Journal of Machine Learning Research 24 (2023) 1-36 Submitted 9/23; Published 9/23

Density Estimation via Random Halfspaces

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

Tianze Jiang TJIANG@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA

Yury Polyanskiy YP@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
32 Vassar St, Cambridge, MA 02142, USA

Rui Sun ERUISUN@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139, USA

Editor: NA

Abstract

Consider two probability densities p, q, separated by ϵ in total variation (TV) and supported on a compact subset of $\mathbb{R}^d$. How large can the difference in mass that they place on a halfspace be? We show that if p, q are β-times differentiable, then there exists a halfspace achieving separation at least $\epsilon^{\frac{2\beta+d+1}{2\beta}}$, and show that this cannot be improved in general. Moreover, the bound is achieved by a *random* half-space, which shows how random, affine test functions can effectively extract information from high dimensional distributions.

We show that average halfspace separation defines a metric, previously considered in the MMD literature under the name of energy distance. Thus, our result implies that a density estimator which even approximately (to within $O(1/\sqrt{n})$) minimizes a generalized energy distance to the empirical distribution of n observations achieves the minimax optimal density estimation rate over the class of compactly supported smooth densities (in TV) and Gaussian mixtures (in L^2), up to $\mathrm{poly}(\log)$ and $\log\log$ factors, resp. This distance enjoys dimension-free concentration, can be estimated fast from samples, and is differentiable. To the best of our knowledge, this is the only metric that possesses both minimax guarantees and is well suited for fitting neural-net based generative models.

Keywords: Density estimation, Linear thresholds, Gaussian mixtures, Sobolev norm

Contents

1 Introduction

Many successful methods in machine learning rely on restricting an intractable optimization in a variational formulation into a simpler function class. Examples are numerous and include ELBO in variational inference (Blei et al., 2017; Zhang et al., 2018), variational autoencoders (Kingma et al., 2019), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017) and diffusion models (Song et al., 2020; Chen et al., 2022).

At the time of their invention, GANs revolutionized the practice of approximately sampling from high dimensional distributions specified through training examples. The breakthrough idea was the use of a double (min-max) optimization where a generator is pitted against a discriminator. Mathematically, we can model this as follows: given i.i.d. data $X_1, \dots, X_n \in \mathbb{R}^d$ from an unknown distribution ν, we choose an estimator $\tilde{\nu}$ that satisfies

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

where $\mathcal{G}$ is a class of distributions representing the generator, and $\mathcal{D}$ is a class of discriminating functions. For future reference, denote $\nu_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ as the empirical measure. In the original formulation one also applied a logarithmic transformation to the output of the discriminator to approximate the Jensen-Shannon divergence; later works extended this to many other distance measures such as the Wasserstein-GAN (Arjovsky et al., 2017) which corresponds to (1).

However, it is unclear how the quality of $\tilde{\nu}$, as an estimate of ν, varies as a function of $\mathcal{G}$ or $\mathcal{D}$. In particular, even if $\tilde{\nu}$ does a good job of fooling the best $f \in \mathcal{D}$, could this be solely because f is too weak? Can $\tilde{\nu}$ be guaranteed to be close to ν in traditional metrics such as total variation even for a simple discriminator class $\mathcal{D}$? In this paper we address these questions in the simplest possible setting, namely, when $\mathcal{D}$ is the class of affine classifiers and $\mathcal{G}$ consists of smooth distributions or Gaussian mixtures. Accordingly, let

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

be the class of affine classifiers. Let $\mathcal{P}_S(\beta, d)$ denote the set of distributions supported on the d-dimensional unit ball $\mathbb{B}(0,1)$ that have a density whose $(\beta, 2)$-Sobolev norm (defined in (4)) is bounded by 1, and let $\mathcal{P}_G(d) = \{\mu * \mathcal{N}(0,1) : \mathrm{supp}(\mu) \subseteq \mathbb{B}(0,1)\}$ be the class of Gaussian mixtures with compactly supported mixing distribution. Finally, write $\mathsf{TV}(\mu, \nu) = \sup_E |\mu(E) - \nu(E)|$ for the total variation distance.

Theorem 1. *Let $\tilde{\nu}, \mathcal{G}, \mathcal{D}$ be as in (1) and assume $\nu \in \mathcal{G}$ i.e. the problem is well-specified.*

1. *If $\mathcal{G} = \mathcal{P}_S(\beta, d)$ and $\mathcal{D} = \mathcal{D}_a$, then there exists a finite C depending on β, d only such that*

$$\mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C n^{-\frac{\beta}{2\beta + d + 1}}. \tag{2}$$

2. *If $\nu \in \mathcal{G} = \mathcal{P}_G(d)$ and $\mathcal{D} = \mathcal{D}_a$, then there exists a finite C depending on d only such that*

$$\mathbb{E}\mathsf{TV}(\tilde{\nu}, \nu) \leq C \frac{(\log(n))^{\frac{2d+2}{4}}}{\sqrt{n}}. \tag{3}$$

Notice that the rate of decay in (2) is *almost* the minimax optimal $n^{-\beta/(2\beta+d)}$; the only difference is that the dimension is effectively inflated by one. Similarly, though to the best of our knowledge, no minimax optimal statement exists in terms of total variation for Gaussian mixtures, in proving (3) we obtain a near-optimal rate in L^2, losing only by a multiplicative factor of $\log(n)^{1/2}$ (Kim and Guntuboyina, 2022). One prevalent explanation of the success of GANs, and other methods that rely on restricting an optimization to a simpler class, is that the discriminators/decoders/score functions that perform well on real-world data are well-approximated by modern neural architectures. In other words, the generator/target class $\mathcal{G} = \{\text{natural distributions}\}$ and the discriminator class $\mathcal{D} = \{\text{ML architectures}\}$ are a perfect pairing. However, Theorem 1 shows that this is not necessarily the case: even when $\mathcal{G}$ has no "human structure" like arbitrary smooth densities on $\mathbb{B}(0,1)$ (which can exhibit wild behavior), and when $\mathcal{D} = \mathcal{D}_a$ is taken to be the most basic, linear architecture, the program still works and is almost optimal in a minimax sense!

The proof of Theorem 1 relies on a comparison inequality between total variation and the maximum halfspace distance

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

The problem becomes easier by relaxing the supremum in the definition of $\overline{d_H}$ to an (unnormalized) average: we mainly work with the distance

$$d_H(\mu,\nu) \triangleq \sqrt{\mathbb{E}_{f\sim\mathrm{Unif}(\mathcal{D}_a)}(\mathbb{E}_\mu f - \mathbb{E}_\nu f)^2}.$$

In particular, Theorem 1 holds also if we replace $\overline{d_H}$ by d_H in (1). Here we abuse notation to say that f is 'drawn' from $\mathrm{Unif}(\mathcal{D}_a)$ if $f(x) = \mathbb{1}\{\langle x, v\rangle \geq b\}$ for a uniformly random unit vector $v \in \mathbb{S}^{d-1}$ and b 'drawn' from the Lebesgue measure on $[-1,1]$. As we will see in later sections, d_H is in fact equal to the *energy distance* (Székely and Rizzo, 2013). For $\overline{d_H}$ and d_H we derive the following separation results (note that $\overline{d_H} \leq \mathsf{TV}$ is trivial since f is a bounded function).

Theorem 2. *Let $\mu,\nu \in \mathcal{G}$ and let C be a small constant depending only on $\mathcal{G}$.*

1. *If $\mathcal{G} = \mathcal{P}_S(\beta,d)$ then one can take C small enough so that*

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

2. *If $\mathcal{G} = \mathcal{P}_G(d)$ then one can take C small enough so that*

$$C\,\mathsf{TV}(\mu,\nu)\log\left(3+\frac{1}{\mathsf{TV}(\mu,\nu)}\right)^{-\frac{2d+1}{4}} \leq d_H(\mu,\nu) \leq \frac{1}{C}\,\overline{d_H}(\mu,\nu)\log\left(3+\frac{1}{\overline{d_H}(\mu,\nu)}\right)^{\frac{1}{4}}.$$

Our main tool for proving Theorem 2 is Fourier analysis and comparison inequalities between total variation and L^2 distance. Before moving on, let us interpret this result. It says that for *any* pair of smooth/Gaussian mixture distributions, there is a halfspace that captures a polynomial fraction of the total variation. Moreover, we can even replace the best halfspace with a *random* halfspace ($\overline{d_H} \to d_H$), and the separation result still holds!

At first look Theorem 2 might be concerning: indeed, if $\overline{d_H} \ll \mathsf{TV}$, how can one expect the minimum-$\overline{d_H}$ density estimator $\tilde{\nu}$ to perform well, as we claim in Theorem 1? The important

observation is that $\overline{d_H}$ between empirical and population measures decays at the parametric rate, i.e. $\sup_{\nu\in\mathcal{G}} \mathbb{E}\overline{d_H}(\nu,\nu_n) \lesssim 1/\sqrt{n}$, since $\mathcal{D} = \mathcal{D}_a$ has finite VC dimension. So, in a sense, by using $\overline{d_H}$ we have traded off the expressivity of the discriminator class for a better concentration of the induced distance. Since overall performance has not degraded significantly (as per Theorem 1), this trade-off is favorable in practical applications where a tractable optimization problem is paramount.

1.1 Contributions

To summarize, our main contributions are as follows. We show that β-smooth distributions and Gaussian mixtures that are far apart in total variation distance possess a large number of halfspaces on which their mass is different (Theorem 2).

We show that the defined average halfspace separation distance d_H has many equivalent expressions: as a weighted L^2-distance between characteristic functions, as the sliced Cramer-2 distance, as an IPM/MMD/energy distance, and as the L^2-norm of the Riesz potential (Proposition 3, Proposition 4, Proposition 5, Section 2.4).

Given an empirical target measure $\mathbb{P}_n$, we show that an approximate ERM density estimator which finds any candidate distribution $\hat{\mathbb{P}}$ such that $d_H(\hat{\mathbb{P}},\mathbb{P}_n) \lesssim \frac{1}{\sqrt{n}}$ attains the minimax optimal TV-estimation rate at dimension $d+1$ (as opposed to d) (Theorem 1).

Generalizing the halfspace distance d_H to include an exponent d_α as (see (6)), we discover that if instead of thresholded linear features $\mathbb{1}\{\langle v,x\rangle > b\}$ we use the non-linearity $|\langle v,x\rangle - b|^\alpha$, $\alpha \in (-\frac{1}{2},\frac{1}{2})$, the distributions can be separated even better (Theorem 7).

Finally, combined with the fact that d_α, similar to d_H, concentrates via samples at the parametric rate (Lemma 6), the ERM for d_α reduces the slack of minimax optimal TV-estimation rate from $d+1$ to $d+2\alpha+1$, thereby almost recovering the optimal rate at dimension d. This result, combined with its strong approximation properties, supports its use in modern generative models (Ho et al. (2020); Goodfellow et al. (2020); Rombach et al. (2022); Ramesh et al.)

