# The Representation Jensen-Shannon Divergence

**Jhoan K. Hoyos-Osorio, Santiago Posso-Murillo & Luis G. Sanchez-Giraldo**
Department of Electrical and Computer Engineering
University of Kentucky
Lexington, USA
{keider.hoyos,spo,luis.sanchez}@uky.edu

## Abstract

Statistical divergences quantify the difference between probability distributions, thereby allowing for multiple uses in machine-learning. However, a fundamental challenge of these quantities is their estimation from empirical samples since the underlying distributions of the data are usually unknown. In this work, we propose a divergence inspired by the Jensen-Shannon divergence which avoids the estimation of the probability density functions. Our approach embeds the data in an reproducing kernel Hilbert space (RKHS) where we associate data distributions with uncentered covariance operators in this representation space. Therefore, we name this measure the representation Jensen-Shannon divergence (RJSD). We provide an estimator from empirical covariance matrices by explicitly mapping the data to an RKHS using Fourier features. This estimator is flexible, scalable, differentiable, and suitable for minibatch-based optimization problems. Additionally, we provide an estimator based on kernel matrices without an explicit mapping to the RKHS. We provide consistency convergence results for the proposed estimator. Moreover, we demonstrate that this quantity is a lower bound on the Jensen-Shannon divergence, leading to a variational approach to estimate it with theoretical guarantees. We leverage the proposed divergence to train generative networks, where our method mitigates mode collapse and encourages samples diversity. Additionally, RJSD surpasses other state-of-the-art techniques in multiple two-sample testing problems, demonstrating superior performance and reliability in discriminating between distributions.

## 1 Introduction

Divergences quantify the difference between probability distributions. In machine-learning, divergences can be applied to a wide range of tasks, including generative modeling (generative adversarial networks, variational auto-encoders), two-sample testing, anomaly detection, and distribution shift detection. The family of $f$-divergences is among the most popular statistical divergences, including the well-known Kullback-Leibler and Jensen-Shannon divergences. A fundamental challenge to using divergences in practice is that the underlying distribution of data is unknown, and thus divergences must be estimated from observations. Several divergence estimators have been proposed (Yang and Barron, 1999; Sriperumbudur et al., 2012; Krishnamurthy et al., 2014; Moon and Hero, 2014; Singh and Póczos, 2014; Li and Turner, 2016; Noshad et al., 2017; Moon et al., 2018; Bu et al., 2018; Berrett and Samworth, 2019; Liang, 2019; Han et al., 2020; Sreekumar and Goldfeld, 2022), most of which fall into four categories: plug-in, kernel density estimation, $k$-nearest neighbors, and neural estimators.

Kernel methods are another approach for measuring the interaction between probability distributions. For example, the maximum mean discrepancy (MMD) (Gretton et al., 2012) is a divergence com-

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

puted as the distance between the mean embeddings (first-order moments) of the two probability distributions in a reproducing kernel Hilbert space (RKHS). However, due to the underlying geometry, MMD lacks a straightforward connection with classical information theory tools (Bach, 2022). On the other hand, covariance operators (second-order moments) in RKHS have been used to propose multiple information theoretic quantities, such as marginal, joint, and conditional entropy (Sanchez Giraldo et al., 2014), as well as mutual information (Yu et al., 2019), and total correlation (Yu et al., 2021). However, strategies for estimating divergences within this framework have been less explored.

To fill this void, we propose a kernel-based information theoretic learning framework for divergence estimation. We make the following contributions:

- A novel divergence, the representation Jensen-Shannon divergence (RJSD), that avoids the estimation of the underlying density functions by mapping the data to an RKHS where distributions can be embedded using uncentered covariance operators acting in this representation space.
- An estimator from empirical covariance matrices that explicitly map data samples to an RKHS using Fourier features. This estimator is flexible, scalable, differentiable, and suitable for minibatch-based optimization problems. Additionally, an estimator based on kernel matrices without an explicit mapping to the RKHS is provided. Consistency results and sample complexity bounds for the proposed estimator are derived.
- A connection between the kernel-based entropy and Shannon's entropy, as well as the relationship between RJSD with the classical Jensen-Shannon divergence. Namely, RJSD emerges as a lower bound on the classical Jensen-Shannon divergence enabling the construction of a variational estimator for the classical Jensen-Shannon divergence with statistical guarantees.

We use RJSD for training generative adversarial networks and show that it prevents mode collapse and encourages diversity, leading to more accurate and heterogeneous results. We also apply RJSD for two-sample testing problems and show that it accurately detects differences between probability distribution functions even for cases where other state-of-the-art measures fall short.

## 2 Background

### 2.1 Covariance operators

Let $\{\mathcal{X}, \mathbf{B}_\mathcal{X}\}$ be a measurable space and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a positive definite kernel. There exists a mapping $\phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_\mathcal{H}$. A probability distribution $\mathbb{P}$ can be mapped to a covariance operator $C_\mathbb{P} : \mathcal{H} \to \mathcal{H}$. The covariance operator is defined as follows:

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

Similarly, for any $f, g \in \mathcal{H}$, $\mathbb{E}_{X \sim \mathbb{P}}[f(X)g(X)] = \langle g, C_\mathbb{P} f \rangle_\mathcal{H}$. For a bounded kernel, the covariance operator is positive semi-definite, Hermitian (self-adjoint), and trace class (Sanchez Giraldo et al., 2014; Bach, 2022). The spectrum of the covariance is discrete and consists of non-negative eigenvalues $\lambda_i$ with $\sum \lambda_i < \infty$.

