Kernel Quantile Embeddings and Associated Probability Metrics

Masha Naslidnyk [1] **Siu Lun Chau** [2] **François-Xavier Briol** [3] **Krikamol Muandet** [4]

Abstract

Embedding probability distributions into reproducing kernel Hilbert spaces (RKHS) has enabled powerful nonparametric methods such as the maximum mean discrepancy (MMD), a statistical distance with strong theoretical and computational properties. At its core, the MMD relies on kernel mean embeddings to represent distributions as mean functions in RKHS. However, it remains unclear if the mean function is the only meaningful RKHS representation. Inspired by generalised quantiles, we introduce the notion of *kernel quantile embeddings (KQEs)*. We then use KQEs to construct a family of distances that: (i) are probability metrics under weaker kernel conditions than MMD; (ii) recover a kernelised form of the sliced Wasserstein distance; and (iii) can be efficiently estimated with near-linear cost. Through hypothesis testing, we show that these distances offer a competitive alternative to MMD and its fast approximations.

1. Introduction

Many machine learning and statistical methods rely on representing, comparing, and measuring the distance between probability distributions. Kernel mean embeddings (KMEs) have been shown to be a mathematically and computationally convenient approach for this task (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Muandet et al., 2016). At its core, a KME represents a distribution as a mean function in a reproducing kernel Hilbert space (RKHS). When the kernel function is sufficiently regular and satisfies a condition called 'characteristic' (Sriperumbudur et al., 2010), the representation of a distribution as a KME is unique, capturing all information about the distribution. The probability metric constructed by comparing KMEs, called maximum mean discrepancy (MMD) (Borgwardt et al., 2006; Gretton et al., 2012), has received significant attention due to its computational tractability. Its most common estimator has cost $\mathcal{O}(n^2)$ and can be estimated with error $\mathcal{O}(n^{-1/2})$ in the number of data points n, but cheaper alternatives have also been proposed (Gretton et al., 2012; Chwialkowski et al., 2015a; Bodenham and Kawahara, 2023; Schrab et al., 2022). For this reason, KMEs and the MMD have been used to tackle a broad range of tasks from hypothesis testing (Gretton et al., 2012) to parameter estimation (Briol et al., 2019; Chérief-Abdellatif and Alquier, 2020), causal inference (Muandet et al., 2021; Sejdinovic, 2024), feature attribution (Chau et al., 2022; 2023), and learning on distributions (Muandet et al., 2012; Szabó et al., 2016).

Nevertheless, the question of whether alternative kernel-based embeddings, particularly nonlinear counterparts, could exhibit desirable properties has long remained underexplored, in part due to the associated computational challenges. Recently, this gap has begun to be addressed, with works investigating kernelised medians (Nienkötter and Jiang, 2023), cumulants (Bonnier et al., 2023), and variances (Makigusa, 2024a). In this paper, we consider an alternative based on the concept of quantiles in an RKHS, which we term *kernel quantile embeddings (KQEs)*. Similarly to the construction of KMEs, KQEs are obtained by considering the directional quantiles of a feature map obtained from a reproducing kernel. KQEs also lead naturally to a family of distances which we call *kernel quantile discrepancies (KQDs)*. This approach is motivated from the statistics and econometrics literature (Kosorok, 1999; Dominicy and Veredas, 2013; Ranger et al., 2020; Stolfi et al., 2022), where matching quantiles has been shown effective in constructing statistical estimators and hypothesis tests.

Our paper identifies several desirable properties of KQEs. Firstly, from a theoretical point of view, we show in Theorem 1 and Theorem 2 that KQEs can represent distributions on any space for which we can define a kernel, and that the conditions to make a kernel *quantile-characteristic*, that is for KQEs to be a one-to-one representation of a probability distribution, are weaker than for the classical notion of characteristic, which we now call *mean-characteristic*. We

[1]Department of Computer Science, University College London, London, UK [2]College of Computing & Data Science, Nanyang Technological University, Singapore [3]Department of Statistical Science, University College London, London, UK [4]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany. Correspondence to: Masha Naslidnyk <masha.naslidnyk.21@ucl.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

then show in Theorem 3 that KQEs can be estimated at a rate of $\mathcal{O}(n^{-1/2})$ in the number of samples n; the same rate as that of the empirical estimator of KMEs (Tolstikhin et al., 2017). As a result, KQDs are probability metrics under much weaker conditions than the MMD (see Theorem 4), while maintaining comparable computational guarantees, including a finite-sample consistency with rate $\mathcal{O}(n^{-1/2})$ (up to log terms) for their empirical estimators (see Theorem 5).

Secondly, we establish a number of connections between KQDs, Wasserstein distances (Kantorovich, 1942; Villani et al., 2009), and generalisations or approximations thereof. In particular, special cases of our KQDs recover existing sliced Wasserstein (SW) distances (Bonneel et al., 2015; Wang et al., 2022; 2024a) and can interpolate between the Wasserstein distance and MMD similarly to Sinkhorn divergences (Cuturi, 2013; Genevay et al., 2019). These results are presented in Connections 1, 2, and 3.

Finally, we consider a specific instance of KQDs based on Gaussian averaging over kernelised quantile directions, which we name the *Gaussian expected kernel quantile discrepancy (e-KQD)*. Beyond the desirable theoretical properties described above, we show that the Gaussian e-KQD also has attractive computational properties. In particular, we show that it has a natural estimator which only requires sampling from a Gaussian measure on the RKHS, and which can be computed with complexity $\mathcal{O}(n \log^2(n))$. It is studied empirically in Section 5 with experiments on two-sample hypothesis testing, where we show that it is competitive with the MMD: it often outperforms estimators of the MMD of the same asymptotic complexity, and in some cases even outperforms MMD at higher computational costs.

2. Background

Let $\mathcal{P}_\mathcal{X}$ denote the set of Borel probability measures on a Borel space $\mathcal{X}$. We begin by reviewing existing definitions of quantiles, followed by a summary of relevant work on probability metrics, including the MMD and SW distances.

2.1. Quantiles

Univariate quantiles. Let $\mathcal{X} \subseteq \mathbb{R}$. For $\alpha \in [0, 1]$, the *α-quantile of $P \in \mathcal{P}_\mathcal{X}$* is defined as $\rho_P^\alpha = \inf\{y \in \mathcal{X} : \Pr_{Y\sim P}[Y \leq y] \geq \alpha\}$. When P has a continuous and strictly monotonic cumulative distribution function F_P, quantiles can also be defined through the inverse of that function $\rho_P^\alpha := F_P^{-1}(\alpha)$. Notable special cases include $\alpha = 0.5$, corresponding to the median, and $\alpha = 0.25, 0.75$, corresponding to lower- and upper-quartiles respectively. Importantly, P is fully characterised by its quantiles $\{\rho_P^\alpha\}_{\alpha\in[0,1]}$.

From a computational viewpoint, univariate quantiles can be straightforwardly estimated using order statistics. Suppose $y_{1:n} = [y_1 \dots y_n]^\top \sim P$, and denote by $[y_{1:n}]_j$ the j^{th} order statistic of $y_{1:n}$ (i.e. the j^{th} largest value in the vector $[y_1 \dots y_n]^\top$). The α-quantile of P, denoted ρ_P^α, can be estimated using $[y_{1:n}]_{\lceil \alpha n \rceil}$ where $\lceil \cdot \rceil$ denotes the ceiling function. This estimator is known to converge at a rate of $\mathcal{O}(n^{-1/2})$ (Serfling, 2009, Section 2.3.2).

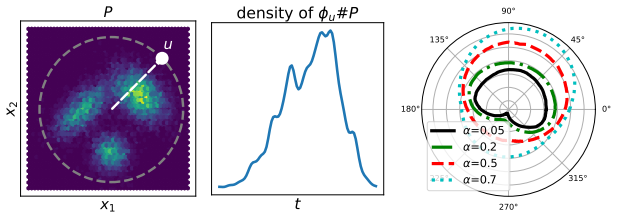

Figure 1: *Illustration of bivariate quantiles. Left:* Bivariate distribution P. *Center:* Density of the projection of P onto direction u on the unit circle, with $\phi_u(x) = \langle u, x\rangle$. *Right:* different quantiles for all possible directions u.

Multivariate quantiles. Suppose now that $\mathcal{X} \subseteq \mathbb{R}^d$ for $d > 1$. The previous definition of quantiles depends on the existence of an ordering in $\mathcal{X}$, and its natural generalisation to $d > 1$ is therefore not unique (Serfling, 2002). In this paper, we will focus on the notion of *α-directional quantile of P along some direction u in the unit sphere S^{d-1}* (Kong and Mizera, 2012),

$$\rho_P^{\alpha,u} := \rho^\alpha_{\phi_u \# P} u, \qquad \phi_u(y) = \langle u, y\rangle.$$

Here, $\phi_u : \mathcal{X} \to \mathbb{R}$ is the projection map onto u, and $\rho^\alpha_{\phi_u\#P}$ is the standard one-dimensional $\alpha-$quantile of $\phi_u\#P$—the law of $\phi_u(X)$ for $X \sim P$. We note that this quantile is now a d-dimensional vector as opposed to a scalar. The α-directional quantiles for $d = 2$ are illustrated in Figure 1, in which the probability measure P is projected onto some line; see the left and middle plots. Once again, we can use quantiles to characterise P, although we must now consider all $\alpha-$quantiles over a sufficiently rich family of projections $\{\rho_P^{\alpha,u} : \alpha \in [0, 1], u \in S^{d-1}\}$; see Theorem 5 of (Kong and Mizera, 2012) for sufficient regularity conditions.

Although these multivariate quantiles satisfy scale equivariance and rotation equivariance, they do not satisfy location equivariance. To remedy this issue, Fraiman and Pateiro-López (2012) introduced a related notion, the *centered* α-directional quantile:

$$\tilde{\rho}_P^{\alpha,u} := \left(\rho^\alpha_{\phi_u\#P} - \phi_u(\mathbb{E}_{X\sim P}[X])\right) u + \mathbb{E}_{X\sim P}[X], \quad (1)$$

Further details are provided in Appendix B.

2.2. Probability Metrics

Kernel mean embeddings and MMD. Let $\mathcal{X}$ be some Borel space, and $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$ be a reproducing kernel Hilbert space (RKHS) induced by a real-valued kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ (Schölkopf and Smola, 2002; Berlinet and Thomas-Agnan, 2004), the *kernel mean embedding* (KME) $\mu_P : \mathcal{X} \to \mathbb{R}$ of any $P \in \mathcal{P}_\mathcal{X}$ is defined as the Bochner integral

$\mu_P(\cdot) = \mathbb{E}_{X\sim P}[k(X,\cdot)] \in \mathcal{H}$. The integral can be shown to exist provided $\mathbb{E}_{X\sim P}[\sqrt{k(X,X)}] < \infty$; this, in turn, holds for all $P \in \mathcal{P}_\mathcal{X}$ if and only if k is bounded (Smola et al., 2007). If the mapping $P \to \mu_P$ is injective, the kernel k is said to be *mean-characteristic*. Many standard kernels—the Matérn family, Gaussian, Laplacian—have been shown to be characteristic on sufficiently regular spaces (Sriperumbudur et al., 2011; Ziegel et al., 2024). KMEs with mean-characteristic kernels lead to the squared *maximum mean discrepancy (MMD)* defined for any $P, Q \in \mathcal{P}_\mathcal{X}$ as

$$\begin{aligned}\mathrm{MMD}^2(P,Q) &:= \|\mu_P - \mu_Q\|^2_\mathcal{H} = \mathbb{E}_{X,X'\sim P}[k(X,X')] \\ &\quad - 2\mathbb{E}_{X\sim P,X'\sim Q}[k(X,X')] + \mathbb{E}_{X,X'\sim Q}[k(X,X')].\end{aligned}$$

This can be computed in rare cases (Briol et al., 2025), but typically needs to be estimated. Given n i.i.d. realisations from P and Q, MMD^2 is most commonly estimated with a U-statistic, which converges to $\mathrm{MMD}^2(P,Q)$ as $\mathcal{O}(n^{-1/2})$ and has computational complexity of $\mathcal{O}(n^2)$—although linear-cost alternatives are also available (Gretton et al., 2012, Lemma 14.). These are discussed in Section 5 and Appendix A.

Wasserstein distances. Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric on $\mathcal{X}$, and $\Gamma(P,Q) \subseteq \mathcal{P}_{\mathcal{X}\times\mathcal{X}}$ denote the space of joint distributions on $\mathcal{X} \times \mathcal{X}$ with first and second marginals P and Q, respectively. The *p-Wasserstein distance* (Kantorovich, 1942; Villani et al., 2009) quantifies the cost of optimally transporting one distribution to another under "cost" $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It is a probability metric under mild conditions (Villani et al., 2009, Section 6), and is defined as

$$W_p(P,Q) = \left(\inf_{\pi\in\Gamma(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}\left[c(X,Y)^p\right]\right)^{1/p}.$$

When $\mathcal{X} \subseteq \mathbb{R}^d$, the metric c is typically taken to be the Euclidean distance $c(x,y) = \|x-y\|_2$. The Wasserstein distance can then be estimated by solving an optimal transport problem using empirical measures constructed through samples of P and Q, an approach that suffers from a high computational cost of $\mathcal{O}(n^3)$ and, when P, Q have at least $2p$ moments, slow convergence of $\mathcal{O}(n^{-1/\max(d,2p)})$ when $\mathcal{X} \subseteq \mathbb{R}^d$ for $d > 1$ (Fournier and Guillin, 2015).

However, when $d = 1$, W_p can be computed at lower cost of $\mathcal{O}(n \log n)$ with convergence of $\mathcal{O}(n^{-1/2p})$ when P, Q have at least $2p$ moments. This motivated the introduction of the *sliced Wasserstein* (SW) distance (Bonneel et al., 2015). Recall that $\phi_u(x) = u^\top x$. The SW distance projects high-dimensional distributions P, Q onto elements on the unit sphere $u \in S^{d-1}$ sampled uniformly, computes the Wasserstein distance between the projected distributions, now in $\mathbb{R}$, and averages over the projections:

$$\mathrm{SW}_p(P,Q) = \left(\mathbb{E}_{u\sim\mathbb{U}(S^{d-1})}\left[W_p^p(\phi_u\#P, \phi_u\#Q)\right]\right)^{1/p}.$$

A further refinement, the *max-sliced Wasserstein (max-SW)* distance (Deshpande et al., 2018), aims to identify the optimal projection that maximises the 1D Wasserstein distance:

$$\text{max-SW}_p(P,Q) = \left(\sup_{u\in S^{d-1}} W_p^p(\phi_u\#P, \phi_u\#Q)\right)^{1/p}.$$

Both slicing distances reduce the computational complexity to $\mathcal{O}(ln \log n)$ and the convergence rate to $\mathcal{O}(l^{-1/2} + n^{-1/2p})$, where l is either the number of projections, or the number of iterations of the optimiser. A further extension is the *Generalised Sliced Wasserstein* (GSW, Kolouri et al. (2019)), which replaces the linear projection ϕ_u with a non-linear mapping. While the conditions for GSW to be a probability metric are highly non-trivial to verify, the authors showed that they hold for polynomials of odd degree.

Another approximation of the Wasserstein distance involves the introduction of an entropic regularisation term (Cuturi, 2013), which reduces the cost to $\mathcal{O}(n^2)$ and can be estimated with sample complexity $\mathcal{O}(n^{-1/2})$ (Genevay et al., 2019). The solution to this regularised problem is referred to as the *Sinkhorn divergence*. Interestingly, Ramdas et al. (2017); Feydy et al. (2019) demonstrated that by varying the strength of the regularisation, the Sinkhorn divergence interpolates between the Wasserstein distance and the MMD with a kernel corresponding to the energy distance.

3. Kernel Quantile Embeddings and Discrepancies

We introduce directional quantiles in the RKHS and the corresponding discrepancies. Unlike in Section 2.1, the measures and their quantiles now live in different spaces: the measures are on $\mathcal{X}$, and the quantiles are in the RKHS $\mathcal{H}$ induced by a kernel on $\mathcal{X}$. This leads to greater flexibility: the approach works for any space a kernel can be defined on. Throughout, we assume the kernel k is measurable.

3.1. Kernel Quantile Embeddings

Let $S_\mathcal{H} = \{u \in \mathcal{H} : \|u\|_\mathcal{H} = 1\}$ be the unit sphere of an RKHS $\mathcal{H}$ induced by the kernel k. For $P \in \mathcal{P}_\mathcal{X}$, we define its *$\alpha$-quantile along RKHS direction* $u \in S_\mathcal{H}$ as a function $\rho_P^{\alpha,u} : \mathcal{X} \to \mathbb{R}$ in $\mathcal{H}$ with

$$\rho_P^{\alpha,u}(x) := \rho^\alpha_{u\#P} u(x) \tag{2}$$

By the reproducing property, it holds that $\rho^\alpha_{u\#P} u(x) = \rho^\alpha_{\phi_u\#[\psi\#P]} u(x)$, where $\psi(x) = k(x,\cdot)$ is the canonical feature map $\mathcal{X} \to \mathcal{H}$, and $\phi_u(h) = \langle u, h\rangle_\mathcal{H}$ is the $\mathcal{H} \to \mathcal{H}$ equivalent of the projection operator onto u defined in Section 2.1. Thus, when $\dim(\mathcal{H}) < \infty$, the RKHS quantiles of P on $\mathcal{X}$ are exactly the multivariate quantiles of the measure of $k(X,\cdot)$, $X \sim P$, on $\mathcal{H}$. In other words, KQEs can be thought of as two-step embeddings: we first embed

$X \sim P \in \mathcal{P}_{\mathcal{X}}$ as an RKHS element and then compute its directional quantiles to obtain the KQEs.

Centered vs uncentered quantiles. Just as done for multivariate quantiles in Equation (1), a centered version of RKHS quantiles can be defined as

$$\tilde{\rho}_P^{\alpha,u}(x) := \left(\rho_{u\#P}^{\alpha} - \langle u, \mu_P \rangle_{\mathcal{H}}\right) u(x) + \mu_P(x),$$

where μ_P is the KME of P. This coincides with Equation (1) for the measure being the law of $k(X,\cdot)$ with $X \sim P$. The impact of centering is examined in detail in Appendix B, but two key observations are relevant here: (1) omitting centering eliminates the computational overhead of calculating means; (2) the only equivariance violated for the uncentered directional quantile is location equivariance: shifting $k(X,\cdot)$ by h shifts the quantile by $\langle h, u\rangle_{\mathcal{H}} u$, rather than by h itself. However, when KQEs are used to compare two distributions, the additional term $\langle h, u\rangle_{\mathcal{H}} u$ cancels out as it does not depend on the measure. For these reasons, we primarily work with the uncentered RKHS quantiles.

