Density estimation using the perceptron

Patrik Róbert Gerber prgerber@mit.edu
Tianze Jiang tjiang@mit.edu
Yury Polyanskiy yp@mit.edu
Rui Sun eruisun@mit.edu

Abstract. We propose a new density estimation algorithm. Given n i.i.d. samples from a distribution belonging to a class of densities on $\mathbb{R}^d$, our estimator outputs any density in the class whose "perceptron discrepancy" with the empirical distribution is at most $O(\sqrt{d/n})$. The perceptron discrepancy between two distributions is defined as the largest difference in mass that they place on any halfspace of $\mathbb{R}^d$. It is shown that this estimator achieves expected total variation distance to the truth that is almost minimax optimal over the class of densities with bounded Sobolev norm and Gaussian mixtures. This suggests that regularity of the prior distribution could be an explanation for the efficiency of the ubiquitous step in machine learning that replaces optimization over large function spaces with simpler parametric classes (e.g. in the discriminators of GANs).

We generalize the above to show that replacing the "perceptron discrepancy" with the generalized energy distance of Székeley-Rizzo [SR13] further improves total variation loss. The generalized energy distance between empirical distributions is easily computable and differentiable, thus making it especially useful for fitting generative models. To the best of our knowledge, it is the first example of a distance with such properties for which there are minimax statistical guarantees.

CONTENTS

1. INTRODUCTION

A standard step in many machine learning algorithms is to replace an (intractable) optimization over a general function space with an optimization over a large parametric class (most often neural networks). People do this in supervised learning for fitting classifiers, in variational inference [BKM17, ZBKM18] for applying ELBO, in variational autoencoders [KW19] for fitting the decoder, in Generative Adversarial Networks (GANs) [GPAM+14, ACB17] for fitting the discriminator, in diffusion models [SSDK+20, CCL+22] for fitting the score function, etc.

To be specific, let us focus on the example of GANs, which brought about the new era of density estimation in high-dimensional spaces. Suppose that we are given i.i.d. data $X_1, \dots, X_n \in \mathbb{R}^d$ from an unknown distribution ν and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$, which is our class of available "generators". The goal of the learner is to find $\arg\min_{\nu' \in \mathcal{G}} D(\nu', \nu)$, where D is some dissimilarity measure ("metric") between probability distributions. In the case of GANs this measure is the Jensen-Shannon divergence $\mathrm{JS}(p,q) \triangleq \mathrm{KL}(p\|\frac{1}{2}p + \frac{1}{2}q) + \mathrm{KL}(q\|\frac{1}{2}p + \frac{1}{2}q)$ and $KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence. As any f-divergence, JS has a variational form, cf. [PW23, Example 7.5]: $\mathrm{JS}(p,q) = \log 2 + \sup_{h:\mathbb{R}^d \to (0,1)} \mathbb{E}_p[h] + \mathbb{E}_q[\log(1-h)]$. With this idea in mind, we can now restate the objective of minimizing $\mathrm{JS}(\nu', \nu)$ as a game between a "generator" ν' and a "discriminator" h, i.e., the GAN's estimator is

$$\tilde{\nu} \in \arg\min_{\nu'} \sup_{h:\mathbb{R}^d \to (0,1)} \frac{1}{n} \sum_{i=1}^{n} h(X_i) + \mathbb{E}_{\nu'}[\log(1-h)], \tag{1.1}$$

where we also replaced the expectation over (the unknown) ν with its empirical version $\nu_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Subsequently, the idea was extended to other types of metrics, most notably the Wasserstein-GAN [ACB17], which defines

$$\tilde{\nu} \in \operatorname*{arg\,min}_{\nu' \in \mathcal{G}} \sup_{f \in \mathcal{D}} \left| \mathbb{E}_{Y \sim \nu'} f(Y) - \frac{1}{n}\sum_{i=1}^n f(X_i) \right| , \tag{1.2}$$

where the set of discriminators $\mathcal{D}$ is a class of Lipschitz functions (corresponding to the variational characterization of the Wasserstein-1 distance).

The final step to turn (1.1) or (1.2) into an algorithm is to relax the domain of the inner maximization ("discriminator") to a parametric class of neural network discriminators $\mathcal{D}$. Note that replacing $\sup_{h:\mathbb{R}^d \to (0,1)}$ with $\sup_{h \in \mathcal{D}}$ effectively changes the objective from minimizing the JS divergence to minimizing a "neural-JS", similar to how MINE [BBR+18] replaces the true mutual information with a "neural" one. This weakening is quite worrisome for a statistician. Indeed, while the JS divergence is a strong statistical distance (for example, it bounds total variation from above and from below [PW23, (7.39)]), the "neural-JS" is unlikely to possess any such properties.

How does one justify this restriction to a simpler class $\mathcal{D}$? A practitioner would say that while taking $\max_{h \in \mathcal{D}}$ restricts the power of the discriminator, the design of $\mathcal{D}$ is fine-tuned to picking up those features of the distributions that are relevant to the human eye[1]. A theoretician, instead, would appeal to universal approximation results about neural networks to claim that restriction to $\mathcal{D}$ is almost lossless.

The purpose of this paper is to suggest (and prove) a third explanation: the answer is in the *regularity* of ν itself. Indeed, we show that the restriction of discriminators to a very small class $\mathcal{D}$ in (1.1) results in almost no loss of (minimax) statistical guarantees, even if $\mathcal{D}$ is far from being a universal approximator. That is, the minimizing distribution $\tilde{\nu}$ selected with respect to a weak form of the distance enjoys almost minimax optimal guarantees with respect to the strong total variation distance, provided that the true distribution ν is regular enough. Phrased yet another way, even though the "neural" distance is very coarse and imprecise, and hence the minimizer selected with respect to it might be expected to only fool very naive discriminators, in reality it turns out to fool all (arbitrarily complex, but bounded) discriminators.

Let us proceed to a more formal statement of our results. One may consult Section 1.3 for notation. We primarily focus on two classes of distributions on $\mathbb{R}^d$: first, $\mathcal{P}_S(\beta, d, C)$ denotes the set of distributions supported on the d-dimensional unit ball $\mathbb{B}(0,1)$ that have a density with finite L^2 norm and whose $(\beta, 2)$-Sobolev norm (defined in (1.8)) is bounded by C; second, $\mathcal{P}_G(d) = \{\mu * \mathcal{N}(0,1) : \operatorname{supp}(\mu) \subseteq \mathbb{B}(0,1)\}$ is a (non-parametric) class of Gaussian mixtures with compactly supported mixing distribution. We remind the reader that the total variation distance has the variational form $\mathsf{TV}(p,q) = \sup_{h:\mathbb{R}^d \to [0,1]} \mathbb{E}_p h - \mathbb{E}_q h$. Our first result concerns the following class of discriminators:

$$\mathcal{D}_1 = \{x \mapsto \mathbb{1}\{x^\top v \geq b\} : v \in \mathbb{R}^d, b \in \mathbb{R}\} ,$$

which is the class of affine classifiers, and which can be seen as a single layer perceptron with a threshold non-linearity.

[1] Implying that whether or not total variation $\mathsf{TV}(\tilde{\nu}, \nu)$ is high is irrelevant as long as the generated images look "good enough" to humans.

THEOREM 1. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sup_{\nu \in \mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C_1 n^{-\frac{\beta}{2\beta+d+1}}\,, \tag{1.3}$$

where the estimator $\tilde{\nu}$ is defined in (1.2) with $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$. Similarly, for any $d \geq 1$ there exists a finite constant C_2 so that

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C_2 \frac{(\log(n))^{\frac{2d+2}{4}}}{\sqrt{n}}\,, \tag{1.4}$$

where the estimator $\tilde{\nu}$ is defined in (1.2) with $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{G} = \mathcal{P}_G(d)$.

Recall the classical result [IK83] which shows that the minimax optimal estimation rate in TV over the class $\mathcal{P}_S(\beta, d, C)$ equals $n^{-\beta/(2\beta+d)}$ (up to constant factors). Thus, the estimator in (1.3) is *almost* optimal, the only difference being that the dimension d is replaced by $d+1$. Similarly, for the Gaussian mixtures we reach the parametric rate up to a polylog factor[2].

The proof of Theorem 1 relies on a comparison inequality between total variation and the "perceptron discrepancy", or maximum halfspace distance, which we define as

$$\overline{d_H}(\mu, \nu) \triangleq \sup_{f \in \mathcal{D}_1} \{\mathbb{E}_\mu f - \mathbb{E}_\nu f\}. \tag{1.5}$$

Note first that $\overline{d_H} \leq \mathsf{TV}$ clearly holds since all functions in the class $\mathcal{D}_1$ are bounded by 1. For the other direction, by proving a generalization of the Gagliardo-Nirenberg-Sobolev inequality we derive the following comparisons.

THEOREM 2. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\mathsf{TV}(\mu, \nu)^{\frac{2\beta+d+1}{2\beta}} \leq C_1 \overline{d_H}(\mu, \nu) \tag{1.6}$$

holds for all $\mu, \nu \in \mathcal{P}_S(\beta, d, C)$. Similarly, for any $d \geq 1$ there exists a finite constant C_2 such that

$$\mathsf{TV}(\mu, \nu) \log\left(3 + \frac{1}{\mathsf{TV}(\mu, \nu)}\right)^{-\frac{d+1}{2}} \leq C_2 \overline{d_H}(\mu, \nu)$$

holds for all $\mu, \nu \in \mathcal{P}_G(d)$.

We remark that we also show (in Theorem 4) that the exponent $\frac{2\beta+d+1}{2\beta}$ in (1.6) is tight up to logarithmic factors.

While we believe that Theorem 1 provides theoretical proof for the efficacy of simple discriminators, it has several serious theoretical and practical deficiencies that we now address. First, the rate for the class $\mathcal{P}_S$ is not minimax optimal. In this regard, we show that by replacing the perceptron class $\mathcal{D}_1$ with a generalized perceptron

$$\mathcal{D}_\gamma = \{x \mapsto |x^\top v - b|^{\frac{\gamma-1}{2}} : v \in \mathbb{R}^d, b \in \mathbb{R}\}\,, \qquad \gamma \in (0, 2)$$

[2] For total variation the precise value of the minimax optimal polylog factor is at present unknown. However, for the L_2 distance the minimax rate is known, and in the course of our proofs (see (3.5)) we show that our estimator only loses a multiplicative factor of $\log(n)^{1/4}$ in loss compared to the optimal rate $\log(n)^{d/4}/\sqrt{n}$ derived in [KG22].

and taking an *average* over $\mathcal{D}_\gamma$ instead of a supremum, we are able to achieve a total variation rate of $n^{-\beta/(2\beta+d+\gamma)}$, thus coming arbitily close to minimax optimality[3] as $\gamma \to 0$. See Theorem 7 for details.

Second, from the implementation point of view, the density estimation algorithm behind Theorem 1 is completely impractical. Indeed, finding the halfspace with maximal separation between even two empirical measures is a nonconvex, non-differentiable problem and takes super-poly time in the dimension d assuming $\mathsf{P} \neq \mathsf{NP}$ [GR09], and $\omega(d^{\omega(\varepsilon^{-1})})$ time for ε-optimal agnostic learning between two densities assuming either SIVP or gapSVP [Tie23].

Even if we disregard the computational complexity, it is unclear how to find the exact minimizer $\tilde{\nu}$ of $\arg\min_{\nu'} \overline{d_H}(\nu', \nu_n)$. This concern is alleviated by the fact that any $\tilde{\nu}$ satisfying $\overline{d_H}(\tilde{\nu}, \nu_n) = \mathcal{O}\left(\sqrt{d/n}\right)$ will work, and thus only an approximate minimizer is needed. Taking this one step further, our proof proceeds by replacing the perceptron discrepancy $\overline{d_H}$ (defined with respect to the best perceptron) with an average version d_H defined in (2.2), for which the comparison in Theorem 2 still holds. Therefore, one doesn't even need to worry about finding an approximately optimal half-space, as a random half-space provides sufficient discriminatory power.

We discover that the average perceptron discrepancy d_H exactly equals Székely-Rizzo's energy distance $\mathcal{E}_1$ (Definition 1, [SR13]), defined as

$$\mathcal{E}_1^2(\mu, \nu) \triangleq \mathbb{E}\left[2\|X - Y\| - \|X - X'\| - \|Y - Y'\|\right], \qquad (X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}, \tag{1.7}$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Here our theoretical result (Theorem 7 with $\gamma = 1$) is that minimizing $\min_{\nu'} \mathcal{E}_1(\nu', \nu_n)$ gives a density estimator with rates over $\mathcal{P}_S$ and $\mathcal{P}_G$ as given in Theorem 1.

From the algorithmic point of view our message is the following. If one has access to a parametric family of generators sampling from ν_θ with $\theta \in \mathbb{R}^p$ being the parameter and assuming that one can compute ∇_θ of the generator forward pass, e.g., via pushforward of a reference distribution under a smooth transport map or neural network-based models [WM22, MRWZ23], then one can fit θ to the empirical sample ν_n by running stochastic gradient descent steps:

- sample m samples from ν_θ and form the empirical distribution ν'_m,
- compute the loss $\mathcal{E}_1(\nu'_m, \nu_n)$ and backpropagate the gradient with respect to θ,
- update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{E}_1(\nu'_m, \nu_n)$.

Note that the computation of $\mathcal{E}_1(\nu'_m, \nu_n)$ according to (1.7) requires $\mathcal{O}(n^2 + m^2)$ steps and is friendly to gradient evaluations.

1.1 Contributions

To summarize, our main contributions are as follows. We show that β-smooth distributions and Gaussian mixtures that are far apart in total variation distance must possess a halfspace on which their mass is substantially different (Theorems 2 and 3). We derive similar results for discrete distributions in Theorem 5.

We apply the separation results to density estimation problems, showing that an ERM density estimator nearly attains the minimax optimal density estimation rate with respect to TV over the class of smooth densities, Gaussian mixtures (Theorems 1 and 7) and discrete distributions (Proposition 8).

[3] More precisely, within a polylog in n.

In Section 2 we show that the average halfspace separation distance d_H is equal up to constant to the energy distance $\mathcal{E}_1$ (Proposition 2), which has many equivalent expressions: as a weighted L^2-distance between characteristic functions (Proposition 3), as the sliced Cramér-2 distance (Proposition 4), as an IPM/MMD/energy distance (Section 2.3), and as the L^2-norm of the Riesz potential (Section 2.5).

We generalize the average halfspace distance d_H to include an exponent $\gamma \in (0, 2)$, corresponding to the generalized energy distance $\mathcal{E}_\gamma$ (Proposition 4). Consequently, we discover that if instead of thresholded linear features $\mathbb{1}\{v^\top x > b\}$ we use the non-linearity $|v^\top x - b|^\gamma$, smooth distributions and Gaussian mixtures can be separated even better (Theorem 3). Combined with the fact that $\mathcal{E}_\gamma$, similarly to d_H, decays between population and sample measures at the parametric rate (Lemma 1), the ERM for $\mathcal{E}_\gamma$ reduces the slack in the density estimation rate, almost achieving minimax optimality. This result, combined with its strong approximation properties, supports its use in modern generative models (e.g. [HJA20, GPAM$^+$14, RBL$^+$22, RDN$^+$]).

Finally, Proposition 10 shows that recent work applying $\overline{d_H}$ for two-sample testing is sub-optimal over the class of smooth distributions in the minimax sense.

1.2 Related work

In the statistics literature, an estimator of the form (1.2) appears in the famous work [Yat85]. Instead of indicators of halfspaces, they consider the class of discriminators $\mathcal{Y} \triangleq \{\mathbb{1}_{\mathrm{d}\nu_i/\mathrm{d}\nu_j \geq 1} : 1 \leq i, j \leq N(\epsilon_n, \mathcal{G})\}$, where $\nu_1, \ldots, \nu_{N(\epsilon_n, \mathcal{G})}$ forms a minimal ϵ_n-TV covering of the class $\mathcal{G}$ and $N(\epsilon_n, \mathcal{G})$ is the so-called covering number. Writing $d_Y(\mu, \mu') = \sup_{f \in \mathcal{Y}}(\mathbb{E}_\mu f - \mathbb{E}_{\mu'} f)$, it is not hard to prove that $|\mathsf{TV} - d_Y| = \mathcal{O}(\epsilon_n)$ on $\mathcal{G} \times \mathcal{G}$ and that $\mathbb{E} d_Y(\nu, \nu_n) \lesssim \sqrt{\log N(\epsilon_n, \mathcal{G})/n}$ by a union bound coupled with a binomial tail inequality. From here $\mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \lesssim \min_{\epsilon > 0}[\sqrt{\log N(\epsilon, \mathcal{G})/n} + \epsilon]$ follows by the triangle inequality (here $\tilde{\nu}$ is defined as in (1.2) with $\mathcal{D} = \mathcal{Y}$). Note that in contrast to our perceptron discrepancy $\overline{d_H}$, Yatracos' estimator attains the optimal rate on $\mathcal{G} = \mathcal{P}_S$, corresponding to the choice $\epsilon_n \asymp n^{-\beta/(2\beta+d)}$.

Several other related works such as [SUL$^+$18, Lia21] study the problem of minimax density estimation over classical smoothness classes with respect to Integral Probability Metrics (IPMs) $d_{\mathcal{D}}(\mathbb{P}, \mathbb{Q}) \triangleq \sup_{f \in \mathcal{D}} |\mathbb{E}_{\mathbb{P}} f - \mathbb{E}_{\mathbb{Q}} f|$. In particular, these works seek estimators $\tilde{\nu}$ such that $d_{\mathcal{D}}(\tilde{\nu}, \nu)$ is small for some discriminator $\mathcal{D}$. Note some crucial differences to our work: first, we evaluate performance with respect to total variation in Theorem 1 which bears more interest both theoretically and empirically; second, we restrict our attention to estimators $\tilde{\nu}$ attained by ERM which is more commonly used in practice, while corresponding results of [SUL$^+$18, Lia21] consider classical orthogonal projection estimators.

A paper closer in spirit to ours is [BMR18] whose authors study comparison inequalities between the Wasserstein distance W_1 and the IPM d_{relu} defined by the discriminator class $\mathcal{D} = \{x \mapsto \mathrm{Relu}(x^\top v + b) : b, \|v\| \leq 1\}$. They show [BMR18, Theorem 3.1] that $\sqrt{\kappa/d} W_1 \lesssim d_{relu} \lesssim W_1$ for Gaussian distributions with mean in the unit ball, where κ is an upper bound on their condition numbers and d is the dimension. They obtain results for other distribution classes (Gaussian mixtures, exponential families), but for each of these they use a different class of discriminators that is adapted to the problem. In contrast, we mainly study the discriminator class $\mathcal{D}_1 = \{x \mapsto \mathbb{1}\{x^\top v \geq b\} : \|v\| \leq 1, b \in \mathbb{R}\}$ and are able to derive novel comparisons to TV for smooth distributions, Gaussian mixtures and discrete distributions. In addition, we prove the (near)-optimality of our results and also derive nonparametric estimation rates for the corresponding GAN density estimators.

Independent of this work, recent results by [PCGT23] investigate the halfspace separability of distributions for the setting of two-sample testing. However, their focus was on the asymptotic power of the test as the number of samples grows to infinity. Our lower bound construction presented in Appendix C proves that their proposed test is sub-optimal in the minimax setting. See Section 5 for a more detailed discussion.

