Journal of Machine Learning Research 26 (2025) 1-51 Submitted 2/24; Revised 6/25; Published 7/25

Density Estimation Using the Perceptron

Patrik Róbert Gerber PRGERBER@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA

Tianze Jiang TZJIANG@PRINCETON.EDU
Operations Research and Financial Engineering
Princeton University
98 Charlton St, Princeton, NJ, 08540, USA

Yury Polyanskiy YP@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
32 Vassar St, Cambridge, MA 02139, USA

Rui Sun ERUISUN@STANFORD.EDU
Department of Statistics
Stanford University
450 Jane Stanford Way, Stanford, CA 94305, USA

Editor: Aapo Hyvarinen

Abstract

We propose a new density estimation algorithm. Given n i.i.d. observations 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 is defined as the largest difference in mass two distribution place on any halfspace. It is shown that this estimator achieves the 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 the 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 (such as discriminators of GANs). We also show that replacing the perceptron discrepancy with the generalized energy distance of Székely and Rizzo (2013) further improves total variation loss. The generalized energy distance between empirical distributions is easily computable and differentiable, which makes it especially useful for fitting generative models. To the best of our knowledge, it is the first "simple" distance with such properties that yields minimax optimal statistical guarantees.

In addition, we shed light on the ubiquitous method of representing discrete data in domain $[k]$ via embedding vectors on a unit ball in $\mathbb{R}^d$. We show that taking $d \asymp \log(k)$ allows one to use simple linear probing to evaluate and estimate total variation distance, as well as recovering minimax optimal sample complexity for the class of discrete distributions on $[k]$.

Keywords: perceptron, halfspaces, GAN, density estimation, one-hot encoding, discrete distributions, total variation

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). This is done in supervised learning for fitting classifiers, in variational inference (Blei et al., 2017; Zhang et al., 2018) for applying ELBO, in variational autoencoders (Kingma and Welling, 2019) for fitting the decoder, in Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017) for fitting the discriminator, in diffusion models (Song et al., 2020; Chen et al., 2022) for fitting the score function, and many other settings.

To be specific, let us focus on the example of GANs, which brought about the new era of density estimation in high-dimensional spaces. The problem setting is the following. We are given access to an i.i.d. data $X_1, \dots, X_n \in \mathbb{R}^d$ sampled from an unknown distribution ν and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$ (the 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 $\mathrm{JS}(p, q) \triangleq \mathrm{KL}(p\|\frac{1}{2}p+\frac{1}{2}q)+\mathrm{KL}(q\|\frac{1}{2}p+\frac{1}{2}q)$ where $KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx$ is the Kullback-Leibler divergence. As any f-divergence, JS has a variational form (see (Polyanskiy and Wu, 2023+, 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

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

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, notably the Wasserstein-GAN (Arjovsky et al., 2017), which defines

$$\tilde{\nu} \in \arg\min_{\nu' \in \mathcal{G}} \sup_{f \in \mathcal{D}} \left| \mathbb{E}_{Y \sim \nu'} f(Y) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|, \tag{2}$$

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) or (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 (Belghazi et al., 2018) replaces the true mutual information with a "neural" one. This weakening is quite worrisome for a statistician. While the JS divergence is a strong statistical distance, as it bounds total variation from above and from below (Polyanskiy and Wu, 2023+, Eq. (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,

1. Implying in other words, 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.

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) 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 any arbitrarily complex, but bounded discriminator.

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 (11), is bounded by C; second, $\mathcal{P}_G(d) = \{\mu * \mathcal{N}(0,1) : \mathrm{supp}(\mu) \subseteq \mathbb{B}(0,1)\}$ is the 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:

$$\mathcal{D}_1 = \{x \mapsto \mathbb{1}\{x^\top v \geq b\} : v \in \mathbb{R}^d, b \in \mathbb{R}\}\,, \tag{3}$$

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

Theorem 1. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant $C_1 = C_1(\beta, d, C)$:*

$$\sup_{\nu \in \mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C_1 n^{-\frac{\beta}{2\beta+d+1}}\,, \tag{4}$$

where the estimator $\tilde{\nu}$ is defined in (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 $C_2 = C_2(d)$ so that*

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

where the estimator $\tilde{\nu}$ is defined in (2) *with $\mathcal{D} = \mathcal{D}_1$ and $\mathcal{G} = \mathcal{P}_G(d)$.*

Recall the classical result (Ibragimov and Khasminskii, 1983) 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 (4) 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]

2. For estimation of Gaussian mixtures in 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 (24)) we show that our estimator only loses a multiplicative factor of $\log(n)^{1/4}$ in loss compared to the optimal L_2-rate $\log(n)^{d/4}/\sqrt{n}$ derived in (Kim and Guntuboyina, 2022).

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

$$\overline{d_H}(\mu,\nu) \triangleq \sup_{f\in\mathcal{D}_1}\{\mathbb{E}_\mu f - \mathbb{E}_\nu f\}. \tag{5}$$

Note first that $\overline{d_H} \le \mathsf{TV}$ 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 $\beta > 0$, $d \ge 1$, there exists a finite constant $C_1 = C_1(\beta, d)$:*

$$\mathsf{TV}(\mu,\nu)^{\frac{2\beta+d+1}{2\beta}} \le C_1 \cdot (\|\mu\|_{\beta,2}^2 + \|\nu\|_{\beta,2}^2)^{\frac{d+1}{4\beta}} \cdot \overline{d_H}(\mu,\nu) \tag{6}$$

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

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

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

We remark that we also show (in Proposition 12) that the exponent $\frac{2\beta+d+1}{2\beta}$ in (6) is tight, i.e. cannot be improved in general.

With Theorem 2 in hand the proof of Theorem 1 is *notably* simple. For example, let us prove (4) (for full details, see Section 4.2). Recall that $X_i \overset{iid}{\sim} \nu$, ν_n is the empirical distribution and $\tilde{\nu} = \arg\min_{\nu'\in\mathcal{P}_S} \overline{d_H}(\nu', \nu_n)$. We then have from the triangle inequality and minimality of $\tilde{\nu}$:

$$\overline{d_H}(\tilde{\nu},\nu) \le \overline{d_H}(\tilde{\nu},\nu_n) + \overline{d_H}(\nu_n,\nu) \le 2\overline{d_H}(\nu_n,\nu)\,.$$

Thus, from Theorem 2 we have

$$\mathsf{TV}(\tilde{\nu},\nu) \lesssim \overline{d_H}(\nu_n,\nu)^{\frac{2\beta}{2\beta+d+1}}\,. \tag{7}$$

Lastly, we recall that $\mathcal{D}_1$ is a class with finite VC-dimension and thus from uniform convergence (Theorem 8.3.23, Vershynin (2018)) we have for some dimension-dependent constant $C(d)$ that

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

Thus, applying expectation and Jensen's inequality to (7) we get

$$\mathbb{E}[\mathsf{TV}(\tilde{\nu},\nu)] \lesssim \mathbb{E}[\overline{d_H}(\nu_n,\nu)^{\frac{2\beta}{2\beta+d+1}}] \lesssim \mathbb{E}[\overline{d_H}(\nu_n,\nu)]^{\frac{2\beta}{2\beta+d+1}} \lesssim n^{-\frac{\beta}{2\beta+d+1}}$$

as claimed.

While we believe that Theorem 1 provides theoretical proof for the efficacy of simple discriminators, it has several theoretical and practical deficiencies that we need to address. First, the guarantee in Theorem 1 for $\mathcal{P}_S$ is strictly worse than the minimax optimal rate, which is $\mathcal{O}\left(n^{\frac{-\beta}{2\beta+d}}\right)$ (see e.g. Ibragimov and Khasminskii (1983)).

Second, from the implementation point of view, computing the distance $\overline{d_H}$ behind Theorem 1 is 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 $\mathsf{P} \neq \mathsf{NP}$ (Guruswami and Raghavendra, 2009), and $\omega(d^{\omega(\varepsilon^{-1})})$ time for ε-optimal agnostic learning between two densities assuming either SIVP or gapSVP (Tiegel, 2023).

Finally, even if we disregard the computational complexity, it is unclear how to minimize $\tilde{\nu}$ concerning $\arg\min_{\nu'} \overline{d_H}(\nu', \nu_n)$ for given samples. This concern is alleviated by the fact that any $\tilde{\nu}$ satisfying $\overline{d_H}(\tilde{\nu}, \nu_n) = \mathcal{O}(\sqrt{d/n})$ will work without degrading our performance guarantee, and thus only an approximate minimizer is needed.

To address the above concerns, we make two changes to improve $\overline{d_H}$ in the min-distance density estimator: (a) we replace the perceptron class $\mathcal{D}_1$ in (3) with a generalized perceptron $\mathcal{D}_\gamma$ for $\gamma \in (0, 2) \setminus \{1\}$ defined as:

$$\mathcal{D}_\gamma = \{x \mapsto |x^\top v - b|^{\frac{\gamma-1}{2}} : v \in \mathbb{R}^d, b \in \mathbb{R}\}, \qquad \gamma \in (0, 2) \setminus \{1\} \tag{8}$$

(b) we replace the perceptron discrepancy $\overline{d_H}$ (defined with respect to the "best" perceptron) with an "*average*" version d_H defined in (13) (see (18) for the definition with general γ). Therefore, one does not even need to find an approximately optimal half-space, as random half-spaces provide sufficient discriminatory power. These changes are made precise in Section 2.

Our Theorem 18 and Corollary 19 show that these two changes allow us to achieve a total variation rate of $n^{-\beta/(2\beta+d+\gamma)}$ for the min distance density estimator. The improved rate comes to (within $\mathrm{polylog}(n)$) minimax optimality as $\gamma \to 0$ adaptively with n, addressing our first concern.

Somewhat unexpectedly, we discover that the average perceptron discrepancy d_H exactly equals Székely and Rizzo's energy distance $\mathcal{E}_1$ (Definition 1, Székely and Rizzo (2013)), defined as

$$\mathcal{E}_1^2(\mu, \nu) \triangleq \mathbb{E}\left[2\|X - Y\| - \|X - X'\| - \|Y - Y'\|\right], \qquad (X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}, \tag{9}$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Thus, our Theorem 18 (with $\gamma = 1$) shows 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. Furthermore, for $\gamma > 1$ the corresponding average over $\mathcal{D}_\gamma$ results in the distance $\mathcal{E}_\gamma$ known as *generalized energy distance*, defined in the same paper. See Section 2 for details.

This discovery addresses our second and final concern in the following sense. Treating our distance $\mathcal{E}_\gamma(\widetilde{\nu}_m^{\text{gen.}}, \nu_n^{\text{target}})$ as a loss function for solving $\widetilde{\nu}_m^{\text{gen.}}$, the closed-form computation of $\mathcal{E}$ via (9) requires only a polynomial $O(n^2 + m^2)$ steps, and is friendly to gradient evaluations.

Overall, our message from an algorithmic point of view is as follows: assuming one has access to a parametric family of generators sampling from ν_θ for parameters $\theta \in \mathbb{R}^p$, and if 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 (Wang and Marzouk (2022); Marzouk et al. (2023)), 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 $\mathcal{E}_\gamma(\nu'_m, \nu_n)$ and backpropagate the gradient with respect to θ,
- update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{E}_\gamma(\nu'_m, \nu_n)$ for some step size η.

Again, computing $\mathcal{E}_\gamma(\nu'_m, \nu_n)$ via (9) requires $O(n^2 + m^2)$ steps and is gradient-friendly. Note, though, that obtaining and certifying good parameterizations for generators with densities satisfying Sobolev norm constraints is nontrivial and may require careful regularization. In this work, however, our focus will be on the discriminator part of GANs, as opposed to the generator.

1.1 Contributions

This work focuses on studying the discriminator classes for adversarially estimating smooth densities, our main contributions are as follows. We show that β-smooth distributions, Gaussian mixtures and discrete distributions that are far apart in total variation distance must possess a halfspace on which their mass is substantially different (Theorems 2, 11 and 14).

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 aforementioned distribution classes (Theorems 1, 18 and 21), suggesting that halfspaces indicators are (almost) sufficient for discriminator classes in GAN (2) in order to achieve (close to) minimax optimal density estimation in the considered classes of distributions.