1.2 Related work

In the statistics literature, an estimator of the form (1) appears in the famous work (Yatracos, 1985). Instead of indicators of halfspaces, they consider the class of discriminators $\mathcal{Y} \triangleq \{\mathbb{1}_{\mathrm{d}\nu_i/\mathrm{d}\nu_j \geq 1} : 1 \leq i,j \leq N(\epsilon_n,\mathcal{G})\}$, where $\nu_1,\ldots,\nu_{N(\epsilon_n,\mathcal{G})}$ forms a minimal ϵ_n-TV covering of the class $\mathcal{G}$ and $N(\epsilon_n,\mathcal{G})$ is the so-called covering number. Writing $d_Y(\mu,\mu') = \sup_{f\in\mathcal{Y}}(\mathbb{E}_\mu f - \mathbb{E}_{\mu'} f)$, it is not hard to prove that $|\mathsf{TV} - d_Y| = \mathcal{O}(\epsilon_n)$ on $\mathcal{G}\times\mathcal{G}$ and that $\mathbb{E}d_Y(\nu,\nu_n) \lesssim \sqrt{\log N(\epsilon_n,\mathcal{G})/n}$ by a union bound coupled with a binomial tail inequality. From here $\mathbb{E}\mathsf{TV}(\tilde{\nu},\nu) \lesssim \min_{\epsilon>0}[\sqrt{\log N(\epsilon,\mathcal{G})/n} + \epsilon]$ follows by the triangle inequality (here $\tilde{\nu}$ is defined as in (1) with $\mathcal{D} = \mathcal{Y}$). Note that in contrast to our maximum halfspace distance $\overline{d_H}$, Yatracos' estimator attains the optimal rate on $\mathcal{G} = \mathcal{P}_S$, corresponding to the choice $\epsilon_n \asymp n^{-\beta/(2\beta+d)}$.

Yatracos' estimator lies at the other end of the discriminator expressiveness-concentration trade-off discussed in the introduction: the distance d_Y is as expressive as total variation when restricted to $\mathcal{G}$, 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 $\mathcal{D}_a$ is oblivious to $\mathcal{G}$ and (ii) the distance d_Y is impractical to compute as it requires a covering of $\mathcal{G}$.

From the above discussion, an interesting question emerges. Namely, 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 $\tilde{\nu}$ of (1) 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? Even more concretely, can partitioning the space using multiple halfspaces, similarly to how ReLU-networks do, help achieve optimality in Theorem 1? The question whether the analysis of Theorem 1 can be improved to attain optimality of $\overline{d_H}$ over the two generator classes also remains open.

Several other related works such as Singh et al. (2018); Liang (2021) study the problem of minimax density estimation over classical smoothness classes with respect to Integral Probability Metrics (IPMs) $d_{\mathcal{D}}(\mathbb{P},\mathbb{Q}) \triangleq \sup_{f\in\mathcal{D}} |\mathbb{E}_{\mathbb{P}} f - \mathbb{E}_{\mathbb{Q}} f|$. In particular, these works seek estimators $\tilde{\nu}$ such that $d_{\mathcal{D}}(\tilde{\nu},\nu)$ is small for some discriminator $\mathcal{D}$. Note some crucial differences to our work: first, we evaluate performance with respect to total variation in Theorem 1 which bears more interest both theoretically and empirically; second, we restrict our attention to estimators $\tilde{\nu}$ attained by ERM which is more commonly used in practice, while corresponding results of (Singh et al., 2018; Liang, 2021) consider classical orthogonal projection estimators.

Independent from this work, recent results by (Paik et al., 2023) investigate the hyperplane separability of distributions under the setting of two-sample testing. However, their focus was on the asymptotic power of separation given samples. In addition, our lower bounds construction of two smooth distributions with a small half-space distance (Appendix E.2) suggests that (at least heuristically) their test is sub-optimal against the known mini-max sample complexity for two-sample testing on smooth distributions with L_2 separation (Arias-Castro et al., 2018).

1.3 Notation

The convolution between two functions or measures is denoted by $*$. Symbols O and Ω follow the convention of the "big-O" notation. We use $\lesssim$ and $\gtrsim$ to hide irrelevant universal multiplicative constants. Given a vector $x \in \mathbb{R}^d$ we write $\|x\|$ for its Euclidean norm. 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 $\mathrm{d}\sigma$ for the unnormalized surface measure on $\mathbb{S}^{d-1}$. We write $\mathcal{P}(\mathbb{R}^d)$ for the space of all probability distributions on $\mathbb{R}^d$. 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. We write $L^p(\mathbb{R}^d)$ for the space of (equivalence classes of) functions $\mathbb{R}^d \to \mathbb{C}$ that satisfy $\|f\|_p \triangleq \left(\int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}x\right)^{\frac{1}{p}} < \infty$. Given a function $f \in L^1(\mathbb{R}^d)$, define its Fourier transform as

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

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 an isometry (up to the constant $(2\pi)^{-d}$) on $L^2(\mathbb{R}^d)$ in the usual way. Given $f \in L^2(\mathbb{R}^d)$ and $\beta > 0$, define the homogenous Sobolev norm of order $(\beta,2)$ of f as

$$\|f\|_{\beta,2}^2 \triangleq \int_{\mathbb{R}^d} \|\omega\|^{2\beta} |\hat{f}(\omega)|^2 \mathrm{d}\omega. \tag{4}$$

Further, we define $\mathcal{P}_S(\beta,d)$ to be the set of distributions on $\mathbb{R}^d$ that have density p with $\mathrm{supp}(p) \subseteq \mathbb{B}(0,1)$ and $\|p\|_{\beta,2} \le 1$. Furthermore, write

$$\mathcal{P}_G(d) \triangleq \{\nu * \mathcal{N}(0,I_d) : \nu \in \mathcal{P}(\mathbb{R}^d), \mathrm{supp}(\nu) \subseteq \mathbb{B}(0,1)\} \tag{5}$$

for the set of all Gaussian mixtures with support in the unit ball.

1.4 Structure

The structure of the paper is as follows. In Secion 2 we introduce the generalized energy distance, the main object of our study. We show how it relates to the maximum halfspace distance $\overline{d_H}$ 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 proves 2 properties of d_α. In Section 4 we analyse the density estimator that minimizes the empirical energy distance, and we prove Theorem 1 and Theorem 2 in Section 4.2. Detailed proofs and auxiliary results are deferred to the Appendix.

2 The energy distance and its alternative formulations

2.1 The Fourier form

Throughout this section, let μ, ν be two probability distributions on $\mathbb{R}^d$. Recall from Section 1 the definition of $\overline{d_H}$ and d_H. In addition to these, for $|\alpha| < 1$ we define

$$d_\alpha(\mu, \nu) \triangleq \sqrt{\frac{1}{\pi} \int_{\mathbb{R}^d} \frac{|\hat{\mu}(\omega) - \hat{\nu}(\omega)|^2}{\|\omega\|^{d+1+\alpha}} \mathrm{d}\omega}. \tag{6}$$

A priori, it is unclear how d_α is related to our problem as described in Section 1. The connection is cleared up by the following.

- The quantity d_α equivalent to the *generalized energy distance* and is a metric on the space of distributions with finite $(1+\alpha)$'th moment.
- The equality $d_0 \equiv d_H$ holds following the definition in the introduction.

The generalized energy distance is a mathematically rich object, and it has multiple equivalent formulations, each with their own interpretation. To avoid getting lost in the details, above we only note one such equivalence, namely the fact that $d_0 \equiv d_H$. In the remainder of this section we derive some properties of the generalized energy distance d_α, which when specialized to $d_0 = d_H$ will yield (with some additional steps) our claims in Section 1.

Throughout this section, let $C_0 = 1$ and constants $C_\alpha = \left(\frac{\pi}{2\cos(\pi\alpha/2)\Gamma(-\alpha)}\right)^2$ for $\alpha \neq 0$ and let D_α be defined by $D_\alpha = \frac{1}{C_\alpha} \frac{\pi^{d/2-1}\Gamma\left(\frac{1-\alpha}{2}\right)}{(1+\alpha)2^{1+\alpha}\Gamma\left(\frac{d+1+\alpha}{2}\right)}$. Due to Proposition 3 and Proposition 4, we are able to show that our distance d_α has at least two equivalent characterizations, as detailed below.

2.2 The standard form

We defined d_α in Equation (6) as a weighted L^2-distance between the Fourier transforms of the input. We did so because this form is the most convenient for our proofs. However, the usual way in which the *energy distance* is introduced is in terms of expected Euclidean distances. The generalized energy distance $\mathcal{E}_{1+\alpha}(\mu, \nu)$ and our $d_\alpha(\mu, \nu)$ are linked by the following proposition.

Proposition 3 (Székely and Rizzo (2013))**.** *Let $|\alpha| < 1$ and let μ, ν be probability measures on $\mathbb{R}^d$ with finite $(1+\alpha)$'th moment. Then, for $(X, X', Y, Y') \sim \mu^2 \otimes \nu^2$,*

$$d_\alpha^2(\mu, \nu) = D_\alpha \mathbb{E}\left[2\|X - Y\|^{1+\alpha} - \|X - X'\|^{1+\alpha} - \|Y - Y'\|^{1+\alpha}\right] = D_\alpha \mathcal{E}_{1+\alpha}^2(\mu, \nu). \tag{7}$$

Another interpretation of Proposition 3 is through the theory of *Maximum Mean Discrepancy (MMD)*. For $|\alpha| < 1$ we can define the kernel

$$k_\alpha(x, y) = \|x\|^{1+\alpha} + \|y\|^{1+\alpha} - \|x - y\|^{1+\alpha}$$

which corresponds to the covariance of fractional Brownian motion. Then, (7) says precisely that the energy distance is (up to a constant) equal to the MMD with kernel k_α. This also implies immediately that d_α can be written as an *Integral Probability Metric (IPM)*, i.e. in the form $d_\alpha(\mu, \nu) \propto \sup_{f \in \mathcal{H}_\alpha : \|f\|_{\mathcal{H}_\alpha} \le 1} \mathbb{E}[f(X) - f(Y)]$, where $\mathcal{H}_\alpha$ is the Reproducing Kernel Hilbert Space corresponding to the kernel k_α. We refer the reader to Schölkopf and Smola (2001); Muandet et al. (2017) for more details on MMD and RKHS with a general kernel.

2.3 The sliced form

We now present yet another equivalent characterization of the energy distance, this time as a *sliced*-distance. Sliced distances are computed by first choosing a random direction on the unit sphere, and then computing a one-dimensional distance in the random direction between the projections of the two input distributions. Define the function

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

The following result, to the best of our knowledge, has not appeared before outside the case $\alpha = 0$.

Proposition 4. *Let $|\alpha| < 1$ and let μ, ν be probability measures on $\mathbb{R}^d$ with finite $(1+\alpha)$'th moment. Then for $(X, Y) \sim \mu \otimes \nu$:*

$$d_\alpha(\mu, \nu) = \sqrt{C_\alpha \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \Big[\mathbb{E}\psi_\alpha(\langle X, v\rangle - b) - \mathbb{E}\psi_\alpha(\langle Y, v\rangle - b)\Big]^2 \mathrm{d}b \mathrm{d}\sigma(v)}. \tag{8}$$

In particular, $d_0 \equiv d_H$ as claimed in Section 3.

The proof of Proposition 4 hinges on computing the Fourier transform of the function ψ_α, which can be interpreted as a tempered distribution.

We point out a special property of the integral on the right hand side of (8). After expanding the square, one finds that the individual terms in the sum are not absolutely integrable. However, due to cancellations, the quantity is still well defined.

One other observation is that d_α is a *sliced* probability divergence in the language of (Nadjahi et al., 2020). Given $v \in \mathbb{S}^{d-1}$ writing $\theta_v = \langle v, \cdot \rangle$ and $\theta_v \# \mathbb{P} = \mathbb{P} \circ \theta_v$ for the pushforward of $\mathbb{P}$ under θ_v, we have

$$d_\alpha^2(\mathbb{P}, \mathbb{Q}) = \int_{\mathbb{S}^{d-1}} d_\alpha^2(\theta_v \# \mathbb{P}, \theta_v \# \mathbb{Q}) \mathrm{d}\sigma(v).$$

More specifically, we observe that d_H is simply the sliced Cramer-2 distance, which has been considered by both theoretical and empirical works (Knop et al., 2018; Kolouri et al., 2022).