### 2.2 Kernel-based information theory

We can define information theoretic quantities on the spectrum of normalized covariance operators with unit trace. This observation was made by Sanchez Giraldo et al. (2014) who proposed the kernel-based entropy functional: $S_\alpha(C_\mathbb{P}) = \frac{1}{1-\alpha} \log \left[\mathrm{Tr}(C_\mathbb{P}^\alpha)\right]$. $\mathrm{Tr}(\cdot)$ denotes the trace operator, and $\alpha > 0$ is the entropy order. In the limit when $\alpha \to 1$, $S_{\alpha \to 1}(C_\mathbb{P}) = -\mathrm{Tr}\left(C_\mathbb{P} \log C_\mathbb{P}\right)$ becomes von Neumann entropy.

**Kernel-based entropy estimator:** The kernel-based entropy estimator relies on the spectrum of the empirical uncentered covariance operator, which is defined as $\boldsymbol{C}_X = \frac{1}{N} \sum_i^N \phi(x_i) \otimes \phi(x_i)$. Let $\mathbf{X} = \{x_i\}_{i=1}^N$ be a set of samples $x \in \mathcal{X}^d$ following an unknown distribution $\mathbb{P}$ defined on $\mathcal{X}^d$. Then, we can construct the Gram matrix $\mathbf{K}_X$, consisting of all normalized pairwise kernel evaluations of the samples in $\mathbf{X}$, that is $(\mathbf{K})_{ij} = \kappa(x_i, x_j)$. If we normalize the matrix $\mathbf{K}_X$ such that, $\mathrm{Tr}\left(\mathbf{K}_X\right) = 1$,

$\boldsymbol{C}_X$ and $\mathbf{K}_X$ have the same non-zero eigenvalues (Sanchez Giraldo et al., 2014). This construction yields the kernel-based entropy estimator:

$$S\left(\mathbf{K}_X\right) = -\operatorname{Tr}\left(\mathbf{K}_X \log \mathbf{K}_X\right) = -\sum_{i=1}^{N} \lambda_i \log \lambda_i, \tag{2}$$

where $\lambda_i$ represents the $i$th eigenvalue of $\mathbf{K}_X$.

**Covariance-based estimator:** Alternatively, we can use an explicit mapping $\phi : \mathcal{X}^d \to \mathcal{H}^D$ to a finite dimensional RKHS. We propose to use Fourier features to construct a mapping function to $\mathcal{H}^D$. Given a shift-invariant kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, the random Fourier features (RFF) (Rahimi and Recht, 2007) is a method to create a smooth feature mapping $\phi_\omega(x) : \mathcal{X}^d \to \mathbb{R}^D$ so that $\kappa(x, x') \approx \langle \phi_\omega(x), \phi_\omega(x') \rangle$. To generate an RFF mapping, we draw $\frac{D}{2}$ i.i.d samples $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{D/2} \in \mathbb{R}^d$. Finally, the mapping is given by $\phi_\omega(\boldsymbol{x}) = \sqrt{\frac{2}{D}}\Big[\cos(\boldsymbol{\omega}_1^\top \boldsymbol{x}), \sin(\boldsymbol{\omega}_1^\top \boldsymbol{x}), \cdots, \cos(\boldsymbol{\omega}_{D/2}^\top \boldsymbol{x}), \sin(\boldsymbol{\omega}_{D/2}^\top \boldsymbol{x})\Big]$.

Letting $\boldsymbol{\Phi}_X \in \mathbb{R}^{N \times D}$ be the matrix containing the mapped samples, we can compute the empirical uncentered covariance matrix as $\boldsymbol{C}_X = \frac{1}{N} \boldsymbol{\Phi}_X^\top \boldsymbol{\Phi}_X$. Finally, we exploit the spectrum of the uncentered covariance matrix to compute the von Neumann entropy of $\boldsymbol{C}_X$ as:

$$S\left(\mathbf{C}_X\right) = -\operatorname{Tr}\left(\mathbf{C}_X \log \mathbf{C}_X\right) = -\sum_{i=1}^{D} \lambda_i \log \lambda_i, \tag{3}$$

where $\lambda_i$ represents the $i$th eigenvalue of $\mathbf{C}_X$.

The kernel-based entropy has been used as a building block for other matrix-based measures, such as joint and conditional entropy, mutual information (Yu et al., 2019), total correlation (Yu et al., 2021), and divergence (Hoyos Osorio et al., 2022). Despite the success of the aforementioned measures, their connection with the classical information theory counterparts remains unclear.

Next, we investigate the relationship between the kernel-based entropy estimator and Shanonn's entropy.

