, we obtain that for $\delta \geq 0$

$$\mathbb{P}\left(\left|\mathsf{D}_{h,\mathcal{G}_k^*}(\mu,\nu) - \hat{\mathsf{D}}_{h,\mathcal{G}_k^*}(X^n, Y^n)\right| \geq \delta + cE_k n^{-\frac{1}{2}}\right) \leq \mathbb{P}\left(\sup_{g_\theta \in \mathcal{G}_k^*} |Z_{g_\theta}| \geq \delta + cE_k n^{-\frac{1}{2}}\right) \leq 2ce^{-\frac{n\delta^2}{V_k}}.$$

Taking supremum over $\mu, \nu \in \mathcal{P}(\mathcal{X})$ such that $\mathsf{D}_{h,\mathcal{G}_k^*}(\mu,\nu) < \infty$ yields (3.3).

By following similar steps with (A.16) and (A.17) in place of (A.18), we have for $\mathcal{G}_k \in \left\{\mathcal{G}_k^{\mathsf{R}}(a), \mathcal{G}_k^{\mathsf{S}}(a)\right\}$ that

$$\mathbb{P}\left(\left|\mathsf{D}_{h,\mathcal{G}_k}(\mu,\nu) - \hat{\mathsf{D}}_{h,\mathcal{G}_k}(X^n, Y^n)\right| \geq \delta + c\bar{E}_{k,a,h,\mathcal{X}} n^{-\frac{1}{2}}\right) \leq 2\, c\, e^{-n\delta^2/\bar{V}_{k,a,h,\mathcal{X}}}, \tag{A.27}$$

where $\bar{V}_{k,a,h,\mathcal{X}} \lesssim \bar{C}\left(|\mathcal{G}_k^{\mathsf{R}}(a)|, \mathcal{X}\right)^2 \left(\bar{C}\left(\left|h' \circ \mathcal{G}_k^{\mathsf{R}}(a)\right|, \mathcal{X}\right) + 1\right)^2$ and $\bar{E}_{k,a,h,\mathcal{X}} \lesssim \sqrt{dk \log k}\, a(\|\mathcal{X}\| + 1)\left(\bar{C}\left(\left|h' \circ \mathcal{G}_k^{\mathsf{R}}(a)\right|, \mathcal{X}\right) + 1\right)$. To establish (A.27), we use

$$\int_0^\delta \sqrt{\log\left(1 + A\epsilon^{-1}\right)} \lesssim \delta \sqrt{\log\left((A+\delta)/\delta\right)}, \tag{A.28}$$

for $A \geq e$ and $0 \leq \delta \leq 1$, which can be shown via integration by parts.

A.1.4 Proof of Theorem 12

We establish a more general upper bound with $\mathcal{G}_k$ replaced by an arbitrary VC-type class $\mathcal{F}_k$ satisfying certain assumptions. This result is also applicable to deep NNs with finite width in each layer, continuous activation and bounded parameters, and hence, may be of independent interest.

Theorem 58 (Estimation error bound) *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $X^n \sim \mu^{\otimes n}$ and $Y^n \sim \nu^{\otimes n}$. Suppose $h : \mathbb{R} \to \bar{\mathbb{R}}$ and $(\mathcal{F}_k)_{k \in \mathbb{N}}$ (with domain $\mathcal{X}$) satisfy the following conditions for each $k \in \mathbb{N}$:*

(i) h is differentiable at every point in $\left[\underline{C}(\mathcal{F}_k, \mathcal{X}), \bar{C}\left(\mathcal{F}_k, \mathcal{X}\right)\right]$ with derivative h';

(ii) $\bar{C}\left(\left|h' \circ \mathcal{F}_k\right|, \mathcal{X}\right) \vee \bar{C}\left(\left|\mathcal{F}_k\right|, \mathcal{X}\right) < \infty$;

(iii) $\mathcal{F}_k$ is a VC-type class with constants $l_{\mathsf{vc}}(\mathcal{F}_k) \geq e$ and $u_{\mathsf{vc}}(\mathcal{F}_k) \geq 1$ satisfying (2.11) w.r.t. a constant envelope M_k (note that this implies $\bar{C}(|\mathcal{F}_k|, \mathcal{X}) \leq M_k$);

(iv) $\mathcal{F}_k$ *is point-wise measurable, i.e., there exists a countable subclass* $\mathcal{F}'_k \subseteq \mathcal{F}_k$ *of measurable functions such that for any* $f \in \mathcal{F}_k$*, there is a sequence of functions* $\{f_j\}_{j\in\mathbb{N}} \subset \mathcal{F}'_k$ *for which* $\lim_{j\to\infty} f_j(x) = f(x),\ \forall x \in \mathcal{X}$.

Then, for every $k, n \in \mathbb{N}$*, we have*

$$\sup_{\substack{\mu,\nu\in\mathcal{P}(\mathcal{X}):\\ \mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu)<\infty}} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{h,\mathcal{F}_k}(X^n,Y^n) - \mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu)\right|\right]$$

$$\lesssim M_k\Big(\bar{C}\left(\left|h'\circ\mathcal{F}_k\right|,\mathcal{X}\right)+1\Big)n^{-\frac{1}{2}}\int_0^1\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(M_k\epsilon,\mathcal{F}_k,\|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon \quad \text{(A.29)}$$

$$\lesssim \big(u_{\mathsf{vc}}(\mathcal{F}_k)\log l_{\mathsf{vc}}(\mathcal{F}_k)\big)^{\frac{1}{2}} M_k\left(\bar{C}\left(\left|h'\circ\mathcal{F}_k\right|,\mathcal{X}\right)+1\right)n^{-\frac{1}{2}}. \quad \text{(A.30)}$$

The proof of this theorem is based on standard maximal inequalities from empirical process theory, and is presented in Appendix A.1.5 below.

To prove (3.4), we first verify that the relevant assumptions given in Theorem 58 hold with $\mathcal{F}_k = \mathcal{G}_k^{\mathsf{R}}(a)$ and a constant envelope $M_k = 3a(\|\mathcal{X}\|+1)$. The proof for $\mathcal{G}_k^{\mathsf{S}}(a)$ is similar, and hence omitted. Conditions (i) and (ii) are satisfied by the hypotheses in the theorem. Condition (iii) holds as

$$\sup_{\gamma\in\mathcal{P}(\mathcal{X})} N\big(M_k\epsilon,\mathcal{G}_k^{\mathsf{R}}(a),\|\cdot\|_{2,\gamma}\big) \le N\big(M_k\epsilon,\mathcal{G}_k^{\mathsf{R}}(a),\|\cdot\|_{\infty,\mathcal{X}}\big) \le \big(1+7\epsilon^{-1}\big)^{(d+2)k+d+1}, \quad \text{(A.31)}$$

for any $0<\epsilon\le 1$, where the last inequality follows from (A.16). To verify condition (iv), note that $g \in \mathcal{G}_k^{\mathsf{R}}(a)$ is measurable since it is a finite linear combination of compositions of an affine function with a continuous activation. Moreover, point-wise measurability of $\mathcal{G}_k^{\mathsf{R}}(a)$ follows by the continuity of activation and the fact that each of the finite number of parameters of $\mathcal{G}_k^{\mathsf{R}}(a)$ can be approximated arbitrary well by rational numbers.

Next, we evaluate the entropy integral term in (A.29) by bounding $N\big(M_k\epsilon,\mathcal{G}_k^{\mathsf{R}}(a),\|\cdot\|_{2,\gamma}\big)$. For this purpose, let $\mathcal{G}_k^{\dagger}(a) := \mathcal{G}_k(1,2k^{-1}a,0,0,\phi_{\mathsf{R}})$. For any $g_\theta, g_{\tilde{\theta}} \in \mathcal{G}_k^{\mathsf{R}}(a)$, where $g_\theta = \sum_{i=1}^k \beta_i\phi_{\mathsf{R}}\left(w_i\cdot x+b_i\right)+w_0\cdot x+b_0$ and $g_{\tilde{\theta}} = \sum_{i=1}^k \tilde{\beta}_i\phi_{\mathsf{R}}\left(\tilde{w}_i\cdot x+\tilde{b}_i\right)+\tilde{w}_0\cdot x+\tilde{b}_0$, we have

$$\left\|g_\theta - g_{\tilde{\theta}}\right\|_{2,\gamma} \le \left\|\sum_{i=1}^k \beta_i\phi_{\mathsf{R}}\left(w_i\cdot x+b_i\right) - \sum_{i=1}^k \tilde{\beta}_i\phi_{\mathsf{R}}(\tilde{w}_i\cdot x+\tilde{b}_i)\right\|_{2,\gamma} + \|w_0-\tilde{w}_0\|_1\|\mathcal{X}\| + |b_0-\tilde{b}_0|.$$

Hence,

$$\begin{aligned} N\big(\epsilon,\mathcal{G}_k^{\mathsf{R}}(a),\|\cdot\|_{2,\gamma}\big) &\le N\big(\epsilon/3,\mathcal{G}_k^{\dagger}(a),\|\cdot\|_{2,\gamma}\big)N(\epsilon/3,B_d^1(a),\|\mathcal{X}\|\,\|\cdot\|_1)N\big(\epsilon/3,B_1(a),|\cdot|\big) \\ &\le N\big(\epsilon/3,\mathcal{G}_k^{\dagger}(a),\|\cdot\|_{2,\gamma}\big)(1+6a(\|\mathcal{X}\|+1)\epsilon^{-1})^{d+1}, \end{aligned} \quad \text{(A.32)}$$

where (A.32) uses (A.19).

Consider $g = \sum_{i=1}^k \beta_i\phi_{\mathsf{R}}\left(w_i\cdot x+b_i\right) \in \mathcal{G}_k^{\dagger}(a)$. Let $\mathcal{F} = 2a\phi_{\mathsf{R}}\circ\tilde{\mathcal{F}}$, where $\tilde{\mathcal{F}} = \{f:\mathcal{X}\to\mathbb{R}: f = w\cdot x+b, w\in\mathbb{R}^d, b\in\mathbb{R}, \|w\|_1\vee|b|\le 1\}$. By considering the d coordinate projections $f_i(x) = x_i$, $1\le i\le d$, and $f_{d+1}(x) = 1$ spanning the finite dimensional vector space $\tilde{\mathcal{F}}$, we have from (Van Der Vaart and Wellner, 1996, Lemma 2.6.15) that $\mathcal{F}$ is a VC subgraph class of index atmost $d+3$. This uses the fact that if $\tilde{\mathcal{F}}$ is a VC subgraph class of index v and ϕ is monotone, then $\phi \circ \tilde{\mathcal{F}}$ is a VC subgraph class of index at most v, which follows from the proof of (Van Der Vaart and Wellner, 1996, Lemma 2.6.18 (viii)). Then, (Van Der Vaart and Wellner, 1996, Theorem 2.6.7) yields $N\big(2a(\|\mathcal{X}\|+1)\epsilon, \mathcal{F}', \|\cdot\|_{2,\gamma}\big) \lesssim (d+3)(16e)^{d+3}\epsilon^{-2(d+2)}$, where $\mathcal{F}' := \mathcal{F} \cup -\mathcal{F}$. Further, by a careful inspection of the proof of (Giné and Nickl, 2015, Theorem 3.6.17), we obtain that $\log N\big(2a(\|\mathcal{X}\|+1)\epsilon, \overline{\mathsf{co}}(\mathcal{F}'), \|\cdot\|_{2,\gamma}\big) \lesssim d\epsilon^{-2(d+2)/(d+3)}$, where $\overline{\mathsf{co}}(\mathcal{F}')$ denotes the sequential closure of the convex hull of $\mathcal{F}'$ given by $\mathsf{co}(\mathcal{F}') := \{\sum_{i=1}^k \lambda_i f_i : f_i \in \mathcal{F}', \sum_{i=1}^k \lambda_i = 1, \lambda_i \geq 0, \ \forall\ 1 \leq i \leq k,\ k \in \mathbb{N}\}$. Since $\mathcal{G}_k^{\dagger}(a) \subseteq \overline{\mathsf{co}}(\mathcal{F}')$, we have $\log N\big(2a(\|\mathcal{X}\|+1)\epsilon, \mathcal{G}_k^{\dagger}(a), \|\cdot\|_{2,\gamma}\big) \lesssim d\epsilon^{-2(d+2)/(d+3)}$. Hence,

$$\begin{aligned}
&\int_0^1 \sqrt{\sup_{\gamma \in \mathcal{P}(\mathcal{X})} \log N\left(M_k\epsilon, \mathcal{G}_k^{\mathsf{R}}(a), \|\cdot\|_{2,\gamma}\right)} \mathrm{d}\epsilon \\
&\overset{(a)}{\leq} \int_0^1 \sqrt{\sup_{\gamma \in \mathcal{P}(\mathcal{X})} \log N\left(M_k\epsilon/3, \mathcal{G}_k^{\dagger}(a), \|\cdot\|_{2,\gamma}\right)} \mathrm{d}\epsilon + \sqrt{d+1}\int_0^1 \sqrt{\log\left(1+2\epsilon^{-1}\right)}\mathrm{d}\epsilon \\
&\lesssim \sqrt{d}\int_0^1 (0.5\epsilon)^{-(d+2)/(d+3)}\mathrm{d}\epsilon + \sqrt{d}\int_0^1 \sqrt{\log\left(1+2\epsilon^{-1}\right)}\mathrm{d}\epsilon \\
&\lesssim d^{\frac{3}{2}},
\end{aligned} \tag{A.33}$$

where (a) uses (A.32) and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$; and the second integral in the penultimate step can be evaluated by applying $\log(1+x) \leq x$ for $x \geq 0$. This completes the proof of (3.4) via (A.29).

A.1.5 Proof of Theorem 58

To simplify notation, we will denote $\bar{C}\left(|\mathcal{F}_k|, \mathcal{X}\right)$ by $\bar{C}\left(|\mathcal{F}_k|\right)$. Fix $\mu, \nu \in \mathcal{P}(\mathcal{X})$ such that $\mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu) < \infty$. Note that

$$\hat{\mathsf{D}}_{h,\mathcal{F}_k}(X^n, Y^n) - \mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu) \leq \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{F}_k} \frac{1}{\sqrt{n}} \sum_{i=1}^n \big(f(X_i) - \mathbb{E}_\mu[f] - h \circ f(Y_i) + \mathbb{E}_\nu[h \circ f]\big).$$

Let μ_n and ν_n denote the empirical measures $n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $n^{-1}\sum_{i=1}^n \delta_{Y_i}$, where δ_x denotes the Dirac measure centered at $x \in \mathcal{X}$. Then, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\mathsf{D}}_{h,\mathcal{F}_k}(X^n, Y^n) - \mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu)\right|\right] \\
&\leq n^{-\frac{1}{2}}\ \mathbb{E}\left[\sup_{f \in \mathcal{F}_k} n^{-\frac{1}{2}} \left|\sum_{i=1}^n \big(f(X_i) - \mathbb{E}_\mu[f]\right|\right] + n^{-\frac{1}{2}}\ \mathbb{E}\left[\sup_{f \in \mathcal{F}_k} n^{-\frac{1}{2}} \left|\sum_{i=1}^n h \circ f(Y_i) - \mathbb{E}_\nu[h \circ f]\big)\right|\right] \\
&\overset{(a)}{\lesssim} n^{-\frac{1}{2}}\mathbb{E}\left[\int_0^\infty \sqrt{\log N\big(\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\mu_n}\big)}\mathrm{d}\epsilon + \int_0^\infty \sqrt{\log N\big(\epsilon, h \circ \mathcal{F}_k, \|\cdot\|_{2,\nu_n}\big)}\mathrm{d}\epsilon\right] \\
&\overset{(b)}{\leq} n^{-\frac{1}{2}} \int_0^\infty \sqrt{\sup_{\gamma \in \mathcal{P}(\mathcal{X})} \log N\left(\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon \\
&\qquad\qquad + n^{-\frac{1}{2}} \int_0^\infty \sqrt{\sup_{\gamma \in \mathcal{P}(\mathcal{X})} \log N\left(\epsilon, \mathcal{F}_k, \bar{C}\left(|h' \circ \mathcal{F}_k|\right)\|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{=} n^{-\frac{1}{2}} \int_0^{2M_k} \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right)} \mathrm{d}\epsilon \\
&\qquad\qquad + n^{-\frac{1}{2}} \int_0^{2M_k \bar{C}(|h'\circ\mathcal{F}_k|)} \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(\epsilon\big(\bar{C}\left(|h'\circ\mathcal{F}_k|\right)\big)^{-1}, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right)} \mathrm{d}\epsilon \\
&\lesssim M_k\Big(\bar{C}\left(\big|h'\circ\mathcal{F}_k\big|\right)+1\Big) n^{-\frac{1}{2}} \int_0^1 \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(M_k\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right)} \mathrm{d}\epsilon \qquad \text{(A.34)}\\
&\overset{(d)}{\leq} M_k (u_{\mathsf{vc}}(\mathcal{F}_k))^{\frac{1}{2}} \Big(\bar{C}\left(\big|h'\circ\mathcal{F}_k\big|\right)+1\Big) n^{-\frac{1}{2}} \int_0^1 \sqrt{\log\left(1+l_{\mathsf{vc}}(\mathcal{F}_k)\epsilon^{-1}\right)} \mathrm{d}\epsilon \\
&\overset{(e)}{\lesssim} M_k\big(u_{\mathsf{vc}}(\mathcal{F}_k)\log l_{\mathsf{vc}}(\mathcal{F}_k)\big)^{\frac{1}{2}} \Big(\bar{C}\left(\big|h'\circ\mathcal{F}_k\big|\right)+1\Big) n^{-\frac{1}{2}}, \qquad \text{(A.35)}
\end{aligned}
$$

where

(a) follows via an application of (Van Der Vaart and Wellner, 1996, Corollary 2.2.8) since for fixed $(X^n, Y^n) = (x^n, y^n)$, Hoeffding's inequality implies that $n^{-\frac{1}{2}} \sum_{i=1}^n \sigma_i f(x_i)$ and $n^{-\frac{1}{2}} \sum_{i=1}^n h\circ f(y_i)\sigma_i$ are sub-Gaussian w.r.t. pseudo-metrics $\|\cdot\|_{2,\mu_n}$ and $\|\cdot\|_{2,\nu_n}$, respectively;

(b) is due to

$$
\begin{aligned}
N\big(\epsilon, h\circ\mathcal{F}_k, \|\cdot\|_{2,\gamma}\big) &\leq N\big(\epsilon, \mathcal{F}_k, \bar{C}\big(\big|h'\circ\mathcal{F}_k\big|\big)\,\|\cdot\|_{2,\gamma}\big) \\
&= N\Big(\epsilon\big(\bar{C}\big(\,\big|h'\circ\mathcal{F}_k\big|\,\big)\big)^{-1}, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\Big), \qquad \text{(A.36)}
\end{aligned}
$$

which in turn follows from (A.25), and taking supremum w.r.t. to $\gamma\in\mathcal{P}(\mathcal{X})$;

(c) follows since $N\left(\epsilon, \mathcal{F}_k, \bar{C}\left(|h'\circ\mathcal{F}_k|\right)\|\cdot\|_{2,\gamma}\right) = 1$ for $\epsilon \geq 2M_k\bar{C}\left(|h'\circ\mathcal{F}_k|\right)$, $N\left(\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right) = 1$ for $\epsilon \geq 2M_k$, and $N\left(\epsilon, \mathcal{F}_k, \bar{C}\left(|h'\circ\mathcal{F}_k|\right)\|\cdot\|_{2,\gamma}\right) = N\left((\bar{C}\left(|h'\circ\mathcal{F}_k|\right))^{-1}\epsilon, \mathcal{F}_k, \|\cdot\|_{2,\gamma}\right)$ (note that both sides equal 1 when $\bar{C}\left(|h'\circ\mathcal{F}_k|\right) = 0$).

(d) is because $\mathcal{F}_k$ is assumed to be a VC-type class with constants $l_{\mathsf{vc}}(\mathcal{F}_k) \geq e$ and $u_{\mathsf{vc}}(\mathcal{F}_k)$ corresponding to envelope M_k;

(e) is since $\int_0^\delta \sqrt{\log(A/\epsilon)}\mathrm{d}\epsilon \lesssim \delta\sqrt{\log(A/\delta)}$ for $A \geq e$ and $0 \leq \delta \leq 1$, which in turn follows via integration by parts.

Taking supremum on both sides of (A.35) over μ, ν such that $\mathsf{D}_{h,\mathcal{F}_k}(\mu,\nu) < \infty$ proves (A.30).

A.2 Proofs for Section 4

A.2.1 Proof of Theorem 14

Let $\mathsf{D}_{\mathcal{G}_k(\mathbf{a}_k,\phi)}(\mu,\nu) := \mathsf{D}_{h_{\mathsf{KL}},\mathcal{G}_k(\mathbf{a}_k,\phi)}(\mu,\nu)$ be the parametrized (by the NN class $\mathcal{G}_k(\mathbf{a}_k,\phi)$) KL divergence. We will use the following lemma which proves consistency of parametrized KL divergence estimator.

Lemma 59 (Parametrized KL divergence estimation) *Let $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$. Then, for any $0<\rho<1$, and n,k_n, such that $k_n^{3/2}(\|\mathcal{X}\|+1)e^{k_n(\|\mathcal{X}\|+1)}=O\left(n^{(1-\rho)/2}\right)$,*

$$\hat{\mathsf{D}}_{\mathcal{G}_k^*(\phi)}(X^n,Y^n)\xrightarrow[n\to\infty]{}\mathsf{D}_{\mathcal{G}_k^*(\phi)}(\mu,\nu),\quad \mathbb{P}-a.s. \tag{A.37}$$

Lemma 59 is proven using Theorem 11; see Appendix B.3 for details.

We proceed with the proof of (4.2). Since $\mathcal{X}$ is compact and $f_{\mathsf{KL}}\in\mathsf{C}(\mathcal{X})$, it follows from (Stinchcombe and White, 1990, Theorem 2.1 and 2.8) that for any $\epsilon>0$, there is a $k_0(\epsilon)\in\mathbb{N}$, such that for any $k\geq k_0(\epsilon)$, there exists a $g_{\theta_k}\in\mathcal{G}_k^*(\phi)$ with

$$\|f_{\mathsf{KL}}-g_{\theta_k}\|_{\infty,\mathcal{X}}\leq\epsilon. \tag{A.38}$$

This implies

$$\lim_{k\to\infty}\mathsf{D}_{\mathcal{G}_k^*(\phi)}(\mu,\nu)=\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right). \tag{A.39}$$

To see this, note that

$$\mathsf{D}_{\mathcal{G}_k^*(\phi)}(\mu,\nu)\leq\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right),\quad\forall\ k\in\mathbb{N}, \tag{A.40}$$

by (2.2) since $g\in\mathcal{G}_k^*(\phi)$ is continuous and bounded ($\|g\|_{\infty,\mathcal{X}}\leq k(\|\mathcal{X}\|+1)+1\leq 2k+1$ for $\mathcal{X}=[0,1]^d$). Moreover, the left-hand side (LHS) of (A.40) is monotonically increasing in k, and being bounded, it has a limit point. Thus, to establish (A.39), it suffices to show that this limit point is $\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)$.

Assume to the contrary that $\lim_{k\to\infty}\mathsf{D}_{\mathcal{G}_k^*(\phi)}(\mu,\nu)<\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)$. Note that $\mathcal{G}_k^*(\phi)$ is a compact set and hence the supremum in the variational form of the LHS of (A.40) is a maximum. Then, defining $D(g):=1+\mathbb{E}_\mu[g]-\mathbb{E}_\nu[e^g]$, it follows that there exists $\delta>0$ and $g_{\bar{\theta}_k}\in\arg\max_{g_\theta\in\mathcal{G}_k^*(\phi)}D(g_\theta)$ such that for all k,

$$\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)-D\left(g_{\bar{\theta}_k}\right)\geq\delta. \tag{A.41}$$

However, we have for all $k\geq k_0(\epsilon)$ that

$$\begin{aligned}
\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)-D(g_{\bar{\theta}_k})&\leq\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)-D(g_{\theta_k})\\
&\leq\mathbb{E}_\mu\left[|f_{\mathsf{KL}}-g_{\theta_k}|\right]+\mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}}-e^{g_{\theta_k}}\right|\right]\\
&\leq\mathbb{E}_\mu\left[|f_{\mathsf{KL}}-g_{\theta_k}|\right]+\mathbb{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right]\left\|1-e^{g_{\theta_k}-f_{\mathsf{KL}}}\right\|_{\infty,\nu}\\
&\leq\epsilon+e^\epsilon-1,
\end{aligned}$$

where the final inequality follows from (A.38) and $\mathbb{E}_\nu\left[\mathrm{d}\mu/\mathrm{d}\nu\right]\leq 1$. Then, taking ϵ sufficiently small contradicts (A.41), thus proving (A.39). From this and (A.37) with $k=k_n\to\infty$, (4.2) follows since $k^{3/2}e^{k(\|\mathcal{X}\|+1)}<e^{k(2+\delta)}$ for $\mathcal{X}=[0,1]^d$, any $\delta>0$, and k sufficiently large.