Assuming for a moment that $d = 1$ and letting $F_\mathbb{P}, F_\mathbb{Q}$ denote the cumulative distribution functions of $\mathbb{P}$ and $\mathbb{Q}$, the Cramer-p distance for $1 \le p \le \infty$ is defined by

$$\mathsf{CR}_p(\mu, \nu) \triangleq \|F_\mu - F_\nu\|_p. \tag{9}$$

With the definition of CR_p in hand and returning to general dimension d, it is straightforward to observe that d_H has the following alternative interpretation: take a uniformly random direction $v \in \mathbb{S}^{d-1}$ and compute the Cramer-2 distance between the projection of $\mathbb{P}$ and $\mathbb{Q}$ onto the line $\{\lambda v : \lambda \in \mathbb{R}\}$.

Proposition 5. *For all μ, ν we have*

$$d_H^2(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathsf{CR}_2^2(\theta_v \# \mu, \theta_v \# \nu) \mathrm{d}\sigma(v). \tag{10}$$

Proof Follows from $F_{\theta_v \# \mu}(b) - F_{\theta_v \# \nu}(b) = \mu(\Sigma_{vb}) - \nu(\Sigma_{vb})$. ■

2.4 Riesz potential

Given $0 < s < d$, the Riesz potential $I_s f$ of a compactly supported measure f on $\mathbb{R}^d$ is defined (in a weak sense) by

$$I_s f = f * K_s,$$

where $K_\alpha(x) = \frac{1}{c_\alpha} \frac{1}{\|x\|^{d-s}}$ and the constant is given by $c_s = \pi^{d/2} 2^s \frac{\Gamma(s/2)}{\Gamma((d-s)/2)}$. The Riesz potential is also a Fourier multiplier, that under sufficient regularity (Landkof, 1972)

$$\widehat{I_s f}(\omega) = \|\omega\|^{-s} \hat{f}(\omega).$$

Therefore, provided $|\alpha + \frac{1}{2}| < \frac{d}{2}$ and under sufficient regularity,

$$d_\alpha^2(\mu,\nu) = 2^d \pi^{d-1} \| I_{\frac{d}{2}+\alpha+\frac{1}{2}}(p-q) \|_2^2.$$

3 Properties of the generalized energy distance

3.1 Convergence of empirical measure under d_α

The following lemma shows that d_α converges at the parametric rate between empirical and population measures. In Section 1 we already sketched the proof of this fact for d_H using the Vapnik-Chervonenkis dimension of the class of halfspaces $\mathcal{A}$. For d_α with $\alpha \neq 0$ we need a more general argument. Recall that $M_t(\nu)$ denotes the t'th absolute moment of the measure ν.

Lemma 6. *Let ν be a probability measure on $\mathbb{R}^d$ and let $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ for an i.i.d. sample $X_1, \ldots, X_n \overset{iid}{\sim} \nu$. Then for any $|\alpha| < 1$,*

$$\mathbb{E} d_\alpha^2(\nu, \nu_n) \leq \frac{8 \operatorname{vol}_{d-1}(\mathbb{S}^{d-1}) M_{1+\alpha}(\nu)}{\pi(1-\alpha^2)} \frac{1}{n}.$$

Proof Note that for all $t \in \mathbb{R}$ the inequality $\sin^2(t) + (\cos(t) - 1)^2 \le 4(t^2 \wedge 1)$ holds. Let $X \sim \nu$. We use Proposition 3 and Tonelli's theorem to interchange integrals to obtain

$$\begin{aligned}
\pi \mathbb{E} d_\alpha^2(\nu, \nu_n) &= \int_{\mathbb{R}^d} \frac{\mathbb{E}|\hat{\nu} - \hat{\nu}_n|^2}{\|\omega\|^{d+\alpha+1}} \mathrm{d}\omega \\
&= \frac{1}{n} \int_{\mathbb{R}^d} \frac{\mathrm{var}(\cos\langle X, \omega\rangle) + \mathrm{var}(\sin\langle X, \omega\rangle)}{\|\omega\|^{d+\alpha+1}} \mathrm{d}\omega \\
&\le \frac{1}{n} \mathbb{E} \int_{\mathbb{R}^d} \frac{(\cos\langle X, \omega\rangle - 1)^2 + (\sin\langle X, \omega\rangle)^2}{\|\omega\|^{d+\alpha+1}} \mathrm{d}\omega \\
&\le \frac{4 \,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})}{n} \mathbb{E} \int_0^\infty \frac{1 \wedge (r^2 \|X\|^2)}{r^{2+\alpha}} \mathrm{d}r \\
&= \frac{4 \,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1})}{n} \mathbb{E} \left\{ \|X\|^2 \int_0^{\|X\|^{-1}} \frac{1}{r^\alpha} \mathrm{d}r + \int_{\|X\|^{-1}}^\infty \frac{1}{r^{\alpha+2}} \mathrm{d}r \right\} \\
&= \frac{8 \,\mathrm{vol}_{d-1}(\mathbb{S}^{d-1}) M_{1+\alpha}(\nu)}{(1-\alpha^2)} \frac{1}{n}.
\end{aligned}$$

■

This settles our claim about the behavior of d_α between empirical and population measures. Lemma 6 also serves as a necessary ingredient for proving Theorem 1 and Theorem 2.

3.2 Separating power of d_α in L^2

We now turn to comparison inequalities between L^2 separation and d_α.

Theorem 7. *Let μ, ν be two absolutely continuous probability measures on $\mathbb{R}^d$.*

1. *Suppose $\mu, \nu \in \mathcal{P}_S(\beta, d)$ for some $\beta > 0$. There exists a constant $C \in (1, \infty)$ depending on d, β, α only such that*

$$\frac{1}{C} \|\mu - \nu\|_2^{\frac{2\beta+d+1+\alpha}{2\beta}} \le d_\alpha(\mu, \nu). \tag{11}$$

   Moreover, for any value of $\epsilon \in (0, 1)$, there exist $\mu_\epsilon, \nu_\epsilon \in \mathcal{P}_S(\beta, d)$ such that $\|\mu_\epsilon - \nu_\epsilon\|_2 / \epsilon \in (1/C, C)$ and

$$d_\alpha(\mu_\epsilon, \nu_\epsilon) \le C \|\mu_\epsilon - \nu_\epsilon\|_2^{\frac{2\beta+d+1+\alpha}{2\beta}} \log\left(3 + \frac{1}{\|\mu_\epsilon - \nu_\epsilon\|_2}\right)^{C \mathbb{1}\{d \ge 2\}}. \tag{12}$$

2. *Suppose $\mu, \nu \in \mathcal{P}_G(d)$. There exists a constant $C < \infty$ depending on d, α only such that*

$$d_\alpha(\mu, \nu) \ge \frac{1}{C} \frac{\|\mu - \nu\|_2}{\log(3 + 1/\|\mu - \nu\|_2)^{\frac{d+1+\alpha}{4}}}. \tag{13}$$

Theorem 7 is our main technical result. It shows that d_α is lower bounded by a polynomial of the L^2 distance for both the smooth distribution class $\mathcal{P}_S$ and Gaussian mixtures $\mathcal{P}_G$. Moreover, 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. We also note that (11) follows from the Gagliardo–Nirenberg-Sobolev interpolation inequality when $d = 1$. However, to our knowledge, the inequality is new for $d > 1$.

The proofs of (11) and (13) are surprisingly straightforward: one simply applies Parseval's theorem and Hölder's inequality to obtain

$$\|\mu - \nu\|_2^2 \leq \||\hat{\mu}(\omega) - \hat{\nu}(\omega)|^{2/p}\|\omega\|^{\gamma}\|_p \cdot \||\hat{\mu}(\omega) - \hat{\nu}(\omega)|^{2/p^*}\|\omega\|^{-\gamma}\|_{p^*}.$$

Provided we choose $\gamma > 0$ and Hölder conjugates p, p^* appropriately, the first term on the right hand side corresponds to the $(\beta, 2)$ Sobolev norm of the difference of the distributions, while the second term corresponds to the generalized energy distance d_α between them. For the Gaussian mixture class $\mathcal{P}_G$ there is an additional step where we must bound the Sobolev norm of the Gaussian density as $\beta \to \infty$. The main theoretical difficulty lies in establishing the tightness of (11) for the smooth class $\mathcal{P}_S$, which we discuss next.

The inspiration for the construction is to have the density difference $f = \frac{\mathrm{d}\nu}{\mathrm{d}x} - \frac{\mathrm{d}\mu}{\mathrm{d}x}$ saturate Hölder's inequality above. The final form of the construction is $f = gh$, where $g(x) \propto \kappa\|x\|^{1-d/2}J_{d/2-1}(\|rx\|)$ with tuning parameters r, κ, and J denotes the Bessel function of the first kind. In particular g is proportional to the inverse Fourier transform of a sphere of radius r (to saturate Hölder's inequality). Then, h is chosen to be a compactly supported bump function that is zero in a neighbourhood of the origin, which kills the spike of g at the origin as we take $r \to \infty$. We also need that $\hat{h}|_{r\mathbb{S}^{d-1}} \equiv 0$ for infinitely large values of $r > 0$ with bounded gaps. The final property that h must satisfy is that the tails of $\hat{h}$ must decay rapidly; towards this end we use the recent result of (Cohen, 2023) who constructs a suitable h with $|\hat{h}(\omega)| \lesssim \exp(-\|\omega\|/\log(3 + \|\omega\|)^2)$.

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

Lemma 8 (Informal)**.** *Let f be constructed as above and let $s > -(d/2 + 1)$. Then there exists a sequence $r_n \to \infty$ with $|r_{n+1} - r_n| = \mathcal{O}(1)$ such that*

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

The utility of Lemma 8 is that we can use it to bound both the energy distance d_α between μ and ν (corresponding to $s = -\frac{d+1+\alpha}{2}$) as well as as the order $(\beta, 2)$ Sobolev norm of f (corresponding to $s = \beta$). Its proof hinges on the near-exponential decay of $\hat{h}$ and the fact that we may choose $h|_{r_n\mathbb{S}^{d-1}} \equiv 0$ for all n which we couple with estimates utilizing Lipschitz continuity.

4 The minimum energy density estimator

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \nu$ for some probability measure ν on $\mathbb{R}^d$. Given a class of distributions $\mathcal{G}$ and $|\alpha| < 1$, define the minimum-d_α estimator as

$$\tilde{\nu} \in \operatorname*{arg\,min}_{\nu' \in \mathcal{G}} d_\alpha(\nu', \nu_n), \tag{14}$$

where $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Note that this doesn't quite agree with our definition of $\tilde{\nu}$ in (1), because the $\alpha = 0$ case corresponds to the *average* halfspace distance d_H and not the maximum halfspace distance $\overline{d_H}$. Nevertheless, the following results in Section 4.1 bound the performance of $\tilde{\nu}$ as defined in (14), in terms of L^2 distance, as an estimator of ν for our two classes of interest. In Section 4.2 we apply these results to prove Theorem 1 and Theorem 2.

4.1 Rate of convergence of $\tilde{\nu}$ in L^2

Theorem 9. *Let $\nu, \tilde{\nu}, \mathcal{G}, \alpha$ be as in (14) and assume $\nu \in \mathcal{G}$, i.e. the problem is well-specified.*

1. *If $\mathcal{G} = \mathcal{P}_S(\beta, d)$ for some $\beta > 0$ then there exists a finite constant C depending on β, d only such that*
$$\mathbb{E}\|\tilde{\nu} - \nu\|_2 \le C(n(1-\alpha^2))^{-\frac{\beta}{2\beta+d+1+\alpha}}.$$
2. *If $\mathcal{G} = \mathcal{P}_G(d)$ then there exists a finite constant C depending on d only such that*
$$\mathbb{E}\|\tilde{\nu} - \nu\|_2 \le C\phi(n(1-\alpha^2)),$$
where $\phi(x) = \frac{\log(3+1/x)^{\frac{d+1+\alpha}{4}}}{\sqrt{x}}$.