1.3 Notation

The symbols $\mathcal{O}, o, \Theta, \Omega, \omega$ follow the conventional "big-O" notation, and $\tilde{\mathcal{O}}, \tilde{o}$ hide polylogarithmic factors. The Gamma function is denoted as $\Gamma(\cdot)$. Given a vector $x \in \mathbb{R}^d$ we write $\|x\|$ for its Euclidean norm and $\langle x, y\rangle \triangleq x^\top y$ for the Euclidean inner product of $x, y \in \mathbb{R}^d$. We write $\mathbb{B}(x, r) \triangleq \{y \in \mathbb{R}^d : \|x-y\| \leq r\}$, $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d : \|x\| = 1\}$ and $\mathrm{d}\sigma$ for the unnormalized surface measure on $\mathbb{S}^{d-1}$. The surface area of a unit $(d-1)$-sphere is also written as $\mathrm{vol}_{d-1}(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma\big(\frac{d}{2}\big)$. The convolution between functions/measures is denoted by $*$. We write $L^p(\mathbb{R}^d)$ for the space of (equivalence classes of) functions $\mathbb{R}^d \to \mathbb{C}$ that satisfy $\|f\|_p \triangleq \big(\int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}x\big)^{1/p} < \infty$. The space of all probability distributions on $\mathbb{R}^d$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. For a signed measure ν we write $\mathrm{supp}(\nu)$ for its support and $M_r(\nu) \triangleq \int \|x\|^r \mathrm{d}|\nu|(x)$ for its r'th absolute moment. Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$ we write $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \triangleq \sup_{A \subseteq \mathbb{R}^d} [\mathbb{P}(A) - \mathbb{Q}(A)]$ for the total variation distance, where the supremum is over all measurable sets.

Given a function $f \in L^1(\mathbb{R}^d)$, define its Fourier transform as

$$\widehat{f}(\omega) \triangleq \mathcal{F}[f](\omega) \triangleq \int_{\mathbb{R}^d} e^{-i\langle x, \omega\rangle} f(x) \mathrm{d}x.$$

Given a finite signed measure ν on $\mathbb{R}^d$, define its Fourier transform as $\mathcal{F}[\nu](\omega) \triangleq \int_{\mathbb{R}^d} e^{-i\langle \omega, x\rangle} \mathrm{d}\nu(x)$. We extend the Fourier transform to $L^2(\mathbb{R}^d)$ and tempered distributions in the standard manner. Given $f \in L^2(\mathbb{R}^d)$ and $\beta > 0$, define its homogenous Sobolev seminorm of order $(\beta, 2)$ as

$$\|f\|_{\beta,2}^2 \triangleq \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\widehat{f}(\omega)|^2 \mathrm{d}\omega. \tag{1.8}$$

Further, we define two specific classes of functions of interest as follows: $\mathcal{P}_S(\beta, d, C)$ is a set of smooth densities while $\mathcal{P}_G(d)$ is a set of all Gaussian mixtures with support in the unit ball, formally

$$\begin{aligned} \mathcal{P}_S(\beta, d, C) &\triangleq \{\mu \in \mathcal{P}(\mathbb{R}^d) : \mathrm{supp}(\mu) \subseteq \mathbb{B}(0,1), \mu \text{ has density } p \text{ with } \|p\|_{\beta,2} \leq C\}, \\ \mathcal{P}_G(d) &\triangleq \{\nu * \mathcal{N}(0, I_d) : \nu \in \mathcal{P}(\mathbb{R}^d), \mathrm{supp}(\nu) \subseteq \mathbb{B}(0,1)\}. \end{aligned} \tag{1.9}$$

Assumption 1. *Throughout the paper we assume that C in the definition of $\mathcal{P}_S(\beta, d, C)$ is large enough relative to β and d, such that $\mathcal{P}_S(\beta, d, C/2)$ is non-empty.*

We use $\lesssim$, $\gtrsim$ and $\asymp$ to hide irrelevant multiplicative constants that depend on d, β, C only.

1.4 Structure

The structure of the paper is as follows. In Section 2 we introduce the generalized energy distance, the main object of our study. We show how it relates to the perceptron discrepancy $\overline{d_H}$ and its relaxation d_H; we record equivalent formulations of the generalized energy distance, one of which is a novel "sliced-distance" form. In Section 3, we present our main technical results on comparison inequalities between total variation and the energy distance. In Section 4 we analyse the density estimator that minimizes the empirical energy distance, and prove Theorem 1 and Theorem 2 in Section 4.1.1. In Section 5 we show that the use of $\overline{d_H}$ for two sample testing results in suboptimal performance. We conclude in Section 6. All omitted proofs and auxiliary results are deferred to the Appendix.

2. THE GENERALIZED ENERGY DISTANCE

Given two probability distributions μ, ν on $\mathbb{R}^d$ with finite γ'th moment, the generalized energy distance of order $\gamma \in (0,2)$ between them is defined as

$$\mathcal{E}_\gamma(\mu,\nu) = \mathbb{E}\Big[2\|X-Y\|^\gamma - \|X-X'\|^\gamma - \|Y-Y'\|^\gamma\Big], \qquad \text{where } (X,X',Y,Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}. \quad (2.1)$$

As we alluded to in the introduction, the proof of Theorems 1 and 2 becomes possible once we relax the supremum in the definition of $\overline{d_H}$ to an *unnormalized* average over halfspaces. In Section 2.1 we discuss this relaxation in more detail and identify a connection to the energy distance $\mathcal{E}_1$ defined above in (2.1). Motivated by this, we study the (generalized) energy distance and give multiple equivalent characterizations of it from Section 2.2 to Section 2.5.

2.1 From perceptron to energy distance

Our first goal is to connect the study of $\overline{d_H}$ to the study of $\mathcal{E}_\gamma$ with $\gamma = 1$. To achieve this, we introduce an intermediary, the "average" perceptron discrepancy d_H. Given two probability distributions μ, ν on $\mathbb{R}^d$, we define

$$d_H(\mu,\nu) \triangleq \sqrt{\int_{v\in\mathbb{S}^{d-1}} \int_{b\in\mathbb{R}} \left(\int_{\langle v,x\rangle \geq b} \mathrm{d}\mu(x) - \mathrm{d}\nu(x)\right)^2 \mathrm{d}b\mathrm{d}\sigma(v)}, \quad (2.2)$$

where σ denotes the surface area measure.

If the two distributions μ, ν are supported on a compact set, then the overall definition can indeed be regarded as an average of the (mean squared) perceptron discrepancy, because the integrals over b and v only range over bounded sets. However, in general, the integrals in the definition of d_H are not normalizable and that is why we put "average" in quotes. Nevertheless, we have the following comparisons between d_H and $\overline{d_H}$.

PROPOSITION 1. *For any $\beta > 0$, $d \geq 1$, $C > 0$, and for all $\mu, \nu \in \mathcal{P}_S(\beta, d, C)$, we have*

$$\sqrt{\frac{\Gamma(d/2)}{4\pi^{d/2}}} d_H(\mu,\nu) \leq \overline{d_H}(\mu,\nu). \quad (2.3)$$

Moreover, for all $d \geq 1$, there exists a finite constant C_1 such that for all $\mu, \nu \in \mathcal{P}_G(d)$,

$$\frac{d_H(\mu,\nu)}{\log(3 + 1/d_H(\mu,\nu))^{1/4}} \leq C_1 \overline{d_H}(\mu,\nu). \quad (2.4)$$

PROOF. The proof of (2.3) is immediate after noting that all distributions in $\mathcal{P}_S(\beta, d, C)$ are supported on the d-dimensional unit ball and that $\int_{v\in\mathbb{S}^{d-1}} \int_{-1}^{1} \mathrm{d}b\mathrm{d}\sigma(v) = 4\pi^{d/2}/\Gamma(d/2)$. Thus, we focus on the Gaussian mixture case. Write $\mu - \nu = \tau * \phi$ where ϕ denotes the density of the standard Gaussian $\mathcal{N}(0, I_d)$ and $\tau \in \mathcal{P}(\mathbb{R}^d)$ is the difference of the two implicit mixing measures. For any $R > 0$, we have

$$\begin{aligned}\overline{d_H}(\mu,\nu) &\geq \sup_{v\in\mathbb{S}^{d-1}, |b|\leq R} \int_{\langle x,v\rangle\geq b} (\tau*\phi)(x)\mathrm{d}x \\ &\geq \sqrt{\frac{1}{2R\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})}\int_{\mathbb{S}^{d-1}}\int_{|b|\leq R}\left(\int_{\langle x,v\rangle\geq b}(\tau*\phi)(x)\mathrm{d}x\right)^2 \mathrm{d}b\mathrm{d}\sigma(v)}.\end{aligned}$$

Now, since τ is supported on a subset of $\mathbb{B}(0,1)$ by definition, for any $v \in \mathbb{S}^{d-1}$ and $R \geq 2$ we have the bound

$$\begin{aligned}\int_{|b|>R}\left(\int_{\langle x,v\rangle\geq b}\int_{\mathbb{R}^d}\phi(x-y)\mathrm{d}\tau(y)\mathrm{d}x\right)^2\mathrm{d}b &\leq \int_{|b|>R}\left(\int_{\langle x,v\rangle\geq |b|}\exp(-(\|x\|-1)^2/2)\mathrm{d}x\right)^2\mathrm{d}b \\ &\leq \int_{|b|>R}\left(\int_{\|x\|\geq|b|}\exp(-\|x\|^2/8)\mathrm{d}x\right)^2\mathrm{d}b \\ &\lesssim \exp(-\Omega(R^2)),\end{aligned}$$

where we implicitly used that $\int \mathrm{d}\tau = 0$ as τ is the difference of two probability distributions. Choosing $R \asymp \sqrt{\log(3 + 1/d_H(\mu,\nu))}$ concludes the proof. □

Proposition 1 implies that to obtain a comparison between TV and $\overline{d_H}$, specifically for lower bounding $\overline{d_H}$, it suffices to consider the relaxation d_H instead. The next observation we make is that d_H is in fact equal, up to constant, to the energy distance.

Proposition 2. *Let μ, ν be probability distributions on $\mathbb{R}^d$ with finite mean. Then*

$$d_H(\mu,\nu) = \frac{\pi^{(d-1)/4}}{\sqrt{\Gamma(\frac{d+1}{2})}}\mathcal{E}_1(\mu,\nu).$$

Proof. This is a direct implication of (2.5) and (2.7). We defer the proof to the more general Proposition 4. □

As will be clear from the rest of the paper, it does pay off to study $\mathcal{E}_\gamma$ for general γ, even though so far we only justified its relevance to the results stated in the introduction for the case of $\gamma = 1$. With this in mind, we proceed to study various properties of the *generalized* energy distances $\{\mathcal{E}_\gamma\}_{\gamma\in(0,2)}$.

2.2 The Fourier form

The formulation of the generalized energy distance that we rely on most heavily in our proofs is the following.

Proposition 3 ([SR13, Proposition 2]). *Let $\gamma \in (0,2)$ and let μ,ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then,*

$$\mathcal{E}_\gamma^2(\mu,\nu) = F_\gamma(d)\int_{\mathbb{R}^d}\frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega, \tag{2.5}$$

where we define $F_\gamma(d) = \frac{\gamma 2^{\gamma-1}\Gamma\left(\frac{d+\gamma}{2}\right)}{\pi^{d/2}\Gamma(1-\frac{\gamma}{2})}$.

Remark 1. Note that $F_\gamma(d) = \Theta\left(\gamma(2-\gamma)\Gamma\left(\frac{d+\gamma}{2}\right)\pi^{-d/2}\right)$ up to a universal constant.

This shows that the generalized energy distance is a weighted L^2 distance in Fourier space. The fact that $\mathcal{E}_\gamma$ is a valid metric on probability distributions with finite γ'th moment is a simple consequence of Proposition 3. Another straightforward consequence of the above characterization is the fact that $\mathcal{E}_\gamma$ decays at the parametric rate between empirical and population measures. Recall that $M_t(\nu)$ denotes the t'th absolute moment of the measure ν.

Lemma 1. *Let ν be a probability distribution on $\mathbb{R}^d$ and let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for an i.i.d. sample $X_1, \dots, X_n \overset{iid}{\sim} \nu$. Then, for any $\gamma \in (0,2)$,*

$$\mathbb{E}\mathcal{E}_\gamma^2(\nu, \nu_n) \le \frac{10 d^{\gamma/2} M_\gamma(\nu)}{n}.$$

Moreover, for a high-probability bound when ν is compactly supported, see Lemma 4.

Proof. Let $X \sim \nu$. We use Proposition 3 as well as Tonelli's theorem to interchange integrals to obtain

$$\begin{aligned}
\frac{1}{F_\gamma(d)}\mathbb{E}\mathcal{E}_\gamma^2(\nu,\nu_n) &= \int_{\mathbb{R}^d} \frac{\mathbb{E}|\widehat{\nu} - \widehat{\nu}_n|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \\
&= \frac{1}{n}\int_{\mathbb{R}^d} \frac{\operatorname{var}(\cos\langle X,\omega\rangle) + \operatorname{var}(\sin\langle X,\omega\rangle)}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \\
&\le \frac{1}{n}\mathbb{E}\int_{\mathbb{R}^d} \frac{(\cos\langle X,\omega\rangle - 1)^2 + (\sin\langle X,\omega\rangle)^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \\
&\le \frac{16\pi^{d/2}M_\gamma(\nu)}{\Gamma(d/2)\gamma(2-\gamma)n},
\end{aligned}$$

where the last line follows by Lemma 6. Multiplying both sides by $F_\gamma(d)$, using that $(2-\gamma)\Gamma(1-\gamma/2) \ge 1.77$ by point 3. of Lemma 10 gives

$$\mathbb{E}\mathcal{E}_\gamma^2(\nu,\nu_n) \le \frac{4.52 \times 2^\gamma M_\gamma(\nu)}{n}\frac{\Gamma\left(\frac{d+\gamma}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

The conclusion now follows by the first point of Lemma 10. □

2.3 The MMD and IPM forms

Another interpretation of Proposition 3 is through the theory of *Maximum Mean Discrepancy* (MMD). Given a set $\mathcal{X}$ and a positive semidefinite kernel $k : \mathcal{X}^2 \to \mathbb{R}$, there is a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ consisting of the closure of the linear span of $\{k(x,\cdot), x \in \mathcal{X}\}$ with respect to the inner product $\langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_k} = k(x,y)$.

For a probability distribution μ on $\mathcal{X}$, define its kernel embedding as $\theta_\mu = \int_{\mathbb{R}^d} k(x,\cdot)\mathrm{d}\mu(x)$. As shown in [MFS+17, Lemma 3.1], the kernel embedding θ_μ exists and belongs to the RKHS $\mathcal{H}_k$ if

$\mathbb{E}[\sqrt{k(X,X')}] < \infty$ for $(X,X') \sim \mu^{\otimes 2}$ — as is the case for our kernel defined later in Equation (2.6). Then, given two probability distributions μ and ν, the MMD measures their distance in the RKHS by

$$\mathrm{MMD}_k(\mu,\nu) \triangleq \|\theta_\mu - \theta_\nu\|_{\mathcal{H}_k}.$$

We refer the reader to [SS01, MFS+17] for more details on the underlying theory. MMD has a closed form thanks to the reproducing property:

$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}\Big[k(X,X') + k(Y,Y') - 2k(X,Y)\Big],$$

where $(X,X',Y,Y') \sim \mu^2 \otimes \nu^2$. Moreover, it also follows that MMD is an *Integral Probability Metric (IPM)* where the supremum is over the unit ball of the RKHS $\mathcal{H}_K$:

$$\mathrm{MMD}_k(\mu,\nu) = \sup_{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \le 1} \mathbb{E}[f(X) - f(Y)].$$

In our case, we can define the kernel

$$k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma \tag{2.6}$$

for $\gamma \in (0,2)$, which corresponds to the covariance operator of fractional Brownian motion. For a proof of the nontrivial fact that k_γ above is positive definite see for example [SSGF13]. With the choice of k_γ it follows trivially from its definition that the generalized energy distance $\mathcal{E}_\gamma$ is equal to the MMD with kernel k_γ, i.e.

$$\mathcal{E}_\gamma(\mu,\nu) = \mathrm{MMD}_{k_\gamma}(\mu,\nu)$$

for all distributions μ,ν with finite γ'th moment. It is noteworthy that while $\overline{d_H}$ is by definition an IPM, so is its averaged version d_H.

2.4 The sliced form

Another equivalent characterization of the generalized energy distance is in the form of a *sliced* distance. Sliced distances are calculated by first choosing a random direction on the unit sphere, and then computing a one-dimensional distance in the chosen direction between the projections of the two input distributions. For $\gamma \in (0,2)$ define the function

$$\psi_\gamma(x) = \begin{cases} |x|^{(\gamma-1)/2} & \text{for } \gamma \neq 1 \\ \mathbb{1}\{x \ge 0\} & \text{otherwise.} \end{cases}$$

The following result, to the best of our knowledge, has not appeared in prior literature except for the case of $\gamma = 1$.

Proposition 4. *Let $\gamma \in (0,2)$ and let μ,ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then for $(X,Y) \sim \mu \otimes \nu$ we have*

$$\mathcal{E}_\gamma^2(\mu,\nu) = \frac{1}{S_\gamma} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X,v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y,v\rangle - b)\Big]^2 \mathrm{d}b\mathrm{d}\sigma(v). \tag{2.7}$$

where $S_\gamma = \frac{\pi^{\frac{d}{2}+1}\Gamma(1-\frac{\gamma}{2})}{\gamma 2^{\gamma+1}\Gamma(\frac{d+\gamma}{2})\cos^2(\frac{\pi(\gamma-1)}{4})\Gamma(\frac{1-\gamma}{2})^2}$ *when* $\gamma \neq 1$ *and* $S_1 = \frac{\pi^{\frac{d-1}{2}}}{4\Gamma(\frac{d+1}{2})}$.

The proof of Proposition 4 hinges on computing the Fourier transform of the function ψ_γ, which can be interpreted as a tempered distribution. We point out a special property of the integral on the right hand side of (2.7). After expanding the square, one finds that the individual terms in the sum are not absolutely integrable for $\gamma \neq 1$. However, due to cancellations, the quantity is still well defined.

As claimed, Proposition 4 allows us to interpret $\mathcal{E}_\gamma$ as a *sliced* probability divergence (using the language of [NDC+20]). Given $v \in \mathbb{S}^{d-1}$, write $\theta_v = \langle v, \cdot \rangle$ and $\theta_v \# \nu = \nu \circ \theta_v$ for the pushforward of ν under θ_v, we have

$$S_\gamma(d)\mathcal{E}_\gamma^2(\mu,\nu) = S_\gamma(1)\int_{\mathbb{S}^{d-1}} \mathcal{E}_\gamma(\theta_v\#\mu, \theta_v\#\nu)\mathrm{d}\sigma(v).$$

We may also observe that $d_H \asymp \mathcal{E}_1$ is equal to the sliced Cramér-2 distance[4] up to constant, which has been studied by both theoretical and empirical works [KTS+18, KNSS22].

2.5 The Riesz potential form

The generalized energy distance can also be linked to the Riesz potential [Lan72, Chapter 1.1], which is the inverse of the fractional Laplace operator. Given $0 < s < d$, the Riesz potential $I_s f$ of a compactly supported signed measure f on $\mathbb{R}^d$ is defined (in a weak sense) by

$$I_s f = f * K_s,$$

where $K_s(x) = \frac{1}{c_s}\frac{1}{\|x\|^{d-s}}$ and $c_s = \pi^{d/2}2^s\frac{\Gamma(s/2)}{\Gamma((d-s)/2)}$. The Fourier transform of the Riesz kernel is given by $\widehat{K_s}(\omega) = \|\omega\|^{-s}$ (as a tempered distribution). Thus the following proposition is derived by setting $s = \frac{d+\gamma}{2}$ and using the Fourier form Proposition 3 of the energy distance.