In Section 2 we show that the average halfspace separation distance d_H is equal up to constant to the energy distance $\mathcal{E}_1$ (Theorem 4), which has many equivalent expressions: as a weighted L^2-distance between characteristic functions (Proposition 5), as the sliced Cramér-2 distance (Theorem 8), as an IPM/MMD/energy distance (Section 2.3), and as the L^2-norm of the Riesz potential (Proposition 9).

We generalize the average halfspace distance d_H to include an exponent $\gamma \in (0, 2)$, corresponding to the generalized energy distance $\mathcal{E}_\gamma$. 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 11). Combined with the fact that $\mathcal{E}_\gamma$, similarly to d_H, decays between population and sample measures at the parametric rate (Lemma 7), the ERM for $\mathcal{E}_\gamma$ reduces the slack in the density estimation rate, achieving minimax log-optimality (Corollary 19). This result, combined with its strong approximation properties, supports its use in modern generative models (e.g. Ho et al. (2020); Goodfellow et al. (2014); Rombach et al. (2022); Ramesh et al.).

Finally, Proposition 24 shows that recent work applying $\overline{d_H}$ for two-sample testing is suboptimal over the class of smooth distributions in the minimax sense.

1.2 Related Work

An important concept related to our work is the Integral Probability Metrics (IPMs)

$$d_{\mathcal{D}}^{\text{IPM}}(\mathbb{P}, \mathbb{Q}) \triangleq \sup_{f \in \mathcal{D}} |\mathbb{E}_{\mathbb{P}} f - \mathbb{E}_{\mathbb{Q}} f| \tag{10}$$

where $\mathcal{D}$ is the discriminator class. Examples include TV when $\mathcal{D}$ is the class of bounded functions and W_1 when $\mathcal{D}$ is the class of Lipschitz functions. IPM distances represent the max separation between two distributions one gets from an adversary powered with $\mathcal{D}$, and can be viewed as the generator objective in GAN (2).

In terms of distance comparison inequalities on smooth density classes similar to our Theorem 2, the closest work we found was Chae and Walker (2020), in which the authors have shown comparison between $\mathsf{TV} \lesssim W_1^{\beta/(\beta+1)}$ for (L_1, β) Sobolev smooth densities on $\mathbb{R}$. This inequality is also optimal in the exponent. Subsequently, Chae (2024) extended comparison inequality to Besov densities on $\mathbb{R}^d$ and W_p metrics. These results[3] serve the same purpose as ours: they show that achieving optimal estimation in a "weak" norm (such as W_1) implies optimal estimation rate under

3. We thank anonymous reviewers for pointing them to us.

a "strong" norm. Since W_1 corresponds to a non-parametric discriminator class of all Lipschitz functions, this result does not explain why simple discriminators work so well in practice.

Another paper related to distance comparison is Bai et al. (2018) whose authors study comparison 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 (Bai et al., 2018, 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 the density estimation literature, an estimator of the form (2) applied on smooth densities first appeared in the famous work of Yatracos (1985). Instead of indicators of halfspaces, they consider the class of discriminators

$$\mathcal{Y}_\epsilon \triangleq \{\mathbb{1}\{\mathrm{d}\nu_i/\mathrm{d}\nu_j \geq 1\} : 1 \leq i, j \leq N(\epsilon, \mathcal{G})\},$$

where $\nu_1, \ldots, \nu_{N(\epsilon,\mathcal{G})}$ forms a minimal ϵ-TV covering of the class $\mathcal{G}$ and $N(\epsilon, \mathcal{G})$ 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 $|\mathsf{TV} - d_Y| = O(\epsilon)$ 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 $\mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \lesssim \sqrt{\log N(\epsilon, \mathcal{G})/n} + \epsilon$ follows by the triangle inequality (here $\tilde{\nu}$ is defined as in (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 $\epsilon = \epsilon(n) \asymp n^{-\beta/(2\beta+d)}$.

In terms of GAN (or more generally any adversarial) density estimation, a series of papers (Singh et al. (2018), Uppal et al. (2019), Chen et al. (2020), Belomestny et al. (2021), Liang (2021), Chae (2022), Stéphanovitch et al. (2023), etc) study the problem of density estimation over smoothness classes with respect to IPM distances. These works typically choose the discriminator class to be a large neural network of size growing with n, β and giving a universal approximation of some non-parametric discriminator class dependent on β. Given such a fine approximation to discriminator class, one obtains IPM error bounds (on e.g. W_1), which is then converted to TV bound via comparison inequalities. In this paper, we stress again, the discriminator class is extremely small and weak (a collection of half-spaces) and does not depend on smoothness β or number of samples n. We provide more detailed literature review and discussion in Section A.

Another paper by Oko et al. (2023) derives minimax density estimation guarantees for a class of diffusion-based estimators. Their result is similar to ours in that it obtains rigorous (near-)optimal guarantees for a method that is a realistic model of what is currently done in practice. However, their analysis also rely crucially on the universal approximation property of neural networks for the score function.

In this present work, we mainly focus on the fixed and interpretable discriminator class $\{x \mapsto f(x^\top v - b) : \|v\| \leq 1, b \in \mathbb{R}\}$ for a set of prescribed f and derive comparisons to TV for smooth distributions, Gaussian mixtures, and discrete distributions. In addition, we prove the (near)-optimality of our results (for smooth densities) and also derive nonparametric estimation rates for the corresponding GAN density estimators. Our work is, to the best of our knowledge, closest to the first that satisfies:

1. Uses a vanilla GAN-type approach ((2)) with ERM (which is close to practice) and obtains total variation rates that are (close to) minimax optimal.

2. Discriminator class is oblivious to the smoothness parameters: parts of our results are also n-oblivious, and the ones that depend on n only concern the activation function. Our proposed discriminator class is also *very* simple: we use a composition of a single non-linearity and a linear function (i.e. it is a VC class with dimension $d+1$).

The closest works to the above objectives are Belomestny et al. (2021) and Stéphanovitch et al. (2023), model 3, both of which derived log-optimal TV risk exponent $\widetilde{O}\left(n^{\frac{-\beta}{2\beta+d}}\right)$ via (2). However, they both relied on the universal approximation of neural networks which has to be larger than some function of β, the smoothness parameter, as well as n, the sample size. Hence another message, in contrast to the above works on discriminator networks, is that on an *extremely* simple class of discriminators parametrized by half-spaces (even with finite VC dimension) without the need for training/optimization, one can get almost optimal density estimation rates by averaging random halfspace distances (Theorem 18). Moreover, if *any* generator fits empirical to within $O(1/\sqrt{n})$ of our distance (this rate is oblivious to β), we get TV with high probability (Proposition 23).

Independent of this work, recent results by Paik et al. (2023) 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 Section E 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 $O, o, \Theta, \Omega, \omega$ follow the conventional "big-O" notation, and $\tilde{O}, \tilde{o}$ hide polylogarithmic factors. We use $\lesssim, \gtrsim$ and $\asymp$ throughout our calculations to hide multiplicative constants that are irrelevant (depending on the context). 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 $x, y \in \mathbb{R}^d$. The Gamma function is denoted by Γ. We write $\mathbb{B}(x, r) \triangleq \{y \in \mathbb{R}^d : \|x - y\| \le r\}$, $\mathbb{S}^{d-1} \triangleq \{x \in \mathbb{R}^d : \|x\| = 1\}$ and σ for the unnormalized surface measure on $\mathbb{S}^{d-1}$. The surface area of a unit $(d-1)$-sphere is also written as $\sigma(\mathbb{S}^{d-1}) = 2\pi^{d/2}/\Gamma\left(\frac{d}{2}\right)$. In particular, if X is a random vector uniformly distributioned on $\mathbb{S}^{d-1}$ then for any h we have

$$\mathbb{E}[h(X)] = \frac{1}{\sigma(\mathbb{S}^{d-1})} \int_{\mathbb{R}^d} h(y) \mathrm{d}\sigma(y)\,.$$

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 $\|f\|_p \triangleq \left(\int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}x\right)^{1/p} < \infty$. 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

$$\|f\|_{\beta,2}^2 \triangleq \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\widehat{f}(\omega)|^2 \mathrm{d}\omega. \tag{11}$$

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

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

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.*

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 4.2. 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

$$\mathcal{E}_\gamma(\mu, \nu) = \mathbb{E}\Big[2\|X - Y\|^\gamma - \|X - X'\|^\gamma - \|Y - Y'\|^\gamma\Big], \qquad \text{where } (X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}. \tag{12}$$

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 (12). Motivated by this, we study the (generalized) energy distance and give multiple equivalent characterizations of it from Section 2.1 to Section 2.5.

2.1 From Perceptron Discrepancy 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

$$d_H(\mu, \nu) \triangleq \sqrt{\int_{v \in \mathbb{S}^{d-1}} \int_{b \in \mathbb{R}} \left(\int_{\langle v, x \rangle \geq b} \mathrm{d}\mu(x) - \mathrm{d}\nu(x) \right)^2 \mathrm{d}b \mathrm{d}\sigma(v)}, \tag{13}$$

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 a 'mean squared' version of perceptron discrepancy, because the integrals over b and v only range over bounded sets. However, in general, the integral over b in the definition of d_H is 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 3. *For any $\beta > 0$, $d \geq 1$, $C > 0$, and for all $\mu, \nu \in \mathcal{P}_S(\beta, d, \infty)$, we have*

$$\sqrt{\frac{\Gamma(d/2)}{4\pi^{d/2}}} d_H(\mu, \nu) \leq \overline{d_H}(\mu, \nu). \tag{14}$$

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

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

Proof The proof of (14) is immediate after noting that all distributions in $\mathcal{P}_S(\beta, d, \infty)$ 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

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

Now, since τ is supported on a subset of $\mathbb{B}(0, 1)$ by definition of the class $\mathcal{P}_G(d)$, 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 3 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.

Theorem 4. *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 (15) and (18). We defer the proof to Theorem 8 which is its direct generalization as the interpretation of $\mathcal{E}_\gamma$ as the "average" $\mathcal{D}_\gamma$ distance for all $\gamma \in (0,2)$. ■

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 5 ((Székely and Rizzo, 2013, Proposition 2))**.** *Let $\gamma \in (0,2)$ and let μ, ν be probability distributions on $\mathbb{R}^d$ with finite γ'th moment. Then,*

$$\mathcal{E}_\gamma^2(\mu,\nu) = F_\gamma(d)\int_{\mathbb{R}^d}\frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{d+\gamma}}\mathrm{d}\omega, \tag{15}$$

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 6. *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 5.

2.3 The MMD and IPM Forms

Another interpretation of the generalized energy distance 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 (Muandet et al., 2017, 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 (16). 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 Schölkopf and Smola (2001); Muandet et al. (2017) 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

$$k_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma \tag{16}$$

for $\gamma \in (0,2)$, which in one dimension 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 Sejdinovic et al. (2013). 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.

A straightforward consequence of the above characterization is the fact that $\mathcal{E}_\gamma$ decays at the parametric rate between empirical and population measures. This is not terribly surprising as analogous results hold for arbitrary MMDs with bounded kernel, see for example (Gretton et al., 2012, Theorem 7). Recall that $M_t(\nu)$ denotes the t'th absolute moment of the measure ν.

Lemma 7. *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 $X_1,\dots,X_n$ from ν. Then, for any $\gamma \in (0,2)$,*

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

For a high-probability bound when ν is compactly supported, see Lemma 22.

