A Kernel Two-Sample Test with the Representation Jensen-Shannon Divergence

Jhoan K. Hoyos-Osorio
Department of Electrical and Computer Engineering
University of Kentucky
Lexington, KY 40508
keider.hoyos@uky.edu

Luis G. Sanchez-Giraldo
Department of Electrical and Computer Engineering
University of Kentucky
Lexington, KY 40508
luis.sanchez@uky.edu

Abstract

We introduce a novel kernel-based information-theoretic framework for two-sample testing, leveraging the representation Jensen-Shannon divergence (RJSD). RJSD captures higher-order information from covariance operators in reproducing Kernel Hilbert spaces and avoids Gaussianity assumptions, providing a robust and flexible measure of divergence between distributions. We develop RJSD-based variants of Maximum Mean Discrepancy (MMD) approaches, demonstrating superior discriminative power in extensive experiments on synthetic and real-world datasets. Our results position RJSD as a powerful alternative to MMD, with the potential to significantly impact kernel-based learning and distribution comparison. By establishing RJSD as a benchmark for two-sample testing, this work lays the foundation for future research in kernel-based divergence estimation and its broad range of applications in machine learning.

1 Introduction

The problem of non-parametric two-sample testing, which aims to detect differences between two data distributions given only observations, remains a fundamental challenge in machine learning. Among the most widely used metrics for two-sample testing is the Maximum Mean Discrepancy (MMD) [Gretton et al., 2012]. MMD consists of mapping the distributions into a reproducing kernel Hilbert space (RKHS) and computing the distance between their mean embeddings. In the past decade, MMD has been a dominant approach in kernel-based two-sample testing, and several MMD-based two-sample tests have been proposed, leading to significant advances in the field [Gretton et al., 2012, Sutherland et al., 2016, Jitkrittum et al., 2016, Liu et al., 2020, Schrab et al., 2023, Biggs et al., 2024]. Despite its widespread adoption, MMD's reliance on first-order moment information has motivated the exploration of alternative methods.

Recent advances in kernel-based divergence estimation offer a promising direction. Specifically, covariance operators (second-order moment information) in RKHS can be used to formulate distribution divergences [Harandi et al., 2014, Minh, 2015, Zhang et al., 2019, Minh, 2021, 2023]. However, these measures often assume Gaussianity in the representation space, which may not hold in practice.

Latinx in AI @ NeurIPS-24 (LXAI 2024).

To address these limitations, a novel kernel-based information-theoretic framework called *the representation Jensen-Shannon divergence* (RJSD) has been proposed as a versatile alternative [Hoyos-Osorio et al., 2023]. RJSD is formulated as the von Neumann Jensen-Shannon divergence between infinite-dimensional covariance operators in reproducing kernel Hilbert spaces (RKHS). This formulation provides a proper divergence between distributions in the input space without relying on density estimation or assuming Gaussianity in the feature space, making it a powerful alternative to existing approaches.

RJSD not only extends the concept of divergence in RKHS but also holds a direct connection to MMD. We show that MMD can be viewed as a particular case of RJSD, while RJSD captures higher-order information, leading to improved performance in tasks like two-sample testing. Additionally, RJSD can be readily estimated from samples in the input space using Gram matrices.

In this work, we leverage RJSD to propose a novel kernel-based information-theoretic framework for two-sample testing. Inspired by three well-known MMD-based tests, including MMD-Split [Sutherland et al., 2016], MMD-Deep [Liu et al., 2020], and MMD-Fuse [Biggs et al., 2024], we develop RJSD-based variants, enabling more powerful and flexible testing procedures. Our work significantly advances kernel-based two-sample testing, providing a robust alternative to MMD. We evaluate the efficacy of our approach through extensive experiments, demonstrating its potential to improve the state-of-the-art in two-sample testing.

2 Preliminaries and background

In this section, we introduce the notation and discuss fundamental concepts.

2.1 Notation

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Let $\mathcal{M}_+^1(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$, and let $P, Q \in \mathcal{M}_+^1(\mathcal{X})$ be two probability measures dominated by a σ-finite measure λ on $(\mathcal{X}, \mathcal{F})$ (Similar notation from Stummer and Vajda [2012]). Then, the densities $p = \frac{\mathrm{d}P}{d\lambda}$ and $q = \frac{\mathrm{d}Q}{d\lambda}$ have common support (the densities are positive on $\mathcal{X}$). $X \sim P$ and $Y \sim Q$ are two random variables distributed according to P and Q.

2.2 Kernel Mean Embedding

Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a positive definite kernel. There exists a mapping $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space, such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. The kernel mean embedding is a mapping μ from $\mathcal{M}_+^1(\mathcal{X})$ to $\mathcal{H}$ defined as follows [Smola et al., 2007]: For $P \in \mathcal{M}_+^1(\mathcal{X})$,

$$\mu_P = \mathbb{E}_{X \sim P}[\phi(X)] = \int_{\mathcal{X}} \phi(x) \, \mathrm{d}\, P(x)$$

For a bounded kernel, $\kappa(x, x) < \infty$ for all $x \in \mathcal{X}$, we have that for any $f \in \mathcal{H}$, $\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}}$.

2.3 Covariance Operator

Another related mapping is the uncentered covariance operator [Baker, 1973], one of the most important and widely used tools in RKHS theory. In this case, $P \in \mathcal{M}_+^1$ is mapped to an operator $C_P : \mathcal{H} \to \mathcal{H}$ given by:

$$C_P = \mathbb{E}_{X \sim P}[\phi(X) \otimes \phi(X)] = \int_{\mathcal{X}} \phi(x) \otimes \phi(x) \, \mathrm{d}\, P(x), \tag{1}$$

where $\otimes$ is the tensor product. Similarly, for any $f, g \in \mathcal{H}$, $\mathbb{E}_{X \sim P}[f(X)g(X)] = \langle g, C_P f \rangle_{\mathcal{H}}$.

The covariance operator is positive semidefinite and Hermitian (self-adjoint). Additionally, if the kernel is bounded, that is $\kappa(x, y) < \infty$, the covariance operator is trace class [Sanchez Giraldo et al., 2014, Bach, 2022]. Therefore, the spectrum of the covariance operator is discrete and consists of non-negative eigenvalues λ_i with $\sum \lambda_i < \infty$, for which we can extend functions on $\mathbb{R}$ such as $t \log(t)$ and t^α to covariance operators via their spectrum [Naoum and Gittan, 2004].