Quantile-characteristic kernels. The kernel k is said to be *quantile-characteristic* if the mapping $P \mapsto \{\rho_P^{\alpha,u} : \alpha \in [0,1], u \in S_{\mathcal{H}}\}$ is injective for $P \in \mathcal{P}_{\mathcal{X}}$. In $\mathbb{R}^d$, the Cramér-Wold theorem (Cramér and Wold, 1936) states that the set of all one-dimensional projections (or, equivalently, all quantiles of all one-dimensional projections) determines the measure. One may therefore recognise our next theorem as an RKHS-specific extension of the Cramér-Wold theorem. Earlier Hilbert space extensions required higher-dimensional projections and imposed restrictive moment assumptions (Cuesta-Albertos et al., 2007). Being concerned with the RKHS case specifically allows us to prove the result under mild assumptions, as stated below.

Assumption A1. *$\mathcal{X}$ is Hausdorff, separable, and σ-compact.*

Being Hausdorff ensures points in $\mathcal{X}$ can be separated, and separability says $\mathcal{X}$ has a countable dense subset. σ-compactness means $\mathcal{X}$ is a union of countably many compact sets. These are mild conditions, notably satisfied by Polish spaces—including discrete topological spaces with at most countably many elements and topological manifolds.

It is possible to drop the σ-compactness and separability. When $\mathcal{X}$ is Hausdorff and completely regular, one can still get quantile-characteristic properties on Radon probability measures—the "non-pathological" Borel probability measures. We discuss this in Appendix C.1 and refer to Willard (1970) for a review of general topological properties.

Assumption A2. *The kernel k is continuous, and separating on $\mathcal{X}$: for any $x \neq y \in \mathcal{X}$, it holds that $k(x,\cdot) \neq k(y,\cdot)$.*

This is a mild condition: most commonly used kernels such as the Matérn, Gaussian, and Laplacian kernels are separating. The constant kernel $k(x,x') = c$ is an example of a non-separating kernel. Trivially, a non-separating kernel

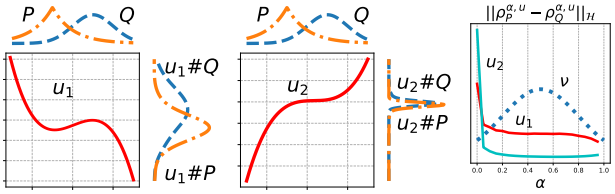

Figure 2: *Illustration of the impact of the slicing direction on KQEs.* Suppose $X \sim P$, the KQEs $\rho_P^{\alpha,u}(x) := \rho_{u\#P}^{\alpha} u(x)$ are obtained by considering the α^{th} quantile of $u(X)$. Clearly, these quantiles might vary significantly depending on the slicing direction used.

for which $k(x,\cdot) = k(y,\cdot)$ will not be able to distinguish between Dirac measures δ_x and δ_y. The proof of the following result uses *characteristic functionals*, an extension of characteristic functions to measures on spaces beyond $\mathbb{R}^d$. Unlike moments, characteristic functionals are defined for any probability measure—which is the key to generality of KQEs. Further discussion and proof are in Appendix C.1.

Theorem 1 (Cramér-Wold Theorem in RKHS). *Under A1 and A2, the kernel k is quantile-characteristic, meaning the mapping $P \mapsto \{\rho_P^{\alpha,u} : \alpha \in [0,1], u \in S_{\mathcal{H}}\}$ is injective.*

The mildness of the assumptions in Theorem 1 naturally raises the question: *is being quantile-characteristic a less restrictive condition than being mean-characteristic?* This indeed holds, as shown in the result below.

Theorem 2. *Every mean-characteristic kernel k is also quantile-characteristic. The converse does not hold.*

This result, proven in Appendix C.2, has a powerful implication. For any discrepancy $D(P,Q)$ that aggregates the KQEs injectively (i.e $D(P,Q) = 0 \iff \rho_P^{\alpha,u} = \rho_Q^{\alpha,u}$ for all α, u), it holds that $\mathrm{MMD}(P,Q) > 0 \Rightarrow D(P,Q) > 0$, but $D(P,Q) > 0 \not\Rightarrow \mathrm{MMD}(P,Q) > 0$. This means D can tell apart every pair of measures MMD can, and sometimes more (see the proof for examples). This is intuitive: MMD is an injective aggregation of means ($\mathrm{MMD}(P,Q) = 0 \iff \mathbb{E}_P[u] = \mathbb{E}_Q[u]$ for all u), and the set of all quantiles captures all the information in the mean, but not vice versa. Before introducing a specific family of quantile discrepancies, we discuss sample versions of KQEs.

Estimating KQEs. For fixed $\alpha \in [0,1]$ and $u \in S_{\mathcal{H}}$, estimating the directional quantile $\rho_P^{\alpha,u}$ with samples $x_{1:n} \sim X$ boils down to estimating the $\mathbb{R}$-quantile $\rho_{u\#P}^{\alpha}$ using samples $u(x_{1:n})$. We employ the classic, model-free approach to estimate a quantile by using the order statistic estimator:

$$\rho_{P_n}^{\alpha,u}(x) := \rho_{u\#P_n}^{\alpha} u(x) = [u(x_{1:n})]_{\lceil \alpha n \rceil} u(x), \qquad (3)$$

where $P_n = {}^1\!/_n \sum_{i=1}^n \delta_{x_i}$. In other words, Equation (3) uses the α-quantile of the set $u(x_{1:n})$—meaning, the $\lceil \alpha n \rceil$-th largest element of $u(x_{1:n})$. We now state an RKHS version of a classic result on convergence of quantile estimators; the proof is provided in Appendix C.3.

Theorem 3 (Finite-Sample Consistency for Empirical KQEs). *Suppose the PDF of $u\#P$ is bounded away from zero, $f_{u\#P}(x) \geq c_u > 0$, and $x_{1:n} \sim P$. Then, with probability at least $1-\delta$, and $C(\delta, u) = \mathcal{O}(\sqrt{\log(2/\delta)})$,*

$$\|\rho^{\alpha,u}_{P_n} - \rho^{\alpha,u}_{P}\|_{\mathcal{H}} \leq C(\delta, u) n^{-1/2}.$$

We do not need to assume A1 and A2 to prove consistency; this was only needed to establish that k is quantile-characteristic, and we may still have a consistent estimator when the kernel is not quantile-characteristic. The condition $f_{u\#P}(x) \geq c_u > 0$ lets us avoid making any assumptions on $\mathcal{X}$, other than the existence of a kernel k on $\mathcal{X}$.

3.2. Kernel Quantile Discrepancies

We propose to quantify the difference between $P, Q \in \mathcal{P}_{\mathcal{X}}$ in unit-norm direction u through a ν-weighted expectation of power-p distance (in the RKHS) between KQEs,

$$\tau_p(P, Q; \nu, u) = \Big(\int_0^1 \big\| \rho^{\alpha,u}_P - \rho^{\alpha,u}_Q \big\|^p_{\mathcal{H}} \nu(\mathrm{d}\alpha) \Big)^{1/p}.$$

Figure 2 illustrates how $u\#P$ and $u\#Q$ vary depending on direction u—and the impact it has on τ_p. The weighting measure ν on $[0,1]$ assigns importance to each α-quantile. For example, the Lebesgue measure $\nu \equiv \mu$ treats all quantiles as equally important, whereas a partial-supported measure would allow us to ignore certain quantiles.

Based on $\tau_p(P, Q; \nu, u)$, we introduce a novel family of *Kernel Quantile Discrepancies (KQDs)* that aggregate the directional differences $\tau_p(P, Q; \nu, u)$ over $u \in S_{\mathcal{H}}$: the L^p-type distance *expected KQD (e-KQD)* that uses the average as the aggregate function, and the L^∞-type distance *supremum KQD (sup-KQD)* that aggregates with the supremum:

$$\begin{aligned} \text{e-KQD}_p(P, Q; \nu, \gamma) &= \left(\mathbb{E}_{u\sim\gamma} \left[\tau_p^p \left(P, Q; \nu, u \right) \right] \right)^{1/p}, \\ \text{sup-KQD}_p(P, Q; \nu) &= \Big(\sup_{u \in S_{\mathcal{H}}} \tau_p^p \left(P, Q; \nu, u \right) \Big)^{1/p}, \end{aligned} \tag{4}$$

where γ is a measure on the unit sphere $S_{\mathcal{H}}$ of the RKHS.

Next, we demonstrate that under mild conditions e-KQD and sup-KQD are indeed distances, and establish connections with existing methods. The proof is in Appendix C.4.

Theorem 4 (KQDs as Probability Metrics). *Under A1, A2, and if ν has full support on $[0,1]$, sup-KQD$_p$ is a distance. Further, if γ has full support on $S_{\mathcal{H}}$, e-KQD$_p$ is a distance.*

As discussed in Section 3, A1 and A2 are minor. The assumptions on the support of ν and γ ensure that no quantile level in $[0,1]$ and no parts of $S_{\mathcal{H}}$ are missed entirely. This is satisfied, for example, for the uniform ν (that considers all quantiles to be equally important), and when $\mathcal{H}$ is separable, for any centered Gaussian $\gamma = \mathcal{N}(0, S)$ with a non-degenerate S by (Kukush, 2020, Corollary 5.3). For example, an $\mathcal{H} \mapsto \mathcal{H}$ covariance operator $S[f](x) = \int_{\mathcal{X}} k(x,y) f(y) \beta(\mathrm{d}y)$ is non-degenerate and well-defined provided (1) β on $\mathcal{X}$ has full support, and (2) $\int_{\mathcal{X}} \sqrt{k(x,x)}\beta(\mathrm{d}x) < \infty$. This choice of γ also happens to be computationally convenient, as discussed in Section 4.

In contrast, while conditions under which MMD is a distance are well-understood for bounded translation-invariant kernels on Euclidean spaces (Sriperumbudur et al., 2011), they are challenging to establish beyond this setting. For instance, it is known that commonly used graph kernels are not characteristic (Kriege et al., 2020).

When ν is chosen as the Lebesgue measure μ, an important connection emerges between e-KQD, sup-KQD, and sliced Wasserstein distances. This connection is formalised in the next result, with a proof provided in Appendix C.6.

Connection 1 (SW). *Suppose P, Q have p-finite moments. Then, e-KQD$_p(P, Q; \nu, \gamma)$ for $\nu \equiv \mu$ corresponds to a kernel expected sliced p-Wasserstein distance, which has not been introduced in the literature. For $\mathcal{X} \subseteq \mathbb{R}^d$, linear $k(x,y) = x^\top y$, and uniform γ, this recovers the expected sliced p-Wasserstein distance (Bonneel et al., 2015).*

Connection 2 (Max-SW). *Suppose P, Q have p-finite moments. Then, sup-KQD$_p(P, Q; \nu)$ for $\nu \equiv \mu$ is the kernel max-sliced p-Wasserstein distance (Wang et al., 2022). For $\mathcal{X} \subseteq \mathbb{R}^d$, linear $k(x,y) = x^\top y$, and uniform γ, it recovers the max-sliced p-Wasserstein (Deshpande et al., 2018).*

For $d = 1$, we recover standard Wasserstein. When k is non-linear but induces a finite-dimensional RKHS, e-KQD is connected to the Generalised Sliced Wasserstein distances of Kolouri et al. (2022)—we explore this in Appendix C.6.

Lastly, we establish a connection to Sinkhorn divergence.

Connection 3 (Sinkhorn). *Sinkhorn divergence (Cuturi, 2013), like e-KQD and sup-KQD, combines the strengths of kernel embeddings and Wasserstein distances. Furthermore, for $p = 2$ and $\nu \equiv \mu$, the centered version of e-KQD and sup-KQD developed in Appendix B can be represented as a sum of MMD and kernelised expected or max-sliced Wasserstein distances, thus positioning these measures as mid-point interpolants between MMD and SW distances.*

It is important to note that the MMD term within the Sinkhorn divergence is restricted to a specific kernel tied to the energy distance—in contrast, e-KQD and sup-KQD offer much greater flexibility in the choice of kernel. Moreover, as will be shown empirically in Section 5, the computational complexity of e-KQD for a particular choice of γ can be made significantly lower than that of Sinkhorn divergences, which have a cost of $\mathcal{O}(n^2)$.

Estimating e-KQD. We propose a Monte-Carlo estimator for e-KQD, and refer to Wang et al. (2022) for an

optimisation-based, $\mathcal{O}(n^3 \log(n))$ estimator for sup-KQD. Let $x_{1:n} \sim P$, $y_{1:n} \sim Q$, the $u_1, \dots, u_l \in S_{\mathcal{H}}$ to be l unit-norm functions sampled from γ, and f_ν to be the density of ν. Denote $P_n = {}^1\!/_n \sum_{i=1}^n \delta_{x_i}, Q_n = {}^1\!/_n \sum_{i=1}^n \delta_{y_i}$. Then, similarly to the order statistic estimator of the quantiles in Equation (3), e-KQD$_p^p(P_n, Q_n; \nu, \gamma_l)$ is the estimator of e-KQD$_p^p(P, Q; \nu, \gamma)$, where

$$\text{e-KQD}_p^p(P_n, Q_n; \nu, \gamma_l) \tag{5}$$
$$= \frac{1}{ln} \sum_{i=1}^{l} \sum_{j=1}^{n} \left(\left[u_i(x_{1:n})\right]_j - \left[u_i(y_{1:n})\right]_j \right)^p f_\nu\left(\lceil j/n \rceil\right)$$

Here, $[u_i(x_{1:n})]_j$ is the j-th order statistics, meaning the j-th smallest element of $u_i(x_{1:n}) = [u_i(x_1), \dots, u_i(x_n)]^\top$. For $p = 1$, we get the following result, proven in Appendix C.5.

Theorem 5 (Finite-Sample Consistency for Empirical KQDs). *Let ν have a density, P, Q be measures on $\mathcal{X}$ s.t. $\mathbb{E}_{X\sim P}\sqrt{k(X,X)} < \infty$ and $\mathbb{E}_{X\sim Q}\sqrt{k(X,X)} < \infty$, and $x_{1:n} \sim P, y_{1:n} \sim Q$. Then, with probability at least $1-\delta$, and $C(\delta) = \mathcal{O}(\sqrt{\log(1/\delta)})$ that depends only on δ, k, ν,*

$$|\textit{e-KQD}_1(P_n, Q_n; \nu, \gamma_l) - \textit{e-KQD}_1(P, Q; \nu, \gamma)| \leq C(\delta)(l^{-1/2} + n^{-1/2}).$$

The rate does not depend on $\dim(\mathcal{X})$—this is a major advantage of projection/slicing-based discrepancies (Nadjahi et al., 2020), which comes at the cost of dependence on the number or projections l. Setting $l = n/\log n$ recovers the MMD rate (up to log-terms), at matching complexity (see Section 4). Here, we do not need e-KQD to be a distance—indeed, we did not assume A1 and A2. The condition of square root integrability of $k(X,X)$ under P, Q is immediately satisfied when k is bounded, and can in fact be further weakened to $\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}\sqrt{k(X,X) - 2k(X,Y) + k(Y,Y)} < \infty$. Requiring that ν has a density is mild and necessary to reduce the problem to CDF convergence—which, by the classic Dvoretzky-Kiefer-Wolfowitz inequality of Dvoretzky et al. (1956) has rate $n^{-1/2}$ under no assumptions on the underlying distributions. The strength of this inequality allows us to assume nothing more of $\mathcal{X}$ than the fact that it is possible to define a kernel on it.

Further, for any integer $p > 1$, the $n^{-1/2}$ rate still holds—if and only if it holds that for $J_p(R) := (F_{u\#R}(t)(1 - F_{u\#R}(t)))^{p/2} / f_{u\#R}^{p-1}(t)$, both $J_p(P)$ and $J_p(Q)$ are integrable over $u \sim \gamma$ and Lebesgue measure on $u(\mathcal{X})$. In turn, this may be reduced to a problem of controlling $d-1$ volumes of level sets of u. We discuss this extension further in Conjecture 1 in Appendix C.5.

4. Gaussian Kernel Quantile Discrepancy

Algorithm 1 Gaussian e-KQD

Input: Data $x_{1:n} \sim P, y_{1:n} \sim Q$, samples from the reference measure $z_{1:m} \sim \xi$, kernel k, density f_ν, number of projections l, power p.
Initialise e-KQD$^p \leftarrow 0$ and $\tau_{p,i}^p \leftarrow 0$ for $i = 1 \dots l$.
for $i = 1$ **to** l **do**
  Sample $\lambda_{1:m} \sim \mathcal{N}(0, \mathrm{Id}_m)$
  Compute $f_i(x_{1:n}) \leftarrow \lambda_{1:m}^\top k(z_{1:m}, x_{1:n})/\sqrt{m}$,
    $f_i(y_{1:n}) \leftarrow \lambda_{1:m}^\top k(z_{1:m}, y_{1:n})/\sqrt{m}$
  Compute $\|f_i\|_{\mathcal{H}} \leftarrow \sqrt{\lambda_{1:m}^\top k(z_{1:m}, z_{1:m})\lambda_{1:m}/m}$
  Compute $u_i(x_{1:n}) \leftarrow f_i(x_{1:n})/\|f_i\|_{\mathcal{H}}$,
    $u_i(y_{1:n}) \leftarrow f_i(y_{1:n})/\|f_i\|_{\mathcal{H}}$
  Sort $u_i(x_{1:n})$ and $u_i(y_{1:n})$
  for $j = 1$ **to** n **do**
    $\tau_{p,i}^p \leftarrow \tau_{p,i}^p + \left([u_i(x_{1:n})]_j - [u_i(y_{1:n})]_j\right)^p f_\nu(\lceil j/n \rceil)$
  end for
  e-KQD$^p \leftarrow$ e-KQD$^p + \tau_{p,i}^p/l$
end for
Return e-KQD^p

We now conduct further empirical study of the squared kernel distance e-KQD$_p$. Unlike its supremum-based counterpart sup-KQD, e-KQD can be approximated simply by drawing samples from γ on $S_{\mathcal{H}}$, avoiding the challenges associated with optimising for the supremum. Although a uniform γ is a natural choice, no such measure exists when $\dim(\mathcal{H})$ is infinite (Kukush, 2020, Section 1.3). Instead, we follow a well-established strategy from the inverse problems literature (Stuart, 2010) and take γ to be the projection onto $S_{\mathcal{H}}$ of a Gaussian measure on $\mathcal{H}$. Using established techniques for sampling Gaussian measures, we then build an efficient estimator for e-KQD$_p(P, Q; \nu, \gamma)$. Gaussian measures on Hilbert spaces are a natural extension of the familiar Gaussian measures on $\mathbb{R}^d$: a measure $\mathcal{N}(0, C)$ on $\mathcal{H}$ is said to be a *centered Gaussian measure* with covariance operator $C : \mathcal{H} \to \mathcal{H}$ if, for every $f \in \mathcal{H}$, the pushforward of $\mathcal{N}(0, C)$ under the $\mathcal{H} \to \mathbb{R}$ projection map $\phi_f(\cdot) = \langle f, \cdot \rangle_{\mathcal{H}}$ is the Gaussian measure $\mathcal{N}(0, \langle C[f], f\rangle_{\mathcal{H}})$ on $\mathbb{R}$. For further details on Gaussian measures in Hilbert spaces, we refer to Kukush (2020).