Proof Let $\tilde{X}_1,\dots,\tilde{X}_n$ be an additional i.i.d. sample from ν, and write $\tilde{\nu}_n$ for the corresponding empirical measure. Using the definition of $\mathcal{E}_\gamma$ in (12), we can compute

$$\begin{aligned}\mathbb{E}\mathcal{E}_\gamma^2(\nu_n,\tilde{\nu}_n) &= \mathbb{E}\Big[\frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - \tilde{X}_j\|^\gamma - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|\tilde{X}_i - \tilde{X}_j\|^\gamma \\ &\qquad - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|X_i - X_j\|^\gamma\Big] \\ &= \frac{2}{n}\mathbb{E}\|X_1 - X_2\|^\gamma.\end{aligned}$$

The conclusion then follows from taking the expectation of the expression

$$\mathbb{E}\Big[\mathcal{E}_\gamma^2(\tilde{\nu}_n,\nu_n)\Big|\nu_n\Big] = \mathcal{E}_\gamma^2(\nu,\nu_n) + \frac{1}{n}\mathbb{E}\|X_1 - X_2\|^\gamma$$

and the inequality $|x+y|^\gamma \le 2^{\max\{0,\gamma-1\}}(|x|^\gamma + |y|^\gamma)$ for all $x, y \in \mathbb{R}$. ■

As a clarification remark, while IPM distances also appear in the formulation of GAN (2) with respect to the discriminator class $\mathcal{D}$, our distance $\mathcal{E}_\gamma$ is not the IPM of $\mathcal{D}_\gamma$, but rather that of the RKHS ball of k_γ defined above. For $\gamma = 1$, the IPM for $\mathcal{D}_\gamma$ returns $\overline{d_H}$, which is lower bounded by (up to constant, see Proposition 3) $d_H = \Theta(\mathcal{E}_1)$. For $\gamma \neq 1$, we leave the study of the IPM for $\mathcal{D}_\gamma$ as well as the characterization of the RKHS ball of k_γ out of scope. Despite not being precisely the IPM of $\mathcal{D}_\gamma$, our next characterization relates $\mathcal{E}_\gamma$ to $\mathcal{D}_\gamma$ in the same way d_H is to $\mathcal{D}_1$ via a "slicing" perspective.

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

$$\psi_\gamma(x) = \begin{cases} |x|^{(\gamma-1)/2} & \text{for } \gamma \neq 1 \\ \mathbb{1}\{x \ge 0\} & \text{otherwise.} \end{cases} \tag{17}$$

The following result, to the best of our knowledge, has not appeared in prior literature except for the case of $\gamma = 1$. This is also the direct generalization of Theorem 4.

Theorem 8. *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*

$$\mathcal{E}_\gamma^2(\mu, \nu) = \frac{1}{S_\gamma} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b)\Big]^2 \mathrm{d}b\mathrm{d}\sigma(v), \tag{18}$$

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 $S_1 = \frac{\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}$.

The proof of Theorem 8 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 (18). 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 within the squared quantity, the integral is finite.

As claimed, using the language of Nadjahi et al. (2020), Theorem 8 allows us to interpret $\mathcal{E}_\gamma$ as a *sliced* probability divergence. 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

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

We may also observe that the energy distance $\mathcal{E}_1$ is equal to the sliced Cramér-2 distance up to constant, which has been studied recently by both theoretical and empirical works (Knop et al., 2018; Kolouri et al., 2022).[4]

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

2.5 The Riesz Potential Form

The generalized energy distance can also be linked to the Riesz potential (Landkof, 1972, 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 $K_s(x) = c_s^{-1}\|x\|^{s-d}$ and $c_s = \pi^{d/2}2^s\Gamma(s/2)\Gamma((d-s)/2)^{-1}$. The Fourier transform of the Riesz kernel is given by $\widehat{K_s}(\omega) = \|\omega\|^{-s}$, interpreted as a tempered distribution. The following proposition is derived by setting $s = \frac{d+\gamma}{2}$ and using the Fourier form (Proposition 5) of the energy distance.

Proposition 9. *Let $\gamma \in (0, \min\{d, 2\})$ and let μ, ν be compactly supported probability distributions on $\mathbb{R}^d$. Then*

$$\mathcal{E}_\gamma(\mu, \nu) = (2\pi)^{d/2}\sqrt{F_\gamma(d)}\|I_{\frac{d+\gamma}{2}}(\mu - \nu)\|_2. \tag{19}$$

3. Main Comparison: TV Versus Energy

After considering the connection of the perceptron discrepancy $\overline{d_H}$ to the 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 derive lower bounds on the generalized energy distance in terms of the total variation distance over 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 — Arbitrary 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 $\mathcal{E}_1$ (Theorem 4), 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 $\gamma = 1$.

Proposition 10. *For any dimension $d \ge 1$ and $\gamma \in (0, 2)$ there exists a finite constant $c = c(d, \gamma)$ such that for any two probability distributions μ, ν supported on the unit ball we have*

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

Proof The proof directly follows from $\|x - y\|^\gamma \le 2^\gamma$ and so $\mathcal{E}_\gamma^2(\mu, \nu) = \mathbb{E}\Big[2\|X - Y\|^\gamma - \|X - X'\|^\gamma - \|Y - Y'\|^\gamma\Big] = \int -\|x - y\|^\gamma(\mathrm{d}\mu(x) - \mathrm{d}\nu(x))(\mathrm{d}\mu(y) - \mathrm{d}\nu(y)) \le 2^\gamma\mathsf{TV}(\mu, \nu)^2$, where $(X, X', Y, Y') \sim \mu^{\otimes 2} \otimes \nu^{\otimes 2}$. ∎

3.2 Lower Bound — Smooth Distributions And Gaussian Mixtures

In Section 3.1 we showed that the energy distance is upper 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 11. *For any $\beta > 0$, $d \geq 1$, there exists a finite constant $C_1 = C_1(\beta, d)$ so that*

$$\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)^{\frac{2\beta+d+\gamma}{2\beta}} \leq C_1 \cdot (\|\mu\|_{\beta,2}^2 + \|\nu\|_{\beta,2}^2)^{\frac{d+\gamma}{4\beta}} \cdot \mathcal{E}_\gamma(\mu,\nu) \tag{20}$$

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

$$\frac{\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}} \leq C_2\mathcal{E}_\gamma(\mu,\nu) \tag{21}$$

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 (20). Jensen's inequality implies that

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

where vol denotes volume and 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

$$\begin{aligned}\|\widehat{\mu}-\widehat{\nu}\|_2^2 &= \int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\mu}(\omega)|^2 \frac{\|\omega\|^\varphi}{\|\omega\|^\varphi}\mathrm{d}\omega \\ &\leq \left(\int_{\mathbb{R}^d} |\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2\|\omega\|^{\varphi r}\mathrm{d}\omega\right)^{1/r}\left(\int_{\mathbb{R}^d}\frac{|\widehat{\mu}(\omega)-\widehat{\nu}(\omega)|^2}{\|\omega\|^{\varphi r^*}}\mathrm{d}\omega\right)^{1/r^*}.\end{aligned} \tag{22}$$

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 5. 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

$$\sqrt{\gamma(2-\gamma)}\|\widehat{\mu}-\widehat{\nu}\|_2^{\frac{2\beta+d+\gamma}{2\beta}} \leq C_1\cdot(\|\mu\|_{\beta,2}^2+\|\nu\|_{\beta,2}^2)^{\frac{d+\gamma}{4\beta}}\cdot\mathcal{E}_\gamma(\mu,\nu),$$

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

Proof of (21). 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 30. Let μ and ν have densities $p*\phi$ and $q*\phi$, where ϕ is the density of $\mathcal{N}(0,I_d)$. Writing $f=(p-q)$, we can extend the proof of (Jia et al., 2023, 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

$$\mathsf{TV}(\mu,\nu)\leq C(d)\|\mu-\nu\|_2\log(3+1/\|\mu-\nu\|_2)^{d/4}. \tag{23}$$

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 5. Using that $|\widehat{f}|\leq|\widehat{\phi}|$ and applying Lemma 30, for $\beta\geq 1$ 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

$$\mathcal{E}_\gamma(\mu,\nu) \geq \frac{C(d)\sqrt{\gamma(2-\gamma)}\|f\|_2}{\log(3+1/\|f\|_2)^{\frac{d+\gamma}{4}}} \geq \frac{C(d)\sqrt{\gamma(2-\gamma)}\mathsf{TV}(\mu,\nu)}{\log(3+1/\mathsf{TV}(\mu,\nu))^{\frac{2d+\gamma}{4}}}, \tag{24}$$

where the second inequality uses (23) and Lemma 26. ■

Theorem 11 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, (20) 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 11, 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.

Proposition 12. *For any $\beta > 0$, $d \geq 1$, $\gamma \in (0,2)$ and $C > 0$ satisfying Assumption 1, there exists a finite constant $C_1 = C_1(\beta, d, \gamma, C)$ 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*

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

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

Proposition 13. *When $\gamma = 1$ we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in Proposition 12.*

Proposition 13 is an improvement over Proposition 12 due to the inequality $d_H \lesssim \overline{d_H}$ over the class $\mathcal{P}_S(\beta, d, C)$, which follows from Proposition 3. 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.

The proofs of both results are presented in Section E. The general idea is to saturate Hölder's inequality in (22), for which the Fourier transform of $f = \frac{\mathrm{d}\nu}{\mathrm{d}x} - \frac{\mathrm{d}\mu}{\mathrm{d}x}$ should be supported on a sphere. However, such f clearly cannot be compactly supported. Thus the actual construction is to multiply the Fourier inverse of the uniform measure on a sphere with a compactly supported mollifier. In $d > 1$ the mollifier that we require must have super-polynomial Fourier spectrum decay, for which we use the recent construction in Cohen (2023).

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 does not 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 14. *Let μ and ν be probability distributions supported on the set $\{x_1, \dots, x_k\} \subseteq \mathbb{R}^d$ 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 (Ball, 1992, Theorem 1) we have

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

■

Remark 15. *Similar results can be proved for the generalized energy distance $\mathcal{E}_\gamma$, using e.g. the work Narcowich and Ward (1992). However, to the best of our knowledge, these estimates degrade significantly in the dimension d in contrast with Ball (1992).*

Notice that by our discussion above, the support set $\{x_1, \dots, x_k\}$ in Theorem 14 is arbitrary and may be chosen by us. Since the scale of the supporting points $x_1, \dots, x_k$ 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 $x_1, \dots, x_k$ 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 16. *There exists a universal constant $C \in (0, \infty)$ such that for any $k \geq 1$ there exists a set of points $x_1, \dots, x_k \in \mathbb{R}^{\lceil C \log(k) \rceil}$ 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 $\mu = (\mu_1, \dots, \mu_k)$ and $\nu = (\nu_1, \dots, \nu_k)$.

The question arises how the set of points $x_1, \dots, x_k$ in Corollary 16 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 d (known as "blocklength" in this context). 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 Justesen (1972) for one such example, and random codes are almost surely good (Barg and Forney, 2002). Clearly the better the code is, the better the constants we obtain in Corollary 16.

Remark 17. *One interesting consequence of Corollary 16 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 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

$$\tilde{\nu}_\gamma \in \operatorname*{arg\,min}_{\nu' \in \mathcal{G}} \mathcal{E}_\gamma(\nu', \nu_n), \tag{25}$$

where $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Note that $\tilde{\nu}_\gamma$ does not quite agree with our definition of $\tilde{\nu}_\gamma$ in (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 (25), 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 4.2 we present the adapation of these to $\overline{d_H}$, thereby proving Theorem 1.

Theorem 18. *Let $\tilde{\nu}_\gamma$ be the estimator defined in* (25)*. For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant $C_1 = C_1(\beta, d, C)$ so that*

$$\sup_{\nu \in \mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_1 (n\gamma(2-\gamma))^{-\frac{\beta}{2\beta+d+\gamma}} \tag{26}$$

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 $C_2 = C_2(d)$ such that

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_2 \phi\left((n\gamma(2-\gamma))^{-1/2}\right) \tag{27}$$

holds for $\mathcal{G} = \mathcal{P}_G(d)$ and any $\gamma \in (0, 2)$, where $\phi(x) = x \cdot \log(3 + 1/x)^{\frac{2d+\gamma}{4}}$.