**Definition 1.** *Let $\phi : \mathcal{X} \to \mathcal{H}$ be a mapping to a reproducing kernel Hilbert space (RKHS), and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a positive definite kernel, such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$, and $\kappa(x, x) = 1$ for all $x \in \mathcal{X}$. Then, the **kernel density function** induced by the mapping $\phi$ is defined as follows:*

$$\mathbb{P}_\phi(x) = \frac{1}{h} \langle \phi(x) | C_{\mathbb{P}} | \phi(x) \rangle = \frac{1}{h} \int_{\mathcal{X}} \kappa^2(x, x') d\mathbb{P}(x'), \tag{4}$$

*where $h = \int_{\mathcal{X}} \langle \phi(x) | C_{\mathbb{P}} | \phi(x) \rangle \, dx$ is the normalizing constant.*

Eqn. 4 can be interpreted as an instance of the Born rule which calculates the probability of finding a state $\phi(x)$ in a system described by the covariance operator $C_{\mathbb{P}}$ (González et al., 2022). Equivalently, the right hand side corresponds to a Parzen density estimator with kernel $\kappa^2(\cdot, \cdot)$.

**Theorem 1.** *Let $\mathbb{P}_\phi(x)$ be the kernel density function induced by a mapping $\phi : \mathcal{X} \to \mathcal{H}$, then, the cross entropy between $\mathbb{P}$ and $\mathbb{P}_\phi$ is:*

$$H(\mathbb{P}, \mathbb{P}_\phi) = S(C_{\mathbb{P}}) + \log(h). \tag{5}$$

*Proof:* See Appendix A.1.

This result relates to kernel-density estimation for entropy calculation; however, the covariance operator bypasses the estimation of the underlying distribution.

## 3 Representation Jensen-Shannon Divergence

For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a measurable space $\{\mathcal{X}, \mathbf{B}_{\mathcal{X}}\}$, the Jensen-Shannon divergence (JSD) is defined as follows:

$$D_{JS}(\mathbb{P}, \mathbb{Q}) = H\left(\frac{\mathbb{P} + \mathbb{Q}}{2}\right) - \frac{1}{2}\left(H(\mathbb{P}) + H(\mathbb{Q})\right), \tag{6}$$

where $\frac{\mathbb{P}+\mathbb{Q}}{2}$ is the mixture of both distributions and $H(\cdot)$ is Shannon's entropy. The Quantum counterpart of the Jensen-Shannon divergence (QJSD) between density matrices $\rho$ and $\sigma$ is defined as $D_{JS}(\rho,\sigma) = S\left(\frac{\rho+\sigma}{2}\right) - \frac{1}{2}\left(S(\rho)+S(\sigma)\right)$, where $S(\cdot)$ is von Neumann's entropy. Similar to the kernel-based entropy, we let the covariance operators play the role of the density matrices to derive a measure of divergence that can be computed directly from data samples.

**Definition 2.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures defined on a measurable space $\{\mathcal{X}, \mathbf{B}_{\mathcal{X}}\}$, and let $\phi : \mathcal{X} \to \mathcal{H}$ be a mapping to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, such that $\langle \phi(x), \phi(x)\rangle_{\mathcal{H}} = 1$ for all $x \in \mathcal{X}$. Then, the **representation Jensen-Shannon divergence** (RJSD) between uncentered covariance operators $C_{\mathbb{P}}$ and $C_{\mathbb{Q}}$ is defined as:*

$$D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) = S\left(\frac{C_{\mathbb{P}} + C_{\mathbb{Q}}}{2}\right) - \frac{1}{2}\left(S(C_{\mathbb{P}}) + S(C_{\mathbb{Q}})\right). \tag{7}$$

### 3.1 Theoretical Properties

RJSD inherits most of the properties of classical and quantum Jensen-Shannon divergence. *Non-negativity*: $D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) \geq 0$. *Positivity:* $D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) = 0$ if and only if $C_{\mathbb{P}} = C_{\mathbb{Q}}$. *Symmetry:* $D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) = D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}})$. *Boundedness:* $D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) \leq \log(2)$. Also, $D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}})^{\frac{1}{2}}$ is a metric on the cone of uncentered covariance matrices in any dimension (Virosztek, 2021).

Below, we introduce key properties of RJSD and the connection with its classical counterpart.

**Theorem 2.** *For all probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on $\mathcal{X}$, and covariance operators $C_{\mathbb{P}}$ and $C_{\mathbb{Q}}$ with RKHS mapping $\phi(\cdot)$ under the conditions of Definition 2, the following inequality holds:*

$$D_{JS}^{\phi}(C_{\mathbb{P}}, C_{\mathbb{Q}}) \leq D_{JS}(\mathbb{P}, \mathbb{Q}) \tag{8}$$

*Proof:* See Appendix A.2.

**Theorem 3.** *let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures defined on $\mathcal{X}$. If there exists a mapping $\phi^*$ such that $\mathbb{P}(x) = \frac{1}{h_{\mathbb{P}}}\langle \phi^*(x)|C_{\mathbb{P}}|\phi^*(x)\rangle$ and $\mathbb{Q}(x) = \frac{1}{h_{\mathbb{Q}}}\langle \phi^*(x)|C_{\mathbb{Q}}|\phi^*(x)\rangle$, then:*

$$D_{JS}(\mathbb{P}, \mathbb{Q}) = D_{JS}^{\phi^*}(C_{\mathbb{P}}, C_{\mathbb{Q}}). \tag{9}$$

*Proof:* See Appendix A.3.