2.4 Information theory with covariance operators

Throughout this paper, unless otherwise stated, we will assume that:

(A1) $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a positive definite kernel with an RKHS mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$, and $\kappa(x, x) = 1 \quad \forall x \in \mathcal{X}$.

Under this assumption, the covariance operator C_P defined in Eqn. 1 is unit-trace. Note that since $\kappa(x, x) = 1$, we have that, $\operatorname{Tr}(\phi(x) \otimes \phi(x)) = \|\phi(x)\|^2 = 1$. Hence, the spectrum of the covariance operator consists of non-negative eigenvalues λ_i with $\sum \lambda_i = 1$, for which we can extend notions of entropy from the spectrum of unit-trace covariance operators.

Definition 1. *Let X be a random variable taking values in $\mathcal{X}$ and probability measure P. Assume $(\mathbf{A1})$ holds, and let C_P be the corresponding unit-trace covariance operator defined in Eqn. 1. Then, the representation (kernel) entropy of X is defined as:*

$$H^{\mathcal{H}}(X) = S(C_P) = -\operatorname{Tr}(C_P \log C_P), \tag{2}$$

where $S(\cdot)$ is a generalization of the von Neumann entropy [Von Neumann, 2018] for trace class operators, and it can be equivalently formulated as $S(C_P) = -\sum \lambda_i \log \lambda_i$.

Although the representation entropy has similar properties to those of Shannon entropy, it is important to emphasize that they are not equal, and thus estimating representation entropy does not amount to estimating Shannon entropy. Instead, the representation entropy incorporates the data representation. Its properties are not only determined by the data distribution but also depend on the representation (kernel).

2.4.1 Empirical estimation of representation entropy

Let $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \sim P$ be n i.i.d samples of a random variable X with probability measure P. An empirical estimate of representation entropy can be obtained based on the spectrum of the empirical uncentered covariance operator $\boldsymbol{C_x}$. Consider the Gram matrix $\boldsymbol{K_x}$, consisting of all pairwise kernel evaluations between data points in the sample $\boldsymbol{X}$, that is, $(\boldsymbol{K_x})_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for $i, j = 1, \ldots, n$. It can be shown that $\boldsymbol{C_x}$ and $\frac{1}{n}\boldsymbol{K_x}$ have the same non-zero eigenvalues [Sanchez Giraldo et al., 2014, Bach, 2022]. Based on this equivalence, the estimator of representation entropy can be expressed in terms of the Gram matrix $\boldsymbol{K_x}$ as follows:

Proposition 1. *The empirical kernel-based representation entropy estimator of X is*

$$\hat{H}^{\mathcal{H}}(X) = S(\boldsymbol{C_x}) = S\left(\tfrac{1}{n}\boldsymbol{K_x}\right) = -\operatorname{Tr}\left(\tfrac{1}{n}\boldsymbol{K_x} \log \tfrac{1}{n}\boldsymbol{K_x}\right) = -\sum_{i=1}^{n} \lambda_i \log \lambda_i, \tag{3}$$

where λ_i denotes the ith eigenvalue of $\frac{1}{n}\boldsymbol{K_x}$. The eigen-decomposition of $\boldsymbol{K_x}$ has $\mathcal{O}(n^3)$ time complexity.

3 The representation Jensen-Shannon divergence

Definition 2. *Let P and Q be two probability measures defined on a measurable space $(\mathcal{X}, \mathcal{F})$, and $(\mathbf{A1})$ is satisfied. Then, the **representation Jensen-Shannon divergence** (RJSD) between P and Q is defined as [Hoyos-Osorio et al., 2023]:*

$$D_{JS}^{\mathcal{H}}(P, Q) = S\left(\frac{C_P + C_Q}{2}\right) - \frac{1}{2}\left(S(C_P) + S(C_Q)\right). \tag{4}$$

3.1 Properties

First, we show that RJSD relates to the maximum mean discrepancy (MMD) with kernel κ^2, where MMD is defined as $\mathrm{MMD}_\kappa^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2$.

Lemma 1. *For all probability measures P and Q defined on $\mathcal{X}$, and covariance operators C_P and C_Q with RKHS mapping $\phi(x)$ such that $\langle \phi(x), \phi(x) \rangle_{\mathcal{H}} = 1 \quad \forall x \in \mathcal{X}$:*

$$D_{JS}^{\mathcal{H}}(P, Q) \geq \frac{1}{8} \|C_P - C_Q\|_{HS}^2 = \frac{1}{8} \operatorname{MMD}_{\kappa^2}^2(P, Q)$$

Proof: See Appendix A.1.

Theorem 1. *Let κ^2 be a characteristic kernel. Then, the representation Jensen-Shannon divergence $D_{JS}^{\mathcal{H}}(P, Q) = 0$ if and only if $P = Q$.*

Proof. It is clear that if $P = Q$ then $D_{JS}^{\mathcal{H}}(P, Q) = 0$. We now prove the opposite. According to Lemma 1, $D_{JS}^{\mathcal{H}}(P, Q) = 0$ implies that $\operatorname{MMD}_{\kappa^2}^2(P, Q) = 0$. Then, if $\operatorname{MMD}_{\kappa^2}^2(P, Q) = 0$ and the kernel κ^2 is characteristic, then $P = Q$ [Gretton et al., 2012], completing the proof. □

This theorem demonstrates that RJSD defines a proper divergence between probability measures in the input space.

Additionally, RJSD has a direct connection with its classical counterpart.

Theorem 2. *[Hoyos-Osorio et al., 2023, Theorem 3] For all probability measures P and Q defined on $\mathcal{X}$, and unit-trace covariance operators C_P and C_Q, the following inequality holds:*

$$D_{JS}^{\mathcal{H}}(P, Q) \leq D_{JS}(P, Q), \tag{5}$$

where $D_{JS}(P, Q)$ is the traditional Jensen-Shannon divergence.