Proposition 5. *Let $\gamma \in (0, \min\{d,2\})$ and let μ, ν be compactly supported probability distributions on $\mathbb{R}^d$, then*

$$\mathcal{E}_\gamma(\mu,\nu) = (2\pi)^{d/2}\sqrt{F_\gamma(d)}\|I_{\frac{d+\gamma}{2}}(\mu-\nu)\|_2. \tag{2.8}$$

3. MAIN COMPARISON: TV VERSUS ENERGY

After considering the connection of the perceptron discrepancy $\overline{d_H}$ to the generalized energy distance in Section 2, we turn to some of our main technical results, which provide novel quantitative comparisons between $\{\mathcal{E}_\gamma\}_{\gamma\in(0,2)}$ and the total variation distance. In Section 3.1 we show that the generalized energy distance is upper bounded by total variation for compactly supported distributions. In Section 3.2 we study the two distribution classes that we have introduced, namely smooth distributions and Gaussian mixtures. Finally, in Section 3.3 we turn to the case of discrete distributions, which requires alternative techniques.

3.1 Upper bound – compactly supported distributions

Note that we (obviously) always have $\overline{d_H}(\mu,\nu) \le \mathsf{TV}(\mu,\nu)$ for arbitrary probability measures μ and ν. Moreover, for distributions supported on a unit ball we also have $d_H(\mu,\nu) \lesssim \overline{d_H}(\mu,\nu)$. Therefore, by the identification of d_H and $\mathcal{E}_1$ (Proposition 2), we can see that for distributions with bounded support, we always have $\mathcal{E}_1(\mu,\nu) \lesssim \mathsf{TV}(\mu,\nu)$. The next result generalizes this estimate for all $\mathcal{E}_\gamma$, not just $\gamma = 1$.

[4] The Cramér-p distance is simply the L^p distance between cumulative distribution functions.

Proposition 6. *For any dimension $d \geq 1$ and $\gamma \in (0, \min\{d, 2\})$ there exists a finite constant c such that for any two probability distributions μ, ν supported on the unit ball we have*

$$\mathcal{E}_\gamma(\mu, \nu) \leq c\mathsf{TV}(\mu, \nu)\,.$$

Proof. Assume first that both μ, ν are absolutely continuous, and let $f(x) = \frac{\mathrm{d}\mu}{\mathrm{d}x} - \frac{\mathrm{d}\nu}{\mathrm{d}x}$ and $\epsilon = \mathsf{TV}(\mu, \nu)$. By Equation (2.8), it suffices to upper bound $\|I_s f\|_2$ for $s = (d+\gamma)/2$. First we decompose $\|I_s f\|_2$ as $\|I_s f\|_2 \leq \|I_s f \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 + \|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2$ by the triangle inequality. To estimate $\|I_s f \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2$, let $f^+(x) = \max\{f(x), 0\}$ and $f^-(x) = \max\{-f(x), 0\}$ so that $\int f^+(x)\mathrm{d}x = \int f^-(x)\mathrm{d}x = \epsilon$. Since $\operatorname{supp}(f) \subseteq \mathbb{B}(0,1)$, for all $\|x\| > 2$ it follows that

$$\begin{aligned}
I_s f(x) &= \frac{1}{c_s}\int \frac{f^+(y) - f^-(y)}{\|x-y\|^{d-s}}\mathrm{d}y \\
&\leq \frac{1}{c_s}\left(\frac{\int f^+(y)\mathrm{d}y}{(\|x\|-1)^{d-s}} - \frac{\int f^-(y)\mathrm{d}y}{(\|x\|+1)^{d-s}}\right) \\
&= \frac{\epsilon}{c_s}\left(\frac{1}{(\|x\|-1)^{d-s}} - \frac{1}{(\|x\|+1)^{d-s}}\right) \\
&\leq \frac{\epsilon}{c_s}\frac{2(d-s)}{(\|x\|-1)^{d-s+1}}
\end{aligned}$$

where the last line follows from the convexity of the function $u \mapsto \frac{1}{u^{d-s}}$ for $u > 0$. Thus, we can upper bound $\|I_s f\, \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2$ by

$$\begin{aligned}
\|I_s f\, \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 &\leq \frac{\epsilon}{c_s}\left(\int_{\mathbb{B}(0,2)^c} \frac{4(d-s)^2}{(\|x\|-1)^{2d-2s+2}}\mathrm{d}x\right)^{\frac{1}{2}} \\
&= \frac{2(d-s)\sqrt{\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})}\epsilon}{c_s}\left(\int_2^\infty \frac{u^{d-1}}{(u-1)^{2d-2s+2}}\mathrm{d}u\right)^{\frac{1}{2}} \\
&\leq \frac{2(d-s)\sqrt{2\pi^{d/2}/\Gamma(d/2)}\epsilon}{c_s}\left(\int_2^\infty \frac{2^{d-1}(u-1)^{d-1}}{(u-1)^{2d-2s+2}}\mathrm{d}u\right)^{\frac{1}{2}} \\
&= C_1\epsilon
\end{aligned}$$

where we set $C_1(d, \gamma) = \frac{2(d-s)\sqrt{2\pi^{d/2}/\Gamma(d/2)}}{c_s}\sqrt{\frac{2^{d-1}}{d+2-2s}}$.

Next we need to estimate $\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2$. Let $q = \frac{2}{1-\gamma/d} > 2$ and let $\|\cdot\|_{q,w}$ denote the weak q-norm. Define the distribution function $\lambda(t) = m\{x \in \mathbb{R}^d : |I_s f(x)\mathbb{1}_{\mathbb{B}(0,2)}| > t\}$ where m is the Lebesgue measure on $\mathbb{R}^d$. Because $\lambda(t) \leq \min\left(\frac{\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_{q,w}^q}{t^q}, m(\mathbb{B}(0,2))\right)$ for any $t \geq 0$, we then have

$$\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2 = \left(2\int_0^\infty t\lambda(t)\mathrm{d}t\right)^{\frac{1}{2}} \leq C_2\|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_{q,w}$$

where the constant $C_2(d, \gamma) = \sqrt{\frac{q}{q-2}}m(\mathbb{B}(0,2))^{\frac{1}{2}-\frac{1}{q}}$. Now by the Hardy-Littlewood-Sobolev lemma [Ste70, Theorem V.1],

$$\|I_s f\|_{q,w} \leq C_3\|f\|_1 = 2C_3\epsilon$$

for some constant $C_3(d,\gamma)$. Combining these inequalities together, we get

$$\|I_s f\|_2 \le \|I_s f \mathbb{1}_{\mathbb{B}(0,2)^c}\|_2 + \|I_s f \mathbb{1}_{\mathbb{B}(0,2)}\|_2 \le (C_1 + 2C_2C_3)\epsilon.$$

Finally, if either μ or ν does not have a density, we pick a positive mollifier $\phi \in C_c^\infty(\mathbb{R}^d)$ such that $\mathrm{supp}(\phi) \subseteq \mathbb{B}(0,1)$, $\phi \ge 0$, and $\int \phi = 1$. Let $\phi_\eta(x) = \eta^{-d}\phi(x-\eta)$, where $\eta > 0$. Consider $\mu_\eta = \mu * \phi_\eta$ and $\nu_\eta = \nu * \phi_\eta$, which are both absolutely continuous and supported on $\mathbb{B}(0, 1+\eta)$, we get $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta) \le (C_1 + 2C_2C_3)(1+\eta)^{\gamma/2}\mathsf{TV}(\mu_\eta,\nu_\eta)$ by the first part of this proof and a simple scaling argument. It is known that $\mu_\eta \to \mu$ and $\nu_\eta \to \nu$ as $\eta \to 0$ in a weak sense, thus by the definition of $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta)$ given in (2.1), we obtain $\mathcal{E}_\gamma(\mu_\eta,\nu_\eta) \to \mathcal{E}_\gamma(\mu,\nu)$ as $\eta \to 0$. Moreover, we have $\mathsf{TV}(\mu_\eta,\nu_\eta) \le \mathsf{TV}(\mu,\nu)$ by the data processing inequality, which concludes the proof.

□

3.2 Lower bound – smooth distributions and Gaussian mixtures

In Section 3.1 we showed that the energy distance is bounded by total variation for compactly supported measures. In this section we look at the reverse direction, namely, we aim to lower bound the energy distance by total variation.

THEOREM 3. *For any $\beta > 0$, $d \ge 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)^{\frac{2\beta+d+\gamma}{2\beta}} \le C_1\mathcal{E}_\gamma(\mu,\nu) \tag{3.1}$$

for any $\mu,\nu \in \mathcal{P}_S(\beta,d,C)$ and $\gamma \in (0,2)$. Similarly, for any $d \ge 1$ there exists a finite constant C_2 such that

$$\frac{\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}} \le C_2\mathcal{E}_\gamma(\mu,\nu) \tag{3.2}$$

for every $\mu,\nu \in \mathcal{P}_G(d)$ and $\gamma \in (0,2)$.

PROOF. Abusing notation, identify μ and ν with their Lebesgue densities. The argument proceeds trough a chain of inequalities:

1. Bound TV by the L^2 distance between densities.
2. Apply Parseval's Theorem to pass to Fourier space.
3. Apply Hölder's inequality with well-chosen exponents.

Proof of (3.1). Jensen's inequality implies that

$$2\mathsf{TV}(\mu,\nu) = \|\mu-\nu\|_1 \le \sqrt{\mathrm{vol}_d(\mathbb{B}(0,1))}\|\mu-\nu\|_2 \lesssim \|\mu-\nu\|_2,$$

where we discard dimension-dependent constants. This completes the first step of our proof. For the second step note that $\mu,\nu \in L^2(\mathbb{R}^d)$ and we may apply Parseval's theorem to obtain

$$\|\mu-\nu\|_2^2 = \frac{1}{(2\pi)^d}\|\widehat{\mu}-\widehat{\nu}\|_2^2.$$

For arbitrary $\varphi > 0$ and $r \in [1,\infty]$, Hölder's inequality with exponents $\frac{1}{r} + \frac{1}{r^*} = 1$ implies that

$$\begin{aligned}\|\widehat{\mu}-\widehat{\nu}\|_2^2 &= \int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\mu}(\omega)|^2 \frac{\|\omega\|^\varphi}{\|\omega\|^\varphi}\mathrm{d}\omega \\ &\le \left(\int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2\|\omega\|^{\varphi r}\mathrm{d}\omega\right)^{1/r}\left(\int_{\mathbb{R}^d} \frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{\varphi r^*}}\mathrm{d}\omega\right)^{1/r^*}.\end{aligned} \tag{3.3}$$

Now, we choose φ and r to satisfy

$$\begin{aligned}\varphi r &= 2\beta \\ \varphi r^* &= d+\gamma.\end{aligned}$$

The first equation ensures that the first integral term is bounded by $\|\mu-\nu\|_{\beta,2}^{2/r}$, which is assumed to be at most a d, β dependent constant. The second equation ensures that the second integral term is equal to $(\mathcal{E}_\gamma(\mu,\nu)^2/F_\gamma(d))^{1/r^*}$ by Proposition 3. The solution to this system of equations is given by $r^* = (2\beta+d+\gamma)/(2\beta)$ and $\varphi = 2\beta \cdot \frac{d+\gamma}{2\beta+d+\gamma}$. Note that clearly $\varphi > 0$ and $r^* \geq 1$. Thus, after rearrangement and using that $F_d(\gamma) = \Theta(\gamma(2-\gamma))$ up to a dimension dependent constant, we obtain

$$\sqrt{\gamma(2-\gamma)}\|\widehat{\mu}-\widehat{\nu}\|_2^{\frac{2\beta+d+\gamma}{2\beta}} \leq C_1\mathcal{E}_\gamma(\mu,\nu),$$

for a finite constant $C_1 = C_1(d,\beta)$, concluding the proof.

Proof of (3.2). We write $C(d) \in (0,\infty)$ for a dimension dependent constant that may change from line to line. The outline of the argument is analogous to the above, with the additional step of having to bound the $(\beta,2)$-Sobolev norm of the Gaussian density as $\beta\to\infty$ for which we rely on Lemma 9. Let μ and ν have densities $p*\phi$ and $q*\phi$, where ϕ is the density of $\mathcal{N}(0,I_d)$. Writing $f=(p-q)*\phi$, we can extend the proof of [JPW23, Theorem 22] to multiple dimensions to find, for any $R>2$, that

$$\begin{aligned}2\mathsf{TV}(\mu,\nu) = \|\mu-\nu\|_1 &= \int_{\|x\|\leq R} |(f*\phi)(x)|\,\mathrm{d}x + \int_{\|x\|>R}\left|\int_{\mathbb{R}^d}\phi(x-y)\mathrm{d}f(y)\right|\mathrm{d}x \\ &\leq \sqrt{\mathrm{vol}_d(\mathbb{B}(0,R))}\sqrt{\int_{\|x\|\leq R}|(f*\phi)(x)|^2\mathrm{d}x} + \int_{\|x\|>R}\exp(-\|x\|^2/8)\mathrm{d}x \\ &\leq C(d)\Big(R^{d/2}\|\mu-\nu\|_2 + \exp(-\Omega(R^2))\Big),\end{aligned}$$

where the second line uses that $\mathrm{supp}(f)\subseteq\mathbb{B}(0,1)$. Taking $R \asymp \sqrt{\log(3+1/\|\mu-\nu\|_2)}$ we obtain the inequality

$$\mathsf{TV}(\mu,\nu) \leq C(d)\|\mu-\nu\|_2\log(3+1/\|\mu-\nu\|_2)^{d/4}. \tag{3.4}$$

By Hölder's inequality we obtain

$$\begin{aligned}\|\widehat{f}\|_2 &\leq \|\|\omega\|^\beta\widehat{f}(\omega)\|_2^{\frac{d+\gamma}{2\beta+d+\gamma}}\left\|\frac{\widehat{f}(\omega)}{\|\omega\|^{\frac{d+\gamma}{2}}}\right\|_2^{\frac{2\beta}{2\beta+d+\gamma}} \\ &= \|\|\omega\|^\beta\widehat{f}(\omega)\|_2^{\frac{d+\gamma}{2\beta+d+\gamma}}\cdot\mathcal{E}_\gamma(\mu,\nu)^{\frac{2\beta}{2\beta+d+\gamma}}\cdot F_\gamma(d)^{-\frac{\beta}{2\beta+d+\gamma}}\end{aligned}$$

by Proposition 3. Using that $|\widehat{f}|\leq|\widehat{\phi}|$ and applying Lemma 9, for $\beta\geq1$ we get

$$F_\gamma(d)^{\frac{\beta}{2\beta+d+\gamma}}\|\widehat{f}\|_2 \leq \mathcal{E}_\gamma(\mu,\nu)^{\frac{2\beta}{2\beta+d+\gamma}}\left(\frac{5\pi^{d/2}}{\Gamma(d/2)}\left(\frac{2\beta+d}{2e}\right)^{\frac{2\beta+d-1}{2}}\right)^{\frac{d+\gamma}{2(2\beta+d+\gamma)}}.$$

Rearranging and using Parseval's Theorem, we get

$$\mathcal{E}_\gamma(\mu,\nu) \geq C(d)\sqrt{\gamma(2-\gamma)}\|f\|_2 \frac{\|f\|_2^{\frac{d+\gamma}{2\beta}}}{\left(\frac{2\beta+d}{2e}\right)^{\frac{(d+\gamma)(2\beta+d-1)}{8\beta}}}$$

for some d-dependent, albeit exponential, constant $C(d) > 0$. Plugging in $\beta = \log(3 + 1/\|f\|_2)$ and assuming that $\|f\|_2$ is small enough in terms of d, we obtain

$$\mathcal{E}_\gamma(\mu,\nu) \geq \frac{C(d)\sqrt{\gamma(2-\gamma)}\|f\|_2}{\log(3+1/\|f\|_2)^{\frac{d+\gamma}{4}}} \geq \frac{C(d)\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}}, \tag{3.5}$$

where the second inequality uses (3.4) and Lemma 5. □

Theorem 3 is our main technical result, which shows that $\mathcal{E}_\gamma$ is lower bounded by a polynomial of the total variation distance for both the smooth distribution class $\mathcal{P}_S$ and Gaussian mixtures $\mathcal{P}_G$. Note also that in one dimension, (3.1) follows from the Gagliardo–Nirenberg-Sobolev interpolation inequality. However, to our knowledge, the inequality is new for $d > 1$. As for the tightness of Theorem 3, we manage to prove that this inequality is the best possible for $\mathcal{P}_S$ in one dimension, and best possible up to a poly-logarithmic factor in dimension 2 and above.

THEOREM 4. *For any $\beta > 0$, $d \geq 1$, $\gamma \in (0,2)$ and $C > 0$, there exists a finite constant C_1 so that for any value of $\epsilon \in (0,1)$, there exist $\mu_\epsilon, \nu_\epsilon \in \mathcal{P}_S(\beta,d,C)$ such that $\mathsf{TV}(\mu_\epsilon,\nu_\epsilon)/\epsilon \in (1/C_1, C_1)$ and*

$$\mathcal{E}_\gamma(\mu_\epsilon,\nu_\epsilon) \leq C_1 \mathsf{TV}(\mu_\epsilon,\nu_\epsilon)^{\frac{2\beta+d+\gamma}{2\beta}} \log\left(3 + \frac{1}{\mathsf{TV}(\mu_\epsilon,\nu_\epsilon)}\right)^{C_1 \mathbb{1}\{d\geq 2\}}. \tag{3.6}$$

In the special case $\gamma = 1$ we obtain an even stronger notion of tightness.

PROPOSITION 7. *When $\gamma = 1$ we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in Theorem 4.*

Proposition 7 is an improvement over Theorem 4 due to the inequality $d_H \lesssim \overline{d_H}$ over the class $\mathcal{P}_S(\beta,d,C)$, which follows from Proposition 1. It shows also that our construction has the property that there does not exist any halfspace that separates μ and ν better than our bounds suggest.

In the remainder of this subsection, we sketch our lower bound construction that we utilize in Theorem 4. The inspiration for the construction is to have the density difference $f = \frac{\mathrm{d}\nu}{\mathrm{d}x} - \frac{\mathrm{d}\mu}{\mathrm{d}x}$ saturate Hölder's inequality in the proof of Equation (3.1). The final form of the construction is $f = gh$, where $g(x) \propto \kappa\|x\|^{1-d/2}J_{d/2-1}(\|rx\|)$ with tuning parameters r, κ, and J denotes the Bessel function of the first kind. In particular g is proportional to the inverse Fourier transform of a sphere of radius r (to saturate Hölder's inequality). Then, h is chosen to be a compactly supported bump function that is zero in a neighbourhood of the origin, which kills the spike of g at the origin as we take $r \to \infty$. We also need that $\widehat{h}|_{r\mathbb{S}^{d-1}} \equiv 0$ for infinitely large values of $r > 0$ with bounded gaps. The final property that h must satisfy is that the tails of $\widehat{h}$ must decay rapidly; towards this end we use the recent result of [Coh23] who constructs a suitable h with $|\widehat{h}(\omega)| \lesssim \exp(-\|\omega\|/\log(3+\|\omega\|)^2)$.