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 $\tilde{\nu}_\gamma$. Writing $C_1 = C_1(\beta, d, C)$

for a finite constant that we relabel freely, the first claim is substantiated by the chain of inequalities

$$\begin{aligned}
\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\overset{\text{Thm. 11}}{\leq} \mathbb{E}\Big[\left(C_1 \frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\Delta-\text{ineq.}}{\leq} \mathbb{E}\Big[\left(C_1 \frac{\mathcal{E}_\gamma(\nu, \nu_n) + \mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\text{Eq. (25)}}{\leq} \mathbb{E}\Big[\left(2C_1 \frac{\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t}\Big] \\
&\overset{\text{Jensen's}}{\leq} \left(2C_1 \frac{\mathbb{E}\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right)^{1/t} \\
&\overset{\text{Lem. 7}}{\leq} \left(\frac{n\gamma(2-\gamma)}{8C_1^2}\right)^{-1/2t}.
\end{aligned}$$

The result for $\mathcal{G} = \mathcal{P}_G$ follows analogously. Define $r(x) = x \cdot \log(3 + 1/x)^{-t}$ where $t = \frac{2d+\gamma}{4} > 0$. By direct calculation, one can check 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. Similarly to the case of smooth distributions covered above, we obtain the chain of inequalities as follows:

$$\begin{aligned}
\mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) &\overset{\text{Thm.11}}{\leq} \mathbb{E}\Big[r^{-1}\left(C_2 \frac{\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right)\Big] \\
&\overset{\text{Jensen's}}{\leq} r^{-1}\left(C_2 \frac{\mathbb{E}\mathcal{E}_\gamma(\tilde{\nu}_\gamma, \nu)}{\sqrt{\gamma(2-\gamma)}}\right) \\
&\overset{\text{Eqn. 25}}{\leq} r^{-1}\left(2C_2 \frac{\mathbb{E}\mathcal{E}_\gamma(\nu, \nu_n)}{\sqrt{\gamma(2-\gamma)}}\right) \\
&\overset{\text{Lem.7}}{\leq} r^{-1}\left(C_2(n\gamma(2-\gamma))^{-1/2}\right).
\end{aligned}$$

The conclusion follows by Lemma 26. ■

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 by adaptively setting γ to be dependent on n is the following.

Corollary 19. *For any $\beta > 0$, $d \geq 1$ and $C > 0$, there exists a finite constant $C_1 = C_1(\beta, d, C)$ such that: for all $n \geq 2$, with $\tilde{\nu}_{\gamma_n} \in \arg\min_{\nu' \in \mathcal{G}} \mathcal{E}_{\gamma_n}(\nu', \nu_n)$ in the setting of* (25) *where $\gamma_n = \log(n)^{-1}$ and $\mathcal{G} = \mathcal{P}_S(\beta, d, C)$, one has:*

$$\sup_{\nu \in \mathcal{P}_S(\beta,d,C)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_{\gamma_n}, \nu) \leq C_1 \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}. \tag{28}$$

Similarly, for any $d \geq 1$ there is a finite constant $C_2 = C_2(d)$ such that for all $n \geq 3$, with $\tilde{\nu}_{\gamma_n} \in \arg\min_{\nu' \in \mathcal{G}} \mathcal{E}_{\gamma_n}(\nu', \nu_n)$ where $\gamma_n = \log\log(2n)^{-1}$ and $\mathcal{G} = \mathcal{P}_G(d)$, one has:

$$\sup_{\nu \in \mathcal{P}_G(d)} \mathbb{E}\mathsf{TV}(\tilde{\nu}_\gamma, \nu) \leq C_2 \log(n)^{d/2} \sqrt{\log\log n}/\sqrt{n}. \tag{29}$$

4.2 Proof of Theorem 1 and Theorem 2

We already have everything needed to deduce Theorem 2. Since it is an exercise in combining results, we simply list the required steps:

1. Use Theorem 11 to get a comparison between TV and $\mathcal{E}_\gamma$.
2. Set $\gamma = 1$ and use Theorem 4 to get the equivalence between $\mathcal{E}_1$ and d_H.
3. Use Proposition 3 to get a comparison between d_H and $\overline{d_H}$.

Turning to the proof of Theorem 1, we find that it is completely analogous to the proof of Theorem 18, with the only difference being that we can no longer rely on Lemma 7 to show that the distance between empirical and population measures decays at the parametric rate, as the latter applies to $\mathcal{E}_\gamma$ instead of $\overline{d_H}$. However, the corresponding result for $\overline{d_H}$ is well known.

Lemma 20. *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 for example from (Vershynin, 2018, 8.3.23) and the fact that $\mathcal{D}_a$, the family of half-space indicators, has VC dimension $d + 1$. ■

With Lemma 20 in hand, completing the argument is straightforward: To deduce Theorem 1 follow the same steps as in the proof of Theorem 18, except use Theorem 2 and Lemma 20 in place of Theorem 11 and Lemma 7 respectively.

4.3 Estimating Discrete Distributions

In many practical machine learning tasks the data is discrete, albeit on a large alphabet $[k] = \{1, 2, \ldots, k\}$: for example, in recommender systems the alphabet could be all possible ads, products or articles. A common idea to apply modern learning pipelines to such data is to use an embedding $E : [k] \to \mathbb{R}^d$, with "one-hot" encoding ($d = k$) being the most popular choice. After such an embedding, the data is effectively made "continuous" and the density estimation methods as discussed previously can be applied. Can such an approach be good in the sense of minimax estimation guarantees? We answer this question positively in this section, provided that embedding E comes from an error-correcting code.

Let $\mathcal{P}_k$ denote the set of all probability distributions on the set $[k]$. Suppose we observe an i.i.d. sample $X_1, \ldots, X_n \sim \nu$ from some unknown distribution $\nu \in \mathcal{P}_k$. The problem of estimating ν is effectively trivial: the empirical distribution provides a minimax optimal estimator. Indeed, it is

a folklore fact, see for example (Canonne, 2020, Theorem 1) or (Polyanskiy and Wu, 2023+, Exc. VI.8), that the optimal rate of estimation is given by

$$\sup_{\nu\in\mathcal{P}_k} \mathbb{E}\mathsf{TV}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \nu\right) \asymp \min\left\{\frac{k}{n}, 1\right\}. \tag{30}$$

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 $i \in [k]$.

Theorem 21. *There exists a universal constant $C < \infty$ with the following property. For any alphabet size k there exist embedding points $a_1, \ldots, a_k \in \mathbb{R}^{\lceil C\log(k)\rceil}$ such that given an i.i.d. sample $X_1, \ldots, X_n \sim \nu$ from an unknown $\nu \in \mathcal{P}_k$, any estimator $\tilde{\nu} \in \mathcal{P}_k$ that satisfies*

$$\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i \delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}\right) \leq \frac{c}{n} \tag{31}$$

for any c enjoys the performance guarantee

$$\sup_{\nu\in\mathcal{P}_k} \mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \leq C\min\left\{(c+1)\frac{k\sqrt{\log(k)}}{n}, 1\right\}. \tag{32}$$

Moreover, we may replace $\mathcal{E}_1$ by $\overline{d_H}$ in (31) *and the result* (32) *remains true with $\sqrt{\log(k)}$ replaced by $\log(k)$.*

Proof Let $a_1, \ldots, a_k \in \mathbb{R}^d$ be the points defined in Corollary 16 (relabeled from $x_1, \ldots, x_k$ for clarity) so that $d \asymp \log(k)$. By the triangle inequality we have

$$\begin{aligned}\mathbb{E}\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i\delta_{a_i}, \sum_{i=1}^k \nu_i\delta_{a_i}\right) &\leq 2\mathbb{E}\mathcal{E}_1^2\left(\sum_{i=1}^k \tilde{\nu}_i\delta_{a_i}, \frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}\right) + 2\mathbb{E}\mathcal{E}_1^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{a_{X_i}}, \sum_{i=1}^k \nu_i\delta_{a_i}\right)\\ &\overset{\text{Lemma 7}}{\lesssim} \frac{c + \max_i \|a_i\|}{n} \lesssim \frac{c+1}{n}.\end{aligned}$$

By Corollary 16, the definition of $\tilde{\nu}$ and the triangle inequality it follows that

$$\mathbb{E}\mathsf{TV}^2(\tilde{\nu}, \nu) \lesssim \frac{k\sqrt{d}(c+1)}{n} \asymp \frac{k\sqrt{\log(k)}(c+1)}{n}.$$

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

$$\frac{\mathsf{TV}}{\sqrt{k}\sqrt[4]{d}} \overset{\text{Cor. 16}}{\lesssim} \mathcal{E}_1 \overset{\text{Prop. 4}}{\asymp} \frac{\sqrt{\Gamma\left(\frac{d+1}{2}\right)}}{\pi^{(d-1)/4}} d_H \overset{\max_i \|a_i\|\leq 1}{\lesssim} \overline{d_H},$$

and Lemma 20 in place of Lemma 7, which is where we loose the $\sqrt{d} \asymp \sqrt{\log(k)}$ factor. ■

As we explained, the empirical distribution $\tilde{\nu} = \frac{1}{n}\sum_{i=1}\delta_{X_i}$ achieves optimality in (30), and clearly also achieves $c = 0$ in (31) i.e. minimizes the empirical risk globally. The point of Theorem 21 is to show that approximate minimizers, such as those found via SGD, are also nearly minimax optimal.

4.3.1 ESTIMATING HÖLDER SMOOTH DENSITIES

Theorem 21 has interesting implications for density estimation over the class of distributions on the cube $[0,1]^d$ with uniformly bounded derivatives up to order $\underline{\beta} \triangleq \lceil \beta - 1 \rceil$ and $(\beta - \underline{\beta})$-Hölder continuous $\underline{\beta}^{th}$ derivative; call such distributions simply β-Hölder smooth.[5] Writing B_j for the cube with center $(j - \frac{1}{2})\epsilon^{1/\beta}$ and sidelength $\epsilon^{1/\beta}$ where $j \in \{1, \ldots, \epsilon^{-1/\beta}\}^d$, it is known that

$$\sum_j \left| \int_{B_j} (f(x) - g(x)) \mathrm{d}x \right| = c \int_{[0,1]^d} |f(x) - g(x)| \mathrm{d}x + O(\epsilon)$$

for any β-Hölder smooth densities f, g and an ϵ-independent constant $c > 0$, see for example (Arias-Castro et al., 2018, Lemma 7.2) or (Ingster and Suslina, 2003, Proposition 2.16). In other words, discretizing such distributions using a regular grid with $\Omega(\epsilon^{-d/\beta})$ cells maintains total-variation distances up to an additive $O(\epsilon)$ error.

Now, consider a 'multilayer perceptron', that is, a fully connected multilayer neural network with activations given by $x \mapsto \mathbb{1}\{x \geq 0\}$. Such a multilayer network with large enough hidden layers can in principle implement the discretization described above, and embed the $\epsilon^{-d/\beta}$ cells as an error correcting code. Thus, due to Theorem 21, the ERM density estimator (2) would achieve the minimax optimal density estimation rate $n^{-\beta/(2\beta+d)}$ over β-Hölder 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 (with one neuron) perceptron then the best possible rate is $n^{-\beta/(2\beta+d+1)}$.

4.4 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.

Lemma 22. *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 $C_1 = C_1(\Omega, \gamma) > 0$ such that*

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

In other words, $\mathcal{E}_\gamma(\nu, \nu_n)$ is $O(1/n)$-subGaussian.

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

$$\sup_{x,x',y,y' \in \mathrm{supp}(\nu)} (k_\gamma(x,y) - k_\gamma(x',y')) \lesssim \mathrm{diam}(\Omega)^\gamma.$$

5. Note that this class is not the same as $\mathcal{P}_S$, although related.

Therefore, by McDiarmid's inequality we know that $\mathcal{E}_\gamma(\nu, \nu_n)$ is $O(1/n)$-subGaussian (note we don't track constants depending on Ω here). From Lemma 7 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 $X_1, \ldots, X_n$ 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 class of generators. 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 $1-\delta$? Note that an additional difficulty here is that our generative model for μ_k is able to generate the samples from μ_k but otherwise gives us no other access to μ_k. The fast (dimension-free) concentration properties of $\mathcal{E}_\gamma$ and the minimax optimality guarantees of its minimizer (whenever ν is smooth) make it an excellent choice for such a stopping criteria.

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 $X_1, \ldots, X_n$. Finally, given the training sample $(X_1, \ldots, X_n)$, 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 $(X_1^{(k)}, \ldots, X_{m_k}^{(k)}) \sim \mu_k^{\otimes m_k}$.

Proposition 23. *For any $\beta > 0, d \geq 1$ and $\gamma \in (0,2)$ there exists a constant $c = c(\beta, d, \gamma)$ 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 $c = C_1$ where C_1 is as in Lemma 22 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 $t_1, t_2, \ldots$, and each $k \geq 1$. By Lemma 22,

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

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

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

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

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

The final conclusion follows from Theorem 11. ■

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 18 that we are a constant factor away from (near-)minimax optimality with probability $1 - \delta$.

5. Suboptimality for Two-Sample Testing

So far in this paper we have shown how the empirical energy distance minimizer, while being mismatched with the target total variation loss, nevertheless achieves nearly minimax optimal performance for density estimation tasks. Unfortunately, this surprising effect does not carry over to other statistical tasks, such as two-sample testing, which we describe in this section.