3.2 Empirical Estimation of the representation Jensen-Shannon divergence

Given two sets of samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subset \mathcal{X}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^m \subset \mathcal{X}$ drawn from two unknown probability measures P and Q, we propose the following RJSD estimator:

Kernel-based estimator: Let κ be a positive definite kernel, $\boldsymbol{Z}$ be the mixture of the samples of $\boldsymbol{X}$ and $\boldsymbol{Y}$, that is, $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{n+m}$ where $\boldsymbol{z}_i = \boldsymbol{x}_i$ for $i \in \{1, \ldots, n\}$ and $\boldsymbol{z}_i = \boldsymbol{y}_{i-n}$ for $i \in \{n+1, \ldots, n+m\}$. Finally, let $\boldsymbol{K}_{\boldsymbol{Z}}$ be the kernel matrix consisting of all normalized pairwise kernel evaluations of the samples in $\boldsymbol{Z}$, that is, the samples from both distributions. Moreover, let $\boldsymbol{K}_{\boldsymbol{X}}$ and $\boldsymbol{K}_{\boldsymbol{Y}}$ be the pairwise kernel matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

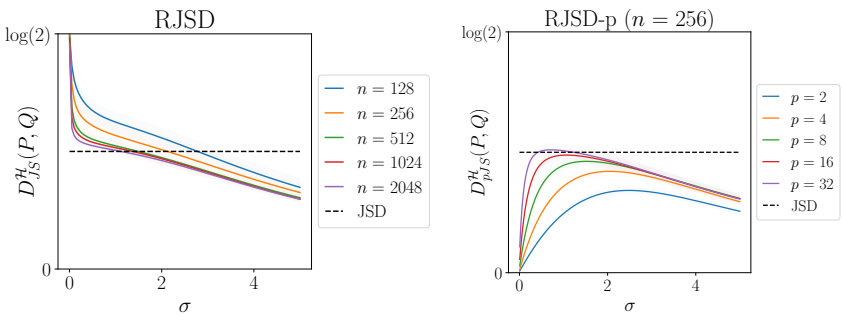
Notice that the sum of uncentered covariance operators in the RKHS corresponds to the covariance operator of the mixture of samples in the input space, that is, $\frac{n}{n+m}\boldsymbol{C}_{\boldsymbol{X}} + \frac{m}{n+m}\boldsymbol{C}_{\boldsymbol{Y}} = \boldsymbol{C}_{\boldsymbol{Z}}$.

Since $\boldsymbol{C}_{\boldsymbol{Z}}, \boldsymbol{C}_{\boldsymbol{X}}, \boldsymbol{C}_{\boldsymbol{Y}}$ and $\frac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}}, \frac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}, \frac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}}$ share the same non-zero eigenvalues respectively, the divergence can be directly computed from samples in the input space as follows.

Proposition 2. *The empirical kernel-based RJSD estimator for a kernel κ is*

$$\widehat{D}_{JS}^{\kappa}(\boldsymbol{X}, \boldsymbol{Y}) = S\left(\tfrac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}}\right) - \left(\tfrac{n}{n+m} S\left(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}\right) + \tfrac{m}{n+m} S\left(\tfrac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}}\right)\right). \tag{6}$$

This estimator, however, presents an upward bias that causes an undesired effect. The kernel RJSD estimator can be trivially maximized when the sample's similarities are negligible, for example, when the kernel bandwidth σ in a Gaussian kernel is close to zero (see Fig. 1(a)). This behavior is caused by the discrepancy between the number of samples used to estimate $S(\frac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}})$ compared to $S(\frac{1}{n}\boldsymbol{K}_{\boldsymbol{X}})$, and $S(\frac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}})$, which causes $S(\frac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}})$ to grow faster and up to $\log(n+m)$ compared to $S(\frac{1}{n}\boldsymbol{K}_{\boldsymbol{X}})$ and $S(\frac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}})$ that can only grow up to $\log(n)$ and $\log(m)$ respectively. To reduce the bias of the estimator in Eqn. 6 and avoid trivial maximization, we need to regularize $S(\frac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}})$ so that it estimates up to similar values of entropy than $S(\frac{1}{n}\boldsymbol{K}_{\boldsymbol{X}})$ and $S(\frac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}})$. We propose the following alternative:



(a) (b)

Figure 1: Comparing RJSD estimators with Gaussian kernel while varying the kernel bandwidth. The figure illustrates the estimated divergences between two Cauchy distributions ($d = 1$) with Jensen-Shannon divergence (JSD) $JSD = 0.5 \times \log(2)$.

Power Series Expansion Approximation: Let $\boldsymbol{A}$ be a positive semidefinite matrix, such that $\|\boldsymbol{A}\|_2 \leq 1$, where $\|\boldsymbol{A}\|_2 = \max_i(\lambda_i)$ denotes the spectral or L^2-norm, (which is the case for all trace-normalized kernel matrices). Then, the following power series expansion converges to $\log(\boldsymbol{A})$ [Higham, 2008]:

$$\log(\boldsymbol{A}) = -\sum_{j=1}^{\infty} \frac{(\boldsymbol{I} - \boldsymbol{A})^j}{j}.$$

We propose approximating the logarithm by truncating this series to a lower order.

Proposition 3. *The power-series kernel entropy estimator of X is:*

$$S_p(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}) = \sum_{j=1}^{p} \frac{1}{j} \operatorname{Tr}\left(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}} \left(\boldsymbol{I} - \tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}\right)^j\right), \tag{7}$$

where p is the order of the approximation.

Proposition 4. *The power-series RJSD estimator is*

$$\widehat{D}^{\kappa}_{pJS}(\boldsymbol{X}, \boldsymbol{Y}) = S_p\left(\tfrac{1}{n+m}\boldsymbol{K}_{\boldsymbol{Z}}\right) - \left(\tfrac{n}{n+m} S_p\left(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}\right) + \tfrac{m}{n+m} S_p\left(\tfrac{1}{m}\boldsymbol{K}_{\boldsymbol{Y}}\right)\right).$$

This approximation has two purposes. First, it avoids the need for eigenvalue decomposition. Second, it indirectly regularizes the three entropy terms of the divergence, where $\boldsymbol{K}_{\boldsymbol{Z}}$ is regularized more strongly due to its larger size. For example, $S_p(\boldsymbol{K}_{\boldsymbol{Z}}) \leq \sum_{j=1}^{p} \frac{1}{j}(1 - \frac{1}{n+m})^j$ while $S_p(\boldsymbol{K}_{\boldsymbol{X}}) \leq \sum_{j=1}^{p} \frac{1}{j}(1 - \frac{1}{n})^j$ and $S_p(\boldsymbol{K}_{\boldsymbol{Y}}) \leq \sum_{j=1}^{p} \frac{1}{j}(1 - \frac{1}{m})^j$.