The key technical lemma of our construction is the following, stated informally.

LEMMA 2 (Informal). *Let f be constructed as above and let $s > -(d/2+1)$. Then there exists a sequence $r_n \to \infty$ with $|r_{n+1} - r_n| = \mathcal{O}(1)$ such that*

$$\|\|\cdot\|^s \widehat{f}\|_2^2 \lesssim \kappa^2 r_n^{2s+d-1} (\log r_n)^{2d+1}.$$

The utility of Lemma 2 is that we can use it to bound both the energy distance $\mathcal{E}_\gamma$ between μ and ν (corresponding to $s = -\frac{d+\gamma}{2}$) as well as as the order $(\beta, 2)$ Sobolev norm of f (corresponding to $s = \beta$). Its proof hinges on the near-exponential decay of $\widehat{h}$ and the fact that we may choose $h|_{r_n \mathbb{S}^{d-1}} \equiv 0$ for all n which we couple with estimates utilizing Lipschitz continuity.

3.3 Lower bound – discrete distributions

Suppose we have two discrete distributions that are supported on a common, finite set of size k. One way to measure the energy distance between them would be to identify their support with the set $\{1, 2, \ldots, k\}$, thereby embedding the two distributions in $\mathbb{R}$, and applying the one-dimensional energy distance.

While the above approach seems reasonable, it is entirely arbitrary. Indeed, there might not be a natural ordering of the support; moreover, why should one choose the integers between 1 and k instead of, say, the set $\{1, 2, 4, \ldots, 2^k\}$? The total variation distance doesn't suffer from such ambiguities, and it is unclear how our choice of embedding affects the relationship to TV. The following result attacks precisely this question.

THEOREM 5. *Let μ and ν be probability distributions supported on the set $\{x_1, \ldots, x_k\} \subseteq \mathbb{R}^d$ and let $\delta = \min_{i \neq j} \|x_i - x_j\|$. Then there exists a universal constant $C > 0$ such that*

$$\mathcal{E}_1^2(\mu, \nu) \geq \frac{C\delta}{k\sqrt{d}} \mathsf{TV}^2(\mu, \nu).$$

PROOF. Let $\mu = \sum_{i=1}^k \mu_i \delta_{x_i}$ and $\nu = \sum_{i=1}^k \nu_i \delta_{x_i}$. Then, by [Bal92, Theorem 1] we have

$$\mathcal{E}_1^2(\mu, \nu) = -\sum_{i,j} (\mu_i - \nu_i)(\mu_j - \nu_j)\|x_i - x_j\| \geq \frac{C\delta}{\sqrt{d}} \sum_{i=1}^k (\mu_i - \nu_i)^2 \geq \frac{C\delta \mathsf{TV}^2(\mu, \nu)}{k\sqrt{d}}$$

as required.

□

REMARK 2. Similar results can be proved for the generalized energy distance $\mathcal{E}_\gamma$, using e.g. the work [NW92]. However, to the best of our knowledge, these estimates degrade significantly in the dimension d in contrast with Ball's result [Bal92].

Notice that by our discussion above, the support set $\{x_1, \ldots, x_k\}$ in Theorem 5 is arbitrary and may be chosen by us. Since the scale of the supporting points $x_1, \ldots, x_k$ is statistically irrelevant, we remove this ambiguity by restricting the points to lie in the unit ball, i.e. requiring that $\max_i \|x_i\| \leq 1$. We see now that the comparison between $\mathcal{E}_1$ and TV *improves* as $\delta/\sqrt{d}$ grows. Given a fixed value of δ, we want to make the dimension d of our embedding as low as possible, which means that the points $x_1, \ldots, x_k$ should form a large δ-packing of the d-dimensional unit ball. Due to well known bounds on the packing number of the Euclidean ball, it follows that the best one can hope for is

$$\log(k) \asymp d \log(1/\delta).$$

Maximizing $\delta/\sqrt{d}$ subject to this constraint yields the choice $d = \Theta(\log(k))$ and $\delta = \Theta(1)$. This gives us the following corollary.

COROLLARY 6. *There exists a universal constant $C \in (0, \infty)$ such that for any $k \geq 1$ there exists a set of points $x_1, \ldots, x_k \in \mathbb{R}^{\lceil C \log(k) \rceil}$ with $\max_i \|x_i\| \leq 1$ such that*

$$\mathcal{E}_1\left(\sum_{i=1}^{k} \mu_i \delta_{x_i}, \sum_{i=1}^{k} \nu_i \delta_{x_i}\right) \geq \frac{\mathsf{TV}(\mu, \nu)}{C\sqrt{k}\sqrt[4]{\log(k)}}$$

for any two probability mass functions $\mu = (\mu_1, \ldots, \mu_k)$ and $\nu = (\nu_1, \ldots, \nu_k)$.

The question arises how the set of points $x_1, \ldots, x_k$ in Corollary 6 should be constructed. One solution is to use an error correcting code (ECC), whereby we take the x_i to be the codewords of an ECC on the scaled hypercube $\frac{1}{\sqrt{d}}\{\pm 1\}^d$ for some dimension/"blocklength" d. An ECC is "asymptotically good" if the message length $\log(k)$ is linear in the blocklength d, that is $d \asymp \log(k)$, and if the minimum Hamming distance between any two codewords is $\Theta(d)$, which translates precisely into $\delta = \min_{i \neq j} \|x_i - x_j\| \asymp 1$. Many explicit constructions of asymptotically good error correcting codes exist, see [Jus72] for one such example, and random codes are almost surely good [BF02]. Clearly the better the code is, the better the constants we obtain in Corollary 6.

REMARK 3. One interesting consequence of Corollary 6 and the preceeding discussion is the following: given a categorical feature with k possible values, the perceptron may obtain better performance by identifying each category with the codewords $x_1, \ldots, x_k$ of an ECC instead of the standard one-hot encoding.

4. DENSITY ESTIMATION

In this section we apply what we've learnt about the generalized energy distance and the perceptron discrepancy $\overline{d_H}$ in prior sections, and analyze multiple problems related to density estimation.

4.1 Estimating smooth distributions and Gaussian mixtures

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ for some probability distribution ν on $\mathbb{R}^d$. Given a class of "generator" distributions $\mathcal{G}$ and $\gamma \in (0, 2)$, define the minimum-$\mathcal{E}_\gamma$ estimator as

$$\tilde{\nu}_\gamma \in \operatorname*{arg\,min}_{\nu' \in \mathcal{G}} \mathcal{E}_\gamma(\nu', \nu_n), \tag{4.1}$$

where $\nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. Note that $\tilde{\nu}_\gamma$ doesn't quite agree with our definition of $\tilde{\nu}_\gamma$ in (1.2), because the $\gamma = 1$ case minimizes the *average* halfspace distance $d_H \asymp \mathcal{E}_1$ and not the perceptron discrepancy $\overline{d_H}$. The following two results bound the performance of $\tilde{\nu}$ as defined in (4.1), as an estimator of ν for the smooth density class $\mathcal{P}_S$ as well as the Gaussian mixture class $\mathcal{P}_G$. In Section 4.1.1 we present the adapation of these to $\overline{d_H}$, thereby proving Theorem 1.

THEOREM 7. *Let $\tilde{\nu}_\gamma$ be the estimator defined in (4.1). For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant C_1 so that*

$$\sup_{\nu \in \mathcal{P}_S(\beta, d, C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_1(n\gamma(2-\gamma))^{-\frac{\beta}{2\beta+d+\gamma}} \tag{4.2}$$

holds for $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$ and any $\gamma \in (0,2)$. Similarly, for any $d \geq 1$ there is a finite constant C_2 such that

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_2 \phi(n\gamma(2-\gamma)) \tag{4.3}$$

holds for $\mathcal{G} = \mathcal{P}_G(d)$ and any $\gamma \in (0,2)$, where $\phi(x) = \frac{\log(3+x)^{\frac{2d+\gamma}{4}}}{\sqrt{x}}$.

Proof. Let us focus on the case $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$ first and let $t = \frac{2\beta+d+\gamma}{2\beta}$. The inequality $\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu_n) \leq \mathcal{E}_\gamma(\nu, \nu_n)$ holds almost surely by the definition of $\tilde{\nu}_\gamma$. Combining Theorem 3 and lemma 1 with the triangle inequality for $\mathcal{E}_\gamma$ and Jensen's inequality, and writing $C_1 = C_1(\beta, d)$ for a finite constant that we relabel freely, the first claim is substantiated by the chain of inequalities

$$\begin{aligned}
\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\leq \mathbb{E}\Big[\left(C_1 \frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\leq \mathbb{E}\Big[\left(2C_1 \frac{\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\leq C_1(n\gamma(2-\gamma))^{-1/2t}.
\end{aligned}$$

The result for $\mathcal{G} = \mathcal{P}_G$ follows analogously. Define the function $r(x) = x\sqrt{\gamma(2-\gamma)}/\log(3+1/x)^{\frac{2d+\gamma}{4}}$. One can check by direct calculation that r is strictly increasing and convex on $\mathbb{R}_+$. As a consequence, its inverse r^{-1} is strictly increasing and concave. Let C_2 be a d-dependent finite constant which we relabel repeatedly. Using Theorem 3 and Jensen's inequality we obtain the chain of inequalities

$$\begin{aligned}
\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\leq \mathbb{E}\Big[r^{-1}\left(C_2 \frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)\Big] \\
&\leq r^{-1}\left(C_2 \frac{\mathbb{E}\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right) \\
&\leq r^{-1}\left(2C_2 \frac{\mathbb{E}\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right) \\
&\leq r^{-1}\left(C_2(n\gamma(2-\gamma))^{-1/2}\right).
\end{aligned}$$

The conclusion follows by Lemma 5. □

Notice that the rate of estimation of the minimum $\mathcal{E}_\gamma$ density estimator improves as $\gamma \downarrow 0$, and in fact seems to approach the optimum. However, simultaneously, the "effective sample size" $n\gamma$ shrinks. The best trade-off that we can derive is the following.

Corollary 8. *The rate in (4.2) (resp. (4.3)) can be improved to $(\log(n)/n)^{\beta/(2\beta+d)}$ (resp. $\log(n)^{d/4}\sqrt{\log\log n}/\sqrt{n}$) by setting $\gamma = \log(n)^{-1}$ (resp. $\gamma = \log\log(n)^{-1}$) adaptively.*

4.1.1 Proof of Theorems 1 and 2 The proofs of Theorems 1 and 2 are completely analogous to the proof of Theorem 7, with the only difference being that we can no longer rely on Lemma 1 to show that the distance between empirical and population measures decays at the parametric rate. However, this fact is well known for the case of $\overline{d_H}$ which provides the missing ingredient for our proof.

LEMMA 3. *Let ν be a probability distribution on $\mathbb{R}^d$ and $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for i.i.d. observations $X_i \sim \nu$. Then, for a finite universal constant C,*

$$\mathbb{E}\overline{d_H}(\nu, \nu_n) \leq C\sqrt{\frac{d}{n}}.$$

PROOF. Follows by e.g. [Ver18, 8.3.23] and the fact that $\mathcal{D}_a$, the family of halfspace indicators, has VC dimension $d+1$. □

With Lemma 3 in hand, the proof of Theorem 1 and Theorem 2 is an exercise in combining results. For the sake of completeness, we describe these steps.

PROOF OF THEOREM 1 AND THEOREM 2. Apply Theorem 3 with $\gamma = 1$, to get a comparison between TV and $\mathcal{E}_1$. Recalling that $\mathcal{E}_1 \asymp d_H$, apply Proposition 1 to get a comparison between $\mathcal{E}_1$ and $\overline{d_H}$. Using these, simply proceed as in the proof of Theorem 7 except use Lemma 3 in place of Lemma 1. □

4.2 Estimating discrete distributions

We now shift gears, and study the problem of estimating a discrete distribution ν with a support of size k. Let $\mathcal{P}_k$ denote the set of all probability distributions on the set $[k] = \{1, 2, \ldots, k\}$. Suppose we observe an i.i.d. sample $X_1, \ldots, X_n \sim \nu$ from some unknown distribution $\nu \in \mathcal{P}_k$ and that we wish to estimate it using $\tilde{\nu}$. It is a folklore fact (see e.g. [Can20, Theorem 1] or [PW23, VI.15]) that the optimal rate of estimation is given by

$$\inf_{\tilde{\nu}} \sup_{\nu \in \mathcal{P}_k} \mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \asymp \min\left\{\frac{k}{n}, 1\right\}$$

and can be achieved by simply taking $\tilde{\nu} = \nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ to be the empirical distribution. Recall from Section 3.3 that we may choose to embed the alphabet $[k]$ into some higher dimensional Euclidean space. Given distinct points $x_1, \ldots, x_k \in \mathbb{R}^d$ for some $d \geq 1$, we can identify any distribution $\mu \in \mathcal{P}_k$ with the probability distribution $\sum_{i=1}^k \mu_i \delta_{x_i}$, where μ_i is the mass that μ puts on $i \in [k]$. Using this embedding, we may then use the analogue of the minimum energy estimator that we proposed in Section 4.1.

PROPOSITION 8. *Let $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ for an unknown $\nu \in \mathcal{P}_k$. Then, there exists a constant $C \in (0, \infty)$ and a set of points $a_1, \ldots, a_k \in \mathbb{R}^{\lceil C\log(k)\rceil}$, independent of the data, such that the minimum energy estimator*

$$\tilde{\nu} \in \operatorname*{arg\,min}_{\nu' \in \mathcal{P}_k} \mathcal{E}_1^2\left(\sum_{i=1}^k \nu_i' \delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}\right)$$

satisfies

$$\sup_{\nu \in \mathcal{P}_k} \mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \leq C \min\left\{\frac{k\log(k)}{n}, 1\right\}. \tag{4.4}$$

Moreover, we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in the definition of $\tilde{\nu}$ and (4.4) remains true with an additional multiplicative $\sqrt{\log k}$ factor.

Proof. Let $a_1, \ldots, a_k \in \mathbb{R}^d$ be the points defined in Corollary 6 (relabeled from $x_1, \ldots, x_k$ for clarity) so that $d \asymp \log(k)$. By Lemma 1 we know that

$$\mathbb{E}\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i\delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}\right) \lesssim \frac{\sqrt{d}\max_i \|a_i\|}{n} \lesssim \frac{\sqrt{d}}{n}.$$

By Corollary 6, the definition of $\tilde{\nu}$ and the triangle inequality it follows that

$$\mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \lesssim \frac{kd}{n} \asymp \frac{k\log(k)}{n}.$$

Noting the trivial fact that $\mathsf{TV} \leq 1$ completes the proof of the first claim.

Suppose now that we replace $\mathcal{E}_1$ by $\overline{d_H}$ in the definition of $\tilde{\nu}$. The proof follows analogously, using the chain of inequalities

$$\frac{\mathsf{TV}}{\sqrt{k}\sqrt[4]{d}} \overset{\text{Cor.6}}{\lesssim} \mathcal{E}_1 \overset{\text{Prop. 2}}{\asymp} \frac{\sqrt{\Gamma\left(\frac{d+1}{2}\right)}}{\pi^{(d-1)/4}} d_H \overset{\max_i \|a_i\| \leq 1}{\lesssim} \overline{d_H},$$

and Lemma 3. □

Remark 4. Proposition 8 has implications for density estimation over the class of distributions on the cube $[0,1]^d$ with uniformly bounded derivatives up to order β [5]. Indeed, it is known (see e.g. [IS03, ACPS18]) that discretizing such distributions using a regular grid with $\Omega(\epsilon^{-d/\beta})$ cells maintains total-variation distances between distributions up to an additive $\mathcal{O}(\epsilon)$ error. Now, consider a 'multilayer perceptron', i.e. a fully connected multilayer neural network with activations given by $x \mapsto \mathbb{1}\{x \geq 0\}$. Such a multilayer network with a hidden layer of size $\mathcal{O}(d\log(1/\epsilon)/\beta)$ can implement the discretization described above, and thus shows (as a consequence of Proposition 8) that the ERM density estimator (1.2) achieves the minimax optimal density estimation rate $n^{-\beta/(2\beta+d)}$ over smooth densities up to polylog factors provided the discriminator class $\mathcal{D}$ includes the aforementioned multilayer perceptron and has VC-dimension at most polylog in $1/\epsilon$. This observation essentially generalizes Theorem 1, which shows that if the discriminator class includes only the *single* layer perceptron then the best possible minimax rate is $n^{-\beta/(2\beta+d+1)}$.

4.3 A stopping criterion for smooth density estimation

As a corollary to our results, we propose a stopping criterion for training density estimators. Before doing so, let us record a result about the concentration properties of the empirical energy distance about its expectation.

[5] Note that this class is not the same as $\mathcal{P}_S$.

LEMMA 4. *Let ν be supported on a compact subset $\Omega \subseteq \mathbb{R}^d$, and let ν_n be its empirical measure based on n i.i.d. observations. For every $\gamma \in (0,2)$ there exists a constant $C_1 \in (0,\infty)$ such that*

$$\mathbb{P}\left(\mathcal{E}_\gamma(\nu,\nu_n) \geq \frac{C_1}{\sqrt{n}} + t\right) \leq 2\exp\left(-\frac{nt^2}{C_1 \operatorname{diam}(\Omega)^\gamma}\right).$$

In other words, $\mathcal{E}_\gamma(\nu,\nu_n)$ is $\mathcal{O}(\operatorname{diam}(\Omega)^\gamma/n)$-sub-Gaussian.

PROOF. Recall the MMD formulation of the generalized energy distance from Section 2.3. The corresponding kernel is given by $k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma$. Clearly

$$\inf_{t\in\mathbb{R}} \sup_{x,y\in\operatorname{supp}(\nu)} |k_\gamma(x,y) - t| \lesssim \operatorname{diam}(\Omega)^\gamma.$$

Therefore, by McDiarmid's inequality we know that $\mathcal{E}_\gamma(\nu,\nu_n)$ is $\mathcal{O}(\operatorname{diam}(\Omega)^\gamma/n)$-subGaussian. From Lemma 1 we know that $\mathbb{E}\mathcal{E}_\gamma(\nu,\nu_n) \lesssim 1/\sqrt{n}$, and the conclusion follows. □

Consider the following scenario: we have i.i.d. training data $X_1,\dots,X_n$ from some distribution ν and we are training an arbitrary generative model to estimate ν. Suppose that this training process gives us a sequence of density estimators $\{\mu_k\}_{k\geq 1}$, which could be the result of, say, subsequent gradient descent steps on our parametric density estimator. Is there any way to figure out after how many steps K we may stop the training process? In other words, can we identify a value of K such that $\mathsf{TV}(\nu,\mu_K)$ is guaranteed to be less than some threshold with probability $1-\delta$? Below we exhibit one such procedure for the class of smooth distributions.

Let $\nu \in \mathcal{P}_S(\beta,d,C)$ and let ν_n be its empirical version based on the n i.i.d. observations. Assume further that $\{\mu_k\}_{k\geq 1} \subseteq \mathcal{P}_S(\beta,d,C)$ is a sequence of density estimators based on the sample $X_1,\dots,X_n$. Finally, given the training sample $(X_1,\dots,X_n)$, for each k let $\mu_{k,m_k} = \frac{1}{m_k}\sum_{i=1}^{m_k}\delta_{X_i^{(k)}}$ be the empirical distribution of the sample $(X_1^{(k)},\dots,X_{m_k}^{(k)}) \sim \mu_k^{\otimes m_k}$.