Next, we prove (4.3). Fix $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$, and with some abuse of notation, let $\mathbf{m}=(m_k)_{k\in\mathbb{N}}$ be a non-decreasing positive divergent sequence, and note that since $c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}},\mathcal{X}\right)\leq$

M, we have from (3.1) that for k such that $m_k \geq M$, there exists $g_{\theta_k} \in \mathcal{G}_k^{\mathsf{R}}(m_k)$ and $c>0$ satisfying

$$\left\|f_{\mathsf{KL}} - g_{\theta_k}\right\|_{\infty,\mathcal{X}} \leq c d^{\frac{1}{2}} M k^{-\frac{1}{2}}. \tag{A.42}$$

Also, since $g_{\theta_k} \in \mathcal{G}_k^{\mathsf{R}}(m_k)$ is bounded, we have that $\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) \geq \mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu)$. Then, the following hold for k such that $m_k \geq M$ and $c^2 d M^2 \leq k/2$:

$$\begin{aligned}\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu)\right| &= \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu)\\ &\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{KL}} - g_{\theta_k}\right|\right] + \left\|1 - e^{g_{\theta_k} - f_{\mathsf{KL}}}\right\|_{\infty,\nu} \mathbb{E}_\nu\left[e^{f_{\mathsf{KL}}}\right]\\ &\lesssim d^{\frac{1}{2}} M k^{-\frac{1}{2}},\end{aligned}$$

where the last bound follows from (A.42), $\mathbb{E}_\nu\left[e^{f_{\mathsf{KL}}}\right] = \mathbb{E}_\nu\left[\mathrm{d}\mu/\mathrm{d}\nu\right] = 1$, and since

$$\left\|1 - e^{g_{\theta_k} - f_{\mathsf{KL}}}\right\|_{\infty,\nu} \leq \sum_{j=1}^{\infty} \frac{\left(c d^{\frac{1}{2}} M k^{-\frac{1}{2}}\right)^j}{j!} \leq \sum_{j=1}^{\infty} \left(c d^{\frac{1}{2}} M k^{-\frac{1}{2}}\right)^j \lesssim d^{\frac{1}{2}} M k^{-\frac{1}{2}}. \tag{A.43}$$

Next, note that $\mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu) \geq 0$ as $g = 0 \in \mathcal{G}_k^{\mathsf{R}}(m_k)$. This implies that for k with $m_k < M$ or $c^2 d M^2 > k/2$, we have $\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu)\right| \leq \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) \leq M$. Consequently

$$\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu)\right| \lesssim_{\mathbf{m},M} d^{\frac{1}{2}} k^{-\frac{1}{2}}, \quad \forall\, k \in \mathbb{N}.$$

On the other hand, since $\bar{C}\left(\left|\mathcal{G}_k^{\mathsf{R}}(m_k)\right|, \mathcal{X}\right) \leq 3 m_k(\|\mathcal{X}\|+1)$ and $\bar{C}\left(\left|h_{\mathsf{KL}} \circ \mathcal{G}_k^{\mathsf{R}}(m_k)\right|, \mathcal{X}\right) \leq e^{3 m_k(\|\mathcal{X}\|+1)}$, it follows from the above, (A.29) and (A.33) that

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]\\ &\qquad \leq \left|\mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right| + \mathbb{E}\left[\left|\mathsf{D}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(\mu,\nu) - \hat{\mathsf{D}}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(X^n,Y^n)\right|\right]\\ &\qquad \lesssim_{\mathbf{m},M} d^{\frac{1}{2}} k^{-\frac{1}{2}} + d^{\frac{3}{2}} m_k(\|\mathcal{X}\|+1) e^{3 m_k(\|\mathcal{X}\|+1)}\, n^{-\frac{1}{2}}.\end{aligned} \tag{A.44}$$

Since $\|\mathcal{X}\| = 1$, choosing $m_k = \log\log k \vee 1$ in (A.44) yields

$$\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k^{\mathsf{R}}(m_k)}(X^n,Y^n) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right] \lesssim_M d^{\frac{1}{2}} k^{-\frac{1}{2}} + d^{\frac{3}{2}} (\log k)^7\, n^{-\frac{1}{2}}.$$

Noting that the above bound holds independent of $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$, the proof is completed by taking supremum w.r.t. such μ,ν. Note that setting $m_k = M$ in (A.44) and taking supremum w.r.t. such μ,ν yields (4.4) from Remark 16.

A.2.2 Proof of Corollary 18

To show that the minimax risk is $\Omega(n^{-1/2})$, it suffices to consider $d=1$. Recall that the minimax risk for differential entropy estimation over the class of one-dimensional Gaussian distributions with unknown variance in a non-empty interval is $\Omega(n^{-1/2})$ (cf., e.g., Appendix

A, Goldfeld et al., 2020b). Take $\mathcal{X}=[a,b]$, for some $a,b\in\mathbb{R}$ with $b>a$, and let $\mathcal{P}_{\mathsf{AC}}(\mathcal{X})$ be the class of (Lebesgue absolutely continuous) distributions on $\mathcal{X}$. The differential entropy of $\mu\in\mathcal{P}_{\mathsf{AC}}(\mathcal{X})$ is defined as $\mathsf{h}(\mu):=-\mathbb{E}_\mu\left[\log\left(\mathrm{d}\mu/\mathrm{d}\lambda\right)\right]$, and can be equivalently written as $\mathsf{h}(\mu)=\log(b-a)-\mathsf{D}_{\mathsf{KL}}\big(\mu\|u_{[a,b]}\big)$, where $u_{[a,b]}$ is the uniform distribution on $\mathcal{X}$. Hence, the minimax rate of KL divergence estimation for distributions in the class $\{(\mu,u_{[a,b]}):\mu\in\mathcal{P}_{\mathsf{AC}}(\mathcal{X})\}$ for any $\tilde{\mathcal{P}}_{\mathsf{AC}}(\mathcal{X})\subseteq\mathcal{P}_{\mathsf{AC}}(\mathcal{X})$ is the same as that of differential entropy estimation for distributions in $\tilde{\mathcal{P}}_{\mathsf{AC}}(\mathcal{X})$.

Let $\mathcal{P}_{\mathsf{TG}}(\mathcal{X})\subset\tilde{P}_{\mathsf{AC}}(\mathcal{X})$ be a class of truncated Gaussians supported on $\mathcal{X}$ with zero mean and variance in an non-empty interval. Note that the minimax rate for differential entropy estimation over $\mathcal{P}_{\mathsf{TG}}(\mathcal{X})$ equals to that over untrucated Gaussian distribution with zero mean and the same variance constraints. This is since both differential entropies are elementary functions of the variance parameter, when the mean (equals zero) and a,b are given. By Proposition 19 (see Remark 20), $\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ contains pairs of truncated Gaussians (with variance and means within an interval that depends on M) and uniform distributions, which implies that the associated KL divergence minimax estimation risk is $\Omega(n^{-1/2})$. The corollary then follows by noting that the NE achieves $O\big(n^{-1/2}\big)$ error rate by setting $k=n$ in (4.4).

A.2.3 Proof of Proposition 19

The proof of Proposition 10 (see (A.14)) shows that there exists extensions $\tilde{p}_{\mathsf{ext}},\tilde{q}_{\mathsf{ext}}\in\mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\big(\mathbb{R}^d\big)\cap\mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big)$ of $\tilde{p},\tilde{q}$, respectively, where $\bar{c}_{b,d,\|\mathcal{X}\|}=(\kappa_d d^{3/2}\left\|\mathcal{X}\right\|\vee 1)b'$, with b' as defined in (A.15). Set $f^{\mathsf{ext}}_{\mathsf{KL}}:=\tilde{p}_{\mathsf{ext}}-\tilde{q}_{\mathsf{ext}}$, and note that since $\tilde{p}_{\mathsf{ext}},\tilde{q}_{\mathsf{ext}}\in\mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big)$, their Fourier transforms exist and the corresponding Fourier inversion formulas hold (see proof of Lemma 55). Also, we have

$$S_2\left(f^{\mathsf{ext}}_{\mathsf{KL}}\right)\left\|\mathcal{X}\right\|\leq S_2\left(\tilde{p}_{\mathsf{ext}}\right)\left\|\mathcal{X}\right\|+S_2\left(\tilde{q}_{\mathsf{ext}}\right)\left\|\mathcal{X}\right\|\leq 2\bar{c}_{b,d,\|\mathcal{X}\|},$$

where the first inequality uses the definition in (2.9) and linearity of the Fourier transform, while the second is because $\tilde{p}_{\mathsf{ext}},\tilde{q}_{\mathsf{ext}}\in\mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\big(\mathbb{R}^d\big)$. Moreover, note that

$$\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)=\mathbb{E}_\mu\left[f_{\mathsf{KL}}\right]=\mathbb{E}_\mu\left[\log p-\log q\right]\leq 2b,$$

where the final inequality is due to $\log p=\tilde{p}|_{\mathcal{X}}$ and $\log q=\tilde{q}|_{\mathcal{X}}$, for $\tilde{p},\tilde{q}\in\mathsf{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$. Lastly, since $f_{\mathsf{KL}}=f^{\mathsf{ext}}_{\mathsf{KL}}\big|_{\mathcal{X}}$, it follows that $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{KL}}(M,\mathcal{X})$ with $M=2\bar{c}_{b,d,\|\mathcal{X}\|}\vee 2b$, and the proposition then follows from Theorem 14.

A.2.4 Proof of Theorem 22

Let $\chi^2_{\mathcal{G}_k(\mathbf{a}_k,\phi)}(\mu,\nu):=\mathsf{D}_{h_{\chi^2},\mathcal{G}_k(\mathbf{a}_k,\phi)}(\mu,\nu)$. We will use the lemma below which proves consistency of parametrized χ^2 divergence estimator (see Appendix B.4 for proof).

Lemma 60 (Parametrized χ^2 divergence estimation) *Let $(\mu,\nu)\in\mathcal{P}^2_{\chi^2}(\mathcal{X})$. Then, for any $0<\rho<1$, and n,k_n such that $k_n^{5/2}(\|\mathcal{X}\|+1)^2=O\left(n^{(1-\rho)/2}\right)$, we have*

$$\hat{\chi}^2_{\mathcal{G}^*_k(\phi)}(X^n,Y^n)\xrightarrow[n\to\infty]{}\chi^2_{\mathcal{G}^*_k(\phi)}(\mu,\nu),\quad\mathbb{P}-a.s.\tag{A.45}$$

Proceeding with the proof of Theorem 22, (4.5) follows from (A.45) using arguments similar to those used to establish (4.2) and steps leading to (A.49) below; details are omitted.

To prove (4.6), fix $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(M,\mathcal{X})$, and let $\mathbf{m} = (m_k)_{k\in\mathbb{N}}$ be a non-decreasing positive divergent sequence. Since $c^{\star}_{\mathsf{KB}}\left(f_{\chi^2},\mathcal{X}\right) \le M$, we have from (3.1) that for k such that $m_k \ge M$, there exists $g_{\theta_k} \in \mathcal{G}^{\mathsf{R}}_k(m_k)$ with

$$\left\|f_{\chi^2} - g_{\theta_k}\right\|_{\infty,\mathcal{X}} \lesssim M d^{\frac{1}{2}} k^{-\frac{1}{2}}. \tag{A.46}$$

Also, $\chi^2\left(\mu\|\nu\right) \ge \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu)$ because $g \in \mathcal{G}^{\mathsf{R}}_k(m_k)$ is bounded. Then, we have

$$\begin{aligned}
\left|\chi^2\left(\mu\|\nu\right) - \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu)\right| &= \chi^2\left(\mu\|\nu\right) - \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu) \\
&\le \chi^2\left(\mu\|\nu\right) - \mathbb{E}_\mu\left[g_{\theta_k}\right] - \mathbb{E}_\nu\left[g_{\theta_k} + 0.25 g^2_{\theta_k}\right] \\
&\le \mathbb{E}_\mu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|f_{\chi^2} - g_{\theta_k}\right| + 0.25\left|f^2_{\chi^2} - g^2_{\theta_k}\right|\right] \qquad \text{(A.47)} \\
&\lesssim M d^{\frac{1}{2}} k^{-\frac{1}{2}} + \mathbb{E}_\nu\left[0.25\left|f_{\chi^2} - g_{\theta_k}\right|\left|f_{\chi^2} + g_{\theta_k}\right|\right] \\
&\lesssim M d^{\frac{1}{2}} k^{-\frac{1}{2}} + \mathbb{E}_\nu\left[0.25\left|f_{\chi^2} - g_{\theta_k}\right|^2 + 0.5\left|f_{\chi^2} - g_{\theta_k}\right|\left|f_{\chi^2}\right|\right] \\
&\lesssim M d^{\frac{1}{2}} k^{-\frac{1}{2}} + M^2 d k^{-1} + 0.5\left\|f_{\chi^2} - g_{\theta_k}\right\|_{\infty,\nu}\mathbb{E}_\nu\left[\left|f_{\chi^2}\right|\right] \qquad \text{(A.48)} \\
&\lesssim M(M+1) d k^{-\frac{1}{2}},
\end{aligned}$$

where the final inequality is due to (A.46) and since $\mathbb{E}_\nu\left[\left|f_{\chi^2}\right|\right] \le \mathbb{E}_\nu\left[2(\mathrm{d}\mu/\mathrm{d}\nu)+2\right] \le 4$.

Since $g = 0 \in \mathcal{G}^{\mathsf{R}}_k(m_k)$, for k such that $m_k < M$, we have

$$\left|\chi^2\left(\mu\|\nu\right) - \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu)\right| = \chi^2\left(\mu\|\nu\right) - \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu) \le \chi^2\left(\mu\|\nu\right) \le M.$$

Hence,

$$\left|\chi^2\left(\mu\|\nu\right) - \chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu)\right| \lesssim_{\mathbf{m},M} d k^{-\frac{1}{2}}, \quad \forall k \in \mathbb{N}. \tag{A.49}$$

Since $\bar{C}\left(\left|\mathcal{G}^{\mathsf{R}}_k(m_k)\right|,\mathcal{X}\right) \le 3m_k(\|\mathcal{X}\|+1)$ and $\bar{C}\left(\left|h'_{\chi^2}\circ\mathcal{G}^{\mathsf{R}}_k(m_k)\right|,\mathcal{X}\right) \le 1.5m_k(\|\mathcal{X}\|+1)+1$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(X^n,Y^n) - \chi^2\left(\mu\|\nu\right)\right|\right] \\
&\qquad \le \left|\chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu) - \chi^2\left(\mu\|\nu\right)\right| + \mathbb{E}\left[\left|\chi^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(\mu,\nu) - \hat{\chi}^2_{\mathcal{G}^{\mathsf{R}}_k(m_k)}(X^n,Y^n)\right|\right] \\
&\qquad \lesssim_{M,\mathbf{m}} d k^{-\frac{1}{2}} + d^{\frac{3}{2}} m_k^2 (\|\mathcal{X}\|+1)^2 n^{-\frac{1}{2}},
\end{aligned} \tag{A.50}$$

where the last inequality uses (A.29), (A.33) and (A.49). Setting $m_k = \log k$ in (A.50) and taking supremum w.r.t. $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(M,\mathcal{X})$ yields (4.6).

A.2.5 Proof of Proposition 25

It follows from (A.14) that there exists extensions $\tilde{p}_{\mathsf{ext}}, \tilde{q}_{\mathsf{ext}} \in \mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\left(\mathbb{R}^d\right) \cap \tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\left(\mathbb{R}^d\right)$ of $\tilde{p},\tilde{q} \in \mathsf{C}^{s_{\mathsf{KB}}}_b(\mathcal{U})$, respectively, where $\tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\left(\mathbb{R}^d\right)$ is defined in (A.13) and $\bar{c}_{b,d,\|\mathcal{X}\|} :=$

$(\kappa_d d^{3/2} \|\mathcal{X}\| \vee 1)b'$, with b' from (A.15). Let $f_{\chi^2}^{\mathsf{ext}} = 2\big(\tilde{p}_{\mathsf{ext}}\tilde{q}_{\mathsf{ext}} - 1\big)$ and recall that $\alpha_{|j}$ denotes a multi-index of order j. We have from the product rule of derivatives that for any $j \in \mathbb{Z}_{\geq 0}$,

$$D^{\alpha_{|j}} f_{\chi^2}^{\mathsf{ext}} = 2 \sum_{\alpha_{|j_1}+\alpha_{|j_2}=\alpha_{|j}} \frac{\alpha_{|j}!}{\alpha_{|j_1}!\,\alpha_{|j_2}!} D^{\alpha_{|j_1}}\tilde{p}_{\mathsf{ext}} D^{\alpha_{|j_2}}\tilde{q}_{\mathsf{ext}} - D^{\alpha_{|j}} 2,$$

where $\alpha! := \prod_{i=1}^d \alpha_i!$. Also, note from (A.11) and (A.12) that for $0 \leq j \leq s_{\mathsf{KB}}$, $\tilde{p}_{\mathsf{ext}}$, $\tilde{q}_{\mathsf{ext}}$ satisfies

$$\left\|D^{\alpha_{|j}}\tilde{p}_{\mathsf{ext}}\right\|_{\infty,\mathbb{R}^d} \vee \left\|D^{\alpha_{|j}}\tilde{q}_{\mathsf{ext}}\right\|_{\infty,\mathbb{R}^d} \leq \hat{b} \leq b',$$

$$\left\|D^{\alpha_{|j}}\tilde{p}_{\mathsf{ext}}\right\|_{2,\mathbb{R}^d} \vee \left\|D^{\alpha_{|j}}\tilde{q}_{\mathsf{ext}}\right\|_{2,\mathbb{R}^d} \leq b'.$$

Combining these observations, we have for $0 \leq j \leq s_{\mathsf{KB}}$ that

$$\begin{aligned}\left\|D^{\alpha_{|j}} f_{\chi^2}^{\mathsf{ext}}\right\|_{2,\mathbb{R}^d} &\leq 2 + 2\left\|\sum_{\alpha_{|j_1}+\alpha_{|j_2}=\alpha_{|j}} \frac{\alpha_{|j}!}{\alpha_{|j_1}!\,\alpha_{|j_2}!} D^{\alpha_{|j_1}}\tilde{p}_{\mathsf{ext}} D^{\alpha_{|j_2}}\tilde{q}_{\mathsf{ext}}\right\|_{2,\mathbb{R}^d} \\ &\leq 2 + 2^{j+1}b'^2. \end{aligned} \tag{A.52}$$

Similarly, we have $\left\|D^{\alpha_{|j}} f_{\chi^2}^{\mathsf{ext}}\right\|_1 < \infty$ for $0 \leq j \leq s_{\mathsf{KB}}$. Hence, $f_{\chi^2}^{\mathsf{ext}} \in \tilde{\mathcal{L}}_{s_{\mathsf{KB}},2+2^{s_{\mathsf{KB}}+1}b'^2}\big(\mathbb{R}^d\big)$. From Lemma 55, it follows that $S_2\big(f_{\chi^2}^{\mathsf{ext}}\big) \leq (2 + 2^{s_{\mathsf{KB}}+1}b'^2)\kappa_d d^{3/2}$. Since $f_{\chi^2} = f_{\chi^2}^{\mathsf{ext}}\big|_{\mathcal{X}}$, this implies that $c^\star_{\mathsf{KB}}\left(f_{\chi^2}, \mathcal{X}\right) \leq (2 + 2^{s_{\mathsf{KB}}+1}b'^2)(\kappa_d d^{3/2}\|\mathcal{X}\| \vee 1)$. Also,

$$\chi^2(\mu\|\nu) = \mathbb{E}_\nu\left[\left(pq^{-1}-1\right)^2\right] \leq \mathbb{E}_\nu\left[p^2q^{-2}+1\right] \leq b^2+1.$$

The claim then follows from Theorem 22 by noting that $b'^2 \leq \bar{c}^2_{b,d,\|\mathcal{X}\|}$ and $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(M,\mathcal{X})$ with $M = \big(2 + 2^{s_{\mathsf{KB}}+1}\bar{c}^2_{b,d,\|\mathcal{X}\|}\big)(\kappa_d d^{3/2}\|\mathcal{X}\| \vee 1) \vee (b^2+1)$.

A.2.6 Proof of Theorem 27

Let $\mathsf{H}^2_{\tilde{\mathcal{G}}_{k,t}(\mathbf{a}_k,\phi)}(\mu,\nu) := \mathsf{D}_{h_{\mathsf{H}^2},\tilde{\mathcal{G}}_{k,t}(\mathbf{a}_k,\phi)}(\mu,\nu)$. We need the following lemma (see Appendix B.5 for proof) which shows that parametrized H^2 distance estimation is consistent.

Lemma 61 (Parametrized H^2 distance estimation) *Let $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$. Then, for any $0<\rho<1$, $t_n \to 0$, and n, k_n, such that $k_n^{3/2}(\|\mathcal{X}\|+1)t_n^{-2} = O\left(n^{(1-\rho)/2}\right)$,*

$$\hat{\mathsf{H}}^2_{\tilde{\mathcal{G}}^*_{k_n,t_n}(\phi)}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k_n,t_n}(\phi)}(\mu,\nu), \quad \mathbb{P}-a.s. \tag{A.53}$$

Continuing with the proof of Theorem 27, we first prove (4.10). Fix $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$. Recall that $f_{\mathsf{H}^2} = 1 - (\mathrm{d}\mu/\mathrm{d}\nu)^{-1/2}$. Since $\|\mathrm{d}\mu/\mathrm{d}\nu\|_{\infty,\eta} \leq M$ by assumption, we have $\|(1-f_{\mathsf{H}^2})\|_{\infty,\eta} \geq M^{-1/2}$. It follows from (Stinchcombe and White, 1990, Theorem 2.1 and 2.8) and the definition of $\tilde{\mathcal{G}}^*_{k,t}(\phi)$ that for any $\epsilon > 0$, there exists $k_0(\epsilon) \in \mathbb{N}$ and $g_{\theta_k} \in \tilde{\mathcal{G}}^*_{k,M^{-1/2}}(\phi)$ such that for all $k \geq k_0(\epsilon)$,

$$\|f_{\mathsf{H}^2} - g_{\theta_k}\|_{\infty,\eta} \leq \epsilon. \tag{A.54}$$

Then, noting that $\mathsf{H}^2(\mu,\nu) \geq \mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,M^{-1/2}}(\phi)}(\mu,\nu)$, we have

$$\begin{aligned}
\left|\mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,M^{-1/2}}(\phi)}(\mu,\nu)\right| &= \mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,M^{-1/2}}(\phi)}(\mu,\nu) \\
&\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{H}^2} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|f_{\mathsf{H}^2}(1-f_{\mathsf{H}^2})^{-1} - g_{\theta_k}\left(1-g_{\theta_k}\right)^{-1}\right|\right] \\
&\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{H}^2} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|\left(f_{\mathsf{H}^2} - g_{\theta_k}\right)(1-f_{\mathsf{H}^2})^{-1}\left(1-g_{\theta_k}\right)^{-1}\right|\right] \\
&\leq \epsilon + M\epsilon, \qquad (A.55)
\end{aligned}$$

where the final inequality uses (A.54), $\left\|1-f_{\mathsf{H}^2}\right\|_{\infty,\eta} \wedge \left\|1-g_{\theta_k}\right\|_{\infty,\eta} \geq M^{-1/2}$. Since $\epsilon > 0$ is arbitrary, this implies (similarly to (A.39) in Theorem 14) that

$$\lim_{k\to\infty} \mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,M^{-1/2}}(\phi)}(\mu,\nu) = \mathsf{H}^2(\mu,\nu).$$

Then, (4.10) follows from (A.53) and (A.55).