This theorem implies that the bound in Eqn. 8 is tight for optimal functions $\phi(x)$ that approximate the true underlying distributions through Eqn. 4.

Finally, we show that RJSD relates to MMD with kernel $\kappa^2(\cdot,\cdot)$.

**Theorem 4.** *For all probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on $\mathcal{X}$, and covariance operators $C_{\mathbb{P}}$ and $C_{\mathbb{Q}}$ with RKHS mapping $\phi(x)$ such that $\forall x \in \mathcal{X}, \langle \phi(x), \phi(x)\rangle_{\mathcal{H}} = 1$:*

$$D_{JS}(C_{\mathbb{P}}, C_{\mathbb{Q}}) \geq \frac{1}{8}\|C_{\mathbb{P}} - C_{\mathbb{Q}}\|_*^2 \geq \frac{1}{8}\|C_{\mathbb{P}} - C_{\mathbb{Q}}\|_{HS}^2 = \frac{1}{8}\,\mathrm{MMD}_{\kappa^2}(\mathbb{P}, \mathbb{Q}) \tag{10}$$

*Proof:* See Appendix A.4.

From this result we should expect RJSD to be at least as efficient as MMD for identifying discrepancies between distributions and that for a characteristic kernel $\kappa$, RJSD to be non zero if $\mathbb{P} \neq \mathbb{Q}$.

### 3.2 Estimating the representation Jensen-Shannon divergence

Given two sets of samples $\mathbf{X} = \{x_i\}_{i=1}^N \subset \mathcal{X}^d$ and $\mathbf{Y} = \{y_i\}_{i=1}^M \subset \mathcal{X}^d$ with unknown distributions $\mathbb{P}$ and $\mathbb{Q}$, we propose two estimators of RJSD.

**Covariance-based estimator:** We propose to use Fourier features to construct a mapping function $\phi_\omega : \mathcal{X}^d \to \mathcal{H}^D$ to a finite dimensional RKHS as explained in Section 2.2. Let $\boldsymbol{\Phi}_X \in \mathbb{R}^{N\times D}$ and $\boldsymbol{\Phi}_Y \in \mathbb{R}^{M\times D}$ be the matrices containing the mapped samples of each distribution. Then, the empirical uncentered covariance matrices are computed as $\boldsymbol{C}_X = \frac{1}{N}\boldsymbol{\Phi}_X^\top\boldsymbol{\Phi}_X$ and $\boldsymbol{C}_Y = \frac{1}{M}\boldsymbol{\Phi}_Y^\top\boldsymbol{\Phi}_Y$. Finally, the covariance-based RJSD estimator is defined as:

$$D_{JS}^{\omega}(\boldsymbol{C}_X, \boldsymbol{C}_Y) = S\left(\pi_1\boldsymbol{C}_X + \pi_2\boldsymbol{C}_Y\right) - \left(\pi_1 S(\boldsymbol{C}_X) + \pi_2 S(\boldsymbol{C}_Y)\right), \tag{11}$$

where $\pi_1 = \frac{N}{N+M}$ and $\pi_2 = \frac{M}{N+M}$ are the sample proportions of each distribution (e.g. $\frac{1}{2}$ if the samples are balanced). Finally, we use Eqn. 3 to estimate the entropies of the covariance matrices.

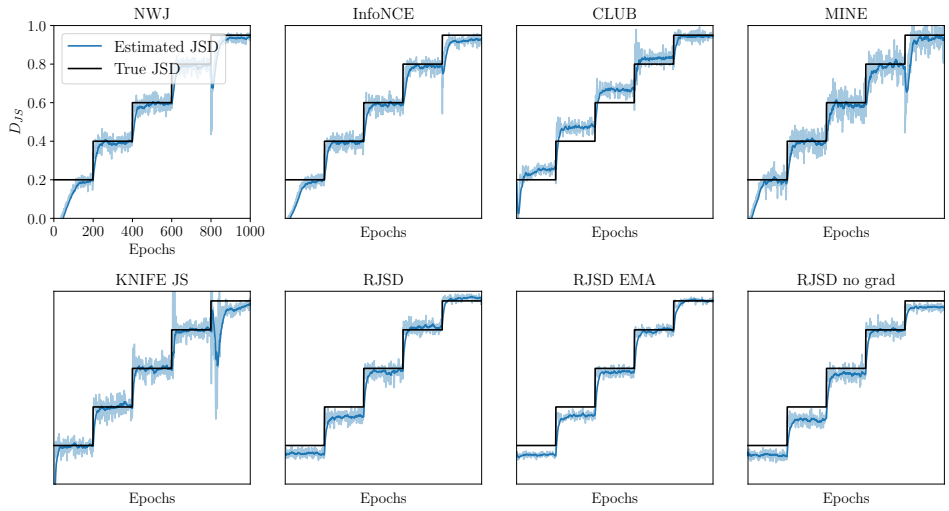



Figure 1: Jensen-Shannon Divergence estimation for two set of samples following Cauchy distributions (N = 512). We compare the following estimators: NWJ (Nguyen et al., 2010), infoNCE (Oord et al., 2018), CLUB (Cheng et al., 2020), MINE (Belghazi et al., 2018), KNIFE (Pichler et al., 2022), RJSD, RJSD with EMA, RJSD for a fixed kernel.