PROPOSITION 9. *For any $\beta > 0, d \geq 1$ and $\gamma \in (0,2)$ there exists a constant $c \in (0,\infty)$ such that*

$$\mathbb{P}\left(\mathsf{TV}(\mu_k,\nu) \leq c\left(\sqrt{\frac{\log(1/\delta)}{n}} + \mathcal{E}_\gamma(\mu_{k,m_k},\nu_n)\right)^{\frac{2\beta}{2\beta+d+\gamma}}, \ \forall\, k \geq 1\right) \geq 1-2\delta$$

provided we take $m_k = cn\log(k^2/\delta)/\log(1/\delta)$.

PROOF. Let $c = C_1 2^\gamma$ where C_1 is as in Lemma 4 and fix $\delta \in (0,1)$. Define the event $A = \left\{\mathcal{E}_\gamma(\nu,\nu_n) \geq \frac{c}{\sqrt{n}} + \sqrt{\frac{c\log(2/\delta)}{n}}\right\}$ and similarly

$$A_k = \left\{\mathcal{E}_\gamma(\mu_{k,m_k},\mu_k) \geq \frac{c}{\sqrt{m_k}} + \sqrt{\frac{ct_k}{m_k}}\right\}$$

for some sequence $t_1,t_2,\dots$, and each $k \geq 1$. By Lemma 4,

$$\begin{aligned}&\mathbb{P}(A) \leq \delta,\\ &\mathbb{P}(A_k) = \mathbb{E}\mathbb{P}(A_k|X_1,\dots,X_n) \leq 2\exp(-t_k).\end{aligned}$$

Taking $t_k = \log(k^2\pi^2/(3\delta))$, the union bound gives

$$\mathbb{P}\left(A \cup \bigcup_{k\geq 1} A_k\right) \leq 2\delta.$$

By the inequality $\mathcal{E}_\gamma(\mu_k, \nu) \leq \mathcal{E}_\gamma(\mu_k, \mu_{k,m_k}) + \mathcal{E}_\gamma(\mu_{k,m_k}, \nu_n) + \mathcal{E}_\gamma(\nu_n, \nu)$ it follows that

$$\mathbb{P}\left(\exists k \geq 1 \,:\, \mathcal{E}_\gamma(\nu, \mu_k) > \mathcal{E}_\gamma(\mu_{k,m_k}, \nu_n) + \frac{c}{\sqrt{n}} + \frac{c}{\sqrt{m_k}} + \sqrt{\frac{c\log(2/\delta)}{n}} + \sqrt{\frac{c\log(k^2\pi^2/(3\delta))}{m_k}}\right) \leq 2\delta.$$

Thus, by choosing $m_k \asymp n\log(k^2/\delta)/\log(1/\delta)$ we can conclude that there exists a constant c' depending only on β, d, γ such that

$$\mathbb{P}\left(\mathcal{E}_\gamma(\nu, \mu_k) \leq c'\sqrt{\frac{\log(1/\delta)}{n}} + \mathcal{E}_\gamma(\mu_{k,m_k}, \nu_n), \forall k \geq 1\right) \geq 1 - 2\delta.$$

The final conclusion follows from Theorem 3. □

Note that our bound on the probability holds for all k simultaneously, which is made possible by the fact that m_k grows as $k \to \infty$. The empirical relevance of such a result is immediate: suppose we have proposed candidate generative models $\mu_1, \mu_2, \ldots$ (e.g. one after each period of training epochs, or from different training models) that is trained on an i.i.d. dataset $X_1, \ldots, X_n$ of size n from $\nu \in \mathcal{P}_S(\beta, d, C)$. A "verifier" only needs to request for m_k independent draws from the k'th candidate, and if we ever achieve $\mathcal{E}_{(\log n)^{-1}}(\mu_{k,m_k}, \nu_n) \lesssim \sqrt{\log(1/\delta)/n}$ we can stop training and claim by Theorem 7 that we are a constant factor away from (near-)minimax optimality with a probability $1 - \delta$.

5. SUBOPTIMALITY FOR TWO-SAMPLE TESTING

The task of two-sample testing over a family of distributions $\mathcal{P}$ is the following. Given two samples $(X, Y) \sim p^{\otimes n} \otimes q^{\otimes m}$ with unknown distribution, we need to distinguish between the hypotheses

$$H_0 : p = q \text{ and } p \in \mathcal{P}, \quad \text{versus} \quad H_1 : \mathsf{TV}(p, q) > \varepsilon, \text{and } p, q \in \mathcal{P}$$

with vanishing type-I and type-II error. The special case of $m = \infty$ is known as *goodness-of-fit* testing, and for the class of smooth distributions it was famously solved by Ingster [Ing87], who showed that in dimension $d = 1$ the problem is solvable if and only if

$$n = \omega(\epsilon^{-\frac{2\beta+d/2}{\beta}}), \tag{5.1}$$

in which case a variant of the χ^2-test works. The case of general m, n and $d \geq 1$ was resolved in [ACPS18] who showed that the problem is also solvable if and only if (5.1) holds with n replaced by $\min\{n, m\}$, using the very same χ^2-test; see also [LY19]. In the remainder of the section we focus on the $m = n$ case for simplicity.

In a recent paper [PCGT23], the following test statistic for two-sample testing was proposed:

$$T_{d,k}(p, q) = \max_{(w,b)\in\mathbb{S}^{d-1}\times[0,\infty)} \left|\mathbb{E}_{X\sim p}\left(w^\top X - b\right)_+^k - \mathbb{E}_{Y\sim q}\left(w^\top Y - b\right)_+^k\right|$$

where the arguments X, Y can be either discrete (e.g. via observed samples) or continuous densities. Note that here we take $(a)_+^0 = \mathbb{1}\{a \geq 0\}$ by convention. Specifically, the test proposed is to reject the null hypothesis when

$$T_{d,k}(p_n, q_n) \geq t_n, \tag{5.2}$$

where $p_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, q_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ are empirical measures and the threshold that satisfies both $t_n = o(1)$ and $t_n = \omega(1/\sqrt{n})$. One of their main technical result [PCGT23, Theorem 6] asserts that the test (5.2) returns the correct hypothesis with probability 1 asymptotically as $n \to \infty$ for any qualifying sequence $\{t_n\}_{n\geq 1}$ and fixed p, q. However, this result leaves open questions about the sample complexity of their test, and in particular, whether it is able to achieve known minimax rates. It turns out that our results imply that their test, at least in the $k = 0$ case, cannot attain the optimal two-sample testing sample complexity (5.1) over the smooth class $\mathcal{P}_S(\beta, d, C)$. To connect to our results, notice that

$$T_{d,0}(p, q) = \overline{d_H}(p, q)\,.$$

PROPOSITION 10. *For all $d, \beta > 0$, there exists constants c, c', c'' such that for all $\varepsilon > 0$, there exists probability density functions p, q supported on the d-dimensional unit ball such that*

(a) $\|p\|_{\beta,2}, \|q\|_{\beta,2} < c$,
(b) $\|p - q\|_1 \asymp \|p - q\|_2 \asymp \varepsilon$, *and*
(c) the expected test statistic satisfies

$$\mathbb{E}[T_{d,0}(p_n, q_n)] \leq \frac{c'}{\sqrt{n}}$$

for any $n \leq (\log\frac{1}{\varepsilon})^{c''}\varepsilon^{-\frac{2\beta+d+1}{\beta}}$.

In other words, consistent testing using the statistic $T_{d,0}$ is impossible with $n = \tilde{o}(\varepsilon^{-\frac{2\beta+d+1}{\beta}})$ samples, which is a far cry from the optimal sample complexity (5.1) attainable by the χ^2 test [ACPS18]. The proof of Proposition 10 is given at Appendix D.

6. CONCLUSION

We analyzed the simple discriminating class of affine classifiers and proved its effectiveness in the ERM-GAN setting (Equation (1.2)) within the Sobolev class $\mathcal{P}_S(\beta, d)$ and Gaussian mixtures $\mathcal{P}_G(d)$ with respect to the L^2 norm (see Theorem 7 and corollary 8) and the total variation distance (see Theorem 1). Our findings affirm the rate's near-optimality for the considered classes of $\mathcal{P}_S$ and $\mathcal{P}_G$. Moreover, we present inequalities that interlink the $\mathcal{E}_\gamma$, TV, and L^2 distances, and demonstrate (in some cases) the tightness of these relationships via corresponding lower bound constructions (Appendix C). We also interpret the generalized energy distance in several ways that help advocate for its use in real applications. This work connects to a broader literature on the theoretical analysis of GAN-style models.

An interesting question emerges about the interaction between the expressiveness and concentration of the discriminator class. We found that the class of affine classifiers $\mathcal{D}_1$ is guaranteed to maintain some (potentially small) proportion of the total variation distance, and that it decays at the parametric rate between population and empirical distributions. Thus, we have traded off expressiveness for better concentration of the resulting IPM. As discussed in Section 1.2, Yatracos' estimator lies at the other end of this discriminator expressiveness-concentration trade-off: the distance d_Y is as expressive as total variation when restricted to the generator class $\mathcal{G}$, but $\sup_{\nu\in\mathcal{G}} \mathbb{E}d_Y(\nu,\nu_n)$ decays strictly slower than $1/\sqrt{n}$ for nonparametric classes $\mathcal{G}$. A downside compared to $\overline{d_H}$ is that (i) the Yatracos class $\mathcal{Y}$ requires knowledge of $\mathcal{G}$ while our $\mathcal{D}_1$ is oblivious to $\mathcal{G}$ and (ii) the distance d_Y is impractical to compute as it requires a covering of $\mathcal{G}$. Our question is: is it possible to find a class of sets $\mathcal{S} \subseteq 2^{\mathbb{B}(0,1)}$ that lies at an intermediate point on this trade-off? In other words, does $\mathcal{S}$ exist such that the ERM $\tilde{\nu}$ (1.2) using the discriminator class $\mathcal{D} = \mathcal{S}$ is optimal over, say, $\mathcal{G} = \mathcal{P}_S$ and the induced distance converges slower than $1/\sqrt{n}$ but faster than $n^{-\beta/(2\beta+d)}$ between empirical and population measures? Would there be desiderata for a sample-efficient discriminator that has neither full expressiveness against total variation and does not concentrate at a parametric rate?

ACKNOWLEDGEMENTS

Supported in part by the NSF grant CCF-2131115 and the MIT-IBM Watson AI Lab.

REFERENCES

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[ACPS18] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471, 2018.

[Bal92] Keith Ball. Eigenvalues of euclidean distance matrices. *Journal of Approximation Theory*, 68(1):74–82, 1992.

[BBR+18] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[BF02] Alexander Barg and G David Forney. Random codes: Minimum distances and error exponents. *IEEE Transactions on Information Theory*, 48(9):2568–2573, 2002.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[BMR18] Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

[Can20] Clément L Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

[CCL+22] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[Coh23] Alex Cohen. Fractal uncertainty in higher dimensions. *arXiv preprint arXiv:2305.05022*, 2023.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[IK83] Ildar A Ibragimov and Rafail Z Khasminskii. Estimation of distribution density. *Journal of Soviet Mathematics*, 21:40–57, 1983.

[Ing87] Yu. I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

[IS03] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2003.

[JPW23] Zeyu Jia, Yury Polyanskiy, and Yihong Wu. Entropic characterization of optimal rates for learning gaussian mixtures. *arXiv preprint arXiv:2306.12308*, 2023.

[Jus72] Jørn Justesen. Class of constructive asymptotically good algebraic codes. *IEEE Transactions on information theory*, 18(5):652–656, 1972.

[KG22] Arlene KH Kim and Adityanand Guntuboyina. Minimax bounds for estimating multivariate gaussian location mixtures. *Electronic Journal of Statistics*, 16(1):1461–1484, 2022.

[KNSS22] Soheil Kolouri, Kimia Nadjahi, Shahin Shahrampour, and Umut Simsekli. Generalized sliced probability metrics. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4513–4517, 2022.

[KTS$^+$18] Szymon Knop, Jacek Tabor, Przemyslaw Spurek, Igor Podolak, Marcin Mazur, and Stanislaw Jastrzebski. Cramer-wold autoencoder. 2018.

[KW19] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[Lan72] Naum S. Landkof. *Foundations of modern potential theory*, volume 180. Springer, 1972.

[Lia21] Tengyuan Liang. How well generative adversarial networks learn distributions. *The Journal of Machine Learning Research*, 22(1):10366–10406, 2021.

[LY19] Tong Li and Ming Yuan. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

[MFS$^+$17] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[MRWZ23] Youssef Marzouk, Zhi Ren, Sven Wang, and Jakob Zech. Distribution learning via neural differential equations: a nonparametric statistical perspective. *arXiv preprint arXiv:2309.01043*, 2023.

[MS64] Henryk Minc and Leroy Sathre. Some inequalities involving (r!) 1/r. *Proceedings of the Edinburgh Mathematical Society*, 14(1):41–46, 1964.

[NDC$^+$20] Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020.

[NW92] Francis J Narcowich and Joseph D Ward. Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices. *Journal of Approximation Theory*, 69(1):84–109, 1992.

[Ole06] Andriy Olenko. Upper bound on $\sqrt{x}j_\nu(x)$ and its applications. *Integral Transforms and Special Functions*, 17(6):455–467, 2006.

[PCGT23] Seunghoon Paik, Michael Celentano, Alden Green, and Ryan J Tibshirani. Maximum mean discrepancy meets neural networks: The radon-kolmogorov-smirnov test. *arXiv preprint arXiv:2309.02422*, 2023.

[PW23] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning.* Cambridge University Press, 2023+.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[RDN+] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125.*

[SR13] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[SS01] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, 06 2001.

[SSDK+20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[SSGF13] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013.

[Ste70] Elias M Stein. *Singular integrals and differentiability properties of functions.* Princeton university press, 1970.

[SUL+18] Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *Advances in Neural Information Processing Systems*, 31, 2018.

[Tie23] Stefan Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3029–3064. PMLR, 12–15 Jul 2023.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[Wat95] George N Watson. *A Treatise on the Theory of Bessel Functions.* Cambridge Mathematical Library. Cambridge University Press, 1995.

[Wen48] James G Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55(9):563, 1948.

[WM22] Sven Wang and Youssef Marzouk. On minimax density estimation via measure transport. *arXiv preprint arXiv:2207.10231*, 2022.

[Yat85] Yannis G Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

[ZBKM18] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

APPENDIX A: TECHNICAL PRELIMINARIES

DEFINITION 1. We define the inner product of two functions $f, g : \mathbb{R}^d \to \mathbb{C}$ as

$$\langle f, g \rangle \triangleq \int_{\mathbb{R}^d} f(x)\overline{g(x)}\mathrm{d}x.$$

DEFINITION 2. Given a function $f \in L^1(\mathbb{R}^d)$ we define its Fourier transform as

$$\mathcal{F}[f](\omega) \triangleq \widehat{f}(\omega) \triangleq \int_{\mathbb{R}^d} e^{-i\langle \omega, x\rangle} f(x)\mathrm{d}x, \tag{A.1}$$

and its inverse Fourier transform as

$$\mathcal{F}^{-1}[f](\omega) \triangleq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{i\langle \omega, x\rangle} f(x)\mathrm{d}x. \tag{A.2}$$

We extend by continuity and use the same notation convention for Fourier transform on $L^2(\mathbb{R}^d)$.

THEOREM 9 (Plancherel theorem). *Let $f, g \in L^2(\mathbb{R}^d)$. Then*

$$\int_{\mathbb{R}^d} f(x)\overline{g(x)}\mathrm{d}x = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\mathrm{d}\omega. \tag{A.3}$$

LEMMA 5. *Suppose $t, x, y > 0$. Then there exist finite t-dependent constants C_1, C_2 such that*

$$x \le \ y\log(3+1/y)^t \implies \frac{x}{\log(3+1/x)^t} \le C_1 \ y \implies x \le C_2 \ y\log(3+1/y)^t.$$

PROOF. Let us focus on the first implication. If $x \le y$, then it clearly holds. If $y \le x \le y\log(3+1/y)^t$ then it suffices to show

$$\left(\frac{x}{\log(3+1/x)^t} \le\right) \frac{y\log(3+1/y)^t}{\log(3+1/(y\log(3+1/y)^t))^t} \overset{!}{\le} C_1 y.$$

The inequality marked by ! is equivalent to

$$3 + 1/y \le (3 + 1/(y\log(3+1/y)^t))^{\sqrt[t]{C_1}}.$$

Now, if $y \ge 1/2$ then clearly taking $C_1 = \log_3(5)^t$ works. Suppose that instead $y \in (0, 1/2)$. Then, since log grows slower than any polynomial, there exists a t-dependent constant $c_t < \infty$ such that $\log(3+1/y) \le c_t y^{-1/(2t)}$ for all $y \in (0, 1/2)$. Therefore, we have

$$3 + \frac{1}{y\log(3+1/y)^t} \ge 3 + \frac{1}{c_t^t y^{1/2}}.$$

It is then clear that

$$3 + \frac{1}{y} \le \left(3 + \frac{1}{c_t^t y^{1/2}}\right)^{\sqrt[t]{C_1}}$$

holds for all $y \in (0, 1/2)$ if we take C_1 large enough in terms of t. The second implication follows analogously and we omit its proof. □

Lemma 6. *Let μ be a probability distribution on $\mathbb{R}^d$ and $\gamma \in (0,2)$. Then*

$$\mathbb{E}_{X\sim\mu}\int_{\mathbb{R}^d}\frac{(\cos\langle\omega,X\rangle-1)^2+\sin^2\langle\omega,X\rangle}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega \le \frac{16\pi^{d/2}M_\gamma(\mu)}{\Gamma(d/2)\gamma(2-\gamma)}.$$

Proof. We use the inequalities $(\cos t-1)^2+\sin^2(t)\le 4(t^2\wedge 1)$ valid for all $t\in\mathbb{R}$. Plugging in and using the Cauchy-Schwarz inequality, the quantity on the left hand side can be bounded as

$$\begin{aligned}4\mathbb{E}\int_{\mathbb{R}^d}\frac{1\wedge(\|\omega\|^2\|X\|^2)}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega &\le 4\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})\mathbb{E}\int_0^\infty\frac{1\wedge(r^2\|X\|^2)}{r^{1+\gamma}}\mathrm{d}r\\ &= 4\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})\mathbb{E}\left\{\|X\|^2\int_0^{\|X\|^{-1}}\frac{1}{r^{\gamma-1}}\mathrm{d}r+\int_{\|X\|^{-1}}^\infty\frac{1}{r^{1+\gamma}}\mathrm{d}r\right\}\\ &= \frac{8\,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})M_\gamma(\nu)}{\gamma(2-\gamma)},\end{aligned}$$

where $\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})=\frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)}$ is the surface area of an unit $(d-1)$-sphere. □

Lemma 7. *For $\gamma\in(0,2)$ define*

$$B_\gamma=\begin{cases}\sup_{0<a<c}\left|\int_a^c\frac{\sin(\omega)}{\omega}\mathrm{d}\omega\right| & \text{if } \gamma=1,\\ \sup_{0<a<c}\left|\int_a^c\frac{\cos(\omega)}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega\right| & \text{if } \gamma\in(0,1),\\ \sup_{0<a<c}\left|\int_a^c\frac{\cos(\omega)-1}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega\right| & \text{if } \gamma\in(1,2).\end{cases}$$

Then $B_\gamma<\infty$.