Next, we prove (4.11). Fix $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$. By some abuse of notation, let $\mathbf{m} = (m_k)_{k\in\mathbb{N}}$ and $\mathbf{t} = (t_k)_{k\in\mathbb{N}}$ denote a non-decreasing positive divergent sequence and a non-increasing sequence tending to zero, respectively. Since $\|\mathrm{d}\mu/\mathrm{d}\nu\|_{\infty,\eta} \leq M$, we have $\|1-f_{\mathsf{H}^2}\|_{\infty,\eta} \geq M^{-1/2}$. Using $t_k \to 0$, it then follows from (3.1) that for k such that $t_k \leq M^{-1/2}$ and $m_k \geq M$, there exists $g_{\theta_k} \in \tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)$ with

$$\|f_{\mathsf{H}^2} - g_{\theta_k}\|_{\infty,\eta} \lesssim M d^{\frac{1}{2}} k^{-\frac{1}{2}}. \qquad (A.56)$$

Then, following the arguments leading to the penultimate step in (A.55), we have

$$\begin{aligned}
&\left|\mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu)\right| \\
&\quad\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{H}^2} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|\left(f_{\mathsf{H}^2} - g_{\theta_k}\right)\left(1-f_{\mathsf{H}^2}\right)^{-1}\left(1-g_{\theta_k}\right)^{-1}\right|\right] \\
&\quad\leq \|f_{\mathsf{H}^2} - g_{\theta_k}\|_{\infty,\mu} + \|f_{\mathsf{H}^2} - g_{\theta_k}\|_{\infty,\nu}\,\mathbb{E}_\nu\left[\left|\left(1-f_{\mathsf{H}^2}\right)^{-1}\left(1-g_{\theta_k}\right)^{-1}\right|\right] \\
&\quad\lesssim_M d^{\frac{1}{2}}(1+t_k^{-1})k^{-\frac{1}{2}},
\end{aligned}$$

where the final inequality is due to (A.56), $1-g_{\theta_k}(x) \geq t_k$ for any $x \in \mathbb{R}^d$, and

$$\mathbb{E}_\nu\left[\left|\left(1-f_{\mathsf{H}^2}\right)^{-1}\right|\right] = \mathbb{E}_\nu\left[\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}}\right] \leq \sqrt{\mathbb{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right]} = 1.$$

Moreover, since $g = 0 \in \tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)$, for k such that $m_k < M$ or $t_k > M^{-1/2}$, we obtain

$$\left|\mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu)\right| = \mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu) \leq \mathsf{H}^2(\mu,\nu) \leq 2,$$

where the last inequality follows from

$$\mathsf{H}^2(\mu,\nu) = \mathbb{E}_\nu\left[\left(\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} - 1\right)^2\right] \leq \mathbb{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu} + 1\right] \leq 2.$$

Thus, for all k, we have

$$\left|\mathsf{H}^2(\mu,\nu)-\mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu)\right|\lesssim_{M,\mathbf{m},\mathbf{t}} d^{\frac{1}{2}}(1+t_k^{-1})k^{-\frac{1}{2}}. \tag{A.57}$$

Noting that $\bar{C}\big(\big|\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)\big|,\mathcal{X}\big)\leq 3m_k(\|\mathcal{X}\|+1)$ and $\bar{C}\big(\big|h'_{\mathsf{H}^2}\circ\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)\big|,\mathcal{X}\big)\leq t_k^{-2}$, it follows from (A.29), (A.33) and (A.57) that

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(X^n,Y^n)-\mathsf{H}^2(\mu,\nu)\right|\right]\\
&\qquad\leq\left|\mathsf{H}^2(\mu,\nu)-\mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu)\right|+\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(X^n,Y^n)-\mathsf{H}^2_{\tilde{\mathcal{G}}^{\mathsf{R}}_{k,t_k}(m_k)}(\mu,\nu)\right|\right]\\
&\qquad\lesssim_{M,\mathbf{m},\mathbf{t}} d^{\frac{1}{2}}(1+t_k^{-1})k^{-\frac{1}{2}}+d^{\frac{3}{2}}(\|\mathcal{X}\|+1)m_k t_k^{-2}n^{-\frac{1}{2}}.
\end{aligned} \tag{A.58}$$

Noting that the above bound holds for any $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$, and setting $m_k=\log k$, $t_k=(\log k)^{-1}$, we obtain (4.11).

A.2.7 Proof of Proposition 30

As in the proof of Proposition 25, (A.14) yields that there exists extensions $\tilde{p}_{\mathsf{ext}},\tilde{q}_{\mathsf{ext}}\in\mathcal{B}_{\bar{c}_{b,d,\|\mathcal{X}\|},2,\mathcal{X}}\big(\mathbb{R}^d\big)\cap\tilde{\mathcal{L}}_{s_{\mathsf{KB}},b'}\big(\mathbb{R}^d\big)$ of $\tilde{p},\tilde{q}$, respectively. Let $f^{\mathsf{ext}}_{\mathsf{H}^2}=1-\tilde{p}_{\mathsf{ext}}\cdot\tilde{q}_{\mathsf{ext}}$. Then, following steps leading to (A.52), we obtain for $0\leq j\leq s_{\mathsf{KB}}$ that

$$\left\|D^{\alpha_{|j}}f^{\mathsf{ext}}_{\mathsf{H}^2}\right\|_2\leq 1+\left\|\sum_{\alpha_{|j_1}+\alpha_{|j_2}=\alpha_{|j}}\frac{\alpha_{|j}!}{\alpha_{|j_1}!\,\alpha_{|j_2}!}D^{\alpha_{|j_1}}\tilde{p}_{\mathsf{ext}}\ D^{\alpha_{|j_2}}\tilde{q}_{\mathsf{ext}}\right\|_2\leq 1+2^j b'^2.$$

Similarly, $\left\|D^{\alpha_{|j}}f^{\mathsf{ext}}_{\mathsf{H}^2}\right\|_1<\infty$ for $0\leq j\leq s_{\mathsf{KB}}$. Hence, $f^{\mathsf{ext}}_{\mathsf{H}^2}\in\tilde{\mathcal{L}}_{s_{\mathsf{KB}},1+2^{s_{\mathsf{KB}}}b'^2}\big(\mathbb{R}^d\big)$, which yields via Lemma 55 that $S_2\big(f^{\mathsf{ext}}_{\mathsf{H}^2}\big)\leq(1+2^{s_{\mathsf{KB}}}b'^2)\kappa_d d^{3/2}$. Since $f_{\mathsf{H}^2}=f^{\mathsf{ext}}_{\mathsf{H}^2}\big|_{\mathcal{X}}$, this implies that $c^\star_{\mathsf{KB}}(f_{\mathsf{H}^2},\mathcal{X})\leq(1+2^{s_{\mathsf{KB}}}b'^2)(\kappa_d d^{3/2}\|\mathcal{X}\|\vee 1)$. Moreover, we have $\|\mathrm{d}\mu/\mathrm{d}\nu\|_{\infty,\eta}=\left\|pq^{-1}\right\|_{\infty,\eta}\leq b^2$. Hence, $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(M,\mathcal{X})$ with $M=(\kappa_d d^{3/2}\|\mathcal{X}\|\vee 1)\big(1+2^{s_{\mathsf{KB}}}\bar{c}^2_{b,d,\|\mathcal{X}\|}\big)\vee b^2$ since $b'^2\leq\bar{c}^2_{b,d,\|\mathcal{X}\|}$. The claim then follows from Theorem 27.

A.2.8 Proof of Theorem 32

Let $\delta_{\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(\mu,\nu):=\mathsf{D}_{h_{\mathsf{TV}},\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(\mu,\nu)$. The proof of Theorem 32 is based on the following lemma which establishes consistency of the parametrized TV distance estimator (see Appendix B.6 for proof).

Lemma 62 (Parametrized TV distance estimation) *Let $\mu,\nu\in\mathcal{P}(\mathcal{X})$. Then, for any $0<\rho<1$, and n,k_n such that $k_n(\|\mathcal{X}\|+1)^{1/2}=O\left(n^{(1-\rho)/2}\right)$,*

$$\hat{\delta}_{\bar{\mathcal{G}}^*_{k_n}(\phi)}(X^n,Y^n)\xrightarrow[n\to\infty]{}\delta_{\bar{\mathcal{G}}^*_{k_n}(\phi)}(\mu,\nu)\,,\quad \mathbb{P}-a.s. \tag{A.59}$$

Equipped with Lemma 62, we first prove (4.14). Since f_{TV} is not continuous, the universal approximation property of NNs used in the consistency proofs until now cannot be used directly in this case. However, we will show that there exists a continuous function approximating f_{TV} to any desired accuracy, which can in turn be approximated by $\bar{\mathcal{G}}_k^*(\phi)$ arbitrary well.

Fix $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Let p and q denote the densities of μ and ν w.r.t. $\eta = 0.5(\mu+\nu) \in \mathcal{P}(\mathcal{X})$, and let $\mathcal{C}^*$ be the set defined in (2.8). Note that $\|p \vee q\|_{\infty,\eta} \leq 2$. Also, observe that $\mathcal{C}^*$ and $\mathcal{X} \setminus \mathcal{C}^*$ are Borel sets, since $p(x)$ and $q(x)$ are Borel measurable by definition, and hence so is $p(x)-q(x)$. Since $\eta \in \mathcal{P}(\mathcal{X})$ is a regular probability measure, for any $\epsilon > 0$, there exists compact sets $\mathcal{C}$, $\bar{\mathcal{C}}$, open sets $\mathcal{U}$,$\bar{\mathcal{U}}$ such that $\mathcal{C} \subseteq \mathcal{C}^* \subseteq \mathcal{U}$, $\bar{\mathcal{C}} \subseteq \mathcal{X} \setminus \mathcal{C}^* \subseteq \bar{\mathcal{U}}$ and

$$\eta(\mathcal{U} \setminus \mathcal{C}) \vee \eta(\bar{\mathcal{U}} \setminus \bar{\mathcal{C}}) \vee \eta(\bar{\mathcal{U}} \cap \mathcal{C}^*) \vee \eta(\mathcal{U} \cap (\mathcal{X} \setminus \mathcal{C}^*)) \leq 0.25\epsilon,$$

along with continuous (Urysohn) functions $\zeta_{\mathcal{C}^*} : \mathbb{R}^d \to [0,1]$, $\zeta_{\mathcal{X}\setminus\mathcal{C}^*} : \mathbb{R}^d \to [0,1]$ such that

$$\zeta_{\mathcal{C}^*}(x) = \begin{cases} 1, & x \in \mathcal{C}, \\ 0, & x \in \mathbb{R}^d \setminus \mathcal{U}, \end{cases}$$

$$\zeta_{\mathcal{X}\setminus\mathcal{C}^*}(x) = \begin{cases} 1, & x \in \bar{\mathcal{C}}, \\ 0, & x \in \mathbb{R}^d \setminus \bar{\mathcal{U}}. \end{cases}$$

Hence,

$$\mathbb{E}_\mu\left[|\mathbb{1}_{\mathcal{C}^*} - \zeta_{\mathcal{C}^*}|\right] \vee \mathbb{E}_\nu\left[|\mathbb{1}_{\mathcal{C}^*} - \zeta_{\mathcal{C}^*}|\right] \leq \mathbb{E}_\eta\left[(p \vee q)\,|\mathbb{1}_{\mathcal{C}^*} - \zeta_{\mathcal{C}^*}|\right] \leq 0.25\,\|p \vee q\|_{\infty,\eta}\,\epsilon, \tag{A.61}$$

$$\begin{aligned} \mathbb{E}_\mu\left[\left|\mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*} - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}\right|\right] \vee \mathbb{E}_\nu\left[\left|\mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*} - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}\right|\right] &\leq \mathbb{E}_\eta\left[(p \vee q)\left|\mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*} - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}\right|\right] \\ &\leq 0.25\,\|p \vee q\|_{\infty,\eta}\,\epsilon. \end{aligned} \tag{A.62}$$

Let $\zeta(x) = \zeta_{\mathcal{C}^*}(x) - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}(x)$. Note that $\zeta(x) \in [-1,1]$, $\zeta(x) = 1$ for $x \in \mathcal{C}\setminus\bar{\mathcal{U}}$ and $\zeta(x) = -1$ for $x \in \bar{\mathcal{C}} \setminus \mathcal{U}$. Since $\zeta(\cdot)$ is a continuous function, it follows from (Stinchcombe and White, 1990, Theorem 2.1 and 2.8) that for any $\epsilon > 0$ and $k \geq k_0(\epsilon)$, there exists a $\tilde{g} \in \mathcal{G}_k^*(\phi)$ such that $\|\zeta - \tilde{g}\|_{\infty,\mathcal{X}} \leq \epsilon$. Since $\|\zeta\|_\infty \leq 1$, it then follows from the definition of $\bar{\mathcal{G}}_k^*(\phi)$ that there exists $g^* \in \bar{\mathcal{G}}_k^*(\phi)$ such that

$$\|\zeta - g^*\|_{\infty,\mathcal{X}} \leq \epsilon. \tag{A.63}$$

Let $\tilde{\delta}_{\mathsf{TV}}(g) := \mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]$. Then, we have for $k \geq k_0(\epsilon)$ that

$$\begin{aligned} &\left|\delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k^*(\phi)}(\mu,\nu)\right| \\ &\quad = \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k^*(\phi)}(\mu,\nu) \\ &\quad \leq \delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}(g^*) \\ &\quad \leq \mathbb{E}_\mu\left[|f_{\mathsf{TV}} - g^*|\right] + \mathbb{E}_\nu\left[|f_{\mathsf{TV}} - g^*|\right] \\ &\quad \leq \mathbb{E}_\mu\left[|f_{\mathsf{TV}} - \zeta| + |\zeta - g^*|\right] + \mathbb{E}_\nu\left[|f_{\mathsf{TV}} - \zeta| + |\zeta - g^*|\right] \\ &\quad \leq \mathbb{E}_\mu\left[|\mathbb{1}_{\mathcal{C}^*} - \zeta_{\mathcal{C}^*}|\right] + \mathbb{E}_\nu\left[|\mathbb{1}_{\mathcal{C}^*} - \zeta_{\mathcal{C}^*}|\right] + \mathbb{E}_\mu\left[\left|\mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*} - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}\right|\right] + \mathbb{E}_\nu\left[\left|\mathbb{1}_{\mathcal{X}\setminus\mathcal{C}^*} - \zeta_{\mathcal{X}\setminus\mathcal{C}^*}\right|\right] \\ &\qquad + \mathbb{E}_\mu\left[|\zeta - g^*|\right] + \mathbb{E}_\nu\left[|\zeta - g^*|\right] \\ &\quad \leq \epsilon\big(\,\|p \vee q\|_{\infty,\eta} + 2\big) \leq 4\epsilon, \end{aligned} \tag{A.64}$$

where (A.64) follows from (A.61), (A.62), (A.63) and $\|p \vee q\|_{\infty,\eta} \leq 2$. Since $\epsilon > 0$ is arbitrary, we have from (A.64) that

$$\lim_{k\to\infty} \delta_{\bar{\mathcal{G}}_k^*(\phi)}(\mu,\nu) = \delta_{\mathsf{TV}}(\mu,\nu)\,.$$

Taking k_n, n satisfying $k_n = O\left(n^{(1-\rho)/2}\right)$, (4.14) follows from the above equation and (A.59).

Next, we prove (4.15). Fix $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{TV}}(M,\mathcal{X})$ such that $f_{\mathsf{TV}} \in \mathsf{Lip}_{s,1,M}(\mathcal{X})$. Since f_{TV} does not belong to the Klusowski-Barron class, we consider approximation of an intermediate function $f^{(t)}_{\mathsf{TV}}$, which is a smoothed version of f_{TV} and belongs to this class. The smoothing parameter t is then decreased as a function of k at an appropriate rate such that the L^1 error between $f^{(t)}_{\mathsf{TV}}$ and f_{TV} vanishes as $k \to \infty$. For this purpose, consider a non-negative smoothing kernel $\Phi \in L^1(\mathbb{R}^d)$, $\Phi \geq 0$, such that $\int_{\mathbb{R}^d}\Phi(x)\mathrm{d}x = 1$. Let $\Phi_t(x) := t^{-d}\Phi(t^{-1}x)$, $t > 0$, and

$$f^{(t)}_{\mathsf{TV}}(x) := f_{\mathsf{TV}} * \Phi_t(x) = \int_{\mathbb{R}^d} f_{\mathsf{TV}}(x-y)\Phi_t(y)\mathrm{d}y,$$

denote the smoothing of f_{TV} using Φ_t.

Recalling that $\tilde{\delta}_{\mathsf{TV}}(f) := \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]$, we have

$$\begin{aligned}\left|\delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(\mu,\nu)\right| &= \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(\mu,\nu)\\ &= \delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}\Big(f^{(t)}_{\mathsf{TV}}\Big) + \tilde{\delta}_{\mathsf{TV}}\Big(f^{(t)}_{\mathsf{TV}}\Big) - \delta_{\bar{\mathcal{G}}_k(\mathbf{a},\phi)}(\mu,\nu)\,, \qquad \text{(A.65)}\end{aligned}$$

The first term in (A.65) can be written as follows:

$$\delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}\Big(f^{(t)}_{\mathsf{TV}}\Big) = \mathbb{E}_\mu\left[f_{\mathsf{TV}} - f^{(t)}_{\mathsf{TV}}\right] - \mathbb{E}_\nu\left[f_{\mathsf{TV}} - f^{(t)}_{\mathsf{TV}}\right]. \qquad \text{(A.66)}$$

Denoting by p, q, the respective densities of μ, ν w.r.t. Lebesgue measure, we have

$$\begin{aligned}\mathbb{E}_\mu\left[f_{\mathsf{TV}} - f^{(t)}_{\mathsf{TV}}\right] &= \int_{\mathbb{R}^d}\left[f_{\mathsf{TV}}(x) - t^{-d}\int_{\mathbb{R}^d} f_{\mathsf{TV}}(y)\Phi\left((x-y)t^{-1}\right)\mathrm{d}y\right]p(x)\mathrm{d}x\\ &= \int_{\mathbb{R}^d}\left[f_{\mathsf{TV}}(x) - \int_{\mathbb{R}^d} f_{\mathsf{TV}}(x-tu)\Phi(u)\mathrm{d}u\right]p(x)\mathrm{d}x\\ &= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left[f_{\mathsf{TV}}(x)\Phi(u) - f_{\mathsf{TV}}(x-tu)\Phi(u)\right]\mathrm{d}u\right]p(x)\mathrm{d}x\\ &\leq \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV}}(x) - f_{\mathsf{TV}}(x-tu)\right|p(x)\mathrm{d}x\right]\Phi(u)\mathrm{d}u\\ &= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV}}(x+tu) - f_{\mathsf{TV}}(x)\right|p(x+tu)\mathrm{d}x\right]\Phi(u)\mathrm{d}u\\ &\leq \|p\|_{\infty,\mathcal{X}}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV}}(x+tu) - f_{\mathsf{TV}}(x)\right|\mathrm{d}x\right]\Phi(u)\mathrm{d}u\\ &\overset{(a)}{\leq} M\int_{\mathbb{R}^d}\xi_{1,1}\left(f_{\mathsf{TV}}, t\,\|u\|\right)\Phi(u)\mathrm{d}u\\ &\overset{(b)}{\leq} M^2\int_{\mathbb{R}^d} t^s\,\|u\|^s\,\Phi(u)\mathrm{d}u, \qquad \text{(A.67)}\end{aligned}$$

where (a) and (b) are due to $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{TV}}(M,\mathcal{X})$ and $f_{\mathsf{TV}}\in\mathsf{Lip}_{s,1,M}(\mathcal{X})$, respectively. Since (A.67) also holds for ν in place of μ, we have from (A.66) that

$$\left|\delta_{\mathsf{TV}}(\mu,\nu)-\tilde{\delta}_{\mathsf{TV}}\Big(f^{(t)}_{\mathsf{TV}}\Big)\right|\leq 2M^2\int_{\mathbb{R}^d}t^s\left\|u\right\|^s\Phi(u)\mathrm{d}u. \tag{A.68}$$

Next, note that

$$\Big\|f^{(t)}_{\mathsf{TV}}\Big\|_1\leq\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left|f_{\mathsf{TV}}(x-y)\Phi_t(y)\right|\mathrm{d}y\mathrm{d}x\overset{(a)}{\leq}\left\|f_{\mathsf{TV}}\right\|_1\left\|\Phi_t\right\|_1\overset{(b)}{\leq}\left\|f_{\mathsf{TV}}\right\|_1\overset{(c)}{<}\infty,$$

where

(a) follows from Minkowski's integral inequality;

(b) is due to $\int_{\mathbb{R}^d}\left|\Phi_t(y)\right|\mathrm{d}y=1$;

(c) is since $f_{\mathsf{TV}}\in L^1(\mathcal{X})$.

Hence, the Fourier transform of $f^{(t)}_{\mathsf{TV}}$ exists, and is given by

$$\mathfrak{F}\left[f^{(t)}_{\mathsf{TV}}\right]=\mathfrak{F}[f_{\mathsf{TV}}]\mathfrak{F}[\Phi_t]. \tag{A.69}$$

Choose Φ to be standard Gaussian kernel, i.e., $\Phi=\Phi^{\mathcal{N}}:=(2\pi)^{-d/2}e^{-0.5\|x\|^2}$. Then, we have

$$\left\|\mathfrak{F}\left[f^{(t)}_{\mathsf{TV}}\right]\right\|_1\overset{(a)}{\leq}\left\|f_{\mathsf{TV}}\right\|_1\int_{\mathbb{R}^d}\left|\mathfrak{F}[\Phi_t](\omega)\right|\mathrm{d}\omega\overset{(b)}{\leq}M\int_{\mathbb{R}^d}\left|\mathfrak{F}[\Phi](t\omega)\right|\mathrm{d}\omega\overset{(c)}{\leq}M\int_{\mathbb{R}^d}e^{-\frac{1}{2}t^2\|\omega\|^2}\mathrm{d}\omega<\infty,$$

where

(a) follows from (A.69) and $\left\|\mathfrak{F}\left[f_{\mathsf{TV}}\right]\right\|_\infty\leq\left\|f_{\mathsf{TV}}\right\|_1$;

(b) is via the formula $\mathfrak{F}\big[\Phi\left(t^{-1}\cdot\right)\big](\omega)=t^d\mathfrak{F}[\Phi]\left(t\omega\right)$, and $\left\|f_{\mathsf{TV}}\right\|_1\leq M$ by the definition of Lipschitz seminorm;

(c) is since $\mathfrak{F}\big[\Phi^{\mathcal{N}}\big](\omega)=e^{-\frac{1}{2}\|\omega\|^2}$.