By increasing the order, the gap between the maximum entropies obtained by the three entropy terms grows, leading to the behavior discussed above. Truncating the power series helps avoid trivial maximization of the divergence at lower kernel bandwidths (see Fig. 1(b)). Consequently, the RJSD power series expansion offers a more robust estimator that goes beyond reducing computational costs.

Next, we show an important connection between the power-series RJSD estimator and MMD:

Theorem 3. *Assume(**A1**) and let $p = 1$ be the order of the power series expansion approximation. Then, given two sets of samples $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \sim P$ and $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^n \sim Q$:*

$$\widehat{D}^{\kappa}_{pJS}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{4}\widehat{\mathrm{MMD}}^2_{\kappa^2}(\boldsymbol{X}, \boldsymbol{Y})$$

Proof: See Appendix A.2.

This theorem establishes that RJSD extends MMD to higher-order statistics of the kernel matrices and the covariance operator. While MMD captures second-order interactions of data projected in the reproducing kernel Hilbert space (RKHS) defined by the kernel function κ, RJSD incorporates higher-order statistics, enhancing the measures' sensitivity to subtle distributional differences.

4 Two-sample Testing with RJSD

We evaluate the discriminatory power of RJSD for two-sample testing. Given two sets of samples, $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ and $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^m$, drawn from P and Q respectively, two-sample testing aims to determine whether P and Q are identical. The null hypothesis H_0 states $P = Q$, while the alternative hypothesis H_1 states $P \neq Q$. A hypothesis test is then performed, rejecting the null hypothesis if $\mathbb{D}(P, Q) > \varepsilon$ for some distance or divergence $\mathbb{D}$ and threshold $\varepsilon > 0$.

Let $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{n+m} = \{\boldsymbol{x}_1, \dots \boldsymbol{x}_n, \boldsymbol{y}_1, \dots, \boldsymbol{y}_m\}$ be the combined sample. One common approach to perform two-sample testing is through permutation tests. These tests apply permutations of the combined data $\boldsymbol{Z}$ to approximate the distribution of the divergence measurement under the null hypothesis. Finally, this distribution determines the rejection threshold ε according to some specified significance level. In this experiment, we employ RJSD as the divergence measure to perform hypothesis testing.

Taking inspiration from 3 well-known MMD-based tests, we designed RJSD-based versions of MMD-Split [Sutherland et al., 2016], MMD-Deep [Liu et al., 2020], and MMD-Fuse [Biggs et al., 2024]. **RJSD-Split** involves splitting the data into training and testing sets to identify the optimal kernel bandwidth on the training set and subsequently evaluate performance on the testing set. Leveraging the lower bound in Eqn. 3, we propose selecting the kernel hyper-parameters that maximize RJSD as these parameters enhance the distinguishability between the two distributions [Sutherland et al., 2016]. Since the kernel-based estimator is not suitable for maximization with respect to the kernel hyperparameters, we use the power-series RJSD estimator.

Similarly, **RJSD-Deep** involves learning the parameters of the following characteristic kernel $\kappa_\theta(x, y)$:

$$\kappa_\theta(x, y) = \left[(1 - \epsilon)\kappa_1(f_\theta(x), f_\theta(y)) + \epsilon\right] \kappa_2(x, y),$$

where $f_\theta : \mathcal{X} \to \mathcal{F}$ represents a deep network that extracts features from the data, thereby enhancing the kernel's flexibility and its ability to capture the structure of complex distributions accurately. Here, $0 < \epsilon < 1$, and κ_1 and κ_2 are Gaussian kernels. Ultimately, we learn the network weights, the kernel bandwidths for κ_1 and κ_2, and the value of ϵ that maximizes RJSD.

On the other hand, **RJSD-Fuse** consists in combining the RJSD estimates of different kernels $\kappa \in \mathcal{K}$ drawn from a distribution $\rho \in \mathcal{M}_+^1(\mathcal{K})$. Then, these different values are passed through a weighted smooth maximum function that considers information from each kernel simultaneously, resulting in a new statistic. The fused statistic with parameter $\lambda > 0$ is defined as:

$$\widehat{\text{FUSE}}_{JS}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{\lambda} \log\left(\mathbb{E}_{\kappa\sim\rho}\left[\exp\left(\lambda \widehat{D}_{pJS}^{\kappa}(\boldsymbol{X}, \boldsymbol{Y})\right)\right]\right).$$

This method does not require data-splitting since the optimal kernel is chosen unsupervised through the log-sum-exponential function. See Appendix B.1 for implementation details.

5 Experiments and Results

We evaluate RJSD discriminatory power using one synthetic dataset and two real-world benchmark datasets for two-sample testing. The **Mixture of Gaussians** dataset [Biggs et al., 2024] consists of 2-dimensional mixtures of four Gaussians P and Q with means at $(\pm\mu, \pm\mu)$ and diagonal covariances. All components of P have unit variance, while only three components of Q have unit variance, with the standard deviation σ in the fourth component being varied. The null hypothesis $H_0 : P = Q$ corresponds to the case where $\sigma = 1$. The **Galaxy MNIST** dataset [Walmsley et al., 2022] consists of four categories of galaxy images captured by a ground-based telescope. P represents uniformly

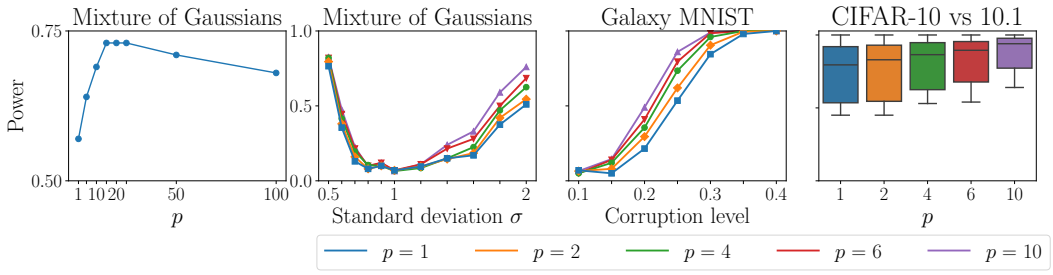


Figure 2: Test Power comparison for different orders of approximation. For the mixture of Gaussians and Galaxy MNIST, we deviate from the null hypothesis for a fixed number of samples of $n = m = 500$. For CIFAR-10 vs 10.1, we show the boxplot of the distribution of the average test power for different training sets.