Proof Let us focus on the case $\mathcal{G} = \mathcal{P}_S(\beta, d)$ first and let $t = \frac{2\beta+d+1+\alpha}{2\beta}$. The inequality $d_\alpha(\tilde{\nu}, \nu_n) \le d_\alpha(\nu, \nu_n)$ holds almost surely by the definition of $\tilde{\nu}$. Combining Theorem 7 and lemma 6 with the triangle inequality for d_α and Jensen's inequality, we obtain for a $\mathcal{G}$-dependent finite constant C, the chain of inequalities

$$\begin{aligned}
\mathbb{E}\|\tilde{\nu} - \nu\|_2 &\le \mathbb{E}\Big[(Cd_\alpha(\tilde{\nu}, \nu))^{1/t}\Big] \\
&\le \mathbb{E}\Big[(2Cd_\alpha(\nu, \nu_n))^{1/t}\Big] \\
&\le (2C\mathbb{E}d_\alpha(\nu, \nu_n))^{1/t} \\
&\lesssim (n(1-\alpha^2))^{-1/2t};
\end{aligned}$$

this substantiates the first claim. The result for $\mathcal{G} = \mathcal{P}_G$ follows analogously. Define the function $r(x) = x/\log(3+1/x)^{\frac{d+1+\alpha}{4}}$. It is easy to check that r is strictly increasing on $\mathbb{R}_+$ and thus so is its inverse function r^{-1}. Moreover, one can check that $r''(s) \ge 0$ for all $s \ge 0$ meaning that r is convex. From Theorem 7, there exists a d-dependent finite constant C such that we obtain the chain of inequalities

$$\begin{aligned}
\mathbb{E}\|\tilde{\nu} - \nu\|_2 &\le \mathbb{E}\Big[r^{-1}(Cd_\alpha(\tilde{\nu}, \nu))\Big] \\
&\le r^{-1}\left(C\mathbb{E}d_\alpha(\tilde{\nu}, \nu)\right) \\
&\le r^{-1}\left(2C\mathbb{E}d_\alpha(\nu, \nu_n)\right) \\
&\le r^{-1}\left(C'(n(1-\alpha^2))^{-1/2}\right)
\end{aligned}$$

for a d-dependent constant $C' < \infty$. The conclusion then follows by Lemma 18. ■

In both cases of Theorem 9 the optimal decay rates are known: for $\mathcal{P}_S$ it is given by $n^{-\beta/(2\beta+d)}$ (Ibragimov and Khasminskii, 1983) and for Gaussian mixture estimation it is $\log(n)^{d/4}/\sqrt{n}$ (Kim and Guntuboyina, 2022). Notice that the rate of estimation of the minimum d_α density estimator improves as $\alpha \downarrow -1$, and in fact seems to approach the optimum. However, simultaneously, the α-dependent constants also degrade in the inequalities we rely on. It turns out that taking $\alpha_n = \log(n)^{-1} - 1$ (resp. $\alpha_n = \log\log(n)^{-1} - 1$) results in the best possible result, which shows that the minimum d_{α_n} density estimator is optimal up to a $\mathrm{polylog}(n)$ (resp. $\mathrm{polyloglog}(n)$) factor over $\mathcal{P}_S$ (resp. $\mathcal{P}_G$).

Corollary 10. *Let $\nu, \tilde{\nu}, \mathcal{G}, \alpha$ be as in (14) and assume $\nu \in \mathcal{G}$, i.e. the problem is well-specified.*

1. *If $\mathcal{G} = \mathcal{P}_S(\beta, d)$ for some $\beta > 0$ and we take $\alpha = (\log n)^{-1} - 1$, then*
$$\mathbb{E}\|\tilde{\nu} - \nu\|_2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}.$$
2. *If $\mathcal{G} = \mathcal{P}_G(d)$ and we take $\alpha = (\log\log n)^{-1} - 1$, then*
$$\mathbb{E}\|\tilde{\nu} - \nu\|_2 \lesssim \frac{\log(n)^{d/4}\sqrt{\log\log n}}{\sqrt{n}}.$$

4.2 Proofs of Theorem 1 and Theorem 2

The final step required to show Theorem 1 and Theorem 2 is to prove results analogous to Theorem 7 and Theorem 9 with the total variation in place of the L^2 norm and for $\overline{d_H}$ instead of $d_H = d_0$. The next technical lemma shows that we can move between $\overline{d_H}$ and d_H as well as between L^1 and L^2 at the cost of a poly-logarithmic factor.

Lemma 11. *There exists a sufficiently large constant $C < \infty$ depending on the dimension d only such that the following hold.*

1. *Let $\mu, \nu \in \mathcal{P}_S(\beta, d)$ for some $\beta > 0$. Then*
$$d_H(\mu,\nu) \le C\overline{d_H}(\mu,\nu) \qquad \text{and} \qquad \|\mu-\nu\|_1 \le C\|\mu-\nu\|_2.$$
2. *Let $\mu, \nu \in \mathcal{P}_G(d)$. Then*
$$\frac{d_H(\mu,\nu)}{\log(3+1/d_H(\mu,\nu))^{1/4}} \le C\overline{d_H}(\mu,\nu) \quad \text{and} \quad \|\mu-\nu\|_1 \le C\|\mu-\nu\|_2 \log(3+1/\|\mu-\nu\|_2)^{d/4}.$$

Proof The claims for $\mathcal{P}_S$ follow immediately due to the bounded support of μ and ν and Hölder's inequality. Thus, we focus on the Gaussian mixture case. Write $\mu - \nu = f * \phi$ where ϕ denotes the density of the standard Gaussian $\mathcal{N}(0, I_d)$ and f is the difference of the two implicit mixing measures. Clearly, for any $R > 0$, we have

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

where $\mathrm{d}\sigma$ denotes the unnormalized surface measure on $\mathbb{S}^{d-1}$. Now, by definition since f is supported on a subset of $\mathbb{B}(0,1)$, for any $v \in \mathbb{S}^{d-1}$ and $R \ge 2$ we have the bound

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

where we implicitly used that $\int \mathrm{d}f = 0$ as f is the difference of two probability measures. Choosing $R \asymp \sqrt{\log(3 + 1/d_H(\mu,\nu))}$ concludes the proof the first claim. For the second claim, connecting L^1 to L^2, we simply extend the proof of (Jia et al., 2023, Theorem 22) to multiple dimensions. Let $\mu - \nu = f * \phi$ be as before. For any $R \geq 2$ we have

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

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 recover the desired result. ■

The final technical fact that we require is that the maximum halfspace distance $\overline{d_H}$ also satisfies a result similar to Lemma 6.

Lemma 12. *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*

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

for a finite universal constant C.

Proof Follows by (Vershynin, 2018, 8.3.23) and the fact that $\mathcal{D}$, the family of halfspace indicators, has VC dimension $d + 1$. ■

In light of Lemmas 11 and 12 and the previous subsection, the proof of Theorem 1 and Theorem 2 is an exercise in combining results. For the sake of completeness, we describe these steps.
Proof of Theorem 1 and Theorem 2 Apply Theorem 7 with $\alpha = 0$ (recalling that $d_0 = d_H$, see Section 2), to get a comparison between L^2 distance and d_H. Apply Lemma 11 to get a comparison between L^1 and L^2 as well as d_H and $\overline{d_H}$. Finally, plug in Lemma 12 (in place of Lemma 6) into the proof of Theorem 9. ■

4.3 A stopping criterion for training density estimators

As a corollary, we propose a stopping criterion for training density estimators.

Lemma 13. *Let $\mathbb{P}$ be compactly supported and let $\mathbb{P}_n$ be its empirical measure based on n i.i.d. observations. For $-1/2 < \alpha \leq 0$ and all $t \geq 0$ we have*

$$\mathbb{P}(d_\alpha(\mathbb{P}, \mathbb{P}_n) \geq \frac{c_1}{\sqrt{n}} + t) \leq 2\exp(-c_2 n t^2),$$

where the constants c_1, c_2 depend on $\alpha, d, \sup_{x \in \mathrm{supp}(\mathbb{P})} \|x\|$.

Proof Recall the definition of the fractional Brownian motion covariance kernel $K_\gamma(x,y) = \|x\|^\gamma + \|y\|^\gamma - \|x-y\|^\gamma$, and set $\gamma = 1+2\alpha$. For $\alpha \in (-\frac{1}{2}, 0]$ we have $0 \le K(x,y) \le 2\sup_{x\in\mathrm{supp}(\mathbb{P})} \|x\|^\gamma$. The conclusion follows by Theorem 7 of (Gretton et al., 2012) and Proposition 3. ■

Let $\mathbb{P}^\star \in \mathcal{P}_S^{\beta,d}(C)$ for a constant C, supported on $[0,1]^d$ for simplicity. Let $\mathbb{P}_n$ be its empirical version based on n i.i.d. observations. Assume further that $\{\mathbb{Q}_k\}_{k\ge1} \subset \mathcal{P}_S^{\beta,d}(C)$ is a sequence of density estimators each with support in $[0,1]^d$. Finally, for each k let $\mathbb{Q}_{k,m_k}$ be the empirical measure of $\mathbb{Q}_k$ based on m_k i.i.d. observations.

Proposition 14. *There exist constants c, c' depending only on α, β, d such for all $\delta \in (0,1)$, taking $m_k = c' n \log(k^2/\delta)/\log(1/\delta)$ ensures that*

$$\mathbb{P}\left(\mathsf{TV}(\mathbb{Q}_k, \mathbb{P}^\star) \le c\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{n}} + d_\alpha(\mathbb{Q}_{k,m_k}, \mathbb{P}_n^\star)\right)^{\frac{2\beta}{2\beta+d+1+2\alpha}},\ \forall\, k \ge 1\right) \ge 1 - 2\delta.$$

Note that m_k grows as $k \to \infty$, which ensures that our bound on the probability holds for all k simultaneously. The empirical relevance of such a result is immediate: suppose one have proposed candidate generative models $\mathbb{Q}_1, \mathbb{Q}_2, \ldots$ (e.g. one after each period of training epochs, or from different training models) that is trained from n inputs in $\mathbb{P}^*$. A verifier only needs to request for m_k inputs from each candidate (and n fresh inputs from P^*), and for $\alpha + \frac{1}{2} \asymp (\log n)^{-1}$, if we ever achieve $d_\alpha\left(\mathbb{Q}_{k,m_k}, \mathbb{P}_n\right) \lesssim \frac{1}{\sqrt{n}}$ we can stop training and claim a constant factor away from the (near-)minimax optimality with a high probability guarantee.

5 Conclusion

In this paper, we analyzed a simple discriminating class of affine classifiers and proved its effectiveness in the ERM-GAN setting (Equation (1)) 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 9 and corollary 10) 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 d_α, TV, and L^2 distances, and demonstrate the (log) tightness of these relationships via corresponding lower bound constructions (Appendix E). We also interpreted our distance in several ways that help advocate for its use in real applications. This work connects to a broader literature analyzing theoretically the success of GAN-styled generative models.

One immediate open direction is the setting where a single half-space is replaced by the intersection of multiple half-spaces, representing e.g. shallow ReLU-networks (whereas our current work focuses on one neuron). Whether or not it helps fill in the $d+1 \to d$ gap in the min-d_H estimation rate is unknown.

As described in Section 1.2, we view our discriminator as having great concentration but sub-optimal expressiveness power, versus e.g. the Yatracos' estimator, where the power of expressiveness comes at the expense of slower concentration. Aside from finding a discriminator class that concentrates at a parametric rate while achieving optimal sample complexity in density learning, another open problem is to characterize what's in between this trade-off. 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? We believe that understanding this expressiveness-concentration tradeoff is necessary for designing more sample-efficient and practical GAN models.