Hence, the Fourier representation in Definition 4 holds via the Fourier inversion formula for $f^{(t)}_{\mathsf{TV}}$. Then, we can bound the spectral norm as

$$\begin{aligned}S_2\Big(f^{(t)}_{\mathsf{TV}}\Big)&:=\int_{\mathbb{R}^d}\|\omega\|_1^2\left|\mathfrak{F}\Big[f^{(t)}_{\mathsf{TV}}\Big](\omega)\right|d\omega\\&\leq\left\|f_{\mathsf{TV}}\right\|_1\int_{\mathbb{R}^d}\|\omega\|_1^2\left|\mathfrak{F}[\Phi_t](\omega)\right|\mathrm{d}\omega\\&\leq Md\int_{\mathbb{R}^d}\|\omega\|^2e^{-\frac{1}{2}t^2\|\omega\|^2}\mathrm{d}\omega.\end{aligned}$$

Evaluating the integral above by converting to hyperspherical coordinates, we obtain

$$\|\mathcal{X}\|\,S_2\Big(f^{(t)}_{\mathsf{TV}}\Big)\leq\|\mathcal{X}\|\,Md\int_{\mathbb{R}^d}\|\omega\|^2e^{-\frac{1}{2}t^2\|\omega\|^2}\mathrm{d}\omega=:c_{d,M,\|\mathcal{X}\|,t}, \tag{A.70}$$

where

$$c_{d,M,\|\mathcal{X}\|,t} := \begin{cases} (2\pi)^{\frac{1}{2}} \|\mathcal{X}\| M t^{-3}, & d=1, \\ 2^{\frac{d+3}{2}} \pi \|\mathcal{X}\| M d t^{-(d+2)} \Gamma((d+2)/2) \prod_{j=1}^{d-2} \int_0^{\pi} \sin^{d-1-j}(\varphi_j) \mathrm{d}\varphi_j, & d \geq 2. \end{cases}$$

Moreover, $\|f_{\mathsf{TV}}\|_\infty \leq 1$ and $\int_{\mathbb{R}^d} |\Phi_t(y)| \, \mathrm{d}y = 1$ implies

$$\left| f_{\mathsf{TV}}^{(t)}(x) \right| \leq \int_{\mathbb{R}^d} |f_{\mathsf{TV}}(x-y)\Phi_t(y)| \, \mathrm{d}y \leq \int_{\mathbb{R}^d} |\Phi_t(y)| \, \mathrm{d}y = 1. \tag{A.71}$$

Since $\left|f_{\mathsf{TV}}^{(t)}(0)\right| \vee \left\|\nabla f_{\mathsf{TV}}^{(t)}(0)\right\| \leq 1 \vee (2d\pi^{-d})^{1/2}\Gamma\big(0.5(d+1)\big)t^{-1}$ and (A.70) holds, there exists $g_{\theta_k} \in \bar{\mathcal{G}}_k^{\mathsf{R}}\left(\hat{c}_{d,M,\|\mathcal{X}\|,t}\right)$ such that for all $0 < t \leq 1$,

$$\left\| f_{\mathsf{TV}}^{(t)} - g_{\theta_k} \right\|_{\infty,\mathcal{X}} \lesssim \hat{c}_{d,M,\|\mathcal{X}\|,t} d^{\frac{1}{2}} k^{-\frac{1}{2}}, \tag{A.72}$$

where $\hat{c}_{d,M,\|\mathcal{X}\|,t} := c_{d,M,\|\mathcal{X}\|,t} \vee 1 \vee (2d\pi^{-d})^{1/2}\Gamma\big(0.5(d+1)\big)t^{-1}$. The existence of g_{θ_k} follows by truncating $g \in \mathcal{G}_k^{\mathsf{R}}\left(\hat{c}_{d,M,\|\mathcal{X}\|,t}\right)$ satisfying (3.1) to $[-1,1]$, and noting that truncation only decreases the approximation error as $\left\|f_{\mathsf{TV}}^{(t)}\right\|_\infty \leq 1$. Hence, we have

$$\begin{aligned} \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV}}^{(t)}\right) - \delta_{\bar{\mathcal{G}}_k^{\mathsf{R}}(\hat{c}_{d,M,\|\mathcal{X}\|,t})}(\mu,\nu) &\leq \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV}}^{(t)}\right) - \tilde{\delta}_{\mathsf{TV}}\big(g_{\theta_k}\big) \\ &\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{TV}}^{(t)} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|f_{\mathsf{TV}}^{(t)} - g_{\theta_k}\right|\right] && \text{(A.73)} \\ &\lesssim \hat{c}_{d,M,\|\mathcal{X}\|,t} d^{\frac{1}{2}} k^{-\frac{1}{2}}. && \text{(A.74)} \end{aligned}$$

Next, observe that (A.68) with $\Phi = \Phi^{\mathcal{N}}$ yields

$$\left| \delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV}}^{(t)}\right) \right| \leq c_{d,M,s} t^s,$$

where $c_{d,M,s} := 2M^2(2\pi)^{-d/2} \int_{\mathbb{R}^d} \|u\|^s e^{-0.5\|u\|^2} \mathrm{d}u$. From this, (A.65) and (A.74), we obtain

$$\left| \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k^{\mathsf{R}}(\hat{c}_{d,M,\|\mathcal{X}\|,t})}(\mu,\nu) \right| = \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k^{\mathsf{R}}(\hat{c}_{d,M,\|\mathcal{X}\|,t})}(\mu,\nu) \lesssim_{d,M,s} t^s + \|\mathcal{X}\| t^{-(d+2)} k^{-\frac{1}{2}}.$$

Setting $t = t_k^* := k^{-1/2(s+d+2)}$ and

$$\tilde{c}_{k,d,s,M,\|\mathcal{X}\|} := \hat{c}_{d,M,\|\mathcal{X}\|,t_k^*} = O_{d,M}\big(\|\mathcal{X}\| k^{(d+2)/2(s+d+2)} \big), \tag{A.75}$$

yields

$$\left| \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\bar{\mathcal{G}}_k^{\mathsf{R}}\left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\right)}(\mu,\nu) \right| \lesssim_{d,M,s} (\|\mathcal{X}\|+1) k^{-s/2(s+d+2)}. \tag{A.76}$$

Finally, we bound the expected empirical estimation error. Note that $\bar{C}\big(\big|\bar{\mathcal{G}}_k^{\mathsf{R}}(a)\big|, \mathcal{X}\big) \leq 1$ and $\bar{C}\big(\big|h'_{\mathsf{TV}} \circ \bar{\mathcal{G}}_k^{\mathsf{R}}(a)\big|, \mathcal{X}\big) = 1$. Then, we have

$$\mathbb{E}\left[\left| \hat{\delta}_{\bar{\mathcal{G}}_k^{\mathsf{R}}\left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\right)}(X^n, Y^n) - \delta_{\bar{\mathcal{G}}_k^{\mathsf{R}}\left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\right)}(\mu,\nu) \right|\right]$$

$$
\begin{aligned}
&\overset{(a)}{\lesssim} n^{-\frac{1}{2}} \int_0^1 \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(\epsilon, \bar{\mathcal{G}}_k^{\mathsf{R}}\big(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\big), \|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon \\
&\overset{(b)}{\leq} n^{-\frac{1}{2}} \int_0^1 \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(\epsilon, \mathcal{G}_k^{\mathsf{R}}\big(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\big), \|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon \\
&\overset{(c)}{\leq} n^{-\frac{1}{2}} \int_0^1 \sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})} \log N\left(\epsilon/3, \mathcal{G}_k^{\dagger}\big(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}\big), \|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon \\
&\qquad\qquad\qquad\qquad + n^{-\frac{1}{2}}(d+1)^{\frac{1}{2}} \int_0^1 \sqrt{\log(1+6\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}(\|\mathcal{X}\|+1))}\mathrm{d}\epsilon \\
&\overset{(d)}{\lesssim} n^{-\frac{1}{2}} d^{\frac{3}{2}} \left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}(\|\mathcal{X}\|+1)\right)^{\frac{d+2}{d+3}} + n^{-\frac{1}{2}} d^{\frac{1}{2}} \left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}(\|\mathcal{X}\|+1)\right)^{\frac{1}{2}}, \\
&\lesssim n^{-\frac{1}{2}} d^{\frac{3}{2}} \left(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|}(\|\mathcal{X}\|+1)+1\right),
\end{aligned} \tag{A.77}
$$

where (a) follows from (A.29); (b) is since the pointwise difference between functions in $\bar{\mathcal{G}}_k^{\mathsf{R}}(a)$ (with range $[-1,1]$) is less than the difference between the corresponding untruncated functions in $\mathcal{G}_k^{\mathsf{R}}(a)$; (c) is due to (A.32); and (d) is via steps leading to (A.33). Then, (A.76) and (A.77) implies that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\hat{\delta}_{\bar{\mathcal{G}}_k^{\mathsf{R}}(\tilde{c}_{k,d,s,M,\|\mathcal{X}\|})}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right] \\
&\qquad\qquad \lesssim_{d,M,s} (\|\mathcal{X}\|+1)k^{-s/2(s+d+2)} + n^{-\frac{1}{2}} k^{(d+2)/2(s+d+2)}\big(\|\mathcal{X}\|^2+1\big)^{\frac{1}{2}}.
\end{aligned} \tag{A.78}
$$

Recalling that $\mathcal{X}=[0,1]^d$ and taking supremum over $(\mu,\nu)\in\mathcal{P}_{\mathsf{TV}}^2(M,\mathcal{X})$ such that $f_{\mathsf{TV}}\in \mathsf{Lip}_{s,1,M}(\mathcal{X})$, we obtain (4.15).

A.2.9 Proof of Proposition 35

Since $p-q\in\mathcal{T}_{b,N}(\mathcal{X})$ and $f_{\mathsf{TV}}=\mathbb{1}_{\{p-q\geq 0\}}-\mathbb{1}_{\{p-q<0\}}$, the definition of $\xi_{1,1}(f_{\mathsf{TV}},t)$ yields

$$
\xi_{1,1}(f_{\mathsf{TV}},t) \leq \begin{cases} 2N\lambda(B_d(t)), & t\leq b \\ 2\left\|f_{\mathsf{TV}}\right\|_1, & \text{otherwise.}\end{cases}
$$

Hence, for any $0<s\leq 1$, it holds that

$$
\begin{aligned}
\|f_{\mathsf{TV}}\|_{\mathsf{Lip}(s,1)} &= \|f_{\mathsf{TV}}\|_1 + \sup_{t>0} t^{-s}\xi_{1,1}(f_{\mathsf{TV}},t) \\
&= \|f_{\mathsf{TV}}\|_1 + \sup_{0<t\leq b} t^{-s}\xi_{1,1}(f_{\mathsf{TV}},t) \vee \sup_{t>b} t^{-s}\xi_{1,1}(f_{\mathsf{TV}},t) \\
&= \|f_{\mathsf{TV}}\|_1 + \sup_{0<t\leq b} t^{-s}2N\lambda(B_d(t)) \vee \sup_{t>b} t^{-s}2\left\|f_{\mathsf{TV}}\right\|_1 \\
&= \lambda(\mathcal{X}) + 2N\pi^{\frac{d}{2}} b^{d-s}\left(\Gamma(0.5d+1)\right)^{-1} \vee 2b^{-s}\lambda(\mathcal{X}),
\end{aligned} \tag{A.79}
$$

where λ denotes the Lebesgue measure and Γ is the gamma function. Hence, $f_{\mathsf{TV}}\in \mathsf{Lip}_{s,1,M}(\mathcal{X})$ with $M=\lambda(\mathcal{X})+2N\big(\Gamma(0.5d+1)\big)^{-1}\pi^{\frac{d}{2}}b^{d-s}\vee 2b^{-s}\lambda(\mathcal{X})$ and any $0<s\leq 1$, thus proving the claim via Theorem 32.

A.3 Proofs for Section 5

A.3.1 Proof of Theorem 39

To prove part (i), fix some $0 < \epsilon < 1$. Let $B_d^c(r) = \mathbb{R}^d \setminus B_d(r)$, and $r(\epsilon)$ be sufficiently large such that $\mathbb{E}_\mu\big[\,|f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r(\epsilon))}\big] \vee \mathbb{E}_\nu\big[|\mathrm{d}\mu/\mathrm{d}\nu - 1|\mathbb{1}_{B_d^c(r(\epsilon))}\big] \leq \epsilon$. Since $f_{\mathsf{KL}} \in \mathsf{C}(\mathbb{R}^d)$, from (Stinchcombe and White, 1990, Theorem 2.1 and 2.8), there is a $k_0(\epsilon, r(\epsilon)) \in \mathbb{N}$, such that for any $k \geq k_0(\epsilon, r(\epsilon))$, there exists a $g_{\theta_k} \in \hat{\mathcal{G}}_k^*(\phi, r(\epsilon))$ with

$$\|f_{\mathsf{KL}} - g_{\theta_k}\|_{\infty, B_d(r(\epsilon))} \leq \epsilon. \tag{A.80}$$

Then, we have

$$\begin{aligned}
&\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\hat{\mathcal{G}}_k^*(\phi, r(\epsilon))}(\mu,\nu)\right| \\
&\leq \mathbb{E}_\mu\left[|f_{\mathsf{KL}} - g_{\theta_k}|\right] + \mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}} - e^{g_{\theta_k}}\right|\right] \\
&= \mathbb{E}_\mu\left[|f_{\mathsf{KL}} - g_{\theta_k}|\,\mathbb{1}_{B_d(r(\epsilon))}\right] + \mathbb{E}_\mu\left[|f_{\mathsf{KL}} - g_{\theta_k}|\,\mathbb{1}_{B_d^c(r(\epsilon))}\right] + \mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}} - e^{g_{\theta_k}}\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}} - e^{g_{\theta_k}}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] \\
&\leq \left\|(f_{\mathsf{KL}} - g_{\theta_k})\mathbb{1}_{B_d(r(\epsilon))}\right\|_{\infty,\mu} + \mathbb{E}_\mu\left[|f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r(\epsilon))}\right] + \mathbb{E}_\nu\left[\left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] \\
&\qquad\qquad + \mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}}\right|\mathbb{1}_{B_d(r(\epsilon))}\right]\left\|\left(1 - e^{g_{\theta_k} - f_{\mathsf{KL}}}\right)\mathbb{1}_{B_d(r(\epsilon))}\right\|_{\infty,\nu} \qquad (A.81)\\
&\lesssim \epsilon, \qquad (A.82)
\end{aligned}$$

where the final inequality is due to (A.80), the choice of $r(\epsilon)$, and $\mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] \leq 1$.

On the other hand, for any $0 < \rho < 1$, and n, k_n, r_n such that $k_n^{3/2}(r_n + 1)e^{k_n(r_n+1)} = O\left(n^{(1-\rho)/2}\right)$, Lemma 59 yields

$$\hat{\mathsf{D}}_{\hat{\mathcal{G}}_{k_n}^*(\phi, r_n)}(X^n, Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\hat{\mathcal{G}}_{k_n}^*(\phi, r_n)}(\mu, \nu), \quad \mathbb{P} - \text{a.s.}$$

This along with (A.82) completes the proof of Part (i).

To prove part (ii), we first state a general error bound for KL neural estimation based on the tail behaviour of random variables $f_{\mathsf{KL}}(X)$ and $h_{\mathsf{KL}} \circ f_{\mathsf{KL}}(Y) := e^{f_{\mathsf{KL}}(Y)} - 1$ outside $B_d(r)$ for $X \sim \mu$ and $Y \sim \nu$. For an increasing positive divergent sequence $\mathbf{r} = (r_k)_{k\in\mathbb{N}}$, $r_k \geq 1$, $(r_k \to \infty)$, a positive non-decreasing sequence $\mathbf{m} = (m_k)_{k\in\mathbb{N}}$, $m_k \geq 1$, and a non-increasing non-negative sequence $\mathbf{v} = (v_k)_{k\in\mathbb{N}}$ with $v_k \to 0$, set

$$\begin{aligned}
&\breve{\mathcal{P}}^2_{\mathsf{KL}}(M, \mathbf{r}, \mathbf{m}, \mathbf{v}) \\
&:= \left\{(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}\big(\mathbb{R}^d\big) : \begin{array}{l} \mathbb{E}_\mu\Big[\,|f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r_k)}\Big] \vee \mathbb{E}_\nu\left[|h_{\mathsf{KL}} \circ f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r_k)}\right] \leq v_k, \\ \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) \leq M,\ c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k,\ k \in \mathbb{N}\end{array}\right\}.
\end{aligned}$$

Then, we have the following lemma.

Lemma 63 (KL divergence neural estimation) *Suppose there exists $M \geq 0$, $0 < \rho < 1$, and $\mathbf{r}, \mathbf{m}, \mathbf{v}$ as above satisfying $1 \leq m_k \lesssim k^{(1-\rho)/2}$, such that $(\mu, \nu) \in \breve{\mathcal{P}}^2_{\mathsf{KL}}(M, \mathbf{r}, \mathbf{m}, \mathbf{v})$. Then, for $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$, we have*

$$\sup_{\substack{(\mu,\nu)\in \\ \breve{\mathcal{P}}^2_{\mathsf{KL}}(M,\mathbf{r},\mathbf{m},\mathbf{v})}} \mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n, Y^n) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right] \lesssim_{d,M,\rho} m_k k^{-\frac{1}{2}} + v_k + m_k r_k e^{3m_k(r_k+1)}\ n^{-\frac{1}{2}}.$$

The proof of the above lemma is based on an application of Theorem 8 to bound the NN approximation error on balls $B_d(r_k)$, leveraging tail integrability assumptions in the definition of $\breve{\mathcal{P}}^2_{\mathsf{KL}}$ to bound the approximation error outside $B_d(r_k)$, and using Theorem 12 to control the empirical estimation error. Its proof is given in Appendix B.7.

Continuing with the proof of the Theorem, we will show that $(\mu, \nu) \in \bar{\mathcal{P}}^2_{\mathsf{KL},\psi}(M, \ell, \mathbf{r}, \mathbf{m})$ implies $(\mu, \nu) \in \breve{\mathcal{P}}^2_{\mathsf{KL}}(M, \mathbf{r}, \mathbf{m}, \mathbf{v})$ for some $\mathbf{v}$ that will be specified below. Then, Part (ii) will follow from Lemma 63.

Note that $\|f_{\mathsf{KL}}\|_{\ell,\mu} \leq M$, where $\ell > 1$ (or equivalently $\ell \geq 2$ since $\ell \in \mathbb{N}$), implies

$$\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) = \mathbb{E}_\mu\left[f_{\mathsf{KL}}\right] \leq \sqrt{\mathbb{E}_\mu\left[f^2_{\mathsf{KL}}\right]} \leq M. \tag{A.83}$$

Also,

$$\begin{aligned} \mathbb{E}_\mu\left[\left|f_{\mathsf{KL}}\right| \mathbb{1}_{B^c_d(r_k)}\right] &\overset{(a)}{\leq} \|f_{\mathsf{KL}}\|_{\ell,\mu}\left(\mu\left(\|X\| > r_k\right)\right)^{\frac{1}{\ell^*}} \\ &\overset{(b)}{\leq} M\mu\Big(\psi\big(\|X\|\, M^{-1}\big) > \psi\big(r_k M^{-1}\big)\Big)^{\frac{1}{\ell^*}} \\ &\overset{(c)}{\leq} M\Big(\mathbb{E}_\mu\left[\psi\big(\|X\|\, M^{-1}\big)\right]\Big)^{\frac{1}{\ell^*}}\Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{\ell^*}} \\ &\overset{(d)}{\leq} M\Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{\ell^*}}, \end{aligned} \tag{A.84}$$

$$\begin{aligned} \mathbb{E}_\nu\left[\left|h_{\mathsf{KL}} \circ f_{\mathsf{KL}}\right| \mathbb{1}_{B^c_d(r_k)}\right] &= \mathbb{E}_\nu\left[\left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right| \mathbb{1}_{B^c_d(r_k)}\right] \\ &\leq \mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu} + 1\right) \mathbb{1}_{B^c_d(r_k)}\right] \\ &= \mu\big(B^c_d(r_k)\big) + \nu\big(B^c_d(r_k)\big) \\ &\overset{(e)}{\leq} \Big(\mathbb{E}_\mu\left[\psi\big(\|X\|\, M^{-1}\big)\right] + \mathbb{E}_\nu\left[\psi\big(\|Y\|\, M^{-1}\big)\right]\Big)\Big(\psi\big(r_k M^{-1}\big)\Big)^{-1} \\ &\overset{(f)}{\leq} 2\Big(\psi\big(r_k M^{-1}\big)\Big)^{-1}, \end{aligned} \tag{A.85}$$

where

(a) follows by Hölder's inequality;

(b) is since $\|f_{\mathsf{KL}}\|_{\ell,\mu} \leq M$ and ψ is increasing;

(c) and (e) are due to Markov's inequality;

(d) and (f) are since $p,q\in L_\psi(M)$ implies that $\mathbb{E}_\mu\left[\psi\big(\|X\|\,M^{-1}\big)\right]\vee\mathbb{E}_\nu\left[\psi\big(\|Y\|\,M^{-1}\big)\right]\leq 1$.

It follows that $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\mathsf{KL}}(M,\mathbf{r},\mathbf{m},\mathbf{v})$ with $v_k\asymp_{M,\psi,\ell}\left(\psi(r_kM^{-1})\right)^{-1/\ell^*}$ since $r_k\geq 1$. Note that $v_k\to 0$ as $r_k\to\infty$. This completes the proof of Part (ii) via Lemma 63.

A.3.2 Proof of Corollary 41

Fix $(\mu,\nu)=\big(\mathcal{N}(\mathsf{m}_p,\sigma_p^2\mathrm{I}_d),\mathcal{N}(\mathsf{m}_q,\sigma_q^2\mathrm{I}_d)\big)\in\bar{\mathcal{P}}^2_{\mathsf{N}}(M)$ and $\mathbf{r}=(r_k)_{k\in\mathbb{N}}=\big(1\vee M+\tilde{r}_k\big)_{k\in\mathbb{N}}$, where $\tilde{r}_k\geq 0$, $k\in\mathbb{N}$, will be specified below. Note that

$$\begin{aligned}
f_{\mathsf{KL}}(x)&=d\log\left(\frac{\sigma_q}{\sigma_p}\right)+\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2},\\
\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)&=d\log\left(\frac{\sigma_q}{\sigma_p}\right)-0.5d+0.5d\frac{\sigma_p^2}{\sigma_q^2}+\frac{\|\mathsf{m}_p-\mathsf{m}_q\|^2}{2\sigma_q^2}.
\end{aligned}$$

Also, f_{KL} is infinitely differentiable on $\mathbb{R}^d$, and it can be seen by computing derivatives that for any multi-index α of dimension d and arbitrary order $\|\alpha\|_1\in\mathbb{N}$,

$$\|D^\alpha f_{\mathsf{KL}}\|_{\infty,B_d(r_k)}\leq b^*_{k,d,M}:=c_{d,M}\left(1+\tilde{r}_k^2\right),$$

for some constant $c_{d,M}$ (polynomial in M). Hence, $f_{\mathsf{KL}}|_{B_d(r_k)}\in\mathsf{C}^{s_{\mathsf{KB}}}_{b^*_{k,d,M}}$, which implies via Proposition 10 that

$$c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)},B_d(r_k)\right)\leq m_k^{\mathsf{KL}}:=c_{d,M}\left(1+\tilde{r}_k^{d+3}\right).$$

By a straightforward calculation by using $1/M<\sigma_p,\sigma_q<M,\|\mathsf{m}_p\|\vee\|\mathsf{m}_q\|\leq M$, it follows from Gaussian integral formulas that there exists some $c_{d,M}$ such that

$$\|f_{\mathsf{KL}}\|_{2,\mu}\vee\|p\|_{\psi_2}\vee\|q\|_{\psi_2}\leq c_{d,M},$$

where $\psi_2(z)=e^{z^2}-1$. Hence, $\bar{\mathcal{P}}^2_{\mathsf{N}}(M)\subseteq\bar{\mathcal{P}}^2_{\mathsf{KL},\psi_2}\left(c_{d,M},2,\mathbf{r},\mathbf{m}^{\mathsf{KL}}\right)$, and we have from Part (ii) of Theorem 39 with $m_k=m_k^{\mathsf{KL}}$ and $\mathcal{G}_k=\hat{\mathcal{G}}_k^{\mathsf{R}}\left(m_k^{\mathsf{KL}},r_k\right)$ that

$$\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n,Y^n)-\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]\lesssim_{d,M,\rho}m_kk^{-\frac{1}{2}}+e^{-\frac{\tilde{r}_k^2}{c^2_{d,M}}}+m_k\tilde{r}_ke^{3m_k\tilde{r}_k}\,n^{-\frac{1}{2}}.$$

Then, setting $r_k=r_k^{\mathsf{KL}}:=1\vee M+\tilde{r}_k$ with $\tilde{r}_k=(c\log k/3c_{d,M})^{1/(d+4)}$ yields for $\mathcal{G}_k=\hat{\mathcal{G}}_k^{\mathsf{R}}\left(m_k^{\mathsf{KL}},r_k^{\mathsf{KL}}\right)$ that

$$\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\mathcal{G}_k}(X^n,Y^n)-\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]\lesssim_{d,M}ck^{-\frac{1}{2}}\log k+e^{-\left(c\log k/c_{d,M}\right)^{\frac{2}{d+4}}}+ck^c\log k\;n^{-\frac{1}{2}}.$$

Solving for the value of c such that the first two terms in the RHS of the equation above are equal (up to logarithmic factors) yields $c=c_{d,M}2^{-(d+4)/2}(\log k)^{(d+2)/2}$. Substituting c and taking supremum over $(\mu,\nu)\in\bar{\mathcal{P}}^2_{\mathsf{N}}(M)$, we obtain the claim in the Corollary.

A.3.3 Proof of Theorem 43

Let $r(\epsilon)$ be sufficiently large such that $\mathbb{E}_\mu\big[|f_{\chi^2}|\mathbb{1}_{B_d^c(r(\epsilon))}\big] \vee \mathbb{E}_\nu\big[|h_{\chi^2} \circ f_{\chi^2}|\mathbb{1}_{B_d^c(r(\epsilon))}\big] \leq \epsilon$. Similar to (A.80), there exists $g_{\theta_k} \in \hat{\mathcal{G}}_k^*(\phi, r(\epsilon))$ satisfying $\left\|f_{\chi^2} - g_{\theta_k}\right\|_{\infty, B_d(r(\epsilon))} \leq \epsilon$ for $k \geq k_0(\epsilon, r(\epsilon))$. Then, we have

$$
\begin{aligned}
&\left|\chi^2\left(\mu\|\nu\right) - \chi^2_{\hat{\mathcal{G}}^*_{k_n}(\phi, r_n)}(\mu,\nu)\right| \\
&\leq \mathbb{E}_\mu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\right] \\
&= \mathbb{E}_\mu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] \\
&\qquad + \mathbb{E}_\mu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] \\
&\leq \left\|\left(f_{\chi^2} - g_{\theta_k}\right)\mathbb{1}_{B_d(r(\epsilon))}\right\|_{\infty,\mu} + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] \\
&\qquad + \mathbb{E}_\mu\left[\left|f_{\chi^2}\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2}\right|\mathbb{1}_{B_d^c(r(\epsilon))}\right] \qquad (A.86) \\
&\overset{(a)}{\lesssim} \epsilon + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] \\
&\overset{(b)}{\lesssim} \epsilon + \mathbb{E}_\nu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\mathbb{1}_{B_d(r(\epsilon))}\right] + \mathbb{E}_\nu\Big[0.25\left|f_{\chi^2} - g_{\theta_k}\right|^2\mathbb{1}_{B_d(r(\epsilon))}\Big] \\
&\qquad + 0.5\mathbb{E}_\nu\Big[\left|f_{\chi^2} - g_{\theta_k}\right|\left|f_{\chi^2}\right|\mathbb{1}_{B_d(r(\epsilon))}\Big] \\
&\lesssim \epsilon + \left\|\left(f_{\chi^2} - g_{\theta_k}\right)\mathbb{1}_{B_d(r(\epsilon))}\right\|_{\infty,\nu}\mathbb{E}_\nu\left[\left|f_{\chi^2}\right|\right] \\
&\overset{(c)}{\lesssim} \epsilon,
\end{aligned}
$$

where (a) follows by definition of $r(\epsilon)$ and g_{θ_k} above; (b) is via steps leading to (A.48); (c) is due to definition of $r(\epsilon)$, g_{θ_k} and $\mathbb{E}_\nu\big[\left|f_{\chi^2}\right|\big] \leq 4$. From this and Lemma 60, Part (i) follows.