**Kernel-based estimator:** Here, we propose an estimator of RJSD from kernel matrices without an explicit mapping to the RKHS.

**Lemma 1.** *Let* $\mathbf{Z}$ *be the mixture of the samples of* $\mathbf{X}$ *and* $\mathbf{Y}$*, that is,* $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N+M}$ *where* $\mathbf{z}_i = \mathbf{x}_i$ *for* $i \in \{1, \dots, N\}$ *and* $\mathbf{z}_i = \mathbf{y}_{i-N}$ *for* $i \in \{N+1, \dots, N+M\}$*. Also, let* $\mathbf{K}_Z$ *be the kernel matrix consisting of all normalized pairwise kernel evaluations of the samples in* $\mathbf{Z}$*, then* $S\left(\pi_1 \boldsymbol{C}_X + \pi_2 \boldsymbol{C}_Y\right) = S(\mathbf{K}_Z)$.

Since the spectrum of $\mathbf{K}_X$ and $\boldsymbol{C}_X$ have the same non-zero eigenvalues, likewise $\mathbf{K}_Y$ and $\boldsymbol{C}_Y$, the divergence can be directly computed from samples in the input space as:

$$D_{JS}^{\kappa}(\mathbf{X}, \mathbf{Y}) = S\left(\mathbf{K}_Z\right) - \left(\pi_1 S(\mathbf{K}_X) + \pi_2 S(\mathbf{K}_Y)\right) \tag{12}$$

## 4 Variational Estimation of classical Jensen-Shannon divergence

We exploit the lower bound in Theorem 2 to derive a variational method for estimating the classical Jensen-Shannon divergence (JSD) given only samples from $\mathbb{P}$ and $\mathbb{Q}$. Accordingly, we choose $\Phi$ to be the family of functions $\phi_\omega : \mathcal{X}^d \to \mathcal{H}^D$ parameterized by $\omega \in \Omega$. Here, we aim to optimize the Fourier features to maximize the lower bound in Eqn. 2. Notice that we can also use a neural network $f_\omega$ with a Fourier features mapping $\phi_\omega$ in the last layer, that is, $\phi_\omega \circ f_\omega = \phi_\omega(f_\omega(x))$. We call this network a *Fourier-features network (FFN)*. Finally, we can compute the divergence based on this representation, leading to a neural estimator of classical JSD.

**Definition 3.** *(Jensen-Shannon divergence variational estimator). Let* $\Phi = \{\phi_\omega \circ f_\omega\}_{\omega \in \Omega}$ *be the set of functions parameterized by a FFN. We define our JSD variational estimator as:*

$$\widehat{D_{JS}}(\mathbb{P}, \mathbb{Q}) = \sup_{\omega \in \Omega} D_{JS}^{\omega}(C_{\mathbb{P}}, C_{\mathbb{Q}}). \tag{13}$$

## 5 Experiments

### 5.1 Variational Jensen-Shannon divergence estimation

First, we evaluate the performance of our variational estimator of Jensen-Shannon divergence (JSD) in a tractable toy experiment. Here, $\mathbb{P} \sim p(x; l_p, s_p)$ and $\mathbb{Q} \sim p(x; l_q, s_q)$ are two Cauchy distributions with location parameters $l_p$ and $l_q$ and scale parameters $s_p = s_q = 1$. We vary the location parameter of $\mathbb{Q}$ over time to increase the divergence. (see Appendix B.1 for more details). Then, we estimate

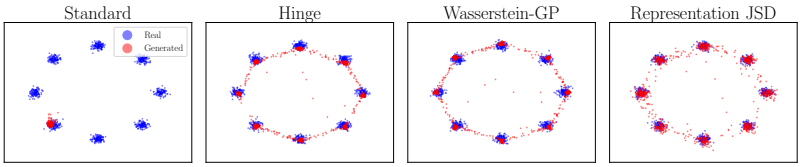



Figure 2: GANs with different loss functions to evaluate mode collapse in eightGaussians dataset. RJSD improves mode coverage and sample diversity.

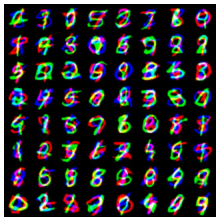


Figure 3: Generated samples using rep JSD.

JSD drawing $N = 512$ samples from both distributions at every epoch. We compare the estimates of divergence against different neural estimators. JSD corresponds to the mutual information between the mixture distribution and a Bernoulli distribution indicating when a sample is drawn from $\mathbb{P}$ or $\mathbb{Q}$. Therefore, we use mutual information estimators to approach the JSD estimation, such as NWJ (Nguyen et al., 2010), infoNCE (Oord et al., 2018), CLUB (Cheng et al., 2020), MINE (Belghazi et al., 2018). We also employ KNIFE (Pichler et al., 2022) to estimate the entropy terms and compute JSD.