Let γ' be a centered Gaussian measure on $\mathcal{H}$ whose covariance function $C : \mathcal{H} \to \mathcal{H}$ is an integral operator with some reference measure ξ on $\mathcal{X}$,

$$\gamma' = \mathcal{N}(0, C), \quad C[f](x) = \int_{\mathcal{X}} k(x,y) f(y) \xi(\mathrm{d}y),$$

and let γ be the pushforward of γ' by the projection $\mathcal{H} \to S_{\mathcal{H}}$ that maps any $f \in \mathcal{H}$ to $f/\|f\|_{\mathcal{H}} \in S_{\mathcal{H}}$. By the change of variables formula for pushforward measures (Bogachev,

2007, Theorem 3.6.1), it holds that

$$\begin{aligned}\text{e-KQD}_p^p(P,Q;\nu,\gamma) &= \mathbb{E}_{u\sim\gamma}\left[\tau_p^p\left(P,Q;\nu,u\right)\right] \\ &= \mathbb{E}_{f\sim\gamma'}\left[\tau_p^p\left(P,Q;\nu,f/\|f\|_{\mathcal{H}}\right)\right].\end{aligned}$$

This equality reduces sampling from γ to sampling from a centered Gaussian measure with an integral operator covariance function. The next proposition reduces sampling from (a finite-sample approximation of) γ to sampling from the standard Gaussian on the real line; proof is in Appendix C.7.

Proposition 1 (Sampling from a Gaussian measure). *Let $z_{1:m} \sim \xi$, and γ'_m to be the estimate of γ' based on the Monte Carlo estimate C_m of the covariance operator C,*

$$\gamma'_m = \mathcal{N}(0, C_m), \quad C_m[g](x) = \frac{1}{m}\sum_{j=1}^{m} k(x, z_j)g(z_j).$$

Suppose $f(x) = m^{-1/2}\sum_{j=1}^{m}\lambda_j k(x, z_j)$ with $\lambda_{1:m} \sim \mathcal{N}(0,1)$. Then, $f \sim \gamma'_m$.

Algorithm 1 brings together the e-KQD estimator in Equation (5), and the procedure for sampling from the Gaussian measure in Proposition 1. The ν choice is left up to the user; the uniform ν remains a default choice. We proceed to analyse the cost. This estimator has complexity $\mathcal{O}(l\max(nm, m^2, n\log n))$: $\mathcal{O}(l)$ for iterating over directions $i \in \{1,\dots,l\}$; $\mathcal{O}(nm)$ for computing $f_i(x_{1:n})$ and $f_i(y_{1:n})$; $\mathcal{O}(m^2)$ for computing $\|f_i\|_{\mathcal{H}}$; and $\mathcal{O}(n\log n)$ for sorting $u_i(x_{1:n})$ and $u_i(y_{1:n})$. For $l := \log n$ and $m := \log n$, the complexity therefore reduces to $\mathcal{O}(n\log^2 n)$; i.e. near-linear (up to log-terms).

5. Experiments

We empirically demonstrate the effectiveness of KQDs for nonparametric two-sample hypothesis testing which aims at determining whether two arbitrary probability distributions, P and Q, differ statistically based on their respective i.i.d. samples. Two-sample testing is widely adopted in scientific discovery fields, such as model verification (Gao et al., 2024), out-of-domain detection (Magesh et al., 2023), and comparing epistemic uncertainties (Chau et al., 2025). Specifically, we test the null hypothesis $H_0 : P = Q$ against the alternative $H_1 : P \neq Q$. In such tests, (estimators of) probability metrics are commonly used as test statistics, including the Kolmogorov-Smirnov distances (Kolmogorov, 1960), Wasserstein distance (Wang et al., 2022), energy-distances (Székely and Rizzo, 2005; Sejdinovic et al., 2013), and most relevant to our work, the MMD (Gretton et al., 2006; 2009; 2012). For an excellent overview of kernel-based two sample testing, we refer readers to Schrab (2025).

Experiments are repeated to calculate the rejection rate, which is the proportion of tests where the null hypothesis is rejected. A high rejection rate indicates better performance at distinguishing between distributions. It is equally important to ensure proper control of Type I error, defined as the rejection rate when the null hypothesis H_0 is true. Specifically, the Type I error rate should not exceed the specified level. Without controlling for Type I error, an inflated rejection rate might not reflect the estimator's ability to detect genuine differences but instead indicate the test rejects more often than it should. We consider a significance level α of 0.05 throughout and report on Type I control in Appendix D.

To determine the rejection threshold for each test statistic, we employ a permutation-based approach: for each trial we pool the two samples, randomly reassign labels 300 times to simulate draws under H_0, compute the test statistic on each permuted split, and take the 95th percentile of this empirical null distribution as our threshold. This fully nonparametric thresholding ensures Type I error control without additional distributional assumptions (Lehmann et al., 1986).

Our experiments aim to demonstrate that, within a comparable computational budget, statistics computed using quantile-characteristic kernels can deliver results competitive with those of MMD tests based on mean-characteristic kernels. Additionally, we seek to explore the inherent trade-offs of the proposed methods. We focus on the nonparametric two-sample testing problem, as it represents one of the most successful applications of the mean-embedding-based MMD and its variants. The code is available at https://github.com/MashaNaslidnyk/kqe.

5.1. Benchmarking

We consider the following distances as test statistics in our experiments. Detail descriptions of these estimators are provided in Appendix A. For KQDs, we take the reference measure ξ (c.f. Proposition 1) to be $1/2P_n + 1/2Q_n$, where P_n corresponds to the empirical distribution $1/n\sum_{i=1}^{n}\delta_{x_i}$, analogously for Q_n. Such ξ is a general choice that is appropriate in the absence of additional information about the space $\mathcal{X}$. We take power $p = 2$ for all KQD-based discrepancies in our experiments; identical experiments for $p = 1$ lead to the same conclusions and are presented for completeness in Appendix D.2. Other than in the second experiment, we use the RBF kernel $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ with σ the bandwidth chosen using the median heuristic method, i.e. $\sigma = \text{Median}(\{\|x_i - x_j\|_2^2,\ \forall i, j \in 1,\dots,n\})$ (Gretton et al., 2012). Due to space constraints, we present all methods on the same plot, regardless of their computational complexity. However, it is important to note that directly comparing test power across methods with varying sampling complexities may be unfair and misleading.

- **e-KQD (ours).** For e-KQD, we set the number of projections to $l = \log n$ and the number of samples drawn from the Gaussian reference to $m = \log n$. Consequently, the overall computational complexity is $\mathcal{O}(n\log^2(n))$.

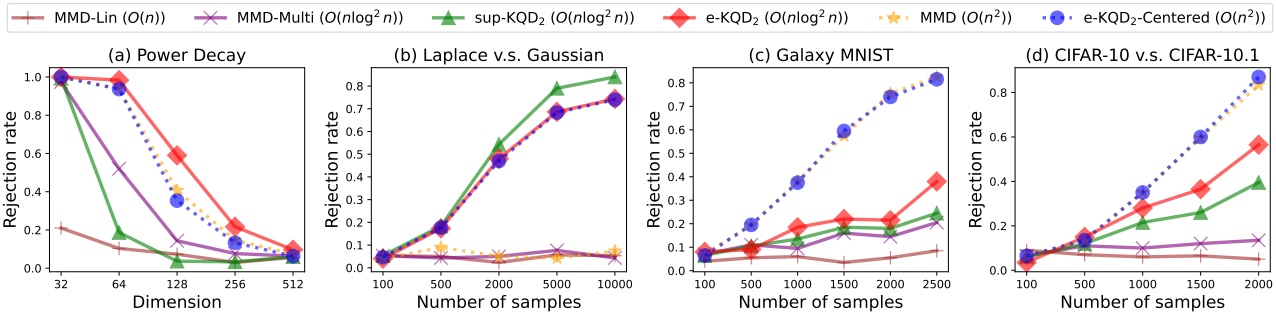

Figure 3: Experimental results comparing our proposed methods with baseline approaches. Methods represented by dotted lines exhibit quadratic complexity for a **single** computation of the test statistic, while the remaining methods achieve near-linear or linear computational efficiency. A higher rejection rate indicates better performance in distinguishing between distributions. **Overall, quadratic-time quantile-based estimators perform comparably to quadratic-time MMD estimators, while near-linear time quantile-based estimators often outperform their MMD-based counterparts.**

- **e-KQD-centered (ours).** The centered version of e-KQD, as discussed in Appendix B, can be expressed as the sum of an e-KQD term and the classical MMD. While the e-KQD component follows the same sampling configuration as above, the MMD computation is the dominant factor in complexity, leading to an overall cost of $\mathcal{O}(n^2)$.
- **sup-KQD (ours).** sup-KQD adopts the same sampling configuration as e-KQD (thus cost $\mathcal{O}(n\log^2(n))$). Instead of averaging over projections, it selects the maximum across all projections. This approach serves as a fast approximation of the kernel max-sliced Wasserstein distance of Wang et al. (2022), where a Riemannian block coordinate descent method is used to optimise an entropic regularised objective at a computational cost of $\mathcal{O}(n^3\log(n))$. In contrast, our approach identifies the largest directional quantile difference across the sampled projections. While we do not claim that this provides an accurate estimate of the true distance, this approach allows for controlled complexity and facilitates comparisons between averaging or taking the supremum.
- **MMD.** The MMD is included as a benchmark to be compared with e-KQD-centered and has complexity $\mathcal{O}(n^2)$. The MMD is estimated using the U-statistic formulation.
- **MMD-Multi.** A fast MMD approximation based on incomplete U-statistic introduced in Schrab et al. (2022) is included to benchmark against our e-KQD distance. Configurations of MMD-Multi are chosen as to match the complexity of e-KQD for a fair comparison.
- **MMD-Lin.** MMD-Linear from Gretton et al. (2012, Lemma 14.) estimates the MMD with complexity $\mathcal{O}(n)$.

5.2. Experimental Setup and Results

We conduct four experiments: two using synthetic data, allowing full control over the simulation environment, and two based on high-dimensional image data to showcase the practicality and competitiveness of our proposed methods. Additional experiments are reported in Appendix D, specifically: studying the impact of changing the measures ν and ξ, comparing with sliced Wasserstein distances, and comparing with MMD based on other KME approximations.

1. Power-decay experiment. This experiment investigates the effect of the curse of dimensionality on our tests, following the setup of Experiment A in Wang et al. (2022). Prior work by Ramdas et al. (2015) has shown that MMD-based methods are particularly vulnerable to the curse of dimensionality. Here, we assess whether our quantile-based test statistic exhibits similar limitations.

We fix $n = 200$ and take P to be an isotropic Gaussian distribution of dimension d. Similarly, we take Q to be a d-dimensional Gaussian distribution with a diagonal covariance matrix $\Sigma = \mathrm{diag}(\{4, 4, 4, 1, \ldots, 1\})$. As we increase the dimension $d \in [32, 64, 128, 256, 512]$, the testing problem becomes increasingly challenging. Figure 3a presents the results. We observe that e-KQD exhibits the slowest decline in test power among all methods, irrespective of their computational complexity. Notably, it maintains its performance significantly better than its $\mathcal{O}(n\log^2(n))$ benchmark, MMD-Multi. These results suggest that quantile-based discrepancies exhibit greater robustness to high-dim data.

2. Laplace v.s. Gaussian. This experiment aims to illustrate Theorem 2 by demonstrating that while a kernel may not be mean-characteristic—meaning it cannot distinguish between two distributions using standard KMEs and MMDs—it can still be quantile-characteristic. In such cases, the distributions can still be effectively distinguished using our KQEs and KQDs. To demonstrate this, we take P to be a standard Gaussian in $d = 1$, and Q to be a Laplace distribution with matching first and second moment. We vary $n \in \{100, 500, 2000, 5000, 10000\}$ and select a polynomial kernel of degree 3, i.e. $k(x, x') = (\langle x, x' \rangle + 1)^3$, for all our

methods. This ensures that k cannot distinguish between the two distributions due to their matching first and second moments, which leads to their KMEs being identical.

Figure 3b shows that our KQDs, irrespective of their computational complexity, exhibit increasing test power as the sample size grows. In contrast, MMD-based methods fail entirely to detect any differences between P and Q. Notably, although e-KQD-centered can be expressed as the sum of an MMD term and an a e-KQD term, the underperformance of the MMD component in this scenario is effectively compensated by the e-KQD term, enabling successful testing.

3. Galaxy MNIST. We examine performance on real-world data through galaxy images (Walmsley et al., 2022) in dimension $d = 3 \times 64 \times 64 = 12288$, following the setting from Biggs et al. (2024). These images consist of four classes. P corresponds to images sampled uniformly from the first three classes, while Q consists of samples from the same classes with probability 0.85 and from the fourth class with probability 0.15. A Gaussian RBF kernel with bandwidth chosen using the median heuristic method is chosen for all estimators. Sample sizes are chosen from $n \in \{100, 500, 1000, 1500, 2000, 2500\}$.

Figure 3c presents the results. e-KQD-centered and MMD exhibit nearly identical performance, suggesting that the MMD term is dominating in the e-KQD-centered estimator. Among the near-linear time test statistics, e-KQD and sup-KQD show a slight advantage over MMD-Multi in distinguishing between the distributions of Galaxy images.

4. CIFAR-10 v.s. CIFAR-10.1. We conclude with an experiment on telling apart the CIFAR-10 (Krizhevsky et al., 2012) and CIFAR-10.1 (Recht et al., 2019) test sets, following again Liu et al. (2020) and Biggs et al. (2024). The dimension is $d = 3 \times 32 \times 32 = 3072$. This is a challenging task, as CIFAR-10.1 was designed to provide new samples from the CIFAR-10 distribution, making it an alternative test set for models trained on CIFAR-10. We conduct the test by drawing n samples from CIFAR-10, and n samples from CIFAR-10.1, with $n \in \{100, 500, 1000, 1500, 2000\}$.

Figure 3d presents the results. Consistent with previous observations, test statistics with quadratic computational complexity exhibit nearly identical performance. However, our quantile discrepancy estimators with near-linear complexity significantly outperform the fast MMD estimators (MMD-Multi) of the same complexity, highlighting the practical advantages of our methods in real-world testing scenarios where computational efficiency is a critical consideration.

An empirical runtime comparison of all methods is presented in Figure 4, which shows the time (in seconds) required to complete this experiment. The empirical results align with our complexity analysis: the near-linear estimators exhibit comparable performance, while the quadratic estimators are significantly slower. The proposed near-linear KQD estimator makes it suitable for larger-scale datasets.

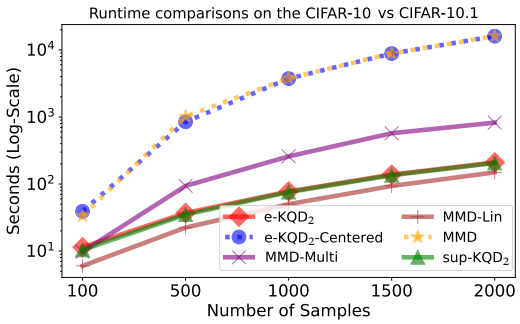

Figure 4: Comparing the time (in seconds) required to complete the CIFAR-10 vs. CIFAR-10.1 experiment, plotted on a logarithmic scale. A shorter time indicates a faster algorithm. These results align with our complexity analysis.

6. Discussion and Future Work

This work explores representations of distributions in a RKHS beyond the mean, using functional quantiles to capture richer distributional characteristics. We introduce kernel quantile embeddings (KQEs) and their associated kernel quantile discrepancies (KQDs), and establish that the conditions required for KQD to define a distance are strictly more general than those needed for MMD to be a distance. Additionally, we propose an efficient estimator for the expected KQD based on Gaussian measures, and demonstrate its effectiveness compared to MMD and its fast approximations through extensive experiments in two-sample testing. Our findings demonstrates the potential of KQEs as a powerful alternative to traditional mean-based representations.

Several promising avenues remain. Firstly, future work could explore more sophisticated methods for improving the empirical estimates of KQEs. The study of optimal kernel selection to maximize test power when using KQD for hypothesis testing, analogous to existing work on MMDs (Jitkrittum et al., 2020; Liu et al., 2020; Schrab et al., 2023) could also be explored. Secondly, considering the demonstrated potential of functional quantiles for representing marginal distributions, it is natural to ask whether they could provide a powerful alternative to conditional mean embeddings (CMEs) (Song et al., 2009; Park and Muandet, 2020), the Hilbert space representation of conditional distributions. These complementary developments will unlock new avenues for enhancing existing applications of KMEs across a wide range of domains, including not only nonparametric two-sample testing, but also (conditional) independence testing, causal inference, reinforcement learning, learning on distributions, generative modeling, robust parameter estimation, and Bayesian representations of distributions via kernel mean embeddings, as explored in Flaxman et al. (2016); Chau et al. (2021a;b), among others.

Acknowledgements

The authors are grateful to Carlo Ciliberto and Antonin Schrab for fruitful discussions on Gaussian measures and MMD two-sample testing respectively. MN acknowledges support from the U.K. Research and Innovation under grant number EP/S021566/1, and from the Helmholtz Information & Data Science Academy (HIDA) for providing financial support enabling a short-term research stay at CISPA (Application No. 14773).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

F. Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2022.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York, 2004.

A. Bharti, M. Naslidnyk, O. Key, S. Kaski, and F.-X. Briol. Optimally-weighted estimators of the maximum mean discrepancy for likelihood-free inference. In *International Conference on Machine Learning*, pages 2289–2312, 2023.

F. Biggs, A. Schrab, and A. Gretton. MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36, 2024.

S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

D. A. Bodenham and Y. Kawahara. euMMD: efficiently computing the MMD two-sample test statistic for univariate data. *Statistics and Computing*, 33(5):1–14, 2023.

V. I. Bogachev. *Measure theory*, volume 1. Springer, 2007.

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

P. Bonnier, H. Oberhauser, and Z. Szabó. Kernelized cumulants: Beyond kernel mean embeddings. *Advances in Neural Information Processing Systems*, 36, 2023.

K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019.

F.-X. Briol, A. Gessner, T. Karvonen, and M. Mahsereci. A dictionary of closed-form kernel mean embeddings. *arXiv:2504.18830*, 2025.

A. Chatalic, N. Schreuder, L. Rosasco, and A. Rudi. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, pages 3006–3024. PMLR, 2022.

S. L. Chau, S. Bouabid, and D. Sejdinovic. Deconditional downscaling with gaussian processes. *Advances in Neural Information Processing Systems*, 34:17813–17825, 2021a.