Next, we prove Part (ii). For sequences $\mathbf{m}$, $\mathbf{r}$ and $\mathbf{v}$ as in Appendix A.3.1, let

$$
\breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v}) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\chi^2}\left(\mathbb{R}^d\right) : \begin{array}{c} \mathbb{E}_\mu\Big[\left|f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\Big] \vee \mathbb{E}_\nu\Big[\left|h_{\chi^2}\circ f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\Big] \leq v_k \\ c^\star_{\mathsf{KB}}\left(f_{\chi^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k, \quad k \in \mathbb{N} \end{array}\right\}.
$$

We will use the following lemma which bounds the χ^2 neural estimation error for distributions satisfying general tail integrability conditions (see Appendix B.8 for proof).

Lemma 64 (χ^2 neural estimation error) *For $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k(m_k, r_k)$, we have*

$$
\sup_{(\mu,\nu)\in\breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v})} \mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n,Y^n) - \chi^2\left(\mu\|\nu\right)\right|\right] \lesssim m_k d^{\frac{1}{2}} k^{-\frac{1}{2}} + m_k^2 d k^{-1} + v_k + d^{\frac{3}{2}} m_k^2 r_k^2 n^{-\frac{1}{2}}. \tag{A.87}
$$

Armed with Lemma 64, we next show that $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\chi^2,\psi}(M,\ell,\mathbf{r},\mathbf{m})$ implies that $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ for some $\mathbf{v}$ that will be identified below. We have

$$
\mathbb{E}_\mu\left[\left|f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\right] \overset{(a)}{\leq} \left\|f_{\chi^2}\right\|_{\ell,\mu}\left(\mu\left(\|X\| > r_k\right)\right)^{\frac{1}{\ell^*}} \overset{(b)}{\leq} M\Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{\ell^*}}, \tag{A.88}
$$

$$\begin{aligned}
\mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\right] &= \mathbb{E}_\nu\left[2\left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\right|\mathbb{1}_{B_d^c(r_k)}\right]+\mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1\right)^2\mathbb{1}_{B_d^c(r_k)}\right]\\
&\leq 2\mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}+1\right)\mathbb{1}_{B_d^c(r_k)}\right]+\mathbb{E}_\nu\left[\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)^2\mathbb{1}_{B_d^c(r_k)}\right]+\nu\big(B_d^c(r_k)\big)\\
&= 2\mu\big(B_d^c(r_k)\big)+3\nu\big(B_d^c(r_k)\big)+\mathbb{E}_\mu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathbb{1}_{B_d^c(r_k)}\right]\\
&= 2\mu\big(B_d^c(r_k)\big)+3\nu\big(B_d^c(r_k)\big)+\mathbb{E}_\mu\left[(0.5f_{\chi^2}+1)\mathbb{1}_{B_d^c(r_k)}\right]\\
&\overset{(c)}{=} 3\mu\big(B_d^c(r_k)\big)+3\nu\big(B_d^c(r_k)\big)+0.5\left\|f_{\chi^2}\right\|_{\ell,\mu}\left(\mu\big(B_d^c(r_k)\big)\right)^{\frac{1}{\ell^*}} \qquad \text{(A.89)}\\
&\overset{(d)}{\leq} 6\Big(\psi\big(r_kM^{-1}\big)\Big)^{-1}+0.5M\Big(\psi\big(r_kM^{-1}\big)\Big)^{-\frac{1}{\ell^*}},
\end{aligned}$$

where

(a) and (c) is by Hölder's inequality;

(b) and (d) follows via Markov's inequality since $\mathbb{E}_\mu\left[\psi\big(\|X\|M^{-1}\big)\right]\vee\mathbb{E}_\nu\left[\psi\big(\|Y\|M^{-1}\big)\right]\leq 1$, and $\left\|f_{\chi^2}\right\|_{\ell,\mu}\leq M$ by assumption.

Hence, $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ with

$$v_k=6\Big(\psi\big(r_kM^{-1}\big)\Big)^{-1}+M\Big(\psi\big(r_kM^{-1}\big)\Big)^{-\frac{1}{\ell^*}}\lesssim_{M,\psi,\ell}\Big(\psi\big(r_kM^{-1}\big)\Big)^{-\frac{1}{\ell^*}}.$$

This implies Part (ii) via Lemma 64 (since $m_kk^{-\frac{1}{2}}+m_k^2k^{-1}\leq 2m_k^2k^{-\frac{1}{2}}$ due to $m_k\geq 1$).

A.3.4 Proof of Corollary 45

We will require the following lemma which bounds the tail probability of an isotropic Gaussian distribution outside a Euclidean ball $B_d(r)$ of radius r. This is a straighforward consequence of Gaussian concentration (Ledoux and Talagrand, 1991, Eqn. 1.4) and the fact that $\|\cdot\|$ is 1-Lipschitz function on the metric space $(\mathbb{R}^d,\|\cdot\|)$.

Lemma 65 (Gaussian tail integral bound) *For any $\mathsf{m}_p\in\mathbb{R}^d$ such that $\|\mathsf{m}_p\|\leq M$, $\sigma^2>0$ and $r\geq M$,*

$$\big(2\pi\sigma^2\big)^{-\frac{d}{2}}\int_{B_d^c(r)}e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma^2}}\mathrm{d}x\leq 2e^{-\frac{(r-M)^2}{2\sigma^2}}. \qquad \text{(A.90)}$$

Proceeding with the proof of the corollary, fix $(\mu,\nu)=\big(\mathcal{N}(\mathsf{m}_p,\sigma^2\mathrm{I}_d),\mathcal{N}(\mathsf{m}_q,\sigma^2\mathrm{I}_d)\big)\in\bar{\mathcal{P}}^2_{\chi^2,\mathsf{N}}(M)$, and $\mathbf{r}=\big(r_k\big)_{k\in\mathbb{N}}=\big(1\vee M+\tilde{r}_k\big)_{k\in\mathbb{N}}$, where $\tilde{r}_k\geq 0$, $k\in\mathbb{N}$, will be specified below. Note that since

$$f_{\chi^2}(x)=2\left(\frac{p(x)}{q(x)}-1\right)=2\left(\left(\frac{\sigma_q}{\sigma_p}\right)^de^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}}-1\right),$$

it is infinitely differentiable on $\mathbb{R}^d$. A straightforward computation shows that for any multi-index $\alpha \in \mathbb{Z}^d_{\geq 0}$ of order $\|\alpha\|_1 \leq s_{\mathsf{KB}}$,

$$\left\|D^\alpha f_{\chi^2}\right\|_{\infty, B_d(r_k)} \leq \tilde{b}^{\mathsf{R}}_k := c_{d,M}\left(1+\tilde{r}_k^{s_{\mathsf{KB}}}\right) e^{2M^2\tilde{r}_k^2}.$$

Hence, $f_{\chi^2}|_{B_d(r_k)} \in \mathsf{C}^{s_{\mathsf{KB}}}_{\tilde{b}^{\mathsf{R}}_k}$, which implies via Proposition 10 that

$$c^\star_{\mathsf{KB}}\left(f_{\chi^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k^{\chi^2} := c_{d,M}\left(1+\tilde{r}_k^{s_{\mathsf{KB}}+d+1}\right) e^{2M^2\tilde{r}_k^2}. \tag{A.91}$$

Furthermore, letting $\tilde{\sigma}^{-2} := \left(\sigma_p^{-2} - 0.5\sigma_q^{-2}\right) \wedge 0.5\sigma_q^{-2} \wedge 0.5\sigma_p^{-2}$ and noting that $\tilde{\sigma}^{-2} \geq 0.5M^{-3}$ by definition of $\bar{\mathcal{P}}^2_{\chi^2,\mathsf{N}}(M)$ and $M > 1$, we have

$$\begin{aligned}
\mathbb{E}_\mu\left[\left|f_{\chi^2}\right| \mathbb{1}_{B^c_d(r_k)}\right] &\leq \frac{2}{(2\pi\sigma_p^2)^{d/2}} \int_{B^c_d(r_k)} \left(\left(\frac{\sigma_q}{\sigma_p}\right)^d e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} + 1\right) e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} \mathrm{d}x \\
&\leq \frac{2}{(2\pi\sigma_p^2)^{d/2}} \int_{B^c_d(r_k)} \left(\frac{\sigma_q}{\sigma_p}\right)^d e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{\sigma_p^2}} \mathrm{d}x + e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} \mathrm{d}x \\
&\overset{(a)}{\lesssim_{d,M}} e^{-\frac{\tilde{r}_k^2}{\tilde{\sigma}^2}} \leq e^{-\frac{\tilde{r}_k^2}{2M^3}},
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_\nu\left[\left|h_{\chi^2} \circ f_{\chi^2}\right| \mathbb{1}_{B^c_d(r_k)}\right] \\
&= \frac{1}{(2\pi\sigma_q^2)^{d/2}}\left(\int_{B^c_d(r_k)} 2\left(\left(\frac{\sigma_q}{\sigma_p}\right)^d e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} - 1\right) e^{-\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}} \mathrm{d}x\right. \\
&\qquad\qquad \left. + \int_{B^c_d(r_k)} \left(\left(\frac{\sigma_q}{\sigma_p}\right)^d e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} - 1\right)^2 e^{-\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}} \mathrm{d}x\right) \\
&\overset{(b)}{\lesssim_{d,M}} \int_{B^c_d(r_k)} e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} \mathrm{d}x + \int_{B^c_d(r_k)} e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{\sigma_p^2}} \mathrm{d}x + \int_{B^c_d(r_k)} e^{-\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}} \mathrm{d}x \\
&\overset{(c)}{\lesssim_{d,M}} e^{-\frac{\tilde{r}_k^2}{\tilde{\sigma}^2}} \leq e^{-\frac{\tilde{r}_k^2}{2M^3}},
\end{aligned}$$

where

(a) and (c) follows by an application of Lemma 65 via completion of squares since $\sigma_p^2 < 2\sigma_q^2$ by assumption;

(b) uses $(ae^x - 1)^2 \leq a^2 e^{2x} + 1$ for $x \in \mathbb{R}^d$ and $a \geq 0$.

Hence, $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\chi^2}\left(\mathbf{r}, \mathbf{m}^{\chi^2}, \mathbf{v}^{\chi^2}\right)$ with $\mathbf{m}^{\chi^2}$ as defined in (A.91) and $v_k^{\chi^2} := c_{d,M} e^{-\tilde{r}_k^2/2M^3}$, and the error bound in (A.87) applies. Setting $r_k = r_k^{\chi^2} := 1 \vee M + \tilde{r}_k = 1 \vee M + 2^{-0.5}M^{-1}\sqrt{c\log k}$ for some constant c in (A.87), optimizing the resulting bound w.r.t. c (achieved at $c = 2M^5/(4M^5+1) < 0.5$), we obtain for $\mathcal{G}_k = \hat{\mathcal{G}}^{\mathsf{R}}_k\left(m_k^{\chi^2}, r_k^{\chi^2}\right)$ that

$$\mathbb{E}\left[\left|\hat{\chi}^2_{\mathcal{G}_k}(X^n, Y^n) - \chi^2\left(\mu\|\nu\right)\right|\right] \lesssim_{d,M} (\log k)^{2(s_{\mathsf{KB}}+d+1)}\left(k^{-\frac{1}{2+8M^5}} + k^{\frac{4M^5}{1+4M^5}} n^{-\frac{1}{2}}\right).$$

Taking supremum over $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\chi^2,\mathsf{N}}(M)$ completes the proof.

A.3.5 Proof of Theorem 47

For sequences $\mathbf{m}$, $\mathbf{r}$ and $\mathbf{v}$ as in Appendix A.3.1, let

$$\breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v}) := \left\{ (\mu,\nu) \in \mathcal{P}^2_{\mathsf{H}^2}\big(\mathbb{R}^d\big) : \begin{array}{l} \mathbb{E}_\mu\left[|f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] \vee \mathbb{E}_\nu\left[|h_{\mathsf{H}^2}\circ f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] \leq v_k, \\ c^\star_{\mathsf{KB}}\left(f_{\mathsf{H}^2}|_{B_d(r_k)}, B_d(r_k)\right) \vee \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty, B_d(r_k)} \leq m_k,\ k\in\mathbb{N} \end{array} \right\}.$$

The following lemma proves consistency of the NE for H^2 estimation and bounds its effective error for distributions satisfying general tail integrability conditions; see Appendix B.9 for proof.

Lemma 66 (H^2 neural estimation) *Let $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$, where $\mathbf{m}$ satisfies $m_k = o(k^{1/4})$. Then, the following hold:*

(i) For k_n, m_{k_n}, r_{k_n}, n satisfying $k_n \to \infty$, $r_{k_n} \to \infty$, $k_n^{1/2} m_{k_n}^2 r_{k_n} = O\left(n^{(1-\rho)/2}\right)$, and $\mathcal{G}_n = \breve{\mathcal{G}}^{\mathsf{R}}_{k_n, m_{k_n}^{-1/2}}(m_{k_n}, r_{k_n})$, we have

$$\hat{\mathsf{H}}^2_{\mathcal{G}_n}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{H}^2(\mu,\nu), \quad \mathbb{P}-a.s. \tag{A.92}$$

(ii) For $\mathcal{G}_k = \breve{\mathcal{G}}^{\mathsf{R}}_{k, m_k^{-1/2}}(m_k, r_k)$, we have

$$\sup_{(\mu,\nu)\in\breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})} \mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{H}^2(\mu,\nu)\right|\right] \lesssim m_k^2 d^{\frac{1}{2}} k^{-\frac{1}{2}} + v_k + d^{\frac{3}{2}} m_k^2 r_k n^{-\frac{1}{2}}. \tag{A.93}$$

To prove the theorem, we will show that $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\mathsf{H}^2,\psi}(M,\mathbf{r},\mathbf{m})$ implies that $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ for some $\mathbf{v}$ stated below. Then, Part (i) and (ii) will follow from the corresponding Parts in the above lemma. We have

$$\begin{aligned} \mathbb{E}_\mu\left[|f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] &= \mathbb{E}_\mu\left[\left|1-\sqrt{qp^{-1}}\right|\mathbb{1}_{B_d^c(r_k)}\right] \\ &\leq \mu\big(B_d^c(r_k)\big) + \mathbb{E}_\mu\left[\sqrt{qp^{-1}}\,\mathbb{1}_{B_d^c(r_k)}\right] \\ &\overset{(a)}{\leq} \mu\big(B_d^c(r_k)\big) + \sqrt{\nu\big(B_d^c(r_k)\big)} \\ &\overset{(b)}{\leq} \Big(\psi\big(r_k M^{-1}\big)\Big)^{-1} + \Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{2}}, \end{aligned} \tag{A.94}$$

$$\begin{aligned} \mathbb{E}_\nu\left[|h_{\mathsf{H}^2}\circ f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] &= \mathbb{E}_\nu\left[\left|\sqrt{pq^{-1}}-1\right|\mathbb{1}_{B_d^c(r_k)}\right] \\ &\overset{(c)}{\leq} \nu\big(B_d^c(r_k)\big) + \sqrt{\mu\big(B_d^c(r_k)\big)} \\ &\overset{(d)}{\leq} \Big(\psi\big(r_k M^{-1}\big)\Big)^{-1} + \Big(\psi\big(r_k M^{-1}\big)\Big)^{-\frac{1}{2}}, \end{aligned} \tag{A.95}$$

where

(a) and (c) follows from Cauchy-Schwarz inequality and $\mathbb{E}_\mu\left[qp^{-1}\right] = \mathbb{E}_\nu\left[pq^{-1}\right] = 1$;

(b) and (d) follows from Markov's inequality as $\mathbb{E}_\mu\left[\psi\big(\|X\|\,M^{-1}\big)\right] \vee \mathbb{E}_\nu\left[\psi\big(\|Y\|\,M^{-1}\big)\right] \leq 1$.

Hence, $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ with $v_k = \big(\psi\big(r_k M^{-1}\big)\big)^{-1} + \big(\psi\big(r_k M^{-1}\big)\big)^{-1/2} \lesssim_{\psi,M} \big(\psi\big(r_k M^{-1}\big)\big)^{-1/2} \to 0$. This completes the proof via Lemma 66.

A.3.6 Proof of Corollary 49

Fix $(\mu,\nu) = \big(\mathcal{N}(\mathsf{m}_p,\sigma^2 \mathrm{I}_d), \mathcal{N}(\mathsf{m}_q,\sigma^2 \mathrm{I}_d)\big) \in \bar{\mathcal{P}}^2_{\mathsf{N}}(M)$, and $\mathbf{r} = (r_k)_{k\in\mathbb{N}} = \big(1 \vee M + \tilde{r}_k\big)_{k\in\mathbb{N}}$, where $\tilde{r}_k \geq 0$, $k\in\mathbb{N}$, will be specified below. Observe that

$$f_{\mathsf{H}^2}(x) = 1 - \left(\frac{p(x)}{q(x)}\right)^{-\frac{1}{2}} = 1 - \left(\frac{\sigma_p}{\sigma_q}\right)^{d/2} e^{\frac{\|x-\mathsf{m}_p\|^2}{4\sigma_p^2} - \frac{\|x-\mathsf{m}_q\|^2}{4\sigma_q^2}},$$

is infinitely differentiable on $\mathbb{R}^d$. Then, for any multi-index $\alpha \in \mathbb{Z}^d_{\geq 0}$ of order $\|\alpha\|_1 \leq s_{\mathsf{KB}}$, it is easy to see by computing partial derivatives that

$$\|D^\alpha f_{\mathsf{H}^2}\|_{\infty,B_d(r_k)} \leq \hat{b}^{\mathsf{R}}_k := c_{d,M}\big(1+\tilde{r}_k^{s_{\mathsf{KB}}}\big)\, e^{M^2\tilde{r}_k^2}.$$

Hence, $f_{\mathsf{H}^2}|_{B_d(r_k)} \in \mathsf{C}^{s_{\mathsf{KB}}}_{\hat{b}^{\mathsf{R}}_k}$, which yields via Proposition 10 that

$$c^\star_{\mathsf{KB}}\left(f_{\mathsf{H}^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq c_{d,M}\left(1+\tilde{r}_k^{s_{\mathsf{KB}}+d+1}\right) e^{M^2\tilde{r}_k^2}.$$

Also, we have

$$\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty,B_d(r_k)} = \sup_{x\in B_d(r_k)} \left(\frac{\sigma_q}{\sigma_p}\right)^d e^{\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} \leq c_{d,M}\left(1+e^{2M^2 r_k^2}\right).$$

Furthermore, defining $\hat{\sigma}^2 := 4\sigma_p^2\sigma_q^2/(\sigma_p^2+\sigma_q^2) \vee 2\sigma_p^2 \vee 2\sigma_q^2 = 2\sigma_p^2 \vee 2\sigma_q^2 \geq 2M^{-1}$, we obtain

$$\begin{aligned}
\mathbb{E}_\mu\left[|f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] &\leq \frac{1}{(2\pi\sigma_p^2)^{d/2}} \int_{B_d^c(r_k)} \left(1 + \left(\frac{\sigma_p}{\sigma_q}\right)^{d/4} e^{\frac{\|x-\mathsf{m}_p\|^2}{4\sigma_p^2} - \frac{\|x-\mathsf{m}_q\|^2}{4\sigma_q^2}}\right) e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}}\,\mathrm{d}x \\
&\leq \frac{1}{(2\pi\sigma_p^2)^{d/2}} \int_{B_d^c(r_k)} \left(e^{-\frac{\|x-\mathsf{m}_p\|^2}{2\sigma_p^2}} + \left(\frac{\sigma_p}{\sigma_q}\right)^{d/4} e^{-\frac{\|x-\mathsf{m}_q\|^2}{4\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{4\sigma_p^2}}\right) \mathrm{d}x \\
&\overset{(a)}{\lesssim_{d,M}} e^{-\frac{\tilde{r}_k^2}{\hat{\sigma}^2}} \leq e^{-0.5M\tilde{r}_k^2},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_\nu\left[|h_{\mathsf{H}^2}\circ f_{\mathsf{H}^2}|\,\mathbb{1}_{B_d^c(r_k)}\right] &= \mathbb{E}_\nu\left[\left|\sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} - 1\right| \mathbb{1}_{B_d^c(r_k)}\right] \\
&\leq \frac{1}{(2\pi\sigma_q^2)^{d/2}} \int_{B_d^c(r_k)} \left(\left(\frac{\sigma_q}{\sigma_p}\right)^{d/4} e^{\frac{\|x-\mathsf{m}_q\|^2}{4\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{4\sigma_p^2}} + 1\right) e^{-\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}}\,\mathrm{d}x \\
&\leq \frac{1}{(2\pi\sigma_q^2)^{d/2}} \int_{B_d^c(r_k)} \left(\frac{\sigma_p}{\sigma_q}\right)^{d/4} e^{-\frac{\|x-\mathsf{m}_q\|^2}{4\sigma_q^2} - \frac{\|x-\mathsf{m}_p\|^2}{4\sigma_p^2}}\,\mathrm{d}x + e^{-\frac{\|x-\mathsf{m}_q\|^2}{2\sigma_q^2}}\,\mathrm{d}x
\end{aligned}$$

$$\overset{(b)}{\lesssim}_{d,M} e^{-\frac{\tilde{r}_k^2}{\tilde{\sigma}^2}} \leq e^{-0.5M\tilde{r}_k^2},$$

where (a) and (b) above follows from Lemma 65. Hence, $\bar{\mathcal{P}}^2_{\mathsf{N}}(M) \subseteq \breve{\mathcal{P}}^2_{\mathsf{H}^2}\big(\mathbf{r}, \mathbf{m}^{\mathsf{H}^2}, \mathbf{v}^{\mathsf{H}^2}\big)$ with $m_k^{\mathsf{H}^2} \asymp_{d,M} \big(1+\tilde{r}_k^{s_{\mathsf{KB}}+d+1}\big)e^{2M^2\tilde{r}_k^2}$ and $v_k^{\mathsf{H}^2} \asymp_{d,M} e^{-0.5M\tilde{r}_k^2}$, and (A.93) applies. Setting $r_k = 1 \vee M + \tilde{r}_k$ with $\tilde{r}_k = \sqrt{2cM^{-1}\log k}$, $c > 0$, and optimizing the resulting bound in (A.93) w.r.t. c (optimum achieved at $c = 0.5/(1+8M)$) yields with $m_k = m_k^{\mathsf{H}^2}$ and $\mathcal{G}_k = \breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k, r_k)$ that

$$\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\mathcal{G}_k}(X^n,Y^n) - \mathsf{H}^2(\mu,\nu)\right|\right] \lesssim_{d,M} (\log k)^{s_{\mathsf{KB}}+d+2} k^{-\frac{1}{2+16M}} \left(1 + k^{\frac{1}{2}} n^{-\frac{1}{2}}\right).$$

Taking supremum over $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\mathsf{N}}(M)$ completes the proof.

A.3.7 Proof of Theorem 51

Fix $\epsilon > 0$ and let $r(\epsilon)$ denote r such that $\mu\big(B_d^c(r)\big) \vee \nu\big(B_d^c(r)\big) \leq \epsilon$. Then, following steps leading to (A.64), there exists $g^* \in \vec{\mathcal{G}}^*_k(\phi, r(\epsilon))$ for $k \geq k_0(\epsilon)$ such that the following holds:

$$\begin{aligned}
&\left|\delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\vec{\mathcal{G}}^*_k(\phi,r(\epsilon))}(\mu,\nu)\right| \\
&\leq \mathbb{E}_\mu\left[|f_{\mathsf{TV}} - g^*|\, \mathbb{1}_{B_d(r(\epsilon))}\right] + \mathbb{E}_\nu\left[|f_{\mathsf{TV}} - g^*|\, \mathbb{1}_{B_d(r(\epsilon))}\right] + \mathbb{E}_\mu\big[\,|f_{\mathsf{TV}} - g^*|\, \mathbb{1}_{B_d^c(r(\epsilon))}\big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \mathbb{E}_\nu\big[\,|f_{\mathsf{TV}} - g^*|\, \mathbb{1}_{B_d^c(r(\epsilon))}\big] \\
&\lesssim \epsilon + \mathbb{E}_\mu\big[\,|f_{\mathsf{TV}}|\, \mathbb{1}_{B_d^c(r(\epsilon))}\big] + \mathbb{E}_\nu\big[\,|f_{\mathsf{TV}}|\, \mathbb{1}_{B_d^c(r(\epsilon))}\big] \\
&\leq \epsilon + \mu\big(B_d^c(r)\big) + \nu\big(B_d^c(r)\big) \lesssim \epsilon,
\end{aligned}$$

This combined with (A.59) proves Part (i).