Proof. In the $\gamma>1$ case, one immediately has $B_\gamma\le\int_0^\infty\frac{\min\{1,\omega^2/2\}}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega<\infty$.
Moving on to the case $\gamma\in(0,1]$, observe that

$$\sup_{0<a<c}\left|\int_a^c f(\omega)\mathrm{d}\omega\right|\le\sum_{n=1}^\infty\left|\int_{2n\pi}^{(2n+2)\pi}f(\omega)\mathrm{d}\omega\right|+2\sup_{a<c<a+2\pi}\left|\int_a^c f(\omega)\mathrm{d}\omega\right| \tag{A.4}$$

for any function $f:\mathbb{R}_+\to\mathbb{R}$, thus it suffices to bound the two terms on the right hand side separately. Since $\cos(x+\pi)=-\cos(x)$ and $\sin(x+\pi)-\sin(x)$, for any $n\ge 1$ one has

$$\left|\int_{2n\pi}^{(2n+2)\pi}\frac{\sin(\omega)}{\omega}\mathrm{d}\omega\right|=\left|\int_{2n\pi}^{(2n+1)\pi}\frac{\pi\sin(\omega)}{\omega(\omega+\pi)}\mathrm{d}\omega\right|\lesssim\left|\int_{2n\pi}^{(2n+1)\pi}\omega^{-2}\mathrm{d}\omega\right|=\mathcal{O}(1/n^2)$$

and

$$\begin{aligned}\left|\int_{2n\pi}^{(2n+2)\pi}\frac{\cos(\omega)}{\omega^{(1+\gamma)/2}}\mathrm{d}\omega\right| &\le\int_{2n\pi}^{(2n+1)\pi}\left|\frac{1}{\omega^{(1+\gamma)/2}}-\frac{1}{(\omega+\pi)^{(1+\gamma)/2}}\right|\mathrm{d}\omega\\ &\lesssim\int_{2n\pi}^{(2n+1)\pi}\frac{1}{\omega^{1+\gamma}}\mathrm{d}\omega=\mathcal{O}(1/n^{1+\gamma}),\end{aligned}$$

both of which converge when summed over n. It thus suffices to bound the second term in the decomposition (A.4). Note that $|\sin(x)/x| \le 1$ for all x which takes care of the case $\gamma = 1$, while $\sup_{a<c<a+2\pi} |\int_a^c \frac{\cos(\omega)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega| \le \int_0^{2\pi} \frac{1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega < \infty$ holds, which completes the proof for $\gamma \in (0,1)$. □

LEMMA 8. *Let* $\int_0^\infty \cdot\, \mathrm{d}\omega \triangleq \lim_{\epsilon\to 0} \int_{1/\epsilon \ge \omega \ge \epsilon} \cdot\, \mathrm{d}\omega$. *Then, for* $x \neq 0$ *the following hold:*

$$\mathbb{1}\{x > 0\} - \frac{1}{2}\mathbb{1}\{x \neq 0\} = C_{\psi_\gamma} \int_0^\infty \frac{\sin(\omega x)}{\omega} \mathrm{d}\omega \qquad \text{for } \gamma = 1, \tag{A.5}$$

$$|x|^{(\gamma-1)/2} = C_{\psi_\gamma} \int_0^\infty \frac{\cos(\omega x)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \qquad \text{for } \gamma \in (0,1), \text{ and} \tag{A.6}$$

$$|x|^{(\gamma-1)/2} = C_{\psi_\gamma} \int_0^\infty \frac{\cos(\omega x) - 1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \qquad \text{for } \gamma \in (1,2), \tag{A.7}$$

where

$$C_{\psi_\gamma} = \begin{cases} \left(\cos(\frac{\pi(\gamma-1)}{4})\Gamma(\frac{1-\gamma}{2})\right)^{-1} & \text{if } \gamma \neq 1, \\ \frac{1}{\pi} & \text{if } \gamma = 1. \end{cases} \tag{A.8}$$

PROOF. For $x \neq 0$ clearly

$$\int_0^\infty \frac{\sin(\omega x)}{w} \mathrm{d}w = \operatorname{sign}(x) \int_0^\infty \frac{\sin(\omega)}{w} \mathrm{d}w = \operatorname{sign}(x) \frac{\pi}{2},$$

which shows the first claim. Assume from here on without loss of generality that $x > 0$. For $\gamma \in (0,1)$, by the residue theorem,

$$\begin{aligned} \int_0^\infty \frac{\cos(\omega x)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega &= x^{(\gamma-1)/2} \int_0^\infty \Re\left(\frac{e^{i\omega}}{\omega^{(1+\gamma)/2}}\right) \mathrm{d}w \\ &= x^{(\gamma-1)/2}\, \Re\left(ie^{-i\frac{\pi}{2}\gamma}\right) \int_0^\infty \frac{e^{-z}}{z^{(1+\gamma)/2}} \mathrm{d}z \\ &= x^{(\gamma-1)/2} \cos\left(\frac{\pi(\gamma-1)}{4}\right) \Gamma((1-\gamma)/2). \end{aligned}$$

Similarly, for $\gamma \in (1,2)$, integration by parts and the residue theorem gives

$$\begin{aligned} \int_0^\infty \frac{\cos(\omega x) - 1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega &= x^{(\gamma-1)/2} \int_0^\infty (\cos(\omega) - 1) \mathrm{d}\left(\frac{-1}{((\gamma-1)/2)\, \omega^{(\gamma-1)/2}}\right) \\ &= -x^{(\gamma-1)/2} \int_0^\infty \frac{\sin(w)}{(\gamma-1)/2\, \omega^{(\gamma-1)/2}} \mathrm{d}\omega \\ &= -x^{(\gamma-1)/2} \frac{2}{\gamma-1} \int_0^\infty \Im\left(\frac{e^{iw}}{\omega^{(\gamma-1)/2}}\right) \mathrm{d}\omega \\ &= -x^{(\gamma-1)/2} \frac{2}{\gamma-1} \Im\left(ie^{-\frac{\pi}{2}(\gamma-1)/2}\right) \int_0^\infty \frac{e^{-z}}{z^{(\gamma-1)/2}} \mathrm{d}z \\ &= -x^{(\gamma-1)/2} \frac{2}{\gamma-1} \cos\left(\frac{\pi(\gamma-1)}{4}\right) \Gamma(1 - (\gamma-1)/2) \\ &= x^{(\gamma-1)/2} \cos\left(\frac{\pi(\gamma-1)}{4}\right) \Gamma((1-\gamma)/2). \end{aligned}$$

□

LEMMA 9. *Let ϕ be the probability density function of $\mathcal{N}(0, \sigma I_d)$ and write $\widehat{\phi}$ for its Fourier transform. Then, for any $\beta \geq 1$,*

$$\|\widehat{\phi}(\omega)\|\omega\|^{\beta}\|_2^2 = \frac{\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\Gamma\left(\frac{2\beta+d}{2}\right) \leq \frac{5\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\left(\frac{2\beta+d}{2e}\right)^{\frac{2\beta+d-1}{2}}.$$

PROOF. It is easy to check that $\widehat{\phi}(\omega) = e^{-\frac{\sigma^2}{2}\|\omega\|^2}$. Using the change of variable $\mathrm{d}w = t^{d-1}\mathrm{d}t\mathrm{d}\sigma(v)$ where $t \in \mathbb{R}^+$ and $v \in \mathbb{S}^{d-1}$, and recalling that the surface area of the $d-1$ dimensional sphere is given by $\sigma(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma(d/2)$, we obtain

$$\begin{aligned}
\|\widehat{\phi}(\omega)\|\omega\|^{\beta}\|_2^2 &= \int_{\mathbb{R}^d} e^{-\sigma^2\|\omega\|^2}\|\omega\|^{2\beta}\mathrm{d}\omega \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)}\int_0^{\infty} e^{-\sigma^2 t^2} t^{2\beta+d-1}\mathrm{d}t \\
&= \frac{2\pi^{d/2}}{\Gamma(d/2)}\sigma^{-2\beta-d}\int_0^{\infty} e^{-t^2} t^{2\beta+d-1}\mathrm{d}t \\
&= \frac{\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\Gamma\left(\frac{2\beta+d}{2}\right).
\end{aligned}$$

The claimed inequality follows by point 2. of Lemma 10 □

LEMMA 10 (Properties of the gamma function). *The Γ function has the following properties.*

1. *For all $x > 0$ and $s \in (0, 1)$ we have $\Gamma(x + s) \leq x^s\Gamma(x)$.*
2. *For all $x > 1$ we have $\Gamma(x) \leq 5(x/e)^{x-1/2}$.*
3. *For all $x > 0$ we have $x\Gamma(x) \geq 0.885$.*

PROOF. For the first claim see [Wen48]. In [MS64] authors showed that $\log\Gamma(x) \leq (x-\frac{1}{2})\log(x) - x + \frac{1}{2}\log(2\pi) + 1$ for all $x \geq 1$, from which the second claim follows as $\exp(\frac{1}{2}\log(2\pi) + 1/2) < 5$. The third claim can be verified numerically. □

APPENDIX B: PROOFS OF SECTION 2

B.1 Proof of Proposition 4

PROOF OF PROPOSITION 4. For $v \in \mathbb{S}^{d-1}$ and $b \in \mathbb{R}$ let $\theta_v(x) = \langle v, x\rangle$ and write $\eta_v \triangleq \theta_v\#(\mu-\nu)$ for the pushforward of the measure $\mu - \nu$ through the map θ_v. To start with, we notice that

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b)\Big]^2\mathrm{d}b\mathrm{d}\sigma(v) \\
&= \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\psi_\gamma(x-b)\mathrm{d}\eta_v(x)\right)^2\mathrm{d}b\mathrm{d}\sigma(v),
\end{aligned} \tag{B.1}$$

For each $v \in \mathbb{S}^{d-1}$, the measure η_v has at most countably many atoms, therefore $b \mapsto \eta_v(\{b\}) = 0$ Leb-almost everywhere. Then, by Tonelli's theorem we can conclude that $\eta_v(\{b\}) = 0$ for $\sigma \otimes$ Leb-almost every (v, b), thus going forward we can focus on the case $x \neq b$. By Lemma 8, and writing

$A_\epsilon = [\epsilon, 1/\epsilon]$ for $\epsilon > 0$, we have

$$\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) = C_{\psi_\gamma} \int_{\mathbb{R}} \lim_{\varepsilon \to 0} \int_{A_\varepsilon} \left\{ \begin{array}{ll} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma = 1 \\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (0,1) \\ \dfrac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (1,2) \end{array} \right\} \mathrm{d}\omega \mathrm{d}\eta_v(x).$$

To exchange the integral over x and the limit over ε, notice that for any $\varepsilon > 0$ and $x \neq b \in \mathbb{R}$,

$$\begin{aligned} &\left| \int_\varepsilon^{1/\varepsilon} \frac{\sin(\omega(x-b))}{\omega} \mathrm{d}\omega \right| \leq B_\gamma && \text{if } \gamma = 1, \\ &\left| \int_\varepsilon^{1/\varepsilon} \frac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \right| \leq B_\gamma |x-b|^{(\gamma-1)/2} && \text{if } \gamma \in (0,1), \\ &\left| \int_\varepsilon^{1/\varepsilon} \frac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega \right| \leq B_\gamma |x-b|^{(\gamma-1)/2} && \text{if } \gamma \in (1,2). \end{aligned}$$

where $B_\gamma < \infty$ depends only on γ and is defined in Lemma 7. We now show that $\int_{\mathbb{R}} |x-b|^{(\gamma-1)/2} \mathrm{d}|\eta_v|(x) < \infty$ for $\sigma \otimes$Leb-almost every b, v. To this end, let $S = \{(b,v) \in \mathbb{R} \times \mathbb{S}^{d-1} : \int_{\mathbb{R}} |x-b|^{(\gamma-1)/2} \mathrm{d}|\eta_v|(x) = \infty\}$ and assume for contradiction $(\sigma \otimes \mathrm{Leb})(S) > 0$. Then $\mathbb{1}_{([-B,B] \times \mathbb{S}^{d-1}) \cap S} \uparrow \mathbb{1}_S$ as $B \to \infty$, and thus by the monotone convergence theorem there exists a finite B such that $\mathrm{Leb}(([-B,B] \times \mathbb{S}^{d-1}) \cap S) > 0$. However, by Tonelli's theorem we have

$$\begin{aligned} \int_{-B}^{B} \left(\int_{\mathbb{R}} |x-b|^{(\gamma-1)/2} \mathrm{d}|\eta_v|(x) \right)^2 \mathrm{d}b &\leq \int_{-B}^{B} \int_{\mathbb{R}} |x-b|^{\gamma-1} \mathrm{d}|\eta_v|(x) \mathrm{d}b \\ &\leq 2 \int_{\mathbb{R}} \int_0^{B+|x|} b^{\gamma-1} \mathrm{d}b \mathrm{d}|\eta_v|(x) \\ &\lesssim \int_{\mathbb{R}} (B+|x|)^\gamma \mathrm{d}|\eta_v|(x) \\ &\lesssim B^\gamma + \mathbb{E}_{X \sim \mu}\left[|\langle v, X \rangle|^\gamma\right] + \mathbb{E}_{Y \sim \nu}\left[|\langle v, Y \rangle|^\gamma\right] \\ &\leq B^\gamma + M_\gamma(\mu+\nu), \end{aligned}$$

which, after integration over $v \in \mathbb{S}^{d-1}$, leads to a contradiction if $M_\gamma(\mu+\nu) < \infty$. Continuing under the assumption $M_\gamma(\mu+\nu) < \infty$, we can apply the dominated convergence theorem to obtain

$$\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) = C_{\psi_\gamma} \lim_{\varepsilon \to 0} \int_{\mathbb{R}} \int_{A_\varepsilon} \left\{ \begin{array}{ll} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma = 1 \\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (0,1) \\ \dfrac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (1,2) \end{array} \right\} \mathrm{d}\omega \mathrm{d}\eta_v(x).$$

Then by Fubini's theorem, we exchange the order of integration to get

$$\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) = C_{\psi_\gamma} \lim_{\varepsilon\to 0} \int_{A_\varepsilon} \int_{\mathbb{R}} \left\{ \begin{matrix} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \gamma = 1 \\ \dfrac{\cos(\omega(x-b))}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (0,1) \\ \dfrac{\cos(\omega(x-b))-1}{\omega^{(1+\gamma)/2}} & \text{if } \gamma \in (1,2) \end{matrix} \right\} \mathrm{d}\eta_v(x)\mathrm{d}\omega.$$

Notice that $\int_{\mathbb{R}} e^{-i\omega x}\mathrm{d}\eta_v(x) = \widehat{\eta}_v(\omega)$, $\widehat{\eta}_v(\omega) = \overline{\widehat{\eta}_v(-\omega)}$ and $\widehat{\eta}_v(0) = 0$,

$$\begin{aligned} \int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) &= C_{\psi_\gamma} \lim_{\varepsilon\to 0} \int_{A_\varepsilon} \frac{1}{\omega^{(1+\gamma)/2}} \left\{ \begin{matrix} \Im(e^{-i\omega b}\overline{\widehat{\eta}_v(\omega)}) & \text{if } \gamma = 1 \\ \Re(e^{-i\omega b}\overline{\widehat{\eta}_v(\omega)}) & \text{if } \gamma \neq 1 \end{matrix} \right\} \mathrm{d}\omega \\ &= C_{\psi_\gamma} \lim_{\varepsilon\to 0} \begin{cases} \Im\left(\widehat{\Psi}_{\gamma,v,\varepsilon}(b)\right) & \text{if } \gamma = 1 \\ \Re\left(\widehat{\Psi}_{\gamma,v,\varepsilon}(b)\right) & \text{if } \gamma \neq 1. \end{cases} \end{aligned} \tag{B.2}$$

where we write

$$\Psi_{\gamma,v,\varepsilon}(\omega) = \frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} \mathbb{1}\{\omega \in A_\varepsilon\}.$$

Notice that $\Psi_{\gamma,v,\varepsilon}$ is bounded and compactly supported and thus lies in $L^p(\mathbb{R})$ for any p, and so in particular

$$\Psi_{\gamma,v,\varepsilon} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}),$$

which ensures that

$$\widehat{\Psi}_{\gamma,v,\varepsilon} \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R}).$$

Finally, let us write

$$\Psi_{\gamma,v}(\omega) = \lim_{\epsilon\to 0} \Psi_{\gamma,v,\varepsilon}(\omega) = \frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} \mathbb{1}\{\omega > 0\}$$

for every ω. We now show that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$ provided $M_\gamma(\mu+\nu) < \infty$, which is assumed throughout. Let $(X,Y) \sim \mu \otimes \nu$. We have

$$\begin{aligned} \int_{\mathbb{R}} |\Psi_{\gamma,v}(\omega)|^2 \mathrm{d}\omega &= \int_0^\infty \frac{|\widehat{\eta}_v(\omega)|^2}{w^{1+\gamma}} \mathrm{d}w \\ &= \int_0^\infty \frac{(\mathbb{E}[\cos\langle\omega, X\rangle - \cos\langle\omega, Y\rangle])^2 + (\mathbb{E}[\sin\langle\omega, X\rangle - \sin\langle\omega, Y\rangle])^2}{\omega^{1+\gamma}} \mathrm{d}\omega. \end{aligned}$$

Using the inequality $(a-b)^2 \leq 2(a-1)^2 + (b-1)^2, \forall\, a, b \in \mathbb{R}$ for the cos term, the inequality $(a+b) \leq 2a^2 + 2b^2, \forall\, a, b \in \mathbb{R}$ for the sin term, and applying Jensen's inequality to take the expectation outside, we can conclude that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$ by Lemma 6. Thus, by the dominated convergence theorem

$$\|\Psi_{\gamma,v,\epsilon} - \Psi_{\gamma,v}\|_2 \to 0$$

as $\epsilon \to 0$. Then, by Parseval's identity

$$\left\| \widehat{\Psi}_{\gamma,v,\epsilon} - \widehat{\Psi}_{\gamma,v} \right\|_2 \to 0 \tag{B.3}$$

as $\epsilon \to 0$. It is well know that convergence in $L^2(\mathbb{R})$ implies that there exists a subsequence $\{\varepsilon_n\}_{n=1}^{\infty}$ with $\varepsilon_n \to 0$ and $\widehat{\Psi}_{\gamma,v,\epsilon} \to \widehat{\Psi}_{\gamma,v}$ almost everywhere[6]. Therefore, by passing to this subsequence, it follows that

$$\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) = C_{\psi_\gamma} \begin{cases} \Im\left(\widehat{\Psi}_{\gamma,v}(b)\right) & \text{if } \gamma = 1 \\ \Re\left(\widehat{\Psi}_{\gamma,v}(b)\right) & \text{if } \gamma \neq 1 \end{cases} \tag{B.4}$$

for $\sigma \otimes$ Leb-almost every $(b,v) \in \mathbb{R} \times \mathbb{S}^{d-1}$. Note that since $\eta_v(\omega) \in \mathbb{R}$,

$$\begin{aligned} \Re\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) + \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2} && \text{(B.5)} \\ &= \frac{1}{2}\int_0^\infty \left(\frac{\widehat{\eta}_v(\omega)}{\omega^{(1+\gamma)/2}} e^{ib\omega} + \frac{\widehat{\eta}_v(-\omega)}{\omega^{(1+\gamma)/2}} e^{-ib\omega} \right) \mathrm{d}\omega && \text{(B.6)} \\ &= \frac{1}{2}\int_{-\infty}^\infty \frac{\widehat{\eta}_v(\omega)\operatorname{sign}(\omega)}{\omega^{(1+\gamma)/2}} e^{ib\omega}\mathrm{d}\omega && \text{(B.7)} \\ &= \mathcal{F}\left[\frac{\widehat{\eta}_v(\omega)\operatorname{sign}(\omega)}{2\omega^{(1+\gamma)/2}} \right](-b), && \text{(B.8)} \end{aligned}$$

$$\begin{aligned} \Im\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) - \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2i} && \text{(B.9)} \\ &= \frac{1}{2i}\int_0^\infty \left(\frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} e^{-ib\omega} - \frac{\overline{\widehat{\eta}_v(-\omega)}}{\omega^{(1+\gamma)/2}} e^{ib\omega} \right) \mathrm{d}\omega && \text{(B.10)} \\ &= \frac{1}{2i}\int_{-\infty}^\infty \frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} \operatorname{sign}(\omega) e^{ib\omega}\mathrm{d}\omega && \text{(B.11)} \\ &= \mathcal{F}\left[\frac{\overline{\widehat{\eta}_v(\omega)}\operatorname{sign}(\omega)}{2i\omega^{(1+\gamma)/2}} \right](-b). && \text{(B.12)} \end{aligned}$$