Acknowledgments and Disclosure of Funding

Thank you so much

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.

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.

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.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

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

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

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, Max Welling, et al. 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 Şimşekli. 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.

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

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.

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

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.

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. URL https://doi.org/10.7551/mitpress/4175.001.0001.

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.

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.

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

George N Watson. *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press, 1995. ISBN 9780521483919.

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.

Appendix A. Technical preliminaries

Definition 15. *We define the inner product of two functions $f, g : \mathbb{R}^d \to \mathbb{C}$ as*

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

Definition 16. *Given a function $f \in L^1(\mathbb{R}^d)$ we define its Fourier transform as*

$$\mathcal{F}[f](\omega) \triangleq \hat{f}(\omega) \triangleq \int_{\mathbb{R}^d} e^{-i\langle\omega,x\rangle} f(x)\mathrm{d}x, \tag{15}$$

and its inverse Fourier transform as

$$\mathcal{F}^{-1}[f](\omega) \triangleq \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i\langle\omega,x\rangle} f(x)\mathrm{d}x. \tag{16}$$

We use the same notation convention for Fourier transform on L^2.

Theorem 17 (Plancherel theorem)**.** *Let $f, g \in L^2(\mathbb{R}^d)$. Then*

$$\int_{\mathbb{R}^d} f(x)\overline{g(x)}\mathrm{d}x = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{f}(\omega)\overline{\hat{g}(\omega)}\mathrm{d}\omega. \tag{17}$$

Appendix B. Auxiliary technical results

Lemma 18. *Suppose $t, x, y > 0$. Then there exist finite t-dependent constants C_1, C_2, C_3 such that*

$$x \le C_1\, y\log(3+1/y)^t \implies \frac{x}{\log(3+1/x)^t} \le C_2\, y \implies x \le C_3\, y\log(3+1/y)^t.$$

Proof Let us focus on the first implication. We use $\lesssim, \gtrsim$ to suppress universal constants in the notation. If $x \lesssim y$ then the first implication clearly holds. If $y \lesssim x \lesssim y\log(3+1/y)^t$ then

$$\frac{x}{\log(3+1/x)^t} \lesssim \frac{y\log(3+1/y)^t}{\log(3+1/(y\log(3+1/y)^t))^t} \lesssim y.$$

The second inequality is equivalent to $\log(3+1/y)^t \lesssim \log(3+1/(y\log(3+1/y)^t))^t$ which is the same as requiring $3 + 1/y \le (3+1/(y\log(3+1/y)^t))^{\sqrt[t]{C}}$ for some t-dependent constant; this clearly holds. The second implication follows analogously. ■

Lemma 19. *Let μ be a probability measure on $\mathbb{R}^d$ and $\alpha \in (-1/2, 1/2)$. Then*

$$\mathbb{E}_{X\sim\mu}\int_{\mathbb{R}^d} \frac{(\cos\langle\omega, X\rangle - 1)^2 + \sin^2\langle\omega, X\rangle}{\|\omega\|^{d+1+2\alpha}}\mathrm{d}\omega \le \frac{4M_{1+2\alpha}(\mu)}{1-4\alpha^2},$$

where the implied constant depends on d only.

Proof We use the inequalities $|\cos t - 1| \le |t| \wedge 1$ and $|\sin t| \le |t| \wedge 1$ 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

$$\begin{aligned} 2\mathbb{E}\int_{\mathbb{R}^d} \frac{1 \wedge (\|\omega\|^2\|X\|^2)}{\|\omega\|^{d+1+2\alpha}}\mathrm{d}\omega &= 2\mathbb{E}\int_0^\infty \frac{1 \wedge (r^2\|X\|^2)}{r^{2+2\alpha}}\mathrm{d}r \\ &= 2\mathbb{E}\Big\{\|X\|^2\int_0^{\|X\|^{-1}} \frac{1}{r^{2\alpha}}\mathrm{d}r + \int_{\|X\|^{-1}}^\infty \frac{1}{r^{2\alpha+2}}\mathrm{d}r\Big\} \\ &= \frac{4M_{1+2\alpha}(\mu)}{1-4\alpha^2}. \end{aligned}$$

■

Lemma 20. *For $\alpha \in (-\frac{1}{2}, \frac{1}{2})$, define*

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

Then $C_\alpha < \infty$.

Proof In the $\alpha > 0$ case, one has immediately $C_\alpha \le \int_0^\infty \frac{2}{\omega^{1+\alpha}}\mathrm{d}\omega < \infty$.

Observe that

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

and hence it suffices to bound the two terms on the right hand side separately. Observe that $h(x+\pi) = -h(x)$ for both sine and cosine, one has that (for $n \ge 1$):

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

and

$$\left|\int_{2n\pi}^{(2n+2)\pi} \frac{\cos(\omega)}{\omega^{1+\alpha}}\mathrm{d}\omega\right| \le \left|\int_{2n\pi}^{(2n+1)\pi} \frac{\omega^{-\alpha}\cos(\omega)}{\omega} - \frac{\omega^{-\alpha}\cos(\omega)}{\omega+\pi}\mathrm{d}\omega\right| \lesssim \left|\int_{2n\pi}^{(2n+1)\pi} \omega^{-2-\alpha}\mathrm{d}\omega\right|$$

both of which converges when summed over n. It thus suffice to bound the $\sup_{a<c<a+2\pi}\left|\int_a^c f(\omega)\mathrm{d}\omega\right|$ term. Note that $|\sin(x)/x| \le 1$ for all x and $|\int_a^c \frac{\cos(\omega)}{\omega^{1+\alpha}}\mathrm{d}\omega| \le \int_0^{2\pi} \frac{1}{\omega^{1+\alpha}}\mathrm{d}\omega < \infty$ holds, the last term $2\sup_{a<c<a+2\pi}|\cdot|$ term is also finite for any α. ■

Lemma 21. *Let* $\int_0^\infty \cdot \,\mathrm{d}\omega \triangleq \lim_{\epsilon\to 0}\int_{1/\epsilon\geq\omega\geq\epsilon}\cdot\,\mathrm{d}\omega$. *Then, for* $x\neq 0$ *the following hold:*

$$\mathbb{1}\{x>0\} = \frac{1}{2} + C_{\psi^{(\alpha)}}\int_0^\infty \frac{\sin(\omega x)}{\omega}\mathrm{d}\omega \qquad \textit{for } \alpha = 0, \tag{18}$$

$$|x|^\alpha = C_{\psi^{(\alpha)}}\int_0^\infty \frac{\cos(\omega x)}{\omega^{1+\alpha}}\mathrm{d}\omega \qquad \textit{for } -1/2<\alpha<0, \textit{ and} \tag{19}$$

$$|x|^\alpha = C_{\psi^{(\alpha)}}\int_0^\infty \frac{\cos(\omega x)-1}{\omega^{1+\alpha}}\mathrm{d}\omega \qquad \textit{for } 0<\alpha<1/2, \tag{20}$$

where

$$C_{\psi^{(\alpha)}} = \begin{cases} \left(2\cos(\frac{\pi\alpha}{2})\Gamma(-\alpha)\right)^{-1} & \textit{if } \alpha\neq 0, \\ \frac{1}{\pi} & \textit{if } \alpha = 0. \end{cases} \tag{21}$$

Proof

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

Assume from here on without loss of generality that $x>0$. For $-\frac{1}{2}<\alpha<0$, by the residue theorem,

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

Similarly, for $0<\alpha<\frac{1}{2}$, integration by parts and residue theorem gives

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

■

Lemma 22. *Let ϕ be the probability density function of $\mathcal{N}(0, \sigma I_d)$ and write $\hat{\phi}$ for its Fourier transform. Then, for any $\beta \geq 0$,*

$$\|\hat{\phi}(\omega)\|\omega\|^\beta\|_2^2 = \frac{\pi^{d/2}}{\Gamma(d/2)\sigma^{2\beta+d}}\Gamma\left(\frac{2\beta+d}{2}\right) \leq \frac{\pi^{d/2}\sqrt{2\pi e}}{\Gamma(d/2)\sigma^{2\beta+d}}\left(\frac{2\beta+d}{2e}\right)^{\frac{2\beta+d-1}{2}} e^{\frac{1}{6(2\beta+d-2)}}.$$

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

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

The claimed inequality follows by Stirling's formula. ■

Appendix C. Proofs in Section 2

C.1 Proof of Proposition 4

For $v \in \mathbb{S}^{d-1}$ and $b \in \mathbb{R}$ let $\psi_{vb}^{(\alpha)}(x) \triangleq \psi^{(\alpha)}(\langle v, x\rangle - b)$. To start with, we notice that

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\left[\mathbb{E}\psi_\alpha(\langle X, v\rangle - b) - \mathbb{E}\psi_\alpha(\langle Y, v\rangle - b)\right]^2 \mathrm{d}b\mathrm{d}\sigma(v) \\
&= \int_{\mathbb{S}^{d-1}}\int_{\mathbb{R}}\left(\int_{\mathbb{R}}\psi^{(\alpha)}(x-b)\mathrm{d}(\theta_v\#(\mu-\nu))(x)\right)^2 \mathrm{d}b\mathrm{d}\sigma(v),
\end{aligned} \tag{22}$$

where $\theta_v(x) = \langle v, x\rangle$ and $\#$ denotes the pushforward. To ease notation, let $A_\epsilon = [\epsilon, 1/\epsilon]$ for $\epsilon > 0$ and write $\mu_v \triangleq \theta_v\#(\mu-\nu)$. For each $v \in \mathbb{S}^{d-1}$, the measure μ_v has at most countably many atoms, therefore $b \mapsto \mu_v(\{b\}) = 0$ Leb-almost everywhere. Then, by Tonelli's theorem we can conclude that $\mu_v(\{b\}) = 0$ for $\sigma \otimes$ Leb-almost every (v, b), thus going forward we can focus on the case $x \neq b$. By Lemma 21,

$$\int_{\mathbb{R}}\psi^{(\alpha)}(x-b)\mathrm{d}\mu_v(x) = C_{\psi^{(\alpha)}}\int_{\mathbb{R}}\lim_{\varepsilon\to 0}\int_{A_\varepsilon}\left\{\begin{array}{ll}\dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \alpha = 0 \\ \dfrac{\cos(\omega(x-b))}{\omega^{1+\alpha}} & \text{if } -\dfrac{1}{2} < \alpha < 0 \\ \dfrac{\cos(\omega(x-b))-1}{\omega^{1+\alpha}} & \text{if } 0 < \alpha < \dfrac{1}{2}\end{array}\right\}\mathrm{d}\omega\mathrm{d}\mu_v(x).$$