S. L. Chau, J.-F. Ton, J. González, Y. Teh, and D. Sejdinovic. Bayesimp: Uncertainty quantification for causal data fusion. *Advances in Neural Information Processing Systems*, 34:3466–3477, 2021b.

S. L. Chau, R. Hu, J. Gonzalez, and D. Sejdinovic. Rkhs-shap: Shapley values for kernel methods. *Advances in neural information processing systems*, 35:13050–13063, 2022.

S. L. Chau, K. Muandet, and D. Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36:50769–50795, 2023.

S. L. Chau, A. Schrab, A. Gretton, D. Sejdinovic, and K. Muandet. Credal two-sample tests of epistemic uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2025.

B.-E. Chérief-Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference (AABI)*, pages 1–21, 2020.

K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pages 1981–1989, 2015a.

K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015b.

H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, 1(4):290–294, 1936.

J. A. Cuesta-Albertos, R. Fraiman, and T. Ransford. A sharp form of the cramér–wold theorem. *Journal of Theoretical Probability*, 20(2):201–209, 2007.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

I. Deshpande, Z. Zhang, and A. Schwing. Generative modeling using the sliced Wasserstein distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

Y. Dominicy and D. Veredas. The method of simulated quantiles. *Journal of Econometrics*, 172(2):235–247, 2013.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, number PMLR 89, pages 2681–2690, 2019.

S. Flaxman, D. Sejdinovic, J. P. Cunningham, and S. Filippi. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Association for Computing Machinery, 2016.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

R. Fraiman and B. Pateiro-López. Quantiles for finite and infinite dimensional data. *Journal of Multivariate Analysis*, 108:1–14, 2012.

I. Gao, P. Liang, and C. Guestrin. Model equality testing: Which model is this api serving? *arXiv preprint arXiv:2410.20247*, 2024.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19: 513–520, 2006.

A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in Neural Information Processing Systems*, 22, 2009.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

M. W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1976.

P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(4):73–101, 1964.

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR, 2020.

L. V. Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR*, 37(7-8):227–229, 1942. ISSN 10723374.

A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Pub Co, 2 edition, 1960.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019.

S. Kolouri, K. Nadjahi, S. Shahrampour, and U. Şimşekli. Generalized Sliced Probability Metrics. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4513–4517, May 2022. doi: 10.1109/ICASSP43922.2022.9746016. URL https://ieeexplore.ieee.org/document/9746016. ISSN: 2379-190X.

L. Kong and I. Mizera. Quantile tomography: Using quantiles with multivariate data. *Statistica Sinica*, 22(4):1589–1610, 2012.

M. R. Kosorok. Two-sample quantile tests under general conditions. *Biometrika*, 86(4):909–921, 1999.

N. M. Kriege, F. D. Johansson, and C. Morris. A survey on graph kernels. *Applied Network Science*, 5:1–42, 2020.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

A. Kukush. *Gaussian measures in Hilbert space: construction and properties*. John Wiley & Sons, 2020.

E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

M. Lerasle, Z. Szabó, T. Mathieu, and G. Lecué. Monk outlier-robust mean embedding estimation by median-of-means. In *International conference on machine learning*, pages 3782–3793. PMLR, 2019.

Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

A. Magesh, V. V. Veeravalli, A. Roy, and S. Jha. Principled out-of-distribution detection via multiple testing. *Journal of Machine Learning Research*, 24(378):1–35, 2023.

N. Makigusa. Two-sample test based on maximum variance discrepancy. *Communications in Statistics - Theory and Methods*, 53(15):5421–5438, 2024a.

N. Makigusa. Two-sample test based on maximum variance discrepancy. *Communications in Statistics-Theory and Methods*, 53(15):5421–5438, 2024b.

S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4), Nov. 2015. ISSN 1350-7265. doi: 10.3150/14-BEJ645. URL http://arxiv.org/abs/1308.1334. arXiv:1308.1334 [math, stat].

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25*, pages 10–18. 2012.

K. Muandet, K. Fukumizu, B. K. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2016.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

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

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli. Statistical and topological properties of sliced probability divergences. In *Neural Information Processing Systems*, 2020.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli. Statistical and Topological Properties of Sliced Probability Divergences, Jan. 2022. URL http://arxiv.org/abs/2003.05783. arXiv:2003.05783 [cs, stat].

A. Nienkötter and X. Jiang. Kernel-based generalized median computation for consensus learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):5872–5888, 2022.

A. Nienkötter and X. Jiang. Kernel-based generalized median computation for consensus learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):5872–5888, 2023.

Z. Niu, J. Meier, and F.-X. Briol. Discrepancy-based inference for intractable generative models using quasi-Monte Carlo. *Electronic Journal of Statistics*, 17(1):1411–1456, 2023.

J. Park and K. Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020.

G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

A. Ramdas, N. García Trillos, and M. Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

J. Ranger, J.-T. Kuhn, and C. Szardenings. Minimum distance estimation of multidimensional diffusion-based item response theory models. *Multivariate Behavioral Research*, 55(6):941–957, 2020.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2002.

A. Schrab. A unified view of optimal kernel hypothesis testing. *arXiv preprint arXiv:2503.07084*, 2025.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems*, pages 18793–18807, 2022.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

D. Sejdinovic. An overview of causal inference using kernel embeddings. *arXiv:2410.22754*, 2024.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

R. Serfling. Quantile functions for multivariate analysis: Approaches and applications. *Statistica Neerlandica*, 56 (2):214–232, 2002.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

S. J. Sheather and J. S. Marron. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410): 410–416, 1990.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, June 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

P. Stolfi, M. Bernardi, and L. Petrella. Sparse simulation-based estimator built on quantiles. *Econometrics and Statistics*, 2022.

A. M. Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

I. Tolstikhin, B. K. Sriperumbudur, K. Mu, et al. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

N. Vakhania, V. Tarieladze, and S. Chobanyan. *Probability distributions on Banach spaces*, volume 14. Springer Science & Business Media, 1987.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

M. Walmsley, C. Lintott, T. Géron, S. Kruk, C. Krawczyk, K. W. Willett, S. Bamford, L. S. Kelvin, L. Fortson, Y. Gal, et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022.

J. Wang, R. Gao, and Y. Xie. Two-Sample Test with Kernel Projected Wasserstein Distance. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 8022–8055. PMLR, 2022.

J. Wang, M. Boedihardjo, and Y. Xie. Statistical and computational guarantees of kernel max-sliced Wasserstein distances. *arXiv:2405.15441*, 2024a.

J. Wang, M. Boedihardjo, and Y. Xie. Statistical and computational guarantees of kernel max-sliced wasserstein distances. *arXiv preprint arXiv:2405.15441*, 2024b.

S. Willard. *General Topology*. Addison-Wesley Series in Mathematics. Addison Wesley Longman Publishing, New York, NY, Jan. 1970.

J. Ziegel, D. Ginsbourger, and L. Dümbgen. Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures. *Bernoulli*, 30(2):1441–1457, 2024.

Supplementary Material

This supplementary material is structured as follows. In Appendix A, we recall existing probability metrics, define alternative KQDs, and then describe their respective finite-sample estimators. In Appendix C, we provide the proofs of all theoretical results in the paper. In Appendix D, we provide additional numerical experiments to complement the main text.

A. Probability Metrics and Their Estimators

A.1. Maximum Mean Discrepancy

We first recall that the MMD is an integral probability metric (Müller, 1997) where the supremum can be obtained in closed-form:

$$\mathrm{MMD}(P,Q) \coloneqq \sup_{\|f\|_{\mathcal{H}}\leq 1} |\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(X)]| = \|\mu_P - \mu_Q\|_{\mathcal{H}}$$

Using the reproducing property, we can then expressed the squared-MMD as

$$\mathrm{MMD}^2(P,Q) \coloneqq \mathbb{E}_{X,X'\sim P}[k(X,X')] - 2\mathbb{E}_{X\sim P,Y\sim Q}[k(X,Y)] + \mathbb{E}_{Y,Y'\sim Q}[k(Y,Y')] \tag{6}$$

This section describes the most widely used estimators for this quantity based on i.i.d. samples $x_{1:n} \sim P$ and $y_{1:n} \sim Q$. The first is a biased V-statistic estimator with computational complexity $\mathcal{O}(n^2)$ and convergence rate $\mathcal{O}(n^{-1/2})$:

$$\mathrm{MMD}^2_V(P,Q) \coloneqq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j) - \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,y_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(y_i,y_j)$$

Alternatively, an unbiased estimator can be constructed through a U-statistic. Such an estimator also has computational complexity $\mathcal{O}(n^2)$ and convergence rate $\mathcal{O}(n^{-1/2})$ (Gretton et al., 2012, Lemma 6;Corollary 16), and is given by

$$\mathrm{MMD}^2_U(P,Q) \coloneqq \frac{1}{n(n-1)}\sum_{i\neq j} k(x_i,x_j) - \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,y_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(y_i,y_j)$$

A cheaper estimator was proposed in Lemma 14 of (Gretton et al., 2012). This estimator has computational complexity $\mathcal{O}(n)$ and convergence rate $\mathcal{O}(n^{-1/2})$, and we will refer to it as *MMD-Lin*. The estimator is given by:

$$\mathrm{MMD}^2_{\mathrm{lin}}(P,Q) \coloneqq \frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{n/2} k(x_{2i-1},x_{2i}) + k(y_{2i-1},y_{2i}) - k(x_{2i-1},y_{2i}) - k(x_{2i},y_{2i-1})$$

Finally, we also studied estimators whose computational complexity are between that of MMD-lin and the U- or V-statistics estimators. These estimators are due to Schrab et al. (2022) and we refer to them as *MMD-Multi*, and takes the following form

$$\mathrm{MMD}^2_{\mathrm{Multi}}(P,Q) \coloneqq \frac{2}{r(2n-r-1)}\sum_{j=1}^{r}\sum_{i=1}^{n-j} k(x_i,x_{i+j}) + k(y_i,y_{i+j}) - k(x_i,y_{i+j}) - k(x_{i+j},y_i)$$

where r is the number of subdiagonal considered. In our experiments, to match the complexity with e-KQD, we set $r = \log^2(n)$. They have computational complexity $\mathcal{O}(rn)$.

Note that several estimators with faster convergence rates exist (Niu et al., 2023; Bharti et al., 2023), but these have computational cost ranging from $\mathcal{O}(n^2)$ to $\mathcal{O}(n^3)$ and require more regularity conditions on k, P and Q, and we therefore omit them from our benchmark. Bodenham and Kawahara (2023) also introduced an estimator with computational complexity of $\mathcal{O}(n\log(n))$ (and convergence $\mathcal{O}(n^{-1/2})$) using slices/projections to $d = 1$. However, their approach is restrictive in that it can only be used for the Laplace kernel, and we therefore also do not compare to it.

A.2. Wasserstein Distance

The p-Wasserstein distance (Kantorovich, 1942; Villani et al., 2009) is defined as

$$W_p(P,Q) \coloneqq \left(\inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi} \left[\rho(X,Y)^p \right] \right)^{1/p}.$$

Given samples $x_{1:n} \sim P$ and $y_{1:n} \sim Q$, this distance can be approximated using a plug-in $W_p(1/n \sum_{i=1}^n \delta_{x_i}, 1/n \sum_{i=1}^n \delta_{y_i})$, which can be computed in closed-form at a cost of $\mathcal{O}(n^3)$, but converges to $W_p(P,Q)$ with a convergence rate $\mathcal{O}(n^{-1/d})$.

When $\mathcal{X} \subseteq \mathbb{R}^d$ and $p = 1$, we obtain the 1-Wasserstein distance which, similarly to the MMD, can be written as an integral probability metric (Müller, 1997):

$$W_1(P,Q) \coloneqq \sup_{\|f\|_{\text{Lip}} \leq 1} |\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)]|$$

where $\|f\|_{\text{Lip}} = \sup_{x,y\in\mathcal{X}, x\neq y} |f(x) - f(y)|/\|x - y\|$ denotes the Lipschitz norm.

When P, Q are distributions on a one dimensional space $\mathcal{X} \subseteq \mathbb{R}$ that have p-finite moments, the p-Wasserstein distance can be expressed in terms of distance between quantiles of P and Q (see for instance Peyré et al. (2019, Remark 2.30))

$$W_p(P,Q) = \left(\int_0^1 |\rho_P^\alpha - \rho_Q^\alpha|^p \mathrm{d}\alpha \right)^{1/p} \tag{7}$$

A natural estimator for the Wasserstein distance is therefore based on approximating these one-dimensional quantiles using order statistics. Given $x_{1:n} \sim P$ and $y_{1:n} \sim Q$, denote by $P_n = 1/n \sum_{i=1}^n \delta_{x_i}$ and $Q_n = 1/n \sum_{i=1}^n \delta_{y_i}$ the corresponding empirical approximations to P and Q. Then, the $\lceil j/n \rceil$-th quantiles of P_n and Q_n are exactly the j-th order statistics $[x_{1:n}]_j$ and $[y_{1:n}]_j$, meaning the j'th smallest elements of $x_{1:n}$ and $y_{1:n}$ respectively. Then $W_p(P_n, Q_n)$ takes the exact form

$$W_p(P_n,Q_n) = \left(\sum_{j=1}^n |[x_{1:n}]_j - [y_{1:n}]_j|^p \right)^{1/p}, \tag{8}$$

and is an estimator of $W_p(P,Q)$. This estimator costs $\mathcal{O}(n \log(n))$ to compute (due to the cost of sorting n data points), and a convergence rate of $\mathcal{O}(n^{-1/2})$ for $p = 1$, and minimax convergence rate $\mathcal{O}(n^{-1/2p})$ for integer $p > 1$ when P, Q have at least $2p$ finite moments. In some cases, the $p > 1$ rate can be improved upon to match the $\mathcal{O}(n^{-1/2})$ rate of $p = 1$: we refer to Bobkov and Ledoux (2019) for a thorough overview.

A.3. Sliced Wasserstein

The *sliced Wasserstein* (SW) distances (Rabin et al., 2011; Bonneel et al., 2015) between two distributions P, Q on $\mathbb{R}^d$ use one-dimensional projections to reduce computational cost.

Expected SW. For an integer $p \geq 1$, expected SW is defined as

$$\text{SW}_p(P,Q) \coloneqq \left(\mathbb{E}_{u\sim\mathbb{U}(S^{d-1})}[W_p^p(\phi_u \# P, \phi_u \# Q)] \right)^{1/p},$$

where $\mathbb{U}(S^{d-1})$ is the uniform distribution on the unit sphere S^{d-1}, the measures $\phi_u \# P$, $\phi_u \# Q$ are pushforwards under the projection operator $\phi_u(x) = \langle u, x \rangle$, and W_p is the one-dimensional p-Wasserstein distance as in Equation (7). Given $x_{1:n} \sim P$ and $y_{1:n} \sim Q$, the integral over the sphere is approximated by Monte Carlo sampling of l directions $u_{1:l}$, which together with the estimator in Equation (8) gives

$$\widehat{\text{SW}}_p^p(P,Q) = \frac{1}{l} \sum_{i=1}^l W_p^p(\phi_{u_j} \# P_n, \phi_{u_j} \# Q_n) = \frac{1}{ln} \sum_{i=1}^l \sum_{j=1}^n \left(\left[\langle u_i, x_{1:n} \rangle \right]_j - \left[\langle u_i, y_{1:n} \rangle \right]_j \right)^p$$

Here, $[\langle u_i, x_{1:n} \rangle]_j$ is the j-th order statistics, meaning the j-th smallest element of $\langle u_i, x_{1:n} \rangle = [\langle u_i, x_1 \rangle, \ldots, \langle u_i, x_n \rangle]^\top$. This estimator can be computed in $\mathcal{O}(ln \log n)$ time (the cost on sorting n samples, for l directions) and was shown to converge at rate $\mathcal{O}(l^{-1/2} + n^{-1/2})$ for $p = 1$ (Nadjahi et al., 2022).

Max SW. The *max-sliced Wasserstein* (max-SW) distance (Deshpande et al., 2018) replaces the average over projections in expected SW with a supremum over directions,

$$\text{max-SW}_p(P,Q) \coloneqq \left(\sup_{u\in S^{d-1}} W_p^p(\phi_u\#P, \phi_u\#Q)\right)^{1/p},$$

where, $\phi_u(x) = \langle u, x\rangle$ is again the projection operator, and W_p is the one-dimensional p-Wasserstein distance of Equation (7). Max-SW emphasizes the direction of greatest dissimilarity between the two measures.

Given $x_{1:n} \sim P$ and $y_{1:n} \sim P$, max-SW is estimated as $W_p(\phi_{u^*}\#P, \phi_{u^*}\#Q)$, for u^* the projection that maximises $W_p^p(\phi_u\#P_n, \phi_u\#Q_n)$ as given in Equation (8). In (Deshpande et al., 2018), u^* was approximated by optimising a heuristic, rather than the actual $W_p^p(\phi_u\#P_n, \phi_u\#Q_n)$. Then, Kolouri et al. (2022) approached the actual problem of

$$u^* = \operatorname*{argmax}_{\|u\|=1} W_p^p(\phi_u\#P_n, \phi_u\#Q_n).$$

by running projected gradient descent on S^{d-1}, where each gradient step requires computing the derivative of the 1D Wasserstein distance w.r.t. u. Concretely, they initialise u_1 randomly and iterate

$$u_{t+1} = \text{Proj}_{S^{d-1}}\Big(\text{Optim}\big(\nabla_u W_p^p(\phi_{u_t}\#P, \phi_{u_t}\#Q), u_{1:t}\big)\Big), \tag{9}$$

where $\text{Proj}_{S^{d-1}}(x) = x/\|x\|$ is the operator projecting onto the unit sphere, and Optim is an optimiser of choice, such as ADAM. Each evaluation of W_p and its gradient in one dimension costs $\mathcal{O}(n\log n)$, so the overall complexity is $\mathcal{O}(Tn\log n)$ for T gradient steps. It is important to point out the optimisation may be noisy, with the value objective getting worse after some iterations. Indeed, if z_{t+1} is the solution to $\text{Optim}\big(\nabla_u W_p^p(\phi_{u_t}\#P, \phi_{u_t}\#Q), u_{1:t}\big)$, is it an improvement over u_t, meaning $W_p^p(\phi_{u_t}\#P, \phi_{u_t}\#Q) \leq W_p^p(\phi_{z_{t+1}}\#P, \phi_{z_{t+1}}\#Q)$. Written out explicitly,

$$\sum_{j=1}^n |[\langle u_t, x_{1:n}\rangle]_j - [\langle u_t, y_{1:n}\rangle]_j|^p \leq \sum_{j=1}^n |[\langle z_{t+1}, x_{1:n}\rangle]_j - [\langle z_{t+1}, y_{1:n}\rangle]_j|^p,$$

Then, $u_{t+1} = \text{Proj}_{S^{d-1}}(z_{t+1}) = z_{t+1}/\|z_{t+1}\|$, and it may happen that $W_p^p(\phi_{u_t}\#P, \phi_{u_t}\#Q) > W_p^p(\phi_{u_{t+1}}\#P, \phi_{u_{t+1}}\#Q)$. The desired $W_p^p(\phi_{u_t}\#P, \phi_{u_t}\#Q) \leq W_p^p(\phi_{u_{t+1}}\#P, \phi_{u_{t+1}}\#Q)$ is guaranteed when $\|z_{t+1}\|^p \leq 1$, which may not happen.