Next, we prove Part (ii). Fix $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\mathsf{TV},\psi}(M, s, \mathbf{r}, \mathbf{m})$. For $t > 0$, let $f_{\mathsf{TV},r_k} := f_{\mathsf{TV}}\mathbb{1}_{B_d(r_k)}$ and $f^{(t)}_{\mathsf{TV},r_k} = f_{\mathsf{TV},r_k} * \Phi_t^{\mathcal{N}}$, where $\Phi_t^{\mathcal{N}}(x) := t^{-d}\Phi^{\mathcal{N}}(t^{-1}x)$ and $\Phi^{\mathcal{N}} = (2\pi)^{-d/2}e^{-0.5\|x\|^2}$. Then, similar to (A.70), we have

$$\begin{aligned}
S_2\Big(f^{(t)}_{\mathsf{TV},r_k}\Big) r_k &:= r_k \int_{\mathbb{R}^d} \|\omega\|_1^2 \left|\mathfrak{F}\left[f^{(t)}_{\mathsf{TV},r_k}\right](\omega)\right| \mathrm{d}\omega \\
&\leq r_k \left\|f_{\mathsf{TV},r_k}\right\|_1 d \int_{\mathbb{R}^d} \|\omega\|^2 \left|\mathfrak{F}[\Phi_t](\omega)\right| \mathrm{d}\omega \\
&= r_k^{d+1} \frac{d\pi^{0.5d}}{\Gamma(0.5d+1)} \int_{\mathbb{R}^d} \|\omega\|^2 e^{-\frac{1}{2}t^2\|\omega\|^2} \mathrm{d}\omega, \\
&=: \check{c}_{d,r_k,t},
\end{aligned}$$

where

$$\check{c}_{d,r_k,t} := \begin{cases} \sqrt{2}\pi r_k^2 t^{-3}\big(\Gamma(3/2)\big)^{-1}, & d = 1, \\ 2^{\frac{d+3}{2}}\pi^{0.5d+1} d r_k^{d+1} t^{-(d+2)} \prod_{j=1}^{d-2} \int_0^\pi \sin^{d-1-j}(\varphi_j)\mathrm{d}\varphi_j, & d \geq 2. \end{cases}$$

Then, noting that $\big|f^{(t)}_{\mathsf{TV}}(0)\big| \vee \big\|\nabla f^{(t)}_{\mathsf{TV}}(0)\big\| \leq 1 \vee (2d\pi^{-d})^{1/2}\Gamma\big(0.5(d+1)\big)t^{-1}$, it follows from (3.1) that there exists $g_{\theta_k} \in \vec{\mathcal{G}}^{\mathsf{R}}_k\big(\check{c}_{d,r_k,t}, r_k\big)$ such that

$$\left|f^{(t)}_{\mathsf{TV},r_k}(x) - g_{\theta_k}(x)\right| \lesssim \begin{cases} \check{c}_{d,r_k,t} d^{\frac{1}{2}} k^{-\frac{1}{2}}, & x \in B_d(r_k), \\ 1, & \text{otherwise}, \end{cases} \tag{A.96}$$

where $\check{c}_{d,r_k,t} := \check{c}_{d,r_k,t} \vee 1 \vee (2d\pi^{-d})^{1/2}\Gamma\big(0.5(d+1)\big)t^{-1}$.

On the other hand, we have similar to steps leading to (A.67) that

$$
\begin{aligned}
\left|\mathbb{E}_\mu\left[f_{\mathsf{TV},r_k} - f_{\mathsf{TV},r_k}^{(t)}\right]\right| &\leq \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV},r_k}(x) - f_{\mathsf{TV},r_k}(x-tu)\right| p(x)\mathrm{d}x\right]\Phi(u)\mathrm{d}u\\
&= \int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV},r_k}(x+tu) - f_{\mathsf{TV},r_k}(x)\right| p(x+tu)\mathrm{d}x\right]\Phi(u)\mathrm{d}u\\
&\leq \|p\|_{\infty,\mathbb{R}^d}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\left|f_{\mathsf{TV},r_k}(x+tu) - f_{\mathsf{TV},r_k}(x)\right| \mathrm{d}x\right]\Phi(u)\mathrm{d}u\\
&\leq M\int_{\mathbb{R}^d}\xi_{1,1}(f, t\left\|u\right\|)\Phi(u)\mathrm{d}u\\
&\leq Mm_k\int_{\mathbb{R}^d} t^s\left\|u\right\|^s\Phi(u)\mathrm{d}u = c_{s,d}^* Mm_k t^s,
\end{aligned}
$$

where $c_{s,d}^* = \int_{\mathbb{R}^d}\left\|u\right\|^s\Phi(u)\mathrm{d}u$. Then, defining $v_k = \mu\big(B_d^c(r_k)\big) \vee \nu\big(B_d^c(r_k)\big)$, we have

$$
\begin{aligned}
\left|\mathbb{E}_\mu\left[f_{\mathsf{TV}} - f_{\mathsf{TV},r_k}^{(t)}\right]\right| &\leq \left|\mathbb{E}_\mu\left[f_{\mathsf{TV}} - f_{\mathsf{TV},r_k}\right]\right| + \left|\mathbb{E}_\mu\left[f_{\mathsf{TV},r_k} - f_{\mathsf{TV},r_k}^{(t)}\right]\right|\\
&\leq 2\mu\big(B_d^c(r_k)\big) + c_{s,d}^* Mm_k t^s\\
&\leq 2v_k + c_{s,d}^* Mm_k t^s.
\end{aligned}
$$

Noting that the above holds with ν in place of μ, we obtain

$$
\left|\delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV},r_k}^{(t)}\right)\right| \leq 4v_k + 2c_{s,d}^* Mm_k t^s. \tag{A.97}
$$

Recalling that $\tilde{\delta}_{\mathsf{TV}}(g) := \mathbb{E}_\mu[g] - \mathbb{E}_\nu[g]$, we have

$$
\begin{aligned}
&\left|\delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\vec{\mathcal{G}}_k^{\mathsf{R}}\left(\check{c}_{d,r_k,t},r_k\right)}(\mu,\nu)\right|\\
&\quad\overset{(a)}{=} \delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\vec{\mathcal{G}}_k^{\mathsf{R}}\left(\check{c}_{d,r_k,t},r_k\right)}(\mu,\nu)\\
&\quad= \delta_{\mathsf{TV}}(\mu,\nu) - \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV},r_k}^{(t)}\right) + \tilde{\delta}_{\mathsf{TV}}\left(f_{\mathsf{TV},r_k}^{(t)}\right) - \delta_{\vec{\mathcal{G}}_k^{\mathsf{R}}(\check{c}_{d,r_k,t})}(\mu,\nu)\\
&\quad\overset{(b)}{\leq} 4v_k + 2c_{s,d}^* Mm_k t^s + \mathbb{E}_\mu\left[\left|f_{\mathsf{TV},r_k}^{(t)} - g_{\theta_k}\right|\right] + \mathbb{E}_\nu\left[\left|f_{\mathsf{TV},r_k}^{(t)} - g_{\theta_k}\right|\right]\\
&\quad\overset{(c)}{\lesssim}_{d,M,s} v_k + m_k t^s + r_k^{d+1}t^{-(d+2)}k^{-\frac{1}{2}} + \mu\big(B_d^c(r_k)\big) + \nu\big(B_d^c(r_k)\big)\\
&\quad\lesssim v_k + m_k t^s + r_k^{d+1}t^{-(d+2)}k^{-\frac{1}{2}},
\end{aligned}
$$

where (a) follows since $\left\|g\right\|_\infty \leq 1$ for $g \in \vec{\mathcal{G}}_k^{\mathsf{R}}\big(\check{c}_{d,r_k,t}, r_k\big)$ and (2.7); (b) uses (A.73) and (A.97); and (c) is due to (A.96). Setting $t = t_{k,s}^* = \big(r_k^{d+1}k^{-1/2}m_k^{-1}\big)^{1/(s+d+2)}$ yields

$$
\left|\delta_{\mathsf{TV}}(\mu,\nu) - \delta_{\vec{\mathcal{G}}_k^{\mathsf{R}}\left(\check{c}_{d,r_k,t_{k,s}^*},r_k\right)}(\mu,\nu)\right| \lesssim_{d,M,s} m_k^{\frac{d+2}{s+d+2}} r_k^{\frac{s(d+1)}{s+d+2}} k^{-\frac{s}{2(s+d+2)}} + v_k.
$$

Then, defining

$$\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}} := \breve{c}_{d,r_k,t^*_{k,s}} = O_d\Big(\big(r_k^{s(d+1)} k^{0.5(d+2)} m_k^{d+2}\big)^{\frac{1}{s+d+2}}\Big), \tag{A.98}$$

we have from the above equation and (A.77) that

$$\mathbb{E}\left[\left|\hat{\delta}_{\vec{\mathcal{G}}_k^{\mathsf{R}}(\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}},r_k)}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right] \lesssim_{d,M,s,\rho} m_k^{\frac{d+2}{s+d+2}} r_k^{\frac{s(d+1)}{s+d+2}} k^{-\frac{s}{2(s+d+2)}} + v_k \\ + n^{-\frac{1}{2}}\Big(m_k r_k^{s+1} k^{\frac{1}{2}}\Big)^{\frac{d+2}{s+d+2}}. \tag{A.99}$$

This completes the proof of Part (*ii*) by taking supremum w.r.t. $(\mu,\nu) \in \bar{\mathcal{P}}^2_{\mathsf{TV},\psi}(M,s,\mathbf{r},\mathbf{m})$ and noting that $v_k \leq \big(\psi\big(r_k M^{-1}\big)\big)^{-1}$ by Markov's inequality.

A.3.8 Proof of Corollary 52

We will use the following relation between sub-Gaussian and norm sub-Gaussian distributions. $\mu \in \mathcal{P}\big(\mathbb{R}^d\big)$ is σ^2-norm sub-Gaussian for $\sigma > 0$ if $X \sim \mu$ satisfies

$$\mu\big(\|X - \mathbb{E}[X]\| > t\big) \leq 2e^{\frac{-t^2}{2\sigma^2}}, \quad \forall\, t \in \mathbb{R}.$$

Lemma 67 *(Jin et al., 2019, Lemma 1) If $\mu \in \mathcal{P}\big(\mathbb{R}^d\big)$ is σ^2-sub-Gaussian, then it is $8d\sigma^2$-norm sub-Gaussian.*

Continuing with the proof of the Corollary, fix $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\mathsf{TV}}(b,M,N)$. From the above lemma, we have for $\mu \in \mathcal{SG}(M)$ and $t \geq M$ that

$$\mu\big(B_d^c(t)\big) \leq \mu\left(\|X - \mathbb{E}_\mu[X]\| + \|\mathbb{E}_\mu[X]\| > t\right) \leq 2e^{\frac{-(t-\|\mathbb{E}_\mu[X]\|)^2}{16d\sigma^2}} \leq 2e^{\frac{-(t-M)^2}{16dM}}. \tag{A.100}$$

Similar bound holds with ν in place of μ. Next, since $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\mathsf{TV}}(b,M,N)$, following the steps leading to (A.79) yields

$$\begin{aligned}\|f_{\mathsf{TV},r_k}\|_{\mathsf{Lip}(s,1)} &= \lambda(B_d(r_k)) + 2N\frac{\pi^{\frac{d}{2}} b^{d-s}}{\Gamma\Big(\frac{d}{2}+1\Big)} \vee 2b^{-s}\lambda(B_d(r_k)) \\ &= \frac{\pi^{\frac{d}{2}} r_k^d}{\Gamma\Big(\frac{d}{2}+1\Big)} + 2N\frac{\pi^{\frac{d}{2}} b^{d-s}}{\Gamma\Big(\frac{d}{2}+1\Big)} \vee 2b^{-s}\frac{\pi^{\frac{d}{2}} r_k^d}{\Gamma\Big(\frac{d}{2}+1\Big)} =: c_{d,s,b,N,r_k}.\end{aligned} \tag{A.101}$$

Then, it follows from (A.99) with $r_k = M \vee 1 + 4\sqrt{dM\log k}$, $v_k = 2e^{-(r_k-M)^2/16dM}$ and $m_k = c_{d,s,b,N,r_k}$ that

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{\delta}_{\vec{\mathcal{G}}_k^{\mathsf{R}}(\vec{c}_{k,d,s,\mathbf{m},\mathbf{r}},r_k)}(X^n,Y^n) - \delta_{\mathsf{TV}}(\mu,\nu)\right|\right] \\ &\qquad \lesssim_{d,s,b,N} (\log k)^{\frac{(s+d)(d+2)}{2(s+d+2)}} k^{-\frac{s}{2(s+d+2)}} + k^{-1} + (\log k)^{\frac{d+2}{2}} k^{\frac{d+2}{2(s+d+2)}} n^{-\frac{1}{2}} \\ &\qquad \lesssim_{d,s,b,N} (\log k)^{\frac{(s+d)(d+2)}{2(s+d+2)}} k^{-\frac{s}{2(s+d+2)}} + (\log k)^{\frac{d+2}{2}} k^{\frac{d+2}{2(s+d+2)}} n^{-\frac{1}{2}}.\end{aligned}$$

This completes the proof by taking supremum w.r.t. $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\mathsf{TV}}(b,M,N)$.

Appendix B. Proofs of Lemmas in Appendix A

B.1 Proof of Lemma 55

Suppose $f \in \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$. Since $f \in L^1(\mathbb{R}^d)$, its Fourier transform $\mathfrak{F}[f] : \mathbb{R}^d \to \mathbb{R}$ is well-defined. Also,

$$\begin{aligned}
\int_{\mathbb{R}^d} |\mathfrak{F}[f](\omega)| \, \mathrm{d}\omega &\overset{(a)}{\leq} \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}\omega}{1 + \|\omega\|^{2s_{\mathsf{KB}}}} \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(1 + \|\omega\|^{2s_{\mathsf{KB}}}\right) |\mathfrak{F}[f](\omega)|^2 \, \mathrm{d}\omega \right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq} \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}\omega}{1 + \|\omega\|^{2s_{\mathsf{KB}}}} \right)^{\frac{1}{2}} \left(\|f\|_2^2 + d^{s_{\mathsf{KB}}} \max_{\alpha:\|\alpha\|_1 = s_{\mathsf{KB}}} \|D^\alpha f\|_2^2 \right)^{\frac{1}{2}} \overset{(c)}{<} \infty,
\end{aligned}$$

where

(a) follows from Cauchy-Schwarz inequality;

(b) is by Plancherel's theorem since $\mathfrak{F}[D^\alpha f](\omega) = \mathfrak{F}[f](\omega) \prod_{j=1}^d (i\omega_j)^{\alpha_j}, \ \forall \ \|\alpha\|_1 \leq s_{\mathsf{KB}}$, where i denotes the imaginary unit $\sqrt{-1}$, and $f \in \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$. The above identity holds because $\|D^\alpha f\|_1 < \infty$ for all $\|\alpha\|_1 \leq s_{\mathsf{KB}}$ by assumption.

(c) follows since the first integral is finite and $f \in \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$.

Hence, $\mathfrak{F}[f] \in L^1(\mathbb{R}^d)$ and the Fourier inversion formula holds (at every $x \in \mathbb{R}^d$ since $f \in \mathcal{L}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$ is necessarily continuous) with $F(d\omega) = \mathfrak{F}[f](\omega)\mathrm{d}\omega$, i.e., $f(x) = \int_0^\infty e^{i\omega\cdot x}\mathfrak{F}[f](\omega)\mathrm{d}\omega$. Then, it follows from $\|\omega\|_1 \leq \sqrt{d}\,\|\omega\|$ that

$$S_2(f) := \int_{\mathbb{R}^d} \|\omega\|_1^2 \left|\mathfrak{F}[f](\omega)\right| \mathrm{d}\omega \leq d \int_{\mathbb{R}^d} \|\omega\|^2 \left|\mathfrak{F}[f](\omega)\right| \mathrm{d}\omega. \tag{B.1}$$

If $\|D^\alpha f\|_2 \leq b$ for all α with $\|\alpha\|_1 \in \{1, s_{\mathsf{KB}}\}$, then we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \|\omega\|^2 \left|\mathfrak{F}[f](\omega)\right| \mathrm{d}\omega &\overset{(a)}{\leq} \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}\omega}{1 + \|\omega\|^{2(s_{\mathsf{B}}-1)}} \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^d} \left(\|\omega\|^4 + \|\omega\|^{2s_{\mathsf{KB}}}\right) \left|\mathfrak{F}[f](\omega)\right|^2 \mathrm{d}\omega \right)^{\frac{1}{2}} \\
&\overset{(b)}{\leq} \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}\omega}{1 + \|\omega\|^{2(s_{\mathsf{B}}-1)}} \right)^{\frac{1}{2}} \left(d^2 + d^{s_{\mathsf{KB}}}\right)^{\frac{1}{2}} b,
\end{aligned} \tag{B.2}$$

where

(a) follows from Cauchy-Schwarz inequality;

(b) is due to Plancherel's theorem and $f \in \mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$.

Combining (B.1) and (B.2) yields $S_2(f) \leq bd^{3/2}\kappa_d$. Following similar steps, it can be shown that if $f \in \mathcal{L}^{\mathsf{B}}_{s_{\mathsf{B}},b}(\mathbb{R}^d)$, then $S_1(f) \leq bd^{1/2}\kappa_d$. The final claims follows from these and definition of the classes $\mathcal{L}^{\mathsf{B}}_{s_{\mathsf{B}},b}(\mathbb{R}^d)$, $\mathcal{L}^{\mathsf{KB}}_{s_{\mathsf{KB}},b}(\mathbb{R}^d)$, $\mathcal{B}_{c,1,\mathcal{X}}(\mathbb{R}^d)$ and $\mathcal{B}_{c,2,\mathcal{X}}(\mathbb{R}^d)$.

B.2 Proof of Lemma 57

Assume that ϕ is monotone increasing. Let $g, \tilde{g} \in \mathcal{G}_k(\mathbf{a}_k, \phi)$ be arbitrary, where $g(x) = \sum_{i=1}^{k} \beta_i \phi\left(w_i \cdot x + b_i\right) + w_0 \cdot x + b_0$ and $\tilde{g}(x) = \sum_{i=1}^{k} \tilde{\beta}_i \phi\left(\tilde{w}_i \cdot x + \tilde{b}_i\right) + \tilde{w}_0 \cdot x + \tilde{b}_0$. Define $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_k)$, $\tilde{\boldsymbol{\beta}} := (\tilde{\beta}_1, \ldots, \tilde{\beta}_k)$, $\mathbf{w} = (w_1, \ldots, w_k)$, $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_k)$, $\mathbf{b} = (b_1, \ldots, b_k)$ and $\tilde{\mathbf{b}} = (\tilde{b}_1, \ldots, \tilde{b}_k)$. Note that $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}, \mathbf{b}, \tilde{\mathbf{b}} \in \mathbb{R}^k$ and $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^{kd}$. For any $x \in \mathcal{X}$, we have

$$
\begin{aligned}
&|g(x) - \tilde{g}(x)| \\
&\leq \left| \sum_{i=1}^{k} \beta_i \phi\left(w_i \cdot x + b_i\right) - \sum_{i=1}^{k} \tilde{\beta}_i \phi\left(\tilde{w}_i \cdot x + \tilde{b}_i\right) \right| + \left|(w_0 - \tilde{w}_0) \cdot x\right| + \left|b_0 - \tilde{b}_0\right| \\
&\leq \left| \sum_{i=1}^{k} \beta_i \phi\left(w_i \cdot x + b_i\right) - \sum_{i=1}^{k} \tilde{\beta}_i \phi\left(w_i \cdot x + b_i\right) \right| + \left\|w_0 - \tilde{w}_0\right\|_1 \|\mathcal{X}\| + \left|b_0 - \tilde{b}_0\right| \\
&\qquad + \left| \sum_{i=1}^{k} \tilde{\beta}_i \phi\left(w_i \cdot x + b_i\right) - \sum_{i=1}^{k} \tilde{\beta}_i \phi\left(\tilde{w}_i \cdot x + \tilde{b}_i\right) \right| \\
&\overset{(a)}{\leq} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right\|_1 \phi\left(\sup_{x \in \mathcal{X}, 1 \leq i \leq k} |w_i \cdot x + b_i| \right) + \left\|w_0 - \tilde{w}_0\right\|_1 \|\mathcal{X}\| + \left|b_0 - \tilde{b}_0\right| \\
&\qquad + L \sum_{i=1}^{k} \left|\tilde{\beta}_i\right| \left|(w_i - \tilde{w}_i) \cdot x + b_i - \tilde{b}_i\right| \\
&\overset{(b)}{\leq} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right\|_1 \phi\big(a_{1,k}(\|\mathcal{X}\| + 1)\big) + \left\|w_0 - \tilde{w}_0\right\|_1 \|\mathcal{X}\| + \left|b_0 - \tilde{b}_0\right| + L a_{2,k} \|\mathcal{X}\| \left\|\mathbf{w} - \tilde{\mathbf{w}}\right\|_1 \\
&\qquad + L a_{2,k} \left\|\mathbf{b} - \tilde{\mathbf{b}}\right\|_1,
\end{aligned}
$$

where

(a) is since ϕ is monotone increasing function with Lipschitz constant bounded by L;

(b) is because $\max_{1 \leq i \leq k} \|w_i\|_1 \vee |b_i| \leq a_{1,k}$ and $\max_{1 \leq i \leq k} |\tilde{\beta}_i| \leq a_{2,k}$.

Defining $u_k = \phi\big(a_{1,k}(\|\mathcal{X}\| + 1)\big)$, it follows by application of (A.19) that

$$
\begin{aligned}
N\big(\epsilon, \mathcal{G}_k(\mathbf{a}_k, \phi), \|\cdot\|_{\infty, \mathcal{X}}\big) &\leq N\big(\epsilon/5, [-a_{2,k}, a_{2,k}]^k, u_k \|\cdot\|_1\big) N\big(\epsilon/5, B_d^1(a_{4,k}), \|\mathcal{X}\| \|\cdot\|_1\big) \\
&\qquad N\big(\epsilon/5, [-a_{3,k}, a_{3,k}], |\cdot|\big) N\big(\epsilon/5, B_{kd}^1(k a_{1,k}), L a_{2,k} \|\mathcal{X}\| \|\cdot\|_1\big) \\
&\qquad N\big(\epsilon/5, B_k^1(k a_{1,k}), L a_{2,k} \|\cdot\|_1\big) \\
&\leq \big(1 + 10 k a_{2,k} u_k \epsilon^{-1}\big)^k \big(1 + 10 a_{4,k} \|\mathcal{X}\| \epsilon^{-1}\big)^d \big(1 + 10 a_{3,k} \epsilon^{-1}\big) \\
&\qquad \big(1 + 10 L k a_{1,k} a_{2,k} \|\mathcal{X}\| \epsilon^{-1}\big)^{dk} \big(1 + 10 L k a_{1,k} a_{2,k} \epsilon^{-1}\big)^k.
\end{aligned}
$$

If ϕ is monotone decreasing, the above holds with $u_k = \phi\big(-a_{1,k}(\|\mathcal{X}\| + 1)\big)$. This proves the first bound in Lemma 57. Specializing to NN classes $\mathcal{G}_k^{\mathsf{R}}(a)$, $\mathcal{G}_k^{\mathsf{S}}(a)$, $\mathcal{G}_k^*(\phi_{\mathsf{R}})$, and $\mathcal{G}_k^*(\phi_{\mathsf{S}})$ by noting that the Lipschitz constant $L \leq 1$ for ϕ_{R} and ϕ_{S}, $|\phi_{\mathsf{R}}(x)| \leq x$, and $|\phi_{\mathsf{S}}(x)| \leq 1$, yields (A.16)-(A.18).