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 (1987), who showed that in dimension $d = 1$ the problem is solvable with probability $1 - o(1)$ if and only if

$$n = \omega(\epsilon^{-\frac{2\beta + d/2}{\beta}}), \tag{33}$$

in which case a variant of the χ^2-test works. The case of general m, n and $d \ge 1$ was resolved in Arias-Castro et al. (2018) who showed that the problem is solvable if and only if (33) holds with n replaced by $\min\{n, m\}$, using the very same χ^2-test; see also Li and Yuan (2019). In the remainder of the section we focus on the $m = n$ case for simplicity.

In a recent paper (Paik et al., 2023), 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

$$T_{d,k}(p_n, q_n) \geq t_n, \tag{34}$$

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 results (Paik et al., 2023, Theorem 6) asserts that the test (34) returns the correct hypothesis with probability $1 - o(1)$ 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 (33) 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 24. *For all $d, \beta > 0$, there exists constants $c = c(d,\beta), c' = c'(d,\beta)$ such that for all $\varepsilon > 0$, there exists probability density functions p, q supported on the d-dim unit ball such that*

1. *$\|p\|_{\beta,2}, \|q\|_{\beta,2} < c$,*
2. *$\|p - q\|_1 \asymp \|p - q\|_2 \asymp \varepsilon$, and*
3. *the expected test statistic satisfies*

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

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

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 (33) attainable by the χ^2 test. The proof of Proposition 24 is given at Section F.

6. Conclusion

We analyzed the simple discriminating class of affine classifiers and proved its effectiveness in the ERM-GAN setting (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 18 and corollary 19) 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 (Section E). 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 $\tilde{\nu}$ defined in (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?

Acknowledgments

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

Appendix A. Related works showing "neural discriminator" implies distance is small

In Section 1.2 we already reviewed at a high level existing results on GANs and density estimation. Here we provide full details and emphasize differences with our work.

A.1 Results on smooth pushforward densities

A range of works has focused on showing results for not smooth densities but smooth pushforwards $g_{\#}U$ (for example $U \sim \mathrm{Unif}([0,1]^d)$ and g is β-Hölder smooth). While the exact setting differs, it can be shown that at least for TV estimation on the open unit cube (or unit ball), smooth pushforward estimations imply smooth density estimation. Indeed, one can first assume without loss of generality that the target density is bounded below via a linear transform $f \to \frac{1}{2}(\mathrm{unif} + f)$ (see e.g. Lemma 32.18 in Polyanskiy and Wu (2023+)) and bounded above from smoothness. Furthermore, for any two (Hölder) β-smooth densities bounded above and below on (open) uniformly convex domains, the unique optimal transport map with respect to the quadratic cost is $(\beta+1)$-smooth (see Lemma 10 in Chae et al. (2023) as well as Theorem 4.14 in Villani (2003)). Therefore, these rates could be compared to our results.

1. (Concurrent with ours) Stéphanovitch et al. (2023), Model 1 and 3.

   This model assumes that the true density is $g_{\#}U$ for U uniform on d-dim torus and g is $(\beta+1)$-Hölder smooth from the torus to $\mathbb{R}^p$. The authors obtain optimal rates with respect to discriminator class consisting of γ-smooth functions for $\gamma \geq 1$ (the case of $\gamma = 1$ matches with Wasserstein W_1). Under the assumption that the target density is β-smooth, results for comparing W_1 and TV (Chae (2024) and Corollary 4.4 in Stéphanovitch et al. (2023)) apply. By substituting in the rate of W_1 distance ($\gamma = 1$) in Theorem 5.8 (Model 3 therein) $\widetilde{O}\left(n^{-\frac{\beta+1}{2\beta+d}}\right)$ with Chae and Walker (2020), one gets the optimal rate $\widetilde{O}\left(n^{-\frac{\beta}{2\beta+d}}\right)$. However, their discriminator crucially rely on knowledge of parameters n, β (see (4.5) therein)

2. Chae (2022)

   The authors obtained results for W_1 distance of smooth pushforwards where the observed samples are Gaussian-noised with variance σ. In the zero-noise case, their convergence results in W_1 on β-Hölder smooth pushforward functions give slightly suboptimal rates of $O\left(n^{-\frac{\beta}{2\beta+d}}\right)$ (Theorem 3.2 therein) compared to the (conjectured) optimal rate $O\left(n^{-\frac{\beta}{2\beta+d-2}}\right)$ (Theorem 4.1 therein). As above, their results on W_1 could be applied to total variation in this setting by known distance comparison inequalities. Furthermore, their discriminator is a large non-parametric class (not a neural network).

3. Belomestny et al. (2021)

   The authors studied the GAN estimator on $U \sim \mathrm{Unif}([0,1]^d)$ and the pushforward map g is Λ-regular and $(\beta+1)$-Hölder smooth (despite having the same minimax rate, this class do *not* cover all bounded β-smooth densities, see discussion after Theorem 3 therein). Their results (Theorem 2) obtained (log)optimal JS divergence rates, equivalent to TV since $\mathsf{TV}^2 \lesssim \mathsf{JS}$. In contrast to our work, their discriminator class is essentially a non-parametric class (of $\frac{p(x)}{q(x)+p(x)}$, where p, q range over pairs of smooth densities) that is approximated by giant

neural networks, for which they assumed existence of an oracle approximating smooth discriminator functions over those networks (Belomestny et al. (2023)). Furthermore, this discriminator network needs to be larger than some function of β and n whereas ours is fixed.

A.2 Results on smooth densities

There is also a large literature on GAN estimation of smooth densities (with slightly varying definitions) that are useful to review for contrasting with our own work.

1. Singh et al. (2018) and Uppal et al. (2019)

   While these works focus on projection density estimators, Theorem 7 in Singh et al. (2018) and Theorem 9 in Uppal et al. (2019) have shown smooth density estimation rates on GANs. The authors have shown that if the discriminator class (functions of Sobolev smoothness and Besov smoothness s, resp.) is replaced with a neural network then the GAN estimator has optimal rate if the neural networks well approximate the smoothness class, for which known results such as Yarotsky (2017) applies. However, bounds on TV or W_1 are not directly covered in their theorems.

2. Chen et al. (2020)

   The authors considered Hölder-smooth densities on convex support that are bounded below. They obtained sub-optimal rates $O\left(n^{-\frac{\gamma}{2\gamma+d}} \vee n^{-\frac{\beta+1}{2(\beta+1)+d}}\right)$ on the IPM for Hölder smooth functions $\mathcal{H}^\gamma, \gamma \geq 1$ (where $\gamma = 1$ transfers to Lipschitz functions with W_1 being the respective IPM) versus the minimax rate $O\left(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}}\right)$. They also had to rely on (oracle) neural network universal approximating discriminators that depends on β.

To conclude this section, we mention that there are also several recent works on non-GAN estimators that we omitted, such as (Stéphanovitch et al., 2023, Model 2), (Liang, 2021, Theorem 4), Divol (2022) and Tang and Yang (2023). We recommend survey in Section 2.3 and Table 1 of Stéphanovitch et al. (2023).

Appendix B. Auxiliary Technical Results

In this section we list some technical lemmas that are used in our later proofs.

Theorem 25 (Plancherel theorem)**.** *Let $f, g \in L^2(\mathbb{R}^d)$. Then*

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

Lemma 26. *Suppose $t, x, y > 0$. Then there exist finite t-dependent constants C_1, C_2 such that*

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

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

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

The second inequality 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 27. *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

$$\begin{aligned} 4\mathbb{E} \int_{\mathbb{R}^d} \frac{1 \wedge (\|\omega\|^2 \|X\|^2)}{\|\omega\|^{d+\gamma}} \mathrm{d}\omega &\le 4 \operatorname{vol}_{d-1}(\mathbb{S}^{d-1}) \mathbb{E} \int_0^\infty \frac{1 \wedge (r^2\|X\|^2)}{r^{1+\gamma}} \mathrm{d}r \\ &= \frac{8\pi^{d/2}}{\Gamma(\frac{d}{2})} \mathbb{E} \left\{ \|X\|^2 \int_0^{\|X\|^{-1}} \frac{1}{r^{\gamma-1}} \mathrm{d}r + \int_{\|X\|^{-1}}^\infty \frac{1}{r^{1+\gamma}} \mathrm{d}r \right\} \\ &= \frac{16\pi^{d/2} M_\gamma(\nu)}{\Gamma(\frac{d}{2})\gamma(2-\gamma)}, \end{aligned}$$

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

Lemma 28. *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 $B_\gamma \leq \int_0^\infty \left| \frac{\min\{1,\omega^2/2\}}{\omega^{(1+\gamma)/2}} \right| \mathrm{d}\omega < \infty$. Now let us consider $\gamma \leq 1$. One has that $B_\gamma = \sup_{0<c_1<c_2} |I_\gamma(c_2) - I_\gamma(c_1)| \leq \sup_{c>0} 2|I_\gamma(c)|$ where

$$I_\gamma(a) = \begin{cases} \int_1^a \frac{\sin(\omega)}{\omega} \mathrm{d}\omega & \text{if } \gamma = 1, \\ \int_1^a \frac{\cos(\omega)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \text{if } \gamma \in (0,1), \end{cases}$$

Clearly $a \mapsto I_\gamma(a)$ is continuous on $(0,\infty)$ for each $\gamma \in (0,2)$. Moreover,

$$|I_\gamma(a)| \leq |a-1| \cdot \max\{a^{-(\gamma+1)/2}, 1\}.$$

Therefore, we only need to show that $\lim_{a\to\infty} |I_\gamma(a)|$ and $\lim_{a\to 0} |I_\gamma(a)|$ are both finite. Let $g_\gamma(x) = x^{-(\gamma+1)/2}$ and $f_\gamma(x) = \sin(x)$ if $\gamma = 1$ and $\cos(x)$ otherwise, so that we may write $I_\gamma(a) = \int_a^1 f_\gamma(\omega) g_\gamma(\omega) \mathrm{d}\omega$.

1. For $a \to \infty$, since $g_\gamma(\infty) = 0$ and $|\int_1^a f_\gamma(x)\mathrm{d}t| \leq 2$ is uniformly bounded, $\int_1^\infty f(\omega)g(\omega)\mathrm{d}\omega$ converges to a finite value by Dirichlet's test for improper integrals (Malik and Arora, 1992, page 391).[6]

2. For $a \to 0$, the conclusion follows by the inequality $|\sin(\omega)| \leq \min\{|\omega|, 1\}$ in the case $\gamma = 1$, and by $|I_\gamma(a)| \leq \int_a^1 \omega^{-(1+\gamma)/2}\mathrm{d}\omega \leq \int_0^1 \omega^{-(1+\gamma)/2}\mathrm{d}\omega = \frac{2}{1-\gamma}$ for $\gamma < 1$.

Therefore, $\sup_{a>0} |I_\gamma(a)| < \infty$ which concludes the proof.

■

Lemma 29. *Let* $\int_0^\infty \cdot \, \mathrm{d}\omega \triangleq \lim_{\epsilon\to 0} \int_{1/\epsilon \geq \omega \geq \epsilon} \cdot \, \mathrm{d}\omega$ *and recall the definition of* ψ_γ *from* (17)*. Then, for* $x \neq 0$ *the following hold:*

$$\psi_\gamma(x) = C_{\psi_\gamma} \begin{cases} \int_0^\infty \frac{\sin(\omega x)}{\omega} \mathrm{d}\omega + \frac{\pi}{2} & \textit{for } \gamma = 1, \\ \int_0^\infty \frac{\cos(\omega x)}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \textit{for } \gamma \in (0,1), \\ \int_0^\infty \frac{\cos(\omega x) - 1}{\omega^{(1+\gamma)/2}} \mathrm{d}\omega & \textit{for } \gamma \in (1,2), \end{cases}$$

where

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

Proof For $x \neq 0$ clearly

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

6. Reference pointed out by user Siminore on math.stackexchange.com.

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 30. *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 well known that $\widehat{\phi}(\omega) = e^{-\frac{\sigma^2}{2}\|\omega\|^2}$. The claimed equality then follows from the formula for the 2β'th moment of the Gaussian distribution with mean 0 and variance $1/(2\sigma^2)$. The inequality follows by Lemma 31.