A.4. Generalised Sliced Wasserstein

The *generalised (max-)sliced Wasserstein* (GSW and max-GSW) distances (Kolouri et al., 2022) extend SW and max-SW by using a family of nonlinear feature maps $\{f_\theta : \mathbb{R}^d \to \mathbb{R}\}_{\theta\in\Theta}$ instead of linear projections. Formally,

$$\text{GSW}_p(P,Q) \coloneqq \big(\mathbb{E}_{\theta\sim\mu} W_p^p(f_\theta\#P, f_\theta\#Q)\big)^{1/p}, \qquad \text{max-GSW}_p(P,Q) \coloneqq \left(\sup_{\theta\in\Theta} W_p^p(f_\theta\#P, f_\theta\#Q)\right)^{1/p},$$

where $f_\theta\#P$ denotes the pushforward of P by f_θ and μ is a probability measure over the parameter space Θ. For $f_\theta(x) = \langle\theta, x\rangle$ and $\Theta = S^{d-1}$ with uniform μ, GSW reduce to the standard SW distances. For expected GSW, sampling $\{\theta_i\}_{i=1}^l \sim \mu$ yields an estimator with the same $\mathcal{O}(ln\log n)$ computational complexity as expected SW (Kolouri et al., 2022). For max-GSW, the projected gradient descent approach of Equation (9) applies, at the same complexity of $\mathcal{O}(Tn\log n)$ as for max-SW.

Statistical and topological properties of GSW depend completely on the choice of the family $\{f_\theta : \theta \in \Theta\}$. Kolouri et al. (2022) consider the specific case of polynomial f_θ, and show GSW is then a metric on probability distributions on $\mathbb{R}^d$.

A.5. Kernel Sliced Wasserstein

A special case of the GSW arises when the feature maps f_θ are drawn from a reproducing kernel Hilbert space (RKHS). Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel that induces the RKHS $\mathcal{H}$ with unit sphere $S_\mathcal{H}$. Then, the *kernel sliced Wasserstein* (KSW) can be introduced as

$$\text{e-KSW}_p(P,Q) \coloneqq \big(\mathbb{E}_{u\sim\gamma} W_p^p(u\#P, u\#Q)\big)^{1/p}, \qquad \text{max-KSW}_p(P,Q) \coloneqq \left(\sup_{u\in S_\mathcal{H}} W_p^p(u\#P, u\#Q)\right)^{1/p},$$

where γ is some probability measure on $S_{\mathcal{H}}$. The expected KSW is a new construct, while max-KSW was introduced in Wang et al. (2022), and studied further in Wang et al. (2024b); in both papers k was assumed to be universal. Finding the optimal u^* for max-KSW was shown to be NP-hard in Wang et al. (2024b); they propose an estimator at cost $\mathcal{O}(T^{3/2}n^2)$. Though still more expensive than computing the V-statistic estimator of MMD, this is an improvement over $\mathcal{O}(Tn^3)$ in the original work of Wang et al. (2022).

As pointed out in the main text, the choice of a uniform γ in e-KSW, while seemingly natural, may not be feasible as there is no uniform or Lebesgue measure in infinite dimensional spaces. In the main paper, we propose a practical choice of γ that facilitates an efficient estimator, and study computational cost. Further, we establish statistical and topological properties that apply to both expected and max-KSW—and do not assume a universal kernel.

A.6. Sinkhorn Divergence

The entropic regularisation of optimal transport leads to the *Sinkhorn divergence* (Cuturi, 2013; Genevay et al., 2019). For distributions P, Q and regularisation parameter $\varepsilon > 0$, the entropic OT cost is defined as

$$W_{p,\varepsilon}(P,Q) \coloneqq \left(\inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}[\|X-Y\|^p] + \varepsilon \mathrm{KL}(\pi \| P \otimes Q) \right)^{1/p}.$$

The Sinkhorn divergence then corrects for the entropic bias:

$$\mathrm{S}_{p,\varepsilon}(P,Q) \coloneqq W_{p,\varepsilon}(P,Q) - W_{p,\varepsilon}(P,P)/2 - W_{p,\varepsilon}(Q,Q)/2. \tag{10}$$

This quantity interpolates between MMD-like behavior for large ε and true Wasserstein for $\varepsilon \to 0$, and can be computed efficiently via Sinkhorn iterations at cost $\mathcal{O}(n^2)$ per iteration (Cuturi, 2013).

A.7. Kernel covariance embeddings

Kernel covariance (operator) embeddings (KCE, Makigusa (2024b)) represent the distribution P as the second-order moment of the function $k(X,\cdot)$, for $X \sim P$, as an alternative to the first-order moment (the kernel mean embedding). Due to being moments of the same distribution, the two share key positives and drawbacks: KCE for kernel k exists if and only if KME for k^2 exists, and the kernel k is covariance characteristic if and only if k^2 is mean-characteristic (Bach, 2022). The divergence proposed in Makigusa (2024b) is the distance between the KCE, and is estimated at $\mathcal{O}(n^3)$ due to the need to compute full eigendecomposition of the KCE in order to compute the norm. In contrast, our proposed kernel quantile embeddings (KQE) embed quantiles, and therefore the relation to the KCE comes down to matching quantiles (which always exist, and come with an efficient estimator), compared to matching the second moment in the infinite-dimensional RKHS (which may not exist, and requires eigenvalue decomposition).

A.8. Kernel median embeddings

The median embedding (Nienkötter and Jiang, 2022) of P is the geometric median of $k(X,\cdot)$, $X \sim P$ in the RKHS, meaning the RKHS element which, on average, is L^1-closest to the point $k(X,\cdot)$. Explicitly put, it is the function $\mathrm{med}_P \in \mathcal{H}$ defined through

$$\mathrm{med}_P = \operatorname*{argmin}_{f\in\mathcal{H}} \int_{\mathcal{H}} \|f(\cdot) - k(x,\cdot)\|_{\mathcal{H}} P(\mathrm{d}x).$$

The median exists for any separable Hilbert space (Minsker, 2015). However, even for an empirical $P_n = {}^1\!/_n \sum_{i=1}^n \delta_{x_i}$, there is no closed-form solution to this L^1-problem, and the median is typically approximated using iterative algorithms like Weiszfeld's algorithm. The estimator proposed in Nienkötter and Jiang (2022) has a computational complexity of $\mathcal{O}(n^2)$. The property of being median-characteristic, as far as the authors are aware, has not been explored, and no theoretical guarantees are available.

The connection to 1D-projected quantiles as done in KQE, even specifically the 1D-projected median, is also unclear. Expanding the understanding of geometric median embeddings is an area for future research.

A.9. Other Related Work

Kernel methods have also been studied in the context of quantile estimation and regression (Sheather and Marron, 1990; Li et al., 2007). These methods, however, focus on using either kernel density estimation or kernel ridge regression to estimate

univariate quantiles. In contrast, our focus lies in exploring directional quantiles in the RKHS, and using them to estimate distances between distributions. We introduce this idea in the following section.

B. Connection between Centered and Uncentered Quantiles

Proposition 2 (Centered e-KQD$_2$)**.** *The e-KQD$_2$ and sup-KQD$_2$ correspondence derived based on centered directional quantiles, now expressed as* $\widetilde{\text{e-KQD}}_2(P,Q;\mu,\gamma)^2$ *and* $\widetilde{\text{sup-KQD}}_2(P,Q;\mu,\gamma)^2$ *can be expressed as follows,*

$$\begin{aligned}
\widetilde{\textit{e-KQD}}_2(P,Q;\mu,\gamma)^2 &= \textit{e-KQD}_2(P,Q;\mu,\gamma) + \mathrm{MMD}^2(P,Q) - \mathbb{E}_{u\sim\gamma}[(\mathbb{E}_{X\sim P}[u(X)] - \mathbb{E}_{Y\sim Q}[u(Y)])^2], \\
&\geq \textit{e-KQD}_2(P,Q;\mu,\gamma) + \mathrm{MMD}^2(P,Q) \\
\widetilde{\textit{sup-KQD}}_2(P,Q;\mu,\gamma)^2 &= \sup_{u\in S_{\mathcal{H}}} \left(\tau_2^2(P,Q,\mu,u) - (\mathbb{E}_{X\sim P}[u(X)] - \mathbb{E}_{Y\sim Q}[u(Y)])^2\right) + \mathrm{MMD}^2(P,Q) \\
&\geq \textit{sup-KQD}_2(P,Q;\mu,\gamma) + \mathrm{MMD}^2(P,Q)
\end{aligned}$$

Proof. Let $P, Q \in \mathcal{P}_{\mathcal{X}}$ be measures on some instance space $\mathcal{X}$. Further, define $\psi : x \mapsto k(x,\cdot)$, and write $P_\psi = \psi\#P$ and $Q_\psi = \psi\#Q$. Now P_ψ and Q_ψ are measures on the RKHS $\mathcal{H}_k$. Recall the definition of centered directional quantiles in Section 2.1,

$$\tilde{\rho}^{\alpha,u}_{P_\psi} = \left(\rho^{\alpha}_{\phi_u\#P_\psi} - \phi_u(\mathbb{E}_{Y\sim P_\psi}[Y])\right) u + \mathbb{E}_{Y\sim P_\psi}[Y]$$

Now since we are working in the RKHS $\mathcal{H}_k$, the expectation term $\mathbb{E}_{Y\sim P_\psi}[Y]$ corresponds to the kernel mean embedding $\mu_P := \mathbb{E}_P[k(X,\cdot)]$, thus we can rewrite the above expression as,

$$\tilde{\rho}^{\alpha,u}_{P_\psi} = \left(\rho_{\phi_u\#P_\psi} - \langle u, \mu_P\rangle\right) u + \mu_P$$

$\tilde{\rho}^{\alpha,u}_{Q_\psi}$ can be defined analogously. Now consider integrating the difference between the two centered directional quantiles along all quantile levels, leading to

$$\tilde{\tau}_2(P,Q,\mu,u) = \left(\int_0^1 \|\tilde{\rho}^{\alpha,u}_{P_\psi} - \tilde{\rho}^{\alpha,u}_{Q_\psi}\|^2_{\mathcal{H}_k}\mu(d\alpha)\right)^{\frac{1}{2}} \tag{11}$$

We now proceed to show $\tilde{\tau}_2^2(P,Q,\mu,u)$,where μ is the Lebesgue measure, can be expressed as a sum between an uncentered e-KQD$_2$ term with the MMD. Starting with expanding the RKHS norm inside the integrand,

$$\begin{aligned}
\|\tilde{\rho}^{\alpha,u}_{P_\psi} - \tilde{\rho}^{\alpha,u}_{Q_\psi}\|^2_{\mathcal{H}_k} &= \| \underbrace{(\rho^{\alpha}_{\phi_u\#P_\psi} - \rho^{\alpha}_{\phi_u\#Q_\psi} - \langle u, \mu_P - \mu_Q\rangle)}_{=:A\in\mathbb{R}} u + \mu_P - \mu_Q\|^2_{\mathcal{H}_k} \\
&= \|Au + (\mu_P - \mu_Q)\|^2_{\mathcal{H}_k} \\
&= 2\langle Au, \mu_P - \mu_Q\rangle + \|Au\|^2_{\mathcal{H}_k} + \|\mu_P - \mu_Q\|^2_{\mathcal{H}_k} \\
&= 2A\langle u, \mu_P - \mu_Q\rangle + A^2 + \mathrm{MMD}^2(P,Q)
\end{aligned} \tag{12}$$

Plugging the expression from Equation (12) into Equation (11), we get the following,

$$\begin{aligned}
\tilde{\tau}_2^2(P,Q,\mu,u) &= \int_0^1 (2A\langle u, \mu_P - \mu_Q\rangle + A^2)\mu(d\alpha) + \mathrm{MMD}^2(P,Q) \\
&= 2\langle u, \mu_P - \mu_Q\rangle \int_0^1 A\mu(d\alpha) + \int_0^1 A^2\mu(d\alpha) + \mathrm{MMD}^2(P,Q)
\end{aligned} \tag{13}$$

For the first term on the right hand side, notice that,

$$\int_0^1 A\mu(d\alpha) = \int_0^1 (\rho^{\alpha}_{\phi_u\#P_\psi} - \rho^{\alpha}_{\phi_u\#Q_\psi} - \langle u, \mu_P - \mu_Q\rangle)\mu(d\alpha) \tag{14}$$

Recall standard results from probability theory that integrating the quantile function between 0 to 1 with the Lebesgue measure returns you the expectation, specifically, that is,

$$\int_0^1 \rho^\alpha_{\phi_u \# P_\psi} \mu(d\alpha) = \mathbb{E}_{X\sim P}[u(X)] = \langle u, \mu_P \rangle.$$

Using this fact, the terms in Equation (14) cancels out, leaving $\int_0^1 A\mu(d\alpha) = 0$. Therefore, continuing from Equation (13), we have,

$$\begin{aligned}
\tilde{\tau}_2^2(P, Q, \mu, u) &= \int_0^1 A^2 \mu(d\alpha) + \mathrm{MMD}^2(P, Q) \\
&= \int_0^1 (\rho^\alpha_{\phi_u \# P_\psi} - \rho^\alpha_{\phi_u \# Q_\psi} - \langle u, \mu_P - \mu_Q \rangle)^2 \mu(d\alpha) + \mathrm{MMD}^2(P, Q) \\
&= \int_0^1 \|(\rho^\alpha_{s_{\mu_P, u} \# (\phi_u \# P_\psi)} - \rho^\alpha_{s_{\mu_Q, u} \# (\phi_u \# Q_\psi)}) u\|^2 \mu(d\alpha) + \mathrm{MMD}^2(P, Q)
\end{aligned}$$

where $s_{\mu_P, u} : \mathbb{R} \to \mathbb{R}$ is a shifting function defined as $s_{\mu_P, u}(r) = r - \langle u, \mu_p \rangle$ for $r \in \mathbb{R}$. Alternatively, after expanding the terms in A^2, we can express $\tilde{\tau}_2^2(P, Q, \mu, u)$ as,

$$\begin{aligned}
\tilde{\tau}_2^2(P, Q, \mu, u) &= \int_0^1 (\rho_{\phi_u \# P_\psi} - \rho_{\phi_u \# Q_\psi})^2 \mu(d\alpha) - (\mathbb{E}[u(X) - u(Y)])^2 + \mathrm{MMD}^2(P, Q) \\
&= \tau_2^2(P, Q, \mu, u) + \mathrm{MMD}^2(P, Q) - (\mathbb{E}[u(X) - u(Y)])^2
\end{aligned}$$

As a result, for γ a measure on the unit sphere of $\mathcal{H}_k$, the centered version of e-KQD$_2$ and sup-KQD$_2$, now expressed as $\widetilde{\text{e-KQD}}_2$ and $\widetilde{\text{sup-KQD}}_2$, are given by,

$$\begin{aligned}
\widetilde{\text{e-KQD}}_2(P, Q; \mu, \gamma)^2 &= \mathbb{E}_{u\sim\gamma}\left[\tilde{\tau}_2^2(P, Q; \mu, u)\right] \\
&= \text{e-KQD}_2(P, Q; \mu, \gamma)^2 + \mathrm{MMD}^2(P, Q) - \mathbb{E}_{u\sim\gamma}[(\mathbb{E}_{X\sim P}[u(X)] - \mathbb{E}_{Y\sim Q}[u(Y)])^2], \\
&\leq \text{e-KQD}_2(P, Q; \mu, \gamma)^2 + \mathrm{MMD}^2(P, Q) \\
\widetilde{\text{sup-KQD}}_2(P, Q; \mu, \gamma)^2 &= \sup_{u\in S_\mathcal{H}} \tilde{\tau}_2^2(P, Q; \mu, u) \\
&= \sup_{u\in S_\mathcal{H}} \left(\tau_2^2(P, Q, \mu, u) - (\mathbb{E}[u(X)] - \mathbb{E}[u(Y)])^2\right) + \mathrm{MMD}^2(P, Q) \\
&\leq \sup_{u\in S_\mathcal{H}} \tau_2^2(P, Q; \mu, u) - \sup_{u\in S_\mathcal{H}} (\mathbb{E}[u(X)] - \mathbb{E}[u(Y)])^2 + \mathrm{MMD}^2(P, Q) \\
&\leq \text{sup-KQD}_2(P, Q; \mu, \gamma)^2 + \mathrm{MMD}^2(P, Q).
\end{aligned}$$

□

When $\nu \equiv \mu$ and the connections to Sliced Wasserstein explored in Connection 1 and Connection 2 emerges, the mean-shifting property of Wasserstein distances allows us to express centered KQD as a sum of uncentered KQD, and MMD—a curious interpretation of centering.

C. Proof of Theoretical Results

This section now provides the proof of all theoretical results in the main text.

C.1. Proof of Theorem 1

The main result in this section, Proposition 3, shows that the set of $\mathbb{R}$ measures $\{u\#P : u \in S_\mathcal{H}\}$ fully determines the distribution P. Since quantiles determine the distribution, Theorem 1 follows immediately.

Being concerned with the RKHS case specifically allows us to prove the result under mild conditions by using *characteristic functionals*, an extension of characteristic functions to measures on spaces beyond $\mathbb{R}^d$. Characteristic functionals describe Borel probability measures as operators acting on some function space $\mathcal{F} : \mathcal{X} \to \mathbb{R}$.

Definition 1 (Vakhania et al. (1987), Section IV.2.1)**.** The *characteristic functional* $\varphi_P : \mathcal{F} \to \mathbb{C}$ of a Borel probability measure P on $\mathcal{X}$ is defined as

$$\varphi_P(f) = \int_{\mathcal{X}} e^{if(x)} P(\mathrm{d}x).$$

Theorem 2.2(a) in Vakhania et al. (1987, Chapter 4) establishes that a P-characteristic functional on $\mathcal{F}$ uniquely determines the distribution P—on the smallest σ-algebra under which all function $f \in \mathcal{F}$ are measurable. Therefore, when $\mathcal{F}$ is such that this σ-algebra coincides with the Borel σ-algebra, the distribution is fully determined by P-characteristic functional on $\mathcal{F}$. We show that, indeed, this holds in our setting, for $\mathcal{F} = \mathcal{H}$.