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

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

where $C_\alpha < \infty$ depends only on α and is defined in Lemma 20. We now show that $\int_{\mathbb{R}} |x-b|^\alpha \mathrm{d}|\mu_v|(x) < \infty$ for $\sigma \otimes \mathrm{Leb}$-almost every b, v. To this end, let $S = \{(b,v) \in \mathbb{R} \times \mathbb{S}^{d-1} : \int_{\mathbb{R}} |x-b|^\alpha \mathrm{d}|\mu_v|(x) = \infty\}$ 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

$$\begin{aligned}
\int_{-B}^{B}\left(\int_{\mathbb{R}} |x-b|^\alpha \mathrm{d}(\theta_v\#\mu)(x)\right)^2 \mathrm{d}b &\leq \int_{-B}^{B}\int_{\mathbb{R}} |x-b|^{2\alpha}\mathrm{d}(\theta_v\#\mu)(x)\mathrm{d}b\\
&\leq 2\int_{\mathbb{R}}\int_0^{B+|x|} b^{2\alpha}\mathrm{d}b\mathrm{d}(\theta_v\#\mu)(x)\\
&\lesssim \int_{\mathbb{R}} (B+|x|)^{1+2\alpha}\mathrm{d}(\theta_v\#\mu)(x)\\
&\lesssim B^{1+2\alpha} + \mathbb{E}_{X\sim\mu}|\langle v, X\rangle|^{1+2\alpha}\\
&\leq B^{1+2\alpha} + \mathbb{E}_{X\sim\mu}\|X\|^{1+2\alpha},
\end{aligned}$$

which, after integration over $v \in \mathbb{S}^{d-1}$, leads to a contradiction if $M_{1+2\alpha}(\mu+\nu) < \infty$. Continuing under the assumption $M_{1+2\alpha}(\mu+\nu) < \infty$, we can apply the dominated convergence theorem to obtain

$$\int_{\mathbb{R}} \psi^{(\alpha)}(x-b)\mathrm{d}\mu_v(x) = C_{\psi^{(\alpha)}} \lim_{\varepsilon\to 0}\int_{\mathbb{R}}\int_{A_\varepsilon} \left\{\begin{array}{ll} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \alpha = 0\\ \dfrac{\cos(\omega(x-b))}{\omega^{1+\alpha}} & \text{if } -\dfrac{1}{2} < \alpha < 0\\ \dfrac{\cos(\omega(x-b))-1}{\omega^{1+\alpha}} & \text{if } 0 < \alpha < \dfrac{1}{2}\end{array}\right\} \mathrm{d}\omega\mathrm{d}\mu_v(x).$$

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

$$\int_{\mathbb{R}} \psi^{(\alpha)}(x-b)\mathrm{d}\mu_v(x) = C_{\psi^{(\alpha)}} \lim_{\varepsilon\to 0}\int_{A_\varepsilon}\int_{\mathbb{R}} \left\{\begin{array}{ll} \dfrac{\sin(\omega(x-b))}{\omega} & \text{if } \alpha = 0\\ \dfrac{\cos(\omega(x-b))}{\omega^{1+\alpha}} & \text{if } -\dfrac{1}{2} < \alpha < 0\\ \dfrac{\cos(\omega(x-b))-1}{\omega^{1+\alpha}} & \text{if } 0 < \alpha < \dfrac{1}{2}\end{array}\right\} \mathrm{d}\mu_v(x)\mathrm{d}\omega.$$

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

$$\begin{aligned}\int_{\mathbb{R}} \psi^{(\alpha)}(x-b)\mathrm{d}\mu_v(x) &= C_{\psi^{(\alpha)}} \lim_{\varepsilon\to 0}\int_{A_\varepsilon} \frac{1}{\omega^{1+\alpha}} \begin{Bmatrix} \Im(e^{-i\omega b}\overline{\hat{\mu}_v(\omega)}) & \text{if } \alpha = 0 \\ \Re(e^{-i\omega b}\overline{\hat{\mu}_v(\omega)}) & \text{if } \alpha \neq 0 \end{Bmatrix} \mathrm{d}\omega \\ &= C_{\psi^{(\alpha)}} \lim_{\varepsilon\to 0} \begin{cases} \Im\left(\hat{\phi}^{(\alpha)}_{\mu_v,\varepsilon}(b)\right) & \text{if } \alpha = 0 \\ \Re\left(\hat{\phi}^{(\alpha)}_{\mu_v,\varepsilon}(b)\right) & \text{if } \alpha \neq 0. \end{cases}\end{aligned} \tag{23}$$

where we write

$$\phi^{(\alpha)}_{\mu_v,\varepsilon}(\omega) = \frac{\overline{\hat{\mu}_v(\omega)}}{\omega^{1+\alpha}}\mathbb{1}_{A_\varepsilon}.$$

Notice that we have $\phi^{(\alpha)}_{\mu_v,\varepsilon}$ is bounded and compactly supported and thus lies in $L^p(\mathbb{R})$ for any p, which means that

$$\phi^{(\alpha)}_{\mu_v,\varepsilon} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}),$$

and so

$$\hat{\phi}^{(\alpha)}_{\mu_v,\varepsilon} \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R}).$$

Finally, let us write

$$\phi^{(\alpha)}_{\mu_v}(\omega) = \lim_{\epsilon\to 0}\phi^{(\alpha)}_{\mu_v,\epsilon}(\omega) = \frac{\overline{\hat{\mu}_v(\omega)}}{\omega^{1+\alpha}}\mathbb{1}\{\omega \geq 0\}$$

for every ω. We now show that $\phi^{(\alpha)}_{\mu_v} \in L^2(\mathbb{R})$ provided $M_{1+2\alpha}(\mu+\nu) < \infty$. Let $(X,Y) \sim \mu\otimes\nu$. We have

$$\begin{aligned}\int_{\mathbb{R}} |\phi^{(\alpha)}_{\mu_v}(\omega)|^2\mathrm{d}\omega &= \int_0^\infty \frac{|\hat{\mu}_v(\omega)|^2}{w^{2+2\alpha}}\mathrm{d}w \\ &= \int_0^\infty \frac{(\mathbb{E}\cos\langle\omega,X\rangle - \cos\langle\omega,Y\rangle)^2 + (\mathbb{E}\sin\langle\omega,X\rangle - \sin\langle\omega,Y\rangle)^2}{\omega^{2+2\alpha}}\mathrm{d}\omega.\end{aligned}$$

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 $\phi^{(\alpha)}_{\mu_v} \in L^2(\mathbb{R})$ by Lemma 19. Thus, by the dominated convergence theorem

$$\left\|\phi^{(\alpha)}_{\mu_v,\varepsilon} - \phi^{(\alpha)}_{\mu_v}\right\|_2 \to 0$$

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

$$\left\|\hat{\phi}^{(\alpha)}_{\mu_v,\varepsilon} - \hat{\phi}^{(\alpha)}_{\mu_v}\right\|_2 \to 0 \tag{24}$$

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 $\hat{\phi}^{(\alpha)}_{\mu_v,\varepsilon_n} \to \hat{\phi}^{(\alpha)}_{\mu_v}$ almost everywhere (we could also conclude this by Carleson's theorem). Therefore,

$$\int_{\mathbb{R}} \psi^{(\alpha)}(x-b)\mathrm{d}(\theta_v\#(\mu-\nu))(x) = \hat{\phi}^{(\alpha)}_{\mu_v}(b) \tag{25}$$

for $\sigma \otimes \mathrm{Leb}$-almost every $b \in \mathbb{R}, v \in \mathbb{S}^{d-1}$. Plug (25) into (22), by Parseval's identity (where we use again that $\phi_{v,h}^{(\alpha)} \in L^2(\mathbb{R})$),

$$\begin{aligned}
\text{RHS} &= C_{\psi^{(\alpha)}}^2 \int_{\mathbb{S}^{d-1}} \|\hat{\phi}_{\mu_v}^{(\alpha)}\|_2^2 \mathrm{d}\sigma(v) \\
&= 2\pi C_{\psi^{(\alpha)}}^2 \int_{\mathbb{S}^{d-1}} \|\phi_{\mu_v}^{(\alpha)}\|_2^2 \mathrm{d}\sigma(v) \\
&= 2\pi C_{\psi^{(\alpha)}}^2 \int_{\mathbb{S}^{d-1}} \int_0^\infty \frac{|\hat{\mu}_v(\omega)|^2}{\omega^{2+2\alpha}} \mathrm{d}\omega \mathrm{d}\sigma(v) \\
&= 2\pi C_{\psi^{(\alpha)}}^2 \int_{\mathbb{S}^{d-1}} \int_0^\infty \frac{|\mathcal{F}[\mu-\nu](\omega v)|^2}{\omega^{2+2\alpha}} \mathrm{d}\omega \mathrm{d}\sigma(v) \\
&= \pi C_{\psi^{(\alpha)}}^2 \int_{\mathbb{R}^d} \frac{|\mathcal{F}[\mu-\nu](\omega)|^2}{\|\omega\|^{d+1+2\alpha}} \mathrm{d}\omega,
\end{aligned}$$

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

$$\text{RHS} = C_{d_\alpha} \int_{\mathbb{R}^d} \frac{|\mathcal{F}[\mu-\nu](\omega)|^2}{\|\omega\|^{d+1+2\alpha}} \mathrm{d}\omega < \infty.$$

Appendix D. Proofs of Upper bounds in Theorem 7

D.1 Sobolev class

Proving separation Let p, q denote the densities of μ, ν. Since the $(\beta, 2)$-Sobolev norm of p and q are finite, we have $p, q \in L^2(\mathbb{R}^d)$ and we may apply Parseval's theorem

$$\|p - q\|_2^2 = \frac{1}{(2\pi)^d} \|\hat{p} - \hat{q}\|_2^2.$$

For any $\gamma > 0$, by Hölder's inequality with exponents $\frac{1}{r} + \frac{1}{r^*} = 1$ we have

$$\begin{aligned}
\|\hat{p} - \hat{q}\|_2^2 &= \int_{\mathbb{R}^d} |\hat{p}(\omega) - \hat{q}(\omega)|^2 \frac{\|\omega\|^\gamma}{\|\omega\|^\gamma} \mathrm{d}\omega \\
&\le \left(\int_{\mathbb{R}^d} |\hat{p}(\omega) - \hat{q}(\omega)|^2 \|\omega\|^{\gamma r} \mathrm{d}\omega \right)^{1/r} \left(\int_{\mathbb{R}^d} \frac{|\hat{p}(\omega) - \hat{q}(\omega)|^2}{\|\omega\|^{\gamma r^*}} \mathrm{d}\omega \right)^{1/r^*}.
\end{aligned} \tag{26}$$

Now, we choose γ and r to satisfy

$$\begin{aligned}
\gamma r &= 2\beta \\
\gamma r^* &= d + 1 + 2\alpha.
\end{aligned}$$

The first equation ensures that the first integral term is bounded by $\|p - q\|_{\beta,2}^{2/r}$, which is assumed constant, and the second equation ensures that the second integral term is equal to $d_\alpha(\mu, \nu)^{2/r^*}$ up to dimension dependent constant. The solution to this system of equations is given by $r^* = (2\beta + d + 2\alpha + 1)/(2\beta)$ and $\gamma = 2\beta \cdot \frac{d+2\alpha+1}{2\beta+d+2\alpha+1}$. Note that clearly $\gamma > 0$ and $r^* \ge 1$. Thus, after rearrangement we obtain

$$\|\hat{p} - \hat{q}\|_2^{\frac{2\beta+d+2\alpha+1}{2\beta}} \lesssim C_{d_\alpha}^{-1/2} d_\alpha(\mu, \nu),$$

concluding the proof.