■

Lemma 31 (Properties of the gamma function)**.** *For all $x > 1$ the inequality $\Gamma(x) \leq 5(x/e)^{x-1/2}$ holds.*

Proof In Minc and Sathre (1964) 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 $\exp(\frac{1}{2}\log(2\pi)+1/2) < 5$. ■

Lemma 32. *Let $b \geq 1$ and $|a| < b$. Then*

$$\int_0^r x^a |\sin(x)|^b \mathrm{d}x = O(r^{a+b})$$

as $r \to \infty$, where we hide constants depending on a, b.

Proof Since we are only interested in the asymptotic behaviour as $r \to \infty$, assume without loss of generality that $r \geq 1$. Then, we have

$$\int_0^r x^a |\sin(x)|^b \mathrm{d}x = \underbrace{\int_0^1 x^a |\sin(x)|^b \mathrm{d}x}_{I} + \underbrace{\int_1^r x^a |\sin(x)|^b \mathrm{d}x}_{II}.$$

Using the inequality $|\sin(x)| \leq x$, we can bound the first term as

$$I \leq \int_0^1 x^{a+b} \mathrm{d}x = \frac{1}{a+b+1} \leq 1 = O(r^{a+b}),$$

since $a + b > 0$. For the second term, we obtain

$$II \leq \int_1^r x^a \mathrm{d}x = \left\{ \begin{matrix} \frac{r^{a+1}-1}{a+1} & \text{if } a \neq -1 \\ \log(r) & \text{if } a = -1 \end{matrix} \right\} \quad = O(r^{a+b}),$$

where the last step uses $a + b > 0$ and $b \geq 1$. ■

Lemma 33. *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

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

Therefore, for large enough r,

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

■

Lemma 34. *Let J_ν be the Bessel function of the first kind of order ν.*

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

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

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

$$J_\nu(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + O(x^{-3/2}). \tag{35}$$

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

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

Proof For Item 1 set $r = \|x\|$ and $s = (2\pi)^{-1}$ in the calculation on page 154 of Stein and Weiss (1971).

For Item 2 see (Watson, 1980, Eq. (1) in Section 7.21).

For Item 3, we can compute

$$\begin{aligned}\int_{\|w\|\le 1} e^{i\langle x,w\rangle}\mathrm{d}w &= \int_{-1}^{1} e^{i\|x\|w_1} \int_{w_2^2+\cdots+w_d^2 \le 1-w_1^2} \mathrm{d}w_2 \ldots \mathrm{d}w_d \mathrm{d}w_1 \\ &= \frac{\pi^{(d-1)/2}}{\Gamma(\frac{d+1}{2})} \int_{-1}^{1} e^{i\|x\|w_1}(1-w_1^2)^{(d-1)/2}\mathrm{d}w_1. \end{aligned} \tag{36}$$

Recall from (Watson, 1980, Section 3.1) the definition of the Bessel function of the first kind as

$$J_\nu(x) = \frac{(x/2)^\nu}{\Gamma(\nu+\frac{1}{2})\Gamma(\frac{1}{2})} \int_0^\infty \cos(x\cos(\theta))\sin^{2\nu}(\theta)\mathrm{d}\theta$$

valid for $\nu > -1/2$; the above is also known as the Poisson representation. Changing variables to $u = \cos(\theta)$, we see that it is equal to

$$J_\nu(x) = \frac{(x/2)^\nu}{\Gamma(\nu+\frac{1}{2})\Gamma(\frac{1}{2})} \int_{-1}^{1} e^{ixu}(1-u^2)^{\nu-1/2}\mathrm{d}u. \tag{37}$$

Comparing (37) with (36) concludes the proof. ∎

Lemma 35. *There exists a radial function $h_0 \in L^2(\mathbb{R}^d)$ such that*

$$\begin{aligned} &\mathrm{supp}(h_0) \subseteq \mathbb{B}(0,1), \\ &|\widehat{h}_0(w)| \le C\exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) && \text{for all } w \in \mathbb{R}^d, \\ &|\widehat{h}_0(w)| \ge \frac{1}{2} && \text{for all } \|w\| \le r_{\min}, \end{aligned}$$

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

Proof Apply Theorem 1.4 in Cohen (2023) using the spherically symmetric weight function $u : \mathbb{R}^d \to \mathbb{R}_{\le 0}$ 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 $a \in \mathbb{R}$. ∎

Appendix C. Proof of Theorem 8

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

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\gamma(\langle X, v\rangle - b) - \mathbb{E}\psi_\gamma(\langle Y, v\rangle - b)\Big]^2 \mathrm{d}b\mathrm{d}\sigma(v) \\ &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \psi_\gamma(x-b)\mathrm{d}\eta_v(x)\right)^2 \mathrm{d}b\mathrm{d}\sigma(v),\end{aligned} \tag{38}$$

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 $\sigma \otimes \mathrm{Leb}$-almost every (v, b), thus going forward we can focus on the case $x \neq b$. By Lemma 29, and writing $A_\epsilon = [\epsilon, 1/\epsilon]$ for $\epsilon > 0$, we have

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

Note that in the $\gamma = 1$ case we implicitly used that $\int \mathrm{d}\eta_v(x) = 0$. 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 28. We now show that $\int_{\mathbb{R}} |x-b|^{(\gamma-1)/2}\mathrm{d}|\eta_v|(x) < \infty$ for $\sigma \otimes \mathrm{Leb}$-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

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

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 $\widehat{\eta}_v(0) = 0$,

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

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 27. 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

$$\left\|\widehat{\Psi}_{\gamma,v,\epsilon} - \widehat{\Psi}_{\gamma,v}\right\|_2 \to 0 \tag{39}$$

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.[7] Therefore, by passing to this subsequence, it follows that

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

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

$$\begin{aligned} \Re\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) + \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2} \\ &= \frac{1}{2}\int_0^\infty \left(\frac{\widehat{\eta}_v(\omega)}{\omega^{(1+\gamma)/2}} e^{ib\omega} + \frac{\widehat{\eta}_v(-\omega)}{\omega^{(1+\gamma)/2}} e^{-ib\omega}\right)\mathrm{d}\omega \\ &= \frac{1}{2}\int_{-\infty}^\infty \frac{\widehat{\eta}_v(\omega)\,\mathrm{sign}(\omega)}{|\omega|^{(1+\gamma)/2}} e^{ib\omega}\mathrm{d}\omega \\ &= \mathcal{F}\left[\frac{\widehat{\eta}_v(\omega)\,\mathrm{sign}(\omega)}{2|\omega|^{(1+\gamma)/2}}\right](-b), \end{aligned} \tag{40}$$

$$\begin{aligned} \Im\left(\widehat{\Psi}_{\gamma,v}(b)\right) &= \frac{\widehat{\Psi}_{\gamma,v}(b) - \overline{\widehat{\Psi}_{\gamma,v}(b)}}{2i} \\ &= \frac{1}{2i}\int_0^\infty \left(\frac{\overline{\widehat{\eta}_v(\omega)}}{\omega^{(1+\gamma)/2}} e^{-ib\omega} - \frac{\overline{\widehat{\eta}_v(-\omega)}}{\omega^{(1+\gamma)/2}} e^{ib\omega}\right)\mathrm{d}\omega \\ &= \frac{1}{2i}\int_{-\infty}^\infty \frac{\overline{\widehat{\eta}_v(\omega)}}{|\omega|^{(1+\gamma)/2}}\,\mathrm{sign}(\omega) e^{ib\omega}\mathrm{d}\omega \\ &= \mathcal{F}\left[\frac{\overline{\widehat{\eta}_v(\omega)}\,\mathrm{sign}(\omega)}{2i|\omega|^{(1+\gamma)/2}}\right](-b). \end{aligned} \tag{41}$$

7. We could also conclude this by Carleson's theorem.

Plugging (40) and (41) into (38), by Parseval's identity (implicitly using that $\Psi_{\gamma,v} \in L^2(\mathbb{R})$), we obtain

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

where the last step uses a polar change of variable. The result follows after comparing with Proposition 5.

Appendix D. Proof of Proposition 9

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 $s = (d+\gamma)/2$. First, note that

$$\int \frac{K_s(x)\mathrm{d}x}{(1+\|x\|^2)^d} < \infty,$$

so by (Landkof, 1972, Theorem 0.10) we have $K_s \in \mathcal{S}'(\mathbb{R}^d)$. By (Landkof, 1972, 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 5.

Appendix E. Proof of Theorem 12 and Proposition 13

In this section we prove both Proposition 12 and Proposition 13. To do so, we give two constructions. The first one, presented in Section E.1, only applies in one dimension and gives optimal results. The second construction is given in Section E.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 $\|f\|_{t,2} = \||\cdot|^t \widehat{f}\|_2$ for potentially negative exponents $t \in \mathbb{R}$. Note that $\mathcal{E}_\gamma(f,0) = \sqrt{F_\gamma(d)}\|\widehat{f}\|_{-\frac{d+\gamma}{2},2}$.

E.1 The Case $d = 1$

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

Lemma 36. *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 $\beta \ge 1$ and let $|t| < \beta$. We have*

$$\|f_\beta\|_{t,2} \asymp r^t, \|f_\beta\|_1 \asymp 1, \text{ and } \overline{d_H}(f_\beta, 0) \asymp \frac{1}{r}, \tag{42}$$

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

Proof The intuition for the estimates (42) 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 $\widehat{f}(\omega) = c\frac{(-1)^r}{i}\frac{r}{\omega^2 - r^2}\sin(\omega\pi)$. Note that because $r \in \mathbb{Z}$ we have $\|\widehat{f}\| \asymp \|\widehat{f_\beta}\| \asymp 1$.

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

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

We decompose the integral into three regimes:

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

$$\int_0^{r/2}(\dots) \asymp r^{-2\beta}\int_0^{r/2}\omega^{2t}\sin^{2\beta}(\omega\pi) \lesssim r^{2t}$$

   by Lemma 32.

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. $\underline{\omega \in [r/2, 3r/2]}$: 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

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

   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 $\|f_\beta\|_1 \ge \|\widehat{f_\beta}\|_\infty \asymp 1$ 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 $\widehat{\operatorname{sign}(x)} = \frac{2}{i\omega}$ and $\overline{d_H}(f_\beta, 0) = \sup_b \frac{1}{2}\int f_\beta(x)\operatorname{sign}(x-b)\mathrm{d}x$, 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 $O(1/r)$ follows analogously to the proof of our bound on $\|f_\beta\|_{t,2}$ so we omit it. This concludes our proof.

■

Proof [Proof of Proposition 12 and Proposition 13 for $d = 1$] We now turn to showing tightness in one dimension, utilizing the density difference constructed in Lemma 36. Given a value of the smoothness $\beta > 0$, set $\overline{\beta} = \lceil\beta\rceil + 1$ and let $f_{\overline{\beta}}$ be as in Lemma 36 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 36, 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 Proposition 12 and Proposition 13 for $d = 1$. ■

E.2 The Case $d > 1$

We move on to the case of general dimension. In Section E.2.1 we outline our approach. Then, in Section E.2.2 we give full details of our construction, following the argument outlined in the prior section.

E.2.1 OVERVIEW

For the discussions below, we will assume that the ambient dimension $d \geq 2$. Our construction here is less straightforward than for $d = 1$ in Section E.1 but shares the same basic idea. Recall that the basic premise is that we want to saturate the Hölder's inequality in eq. (22), which requires the density difference $f = \mu - \nu$ to have Fourier transform be (almost) supported on a sphere. For $d = 1$ we took f to be a pure sinusoid. However, of course such f is not compactly supported and that is why we multiplied the sinusoid by a rectangle (and then convolved many times to gain smoothness), which served as a mollifier.

For $d > 1$ let us attempt to follow the same strategy and take

$$f_r(x) = g_r(x)h(x)\,,$$

where $r > 0$ is a parameter, h is some compactly supported smooth mollifier and $g_r(x)$ is defined implicitly via

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

where here and below we denote, a bit informally, by $\delta(\|\cdot\| - r)$ a distribution that integrates any smooth compactly supported function ϕ as follows:

$$\int_{\mathbb{R}^d} \phi(\omega)\delta(\|\omega\| - r)\mathrm{d}\omega \triangleq r^{d-1} \int_{\mathbb{R}^d} \phi(r\omega)\mathrm{d}\sigma(\omega) = \frac{2\pi^{d/2} r^{d-1}}{\Gamma(\frac{d}{2})} \mathbb{E}[\phi(rX)]\,,$$

where σ is the unnormalized surface measure of $\mathbb{S}^{d-1}$ and X is a random vector uniformly distributed on $\mathbb{S}^{d-1}$. Explicit computation shows

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

where J_ν denotes Bessel functions of the first kind of order ν. Notice that g is spherically symmetric and real-valued (some further properties of it are collected below in Lemma 34).