Fig. 1 shows the estimation results. All compared methods approximate JSD; however, some of them struggle to adapt to distribution changes. These abrupt adjustments could lead to instabilities during training. In contrast to the compared methods, the RJSD estimator accurately estimates divergence with a lower variance, adjusting itself smoothly to changes in the distributions. Additionally, by using Exponential Moving averages (EMA) of the covariance matrices, the estimation variance decreases further yielding a smoother estimation. Finally, we compute RJSD for a fixed set of Fourier features without any optimization (no gradients backpropagated), and we can observe that RJSD still approximates the true divergence. This result agrees with theorem 3 suggesting that the computed kernel implicitly approximates the underlying distributions of the data.

### 5.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a family of models to generate images/audio by minimizing the divergence between the generated and the real data distributions (Farnia and Ozdaglar, 2020).

Below, we propose a methodology for training GANs using RJSD in the objective function. The RJSD-GAN is formulated as follows:

$$\min_{\theta\in\Theta}\max_{\omega\in\Omega} D^{\omega}_{JS}(\mathbf{X}, \mathbf{Y}^{\theta}), \tag{14}$$

where $\mathbf{X}$ are samples from the real data, and $\mathbf{Y}^{\theta}$ are samples created by a generator $G_{\theta}$. Instead of classifying real and fake samples, we use a *Fourier-features network* $\{\phi_{\omega} \circ f_{\omega}\}_{\omega\in\Omega}$ (FFN, see Section 4) to learn a multidimensional representation in an RKHS where the divergence is maximized. Subsequently, the generator $\{G_{\theta}\}_{\theta\in\Theta}$ attempts to minimize RJSD. We follow a single-step alternating gradient method. We assess our GAN formulation in two well-known mode-collapse experiments: eight Gaussians dataset and stacked MNIST.

#### 5.2.1 Synthetic experiments

We apply RJSD to train a GAN in a synthetic experiment. The target distribution is a mixture of eight Gaussian distributions arranged in a circle. Fig. 2 shows the real data and the samples generated by various learning functions to train GANs. As expected, the standard (vanilla) GAN fails to generate samples from all modes (Fig. 2(a)). The Hinge (Lim and Ye, 2017) and Wasserstein-GP GANs (Gulrajani et al., 2017) successfully produce samples representing all eight modes, yet Figs. 2(b) and 2(c) exhibit generated samples with reduced variance/diversity (lower entropy) within each mode: a phenomenon termed intra-class collapse. As we observe, the generated samples fail to cover the entire support of each Gaussian mode clustering towards the center. In contrast to the compared methods, the samples generated by the RJSD-GAN show improved mode coverage and higher diversity. This is visually noticeable in Fig. 2(d). Additionally, we perform the following quantitative analysis. We cluster the eight modes generated by each method and estimate their mean and covariance matrices.

Then, we calculate the Kullback-Leibler (KL) divergence between the real Gaussian modes and their generated counterparts. Finally, we average the divergence among the eight modes. Table 1 highlights the superiority of RJSD in terms of KL divergence when contrasted with the baseline methods. This empirical evidence supports the efficacy of RJSD to avoid mode collapse and to generate samples matching the target distribution beyond visual comparability.

### 5.2.2 Stacked MNIST

We conduct a quantitative evaluation to assess the efficacy of RJSD in reducing mode collapse on the stacked MNIST dataset. This dataset consists of three randomly sampled MNIST digits stacked along different color channels. This procedure results in 1000 possible classes (modes) corresponding to all combinations of the 10 digits. We use the standard DCGAN generator architecture (Radford et al., 2015), and modify the discriminator architecture to include a Fourier-features mapping (see implementation details in Appendix B.2.2). We compare our method against a considerable number of GAN algorithms using the same generator and following the same evaluation protocol. We utilize a pre-trained classifier to quantify the number of distinct generated modes. Additionally, we calculate the Kullback-Leibler (KL) divergence between the distribution of the generated modes and the real mode distribution. Finally, we average the results over five runs. Table 2 shows the results, and RJSD captures all modes and steadily generates samples from all classes achieving the lowest KL-divergence compared to the baseline approaches. Although our algorithm is a standard GAN that explicitly minimizes the Jensen-Shannon divergence, RJSD does not require the incorporation of entropy regularizers or mode-collapse prevention mechanisms beyond the learning function itself.