Lemma 1. *Suppose A1 and A2 holds. Then, the Borel σ-algebra $\mathcal{B}(\mathcal{X})$ is the smallest σ-algebra on $\mathcal{X}$ under which all functions $f \in \mathcal{H}$ are measurable.*

Proof. Denote by $\hat{C}(\mathcal{X}, \mathcal{H})$ the smallest σ-algebra on $\mathcal{X}$ under which all functions $f \in \mathcal{H}$ are measurable, and recall that the Borel σ-algebra is the σ-algebra that contains all closed sets. Therefore, we need to show that $\hat{C}(\mathcal{X}, \mathcal{H})$ contains every closed set in $\mathcal{X}$. We split the proof into two parts: (1) show that $\mathcal{H}$ contains a countable separating subspace, and (2) show that this implies that every closed set lies in $\hat{C}(\mathcal{X}, \mathcal{H})$.

$\mathcal{H}$ contains a countable separating subspace. Recall that a function space $\mathcal{F}$ on $\mathcal{X}$ is said to be separating when for any $x_1 \neq x_2 \in \mathcal{X}$, there is a function $f \in \mathcal{F}$ such that $f(x_1) \neq f(x_2)$. Since k is separating, $\mathcal{H}$ is separating. Since $\mathcal{H}$ is separable, it contains a countable dense subspace $\mathcal{H}_0 \subseteq \mathcal{H}$. By $\mathcal{H}_0$ being dense in $\mathcal{H}$, it must also be separating.

Every closed set lies in $\hat{C}(\mathcal{X}, \mathcal{H})$. By Vakhania et al. (1987, Section I.1, Exercise 9), all compact sets in $\mathcal{X}$ lie in $\hat{C}(\mathcal{X}, \mathcal{H}_0)$, by $\mathcal{H}_0$ being countable, continuous, separating space of real-valued functions. By definition, $\hat{C}(\mathcal{X}, \mathcal{H}_0) \subseteq \hat{C}(\mathcal{X}, \mathcal{H})$, and so $\hat{C}(\mathcal{X}, \mathcal{H})$ contains all compact sets. We now show this means every closed set must also lie in $\hat{C}(\mathcal{X}, \mathcal{H})$.

By $\mathcal{X}$ being σ-compact, there is a family of compact sets $\{\mathcal{X}_i\}_{i=1}^{\infty}$ such that $\mathcal{X} = \cup_{i=1}^{\infty} \mathcal{X}_i$. Take any closed $K \subseteq X$; then, $K = \cup_{i=1}^{\infty} (\mathcal{X}_i \cap K)$. Since $\mathcal{X}_i \cap K$ is compact as the intersection of a compact set and a closed set, and σ-algebras are closed under countable unions, K must lie in $\hat{C}(\mathcal{X}, \mathcal{H})$. As this holds for every closed K, we conclude $\mathcal{B}(\mathcal{X}) = \hat{C}(\mathcal{X}, \mathcal{H})$. □

We now restate the RKHS-specific version of the Vakhania result here for completeness.

Theorem 6 (Theorem 2.2(a) in Vakhania et al. (1987) for RKHS)**.** *Suppose A1 and A2 holds, and for Borel probability measures P, Q on $\mathcal{X}$, it holds that $\varphi_P(f) = \varphi_Q(f)$ for every $f \in \mathcal{H}$. Then, $P = Q$.*

We are now ready to prove the distribution of projections uniquely determines the distribution.

Proposition 3. *Under A1 and A2, it holds that*

$$u\#P = u\#Q \text{ for all } u \in S_{\mathcal{H}} \iff P = Q.$$

Proof. The main idea of the proof is to show that equality of $u\#P$ and $u\#Q$ implies equality of characteristic functionals, $\varphi_P(f) = \varphi_Q(f)$ for all $f \in \mathcal{H}$ such that $f(x) = tu(x)$ for some $t \in \mathbb{R}$ and u in the unit sphere. Since such f form the entire $\mathcal{H}$, the result immediately follows.

First, recall that $u\#P = u\#Q$ for all u if and only if their characteristic functions coincide, meaning

$$\int_{\mathbb{R}} e^{itz} u\#P(\mathrm{d}z) = \int_{\mathbb{R}} e^{itz} u\#Q(\mathrm{d}z) \quad \forall u \in S_{\mathcal{H}}, \forall t \in \mathbb{R}. \tag{15}$$

Notice that the measure $u\#P$ is a pushforward of P under the map $x \to u(x)$. Then, for any measurable g it holds that

$$\int_{\mathcal{X}} g(u(x)) P(\mathrm{d}x) = \int_{\mathbb{R}} g(z) u\#P(\mathrm{d}z) \quad \forall u \in S_{\mathcal{H}}. \tag{16}$$

Take $g(z) = e^{itz}$, for some $t \in \mathbb{R}$. Then, for all u it holds that $\int_{\mathbb{R}} e^{itz} u\#P(\mathrm{d}z) = \int_{\mathbb{R}} e^{itz} u\#Q(\mathrm{d}z)$, and consequently by (15) we have that

$$\int_{\mathcal{X}} e^{itu(x)} P(\mathrm{d}x) = \int_{\mathcal{X}} e^{itu(x)} Q(\mathrm{d}x) \quad \forall u \in S_{\mathcal{H}}, \forall t \in \mathbb{R}. \tag{17}$$

Finally, let us pick an $f \in \mathcal{H}$ and show that $\varphi_P(f) = \varphi_Q(f)$. Define $u = f/\|f\|$, and $t = \|f\|$; then,

$$\varphi_P(f) = \int_{\mathcal{X}} e^{if(x)} P(\mathrm{d}x) = \int_{\mathcal{X}} e^{itu(x)} P(\mathrm{d}x),$$

and by (17), we arrive at the equality of characteristic functionals, $\varphi_P(f) = \varphi_Q(f)$. By Theorem 6 characteristic functionals uniquely determine the underlying distribution, meaning $P = Q$. □

For the sake of clarity, we give the proof of the original result.

Proof of Theorem 1. Suppose $\{\rho_P^{\alpha,u} : \alpha \in [0,1], u \in S_{\mathcal{H}}\} = \{\rho_Q^{\alpha,u} : \alpha \in [0,1], u \in S_{\mathcal{H}}\}$ for some Borel probability measures P, Q. For any fixed u, since every quantile of of $u\#P$ and $u\#Q$ coincide, the measures coincide as well, $u\#P = u\#Q$. As that holds for every u, by Proposition 3, $P = Q$. □

Lastly, we point out A1 may be relaxed. Provided $\mathcal{X}$ is a Tychonoff space—meaning, a completely regular Hausdorff space—part (b) of Theorem 2.2 in Vakhania et al. (1987) says the following.

Theorem 7 (Theorem 2.2(b) in Vakhania et al. (1987) for RKHS)**.** *Suppose $\mathcal{X}$ is Tychonoff, A2 holds, and for Radon probability measures P, Q on $\mathcal{X}$, it holds that $\varphi_P(f) = \varphi_Q(f)$ for every $f \in \mathcal{H}$. Then, $P = Q$.*

Therefore, when A1 is replaced with $\mathcal{X}$ being Tychonoff, Theorem 1 continues to hold—but only for Radon P, Q, not any Borel P, Q. Radon probability measures can be intuitively seen as the "non-pathological" Borel measures—a restriction employed in order to drop the regularity assumptions of $\mathcal{X}$ being separable and σ-compact.

C.2. Proof of Theorem 2

We prove that every mean-characteristic kernel is quantile-characteristic, and give an example quantile-characteristic kernel that is not mean-characteristic.

mean-characteristic ⇒ quantile-characteristic. Suppose k on $\mathcal{X}$ is mean-characteristic, and $P \neq Q$ are any probability measures on $\mathcal{X}$. We will identify a unit-norm u for which the sets of quantiles of $u\#P$ and $u\#Q$ differ.

Since k is mean characteristic, $\mu_P \neq \mu_Q$, and $\mathrm{MMD}^2(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 > 0$. Recall that MMD can be expressed as

$$\mathrm{MMD}^2(P,Q) = \sup_{u \in \mathcal{H}, \|u\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_{X \sim P} u(X) - \mathbb{E}_{Y \sim Q} u(Y) \right|,$$

and the supremum is attained at $u^* = (\mu_P - \mu_Q)/\|\mu_P - \mu_Q\|_{\mathcal{H}}$ (Gretton et al., 2012). In other words, $\mathbb{E}_{X \sim P} u^*(X) \neq \mathbb{E}_{Y \sim Q} u^*(Y)$—the means of $u^*\#P$ and $u^*\#Q$ don't coincide. Therefore, the measures $u^*\#P$ and $u^*\#Q$ don't coincide, or equivalently $\{\rho^{\alpha}_{u^*\#P} : \alpha \in [0,1]\} \neq \{\rho^{\alpha}_{u^*\#Q} : \alpha \in [0,1]\}$. Then, $\{\rho_P^{u,\alpha} : \alpha \in [0,1], u \in S_{\mathcal{H}}\} \neq \{\rho_Q^{\alpha} : \alpha \in [0,1], u \in S_{\mathcal{H}}\}$. And since this holds for any arbitrary $P \neq Q$, the kernel k is quantile-characteristic.

quantile-characteristic ⇏ mean-characteristic. To show the converse implication does not hold, we provide an example when k is quantile-characteristic but not mean-characteristic. Take $\mathcal{X} = \mathbb{R}^d$, and let k be a degree T polynomial kernel, $k(x, x') = (x^\top x' + 1)^T$. Since A1 and A2 hold—$\mathbb{R}^d$ is Polish, and k is trivially continuous and separating—by Theorem 1 the kernel k is quantile-characteristic.

Now, we show k is not mean-characteristic. Suppose P and Q are such that $\mathbb{E}_{X \sim P} X^i = \mathbb{E}_{Y \sim P} Y^i$ for $i \in \{1, \ldots, T\}$—for example, the Gaussian and Laplace distribution with matching expectation and variance and $T = 2$, as is done in Section 5.2. Then, $\mathbb{E}_{X \sim P}(X^\top x')^i = \mathbb{E}_{Y \sim P}(Y^\top x')^i$ for any $x' \in \mathbb{R}^d$, and since

$$\mu_P(x') \coloneqq \mathbb{E}_{X \sim P} k(X, x') = \mathbb{E}_{X \sim P}[(X^\top x' + 1)^T] = \mathbb{E}_{X \sim P}\left[\sum_{i=0}^{T} \binom{T}{i} (X^\top x')^i\right] = \sum_{i=0}^{T} \binom{T}{i} \mathbb{E}_{X \sim P}\left[(X^\top x')^i\right],$$

it holds that $\mu_P = \mu_Q$. The kernel is not mean-characteristic.

C.3. Proof of Theorem 3

By the Theorem in Serfling (2009, Section 2.3.2), for any $\varepsilon > 0$ it holds that

$$P(|\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}| > \varepsilon) \leq 2e^{-2n\delta_\varepsilon^2}, \qquad \text{for} \qquad \delta_\varepsilon := \min\left\{\int_{\rho^\alpha_{u\#P}}^{\rho^\alpha_{u\#P}+\varepsilon} f_{u\#P}(t)\mathrm{d}t, \int_{\rho^\alpha_{u\#P}-\varepsilon}^{\rho^\alpha_{u\#P}} f_{u\#P}(t)\mathrm{d}t\right\}.$$

Since it was assumed $f_{u\#P}(x) \geq c_u > 0$, it holds that $\delta_\varepsilon \geq c_u\varepsilon$, and $P(|\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}| > \varepsilon) \leq 2e^{-2nc_u^2\varepsilon^2}$, or equivalently,

$$P(|\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}| \leq \varepsilon) \geq 1 - 2e^{-2nc_u^2\varepsilon^2}.$$

Take $\delta := 2e^{-2nc_u^2\varepsilon^2}$. Then,

$$P(|\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}| \leq C(\delta, u)n^{-1/2}) \geq 1 - \delta, \qquad \text{for} \qquad C(\delta, u) = \sqrt{\frac{\log(2/\delta)}{2c_u^2}}.$$

Since $\|\rho^{\alpha,u}_{P_n} - \rho^{\alpha,u}_P\|_\mathcal{H} = |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|$, the proof is complete.

C.4. Proof of Theorem 4

We prove e-KQD and sup-KQD, defined in Equation (4) as

$$\text{e-KQD}_p(P, Q; \nu, \gamma) = \left(\mathbb{E}_{u\sim\gamma}\tau_p^p\left(P, Q; \nu, u\right)\right)^{1/p},$$
$$\text{sup-KQD}_p(P, Q; \nu) = \left(\sup_{u\in S_\mathcal{H}} \tau_p^p\left(P, Q; \nu, u\right)\right)^{1/p},$$

are probability metrics on the set of Borel probability measures on $\mathcal{X}$. Symmetry and non-negativity hold trivially.

Triangle inequality. By Minkowski inequality, for any P, P', Q,

$$\int_0^1 |\rho^\alpha_P - \rho^\alpha_{P'}|^p \nu(\mathrm{d}\alpha) \leq \left(\left(\int_0^1 |\rho^\alpha_P - \rho^\alpha_Q|^p \nu(\mathrm{d}\alpha)\right)^{1/p} + \left(\int_0^1 |\rho^\alpha_Q - \rho^\alpha_{P'}|^p \nu(\mathrm{d}\alpha)\right)^{1/p}\right)^p.$$

Plugging this in and using Minkowski inequality again on the outermost integral, we get

$$\begin{aligned}
\text{e-KQD}_p(P, P'; \nu, \gamma) &= \left(\mathbb{E}_{u\sim\gamma}\int_0^1 |\rho^\alpha_P - \rho^\alpha_{P'}|^p \nu(\mathrm{d}\alpha)\right)^{1/p} \\
&\leq \left(\mathbb{E}_{u\sim\gamma}\left(\left(\int_0^1 |\rho^\alpha_P - \rho^\alpha_Q|^p \nu(\mathrm{d}\alpha)\right)^{1/p} + \left(\int_0^1 |\rho^\alpha_Q - \rho^\alpha_{P'}|^p \nu(\mathrm{d}\alpha)\right)^{1/p}\right)^p\right)^{1/p} \\
&\leq \left(\mathbb{E}_{u\sim\gamma}\int_0^1 |\rho^\alpha_P - \rho^\alpha_Q|^p \nu(\mathrm{d}\alpha)\right)^{1/p} + \left(\mathbb{E}_{u\sim\gamma}\int_0^1 |\rho^\alpha_Q - \rho^\alpha_{P'}|^p \nu(\mathrm{d}\alpha)\right)^{1/p} \\
&= \text{e-KQD}_p(P, Q; \nu, \gamma) + \text{e-KQD}_p(Q, P'; \nu, \gamma).
\end{aligned}$$

Similarly, since $\sup_x f^p(x) = (\sup_x |f(x)|)^p$ for any f,

$$\begin{aligned}
\text{sup-KQD}_p(P, P'; \nu, \gamma) &= \left(\sup_{u \in S_{\mathcal{H}}} \int_0^1 \left| \rho_P^\alpha - \rho_{P'}^\alpha \right|^p \nu(\mathrm{d}\alpha) \right)^{1/p} \\
&\leq \left(\sup_{u \in S_{\mathcal{H}}} \left(\left(\int_0^1 \left| \rho_P^\alpha - \rho_Q^\alpha \right|^p \nu(\mathrm{d}\alpha) \right)^{1/p} \right. \right. \\
&\qquad \left. \left. + \left(\int_0^1 \left| \rho_Q^\alpha - \rho_{P'}^\alpha \right|^p \nu(\mathrm{d}\alpha) \right)^{1/p} \right)^p \right)^{1/p} \\
&= \left(\sup_{u \in S_{\mathcal{H}}} \int_0^1 \left| \rho_P^\alpha - \rho_Q^\alpha \right|^p \nu(\mathrm{d}\alpha) \right)^{1/p} \\
&\qquad + \left(\sup_{u \in S_{\mathcal{H}}} \int_0^1 \left| \rho_Q^\alpha - \rho_{P'}^\alpha \right|^p \nu(\mathrm{d}\alpha) \right)^{1/p} \\
&= \text{sup-KQD}_p(P, Q; \nu, \gamma) + \text{sup-KQD}_p(Q, P'; \nu, \gamma).
\end{aligned}$$

Identity of indiscernibles. In the rest of this section, we show that

$$\text{e-KQD}_p(P, Q; \nu, \gamma) = 0 \iff P = Q; \qquad \text{and} \qquad \text{sup-KQD}_p(P, Q; \nu, \gamma) = 0 \iff P = Q.$$

Necessity (meaning the $\Leftarrow$ direction) holds trivially—quantiles of identical measure are identical. To prove sufficiency, we only need to show that both discrepancies aggregate the directions in a way that preserves injectivity, meaning

$$\text{e-KQD}_p(P, Q) = 0 \Rightarrow \rho_P^{\alpha,u} = \rho_Q^{\alpha,u} \text{ for all } \alpha, u; \qquad \text{and} \qquad \text{sup-KQD}_p(P, Q) = 0 \Rightarrow \rho_P^{\alpha,u} = \rho_Q^{\alpha,u} \text{ for all } \alpha, u.$$

Together with Theorem 1, this will complete the proof of sufficiency.

First, we show that for any pair of probability measures, a ν-aggregation over the quantiles is injective.

Lemma 2. *Let ν have full support, meaning $\nu(A) > 0$ for any open $A \subset [0, 1]$. For any Borel probability measures P', Q',*

$$\int_0^1 |\rho_{P'}^\alpha - \rho_{Q'}^\alpha|^2 \nu(d\alpha) = 0 \qquad \Rightarrow \qquad \rho_{P'}^\alpha = \rho_{Q'}^\alpha \quad \textit{for all} \quad \alpha \in [0, 1].$$

Proof. Suppose $\int_0^1 |\rho_{P'}^\alpha - \rho_{Q'}^\alpha|^2 \nu(\mathrm{d}\alpha) = 0$, but there is an α_0 such that $\rho_{P'}^{\alpha_0} = \rho_{Q'}^{\alpha_0}$. We will show that this implies the existence of an open set (containing α_0) over which $|\rho_{P'}^{\alpha_0} - \rho_{Q'}^{\alpha_0}|^2 > 0$—which will contradict ν having full support.

Since $|\rho_{P'}^{\alpha_0} - \rho_{Q'}^{\alpha_0}|^2 > 0$ and the quantile function $\alpha \mapsto q_P^\alpha$ is left-continuous (by definition) for any probability measure P, there is a $\alpha_1 < \alpha_0$ such that $|\rho_{P'}^\alpha - \rho_{Q'}^\alpha|^2 > 0$ for all $\alpha \in (\alpha_1, \alpha_0]$. Take some $\alpha_2 \in (\alpha_1, \alpha_0)$. Then, for all $\alpha \in (\alpha_1, \alpha_2)$, we have $|\rho_{P'}^\alpha - \rho_{Q'}^\alpha|^2 > 0$. We arrive at a contradiction. Such α_0 cannot exist, and therefore $\rho_{P'}^\alpha = \rho_{Q'}^\alpha$ for all $\alpha \in [0, 1]$. □

This result applies directly to the directional differences τ_p. Provided ν has full support,

$$\tau_p(P, Q; \nu, u) = 0 \qquad \Rightarrow \qquad \rho_{P'}^\alpha = \rho_{Q'}^\alpha \quad \text{for all} \quad \alpha \in [0, 1].$$

Since supremum aggregation simply considers u that corresponds to the largest $\tau_p^p(P, Q; \nu, u)$, this concludes the proof for sup-KQD. Expectation aggregation over the directions u needs an extra result, given below.