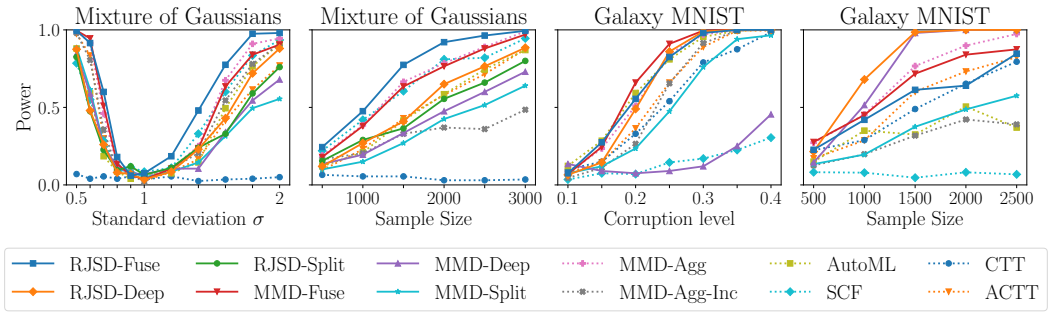


Figure 3: Test Power comparison different methods.

sampled images from the first three categories, while Q represents samples drawn from the first three categories with probability $1 - c$ and from the fourth category with probability $c \in [0, 1]$. We vary the corruption level c, with the null hypothesis corresponding to the case where $c = 0$. Finally, the **CIFAR 10 vs 10.1** dataset [Liu et al., 2020] compares the distribution P of the original CIFAR-10 dataset [Krizhevsky et al., 2009] with the distribution Q of CIFAR-10.1, which was collected as an alternative test set for models trained on CIFAR-10.

We compare the test power of RJSD-Split, RJSD-Deep, and RJSD-Fuse against various MMD-based tests: data splitting (MMD-Split)[Sutherland et al., 2016], Smooth Characteristic Functions (SCF) [Jitkrittum et al., 2016], the MMD Deep kernel (MMD-Deep) [Liu et al., 2020], Automated Machine Learning (AutoTST) [Kübler et al., 2022], kernel thinning to (Aggregate) Compress Then Test (CTT & ACTT)[Domingo-Enrich et al., 2023], and MMD Aggregated (Incomplete) tests (MMDAgg & MMDAggInc) [Schrab et al., 2023] and MMD-FUSE [Biggs et al., 2024].

5.1 Results

We first investigate the impact of increasing the approximation order p in the power-series expansion on test performance. Fig. 2 illustrates this effect across various datasets and scenarios. For the mixture of Gaussians with a fixed standard deviation $\sigma = 2$ and $n = m = 500$, we analyze the test power of RJSD-Split as p increases (leftmost). The results indicate a monotonic increase in test power up to a particular order, after which it declines. This pattern was consistently observed across different standard deviations. Similarly, for the Galaxy MNIST ($n = m = 500$) and CIFAR-10 vs. 10.1 ($n = m = 2021$) datasets, we evaluate RJSD-Deep with varying approximation orders. The trend was consistent across all scenarios, with higher-order approximations outperforming lower ones. Notably, $p = 10$ achieved the highest test power in each case. It is important to note that $p = 1$ corresponds to MMD, highlighting that RJSD consistently exhibits superior test power compared to MMD.

Table 1: Average test power for CIFAR-10 vs. CIFAR-10.1.

Tests	Power
RJSD-Fuse	**1.000**
MMD-Fuse	0.937
MMD-Agg	0.883
RJSD-Deep	<u>0.868</u>
MMD-Deep	0.744
CTT	0.711
ACTT	0.678
AutoML	0.544
MMD-Split	0.316
MMD-Agg-Inc	0.281
SCF	0.171

Bold: Best approach
<u>Underline:</u> Best data-splitting approach

Fig. 3 compares the test power of various approaches across the tested datasets. In most scenarios, RJSD-Fuse ($p = 10$) consistently outperforms or matches the performance of state-of-the-art methods like MMD-Fuse and MMD-Agg. Similarly, RJSD-Deep and RJSD-Split also demonstrate superior test power compared to their MMD counterparts in most cases. However, in the Galaxy MNIST dataset, when the sample size is increased, RJSD-Deep leads in performance, while RJSD-Fuse slightly falls behind MMD-Fuse. This discrepancy may be attributed to our estimator's lack of bias correction, which could affect certain cases.

Additionally, Table 1 presents the average power test for CIFAR-10 vs. CIFAR-10.1 computed over ten distinct training sets and 100 testing sets per training set (total of 1000 repetitions). Again, RJSD-Fuse ($p = 10$) achieves the highest test power, outperforming all other methods. Also, RJSD-Deep achieves the maximum power among data-splitting techniques, significantly surpassing MMD-Deep. These results highlight the robustness and efficacy of RJSD in measuring and detecting differences in distributions, demonstrating its potential as a powerful alternative to MMD for both statistical testing and broader machine-learning applications.

6 Conclusions

In this work, we introduced a novel kernel-based information-theoretic framework for two-sample testing, leveraging the representation Jensen-Shannon divergence (RJSD). We presented a method that extends beyond traditional MMD-based approaches by incorporating higher-order information from kernel matrices. Our framework offers a more robust and flexible measure of divergence between distributions without assuming Gaussianity. Moreover, we developed RJSD-based variants of well-known MMD tests, including MMD-Split, MMD-Deep, and MMD-Fuse, offering more flexible and powerful testing procedures.

Empirical results demonstrate the superior discriminative power of RJSD in two-sample testing tasks, positioning it as a robust alternative to MMD. RJSD's ability to capture more nuanced differences between distributions showcases its potential as a foundational tool for future machine learning research and applications. Given its versatility, ease of estimation from samples, and performance improvements, RJSD holds promise to significantly impact the field of kernel-based learning and contribute to advancing state-of-the-art methodologies in distribution comparison.