Table 1: KL divergence between real and generated distributions on eightmodes dataset.

| Average KL divergence | | |
|---|---|---|
| RJSD | Wasserstein-GP | Hinge |
| $\mathbf{0.699 \pm 0.245}$ | $0.981 \pm 0.701$ | $1.623 \pm 1.000$ |

Table 2: Number of modes and KL divergence between real and generated distributions on stacked MNIST.

| | Modes (Max 1000) | KL |
|---|---|---|
| DCGAN (Radford et al., 2015) | 99.0 | 3.40 |
| ALI (Dumoulin et al., 2016) | 16.0 | 5.40 |
| Unrolled GAN (Metz et al., 2016) | 48.7 | 4.32 |
| VEEGAN (Srivastava et al., 2017) | 150 | 2.95 |
| WGAN-GP (Gulrajani et al., 2017) | 959.0 | 0.72 |
| PresGAN (Dieng et al., 2019) | $999.6 \pm 0.4$ | $0.11 \pm 7.0\mathrm{e}{-2}$ |
| PacGAN (Lin et al., 2018) | $1000.0 \pm 0$ | $0.06 \pm 1.0\mathrm{e}{-2}$ |
| GAN+MINE (Belghazi et al., 2018) | $1000.0 \pm 0$ | $0.05 \pm 6.9\mathrm{e}{-3}$ |
| **GAN + rep JSD** | $1000.0 \pm 0$ | $\mathbf{0.04 \pm 1.2\mathrm{e}{-3}}$ |

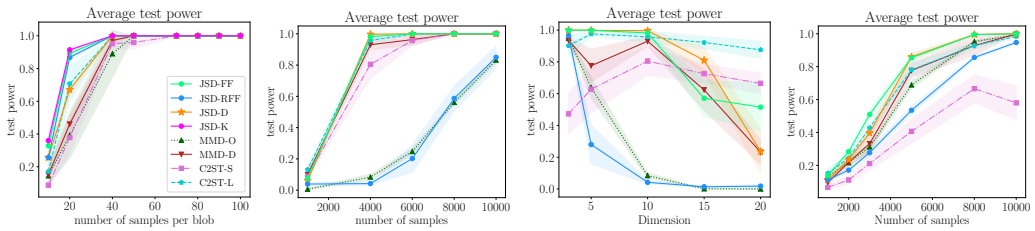
### 5.3 Two sample testing

We evaluate the performance of the proposed divergence for two-sample testing on different datasets and compare it against different state-of-the-art (SOTA) methods. We perform the following tests: (a) RJSD-FF: Two-sample test based on RJSD, optimizing the Fourier features applied to the input data. (b) RJSD-RFF: Two-sample test based on RJSD using random Fourier features, optimizing just the length-scale of the associated Gaussian kernel. (c) RJSD-D: Two-sample test based on RJSD using a deep Fourier-features network as explained in section 4. (d) RJSD-K[1]: Two-sample test based on the kernel RJSD estimator, optimizing the length-scale of a Gaussian kernel. (e) MMD-O: Two-sample test based on MMD, optimizing the length-scale of the Gaussian kernel (Liu et al., 2020). (f) MMD-D: Two-sample test based on MMD with a deep kernel (Liu et al., 2020). (g) C2ST-L: a classifier two-sample test based on the output classification scores (Cheng and Cloninger, 2022). (h) C2ST-S: a classifier two-sample test based on the sign of the output classification scores (Lopez-Paz and Oquab, 2016). We perform two-sample testing on two synthetic and two real-world datasets.

**Blobs dataset (Liu et al., 2020):** In this dataset, $\mathbb{P}$ and $\mathbb{Q}$ are mixtures of nine Gaussians with the same modes. Each mode in $\mathbb{P}$ is an isotropic Gaussian; however, the modes in $\mathbb{Q}$ have different covariances. Here, we perform two-sample testing increasing the number of samples per blob ($N = 9 \times$ samples per blob). Fig 4(a) presents the results. We can clearly see that RJSD-FF, RJSD-D, and JSD outperform all SOTA methods. We can conclude that even for a small number of samples the RJSD-based methods exhibit high test power.

**High-Dimensional Gaussian Mixtures (Liu et al., 2020):** We assess the performance of RJSD at high dimensions on a bimodal multidimensional Gaussian dataset. In this dataset, $\mathbb{P}$ and $\mathbb{Q}$ have the same modes, and their covariances differ only on a single dimension. See Liu et al. (2020) for details.

[1]We did not perform this test for large size datasets due to computational restrictions




(a) (b) (c) (d)

Figure 4: Average test power ($\alpha = 0.05$) over 10 trials on the (a) Blobs dataset. (b) High dimensional Gaussian mixture, fixed $d = 10$. (c) High dimensional Gaussian mixture, fixed $N + M = 4000$ (d) Higgs dataset

Table 3: MNIST average test power ($\alpha = 0.05$). Bold represents higher mean per column.

| $N + M$ | 200 | 300 | 400 | 500 | 600 |
|---|---|---|---|---|---|
| RJSD-FF | 0.374 ± 0.100 | 0.811± 0.012 | 0.996± 0.001 | **1.000 ±0.000** | **1.000±0.000** |
| RJSD-RFF | 0.184 ± 0.025 | 0.320± 0.029 | 0.436 ±0.030 | 0.644 ±0.037 | 0.800 ±0.051 |
| RJSD-D | 0.352± 0.084 | **0.898 ±0.108** | **1.000 ±0.000** | **1.000± 0.000** | **1.000± 0.000** |
| MMD-O | 0.148± 0.035 | 0.221± 0.042 | 0.283± 0.042 | 0.398± 0.050 | 0.498± 0.035 |
| MMD-D | **0.449± 0.124** | 0.704± 0.182 | 0.985 ± 0.010 | 0.999± 0.003 | **1.000± 0.000** |
| C2ST-L | 0.254± 0.126 | 0.424± 0.113 | 0.818± 0.102 | 0.967± 0.029 | 0.994± 0.010 |
| C2ST-S | 0.181± 0.112 | 0.364± 0.104 | 0.759± 0.121 | 0.945± 0.042 | 0.986± 0.014 |