Lemma 3. *Let γ have full support on $S_{\mathcal{H}}$, and ν have full support on $[0, 1]$. For any Borel probability measures P, Q on $\mathcal{X}$,*

$$\mathbb{E}_{u \sim \gamma} \tau_p^p(P, Q; \nu, u) = 0 \qquad \Rightarrow \qquad P = Q.$$

Proof. Same as in the proof Theorem 1, we will use the technique of characteristic functionals φ_P, φ_Q, to carefully prove equality almost everywhere with respect to a full support measure γ implies full equality. Consider the function

$$f \mapsto \varphi_P(f) - \varphi_Q(f),$$

which is continuous by continuity of characteristic functionals. Define $f_0 \equiv 0$, the zero function in $\mathcal{H}$. The set

$$\mathcal{H}^{\setminus 0} := \{f \in \mathcal{H} \setminus \{f_0\} : \varphi_P(f) - \varphi_Q(f) \in \mathbb{R} \setminus \{0\}\} = \{f \in \mathcal{H} \setminus \{f_0\} : \varphi_P(f) \neq \varphi_Q(f)\}$$

is open, as a preimage of an open set $\mathbb{R} \setminus \{0\}$, intersected with an open set $\{\mathcal{H} \setminus \{f_0\}\}$. Since the projection map $f \mapsto f/\|f\|_{\mathcal{H}}$ is open on $\mathcal{H} \setminus \{f_0\}$, the projection of $\mathcal{H}_{\setminus 0}$ onto $S_{\mathcal{H}}$ is open. In other words, the set

$$S_{\mathcal{H}}^{\setminus 0} := \{u \in S_{\mathcal{H}} : \varphi_P(t_u u) \neq \varphi_Q(t_u u) \text{ for some } t_u \in \mathbb{R}\}$$

is open in $S_{\mathcal{H}}$. Then, by definition of characteristic functionals, for $u \in S_{\mathcal{H}}^{\setminus 0}$ it holds that

$$\varphi_{u\#P}(t_u) = \varphi_P(t_u u) \neq \varphi_Q(t_u u) = \varphi_{u\#Q}(t_u),$$

meaning the characteristic functions of $u\#P$ and $u\#Q$ are not identical, and therefore $u\#P \neq u\#Q$. Since ν has full support on $[0,1]$, it follows that

$$\tau_p^p(P, Q; \nu, u) = \int_0^1 |\rho_{u\#P}^\alpha - \rho_{u\#Q}^\alpha|^p \nu(\mathrm{d}\alpha) > 0, \qquad \text{for all } u \in S_{\mathcal{H}}^{\setminus 0}$$

We arrive at a contradiction: since γ has full support on $S_{\mathcal{H}}$ and $S_{\mathcal{H}}^{\setminus 0} \subseteq S_{\mathcal{H}}$ was shown to be an open set, it holds that

$$\mathbb{E}_{u\sim\gamma} \tau_p^p(P, Q; \nu, u) \geq \int_{S_{\mathcal{H}}^{\setminus 0}} \tau_p^p(P, Q; \nu, u) \gamma(\mathrm{d}u) > 0.$$

Therefore, for $\mathbb{E}_{u\sim\gamma} \tau_p^p(P, Q; \nu, u)$ to be zero, $S_{\mathcal{H}}^{\setminus 0}$ must be empty—which, by construction, can only happen when $\mathcal{H}^{\setminus 0}$ is empty, i.e. $\varphi_P(f) = \varphi_Q(f)$ for all $f \in \mathcal{H} \setminus f_0$, where $f_0 \equiv 0$. Since $\varphi_P(f_0) = \varphi_Q(f_0)$ holds trivially for any P, Q, the characteristic functionals of P and Q are identical. By Theorem 6, $P = Q$. This concludes the proof. □

C.5. Proof of Theorem 5

We start with two auxiliary lemmas that, when combined, bound e-KQD approximation error due to replacing P, Q with P_n, Q_n in $n^{-1/2}$. This will be crucial in showing convergence of the approximate e-KQD to the true e-KQD.

Lemma 4. *For any measure ν on $[0,1]$ and any measure γ on $S_{\mathcal{H}}$, it holds that*

$$|\textit{e-KQD}_1(P_n, Q_n; \nu, \gamma) - \textit{e-KQD}_1(P, Q; \nu, \gamma)| \leq \textit{e-KQD}_1(P_n, P; \nu, \gamma) + \textit{e-KQD}_1(Q_n, Q; \nu, \gamma)).$$

Proof. By the definition of e-KQD$_1$ and Jensen inequality for the absolute value,

$$\begin{aligned} |\text{e-KQD}_1(P_n, Q_n; \nu, \gamma) - \text{e-KQD}_1(P, Q; \nu, \gamma)| &= \left| \mathbb{E}_{u\sim\gamma} \left[\int_0^1 \left(|\rho_{u\#P_n}^\alpha - \rho_{u\#Q_n}^\alpha| - |\rho_{u\#P}^\alpha - \rho_{u\#Q}^\alpha| \right) \mathrm{d}\alpha \right] \right| \\ &\leq \mathbb{E}_{u\sim\gamma} \left[\int_0^1 \left| |\rho_{u\#P_n}^\alpha - \rho_{u\#Q_n}^\alpha| - |\rho_{u\#P}^\alpha - \rho_{u\#Q}^\alpha| \right| \mathrm{d}\alpha \right] \end{aligned}$$

By the reverse triangle inequality followed by the triangle inequality,

$$\begin{aligned} \left| |\rho_{u\#P_n}^\alpha - \rho_{u\#Q_n}^\alpha| - |\rho_{u\#P}^\alpha - \rho_{u\#Q}^\alpha| \right| &\leq |\rho_{u\#P_n}^\alpha - \rho_{u\#P}^\alpha + \rho_{u\#Q}^\alpha - \rho_{u\#Q_n}^\alpha| \\ &\leq |\rho_{u\#P_n}^\alpha - \rho_{u\#P}^\alpha| + |\rho_{u\#Q_n}^\alpha - \rho_{u\#Q}^\alpha|, \end{aligned} \tag{18}$$

and the statement of the lemma follows. □

Lemma 5. *Let ν be a measure on $[0,1]$ with density f_ν bounded above by $C_\nu > 0$. With probability at least $1 - \delta/4$, for $C'(\delta) = 2C_\nu\sqrt{\log(8/\delta)/2}$, it holds that*

$$\textit{e-KQD}_1(P_n, P; \nu, \gamma) \leq \frac{C'(\delta)}{2} n^{-1/2}$$

Proof. Recall that

$$\text{e-KQD}_1(P_n, P; \nu, \gamma) = \mathbb{E}_{u\sim\gamma}\left[\tau_1(P_n, P; \nu, u)\right], \qquad \tau_1(P_n, P; \nu, u) = \int_0^1 |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|\nu(\mathrm{d}\alpha).$$

Let $F_{u\#P}$ and $F_{u\#P_n}$ be the CDFs of $u\#P$ and $u\#P_n$ respectively. Then,

$$\begin{aligned}\int_0^1 |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|\nu(\mathrm{d}\alpha) \le C_\nu \int_0^1 |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|\mathrm{d}\alpha &= C_\nu \int_{u(\mathcal{X})} |F_{u\#P_n}(t) - F_{u\#P}(t)|\mathrm{d}t \\ &\le C_\nu \sup_{t\in u(\mathcal{X})} |F_{u\#P_n}(t) - F_{u\#P}(t)|,\end{aligned}$$

where the last equality is the well known fact that integrated difference between quantiles is equal to integrated difference between CDFs (see, for instance, Bobkov and Ledoux (2019, Theorem 2.9)). By the Dvoretzky-Kiefer-Wolfowitz inequality, with probability at least $1 - \delta/4$ it holds that,

$$\sup |F_{u\#P_n}(t) - F_{u\#P}(t)| < \sqrt{\log(8/\delta)/2}n^{-1/2},$$

and therefore, with probability at least $1 - \delta/4$ for $C'(\delta) = 2C_\nu\sqrt{\log(8/\delta)/2}$,

$$\tau_1(P_n, P; \nu, u) = \int_0^1 |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|\nu(\mathrm{d}\alpha) \le \frac{C'(\delta)}{2}n^{-1/2}.$$

In other words, the random variable $\tau_1(P_n, P; \nu, u)$ is sub-Gaussian with sub-Gaussian constant $C_\tau := C_\nu^2/(2n)$, meaning

$$\Pr\left[\tau_1(P_n, P; \nu, u) \ge \varepsilon\right] \le 2\exp\{-\varepsilon^2/C_\tau^2\}$$

One of the equivalent definitions for a sub-Gaussian random variable is the moment condition: for any $p \ge 1$,

$$\mathbb{E}_{x_{1:n}}\left[\tau_1(P_n, P; \nu, u)^p\right] \le 2C_\tau^p\Gamma(p/2 + 1).$$

An application of Jensen inequality and Fubini's theorem shows that the moment condition holds for $\mathbb{E}_{u\sim\gamma}\tau_1(P_n, P; \nu, u)$,

$$\mathbb{E}_{x_{1:n}}\left[(\mathbb{E}_{u\sim\gamma}\tau_1(P_n, P; \nu, u))^p\right] \le \mathbb{E}_{x_{1:n}}\mathbb{E}_{u\sim\gamma}\left[\tau_1(P_n, P; \nu, u)^p\right] = \mathbb{E}_{u\sim\gamma}\mathbb{E}_{x_{1:n}}\left[\tau_1(P_n, P; \nu, u)^p\right] \le 2C_\tau^p\Gamma(p/2 + 1).$$

Therefore, $\mathbb{E}_{u\sim\gamma}\tau_1(P_n, P; \nu, u)$ is sub-Gaussian with constant $C_\tau = C_\nu^2/(2n)$, meaning it holds with probability at least $1 - \delta/4$ that

$$\text{e-KQD}_1(P_n, P; \nu, \gamma) = \mathbb{E}_{u\sim\gamma}\tau_1(P_n, P; \nu, u) \le \frac{C'(\delta)}{2}n^{-1/2}.$$

□

We are now ready to prove the full result.

Proof of Theorem 5. Let C_ν be an upper bound on the density of ν. By triangle inequality, the full error can be upper bounded by R_l, the error due to approximation of γ with γ_l, plus R_n, the error due to approximation of P, Q with P_n, Q_n,

$$\begin{aligned}|\text{e-KQD}_1(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_1(P, Q; \nu, \gamma)| &\le |\text{e-KQD}_1(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_1(P_n, Q_n; \nu, \gamma)| \\ &\quad + |\text{e-KQD}_1(P_n, Q_n; \nu, \gamma) - \text{e-KQD}_1(P, Q; \nu, \gamma)| \\ &=: R_l + R_n.\end{aligned}$$

We bound R_l in $l^{-1/2}$, and R_n in $n^{-1/2}$, with high probability.

Bounding R_l. Recall that e-KQD$_1(P_n, Q_n; \nu, \gamma) = \mathbb{E}_{u\sim\gamma}\left[\int_0^1 |\rho^\alpha_{u\#P} - \rho^\alpha_{u\#Q}|\nu(\mathrm{d}\alpha)\right]$. Therefore, we may apply McDiarmid's inequality provided for any $u, u' \in S_\mathcal{H}$ we upper bound the difference

$$\left|\int_0^1 \left|\rho^\alpha_{u\#P} - \rho^\alpha_{u\#Q}\right| - \left|\rho^\alpha_{u'\#P} - \rho^\alpha_{u'\#Q}\right| \nu(\mathrm{d}\alpha)\right|.$$

We have that

$$\begin{aligned}
\left|\int_0^1 \left|\rho^\alpha_{u\#P} - \rho^\alpha_{u\#Q}\right| - \left|\rho^\alpha_{u'\#P} - \rho^\alpha_{u'\#Q}\right| \nu(\mathrm{d}\alpha)\right| &\overset{(A)}{\leq} \int_0^1 \left|\rho^\alpha_{u\#P} - \rho^\alpha_{u\#Q}\right| \nu(\mathrm{d}\alpha) + \int_0^1 \left|\rho^\alpha_{u'\#P} - \rho^\alpha_{u'\#Q}\right| \nu(\mathrm{d}\alpha) \\
&\overset{(B)}{\leq} 2C_\nu \sup_{u\in S_\mathcal{H}} W_1(u\#P, u\#Q) \\
&\overset{(C)}{\leq} 2C_\nu \sup_{u\in S_\mathcal{H}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}|u(X) - u(Y)| \\
&\overset{(D)}{\leq} 2C_\nu \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}\sqrt{k(X,X) - 2k(X,Y) + k(Y,Y)}
\end{aligned}$$

where (A) holds by Jensen's and triangle inequalities; (B) uses boundedness of the density of ν by C_ν and the property of the Wasserstein distance in $\mathbb{R}$ from Equation (7); (C) uses the infimum definition of the Wasserstein distance; and (D) holds by the reasoning we employed multiple times through the paper, via reproducing property, Cauchy-Schwarz, and having $u, u' \in S_\mathcal{H}$. So we arrive at a bound

$$\left|\int_0^1 \left|\rho^\alpha_{u\#P} - \rho^\alpha_{u\#Q}\right| - \left|\rho^\alpha_{u'\#P} - \rho^\alpha_{u'\#Q}\right| \nu(\mathrm{d}\alpha)\right| \leq 2C_\nu \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}\sqrt{k(X,X) - 2k(X,Y) + k(Y,Y)} =: 2C_\nu C_k.$$

Now that boundedness of the difference has been established, by McDiarmid's inequality, with probability at least $1 - \delta/2$ and for $C''(\delta) = \sqrt{2C_\nu C_k \log(4/\delta)}$ it holds that

$$|\text{e-KQD}_1(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_1(P_n, Q_n; \nu, \gamma)| \leq C''(\delta) l^{-1/2}.$$

Bounding R_n. By Lemma 4,

$$|\text{e-KQD}_1(P_n, Q_n; \nu, \gamma) - \text{e-KQD}_1(P, Q; \nu, \gamma)| \leq \text{e-KQD}_1(P_n, P; \nu, \gamma) + \text{e-KQD}_1(Q_n, Q; \nu, \gamma))$$

By Lemma 5 and the union bound, with probability at least $1 - \delta/2$ and for $C'(\delta) = 2C_\nu\sqrt{\log(8/\delta)/2}$, it holds that

$$R_n = |\text{e-KQD}_1(P_n, Q_n; \nu, \gamma) - \text{e-KQD}_1(P, Q; \nu, \gamma)| \leq C'(\delta) n^{-1/2}.$$

Combining bounds. By applying the union bound again, to $R_l + R_n$, we get that, with probability at least $1 - \delta$,

$$|\text{e-KQD}_1(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_1(P, Q; \nu, \gamma)| \leq R_l + R_n \leq C''(\delta) l^{-1/2} + C'(\delta) n^{-1/2} \leq C(\delta)(l^{-1/2} + n^{-1/2}),$$

for $C(\delta) = \max\{C'(\delta), C''(\delta)\} = \mathcal{O}(\sqrt{\log(1/\delta)})$. This completes the proof □

As pointed out in the main text, $\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}\sqrt{k(X,X) - 2k(X,Y) + k(Y,Y)} < \infty$ holds immediately when $\mathbb{E}_{X\sim P}\sqrt{k(X,X)}$ and $\mathbb{E}_{X\sim Q}\sqrt{k(X,X)}$ are finite, and even more specifically, when the kernel k is bounded. Unbounded k and finite expectations, for example, happens when the tails of both P and Q decay fast enough to "compensate" for the growth of $k(x,x)$. For instance, when k is a polynomial kernel of any order (which is unbounded), and P and Q are laws of sub-exponential random variables. For clarity, note that $\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim Q}\sqrt{k(X,X) - 2k(X,Y) + k(Y,Y)}$ does not compare to MMD, which integrates $k(X, X')$ rather than $k(X, X)$ (see Equation (6)).

For integer $p > 1$, proving the $n^{-1/2}$ convergence rate is feasible if more involved—primarily because we can no longer reduce the problem to convergence of empirical CDFs to true CDFs. In general, for $p > 1$,

$$\int_0^1 |\rho^\alpha_{u\#P_n} - \rho^\alpha_{u\#P}|^p \mathrm{d}\alpha \neq \int_{u(\mathcal{X})} |F_{u\#P_n}(t) - F_{u\#P}(t)|^p \mathrm{d}t.$$

The following result, restated in our notation, makes the added complexity explicit.

Lemma 6 (Theorem 5.3 in Bobkov and Ledoux (2019)). *Suppose $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a bounded kernel, and ν has a density $0 < c_\nu \leq f_\nu \leq C_\nu$ on $[0, 1]$. Then, for any $u \in S_\mathcal{H}$, and for any $p \geq 1$ and $n \geq 1$,*

$$\mathbb{E}_{x_{1:n}\sim P}\left[\tau_p^p(P_n, P; \nu, u)\right] \leq \left(\frac{5pC_\nu}{\sqrt{n+2}}\right)^p J_p(u\#P), \quad \text{for} \quad J_p(u\#P) = \int_{u(\mathcal{X})} \frac{(F_{u\#P}(t)(1-F_{u\#P}(t)))^{p/2}}{f_{u\#P}^{p-1}(x)} dt.$$

Further, it holds that $\mathbb{E}_{x_{1:n}\sim P}\left[\tau_p^p(P_n, P; \nu, u) = \mathcal{O}(n^{-p/2})\right]$ if and only if $J_p(u\#P) < \infty$.

We now state a likely result for $p > 1$ as a conjecture, and outline the proof.