Further research is needed to analyze the bias and variance of the representation Jensen-Shannon divergence estimators under both null and alternative hypotheses. This analysis will offer important insights into the reliability of our methods for two-sample testing and lead to more principled approaches to bias correction in our estimators.

References

Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 2022.

Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, 36, 2024.

Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. *arXiv preprint arXiv:2301.05974*, 2023.

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

Mehrtash Harandi, Mathieu Salzmann, and Fatih Porikli. Bregman divergences for infinite dimensional covariance matrices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1003–1010, 2014.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Jhoan K Hoyos-Osorio, Santiago Posso-Murillo, and Luis G Sanchez-Giraldo. The representation jensen-shannon divergence. *arXiv preprint arXiv:2305.16446*, 2023.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Jonas M Kübler, Vincent Stimper, Simon Buchholz, Krikamol Muandet, and Bernhard Schölkopf. Automl two-sample test. *Advances in Neural Information Processing Systems*, 35:15929–15941, 2022.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

Hà Quang Minh. Affine-invariant riemannian distance between infinite-dimensional covariance operators. In *International Conference on Geometric Science of Information*, pages 30–38. Springer, 2015.

Hà Quang Minh. Regularized divergences between covariance operators and gaussian measures on hilbert spaces. *Journal of Theoretical Probability*, 34:580–643, 2021.

Ha Quang Minh. Entropic regularization of wasserstein distance between infinite-dimensional gaussian measures and gaussian processes. *Journal of Theoretical Probability*, 36(1):201–296, 2023.

Adil G. Naoum and Asma I. Gittan. A note on compact operators. *Publikacije Elektrotehničkog fakulteta. Serija Matematika*, (15):26–31, 2004. ISSN 03538893, 24060852. URL `http://www.jstor.org/stable/43666591`.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.

Wolfgang Stummer and Igor Vajda. On bregman distances and divergences of probability measures. *IEEE Transactions on Information Theory*, 58(3):1277–1288, 2012.

Danica J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

John Von Neumann. *Mathematical foundations of quantum mechanics: New edition*, volume 53. Princeton university press, 2018.

Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin 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.

Zhen Zhang, Mianzhi Wang, and Arye Nehorai. Optimal transport in reproducing kernel hilbert spaces: Theory and applications. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1741–1754, 2019.

A Appendix / supplemental material

A.1 Proof Lemma 1

Proof. To prove this Lemma, we use (Proposition 4.e) in Bach [2022]. We have that

$$D_{KL}\left(C_P, C_Q\right) \geq \frac{1}{2}\|C_P - C_Q\|_*^2 \geq \frac{1}{2}\|C_P - C_Q\|_{\text{HS}}^2,$$

where D_{KL} is the kernel Kullback-Leibler divergence and $\|\cdot\|_*$ and $\|\cdot\|_{HS}$ denote the nuclear and Hilbert-Schmidt norms respectively. Let $C_M = \frac{C_P + C_Q}{2}$, then:

$$\begin{aligned}
D_{JS}(C_P, C_Q) =& \frac{1}{2}D_{KL}(C_P, C_M) + \frac{1}{2}D_{KL}(C_P, C_M) \\
\geq& \frac{1}{4}\left\|C_P - \frac{1}{2}(C_P + C_Q)\right\|_*^2 + \frac{1}{4}\left\|C_Q - \frac{1}{2}(C_P + C_Q)\right\|_*^2 \\
\geq& \frac{1}{4}\left\|\frac{1}{2}C_P - \frac{1}{2}C_Q\right\|_*^2 + \frac{1}{4}\left\|\frac{1}{2}C_Q - \frac{1}{2}C_P\right\|_*^2 = \frac{1}{8}\left\|C_P - C_Q\right\|_*^2
\end{aligned}$$

and thus, $D_{JS}(C_P, C_Q) \geq \frac{1}{8}\left\|C_P - C_Q\right\|_*^2 \geq \frac{1}{8}\left\|C_P - C_Q\right\|_{\text{HS}}^2$.
Now, let $\phi : \mathcal{X} \mapsto \mathcal{H}$ then, and $\{e_\alpha\}$ be an orthonormal basis in $\mathcal{H}$, we have that

$$\begin{aligned}
\operatorname{Tr}\left(\phi(x) \otimes \phi(x)\phi(y) \otimes \phi(y)\right) =& \sum_\alpha \langle \phi(x) \otimes \phi(x)\phi(y) \otimes \phi(y) e_\alpha, e_\alpha\rangle \\
=& \sum_\alpha \langle \phi(x)\langle \phi(x), \phi(y) \otimes \phi(y) e_\alpha\rangle, e_\alpha\rangle \\
=& \sum_\alpha \langle \phi(x)\langle \phi(x), \phi(y)\langle \phi(y), e_\alpha\rangle\rangle, e_\alpha\rangle \\
=& \sum_\alpha \langle \phi(x)\langle \phi(x), \phi(y)\rangle\langle \phi(y), e_\alpha\rangle, e_\alpha\rangle \\
=& \sum_\alpha \langle \phi(x), e_\alpha\rangle\langle \phi(x), \phi(y)\rangle\langle \phi(y), e_\alpha\rangle \\
=& \langle \phi(x), \phi(y)\rangle \sum_\alpha \langle \phi(x), e_\alpha\rangle\langle \phi(y), e_\alpha\rangle = \langle \phi(x), \phi(y)\rangle\langle \phi(x), \phi(y)\rangle \\
=& \langle \phi(x), \phi(y)\rangle^2 = \kappa(x, y)^2
\end{aligned}$$