D.2 Gaussian mixtures

Proof Let μ and ν have densities $p * \phi$ and $q * \phi$, where ϕ is the density of $\mathcal{N}(0, I_d)$. Writing $f = (p - q) * \phi$, and applying Hölder's inequality analogously to the proof of the Sobolev class, we obtain

$$\begin{aligned}\|\hat{f}\|_2 &\le \|\|\omega\|^\beta \hat{f}(\omega)\|_2^{\frac{d+1+2\alpha}{d+1+2\alpha+2\beta}} \left\| \frac{\hat{f}(\omega)}{\|\omega\|^{\frac{d+1+2\alpha}{2}}} \right\|_2^{\frac{2\beta}{d+1+2\alpha+2\beta}} \\ &= C_{d_\alpha}^{\frac{\beta}{d+1+2\alpha+2\beta}} \cdot \|\|\omega\|^\beta \hat{f}(\omega)\|_2^{\frac{d+1+2\alpha}{d+1+2\alpha+2\beta}} \cdot d_\alpha(\mu,\nu)^{\frac{2\beta}{d+1+2\alpha+2\beta}}\end{aligned}$$

Using that $|\hat{f}(\omega)| \le |\hat{\phi}(\omega)|$ and applying Lemma 22 we obtain

$$\|\hat{f}\|_2 \le C_{d_\alpha}^{\frac{\beta}{d+1+2\alpha+2\beta}} d_\alpha(\mu,\nu)^{\frac{2\beta}{d+1+2\alpha+2\beta}} \left(\frac{\pi^{d/2}\sqrt{2\pi e}}{\Gamma(d/2)} \left(\frac{2\beta+d}{2e}\right)^{\frac{2\beta+d-1}{2}} e^{\frac{1}{6(2\beta+d-2)}} \right)^{\frac{d+1+2\alpha}{2(d+1+2\alpha+2\beta)}}.$$

Going forward we treat d as a fixed, and take $\beta = \log(1/\|f\|_2)$ and assume that $\beta \ge d$. We write $c_{d\alpha}$ for a d, α dependent constant that may change line to line. Rearranging and using Parseval's Theorem, we get

$$d_\alpha(\mu,\nu) \ge c_{d\alpha}\|f\|_2 \frac{\|f\|_2^{\frac{d+1+2\alpha}{2\beta}}}{\left(\frac{2\beta+d}{2e}\right)^{\frac{(d+1+2\alpha)(2\beta+d-1)}{8\beta}}}.$$

Plugging in for β and using that $0 \le x \mapsto x^{1/x}$ is uniformly bounded, we get

$$d_\alpha(\mu,\nu) \ge \frac{c_{d\alpha}\|f\|_2}{\log(1/\|f\|_2)^{\frac{d+1+2\alpha}{4}}}.$$

■

Appendix E. Tightness for smooth densities

E.1 Special case: tightness for one-dimension

Lemma 23. *Let $f(x) = 1\{|x| \le \pi\}\sin(rx)$ and write $f_\beta = f * \cdots * f$ for f convolved with itself $\beta - 1$ times ($f_1 = f, f_2 = f * f \ldots$). Fix $\beta \in \mathbb{N}, \alpha \in (-1/2, 1/2)$. As $r \to \infty$, we have:*

$$\|f_\beta\|_2 \asymp 1, \|f_\beta\|_{\beta,2} \asymp r^\beta, \|f_{\beta+1}\|_{\beta,2} \asymp r^\beta, d_\alpha(f_\beta) \asymp \frac{1}{r^{1+\alpha}}. \tag{27}$$

where constants may depend on α, β.

Proof The intuition for estimates (27) is simple: most of the energy of f (and hence f_β) is at frequencies around $|\omega| \approx r$ and thus β-times differentiation boosts L_2-energy by r^β, while d_H distance suppresses energy by r^{-1}. We proceed to computation.

A simple computation shows $\hat{f}(\omega) = c\frac{(-1)^r}{i}\frac{r}{\omega^2-r^2}\sin(\omega\pi)$. Thus, we have

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

We decompose integral into three regimes:

1. $\omega < r/2$: here $(\omega^2 - r^2) \asymp r^2$ and thus
$$\int_0^{r/2} \asymp r^{-2\beta}\int_0^{r/2}\sin^{2\beta}(\omega\pi) \asymp r^{1-2\beta}$$
2. $\omega > 3/2r$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus
$$\int_{3r/2}^\infty \asymp r^{2\beta}\int_{3r/2}^\infty \frac{\sin^{2\beta}(\omega\pi)}{\omega^{4\beta}} \asymp r^{2\beta}r^{-4\beta+1} = r^{1-2\beta}$$
3. $\omega \in [1/2r, 3/2r]$: here $(\omega^2 - r^2) \asymp yr$, where $y = \omega - r$. Note also $\sin(\omega\pi) = \sin(r\pi + y\pi) = (-1)^r\sin(y\pi)$. Thus
$$\int_{r/2}^{3r/2} d\omega = \int_{-r/2}^{r/2} dy \asymp r^{2\beta}\int_{-r/2}^{r/2} dy\frac{\sin^{2\beta}(y\pi)}{(yr)^{2\beta}} \asymp \int_{-\infty}^{\infty}\left(\frac{\sin(y\pi)}{y}\right)^{2\beta} \asymp 1\,,$$
where the last inequality can be seen by the fact that integrand function is bounded at 0 and has $\frac{1}{y^{2\beta}}$ tail.

This proves the first estimate of (27).

For the second integral we have

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

We decompose integral into three regimes:

1. $\omega < r/2$: here $(\omega^2 - r^2) \asymp r^2$ and thus
$$\int_0^{r/2} \asymp r^{-2\beta}\int_0^{r/2}\omega^{2\beta}\sin^{2\beta}(\omega\pi) \asymp r^{2\beta+1-2\beta}$$
2. $\omega > 3/2r$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus
$$\int_{3r/2}^\infty \asymp r^{2\beta}\int_{3r/2}^\infty \frac{\sin^{2\beta}(\omega\pi)\omega^{2\beta}}{\omega^{4\beta}} \asymp r^{2\beta}r^{-2\beta+1} = r$$
3. $\omega \in [1/2r, 3/2r]$: 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
$$\int_{r/2}^{3r/2} d\omega = \int_{-r/2}^{r/2} dy \asymp r^{2\beta}\int_{-r/2}^{r/2} dy\frac{\sin^{2\beta}(y\pi)r^{2\beta}}{(yr)^{2\beta}} \asymp r^{2\beta}\int_{-\infty}^{\infty}\left(\frac{\sin(y\pi)}{y}\right)^{2\beta} \asymp r^{2\beta}\,,$$
where the last inequality can be seen by the fact that integrand function is bounded at 0 and has $\frac{1}{y^{2\beta}}$ tail.

This proves the second estimate in (27). The third follows similarly from:

$$\|f_{\beta+1}\|_{\dot{\beta},2}^2 \asymp \int_0^\infty |\hat{f}(\omega)|^{2\beta+2}\omega^{2\beta} \asymp \int_0^\infty \frac{r^{2\beta+2}}{(\omega^2-r^2)^{2\beta+2}}\omega^{2\beta}\sin^{2\beta+2}(\omega\pi)\,.$$

We decompose integral into three regimes:

1. $\omega < r/2$: here $(\omega^2 - r^2) \asymp r^2$ and thus
$$\int_0^{r/2} \asymp r^{-2\beta-2}\int_0^{r/2}\omega^{2\beta}\sin^{2\beta+2}(\omega\pi) \asymp r^{-1}$$
2. $\omega > 3/2r$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus
$$\int_{3r/2}^\infty \asymp r^{2\beta+2}\int_{3r/2}^\infty \frac{\sin^{2\beta+2}(\omega\pi)\omega^{2\beta}}{\omega^{4\beta+4}} \asymp r^{2\beta+2}r^{-2\beta-3} = r^{-1}$$
3. $\omega \in [1/2r, 3/2r]$: 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
$$\int_{r/2}^{3r/2} d\omega = \int_{-r/2}^{r/2} dy \asymp r^{2\beta}\int_{-r/2}^{r/2} dy\frac{\sin^{2\beta+2}(y\pi)r^{2\beta}}{(yr)^{2\beta+2}} \asymp r^{2\beta}\int_{-\infty}^{\infty}\left(\frac{\sin(y\pi)}{y}\right)^{2\beta+2} \asymp r^{2\beta}$$
where the last inequality follows by that the integrand is bounded at 0 and has $\frac{1}{y^{2\beta+2}}$ tail.

For the last integral, we have

$$d_H(f)^2 \asymp \int_0^\infty |\hat{f}(\omega)|^{2\beta}\omega^{-2} \asymp \int_0^\infty \frac{r^{2\beta}}{(\omega^2-r^2)^{2\beta}}\omega^{-2-2\alpha}\sin^{2\beta}(\omega\pi)\,.$$

We decompose integral into three regimes:

1. $\omega < r/2$: here $(\omega^2 - r^2) \asymp r^2$ and thus
$$\int_0^{r/2} \asymp r^{-2\beta}\int_0^{r/2}\omega^{-2-2\alpha}\sin^{2\beta}(\omega\pi) \le r^{-2\beta}\int_0^\infty \omega^{-2-2\alpha}\sin^2(\omega\pi) \asymp r^{-2\beta}$$
2. $\omega > 3/2r$: here $(\omega^2 - r^2) \asymp \omega^2$ and thus
$$\int_{3r/2}^\infty \asymp r^{2\beta}\int_{3r/2}^\infty \frac{\sin^{2\beta}(\omega\pi)}{\omega^{4\beta+2+2\alpha}} \asymp r^{2\beta}r^{-4\beta-1-2\alpha} = r^{-2\beta-1-2\alpha}$$
3. $\omega \in [1/2r, 3/2r]$: 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
$$\int_{r/2}^{3r/2} d\omega = \int_{-r/2}^{r/2} dy \asymp r^{2\beta}\int_{-r/2}^{r/2} dy\frac{\sin^{2\beta}(y\pi)r^{-2-2\alpha}}{(yr)^{2\beta}} \asymp r^{-2-2\alpha}\int_{-\infty}^{\infty}\left(\frac{\sin(y\pi)}{y}\right)^{2\beta} \asymp r^{-2-2\alpha}\,,$$
where the last inequality can be seen by the fact that the integrand function is bounded at 0 and has $\frac{1}{y^{2\beta}}$ tail.

This proves the last estimate in (27). ■

Proving tightness in one dimension We now turn to showing tightness of the inequality, first in 1-dimension and for $\beta \in \mathbb{N}$. Let f_β be as in Lemma 23 with $r = \epsilon^{-1/\beta}$ for $\epsilon \in (0,1)$. Let p_0 a smooth, compactly supported density with $p_0(x) = \frac{1}{4\pi}$ for all $x \in [-\pi, \pi]$. Define

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

By Lemma 23 they satisfy

$$\|p_\epsilon - q_\epsilon\|_2 \asymp \epsilon,\ \|p_\epsilon - q_\epsilon\|_{\beta,2} \asymp 1,\ d_\alpha(p_\epsilon, q_\epsilon) \asymp \epsilon^{\frac{1+\alpha+\beta}{\beta}}.$$

Note that for $\epsilon < 1/4\pi$ both p_ϵ, q_ϵ are probability densities with

$$\|p_\epsilon\|_{\beta,2} + \|q_\epsilon\|_{\beta,2} \le 2\|p_0\|_{\beta,2} + \epsilon\|f_\beta\|_{\beta,2} \asymp 1.$$

This proves our tightness claim for $d = 1, \beta \in \mathbb{N}$. By showing that $\|f_{\beta+1}\|_{\beta,2} \asymp \epsilon^{1/\beta}$ and using the fact that

$$\|f_{\beta+1}\|_{\beta+s,2} \le \|f_{\beta+1,2}\|_\beta^{1-s}\|f_{\beta+1}\|_{\beta+1,2}^s$$

for all $s \in (0,1)$ which follows from Hölder's inequality, we can extend this result to general $\beta > 0$.

E.2 Compact construction: log-scale tight

For the discussions below, we will assume that the ambient dimension $d \ge 2$.