Plugging these expressions into (B.1), by Parseval's identity (where we use again that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$), we obtain

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b) \Big]^2 \mathrm{d}b\mathrm{d}\sigma(v) &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x) \right)^2 \mathrm{d}b\mathrm{d}\sigma(v), \\ &= 2\pi C_{\psi_\gamma}^2 \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \frac{|\widehat{\eta}_v(\omega)|^2}{4|\omega|^{1+\gamma}} \mathrm{d}w\mathrm{d}\sigma(v) \\ &= \pi C_{\psi_\gamma}^2 \int_{\mathbb{S}^{d-1}} \int_0^\infty \frac{|\widehat{\eta}_v(\omega)|^2}{|\omega|^{1+\gamma}} \mathrm{d}w\mathrm{d}\sigma(v) \\ &= \pi C_{\psi_\gamma}^2 \int_{\mathbb{R}^d} \frac{|\mathcal{F}[\mu - \nu](\omega)|^2}{\|\omega\|^{d+\gamma}} \mathrm{d}\omega < \infty, \end{aligned}$$

where in the last step we use the change of variable $\mathrm{d}\omega = \|\omega\|^{d-1}\mathrm{d}\|\omega\|\mathrm{d}\sigma(v)$ for $\omega = \|\omega\| v \in \mathbb{R}^d$. □

[6] we could also conclude this by Carleson's theorem

B.2 Proof of Proposition 5

Proof. Let $\mathcal{S}(\mathbb{R}^d)$ be the Schwartz space and $\mathcal{S}'(\mathbb{R}^d)$ be the space of all tempered distributions on $\mathbb{R}^d$. Let $\tau = \mu - \nu$ and $s = \frac{d+\gamma}{2}$. First, note that $\int \frac{K_s(x)\mathrm{d}x}{(1+\|x\|^2)^d} < \infty$, so by [Lan72, Theorem 0.10] we indeed have $K_s \in \mathcal{S}'(\mathbb{R}^d)$. By [Lan72, Theorem 0.12], since $K_s \in \mathcal{S}'(\mathbb{R}^d)$ and τ has compact support,

$$\widehat{I_s f} = \widehat{K_s * \tau} = \widehat{K_s}\widehat{\tau}.$$

By Plancherel's identity,

$$(2\pi)^{\frac{d}{2}}\|I_s\tau\|_2 = \|\widehat{I_s\tau}\|_2 = \|\widehat{K_s}\widehat{\tau}\|_2 = \frac{1}{\sqrt{F_\gamma(d)}}\mathcal{E}_\gamma(\mu,\nu),$$

where the last equality follows from Proposition 3. □

APPENDIX C: PROOF OF THEOREM 4 AND PROPOSITION 7

In this section we prove both Theorem 4 and Proposition 7. To do so, we give two constructions. The first one, presented in Appendix C.1, only applies in one dimension and gives optimal results. The second construction is given in Appendix C.2 applies in all dimensions, but loses a polylogarithmic factor.

Notation: Abusing notation, in what follows we write $\mathcal{E}_\gamma(f,g)$ and $\overline{d_H}(f,g)$ even when f and g are not necessarily probability measures or probability densities. We will also write $\|f\|_{t,2} = \|\|\cdot\|^t\hat{f}\|_2$ for potentially negative exponents $t \in \mathbb{R}$. Note that $\mathcal{E}_\gamma(f,0) = \|\hat{f}\|_{-\frac{d+\gamma}{2},2}$.

C.1 Tightness in one dimension

The Lemma below constructs the *difference* of two densities that has favorable properties.

Lemma 11. *Let $f(x) = 1\{|x| \le \pi\}\sin(rx)$ with $r \in \mathbb{Z}$ and write $f_\beta = f * \cdots * f$ for f convolved with itself $\beta - 1$ times, i.e. $f_1 = f$, $f_2 = f * f$ and so on. Fix an integer $\beta \ge 1$, $|t| \le \beta$ and $\gamma \in (0,2)$. We have*

$$\|f_\beta\|_{t,2} \asymp r^t, \|f_\beta\|_1 \asymp 1, \text{ and } \overline{d_H}(f_\beta,0) \asymp \frac{1}{r}, \tag{C.1}$$

as $r \to \infty$ where the constants may depend on γ, β, t.

Proof. The intuition for the estimates (C.1) is simple: most of the energy of f (and hence f_β) is at frequencies around $|\omega| \approx r$ and thus differentiating t times boosts the L_2-energy by r^t. A simple computation shows $\widehat{f}(\omega) = c\frac{(-1)^r}{i}\frac{r}{\omega^2 - r^2}\sin(\omega\pi)$.

Estimating $\|f_\beta\|_{t,2}$. By definition we have

$$\|f_\beta\|_{t,2}^2 \asymp \int_0^\infty |\widehat{f}(\omega)|^{2\beta}\omega^{2t} \asymp \int_0^\infty \frac{r^{2\beta}}{(\omega^2 - r^2)^{2\beta}}\omega^{2t}\sin^{2\beta}(\omega\pi)\,.$$

We decompose the integral into three regimes:

1. $\underline{\omega < r/2}$: here $(\omega^2 - r^2) \asymp r^2$ and thus

$$\int_0^{r/2} (\dots) \asymp r^{-2\beta} \int_0^{r/2} \omega^{2t} \sin^{2\beta}(\omega\pi) \asymp \left\{ \begin{matrix} r^{-2\beta}, & \text{if } 2t < -1, \\ r^{-2\beta}\log(r), & \text{if } 2t = -1, \\ r^{1+2t-2\beta}, & \text{if } 2t > -1, \end{matrix} \right\} \lesssim r^{2t}.$$

2. $\underline{\omega > 3r/2}$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus

$$\int_{3r/2}^{\infty} (\dots) \asymp r^{2\beta} \int_{3r/2}^{\infty} \frac{\sin^{2\beta}(\omega\pi)\omega^{2t}}{\omega^{4\beta}} \asymp r^{2\beta} r^{2t-4\beta+1} = r^{1+2t-2\beta} \ll r^{2t}.$$

3. $\underline{\omega \in [1/2r, 3r/2]}$: here $(\omega^2 - r^2) \asymp yr$, where $y = \omega - r$. Note also $\sin(\omega\pi) = \sin(r\pi + y\pi) = (-1)^r \sin(y\pi)$, and $\omega \asymp r$. Thus

$$\int_{r/2}^{3r/2} (\dots) \mathrm{d}\omega = \int_{-r/2}^{r/2} (\dots) dy \asymp r^{2\beta} \int_{-r/2}^{r/2} \mathrm{d}y \frac{\sin^{2\beta}(y\pi) r^{2t}}{(yr)^{2\beta}} \asymp r^{2t} \int_{\mathbb{R}} \left(\frac{\sin(y\pi)}{y} \right)^{2\beta} \mathrm{d}y \asymp r^{2t}.$$

where the last inequality follows by that the integrand is bounded at 0 and has $y^{-2\beta} \lesssim y^{-2}$ tail.

Estimating $\|f_\beta\|_1$. Follows from $\|f_\beta\|_1 \lesssim \|f_\beta\|_2 \asymp 1$ by the Cauchy-Schwartz inequality and $\|f_\beta\|_1 \geq \|\widehat{f_\beta}\|_\infty \asymp 1$ by the Hausdorff–Young inequality.

Estimating $\overline{d_H}$. We get $\overline{d_H}(f_\beta, 0) \gtrsim \mathcal{E}_1(f_\beta, 0) \asymp \|f_\beta\|_{-1,2} \asymp \frac{1}{r}$ from the first estimate. For the upper bound, note that $\widehat{\mathrm{sign}(x)} = \frac{2}{i\omega}$ and $\overline{d_H}(f_\beta, 0) = \sup_b \frac{1}{2} \int f_\beta(x) \,\mathrm{sign}(x-b) \mathrm{d}x$, so by Plancherel's identity,

$$\overline{d_H}(f_\beta, 0) \lesssim \sup_b \int \left| \widehat{f_\beta}(\omega) \frac{e^{ib\omega}}{\omega} \right| \mathrm{d}\omega \lesssim \int_0^\infty \frac{r^\beta}{(\omega^2 - r^2)^\beta} \omega^{-1} \sin^\beta(\omega\pi).$$

The fact that the above is $\mathcal{O}(1/r)$ follows analogously to the proof of our bound on $\|f_\beta\|_{t,2}$ so we omit it. This concludes our proof.

□

Proof of Theorem 4 and Proposition 7 in one dimension. We now turn to showing tightness in one dimension, utilizing the density difference constructed in Lemma 11. Let $\beta > 0$, $\overline{\beta} = \lceil \beta \rceil + 1$ and $f_{\overline{\beta}}$ be as in Lemma 11 with $r = \epsilon^{-1/\beta}$ for some $\epsilon \in (0, 1)$. Let p_0 be a smooth, compactly supported density with $\inf_{x \in [-\pi,\pi]} p_0(x) > 0$. Define

$$p_\epsilon(x) = p_0(x) + \epsilon f_{\overline{\beta}}(\overline{\beta}x)/2 \qquad \text{and} \qquad q_\epsilon(x) = p_0(x) - \epsilon f_{\overline{\beta}}(\overline{\beta}x)/2.$$

Clearly both p_ϵ, q_ϵ are compactly supported probability densities for sufficiently small ϵ, since $\|f_{\overline{\beta}}\|_\infty < \infty$ and is supported on $[-\overline{\beta}\pi, \overline{\beta}\pi]$. By Lemma 11, for each $\gamma \in (0, 2)$ the two densities satisfy

$$\|p_\epsilon - q_\epsilon\|_1 \asymp \epsilon,\ \|p_\epsilon\|_{\beta,2} \asymp \|q_\epsilon\|_{\beta,2} \asymp 1,\ \mathcal{E}_\gamma(p_\epsilon, q_\epsilon) \asymp \|p_\epsilon - q_\epsilon\|_{-(1+\gamma)/2,2} \asymp \epsilon^{\frac{2\beta+\gamma+1}{2\beta}}, \overline{d_H}(p_\epsilon, q_\epsilon) \asymp \epsilon^{\frac{\beta+1}{\beta}}.$$

This proves both Theorem 4 for $d = 1$ and Proposition 7 for all $\beta > 0$. □

C.2 Tightness in general dimension

For the discussions below, we will assume that the ambient dimension $d \geq 2$. Our construction here is less straightforward and explicit as in Appendix C.1. Before proceeding to said construction, we record some technical results that we rely on.

C.2.1 Technical preliminaries

LEMMA 12. *Let $a, b, c \in \mathbb{R}$ with $b > 0$ be constants. For all large enough r one has*

$$\int_r^\infty x^a \exp\left(-\frac{bx}{\log^2(x+2)}\right) \mathrm{d}x < r^{-c}.$$

PROOF. Assume, without loss of generality, that $c \geq 0$. For all large enough x one has $\exp(-\frac{bx}{\log^2(x+2)}) < x^{-a-c-2}$, therefore

$$\int_r^\infty x^a \exp\left(-\frac{bx}{\log^2(x+2)}\right) \mathrm{d}x < \int_r^\infty x^{-c-2} \mathrm{d}x \asymp r^{-c-1} < r^{-c}$$

when $c > 0$. □

LEMMA 13. *Let J_ν be the Bessel function of the first kind of order ν.*

1. For all $x \in \mathbb{R}^d$,

$$\int_{\mathbb{S}^{d-1}} e^{i\langle x, v\rangle} \mathrm{d}\sigma(v) = (2\pi)^{d/2} \|x\|^{1-d/2} J_{d/2-1}(\|x\|).$$

2. The inequality $J_\nu(x) \leq \frac{x^\nu}{2^{\nu-1}\Gamma(\nu+1)}$ holds provided $\nu \geq -1/2$ and $|x| \leq 1$.

3. For any $\nu \in \mathbb{R}$, as $x \to \infty$

$$J_\nu(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + O(x^{-3/2}). \tag{C.2}$$

4. For any $\nu \in \mathbb{R}$, $C_{J_\nu} := \sup_{x \geq 0} \sqrt{x}|J_\nu(x)|$ is a finite constant.

5. For all $w \in \mathbb{R}^d$,

$$\int_{\mathbb{B}^d(0,1)} e^{i\langle x, w\rangle} \mathrm{d}x = (2\pi)^{d/2} \|w\|^{-d/2} J_{d/2}(\|w\|).$$

PROOF. The second claim follows easily from the series representation

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{k=0}^\infty \frac{(-1)^k}{\Gamma(k+1)\Gamma(k+\nu+1)} \left(\frac{x}{2}\right)^{2k}$$

which is valid for all $\nu \geq -1$ and $x \geq 0$. For the first, third, and fifth claims see standard references such as [Wat95]. For the fourth claim, see [Ole06]. □

LEMMA 14. *There exists a radial function $h_0 \in L^2(\mathbb{R}^d)$ such that*

$$\operatorname{supp}(h_0) \subseteq \mathbb{B}(0,1), \tag{C.3}$$

$$|\widehat{h}_0(w)| \leq C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \qquad \textit{for all } w \in \mathbb{R}^d, \tag{C.4}$$

$$|\widehat{h}_0(w)| \geq \frac{1}{2} \qquad \textit{for all } \|w\| \leq r_{\min}, \tag{C.5}$$

where $C, c, r_{\min} > 0$.

PROOF. Apply Theorem 1.4 in [Coh23] using the spherically symmetric weight function $u : \mathbb{R}^d \to \mathbb{R}_{\leq 0}$ defined by

$$u(w) = u(\|w\|) = -\frac{\|w\|}{\log(\|w\|+2)^2}\left(\frac{(\|w\|-2)_+}{\|w\|+2}\right)^4,$$

where $(a)_+ := \max(a, 0)$ for $a \in \mathbb{R}$. □

C.2.2 The construction We are ready to present our construction of the density difference that we utilize for our proof of Theorem 4 and Proposition 7 for dimension $d \geq 2$.

LEMMA 15. *There exists a radial function $h \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ and a sequence $\{r_n\}_{n=1}^\infty$ satisfying $0 < r_1 < r_2 < \cdots < r_n \to \infty$ and $\sup_k |r_{k+1} - r_k| = \mathcal{O}(1)$ such that*

$$\operatorname{supp}(h) \subset \mathbb{B}(0,1), \tag{C.6}$$

$$\operatorname{supp}(h) \subset \mathbb{R}^d \setminus \mathbb{B}(0,r_0), \tag{C.7}$$

$$|\widehat{h}(w)| \leq C\exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \qquad \textit{for all } w \in \mathbb{R}^d, \tag{C.8}$$

$$\widehat{h}|_{\partial\mathbb{B}(0,r_n)} \equiv 0, \qquad \textit{for } n \in \mathbb{Z}^+. \tag{C.9}$$

for constants $C, c, r_0 > 0$.

PROOF. First, we construct h_0 as per the requirements in Lemma 14. It already satisfies Equation (C.6) and Equation (C.8). To address the other two requirements, we modify h_0 by convolving it with two additional terms:

$$h(x) := (A_0(\cdot) * h_0(8\cdot) * \rho_0(\cdot))(x),$$

where A_0 and ρ_0 aim to address Equation (C.7) and Equation (C.9), respectively, and are defined as

$$A_0(x) = \exp\left(-\frac{1}{1/64 - (\|x\| - 1/2)^2}\right)\mathbb{1}\{\|x\| \in (3/8, 5/8)\}, \qquad \rho_0(x) = \mathbb{1}\{\|x\| < 1/8\}.$$

Now, let's verify that h indeed satisfies the four requirements. Note that A_0 is an "annulus" supported on $\mathbb{B}(0,5/8)\backslash\mathbb{B}(0,3/8)$, and both $h_0(8\cdot)$ and ρ_0 are supported on $\mathbb{B}(0,1/8)$. Therefore, $\operatorname{supp}(h) \subset \mathbb{B}(0,7/8)\backslash\mathbb{B}(0,1/8)$, which implies Equations (C.6) and (C.7). We now turn to the other two conditions in Fourier space. Note that

$$\widehat{h}(w) = (1/8)^d \cdot \widehat{A}_0(w) \cdot \widehat{h}_0(w/8) \cdot \widehat{\rho}_0(w).$$

From Lemma 13,

$$\mathcal{F}[\mathbb{1}\{\|x\| < 1\}](w) = (2\pi)^{\frac{d}{2}}\frac{J_{\frac{d}{2}}(\|w\|)}{\|w\|^{\frac{d}{2}}} \sim (2\pi)^{\frac{d}{2}}\sqrt{\frac{2}{\pi}}\frac{\cos(\|w\| - \frac{(d+1)\pi}{4})}{\|w\|^{\frac{d+1}{2}}}$$

as $\|w\| \to \infty$ and hence $\widehat{\rho}_0(w) = (1/8)^d\mathcal{F}[\mathbb{1}\{\|x\| < 1\}](w/8)$ has infinitely many zeros around $\|w\| = 8(2n\pi + \frac{(d+1)\pi}{4})$ for sufficiently large $n \in \mathbb{Z}^+$, which implies Equation (C.9).

Finally, for Equation (C.8), note that since both A_0 and ρ_0 are Schwartz functions, so are their Fourier transforms $\widehat{A}_0$ and $\widehat{\rho}_0$ so that

$$\widehat{h}(w) \leq (1/8)^d \|\widehat{A}_0\|_\infty \|\widehat{\rho}_0\|_\infty \cdot \widehat{h}_0(w/8) \lesssim \widehat{h}_0(w/8),$$

concluding the proof. □

Given that h is bounded with compact support, we know that $\widehat{h}$ is Lipschitz with some finite parameter L. This leads to the following Lemma, which records a crucial property of the function h constructed in Lemma 15 above.

LEMMA 16. *Suppose $\widehat{h}$ is radial, L-Lipschitz and that there exist $C, c > 0$ with $|\widehat{h}(w)| < C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right)$ for all $w \in \mathbb{R}^d$. Further suppose that $\widehat{h}$ vanishes on $\partial\mathbb{B}(0, r_n)$ for a sequence $1 < r_1 < r_2 < \cdots < r_n \to \infty$. Then, for any $s > -d/2 - 1$ there exists n_0 such that*

$$\left\| \| \cdot \|^s \delta(\| \cdot \| - r_n) * \widehat{h} \right\|_2^2 \lesssim r_n^{d-1+2s} (\log r_n)^{2d+\eta}$$

for all $n \geq n_0$, hiding finite multiplicative factors involving C, c, d, L, s and where η denotes an arbitrarily small positive number.