B.3 Proof of Lemma 59

We will use Theorem 11 for the proof. Fix any $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$. Note that for $h_{\mathsf{KL}}(x) = e^x - 1$, we have $\bar{C}\big(|\mathcal{G}^*_k(\phi)|, \mathcal{X}\big) \leq k(\|\mathcal{X}\|+1)+1$,

$$\begin{aligned} &\bar{C}\left(\left|h'_{\mathsf{KL}} \circ \mathcal{G}^*_k(\phi)\right|, \mathcal{X}\right) \leq e^{k(\|\mathcal{X}\|+1)+1}, \\ &V_{k,h,\phi,\mathcal{X}} \lesssim \big(k(\|\mathcal{X}\|+1)+1\big)^2 \big(e^{k(\|\mathcal{X}\|+1)+1}+1\big)^2, \end{aligned} \tag{B.3}$$

where h'_{KL} denotes the derivative of h_{KL} and $V_{k,h,\phi,\mathcal{X}}$ is given in (A.20). Also, observe that since $g \in \mathcal{G}^*_k(\phi)$ is continuous and bounded, $\mathsf{D}_{\mathcal{G}^*_k(\phi)}(\mu,\nu) \leq \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) < \infty$. Then, since

$$E_{k,h,\phi,\mathcal{X}}\, n^{-\frac{1}{2}} \lesssim n^{-\frac{1}{2}} k\sqrt{d(\|\mathcal{X}\|+1)}\sqrt{k(\|\mathcal{X}\|+1)+1}\left(e^{k(\|\mathcal{X}\|+1)+1}+1\right) \xrightarrow[n\to\infty]{} 0,$$

for k such that $k^{3/2}(\|\mathcal{X}\|+1)e^{k(\|\mathcal{X}\|+1)} = O\left(n^{(1-\rho)/2}\right)$ for $0<\rho<1$, it follows from (3.3) that for any $k \in \mathbb{N}$, $\delta>0$, and n sufficiently large,

$$\mathbb{P}\left(\left|\mathsf{D}_{\mathcal{G}^*_k(\phi)}(\mu,\nu) - \hat{\mathsf{D}}_{\mathcal{G}^*_k(\phi)}(X^n,Y^n)\right| \geq \delta\right) \leq c e^{-\frac{n\left(\delta - E_{k,h,\phi,\mathcal{X}}n^{-1/2}\right)^2}{V_{k,h,\phi,\mathcal{X}}}}.$$

Hence, for k_n such that $k_n^{3/2}(\|\mathcal{X}\|+1)e^{k_n(\|\mathcal{X}\|+1)} = O\left(n^{(1-\rho)/2}\right)$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\mathsf{D}_{\mathcal{G}^*_k(\phi)}(\mu,\nu) - \hat{\mathsf{D}}_{\mathcal{G}^*_k(\phi)}(X^n,Y^n)\right| \geq \delta\right) \leq c\sum_{n=1}^{\infty} e^{\frac{-n\left(\delta - E_{k_n,h,\phi,\mathcal{X}}n^{-1/2}\right)^2}{V_{k_n,h,\phi,\mathcal{X}}}} < \infty, \tag{B.4}$$

where the final inequality in (B.4) can be established via integral test for sum of series. This implies (A.37) via the first Borel-Cantelli lemma by taking supremum w.r.t. $(\mu,\nu) \in \mathcal{P}^2_{\mathsf{KL}}(\mathcal{X})$.

B.4 Proof of Lemma 60

Fix $(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathcal{X})$. Recalling the quantities defined in Theorem 11, we have for $h_{\chi^2}(x) = x + 0.25x^2$ that

$$\begin{aligned} &\bar{C}\big(\left|h'_{\chi^2} \circ \mathcal{G}^*_k(\phi)\right|, \mathcal{X}\big) \leq 0.5k(\|\mathcal{X}\|+1)+1.5, \\ &V_{k,h,\phi,\mathcal{X}} \lesssim (k(\|\mathcal{X}\|+1)+1)^2(0.5k(\|\mathcal{X}\|+1)+1.5)^2, \\ &E_{k,h,\phi,\mathcal{X}} \lesssim k\sqrt{d(\|\mathcal{X}\|+1)}\big(0.5k(\|\mathcal{X}\|+1)+1.5\big)\sqrt{k(\|\mathcal{X}\|+1)+1}, \end{aligned} \tag{B.5}$$

where h'_{χ^2} denotes the derivative of h_{χ^2}. Also, note that $\chi^2_{\mathcal{G}^*_k(\phi)}(\mu,\nu) \leq \chi^2\left(\mu\|\nu\right) < \infty$. Then, since

$$0 \leq E_{k,h,\phi,\mathcal{X}}\, n^{-\frac{1}{2}} \lesssim k^{\frac{5}{2}}(\|\mathcal{X}\|+1)^2 n^{-\frac{1}{2}} \xrightarrow[n\to\infty]{} 0,$$

for $k^{5/2}(\|\mathcal{X}\|+1)^2 = O\left(n^{(1-\rho)/2}\right)$, $0<\rho<1$, it follows from (3.3) that for any $k \in \mathbb{N}$, $\delta > 0$, and n sufficiently large,

$$\mathbb{P}\left(\left|\hat{\chi}^2_{\mathcal{G}^*_k(\phi)}(X^n,Y^n) - \chi^2_{\mathcal{G}^*_k(\phi)}(\mu,\nu)\right| \geq \delta\right) \leq c e^{-\frac{n\left(\delta - E_{k,h,\phi,\mathcal{X}}n^{-1/2}\right)^2}{V_{k,h,\phi,\mathcal{X}}}}.$$

Then, (A.45) follows using similar steps used to prove (A.37) (see (B.4)). This completes the proof.

B.5 Proof of Lemma 61

Fix $(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}(\mathcal{X})$. Note that $h_{\mathsf{H}^2}(x)=x/(1-x)$ and

$$\bar{C}\big(\big|h'_{\mathsf{H}^2}\circ\tilde{\mathcal{G}}^*_{k,t}(\phi)\big|\big)=\sup_{g_\theta\in\tilde{\mathcal{G}}^*_{k,t}(\phi),x\in\mathcal{X}}(1-g_\theta(x))^{-2}\le t^{-2},$$

where h'_{H^2} denotes derivative of h_{H^2}. By examining the proof, it can be seen that Theorem 11 continues to hold with $\mathcal{G}^*_k(\phi)$ in (3.2) and (3.3) replaced with $\tilde{\mathcal{G}}^*_{k,t}(\phi)$. We have $V_{k,h,\phi,\mathcal{X}}\lesssim(k(\|\mathcal{X}\|+1)+1)^2\left(t_k^{-2}+1\right)^2$, and

$$0\le E_{k,h,\phi,\mathcal{X}}n^{-\frac{1}{2}}\lesssim n^{-\frac{1}{2}}k\sqrt{d(\|\mathcal{X}\|+1)}\big(t_k^{-2}+1\big)\sqrt{k(\|\mathcal{X}\|+1)+1}\xrightarrow[n\to\infty]{}0,$$

for k,t_k such that $k^{3/2}(\|\mathcal{X}\|+1)t_k^{-2}=O\left(n^{(1-\rho)/2}\right)$. Further, $\mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,t}(\phi)}(\mu,\nu)<\mathsf{H}^2(\mu,\nu)\le 2$. It then follows from (3.3) that for any $k\in\mathbb{N}$, $\delta>0$, and n sufficiently large,

$$\mathbb{P}\left(\left|\hat{\mathsf{H}}^2_{\tilde{\mathcal{G}}^*_{k,t_k}(\phi)}(X^n,Y^n)-\mathsf{H}^2_{\tilde{\mathcal{G}}^*_{k,t_k}(\phi)}(\mu,\nu)\right|\ge\delta\right)\le ce^{-\frac{n\left(\delta-E_{k,h,\phi,\mathcal{X}}n^{-1/2}\right)^2}{V_{k,h,\phi,\mathcal{X}}}}.$$

Then, (A.53) follows via similar steps used to prove (A.37) (see (B.4)).

B.6 Proof of Lemma 62

Fix $\mu,\nu\in\mathcal{P}(\mathcal{X})$. We have $\delta_{\bar{\mathcal{G}}^*_k(\phi)}(\mu,\nu)\le\delta_{\mathsf{TV}}(\mu,\nu)\le 2$, $\bar{C}\left(\left|\bar{\mathcal{G}}^*_k(\phi)\right|\right)\le 1$, and

$$\bar{C}\left(\left|h'_{\mathsf{TV}}\circ\bar{\mathcal{G}}^*_k(\phi)\right|\right)=1,$$

where h'_{TV} denotes the derivative of h_{TV}. Also, it can be seen from the proof of Theorem 11 that it holds with $\mathcal{G}^*_k(\phi)$ in (3.2) and (3.3) replaced by $\bar{\mathcal{G}}^*_k(\phi)$. Further, $V_{k,h,\phi,\mathcal{X}}\lesssim 1$, and

$$0\le E_{k,h,\phi,\mathcal{X}}n^{-\frac{1}{2}}\lesssim n^{-\frac{1}{2}}k\sqrt{d(\|\mathcal{X}\|+1)}\xrightarrow[n\to\infty]{}0,$$

for k,n such that $k(\|\mathcal{X}\|+1)^{1/2}=O\left(n^{(1-\rho)/2}\right)$. It follows from (3.3) that for any $k\in\mathbb{N}$, $\delta>0$, and n sufficiently large,

$$\mathbb{P}\left(\left|\hat{\delta}_{\bar{\mathcal{G}}^*_k(\phi)}(X^n,Y^n)-\delta_{\bar{\mathcal{G}}^*_k(\phi)}(\mu,\nu)\right|\ge\delta\right)\le ce^{-\frac{n\left(\delta-E_{k,h,\phi,\mathcal{X}}n^{-1/2}\right)^2}{V_{k,h,\phi,\mathcal{X}}}}.$$

Then, (A.59) follows using similar steps used to prove (A.37). This completes the proof.

B.7 Proof of Lemma 63

Fix $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\mathsf{KL}}(M,\mathbf{r},\mathbf{m},\mathbf{v})$. Recall that $\hat{\mathcal{G}}^{\mathsf{R}}_k(a,r)=\{g\mathbb{1}_{B_d(r)}:g\in\mathcal{G}^{\mathsf{R}}_k(a)\}$. Since $c^\star_{\mathsf{KB}}\left(f_{\mathsf{KL}}|_{B_d(r_k)},B_d(r_k)\right)\le m_k$, it follows from (3.1) that there exists $g_{\theta_k}\in\hat{\mathcal{G}}^{\mathsf{R}}_k(m_k,r_k)$ and $c>0$ such that

$$\left\|f_{\mathsf{KL}}-g_{\theta_k}\right\|_{\infty,B_d(r_k)}\le cd^{\frac{1}{2}}m_kk^{-\frac{1}{2}}. \tag{B.6}$$

Then, following steps leading to (A.81), we have for k with $c^2dm_k^2 < 0.5k$ that

$$\begin{aligned}
&\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu)\right| \\
&\leq \left\|(f_{\mathsf{KL}} - g_{\theta_k})\mathbb{1}_{B_d(r_k)}\right\|_{\infty,\mu} + \mathbb{E}_\mu\left[|f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r_k)}\right] + \mathbb{E}_\nu\left[\left|\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1\right|\mathbb{1}_{B_d^c(r_k)}\right] \\
&\qquad\qquad + \mathbb{E}_\nu\left[\left|e^{f_{\mathsf{KL}}}\right|\mathbb{1}_{B_d(r_k)}\right]\left\|\left(1 - e^{g_{\theta_k} - f_{\mathsf{KL}}}\right)\mathbb{1}_{B_d(r_k)}\right\|_{\infty,\nu} \\
&\lesssim m_k d^{\frac{1}{2}} k^{-\frac{1}{2}} + v_k,
\end{aligned}$$

where the final inequality is due to (B.6), $e^{cd^{\frac{1}{2}}m_k k^{-1/2}} - 1 \leq cd^{\frac{1}{2}}m_k k^{-1/2}$ which follows similar to (A.43) (since $c^2dm_k^2 < 0.5k$), $\mathbb{E}_\mu\big[|f_{\mathsf{KL}}|\mathbb{1}_{B_d^c(r_k)}\big] \vee \mathbb{E}_\nu\big[|(\mathrm{d}\mu/\mathrm{d}\nu) - 1|\mathbb{1}_{B_d^c(r_k)}\big] \leq v_k$, and $\mathbb{E}_\nu\big[|e^{f_{\mathsf{KL}}}|\mathbb{1}_{B_d(r_k)}\big] \leq 1$.

On the other hand, for k such that $c^2dm_k^2 \geq 0.5k$, $g = 0 \in \hat{\mathcal{G}}_k^{\mathsf{R}}(m_k, r_k)$ implies that

$$\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu)\right| = \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu) \leq \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) \leq M.$$

Since $m_k^2 \lesssim k^{1-\rho}$, k such that $c^2dm_k^2 \geq 0.5k$ necessarily satisfies $k^\rho \lesssim d$. Thus, for all $k \in \mathbb{N}$,

$$\left|\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right) - \mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu)\right| \lesssim_{d,M,\rho} m_k k^{-\frac{1}{2}} + v_k. \tag{B.7}$$

Note that the RHS above tends to zero as $k \to \infty$ since $v_k \to 0$ and $m_k^2 \lesssim k^{1-\rho}$.

Next, it follows from (A.29), (A.33), and (B.7) that for k, m_k satisfying $m_k^2 \lesssim k^{1-\rho}$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(X^n, Y^n) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right] \\
&\quad \leq \left|\mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu) - \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right| + \mathbb{E}\left[\left|\mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu) - \hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(X^n, Y^n)\right|\right] \\
&\quad \lesssim_{d,M,\rho} m_k k^{-\frac{1}{2}} + v_k + m_k r_k e^{3m_k(r_k+1)} n^{-\frac{1}{2}}.
\end{aligned}$$

Taking supremum w.r.t. $(\mu,\nu) \in \breve{\mathcal{P}}_{\mathsf{KL}}^2(M, \mathbf{r}, \mathbf{m}, \mathbf{v})$ completes the proof.

B.8 Proof of Lemma 64

Fix $(\mu,\nu) \in \breve{\mathcal{P}}_{\chi^2}^2(\mathbf{r}, \mathbf{m}, \mathbf{v})$. Since $c_{\mathsf{KB}}^\star\left(f_{\chi^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq m_k$, there exists $g_{\theta_k} \in \hat{\mathcal{G}}_k^{\mathsf{R}}(m_k, r_k)$ such that

$$\left\|f_{\chi^2} - g_{\theta_k}\right\|_{\infty,B_d(r_k)} \lesssim d^{\frac{1}{2}} m_k k^{-\frac{1}{2}}. \tag{B.8}$$

Then, following steps leading to (A.86), we have for all $k \in \mathbb{N}$ that

$$\begin{aligned}
&\left|\chi^2\left(\mu\|\nu\right) - \chi^2_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu)\right| \\
&\leq \left\|\left(f_{\chi^2} - g_{\theta_k}\right)\mathbb{1}_{B_d(r_k)}\right\|_{\infty,\mu} + \mathbb{E}_\mu\left[\left|f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\right] + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d(r_k)}\right] \\
&\qquad\qquad + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2}\right|\mathbb{1}_{B_d^c(r_k)}\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\lesssim} d^{\frac{1}{2}} m_k k^{-\frac{1}{2}} + v_k + \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2} - h_{\chi^2}\circ g_{\theta_k}\right|\mathbb{1}_{B_d(r_k)}\right]\\
&\overset{(b)}{\lesssim} d^{\frac{1}{2}} m_k k^{-\frac{1}{2}} + v_k + \mathbb{E}_\nu\left[\left|f_{\chi^2} - g_{\theta_k}\right|\mathbb{1}_{B_d(r_k)}\right] + \mathbb{E}_\nu\Big[0.25\left|f_{\chi^2} - g_{\theta_k}\right|^2\mathbb{1}_{B_d(r_k)}\Big]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + 0.5\mathbb{E}_\nu\Big[\left|f_{\chi^2} - g_{\theta_k}\right|\left|f_{\chi^2}\right|\mathbb{1}_{B_d(r_k)}\Big]\\
&\lesssim d^{\frac{1}{2}} m_k k^{-\frac{1}{2}} + v_k + d m_k^2 k^{-1} + \left\|(f_{\chi^2} - g_{\theta_k})\mathbb{1}_{B_d(r_k)}\right\|_{\infty,\nu}\mathbb{E}_\nu\left[\left|f_{\chi^2}\right|\right]\\
&\overset{(c)}{\lesssim} d^{\frac{1}{2}} m_k k^{-\frac{1}{2}} + d m_k^2 k^{-1} + v_k,
\end{aligned}
$$

where

(a) follows from (B.8) and since $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v})$;

(b) is via steps leading to (A.48);

(c) is due to (B.8) and $\mathbb{E}_\nu\left[\left|f_{\chi^2}\right|\right]\leq 4$.

Then, it follows from the above equation, (A.29) and (A.33) that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\hat{\chi}^2_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(X^n,Y^n) - \chi^2\left(\mu\|\nu\right)\right|\right]\\
&\qquad\qquad \leq \left|\chi^2_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu) - \chi^2\left(\mu\|\nu\right)\right| + \mathbb{E}\left[\left|\chi^2_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(\mu,\nu) - \hat{\chi}_{\hat{\mathcal{G}}_k^{\mathsf{R}}(m_k,r_k)}(X^n,Y^n)\right|\right]\\
&\qquad\qquad \lesssim d^{\frac{1}{2}} m_k k^{-\frac{1}{2}} + d m_k^2 k^{-1} + v_k + d^{\frac{3}{2}} m_k^2 r_k^2 n^{-\frac{1}{2}}.
\end{aligned}
$$

Taking supremum w.r.t. $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\chi^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ completes the proof.

B.9 Proof of Lemma 66

Fix $(\mu,\nu)\in\breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$. Since $\left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty,B_d(r_k)}\leq m_k$, we have

$$
1 - f_{\mathsf{H}^2}(x) = \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)^{-\frac{1}{2}} \geq m_k^{-\frac{1}{2}},\ x\in B_d(r_k). \tag{B.9}
$$

Hence, $c^\star_{\mathsf{KB}}\left(f_{\mathsf{H}^2}|_{B_d(r_k)}, B_d(r_k)\right)\leq m_k$ implies via (3.1) and (5.2) that there exists $g_{\theta_k}\in\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)$ such that

$$
\left\|f_{\mathsf{H}^2} - g_{\theta_k}\right\|_{\infty,B_d(r_k)} \lesssim m_k d^{\frac{1}{2}} k^{-\frac{1}{2}}. \tag{B.10}
$$

Following the derivation leading to the penultimate step in (A.55), we have

$$
\begin{aligned}
&\left|\mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(\mu,\nu)\right|\\
&\leq \mathbb{E}_\mu\left[\left|f_{\mathsf{H}^2} - g_{\theta_k}\right|\mathbb{1}_{B_d(r_k)}\right] + \mathbb{E}_\nu\left[\left|\frac{f_{\mathsf{H}^2} - g_{\theta_k}}{(1-f_{\mathsf{H}^2})(1-g_{\theta_k})}\right|\mathbb{1}_{B_d(r_k)}\right] + \mathbb{E}_\mu\left[\left|f_{\mathsf{H}^2}\right|\mathbb{1}_{B_d^c(r_k)}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \mathbb{E}_\nu\left[\left|\frac{f_{\mathsf{H}^2}}{(1-f_{\mathsf{H}^2})}\right|\mathbb{1}_{B_d^c(r_k)}\right]
\end{aligned}
$$

$$\overset{(a)}{\lesssim} m_k d^{\frac{1}{2}} k^{-\frac{1}{2}} + m_k^2 d^{\frac{1}{2}} k^{-\frac{1}{2}} + v_k$$
$$\overset{(b)}{\lesssim} m_k^2 d^{\frac{1}{2}} k^{-\frac{1}{2}} + v_k, \tag{B.11}$$

where (a) follows from (B.9), (B.10), $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$, and $1 - g_{\theta_k}(x) \geq m_k^{-1/2}$ by the definition of $\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k, r_k)$, while (b) is due to $m_k \geq 1$.

Next, using (A.27) and following steps similar to proof of Lemma 59, we obtain that for k, m_k, r_k, n such that $k^{1/2} m_k^2 r_k = O\left(n^{(1-\rho)/2}\right)$,

$$\hat{\mathsf{H}}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(X^n, Y^n) \xrightarrow[n\to\infty]{} \mathsf{H}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(\mu,\nu), \quad \mathbb{P}-\text{a.s.}$$

Then, (A.92) follows from this and (B.11) since $m_k = o(k^{1/4})$ and $v_k \to 0$ by assumption.

Also,

$$\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(X^n,Y^n) - \mathsf{H}^2(\mu,\nu)\right|\right]$$
$$\leq \left|\mathsf{H}^2(\mu,\nu) - \mathsf{H}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(\mu,\nu)\right| + \mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(X^n,Y^n) - \mathsf{H}^2_{\breve{\mathcal{G}}^{\mathsf{R}}_{k,m_k^{-1/2}}(m_k,r_k)}(\mu,\nu)\right|\right]$$
$$\lesssim m_k^2 d^{\frac{1}{2}} k^{-\frac{1}{2}} + v_k + d^{\frac{3}{2}} m_k^2 r_k n^{-\frac{1}{2}},$$

where the final inequality uses (A.29), (A.33) and (B.11) to bound the last term. Taking supremum over $(\mu,\nu) \in \breve{\mathcal{P}}^2_{\mathsf{H}^2}(\mathbf{r},\mathbf{m},\mathbf{v})$ yields (A.93).

Appendix C. Consistency and effective error bounds for DV-NE

Defining $\mathsf{D}_{\mathsf{DV},\mathcal{G}}(\mu,\nu) := \sup_{g\in\mathcal{G}} \left(\mathbb{E}_\mu[g] - \log \mathbb{E}_\nu[e^g]\right)$ and

$$\tilde{Z}_g := \frac{1}{n}\sum_{i=1}^n g(X_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}\right) - \mathbb{E}_\mu\big[g\big] + \log \mathbb{E}_\nu\big[e^g\big],$$

we have similar to (A.22) that

$$\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}}(X^n,Y^n) - \mathsf{D}_{\mathsf{DV},\mathcal{G}}(\mu,\nu) \leq \sup_{g\in\mathcal{G}} \tilde{Z}_g.$$

Moreover, since the Lipschitz constant of logarithm is bounded by $e^{\bar{C}(|\mathcal{G}|,\mathcal{X})}$ in $\left[e^{-\bar{C}(|\mathcal{G}|,\mathcal{X})}, e^{\bar{C}(|\mathcal{G}|,\mathcal{X})}\right]$, we have almost surely that

$$|Z_g - Z_{\tilde{g}}| \leq n^{-1}\sum_{i=1}^n \left|g(X_i) - \tilde{g}(X_i) - \mathbb{E}_\mu\big[g-\tilde{g}\big]\right| + e^{\bar{C}(\mathcal{G},\mathcal{X})}\left|e^{g(Y_i)} - e^{\tilde{g}(Y_i)} - \mathbb{E}_\nu\left[e^g - e^{\tilde{g}}\right]\right|,$$

where each term inside the summation is bounded by $2\left(e^{2\bar{C}(|\mathcal{G}|)}+1\right)\left\|g_\theta - g_{\tilde{\theta}}\right\|_{\infty,\mathcal{X}}$ similar to (A.26). Then, following the steps in the proof of Theorem 11, we have

$$\sup_{\substack{\mu,\nu\in\mathcal{P}(\mathcal{X}):\\ \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(\mu,\nu)<\infty}} \mathbb{P}\left(\left|\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(X^n,Y^n) - \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(\mu,\nu)\right| \geq \delta + \tilde{E}_{k,h,\phi,\mathcal{X}} n^{-\frac{1}{2}}\right) \leq c\, e^{-\frac{n\delta^2}{\tilde{V}_{k,h,\phi,\mathcal{X}}}},$$

where $\tilde{V}_{k,h,\phi,\mathcal{X}} \lesssim (k(\|\mathcal{X}\|+1)+1)^2 e^{4k(\|\mathcal{X}\|+1)}$ and $\tilde{E}_{k,h,\phi,\mathcal{X}} \lesssim k^{3/2} d^{1/2}(\|\mathcal{X}\|+1)e^{2k(\|\mathcal{X}\|+1)}$. Then, similar to Lemma 59, we obtain that for any $0<\rho<1$, and n, k_n such that $k_n^{3/2}(\|\mathcal{X}\|+1)e^{2k_n(\|\mathcal{X}\|+1)} = O\left(n^{(1-\rho)/2}\right)$,

$$\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(\mu,\nu), \quad \mathbb{P}-\text{a.s.}$$

Moreover, $\lim_{n\to\infty} \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(\mu,\nu) = \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)$ follows identical to (A.39) provided $f_{\mathsf{KL}} \in \mathsf{C}(\mathcal{X})$. Hence, for $\mathcal{X}=[0,1]^d$, we obtain that for any $0<\rho<1$, $(k_n)_{n\in\mathbb{N}}$ with $k_n \to\infty$ and $k_n \leq \frac{1}{8}(1-\rho)\log n$, we have

$$\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}_k^*(\phi)}(X^n,Y^n) \xrightarrow[n\to\infty]{} \mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right), \quad \mathbb{P}-\text{a.s.} \tag{C.1}$$

Next, we bound the expected error of the DV-NE estimator. Note that

$$\begin{aligned}
&\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}_k^{\mathsf{R}}(a)}(X^n,Y^n) - \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^{\mathsf{R}}(a)}(\mu,\nu)\\
&= \sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)} \frac{1}{n}\sum_{i=1}^n g(X_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}\right) - \sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)} \left(\mathbb{E}_\mu[g]-\log\mathbb{E}_\nu[e^g]\right)\\
&\leq \sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)} \frac{1}{n}\sum_{i=1}^n g(X_i) - \log\left(\frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}\right) - \left(\mathbb{E}_\mu[g]-\log\mathbb{E}_\nu[e^g]\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left|\check{\mathsf{D}}_{\mathsf{DV},\mathcal{G}_k^{\mathsf{R}}(a)}(X^n,Y^n) - \mathsf{D}_{\mathsf{DV},\mathcal{G}_k^{\mathsf{R}}(a)}(\mu,\nu)\right|\right]\\
&\leq \mathbb{E}\left[\sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)}\left|\frac{1}{n}\sum_{i=1}^n g(X_i)-\mathbb{E}_\mu[g]\right|\right] + \mathbb{E}\left[\sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)}\left|\log\left(\frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}\right)-\log\mathbb{E}_\nu[e^g]\right|\right]\\
&\overset{(a)}{\leq} \mathbb{E}\left[\sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)}\left|\frac{1}{n}\sum_{i=1}^n g(X_i)-\mathbb{E}_\mu[g]\right|\right] + e^{3a(\|\mathcal{X}\|+1)}\mathbb{E}\left[\sup_{g\in\mathcal{G}_k^{\mathsf{R}}(a)}\left|\frac{1}{n}\sum_{i=1}^n e^{g(Y_i)}-\mathbb{E}_\nu[e^g]\right|\right]\\
&\overset{(b)}{\lesssim} a(\|\mathcal{X}\|+1)\big(e^{6a(\|\mathcal{X}\|+1)}+1\big)n^{-\frac{1}{2}}\int_0^1\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(3a(\|\mathcal{X}\|+1)\epsilon,\mathcal{G}_k^{\mathsf{R}}(a),\|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon,\\
&\overset{(c)}{\lesssim} a(\|\mathcal{X}\|+1)\big(e^{6a(\|\mathcal{X}\|+1)}+1\big)d^{\frac{3}{2}}n^{-\frac{1}{2}},
\end{aligned} \tag{C.2}$$

where

(a) is since $\bar{C}(|\mathcal{G}_k^{\mathsf{R}}(a)|,\mathcal{X}) \leq 3a(\|\mathcal{X}\|+1)$ and the Lipschitz constant of $\log x$ is bounded by $e^{3a(\|\mathcal{X}\|+1)}$ in $\left[e^{-\bar{C}(|\mathcal{G}_k^{\mathsf{R}}(a)|,\mathcal{X})}, e^{\bar{C}(|\mathcal{G}_k^{\mathsf{R}}(a)|,\mathcal{X})}\right]$;

(b) follows using steps akin to (A.34) and (Van Der Vaart and Wellner, 1996, Corollary 2.2.8);

(c) is due to (A.33).