Lemma 24. *Let $a, b, c \in \mathbb{R}$ with $b > 0$ be constants and let η denote an arbitrarily small positive number. 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 $c \ge 0$. For all large enough x one has $\exp(-\frac{bx}{\log^2(x+2)}) < x^{-a-c-2}$, therefore

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

when $c > 0$. ■

Lemma 25. *Let J_ν be the Bessel function of the first kind of order ν.*

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

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

2. *Provided $\nu \ge -1/2$ and $|x| \le 1$, the inequality $J_\nu(x) \le \frac{x^\nu}{2^{\nu-1}\Gamma(\nu+1)}$ holds.*

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

$$J_\nu(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4}\right) + O(x^{-3/2}). \tag{28}$$

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

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

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

Proof The second claim follows easily from the series representation

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

which is valid for all $\nu \geq -1$ and $x \geq 0$. For the first, third, and fifth claims see standard references such as Watson (1995). For the fourth claim, see Olenko (2006). ■

Lemma 26. *There exists a function $h_0 \in L^2(\mathbb{R}^d)$ such that*

$$\mathrm{supp}(h_0) \in \mathbb{B}(0,1), \tag{29}$$

$$|\hat{h}_0(w)| \leq C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \qquad \text{for all } w \in \mathbb{R}^d, \tag{30}$$

$$|\hat{h}_0(w)| \geq \frac{1}{2} \qquad \text{for all } \|w\| \leq r_{\min}, \tag{31}$$

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

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

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

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

Lemma 27. *There exists a function $h \in L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d)$ and a sequence $\{r_n\}_{n=1}^\infty$ satisfying $0 < r_1 < r_2 < \cdots < r_n \to \infty$ and $\sup_k |r_{k+1} - r_k| = \mathcal{O}(1)$ such that*

$$\mathrm{supp}(h) \subset \mathbb{B}(0,1), \tag{32}$$

$$\mathrm{supp}(h) \subset \mathbb{R}^d \setminus \mathbb{B}(0, r_0), \tag{33}$$

$$|\hat{h}(w)| \leq C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \qquad \text{for all } w \in \mathbb{R}^d, \tag{34}$$

$$\hat{h}|_{\partial\mathbb{B}(0,r_n)} \equiv 0, \qquad \text{for } n \in \mathbb{Z}^+. \tag{35}$$

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

Proof First, we construct h_0 as per the requirements in Lemma 26. It already satisfies Equation (32) and Equation (34). To address the other two requirements, we modify h_0 by convolving it with two additional terms:

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

where A_0 and ρ_0 aim to address Equation (33) and Equation (35), respectively, and are defined as

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

Now, let's verify that h indeed satisfies the four requirements. Note that A_0 is an "annulus" supported on $\mathbb{B}(0, 5/8)\backslash\mathbb{B}(0, 3/8)$, and both $h_0(8x)$ and ρ_0 are supported on $\mathbb{B}(0, 1/8)$. Therefore, $\text{supp}(h) \subset \mathbb{B}(0, 7/8)\backslash\mathbb{B}(0, 1/8)$, which implies Equations (32) and (33). We now turn to the other two conditions in Fourier space. Note that

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

From Lemma 25,

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

as $\|w\| \to \infty$ and hence $\hat{\rho}_0(w) = (1/8)^d \mathcal{F}[\mathbb{1}\{\|x\| < 1\}](w/8)$ has infinitely many zeros around $\|w\| = 8(2n\pi + \frac{(d+1)\pi}{4})$ for sufficiently large $n \in \mathbb{Z}^+$, which implies Equation (35).

Finally, for Equation (34), note that since A_0 is a Schwartz function, so is $\hat{A}_0$ and thus

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

concluding the proof. ■

Given that h is bounded with compact support, we know that $\hat{h}$ is Lipschitz with some finite parameter L. This leads to the following Lemma.

Lemma 28. *Suppose $\hat{h}$ is L-Lipschitz and that there exist $C, c > 0$ with $|\hat{h}(w)| < C \exp\left(-\frac{c\|w\|}{\log(\|w\|+2)^2}\right) \triangleq H(\|w\|)$ for all $w \in \mathbb{R}^d$. Further suppose that $\hat{h}$ vanishes on $\partial\mathbb{B}(0, r_n)$ for a sequence $1 < r_1 < r_2 < \cdots < r_n \to \infty$. Then, for any $s > -d/2 - 1$ there exists n_0 such that*

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

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

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

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

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

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

Consider now the tail outside of the ball $\mathbb{B}(0, D\log^{1+\eta} r)$. We upper bound the integral noticing that $|\hat{h} \mathbb{1}_{\overline{\mathbb{B}(0, D\log^{1+\eta} r)}}(w)| < H(D\log^{1+\eta}(r) \vee \|w\|)$. For some $\gamma > 0$ to be specified later we write

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

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

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

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

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

where the last equality follows for any $\gamma \ge \frac{2s+d-1}{2s+d+2}$. The second term can be bounded since $\int_{r^{-\gamma}}^{2r} r^{d-1} \|w\|^{2s+d-1} \mathrm{d}\|w\|$ is clearly upper bounded by a polynomial of r, but $H(D\log^{1+\eta}(r))$ vanishes quicker than any polynomial of r. Finally, the last term also vanishes according to Lemma 24 due to the near-exponential decay

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

This concludes our proof. ■

In the construction below, we fix h, d, β and α effectively treating them as constants, while letting a single parameter $r \to \infty$ to get the desired asymptotics. Whenever we write $\lesssim, \gtrsim, \asymp$ we hide multiplicative constants that may depend on h, d, β and α.

Proposition 29. *Let $h, \{r_n\}_{n=0}^{\infty}$ be constructed following Lemma 27 for some $\eta > 0$ and take $\alpha \in (-1/2, 1/2)$. For every n large enough, there exists a function $g = g_n$, such that $f = gh$ satisfies the following properties for $r = r_n$:*

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

Proof Let $\kappa = r^{1/2-\beta}$. Let g be given by

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

so that it is proportional to the inverse Fourier transform of the surface measure of the sphere $\mathbb{B}(0,r)$. Writing g more explicitly from Lemma 25 we have

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

where J_ν denotes Bessel functions of the first kind of order ν. Notice that g is spherically symmetric and real-valued. We will later require some properties of J_ν, which are collected in Lemma 25. Let $f = gh$ so that $\hat{f} = \hat{g} * \hat{h}$; this choice is inspired by the equality case of Hölder's inequality in Equation (26). Next we verify the claimed properties one by one.

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

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

Showing $\|f\|_\infty \lesssim r^{-\beta}$. Recall that $h|_{\mathbb{B}(0,r_0)} = 0$, it thus suffices to bound $\|g|_{\overline{\mathbb{B}(0,r_0)}}\|_\infty$. By Lemma 25 we know that as $r = r_n \to \infty$, for any fixed x,

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

with constants that do not depend on n. Since h is independent of n, we get $\|f\|_\infty \lesssim r^{-\beta} \to 0$ as required.

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

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

Again, we are done by (28).

Showing $\|f\|_{\beta,2} \lesssim \kappa r^{\beta-1/2}(\log r)^{(d+1)/2}$ and $d_\alpha(f) \lesssim \kappa r^{-d/2-1-\alpha}(\log r)^{(d+1)/2}$. They follow from Lemma 28 in the following sense:

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

and

$$\begin{aligned}
d_\alpha(\mu,\nu) &= \left\| \frac{\hat{f}}{\omega^{\frac{d+1+2\alpha}{2}}} \right\|_2 \\
&\asymp \kappa r^{-d/2} \cdot \left\| \omega^{-\frac{d+1+2\alpha}{2}} \cdot \left(\delta(\|\omega\| - r) * \hat{h} \right)(w) \right\|_2 \\
&\lesssim \kappa r^{-\frac{d}{2}-1-\alpha} \log(r)^{d+\eta}.
\end{aligned}$$

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

$$\max_{v\in\mathbb{S}^{d-1}, b\in\mathbb{R}} \left| \int \psi^{(\alpha)}(\langle x, v\rangle - b) f(x) \mathrm{d}x \right| \overset{!}{\lesssim} \kappa r^{-d/2-1-\alpha}. \tag{36}$$

where the constants only depend on h, d. Given arbitrary $b \in \mathbb{R}$ and $v \in \mathbb{S}^{d-1}$ let

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

We split the argument into two cases. Suppose first that $\alpha \neq 0$. Then, using Lemmas 21 and 20, we know by dominated convergence that

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

Let D be a large constant independent of n, following similar steps to the proof of Lemma 28, we split

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

Consider the first term,

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

Consider the second term, split

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

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

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

Suppose now that $\alpha = 0$. By an analogous argument to the above, we obtain

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

As before, we obtain

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

as required, and the proof of (36) is complete.

These conclude our proof. ■

Corollary 30 (Lower bounds for compact distributions). *Fix $\alpha \in (-1/2, 1/2), \beta > 0, d \geq 1$, there exists a constant $C_0(\beta, d)$, such that for any $C > C_0$, for all small enough $\varepsilon > 0$, there exist a pair of probability densities $p = p_\varepsilon, q = q_\varepsilon$ in $\mathcal{P}(\mathbb{R}^d)$:*

1. *The densities $p_\varepsilon, q_\varepsilon$ are supported on the unit ball. Moreover, $\|p\|_{\beta,2}, \|q\|_{\beta,2} < C$.*
2. *The following inequality holds:*

$$\begin{aligned}
\mathsf{TV}(p, q) &\gtrsim \varepsilon \left(\log(1/\varepsilon)\right)^{-\frac{d+1}{2}}, \\
d_\alpha(p, q) &\lesssim \max_{v \in \mathbb{S}^{d-1}, b \in \mathbb{R}} \left| \int \psi^{(\alpha)}(\langle x, v\rangle - b)(p - q)(x) \mathrm{d}x \right| \\
&\lesssim \varepsilon^{\frac{2\beta + 2\alpha + d + 1}{2\beta}}.
\end{aligned}$$

hiding multiplicative factors involving α, β, d.

Proof Without loss of generality (losing only a constant factor), we assume that

$$r_n = \varepsilon^{-\frac{1}{\beta}}$$

for some $n \in \mathbb{Z}^+$, where $\{r_n\}_{n=0}^\infty$ is the sequence constructed in Proposition 29. The overarching idea is that it suffices for $p(x) - q(x)$ to satisfy Proposition 29. Let $f(x)$ be the function established in Proposition 29 when $r = r_n$. Let $m(x)$ represent any given non-negative smooth function with compact support $\mathbb{B}(0, 1)$ such that $\int m(x) \mathrm{d}x = 1$. Denote $C_0 = \|m\|_{\beta,2}$. Define:

$$p(x) = m(x) + cf(2x)$$

and

$$q(x) = m(x) - cf(2x)$$

To ensure that both p and q are valid densities, it suffices to have: $|m(x)| \geq |cf(2x)|$. for all $x \in \mathbb{R}^d$. This can be achieved by setting: $c \leq \inf_{\|x\| < 1/2} \frac{m(x)}{\|f\|_\infty}$. As demonstrated in Proposition 29, $\|f\|_\infty < \infty$. so such a c does exist. Considering the Sobolev norms, we obtain:

$$\|p\|_{\beta,2}, \|q\|_{\beta,2} \leq \|m\|_{\beta,2} + c\|f\|_{\beta,2} \leq \|m\|_{\beta,2} + c(\log r)^{\frac{d+1}{2}}$$

For the total variation norm:

$$\mathsf{TV}(p, q) = 2c\|f\|_1 \gtrsim cr^{-\frac{1}{\beta}} \gtrsim c\varepsilon$$

Regarding the generalized hyperplane distance:

$$d_\alpha(p, q) = d_\alpha(2cf) = 2cd_\alpha(f) \lesssim 2cr^{-1/2-\beta-d/2-\alpha}(\log r)^{(d+1)/2}$$

Therefore, by taking $c = \Theta((\log r)^{-\frac{d+1}{2}})$. we derive:

$$\begin{aligned}
\mathsf{TV}(p, q) &\gtrsim \varepsilon(\log \varepsilon^{-1})^{-\frac{d+1}{2}}, \\
d_\alpha(p, q) &\lesssim \varepsilon^{\frac{2\beta + 2\alpha + d + 1}{2\beta}}.
\end{aligned}$$

■