PROOF. To ease notation, write r instead of r_n, dropping the dependence on n, which is assumed throughout. Moreover, we repeatedly relabel η, treating it essentially as an $o(1)$ term. Let D be a large constant independent of n, and decompose $\widehat{h}$ as $\widehat{h} = \widehat{h}\mathbb{1}_{\mathbb{B}(0, D\log^{1+\eta} r)} + \widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0, D\log^{1+\eta} r)}}$. By the triangle inequality

$$\begin{aligned}\left\| \| \cdot \|^s \delta(\| \cdot \| - r) * \widehat{h} \right\|_2^2 \leq 2 \left\| \| \cdot \|^s \delta(\| \cdot \| - r) * (\widehat{h}\mathbb{1}_{\mathbb{B}(0, D\log^{1+\eta} r)}) \right\|_2^2 \\ + 2 \left\| \| \cdot \|^s \delta(\| \cdot \| - r) * (\widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0, D\log^{1+\eta} r)}}) \right\|_2^2\end{aligned}$$

it suffices to consider the two terms separately. Let $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$ and note that $\|\widehat{h}\|_\infty \leq C$. Focusing on the first term, we can bound it as

$$\begin{aligned}&\left\| \| \cdot \|^s \delta(\| \cdot \| - r) * (\widehat{h}\mathbb{1}_{\mathbb{B}(0, D\log^{1+\eta} r)}) \right\|_2^2 \\ &\leq \operatorname{vol}(\mathbb{S}^{d-1}) \int_0^\infty \|w\|^{2s+d-1} \left(\left(\delta(\| \cdot \| - r) * (C\mathbb{1}_{\mathbb{B}(0, D\log^{1+\eta} r)}) \right)(\|w\| e_1) \right)^2 \mathrm{d}\|w\| \\ &\lesssim \int_{r - D\log^{1+\eta} r}^{r + D\log^{1+\eta} r} \|w\|^{2s+d-1} \left(\left(\delta(\| \cdot \| - r) * \mathbb{1}_{\mathbb{B}(0, D\log^{1+\eta} r)} \right)(\|w\| e_1) \right)^2 \mathrm{d}\|w\| \\ &\lesssim \int_{r - D\log^{1+\eta} r}^{r + D\log^{1+\eta} r} \|w\|^{2s+d-1} \log^{2(d-1)+\eta}(r) \mathrm{d}\|w\| \lesssim r^{2s+d-1} \log^{2d+\eta}(r).\end{aligned}$$

Consider now the tail outside of the ball $\mathbb{B}(0, D\log^{1+\eta} r)$. We upper bound the integral noticing that $|\widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0, D\log^{1+\eta} r)}}(w)| < H(D\log^{1+\eta}(r) \vee \|w\|)$, where $H(\|w\|) \triangleq C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right)$ is our upper bound on $|\widehat{h}|$. For some $\gamma > 0$ to be specified later we write

$$\begin{aligned}
&\Big\| \|\cdot\|^s \delta(\|\cdot\| - r) * (\widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0, D\log^{1+\eta} r)}}) \Big\|_2^2 \\
&\lesssim \int_0^{r^{-\gamma}} \|w\|^{2s+d-1} \Big(\Big(\delta(\|\cdot\| - r) * |\widehat{h}|\Big)(\|w\|e_1) \Big)^2 \mathrm{d}\|w\| \\
&\quad + \int_{r^{-\gamma}}^{2r} \|w\|^{2s+d-1} \Big(\Big(\delta(\|\cdot\| - r) * H(D\log^{1+\eta} r)\Big)(\|w\|e_1) \Big)^2 \mathrm{d}\|w\| \\
&\quad + \int_{2r}^{\infty} \|w\|^{2s+d-1} \Big(\Big(\delta(\|\cdot\| - r) * H(\|\cdot\|)\Big)(\|w\|e_1) \Big)^2 \mathrm{d}\|w\|.
\end{aligned}$$

Since implicitly $r = r_n$ for some $n \in \mathbb{Z}$, the first term can be bounded using the fact that

$$\begin{aligned}
\Big| \Big(\delta(\|\cdot\| - r) * |\widehat{h}|\Big)(\|w\|e_1) \Big| &= \Big| \int_{\mathbb{S}^{d-1}} \widehat{h}(\|\omega\|e_1 + ru) \mathrm{d}\sigma(u) \Big| \\
&= \Big| \int_{\mathbb{S}^{d-1}} (\widehat{h}(\|\omega\|e_1 + ru) - \widehat{h}(ru)) \mathrm{d}\sigma(u) \Big| \\
&\lesssim L\|w\| \asymp \|w\|
\end{aligned}$$

since $\widehat{h}$ vanishes on $\partial\mathbb{B}(0, r_n)$ and is L-Lipschitz. Therefore,

$$\begin{aligned}
&\int_0^{r^{-\gamma}} \|w\|^{2s+d-1} \Big(\Big(\delta(\|\cdot\| - r) * |\widehat{h}|\Big)(\|w\|e_1) \Big)^2 \mathrm{d}\|w\| \\
&\lesssim \int_0^{r^{-\gamma}} \|w\|^{2s+d+1} \mathrm{d}\|w\| \asymp r^{-\gamma(2s+d+2)} = \mathcal{O}(r^{2s+d-1}),
\end{aligned}$$

where the last equality follows for any $\gamma \geq \frac{2s+d-1}{2s+d+2}$. The second term can be bounded since $\int_{r^{-\gamma}}^{2r} r^{d-1}\|w\|^{2s+d-1}\mathrm{d}\|w\|$ is clearly upper bounded by a polynomial of r, but $H(D\log^{1+\eta}(r))$ vanishes quicker than any polynomial of r. Finally, the last term also vanishes according to Lemma 12 due to the near-exponential decay

$$\Big(\delta(\|\cdot\| - r) * H(\|\cdot\|)\Big)(\|w\|v) \lesssim r^{d-1}H(\|w\|/2).$$

This concludes our proof. □

With Lemmas 15 and 16 we are finally ready to present our construction. In the result below, we fix h, d, β and γ effectively treating them as constants, while letting a single parameter $r \to \infty$ to get the desired asymptotics. Whenever we write $\lesssim, \gtrsim, \asymp$ we hide multiplicative constants that may depend on h, d, β and γ.

PROPOSITION 11. *Let $h, \{r_n\}_{n=0}^{\infty}$ be constructed following Lemma 15 for some $\eta > 0$ and take $\gamma \in (0, 2)$. For every n large enough, there exists a function $g = g_n$, such that $f = gh$ satisfies the following properties for $r = r_n$:*

1. $\int f(x)\mathrm{d}x = 0$.

2. $\mathrm{supp}(f) \subset \mathbb{B}(0,1)$.
3. $\|f\|_\infty \asymp \|f\|_2 \asymp \|f\|_1 \asymp r^{-1/2}$.
4. $\|f\|_{\beta,2} \lesssim r^{\beta-1/2}(\log r)^{d+\eta}$.
5. $\mathcal{E}_\gamma(f,0) \lesssim r^{-d/2-1/2-\gamma/2}(\log r)^{d+\eta}$.
6. $\max_{v\in\mathbb{S}^{d-1},b\in\mathbb{R}} \left|\int \psi_\gamma(\langle x,v\rangle - b)f(x)\mathrm{d}x\right| \lesssim r^{-d/2-1/2-\gamma/2}(\log r)^{2d}$.

Proof. As in the proof of Lemma 16 we may relabel η and essentially treat it as a positive $o(1)$ term. Let g be given by

$$\widehat{g}(\omega) = \frac{1}{r^{d/2}}\delta(\|\omega\| - r)$$

so that it is proportional to the inverse Fourier transform of the surface measure of the sphere $\mathbb{B}(0,r)$. Writing g more explicitly from Lemma 13 we have

$$\begin{aligned} g(x) = \mathcal{F}^{-1}[\widehat{g}](x) &= \frac{1}{(2\pi)^d r^{d/2}}\int_{\mathbb{R}^d} e^{i\langle\omega,x\rangle}\delta(\|\omega\|-r)\mathrm{d}\omega \\ &= \frac{1}{(2\pi)^{d/2}}\|x\|^{1-d/2} J_{d/2-1}(\|rx\|), \end{aligned}$$

where J_ν denotes Bessel functions of the first kind of order ν. Notice that g is spherically symmetric and real-valued. We will later require some properties of J_ν, which are collected in Lemma 13. Let $f = gh$ so that $\widehat{f} = \widehat{g} * \widehat{h}$; this choice is inspired by the equality case of Hölder's inequality in Equation (3.3). Next we verify the claimed properties one by one.

Showing $\int f(x)\mathrm{d}x = 0$. This is equivalent to $\widehat{f}(0) = 0$, which is guaranteed by $\widehat{h}|_{\partial\mathbb{B}(0,r)} = 0$ and that $\widehat{g}$ is supported on $\partial\mathbb{B}(0,r)$.

Showing $\mathrm{supp}(f) \subset \mathbb{B}(0,1)$. This follows from $\mathrm{supp}(h) \subset \mathbb{B}(0,1)$.

Showing $\|f\|_\infty \lesssim r^{-1/2}$. Recall that $h|_{\mathbb{B}(0,r_0)} = 0$, it thus suffices to bound $\|g|_{\overline{\mathbb{B}(0,r_0)}}\|_\infty$. By Lemma 13 we know that as $r = r_n \to \infty$, for any fixed x with $\|x\| \geq r_0$,

$$g(x) \lesssim \|x\|^{1-d/2}\frac{1}{\sqrt{r\|x\|}} \lesssim r^{-1/2}.$$

Showing $\|f\|_2 \gtrsim \|f\|_1 \gtrsim r^{-1/2}$. The first half of the inequality follows from the Cauchy-Schwartz inequality (since f has a compact support). For the second inequality, recall that h is uniformly continuous and nontrivial, hence $\int_{\mathbb{S}^{d-1}} |h(cx)|\mathrm{d}\,\mathrm{vol}_{d-1}(x) \neq 0$ for some radius c^* and thus for all $c \in (c_0, c_1)$ for some constants $c_0, c_1 \in (0,1)$. Recall that g is spherically symmetric, therefore

$$\|f\|_1 \gtrsim \int_{c_0}^{c_1} \left|J_{d/2-1}(rx)\right| \mathrm{d}x$$

Again, we are done by (C.2).

Showing $\|f\|_{\beta,2} \lesssim r^{\beta-1/2}(\log r)^{d+\eta}$ and $\mathcal{E}_\gamma(f,0) \lesssim r^{-d/2-\gamma/2-1/2}(\log r)^{d+\eta}$. They follow from Lemma 16 in the following sense:

$$\begin{aligned}\|f\|_{\beta,2} &= \|\|\cdot\|^\beta \widehat{f}\|_2\\ &= \left\|\|\cdot\|^\beta\left(\frac{(2\pi)^{\frac{d}{2}}}{r^{\frac{d}{2}}}\delta(\|\cdot\|-r)*\widehat{h}\right)\right\|_2\\ &\lesssim r^{\beta-\frac{1}{2}}\log(r)^{d+\eta},\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}_\gamma(\mu,\nu) &= \left\|\frac{\widehat{f}}{\|\cdot\|^{\frac{d+\gamma}{2}}}\right\|_2\\ &\asymp r^{-d/2}\cdot\left\|\|\cdot\|^{-\frac{d+\gamma}{2}}\left(\delta(\|\cdot\|-r)*\widehat{h}\right)\right\|_2\\ &\lesssim r^{-d/2-\gamma/2-1/2}\log(r)^{d+\eta}.\end{aligned}$$

Showing $\max_{v\in\mathbb{S}^{d-1},b\in\mathbb{R}}\left|\int\psi_\gamma(\langle x,v\rangle-b)f(x)\mathrm{d}x\right| \lesssim r^{-d/2-\gamma/2-1/2}(\log(r))^{2d}$. Finally, we turn to showing that the max-sliced distance is also small, namely, we want to show that

$$\max_{v\in\mathbb{S}^{d-1},b\in\mathbb{R}}\left|\int\psi_\gamma(\langle x,v\rangle-b)f(x)\mathrm{d}x\right| \overset{!}{\lesssim} r^{-d/2-\gamma/2-1/2}. \tag{C.10}$$

Given arbitrary $b\in\mathbb{R}$ and $v\in\mathbb{S}^{d-1}$ let

$$F_v(b) := \int_{\mathbb{R}^d}\psi_\gamma(\langle v,x\rangle-b)f(x)\mathrm{d}x.$$

We split the argument into two cases. Suppose first that $\gamma\neq 1$. Then, using Lemmas 8 and 7, we know by dominated convergence that

$$\begin{aligned}F_v(b) &= \int_{\mathbb{R}^d}\lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon}C_{\psi_\gamma}\frac{\cos(t(\langle v,x\rangle-b))-\mathbb{1}\{\gamma>1\}}{t^{(1+\gamma)/2}}f(x)\mathrm{d}t\mathrm{d}x\\ &= C_{\psi_\gamma}\lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon}\Re\left\{\int_{\mathbb{R}^d}\frac{e^{it(\langle v,x\rangle-b)}f(x)}{t^{(1+\gamma)/2}}\mathrm{d}x\right\}\mathrm{d}t\\ &= C_{\psi_\gamma}\lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon}\frac{\cos(tb)\widehat{f}(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t.\end{aligned}$$

Let D be a large constant independent of n, following similar steps to the proof of Lemma 16, we split

$$F_v(b) = C_{\psi_\gamma}\underbrace{\int_0^\infty\frac{\cos(tb)(\widehat{g}*\widehat{h}\mathbb{1}_{\mathbb{B}(0,D\log^2 r)})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t}_{I} + C_{\psi_\gamma}\underbrace{\int_0^\infty\frac{\cos(tb)(\widehat{g}*\widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0,D\log^2 r)}})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t}_{II}.$$

Consider the first term,

$$\begin{aligned}
I &= \int_{r-D\log(r)^2}^{r+D\log(r)^2} \frac{\cos(tb)(\widehat{g} * \widehat{h}\mathbb{1}_{\mathbb{B}(0,D\log^2 r)})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\leq \|\widehat{h}\|_\infty \int_{r-D\log(r)^2}^{r+D\log(r)^2} \frac{\cos(tb)(\widehat{g} * \mathbb{1}_{\mathbb{B}(0,D\log^2 r)})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\lesssim \frac{1}{r^{(1+\gamma)/2}}\frac{1}{r^{d/2}} \int_{r-D\log(r)^2}^{r+D\log(r)^2} (\delta(\|\cdot\| - r) * \mathbb{1}_{\mathbb{B}(0,D\log^2 r)})(tv)\mathrm{d}t \\
&\lesssim \frac{1}{r^{(1+\gamma)/2}}\frac{1}{r^{d/2}}(\log r)^{2d}.
\end{aligned}$$

Consider the second term, split

$$II = \int_0^{2r} \frac{\cos(tb)(\widehat{g} * \widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0,D\log^2 r)}})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t + \int_{2r}^{\infty} \frac{\cos(tb)(\widehat{g} * \widehat{h}\mathbb{1}_{\overline{\mathbb{B}(0,D\log^2 r)}})(tv)}{t^{(1+\gamma)/2}}\mathrm{d}t$$

Both terms vanish faster than any polynomial of r. Plugging in, we obtain

$$|F_v(b)| \lesssim I \lesssim r^{-d/2-\gamma/2-1/2}\log(r)^{2d}.$$

Suppose now that $\gamma = 1$. By an analogous argument to the above, we obtain

$$\begin{aligned}
F_v(b) &= \int_{\mathbb{R}^d} \lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon} C_{\psi_\gamma}\frac{\sin(t(\langle v, x\rangle - b))}{t} f(x)\mathrm{d}t\mathrm{d}x \\
&= C_{\psi_\gamma}\lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon} \Im\left\{\int_{\mathbb{R}^d} \frac{e^{it(\langle v,x\rangle - b)}f(x)}{t}\mathrm{d}x\right\}\mathrm{d}t \\
&= C_{\psi_\gamma}\lim_{\epsilon\to 0}\int_\epsilon^{1/\epsilon} \frac{\sin(-tb)\widehat{f}(t\langle v, x\rangle v)}{t}\mathrm{d}t.
\end{aligned}$$

As before, we obtain

$$|F_v(b)| \lesssim \frac{1}{r^{d/2}}\left|\int_{r-D\log(r)^2}^{r+D\log(r)^2} \frac{\sin(-tb)}{t}\mathrm{d}t\right| \lesssim r^{-d/2-1}\log(r)^{2d}.$$

as required, and the proof of (C.10) is complete. □

Proof of Theorem 4 in general dimension. Let $\kappa > 0$ and $n \in \mathbb{N}$ be such that

$$\kappa r_n^{-1/2} \asymp \epsilon \qquad \text{and} \qquad \kappa r_n^{\beta - 1/2}(\log r_n)^{2d} \leq 1,$$

where $\asymp$ hides a numerical constant. Let f and $r = r_n$ be as constructed in Proposition 11 and with n chosen as the solution of the above display. Let p_0 be some fixed probability density with $\|p_0\|_{\beta,2} < \infty$, compact support, and satisfying $\inf_{\|x\|\leq 1} p_0(x) > 0$. Define

$$p(x) = p_0(x) + \kappa f(x)/2 \qquad \text{and} \qquad q(x) = p_0(x) - \kappa f(x)/2,$$

noting that both p and q are valid compactly supported probability densities provided ϵ is small enough since $\|\kappa f\|_\infty \lesssim \epsilon$ by construction. Moreover, we have

$$\begin{aligned}&\|p-q\|_1 \asymp \epsilon \quad \text{and} \quad \|p\|_{\beta,2} \asymp \|q\|_{\beta,2} \asymp 1 \quad \text{and}\\ &\mathcal{E}_\gamma(p,q) \asymp \epsilon^{-\frac{2\beta+d+\gamma}{2\beta}} \log(1/\epsilon)^{\Theta(d)} \quad \text{and} \quad \overline{d_H}(p,q) \asymp \epsilon^{-\frac{2\beta+d+1}{2\beta}} \log(1/\epsilon)^{\Theta(d)}.\end{aligned}$$

This completes the proof of Theorem 4. □

APPENDIX D: PROOF OF PROPOSITION 10

PROOF. Note that $\overline{d_H} = T_{d,0}$. Let p, q be the compactly supported densities constructed in the proof of Theorem 4 in the general dimensional case. Then by construction

$$\varepsilon \asymp \mathsf{TV}(p,q) \asymp \|p-q\|_2 \quad \text{and} \quad \|p\|_{\beta,2} + \|q\|_{\beta,2} \leq c \quad \text{and} \quad \overline{d_H}(p,q) \leq c\epsilon^{\frac{2\beta+d+1}{\beta}} \log(1/\epsilon)^{\Theta(d)},$$

where c is some ϵ-independent finite constant. We obtain

$$\begin{aligned}\mathbb{E}\overline{d_H}(p_n,q_n) &\leq \mathbb{E}\overline{d_H}(p_n,p) + \overline{d_H}(p,q) + \mathbb{E}\overline{d_H}(q,q_n)\\ &\lesssim 1/\sqrt{n} + \epsilon^{\frac{2\beta+d+1}{2\beta}} \log(1/\epsilon)^{\Theta(d)},\end{aligned}$$

where the second line follows by Lemma 3. This completes the proof. □