We test both, changing the number of samples while keeping the dimension constant ($d = 10$) and maintaining the number of samples ($N = 4000$) while modifying the dimensionality. Figs. 4(b) and 4(c) display the results. RJSD-D and RJSD-FF are the winners in most settings, although C2ST-L performs better at higher dimensions.

**Higgs dataset (Baldi et al., 2014):** Following Liu et al. (2020) we perform two-sample testing on the Higgs dataset ($d = 4$) as we increase the number of samples. Fig. 4(d) shows the results. Once again, RJSD-D and RJSD-FF outperform the baselines in almost all scenarios.

**MNIST generative model:** Here, we train RJSD models to distinguish between the distribution $\mathbb{P}$ of MNIST digits and the distribution $\mathbb{Q}$ of generated samples from a pretrained deep convolutional generative adversarial network (DCGAN) (Radford et al., 2015). Table 3 reports the average test power for all methods as we increase the number of samples. RJSD-D consistently outperforms the compared methods, except with the lowest number of observations.

## 6 Conclusions

We introduce the representation Jensen-Shannon divergence (RJSD), a novel measure based on embedding distributions in a feature space allowing the construction of non-parametric estimators based on Fourier features. Notably, this estimator demonstrates scalability, differentiability, making it suitable for diverse machine-learning problems. We demonstrated that RJSD provides a lower bound on the classical Jensen-Shannon divergence leading to a variational estimator of high precision compared to related approaches. We leveraged this novel divergence to train generative networks, and the empirical results show that RJSD effectively mitigates mode collapse yielding generative models that produce more accurate and diverse results. Furthermore, when applied to two-sample testing, RJSD surpassed other SOTA techniques demonstrating superior performance and reliability to discriminate between distributions. These findings highlight the significant practical implications of our divergence measure.

## References

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

P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):4308, 2014.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

T. B. Berrett and R. J. Samworth. Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv preprint arXiv:1904.09347*, 2019.

Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.

P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

X. Cheng and A. Cloninger. Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, 68(10):6631–6662, 2022.

A. B. Dieng, F. J. Ruiz, D. M. Blei, and M. K. Titsias. Prescribed generative adversarial networks. *arXiv preprint arXiv:1910.04302*, 2019.

V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

F. Farnia and A. Ozdaglar. Gans may have no nash equilibria. *arXiv preprint arXiv:2002.09124*, 2020.

F. A. González, A. Gallego, S. Toledo-Cortés, and V. Vargas-Calderón. Learning with density matrices and random features. *Quantum Machine Intelligence*, 4(2):23, 2022.

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

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Y. Han, J. Jiao, T. Weissman, and Y. Wu. Optimal rates of entropy estimation over lipschitz balls. 2020.

J. K. Hoyos Osorio, O. Skean, A. J. Brockmeier, and L. G. Sanchez Giraldo. The representation jensen-rényi divergence. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4313–4317. IEEE, 2022.

A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927. PMLR, 2014.

Y. Li and R. E. Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

T. Liang. Estimating certain integral probability metric (ipm) is as hard as estimating under the ipm. *arXiv preprint arXiv:1911.00730*, 2019.

J. H. Lim and J. C. Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

Z. Lin, A. Khetan, G. Fanti, and S. Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

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.

D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

K. Moon and A. Hero. Multivariate f-divergence estimation with confidence. *Advances in neural information processing systems*, 27, 2014.

K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero III. Ensemble estimation of information divergence. *Entropy*, 20(8):560, 2018.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero. Direct estimation of information divergence using nearest neighbor ratios. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 903–907. IEEE, 2017.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

G. Pichler, P. J. A. Colombo, M. Boudiaf, G. Koliander, and P. Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pages 17691–17715. PMLR, 2022.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

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

S. Singh and B. Póczos. Generalized exponential concentration inequality for rényi divergence estimation. In *International Conference on Machine Learning*, pages 333–341. PMLR, 2014.

S. Sreekumar and Z. Goldfeld. Neural estimation of statistical divergences. *Journal of machine learning research*, 23(126), 2022.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On the empirical estimation of integral probability metrics. 2012.

A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

D. Virosztek. The metric property of the quantum jensen-shannon divergence. *Advances in Mathematics*, 380:107595, 2021.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

S. Yu, L. G. Sanchez Giraldo, R. Jenssen, and J. C. Principe. Multivariate extension of matrix-based rényi's $\alpha$-order entropy functional. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2960–2966, 2019.

S. Yu, F. Alesiani, X. Yu, R. Jenssen, and J. Principe. Measuring dependence with matrix-based entropy functional. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10781–10789, 2021.