Note that $|g_r(x)| = O(1)$ as $r \to \infty$ for any fixed $x \neq 0$ (Lemma 34), while at the origin we have $|g_r(0)| = \Omega(r^{(d-1)/2})$, which follows from the series expansion of the Bessel function given in for example (Watson, 1980, Section 3.1-3.11). This causes an issue for $d > 1$, as g_r is too large at the origin as $r \to \infty$ compared to its tails, which makes it difficult to use it as the difference between two probability densities. Hence, we choose our mollifier h to be supported on an annulus instead of on a ball. In addition, it will also be convenient for it to have a super-polynomially decaying Fourier transform, i.e.

$$|\widehat{h}(w)| \le H(\|w\|) \triangleq C \exp\left(-\frac{c\|w\|}{\log(\|w\| + 2)^2}\right) \qquad \forall w \in \mathbb{R}^d\,.$$

The existence of the desired function h is proven Lemma 37.

Note that all of the Fourier energy of g_r lies at frequencies $\|\omega\| = r$ by construction. However, after multiplying by h the energy spills over to adjacent frequencies as well and we need to estimate the amount of the spill. Due to the fast decay of $\widehat{h}$ we will show, roughly, the following estimates on the behavior of $\widehat{f_r}$:

$$\begin{aligned} &|\widehat{f_r}(\omega)| \lesssim \tilde{O}(r^{(1-d)/2}) 1\{\|\omega - r\| \le \log^2(r)\} + r^{(d-1)/2} H(\max(\|\omega\| - r, \log^2 r)), \quad \text{and} \\ &|\widehat{f_r}(\omega)| \lesssim r^{(d-1)/2}\|\omega\| \end{aligned}$$

as $r \to \infty$. Note that the first bound above is super-polynomially decaying in both $\|\omega\|$ and r, which allows us to show that

$$\|f_r\|_{t,2} \le \tilde{O}(r^t)$$

for $t > -\frac{d+2}{2}$, recalling the notation $\|f\|_{t,2} = \||\cdot|^t \widehat{f}\|_2$. A direct calculation will also show

$$\|f_r\|_1 \asymp \|f_r\|_\infty \asymp \|f_r\|_2 \asymp 1\,.$$

For a desired total-variation separation ϵ, we will set $\mu - \nu = \epsilon f_r$ and choose $r = \epsilon^{-1/\beta}$ to ensure that $\epsilon\|f_r\|_{\beta,2} = \tilde{O}(1)$. For the energy distance between μ and ν these choices yield

$$\mathcal{E}_\gamma(\mu, \nu) \asymp \|\epsilon f_{\epsilon^{-1/\beta}}\|_{-\frac{d+\gamma}{2},2} = \tilde{O}(\epsilon^{1+\frac{d+\gamma}{2\beta}}) = \tilde{O}(\mathsf{TV}^{\frac{d+2\beta+\gamma}{2\beta}})\,,$$

as required.

We now proceed to rigorous details.

E.2.2 THE CONSTRUCTION

First, we must construct the mollifier h with the properties outlined in Section E.2.1. Recall that a function f is radial (also known as spherically symmetric) if its value at $x \in \mathbb{R}^d$ depends only on $\|x\|$. In other words, $f(x) = f(y)$ holds for all $x, y \in \mathbb{R}^d$ with $\|x\| = \|y\|$.

Lemma 37. *There exists a compactly supported radial Schwartz function h, and a positive sequence $\{r_n\}_{n=1}^{\infty}$ satisfying $r_n = \Theta(n)$, such that*

$$\operatorname{supp}(h) \subset \mathbb{B}(0,1), \tag{43}$$

$$\operatorname{supp}(h) \subset \mathbb{R}^d \setminus \mathbb{B}(0, r_0), \tag{44}$$

$$|\widehat{h}(w)| \le C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \quad \textit{for all } w \in \mathbb{R}^d, \textit{ and} \tag{45}$$

$$\widehat{h}(r_n u) = 0 \quad \textit{for all } u \in \mathbb{S}^{d-1}, \tag{46}$$

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

Proof First, let h_0 be as constructed in Lemma 35, which already satisfies eq. (43) and Equation (45). 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 (44) and Equation (46), respectively, and are defined as

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

Before proceeding, note that clearly h is a radial Schwartz function. Let us now 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 (43) and (44). 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 Item 3 of Lemma 34 we know that

$$\mathcal{F}[\mathbb{1}\{\|\cdot\| < 1\}](w) = \left(\frac{2\pi}{\|w\|}\right)^{\frac{d}{2}} J_{\frac{d}{2}}(\|w\|).$$

Hence, by Item 2 of Lemma 34, the function $\widehat{\rho}_0(w) = (1/8)^d \mathcal{F}[\mathbb{1}\{\|\cdot\| < 1\}](w/8)$ has infinitely many zeros near the values of $\|w\| = 8(2n\pi + \frac{(d+1)\pi}{4})$ for sufficiently large $n \in \mathbb{Z}^+$, which implies Equation (46).

Finally, for Equation (45), 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

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

concluding the proof. ■

Let h be as constructed in Lemma 37, and define

$$f_r = g_r h \tag{47}$$

for $r > 0$ and $\widehat{g}(\omega) \triangleq r^{(1-d)/2}\delta(\|\omega\| - r)$. Recall from the overview of our construction that we gave in Section E.2.1 that f_r is our proposed density difference which we claim (approximately) saturates Hölder's inequality in (22). The next Lemma records the properties of f_r which will enable us to complete our proof.

Lemma 38. *Let f_r be as in* (47) *and let $\{r_n\}_{n=1}^{\infty}$ be the sequence constructed in Lemma 37. The following hold.*

(i) For all $n \in \mathbb{N}$ we have

$$\int_{\mathbb{R}^d} f_{r_n}(x)\mathrm{d}x = 0 \qquad \textit{and} \qquad \mathrm{supp}(f_{r_n}) \subset \mathbb{B}(0,1).$$

(ii) We have

$$\|f_{r_n}\|_\infty \asymp \|f_{r_n}\|_2 \asymp \|f_{r_n}\|_1 \asymp 1,$$

hiding constants independent of n.

(iii) For any $t > -\frac{d+2}{2}$ we have

$$\|f_{r_n}\|_{t,2} = O(r_n^t \log^d(r_n))$$

as $n \to \infty$, hiding constants independent of n.

(iv) Recall the definition of ψ_γ from (17)*. For any $\gamma \in (0,2)$ we have*

$$\sup_{v \in \mathbb{S}^{d-1}, b \in \mathbb{R}} \left| \int_{\mathbb{R}^d} \psi_\gamma(\langle x, v\rangle - b) f_{r_n}(x)\mathrm{d}x \right| = O(r_n^{-(d+\gamma)/2} \log(r_n)^d)$$

hiding constants independent of n.

Proof Let us drop the dependence of r_n to simplify notation.
Showing (i). Note that $\int_{\mathbb{R}^d} f_r(x)\mathrm{d}x = \widehat{f_r}(0)$. Then, $\widehat{f_r}(0) = 0$ follows from the construction of h and g_r. Indeed, $\widehat{g}_r$ is supported on $r\mathbb{S}^{d-1}$ while $\widehat{h}|_{r\mathbb{S}^{d-1}} \equiv 0$. The fact that $\mathrm{supp}(f_r) \subset \mathbb{B}(0,1)$ follows from $\mathrm{supp}(h) \subset \mathbb{B}(0,1)$.

Showing (ii). Since f_r has compact support, we immediately have

$$\|f_r\|_1 \lesssim \|f_r\|_2 \lesssim \|f_r\|_\infty.$$

As h is continuous and supported on the annulus $\{x : r_0 \le \|x\| \le 1\}$ by construction, it suffices to bound g_r on said annulus. Now, for any x with $r_0 \le \|x\| \le 1$, we have by Lemma 34 that

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

which shows that $\|f_r\|_\infty \lesssim 1$.

We now turn to lower bounding $\|f_r\|_1$. Recall that h is uniformly continuous and nontrivial, hence $\int |h(u^* v)| \mathrm{d}\sigma(v) \neq 0$ for some radius u^*, and thus for all $u \in (u_0, u_1) \subseteq (0,1)$ for some constants u_0, u_1. Using that g_r is spherically symmetric, we compute

$$\begin{aligned}
\|f\|_1 &= \int_{\mathbb{R}^d} |g_r(x)||h(x)| \mathrm{d}x \\
&= \int_0^\infty u^{d-1} g_r(u, 0, \ldots, 0) \int h(uv) \mathrm{d}\sigma(v) \mathrm{d}u \\
&\gtrsim \sqrt{r} \int_{u_0}^{u_1} \left| J_{d/2-1}(ru) \right| \mathrm{d}u \gtrsim 1,
\end{aligned}$$

where the last line follows by (35) once again.

Showing (iii). Let $0 < s < r$, whose precise value will be set later. For convenience, set $B_s = \{x \in \mathbb{R}^d : \|x\| \leq s\}$ and $B_s^c = \mathbb{R}^d \setminus B_s$. Recall that by definition

$$\begin{aligned}
\widehat{f_r}(\omega) &= r^{(1-d)/2} \int_{\mathbb{R}^d} \widehat{h}(\omega + x) \delta(\|x\| - r) \mathrm{d}x \\
&= \underbrace{r^{(1-d)/2} \int_{\mathbb{R}^d} (\widehat{h} \mathbb{1}_{B_s})(\omega + x) \delta(\|x\| - r) \mathrm{d}x}_{I} + \underbrace{r^{(1-d)/2} \int_{\mathbb{R}^d} (\widehat{h} \mathbb{1}_{B_s^c})(\omega + x) \delta(\|x\| - r) \mathrm{d}x}_{II}.
\end{aligned} \tag{48}$$

Let C, c be as in Lemma 37, and $H(x) = C \exp(-c\|x\| / \log^2(\|x\| + 2))$. Note that $\|\widehat{h}\|_\infty \leq C$. Therefore, the first term in the decomposition (48) can be bounded by

$$\begin{aligned}
|I| &\leq C r^{(1-d)/2} \int_{\mathbb{R}^d} \mathbb{1}\{\|\omega + x\| \leq s\} \delta(\|x\| - r) \mathrm{d}x \\
&= C r^{(1-d)/2} \mathbb{1}\{\|\omega\| \in [r - s, r + s]\} \int_{\mathbb{R}^d} \mathbb{1}\{\|\omega + x\| \leq s\} \delta(\|x\| - r) \mathrm{d}x \\
&\lesssim r^{(1-d)/2} s^{d-1} \mathbb{1}\{\|\omega\| \in [r - s, r + s]\}.
\end{aligned}$$

The second line uses that if $\|\omega\| \notin [r - s, r + s]$ then the integral becomes zero. The third line uses the fact that the surface area of the intersection of B_s with any sphere of any radius (and the one centered at ω with radius r in particular) is at most $O(s^{d-1})$.

Moving on to the second term, we have

$$|II| = r^{(d-1)/2} \int |(\widehat{h} \mathbb{1}_{B_s^c})(\omega + ru)| \mathrm{d}\sigma(u) \lesssim r^{(d-1)/2} H(\max\{\|\omega\| - r, s\})$$

using that $H : [0, \infty) \to (0, C]$ is decreasing and that $|\widehat{h}(y) \mathbb{1}_{B_s^c}(y)\| \leq H(\max\{y, s\})$ for all $y \in \mathbb{R}^d$. Summarizing, we have the pointwise estimate

$$|\widehat{f_r}(\omega)| \lesssim r^{(1-d)/2} s^{d-1} \mathbb{1}\{\|\omega\| \in [r - s, r + s]\} + r^{(d-1)/2} H(\max\{\|\omega\| - r, s\}) \tag{49}$$

for all $\omega \in \mathbb{R}^d$ and $0 < s < r$.