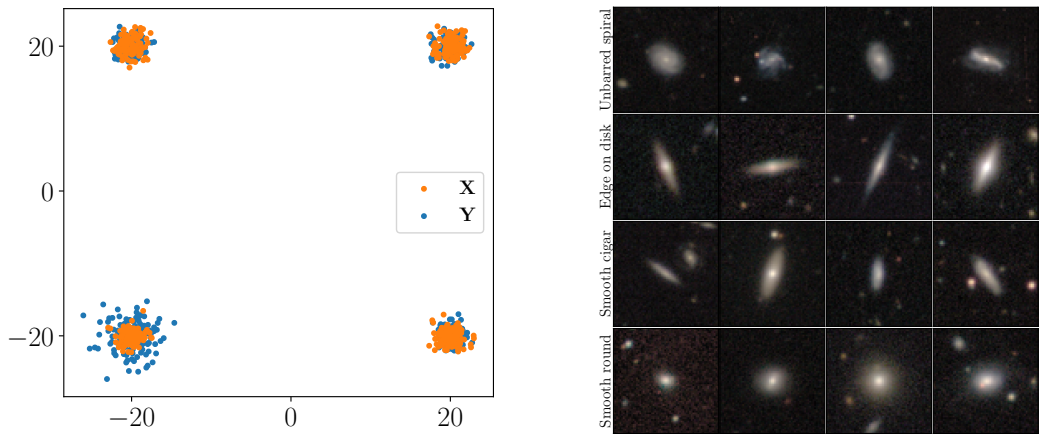
Note that for $T : \mathcal{H} \mapsto \mathcal{H}$, $\operatorname{Tr}(T^*T) = \sum_\alpha \langle Te_\alpha, Te_\alpha\rangle = \|T\|_{\text{HS}}^2$. In particular, if we have that $T = \phi(x) \otimes \phi(x) - \phi(y) \otimes \phi(y)$,

$$\begin{aligned}
\|\phi(x) \otimes \phi(x) - \phi(y) \otimes \phi(y)\|_{\text{HS}}^2 =& \operatorname{Tr}(\phi(x) \otimes \phi(x)\phi(x) \otimes \phi(x)) - 2\operatorname{Tr}(\phi(x) \otimes \phi(x)\phi(y) \otimes \phi(y)) \\
&+ \operatorname{Tr}(\phi(y) \otimes \phi(y)\phi(y) \otimes \phi(y)) \\
=& \kappa^2(x, x) - 2\kappa^2(x, y) + \kappa^2(y, y)
\end{aligned}$$

Finally, note that

$$\begin{aligned}
\|C_P - C_Q\|_{\text{HS}}^2 =& \operatorname{Tr}(\mathbb{E}_P[\phi(x) \otimes \phi(x)]\mathbb{E}_{P'}[\phi(x) \otimes \phi(x)]) - 2\operatorname{Tr}(\mathbb{E}_P[\phi(x) \otimes \phi(x)]\mathbb{E}_Q[\phi(y) \otimes \phi(y)]) \\
&+ \operatorname{Tr}(\mathbb{E}_Q[\phi(y) \otimes \phi(y)]\mathbb{E}_{Q'}[\phi(y) \otimes \phi(y)]) \\
=& \operatorname{Tr}(\mathbb{E}_{P,P'}[\phi(x) \otimes \phi(x)\phi(x') \otimes \phi(x')]) - 2\operatorname{Tr}(\mathbb{E}_{P,Q}[\phi(x) \otimes \phi(x)\phi(y) \otimes \phi(y)]) \\
&+ \operatorname{Tr}(\mathbb{E}_{Q,Q'}[\phi(y) \otimes \phi(y)\phi(y') \otimes \phi(y')]) \\
=& \mathbb{E}_{P,P'}[\kappa^2(x, x')] - 2\mathbb{E}_{P,Q}[\kappa^2(x, y)] + \mathbb{E}_{Q,Q'}[\kappa^2(y, y')],
\end{aligned}$$

which corresponds to squared MMD with kernel $\kappa^2(\cdot, \cdot)$. □



(a) Mixture of Gaussians (b) Galaxy MNIST

Figure 4: Mixture of Gaussians and Galaxy MNIST datasets.

A.2 Proof Theorem 3

Proof.

$$
\begin{aligned}
\widehat{D}^{\kappa}_{pJS}(\boldsymbol{X},\boldsymbol{Y}) &= \operatorname{Tr}\left(\tfrac{1}{2n}\boldsymbol{K}_{\boldsymbol{Z}}(\boldsymbol{I}-\tfrac{1}{2n}\boldsymbol{K}_{\boldsymbol{Z}})\right) - \frac{1}{2}\operatorname{Tr}\left(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}}(\boldsymbol{I}-\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{X}})\right) - \frac{1}{2}\operatorname{Tr}\left(\boldsymbol{I}-\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{Y}}(\tfrac{1}{n}\boldsymbol{K}_{\boldsymbol{Y}})\right) \\
&= -\operatorname{Tr}\left(\tfrac{1}{4n^2}\boldsymbol{K}_{\boldsymbol{Z}}\boldsymbol{K}_{\boldsymbol{Z}}\right) + \frac{1}{2}\operatorname{Tr}\left(\tfrac{1}{n^2}\boldsymbol{K}_{\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X}}\right) + \frac{1}{2}\operatorname{Tr}\left(\tfrac{1}{n^2}\boldsymbol{K}_{\boldsymbol{Y}}\boldsymbol{K}_{\boldsymbol{Y}}\right) \\
&= -\frac{1}{4n^2}\|\boldsymbol{K}_{\boldsymbol{Z}}\|_F^2 + \frac{1}{2n^2}\|\boldsymbol{K}_{\boldsymbol{X}}\|_F^2 + \frac{1}{2n^2}\|\boldsymbol{K}_{\boldsymbol{Y}}\|_F^2 \\
&= -\frac{1}{4n^2}\sum_{i,j}^{2n}\kappa^2(\boldsymbol{z}_i,\boldsymbol{z}_j) + \frac{1}{2n^2}\sum_{i,j}^{n}\kappa^2(\boldsymbol{x}_i,\boldsymbol{x}_j) + \frac{1}{2n^2}\sum_{i,j}^{n}\kappa^2(\boldsymbol{y}_i,\boldsymbol{y}_j) \\
&= \frac{1}{4n^2}\sum_{i,j}^{n}\kappa^2(\boldsymbol{x}_i,\boldsymbol{x}_j) + \frac{1}{4n^2}\sum_{i,j}^{n}\kappa^2(\boldsymbol{y}_i,\boldsymbol{y}_j) - \frac{2}{4n^2}\sum_{i,j}^{n}\kappa^2(\boldsymbol{x}_i,\boldsymbol{y}_j) \\
&= \frac{1}{4}\widehat{\mathrm{MMD}}^2_{\kappa^2}(\boldsymbol{X},\boldsymbol{Y})
\end{aligned}
$$