Appendix D. CoD-Free Error Rate in the Unbounded Support Case

D.1 KL Divergence

Consider the NN class $\hat{\mathcal{G}}_k^{\mathsf{S}}(a,r)=\left\{g\mathbb{1}_{B_d(r)}:g\in\mathcal{G}_k^{\mathsf{S}}(a)\right\}$ (see Definition 7) and the following class of sub-Gaussian distributions:

$$\begin{aligned}&\hat{\mathcal{P}}_{\mathsf{KL}}^2(M,\ell):=\Big\{(\mu,\nu)\in\mathcal{P}_{\mathsf{KL}}^2\big(\mathbb{R}^d\big):\mu,\nu\in\mathcal{SG}(M),\ f_{\mathsf{KL}}\in\hat{\mathcal{I}}(M),\|f_{\mathsf{KL}}\|_{\ell,\mu}\leq M\Big\},\\&\hat{\mathcal{I}}(M):=\{f:S_1(f)\vee|f(0)|\leq M\}.\end{aligned}\tag{D.1}$$

Proposition 68 (KL CoD-free error bound) *Let $M\geq 0$, $\ell>1$ and $\ell^*=\ell/(\ell-1)$. Then, for $z_k=12\sqrt{\ell^* d}M^{3/2}(\log k)^{-1/2}$ and $r_k=M\vee 1+4\sqrt{dM\ell^*\log k}$,*

$$\sup_{(\mu,\nu)\in\hat{\mathcal{P}}_{\mathsf{KL}}^2(M,\ell)}\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k)}(X^n,Y^n)-\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]\lesssim_{d,M,\ell}(\log k)^{\frac{3}{2}}\Big(k^{-\frac{1}{2}}+k^{z_k}\ n^{-\frac{1}{2}}\Big).$$

Setting $k=n$ in the above bound gives an effective error bound $O\big(n^{-1/3}\big)$.

Proof Fix $(\mu,\nu)\in\hat{\mathcal{P}}_{\mathsf{KL}}^2(M,\ell)$. From (A.100), we have for $\mu,\nu\in\mathcal{SG}(M)$ and $r\geq M$ that

$$\mu\big(B_d^c(r)\big)\vee\nu\big(B_d^c(r)\big)\leq 2e^{\frac{-(r-M)^2}{16dM}}.\tag{D.2}$$

Then, it follows from (A.84) and (A.85) that for $r_k\geq M$,

$$\mathbb{E}_\mu\left[|f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r_k)}\right]\vee\mathbb{E}_\nu\left[|h_{\mathsf{KL}}\circ f_{\mathsf{KL}}|\,\mathbb{1}_{B_d^c(r_k)}\right]\lesssim_M e^{\frac{-(r_k-M)^2}{16dM\ell^*}}.$$

Moreover, $f_{\mathsf{KL}}\in\hat{\mathcal{I}}(M)$ implies $c_{\mathsf{B}}^\star\left(f_{\mathsf{KL}}|_{B_d(r_k)},B_d(r_k)\right)\leq Mr_k$ for $r_k\geq 1$.

Next, note that $\bar{C}(|\hat{\mathcal{G}}_k^{\mathsf{S}}(m_k,r_k)|,B_d(r_k))\leq 3Mr_k$, and $\bar{C}\big(\big|h'_{\mathsf{KL}}\circ\hat{\mathcal{G}}_k^{\mathsf{S}}(m_k,r_k)\big|,B_d(r_k)\big)\leq e^{3Mr_k}$. Also, similar to (A.33), we have

$$\int_0^1\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(3Mr_k\epsilon,\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k),\|\cdot\|_{2,\gamma}\right)}\mathrm{d}\epsilon\lesssim d^{\frac{3}{2}},\tag{D.3}$$

Then, (A.29) implies

$$\mathbb{E}\left[\left|\mathsf{D}_{\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k)}(\mu,\nu)-\hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k)}(X^n,Y^n)\right|\right]\lesssim_M d^{\frac{3}{2}}r_k e^{3Mr_k}n^{-\frac{1}{2}}.$$

Thus, we have similar to Lemma 63 (by using Theorem 8 for the sigmoid NN class) for $1\leq Mr_k\lesssim k^{(1-\rho)/2}$ for some $\rho>0$ that

$$\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k)}(X^n,Y^n)-\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]\lesssim_{d,M,\rho}r_k k^{-\frac{1}{2}}+r_k e^{3Mr_k}n^{-\frac{1}{2}}+e^{\frac{-(r_k-M)^2}{16dM\ell^*}}.$$

Taking $r_k=M\vee 1+4\sqrt{dM\ell^*\log k}$ and noting that $1\leq Mr_k\lesssim k^{1/4}$ (say), we obtain

$$\mathbb{E}\left[\left|\hat{\mathsf{D}}_{\hat{\mathcal{G}}_k^{\mathsf{S}}(Mr_k,r_k)}(X^n,Y^n)-\mathsf{D}_{\mathsf{KL}}\left(\mu\|\nu\right)\right|\right]$$

$$\lesssim_{d,M,\ell} k^{-\frac{1}{2}}(\log k)^{\frac{1}{2}} + k^{-1} + (\log k)^{\frac{3}{2}}\ e^{12M^{3/2}\sqrt{\ell^* d\log k}}\ n^{-\frac{1}{2}}$$
$$\lesssim_{d,M,\ell} k^{-\frac{1}{2}}(\log k)^{\frac{1}{2}} + k^{\frac{12\sqrt{\ell^*}dM^{3/2}}{\sqrt{\log k}}}(\log k)^{\frac{3}{2}}\ n^{-\frac{1}{2}}.$$

Taking supremum w.r.t. $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\mathsf{KL}}(M,\ell)$ yields the claim. ■

Remark 69 (CoD-free rate) *$\hat{\mathcal{P}}^2_{\mathsf{KL}}(M,\ell)$, for example, includes M-sub-Gaussian distributions (μ,ν) such that $\|f_{\mathsf{KL}}\|_{\ell,\mu} \leq M$ and $f_{\mathsf{KL}} \in \mathcal{L}^{\mathsf{B}}_{s_{\mathsf{B}},b}(\mathbb{R}^d)$ (for appropriate value of b), where $s_{\mathsf{B}} = \lfloor d/2 \rfloor + 2$ and $\mathcal{L}^{\mathsf{B}}_{s_{\mathsf{B}},b}(\mathbb{R}^d)$ is given in* (A.2)*. It also contains certain M-sub-Gaussian distributions (μ,ν) such that $f_{\mathsf{KL}} = c + f$ for some $c \in \mathbb{R}$ and $f \in \mathcal{S}(\mathbb{R}^d)$, where $\mathcal{S}(\mathbb{R}^d) = \left\{f \in \mathsf{C}^\infty(\mathbb{R}^d) : \sup_{x\in\mathbb{R}^d}\left|x^\alpha D^{\tilde{\alpha}} f(x)\right| < \infty,\ \forall \alpha,\tilde{\alpha} \in \mathbb{Z}^d_{\geq 0}\right\}$ is the Schwartz space of rapidly decreasing functions and $\alpha,\tilde{\alpha}$ are multi-indices of dimension d. An example would be some M-sub-Gaussian distributions (μ,ν) with $pq^{-1} = ce^{e^{-x^2}}$, where c is normalization constant (e.g., take q to be multivariate Gaussian, $p(x) = ce^{e^{-x^2}}q(x)$ and c such that $\int_{\mathbb{R}^d} q(x)\mathrm{d}x = 1$). We note that $f \in \mathcal{S}(\mathbb{R}^d)$ implies existence of Fourier transforms and Fourier inversion formula such that $S_1(f) < \infty$.*

D.2 χ^2 Divergence

With $\hat{\mathcal{I}}(M)$ as defined in (D.1), let

$$\hat{\mathcal{P}}^2_{\chi^2}(M,\ell) := \left\{(\mu,\nu) \in \mathcal{P}^2_{\chi^2}(\mathbb{R}^d) : \mu,\nu \in \mathcal{SG}(M),\ f_{\chi^2} \in \hat{\mathcal{I}}(M), \left\|f_{\chi^2}\right\|_{\ell,\mu} \leq M\right\}.$$

Proposition 70 (χ^2 CoD-free error bound) *Let $M \geq 0$, $\ell > 1$ and $\ell^* = \ell/(\ell-1)$. Then, for $r_k = M \vee 1 + 4\sqrt{dM\ell^*\log k}$,*

$$\sup_{(\mu,\nu)\in\hat{\mathcal{P}}^2_{\chi^2}(M,\ell)} \mathbb{E}\left[\left|\hat{\chi}^2_{\hat{\mathcal{G}}^{\mathsf{S}}_k(Mr_k,r_k)}(X^n,Y^n) - \chi^2(\mu\|\nu)\right|\right] \lesssim_{d,M,\ell} k^{-\frac{1}{2}}(\log k)^{\frac{1}{2}} + \log k\ n^{-\frac{1}{2}}.$$

Setting $k = n$ yields an effective error bound $\tilde{O}(n^{-1/2})$.

Proof Fix $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\chi^2}(M,\ell)$. From (D.2), (A.88) and (A.89), we have

$$\mathbb{E}_\mu\left[\left|f_{\chi^2}\right| \mathbb{1}_{B^c_d(r_k)}\right] \vee \mathbb{E}_\nu\left[\left|h_{\chi^2}\circ f_{\chi^2}\right| \mathbb{1}_{B^c_d(r_k)}\right] \lesssim_M e^{\frac{-(r_k-M)^2}{16dM\ell^*}}.$$

Note that $c^\star_{\mathsf{B}}\left(f_{\chi^2}|_{B_d(r_k)}, B_d(r_k)\right) \leq Mr_k$ for $r_k \geq 1$, $\bar{C}(|\hat{\mathcal{G}}^{\mathsf{S}}_k(m_k,r_k)|, B_d(r_k)) \leq 3Mr_k$, and $\bar{C}\left(\left|h'_{\chi^2}\circ\bar{\mathcal{G}}^{\mathsf{S}}_k(m_k,r_k)\right|, B_d(r_k)\right) \leq 1.5Mr_k+1$. Then we obtain similar to (A.87) using Theorem 8 and (D.3) that

$$\mathbb{E}\left[\left|\hat{\chi}^2_{\hat{\mathcal{G}}^{\mathsf{S}}_k(Mr_k,r_k)}(X^n,Y^n) - \chi^2(\mu\|\nu)\right|\right] \lesssim_M r_k d^{\frac{1}{2}}k^{-\frac{1}{2}} + r_k^2 dk^{-1} + e^{\frac{-(r_k-M)^2}{16dM\ell^*}} + d^{\frac{3}{2}}r_k^2 n^{-\frac{1}{2}}.$$

Setting $r_k = M \vee 1 + 4\sqrt{dM\ell^*\log k}$, and taking supremum w.r.t. $(\mu,\nu) \in \hat{\mathcal{P}}^2_{\chi^2}(M,\ell)$ proves the claim. ■

Remark 71 (CoD-free rate) *$\hat{\mathcal{P}}^2_{\chi^2}(M,\ell)$ contains certain M-sub-Gaussian distributions (μ,ν) such that $\left\|f_{\chi^2}\right\|_{\ell,\mu} \leq M$, and $f_{\chi^2} \in \mathcal{S}\big(\mathbb{R}^d\big) \cup \mathcal{L}^{\mathsf{B}}_{s_{\mathsf{B}},b}\big(\mathbb{R}^d\big)$ for appropriate value of b. In particular, this includes certain Gaussian distributions pairs $\big(\mathcal{N}(\mathsf{m}_p,\sigma_p^2\mathrm{I}_d),\mathcal{N}(\mathsf{m}_q,\sigma_q^2\mathrm{I}_d)\big)$ with $0<\sigma_p<\sigma_q\leq M$ and $\|\mathsf{m}_p\|\vee\|\mathsf{m}_q\|\leq M$. To see this, recall $f_{\chi^2}=2(pq^{-1}-1)$, and note $\sigma_q>\sigma_p$ ensures that $\left\|f_{\chi^2}\right\|_{\infty,\mathbb{R}^d}<\infty$ implying that $\left\|f_{\chi^2}\right\|_{\ell,\mu}<\infty$. Also, since pq^{-1} is again (upto constants) a Gaussian density, $\mathfrak{F}[pq^{-1}]$ exist which is again a Gaussian density (upto constants). Hence, $\mathfrak{F}[pq^{-1}]$ is integrable and this implies the Fourier inversion formula holds. Moreover, it is easy to verify that $S_1\big(pq^{-1}\big)<\infty$. Hence, such Gaussian pairs satisfies the conditions defining $\hat{\mathcal{P}}^2_{\chi^2}(M,\ell)$ for large enough M, and the claim follows.*

D.3 Squared Hellinger Distance

Let $\check{\mathcal{G}}^{\mathsf{S}}_{k,t}(a,r) := \big\{g\mathbb{1}_{B_d(r)} : g \in \mathcal{G}_k\big(k^{1/2}\log k, 2k^{-1}a, a, 0, \phi_{\mathsf{S}}\big)\big\}$, and

$$\hat{\mathcal{P}}^2_{\mathsf{H}^2}(M) := \left\{(\mu,\nu)\in\mathcal{P}^2_{\mathsf{H}^2}\big(\mathbb{R}^d\big) : \mu,\nu\in\mathcal{SG}(M),\ f_{\mathsf{H}^2}\in\hat{\mathcal{I}}(M), \left\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|_{\infty,\mathbb{R}^d}\leq M\right\},$$

where $\hat{\mathcal{I}}(m)$ is given in (D.1).

Proposition 72 (H^2 CoD-free error bound) *For $M\geq 0$, $m_k = Mr_k$ and $r_k = M\vee 1 + \sqrt{32dM\log k}$,*

$$\sup_{(\mu,\nu)\in\hat{\mathcal{P}}^2_{\mathsf{H}^2}(M)} \mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\check{\mathcal{G}}^{\mathsf{S}}_{k,m_k^{-1/2}}(m_k,r_k)}(X^n,Y^n)-\mathsf{H}^2(\mu,\nu)\right|\right] \lesssim_{d,M} k^{-\frac{1}{2}}\log k + n^{-\frac{1}{2}}\log k.$$

Setting $k=n$ yields an effective error bound $\tilde{O}(n^{-1/2})$.

Proof Fix $(\mu,\nu)\in\hat{\mathcal{P}}^2_{\mathsf{H}^2}(M)$. From (D.2), (A.94) and (A.95), we obtain

$$\mathbb{E}_\mu\left[|f_{\mathsf{H}^2}|\,\mathbb{1}_{B^c_d(r_k)}\right]\vee\mathbb{E}_\nu\left[|h_{\mathsf{H}^2}\circ f_{\mathsf{H}^2}|\,\mathbb{1}_{B^c_d(r_k)}\right]\leq e^{\frac{-(r_k-M)^2}{32dM}}.$$

We have $c^\star_{\mathsf{B}}\left(f_{\mathsf{H}^2}|_{B_d(r_k)}, B_d(r_k)\right)\leq Mr_k$ for $r_k\geq 1$, $\bar{C}(|\check{\mathcal{G}}^{\mathsf{S}}_{k,t}(m_k,r_k)|, B_d(r_k))\leq 3Mr_k$, and $\bar{C}\big(\big|h'_{\mathsf{H}^2}\circ\check{\mathcal{G}}^{\mathsf{S}}_{k,t}(m_k,r_k)\big|, B_d(r_k)\big)\leq t^{-2}$. Then, for k, r_k satisfying $r_k = o(k^{1/4})$, we have similar to (A.93) using Theorem 8 and (D.3) that

$$\mathbb{E}\left[\left|\hat{\mathsf{H}}^2_{\check{\mathcal{G}}^{\mathsf{S}}_{k,m_k^{-1/2}}(m_k,r_k)}(X^n,Y^n)-\mathsf{H}^2(\mu,\nu)\right|\right]\lesssim_M r_k^2 d^{\frac{1}{2}}k^{-\frac{1}{2}} + e^{\frac{-(r_k-M)^2}{32dM}} + d^{\frac{3}{2}}r_k^2 n^{-\frac{1}{2}}.$$

Setting $r_k = M\vee 1+\sqrt{32dM\log k}$ and taking supremum w.r.t. $(\mu,\nu)\in\hat{\mathcal{P}}^2_{\mathsf{H}^2}(M)$, we obtain the claim in the Proposition. ■

Remark 73 (CoD-free rate) *$\hat{\mathcal{P}}^2_{\mathsf{H}^2}(M)$ includes certain M-sub-Gaussian pairs (μ,ν) such that $\left\|pq^{-1}\right\|_{\infty,\mathbb{R}^d}\leq M$ and $qp^{-1}=(e^f+c)^2$ for some $f\in\mathcal{S}\big(\mathbb{R}^d\big)$, where c is the normalization constant to ensure that p and q are probability densities. To see this, note that $\sqrt{qp^{-1}}$ and $\sqrt{pq^{-1}}$ are both bounded on $\mathbb{R}^d$. Moreover, $f_{\mathsf{H}^2}=1-\sqrt{qp^{-1}}=-c+1-e^f$. Noting that $1-e^f\in\mathcal{S}\big(\mathbb{R}^d\big)$ if $f\in\mathcal{S}\big(\mathbb{R}^d\big)$, it follows as discussed in Remark 69 that $S_1(f_{\mathsf{H}^2})<\infty$, thus implying the claim.*

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 34th International Conference on Machine Learning*, 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 34th International Conference on Machine Learning*, pages 224–232, Sydney, Australia, Jul. 2017.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, Apr. 2017.

A. R. Barron. Neural net approximation. In *Proceedings of 7th Yale Workshop on Adaptive and Learning Systems*, CT, USA, May 1992.

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

A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 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, Jul. 2018.

T. B. Berrett and R. J. Samworth. Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv:1904.09347*, Apr. 2019.

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, Feb. 2019.

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

C. Domingo-Enrich and Y. Mroueh. Tighter sparse approximation bounds for ReLU neural networks. *arXiv:2110.03673*, Oct. 2021.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015.

Z. Goldfeld, K. Greenewald, and K. Kato. Asymptotic guarantees for generative modeling based on the smooth Wasserstein distance. In *Proceedings of the 34th Annual Conference on Advances in Neural Information Processing Systems*, Dec. 2020a.

Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, Jul. 2020b.

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 28th Annual Conference on Advances in Neural Information Processing Systems*, pages 2672–2680, Montréal, Canada, Dec. 2014.

M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328–1371, Mar. 2021.

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, Dec. 2020.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736*, Feb. 2019.

M. Kac, J. Kiefer, and J. Wolfowitz. On tests of normality and other tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*, 26(2):189–211, Jun. 1955.

B. Kalantari. An infinite family of bounds on zeros of analytic functions and relationship to Smale's bound. *Mathematics of Computation*, 74(250):841–852, May 2004.

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 29th Annual Conference on Advances in Neural Information Processing Systems*, pages 397–405, Montréal, Canada, Dec. 2015.

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

J. M. Klusowski and A. R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12):7649–7656, Oct. 2018.

A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman. Nonparametric estimation of Rényi divergence and friends. In *Proceedings of the 31st International Conference on Machine Learning*, pages 919–927, Beijing, China, Jun. 2014.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag Berlin Heidelberg, 1991.

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

Y. Makovoz. Uniform approximation by neural networks. *Journal of Approximation Theory*, 95(2):215–228, 1998.

K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero. Ensemble estimation of information divergence. *Entropy*, 20(8), Aug. 2018.

Y. Mroueh, I. Melnyk, P. Dognin, J. Ross, and T. Sercu. Improved mutual information estimation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 9009–9017, May 2021.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, Jun. 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, Oct. 2010.

S. Nietert, Z. Goldfeld, and K. Kato. Smooth p-Wasserstein distance: Structure, empirical approximation, and statistical applications. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8172–8183, Jul. 2021.

M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero. Direct estimation of information divergence using nearest neighbor ratios. In *Proceedings of the 2017 IEEE International Symposium on Information Theory*, pages 903–907, Jun. 2017.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th Annual Conference on Advances in Neural Information Processing Systems*, pages 271–279, Barcelona, Spain, Dec. 2016.

G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *Proceedings of the 8th International Conference on Learning Representations*, Apr.-May 2020.

F. Perez-Cruz. Kullback-Leibler divergence estimation of continuous distributions. In *Proceedings of the 2008 IEEE International Symposium on Information Theory*, pages 1666–1670, Toronto, ON, Canada, Jul. 2008.

B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 599–608, Barcelona, Spain, Jul. 2011.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, May 2019.

S. Singh and B. Póczos. Exponential concentration of a density functional estimator. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 3032–3040, Montreal, Canada, Dec. 2014a.

S. Singh and B. Póczos. Generalized exponential concentration inequality for renyi divergence estimation. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 333–341, Beijing, China, Jun. 2014b.

S. Singh and B. Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1225–1233, Barcelona, Spain, Dec. 2016.

S. Smale. Newton's method estimates from data at one point. In *Proceedings of The Merging of Disciplines: New Directions in Pure, Applied, and Computational Mathematics*. Springer New York, 1986.

S. Sreekumar, Z. Zhang, and Z. Goldfeld. Non-asymptotic performance guarantees for neural estimation of f-divergences. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 3322–3330, Apr. 2021.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550 – 1599, 2012.

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*, pages 7–16, San Diego, CA, US, Jun. 1990.

T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *Proceedings of 7th International Conference on Learning Representations*, New Orleans,US, May. 2019.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, Apr.-May 2018.

A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation and convergence of GANs under Besov IPM losses. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, volume 32, Vancouver,Canada, Dec. 2019.

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, Sep. 2005.

A. Wisler, K. Moon, and V. Berisha. Direct ensemble estimation of density functionals. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2866–2870, Apr. 2018.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, Oct. 1999.

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

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, Jul. 1995.

H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar. Adversarially learned anomaly detection. In *Proceedings of the 2018 IEEE International Conference on Data Mining*, pages 727–736, Nov. 2018.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discrimination-generalization tradeoff in GANs. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, Canada, Apr.-May 2018a.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28:113–130, Jan. 2018b.

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