We now show that f_r is Lipschitz continuous. Recall from the construction of h (Lemma 37) that $h|_{r_n\mathbb{S}^{d-1}} \equiv 0$. Then, we observe that for any $\omega \in \mathbb{R}^d$

$$\begin{aligned}
|\widehat{f}_r(\omega)| &= r^{(d-1)/2}\left|\int \widehat{h}(\omega + ru)\mathrm{d}\sigma(u)\right| \\
&= r^{(d-1)/2}\left|\int \{\widehat{h}(\omega + ru) - \widehat{h}(ru)\}\mathrm{d}\sigma(u)\right| \\
&= r^{(d-1)/2}\|\widehat{h}\|_{\mathrm{Lip}}\frac{2\pi^{d/2}\|\omega\|}{\Gamma(\frac{d}{2})} \\
&\lesssim r^{(d-1)/2}\|\omega\|, \qquad (50)
\end{aligned}$$

where we use that $\widehat{h}$ is Schwartz by construction, and thus has finite Lipschitz constant $\|\widehat{h}\|_{\mathrm{Lip}}$.

With (49) and (50) in hand we can proceed to bounding the norm of f_r. Let $s = D\log(r)^2$ for a large constant D independent of r, and assume that r is large enough so that $s < r/2$. Also set $\theta > 0$, whose precise value is specified later. We have

$$\begin{aligned}
\|f_r\|_{t,2}^2 &= \int_{\mathbb{R}^d}\|\omega\|^{2t}|\widehat{f}_r(\omega)|^2\mathrm{d}\omega \\
&\overset{(50)}{\lesssim} r^{d-1}\int_{\|\omega\|\le r^{-\theta}}\|\omega\|^{2t+2}\mathrm{d}\omega + \int_{\|\omega\|>r^{-\theta}}\|\omega\|^{2t}|\widehat{f}_r(\omega)|^2\mathrm{d}\omega \\
&\overset{(49)}{\lesssim} r^{d-1-\theta(2t+d)} \\
&\qquad + r^{1-d}\log(r)^{2(d-1)}\int_{\|\omega\|>r^{-\theta}}\|\omega\|^{2t}\mathbb{1}\{\|\omega\| \in [r-s, r+s]\}\mathrm{d}\omega \\
&\qquad + r^{d-1}\int_{\|\omega\|>r^{-\theta}}\|\omega\|^{2t}H^2(\max\{\|\omega\| - r, s\})\mathrm{d}\omega \\
&\lesssim r^{d-1-\theta(2t+d)} + r^{2t}\log(r)^{2d-1} + r^{d-1}\int_{r^{-\theta}}^{\infty} u^{2t+d-1}H^2(\max\{u-r, s\})\mathrm{d}u.
\end{aligned}$$

Note that in the derivation above we changed to polar coordinates freely, and that in the second inequality we used the assumption $t > -d/2 - 1$. Setting θ to any positive value greater than $(d-1-2t)/(2t+d)$ ensures that the first term in the final line is $O(r^{2t})$. As for the integral term, we can bound it by

$$\lesssim r^{d-1}H^2(s)\int_{r^{-\theta}}^{2r} u^{2t+d-1}\mathrm{d}u + r^{d-1}\int_{2r}^{\infty} H^2(u/2)\mathrm{d}u \overset{\text{Lemma 33}}{\lesssim} \mathrm{poly}(r)\times H^2(s) + r^{2t}.$$

By taking D large enough (independently of r) in the definition of $s = D\log^2(r)$ we can make also the first term $\mathrm{poly}(r)\times H^2(s)$ less than $O(r^{2t})$, which concludes the proof of (iii).

Showing (iv). The bounds that we develop below are analogous to those given in the proof of (iii). Fix $b \in \mathbb{R}$ and $v \in \mathbb{S}^{d-1}$ and define

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

Suppose first that $\gamma \neq 1$. Then, using Lemmas 28 and 29, we know by dominated convergence that

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

Similarly, for $\gamma = 1$ we can compute

$$\dagger = C_{\psi_\gamma} \lim_{\epsilon\to 0} \int_\epsilon^{1/\epsilon} \frac{\sin(-tb)\widehat{f}_r(tv)}{t} \mathrm{d}t.$$

In either case, we have $|\dagger| \lesssim \int_0^\infty |\widehat{f}_r(tv)|/t^{(1+\gamma)/2}\mathrm{d}t$.

Let $s = D\log^2(r)$ for large D independent of r as in the proof of (iii), and let $\theta > 0$ whose precise value is specified later. Assuming that r is large enough so that $s < r/2$, for any $\gamma \in (0,2)$ we have

$$\begin{aligned}
|\dagger| &\leq \int_0^{r^{-\theta}} \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t + \int_{r^{-\theta}}^\infty \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{(50)}{\lesssim} r^{(d-1)/2} \int_0^{r^{-\theta}} t^{(1-\gamma)/2}\mathrm{d}t + \int_{r^{-\theta}}^\infty \frac{|\widehat{f}_r(tv)|}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{(49)}{\lesssim} r^{\frac{d-1}{2} - \theta\frac{3-\gamma}{2}} \\
&\qquad + \int_{r^{-\theta}}^\infty \frac{1}{t^{(1+\gamma)/2}} \left(r^{(1-d)/2} s^{d-1} \mathbb{1}\{t \in [r-s, r+s]\} + r^{(d-1)/2} H(\max\{t-r, s\}) \right) \mathrm{d}t \\
&\lesssim r^{\frac{d-1}{2} - \theta\frac{3-\gamma}{2}} + r^{-(d+\gamma)/2}\log^d(r) + H(s) r^{(d-1)/2} \int_{r^{-\theta}}^{2r} \frac{\mathrm{d}t}{t^{(1+\gamma)/2}} + \int_{2r}^\infty \frac{H(t/2)}{t^{(1+\gamma)/2}}\mathrm{d}t \\
&\overset{\text{Lemma 33}}{\lesssim} r^{\frac{d-1}{2} - \theta\frac{3-\gamma}{2}} + r^{-(d+\gamma)/2}\log^d(r) + H(s) \times \mathrm{poly}(r) + r^{-100d},
\end{aligned} \tag{51}$$

Set θ to any value greater than $(2d+\gamma-1)/(3-\gamma)$, which ensures that the first term in (51) is $O(r^{-(d+\gamma)/2})$. By taking D large enough in the definition of $s = D\log^2(r)$, we can make $H(s)$ smaller than any polynomial in r, which ensures that the third term in (51) is also $O(r^{-(d+\gamma)/2})$. We thus obtain the final bound $|\dagger| \lesssim r^{-(d+\gamma)/2}\log^d(r)$, concluding our proof. ■

Proof [Proof of Theorem 12 and Proposition 13 for $d > 1$] Using the functions $\{f_{r_n}\}_{n=1}^\infty$ we constructed in Lemma 38, we are ready to prove Propositions 12 and 13 for $d > 1$.

Let p_0 be a compactly supported probability density with $\inf_{\|x\|\leq 1} p_0(x) > 0$. Fix the smoothness $\beta > 0$. Given any desired total variation separation $\epsilon \in (0,1)$, we can find $n_0 \in \mathbb{N}$ such that $\epsilon^{-1/\beta} \asymp r_{n_0}$, where we hide an ϵ-independent multiplicative constant. Define

$$p_\epsilon = p_0 + \epsilon f_{r_{n_0}}/2 \qquad \text{and} \qquad q_\epsilon = p_0 - \epsilon f_{r_{n_0}}/2.$$

Clearly p_ϵ and q_ϵ are compactly supported probability densities for all small enough ϵ. Moreover, by Lemma 38 they satisfy

$$\begin{aligned} \|p_\epsilon - q_\epsilon\|_1 &\asymp \epsilon \qquad \text{and} \qquad \|p_\epsilon\|_{\beta,2} \asymp \|q_\epsilon\|_{\beta,2} \asymp 1 \qquad \text{and} \\ \mathcal{E}_\gamma(p_\epsilon, q_\epsilon) &\lesssim \epsilon^{\frac{2\beta+d+\gamma}{2\beta}} \log(1/\epsilon)^d \qquad \text{and} \qquad \overline{d_H}(p_\epsilon, q_\epsilon) \lesssim \epsilon^{\frac{2\beta+d+1}{2\beta}} \log(1/\epsilon)^d \end{aligned}$$

for all fixed $\gamma \in (0, 2)$. This concludes our proof. ■

Appendix F. Proof of Proposition 24

Proof Note that $\overline{d_H} = T_{d,0}$. Let p_ϵ, q_ϵ be the compactly supported densities constructed in the proof of Proposition 12 in the general dimensional case. Then by construction

$$\varepsilon \asymp \mathsf{TV}(p_\epsilon, q_\epsilon) \asymp \|p_\epsilon - q_\epsilon\|_2 \quad \text{and} \quad \|p_\epsilon\|_{\beta,2} + \|q_\epsilon\|_{\beta,2} \lesssim 1 \quad \text{and} \quad \overline{d_H}(p_\epsilon, q_\epsilon) \lesssim \epsilon^{\frac{2\beta+d+1}{\beta}} \log(1/\epsilon)^d.$$

Write $p_{\epsilon,n}$ and $q_{\epsilon,n}$ for the empirical measures of p_ϵ and q_ϵ respectively, based on n i.i.d. observations each. By the triangle inequality we have

$$\begin{aligned} \mathbb{E}\overline{d_H}(p_{\epsilon,n}, q_{\epsilon,n}) &\le \mathbb{E}\overline{d_H}(p_{\epsilon,n}, p_\epsilon) + \overline{d_H}(p_\epsilon, q_\epsilon) + \mathbb{E}\overline{d_H}(q_\epsilon, q_{\epsilon,n}) \\ &\overset{\text{Lemma 20}}{\lesssim} 1/\sqrt{n} + \epsilon^{\frac{2\beta+d+1}{2\beta}} \log(1/\epsilon)^d. \end{aligned}$$

This completes the proof. ■

References

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.

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

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

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

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

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.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of convergence for density estimation with generative adversarial networks. *arXiv preprint arXiv:2102.00199*, 2021.

Denis Belomestny, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161:242–253, 2023.

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.

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

Minwoo Chae. Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *arXiv preprint arXiv:2202.02890*, 2022.

Minwoo Chae. Wasserstein upper bounds of lp-norms for multivariate densities in besov spaces. *Statistics & Probability Letters*, 210:110131, 2024.

Minwoo Chae and Stephen G. Walker. Wasserstein upper bounds of the total variation for smooth densities. *Statistics & Probability Letters*, 163:108771, 2020. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2020.108771. URL https://www.sciencedirect.com/science/article/pii/S0167715220300742.

Minwoo Chae, Dongha Kim, Yongdai Kim, and Lizhen Lin. A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. *Journal of machine learning research*, 24(77):1–42, 2023.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

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.

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

Vincent Divol. Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183(1):581–647, 2022.

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.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

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. doi: 10.1137/1131042.

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

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

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

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

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

Szymon Knop, Jacek Tabor, Przemyslaw Spurek, Igor Podolak, Marcin Mazur, and Stanislaw Jastrzebski. Cramer-wold autoencoder. 2018. doi: 10.48550/ARXIV.1805.09235.

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. doi: 10.1109/ICASSP43922.2022.9746016.

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

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

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

Subhash Chandra Malik and Savita Arora. *Mathematical analysis*. New Age International, 1992.

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.

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

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.

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.

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.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. *arXiv preprint arXiv:2303.01861*, 2023.

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.

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

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

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.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 06 2001. ISBN 9780262256933. doi: 10.7551/mitpress/4175.001.0001.

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.

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.

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.

Elias M Stein and Guido Weiss. *Introduction to Fourier analysis on Euclidean spaces*, volume 1. Princeton university press, 1971.

Arthur Stéphanovitch, Eddie Aamari, and Clément Levrard. Wasserstein gans are minimax optimal distribution estimators. *arXiv preprint arXiv:2311.18613*, 2023.

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.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282–1308, 2023.

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. URL https://proceedings.mlr.press/v195/tiegel23a.html.

Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. *Advances in neural information processing systems*, 32, 2019.

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

C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003. ISBN 9780821833124. URL https://books.google.com/books?id=idyFAwAAQBAJ.

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

George N Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge University Press, 1980. ISBN 052106743X.

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

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

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.