Conjecture 1 (**Finite-Sample Consistency for Empirical KQDs for $p > 1$**). *Let $\mathcal{X} \subseteq \mathbb{R}^d$, ν have a density, P, Q be measures on $\mathcal{X}$ with densities bounded away from zero, $f_P(x) \geq c_P > 0$ and $f_P(x) \geq c_Q > 0$. Suppose $\mathbb{E}_{X\sim P}[k(X,X)^{p/2}] < \infty$ and $\mathbb{E}_{X\sim Q}[k(X,X)^{p/2}] < \infty$, and $x_{1:n} \sim P, y_{1:n} \sim Q$. Then,*

$$\mathbb{E}_{\substack{x_{1:n}\sim P \\ y_{1:n}\sim Q}} |\textit{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) - \textit{e-KQD}_p(P, Q; \nu, \gamma)| = \mathcal{O}(l^{-1/2} + n^{-1/2}).$$

Sketch proof. Analogously to the proof of Theorem 5, we can decompose the term of interest as

$$\begin{aligned}
&\mathbb{E}_{\substack{x_{1:n}\sim P \\ y_{1:n}\sim Q}} |\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_p(P, Q; \nu, \gamma)| \\
&\qquad \leq \mathbb{E}_{\substack{x_{1:n}\sim P \\ y_{1:n}\sim Q}} |\text{e-KQD}_p(P_n, Q_n; \nu, \gamma_l) - \text{e-KQD}_p(P_n, Q_n; \nu, \gamma)| \\
&\qquad\qquad + \left(\mathbb{E}_{x_{1:n}\sim P} \text{e-KQD}_p^p(P_n, P; \nu, \gamma)\right)^{1/p} + \left(\mathbb{E}_{y_{1:n}\sim Q} \text{e-KQD}_p^p(Q_n, Q; \nu, \gamma)\right)^{1/p}
\end{aligned}$$

The first term can be, same as in the proof of Theorem 5, bounded by McDiarmid's inequality. The second term (to the power p) takes the form

$$\mathbb{E}_{x_{1:n}\sim P} \text{e-KQD}_p^p(P_n, P; \nu, \gamma) = \mathbb{E}_{x_{1:n}\sim P} \mathbb{E}_{u\sim\gamma} \tau_p^p(P_n, P; \nu, u).$$

Then, by Lemma 6 (possibly modified to account for an extra expectation), to get the result we will need to show that $\mathbb{E}_{u\sim\gamma} J_p(u\#P) < \infty$,

$$\mathbb{E}_{u\sim\gamma} J_p(u\#P) = \mathbb{E}_{u\sim\gamma}\left[\int_{u(\mathcal{X})} \frac{(F_{u\#P}(t)(1-F_{u\#P}(t)))^{p/2}}{f_{u\#P}^{p-1}(x)} \mathrm{d}t\right] < \infty$$

The nominator is upper bounded by 2^{-p}. The denominator may get arbitrarily small without the nominator getting arbitrarily small: when the PDF $f_{u\#P}^{p-1}(x)$ is small, the CDF $F_{u\#P}(x)$ need not be close to zero or one. Therefore, it is necessary and sufficient to show

$$\mathbb{E}_{u\sim\gamma}\left[\int_{u(\mathcal{X})} \frac{1}{f_{u\#P}^{p-1}(x)} \mathrm{d}t\right] < \infty. \tag{19}$$

We proceed to outline key elements of the proof of such result, and leave a rigorous proof for future work. By the coarea formula, and since $f_P(x) \geq c_P > 0$,

$$f_{u\#P}(t) = \int_{u^{-1}(t)} \frac{f_P(x)}{|\nabla u(x)|} H^{d-1}(\mathrm{d}x) \geq c_0 \int_{u^{-1}(t)} \frac{1}{|\nabla u(x)|} H^{d-1}(\mathrm{d}x), \qquad \text{for} \qquad |\nabla u(x)| = \sqrt{\sum_{i=1}^d \left(\frac{\partial u(x)}{\partial x_i}\right)^2}$$

where $u^{-1}(t) = \{x \in \mathcal{X} : u(x) = t\}$, and H^{d-1} is the $d-1$-dimensional Hausdorff measure, which within $\mathcal{X} \subseteq \mathbb{R}^d$ is equal to $d-1$ dimensional Lebesgue measure, scaled by a constant that only depends on $d-1$.

Therefore, the integral in Equation (19) may diverge if the integral

$$\int_{u^{-1}(t)} \frac{1}{|\nabla u(x)|} H^{d-1}(\mathrm{d}x) \tag{20}$$

gets very small over "large" parts of $u(\mathcal{X})$—on average over $u \sim \gamma$. Trivially, if u is constant over some interval—or more generally, u has infinitely many critical points—the integral diverges. Fortunately, the more general condition is easy to

control: if u is a *Morse function* and $\mathcal{X}$ is compact, then u has only a finite number of critical points. It is a classic result (see, for instance, Hirsch (1976, Theorem 1.2)) that Morse functions form a dense open subset of twice differentiable real-valued functions on R^d, denoted $C^2(\mathbb{R}^d)$. Therefore, if $\mathcal{H} \subset C^2(\mathcal{X})$ (which can be reduced to smoothness of the kernel k—it holds for instance, for the Matérn-5/2 kernel), we get that $u \sim \gamma$ has a finite number of critical points almost surely under mild regularity assumptions on γ.

The final ingredient is to use the Morse lemma to lower bound Equation (20) in the epsilon-ball of each critical point. Morse lemma says u is quadratic around each critical point—which yields bounds on both the volume of $u^{-1}(t)$, and $1/|\nabla u(x)|$ in terms of the eigenvalues of the Hessian. Careful analysis of the eigenvalues will be needed to ensure the expectation with respect to $u \sim \gamma$ is finite. □

C.6. Proof of Connections 1 and 2

The equality in Equation (7) immediately gives the connection of e-KQD and sup-KQD to the expected-SW and max-SW respectively—previously only defined on $\mathcal{X} = \mathbb{R}^d$.

Further, for $\mathcal{X} = \mathbb{R}^d$, viewing $x \mapsto k(x, \cdot)$ as a transformation on $\mathcal{X}$ reveals a connection to Generalised Sliced Wasserstein (GSW, Kolouri et al. (2022)). In particular, the polynomial kernel $k(x, x') = (x^\top x' + 1)^T$ of odd degree T recovers the polynomial transformation for which GSW was proven to be a probability metric. Outside of the case of the polynomial case, proving that GSW is a metric is highly challenging. This is easier under the kernel framework, as we showed in Theorem 4. In Kolouri et al. (2022), the authors investigate learning transformations with neural networks (NNs). An interesting direction for future work is the relationship between said NNs and the kernels they induce.

C.7. Proof of Proposition 1

Recall that by definition of Gaussian measures in Hilbert spaces (Kukush, 2020), a random element $f \in \mathcal{H}$ has the law of a Gaussian measure $\mathcal{N}(0, C_m)$ on $\mathcal{H}$ when for any $g \in \mathcal{H}$,

$$\langle f, g\rangle_{\mathcal{H}} \sim \mathcal{N}(0, \langle C_m[g], g\rangle). \tag{21}$$

Since $C_m[g](x) = {}^1\!/_m \sum_{j=1}^m g(z_j)k(z_j, x)$, by the reproducing property,

$$\langle C_m[g], g\rangle = \frac{1}{m}\sum_{j=1}^m g(z_i)^2. \tag{22}$$

Take $f(x) = {}^1\!/_{\sqrt{m}} \sum_{j=1}^m \lambda_j k(z_j, x)$, for $\lambda_1, \ldots, \lambda_m \sim \mathcal{N}(0, \mathrm{Id})$. Then, for any $g \in \mathcal{H}$, by the reproducing property it holds that

$$\langle f, g\rangle_{\mathcal{H}} = \frac{1}{\sqrt{m}}\sum_{j=1}^m \lambda_j g(z_j) \sim \mathcal{N}\left(0, \frac{1}{m}\sum_{i=1}^m g(z_i)^2\right),$$

which is exactly the Gaussian measure with covariance operator C_m, as per Equations (21) and (22).

D. Additional Numerical Results

D.1. Type I control

We report the Type I control experiments for the CIFAR-10 v.s. CIFAR-10.1 experiment. Results are shown in Figure 5.

D.2. Figure 3 for e-KQD$_1$

It is common in power p-parametrised methods to select $p = 2$, to balance out sensitivity to outliers (which is higher for larger p, to the point of methods becoming brittle for $p > 2$), and robustness (which tends to be highest for $p = 1$); this trade-off, for instance, inspired the introduction of the Huber loss (Huber, 1964). However, for completeness, we now repeat experiments in the main paper for $p = 1$. The relationship to baseline approaches—MMD, MMD-Multi, and MMD-Lin—remains the same as observed for $p = 2$. However, it is evident that e-KQD$_1$ performed better than e-KQD$_2$ at the power decay and galaxy MNIST experiments, but the centered e-KQD$_1$ performed worse than centered e-KQD$_2$ at the Laplace v.s. Gaussian experiment. The implications of choosing p warrants a deeper investigation, left to future work.

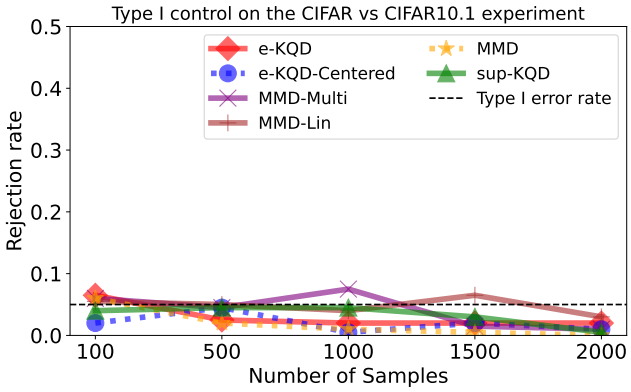

Figure 5: Type I control results for our experiment on CIFAR-10 v.s. CIFAR-10.1. We see all methods control their Type I error around or below the specified Type I error rate 0.05, thus confirming our tests in the main text are valid testing procedures.

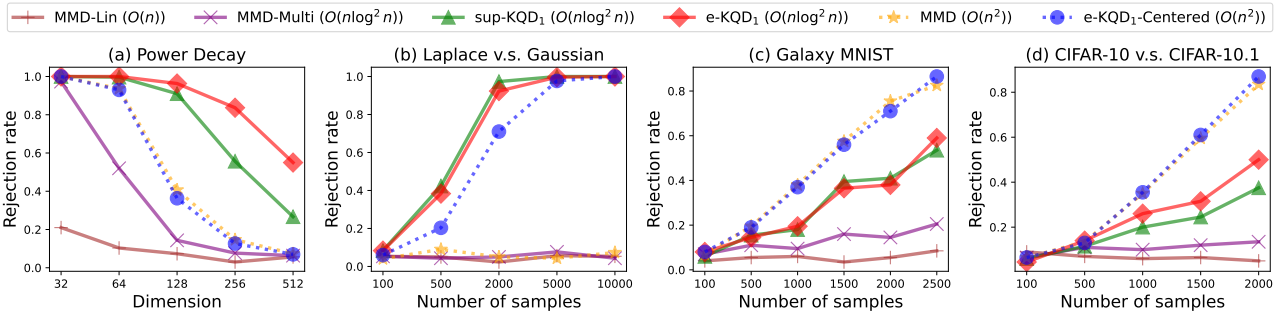

Figure 6: The experiments in Figure 3 repeated for $p = 1$. Experimental results comparing our proposed methods with baseline approaches. A higher rejection rate indicates better performance in distinguishing between distributions. **Same as for $p = 2$, quadratic-time quantile-based estimators perform comparably to quadratic-time MMD estimators, while near-linear time quantile-based estimators often outperform their MMD-based counterparts.**

D.3. Comparison of weighting measures

The Gaussian Kernel Quantile Discrepancy introduced in Section 4 has multiple weighting measures that determine properties of the distance: the measure ν on the quantile levels, the measure ξ within the covariance operator, and the measure γ on the unit sphere $S_{\mathcal{H}}$. We investigate the impact of varying these.

Varying ν. We conducted the following experiment using the Galaxy MNIST and CIFAR datasets. We varied ν, from assigning more weight to the extreme quantiles to down-weighting them. The results are presented in Figure 7, where the reverse triangle $\backslash/$ stands for up-weighing extreme quantiles, and the triangle $/\backslash$ stands for down-weighing them. We observed some improvement over the uniform ν: for Galaxy MNIST, test power improved when ν assigned less weight to extremes, whereas for CIFAR, the opposite was true, with higher test power when more weight was given to extremes. Uniform weighting of the quantiles remained a good choice. This suggests that tuning ν beyond the uniform is problem-dependent and can enhance performance. The difference likely arises from the nature of the problems: CIFAR datasets, where samples are expected to be similar, benefit from emphasising extremes, while Galaxy MNIST, which contains fundamentally different galaxy images, performs better when "robustified," i.e., focusing on differences away from the tails. Exploring this further presents an exciting avenue for future work.

Varying ξ. The reference measure ξ in the covariance operator C serves to "cover the input space" and is typically set to a "default" measure on the space—for $\mathbb{R}^d$, the standard Gaussian measure. The choice $(P_n + Q_n)/2$ made in the main body of the paper is aiming to adhere to the most general setting, when no default measure may be available—only P_n and Q_n.

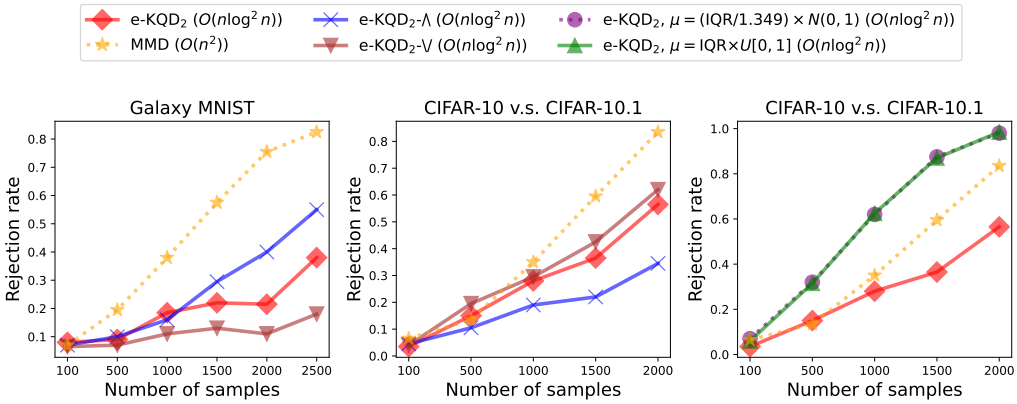

Figure 7: Gaussian KQD test power under different weighting measures. *Left, middle:* Varying measure ν: down-weighing ($/\backslash$) extremes boosts power on Galaxy MNIST, while up-weighing ($\backslash/$) them helps on CIFAR. Uniform weighting remains a strong default, with optimal ν depending on the dataset. *Right:* Varying measure ξ: using an IQR-scaled Gaussian or uniform default reference measure ξ both outperform MMD—indicating potential advantage of a "default" ξ over the problem-based $\xi = (P_n + Q_n)/2$.

We report a comparison on performance when the reference measure is: (1) $(P_n + Q_n)/2$; (2) a standard Gaussian measure, scaled by IQR/1.349 to match the spread of the data, where IQR is the interquantile range of $P_n + Q_n$, and 1.349 is the interquantile range of the standard Gaussian; and (3) a uniform measure on $[-1, 1]^d$, scaled by IQR.

The results, presented in Figure 7, show performance superior to MMD for the standard/uniform ξ. This indicates value in picking a "default" measure when one is available.

Varying γ Varying the measure on the sphere beyond a Gaussian is extremely challenging in infinite-dimensional spaces due to the complexity of both its theoretical definition and practical sampling. Since no practically relevant alternative has been proposed, we leave this direction unexplored.

D.4. Comparison to sliced Wasserstein distances

We extend the power decay experiment to include sliced Wasserstein and max-sliced Wasserstein distances, with directions (1) sampled uniformly on the sphere, and (2) sampled from $(P_n + Q_n)/2$ and projected onto the sphere. The results are plotted in Figure 8, and show that sliced Wasserstein distances perform significantly worse than e-KQD. This outcome is expected—as noted in Connections 1 and 2, sliced Wasserstein is equivalent to e-KQD with the linear kernel, which is less expressive than the Gaussian kernel.

D.5. Comparison with MMD based on Other KME Approximations

There are several efficient kernel mean embedding methods available in the literature, and no single approach has emerged as definitively superior. To complement experiments in the main body of the paper, we compare the e-KQD (at matching cost) with (1) The Mean Embedding (ME) approximation of MMD of Chwialkowski et al. (2015b), which was identified as the best-performing method in their numerical study; (2) the Nyström-MMD method of Chatalic et al. (2022), and (3) the Median-of-Means (MOM) approximation of Lerasle et al. (2019), specifically, their faster method (MONK BCD-Fast) that achieves matching cost to our e-KQD at the number of blocks $Q = n/\log n$.

The results are presented in Figure 8. ME performs at the level of MMD-multi, while Nyström has extremely high Type II error, likely due to sensitivity to hyperparameters. Due to Median-of-Means still being considerably slower than e-KQD (with the number of optimiser iterations set to $T = 100$), we apply it to a cheaper Power Decay problem (rather than the larger and more complicated Galaxy MNIST), where it performs at the level of the linear approximation of MMD. This may be due to MOM primarily being robustness-enforcing method, rather than a method aiming to build an efficient approximation of MMD.

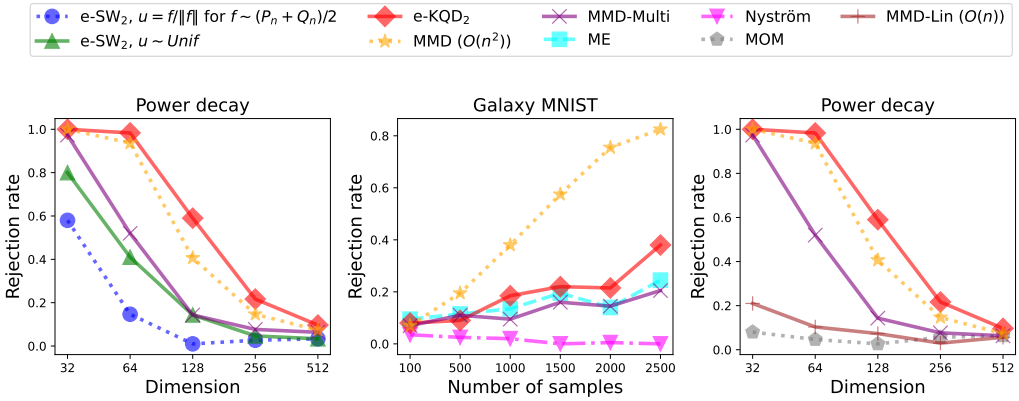

Figure 8: All methods are cost $\mathcal{O}(n \log^2 n)$ unless specified otherwise. *Left:* Gaussian KQD compared with sliced Wasserstein with uniform or data-driven directions, on the power decay problem. Sliced Wasserstein fall well below KQD—consistent with their equivalence to KQD using a less expressive linear kernel. *Middle:* Comparison with alternative approximate KME methods, at matching cost. ME matches MMD-multi power, while Nyström-MMD suffers high Type II error. *Right:* Comparison with Median-of-Means (MOM) KME approximation, at matching cost. MOM is primarily a robustness-enforcing method, not a cheap-approximation method, and doesn't perform well at set cost of $\mathcal{O}(n \log^2 n)$.