□

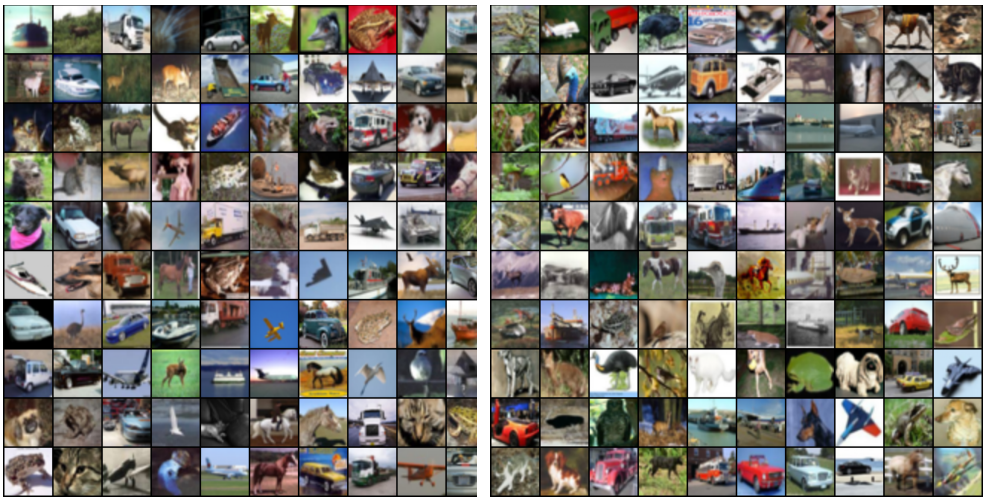
B Two-sample testing implementation details

B.1 RJSD-Fuse

Biggs et al. [2024] proposes MMD-Fuse, which computes a weighted smooth maximum of different MMD values from different kernels $\kappa \in \mathcal{K}$ drawn from a distribution $\rho \in \mathcal{M}_+^1(\mathcal{K})$. The proposed statistic is defined as:

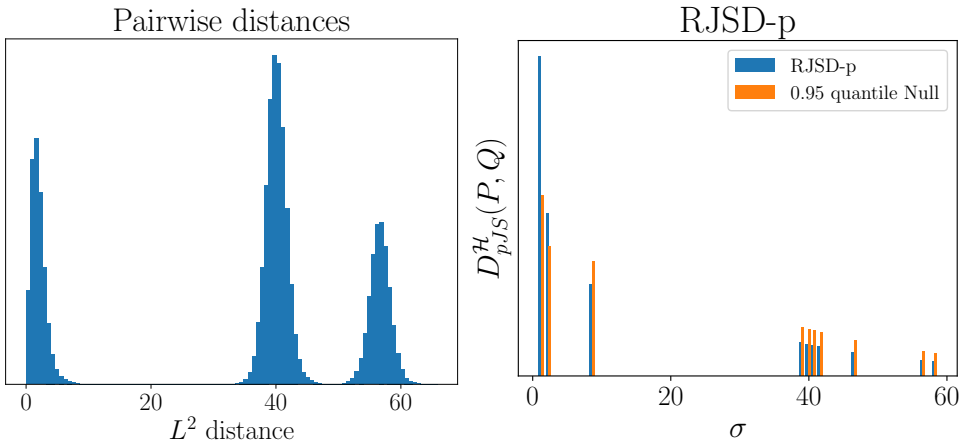
$$
\widehat{\mathrm{FUSE}}_{MMD}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{\lambda}\log\left(\mathbb{E}_{\kappa\sim\rho}\left[\exp\left(\lambda\frac{\widehat{\mathrm{MMD}}^2_{\kappa}(\boldsymbol{X},\boldsymbol{Y})}{N_{\kappa}(\boldsymbol{Z})}\right)\right]\right).
$$

Here, the different MMD estimates are normalized by a permutation invariant factor $N_\kappa(\boldsymbol{Z}) := \sqrt{\frac{1}{n\times(n-1)}\sum_{i\neq j}\kappa(z_i,z_j)^2}$ to account for the different scales and variances of distinct kernels before computing the "maximum". To include this term within our approach, instead of normalizing the divergence estimates, we normalize the kernels by $N_\kappa(\boldsymbol{Z})$, which in the case of $p=1$ is equivalent to MMD-Fuse. That is:



(a) CIFAR 10 (b) CIFAR-10.1

Figure 5: CIFAR 10 vs 10.1 images.



(a) (b)

Figure 6: L^2 distance distribution for the mixture of Gaussians and RJSD estimates for ten different bandwidths tested.

$$\hat{D}^{\mathcal{H}}_{pJS}(P,Q) = S_p\left(\frac{1}{n+m}\frac{\boldsymbol{K_Z}}{\sqrt{N_\kappa(\boldsymbol{Z})}}\right) - \left(\frac{n}{n+m}S_p\left(\frac{1}{n}\frac{\boldsymbol{K_X}}{\sqrt{N_\kappa(\boldsymbol{Z})}}\right) + \frac{m}{n+m}S_p\left(\frac{1}{m}\frac{\boldsymbol{K_Y}}{\sqrt{N_\kappa(\boldsymbol{Z})}}\right)\right).$$

Notice that for $p = 1$, this is equivalent to MMD-Fuse, where the measurement is normalized. However, normalizing the kernel allows the normalization to account for higher-order interactions between the kernel matrices for $p > 1$.

Distribution over kernels: Similarly to MMD-Fuse, we use a collection of Laplacian $\kappa^l_\sigma(x,x') = \exp\left(-\frac{\|x-x'\|_1}{\sigma}\right)$ and Gaussian $\kappa^g_\sigma(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$ kernels with distinct bandwidths $\sigma > 0$. In our implementation, we choose the bandwidths as the $5\%, 15\%, 25\%, \ldots 95\%$ quantiles of $\{\|z - z'\|_r : z, z' \in \boldsymbol{Z}\}$, with $r \in 1, 2$ for the Laplace and Gaussian kernels respectively. This choice is similar to MMD-Fuse, where ten bandwidths per kernel type are also selected. See Fig. 6.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The last paragraph of the conclusions discusses some of the current theoretical limitations of the method.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proofs of the proposed theoretical results are included in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details of the two samples testing experiments are in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be provided upon acceptance of the article.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Most implementation details are provided in the appendix. Other subtle details can be found in the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper presents the average power test over multiple randomized realizations of all the experiments, with a significance level of 0.05.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This information will be released upon acceptance in the provided code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.