An Information-Theoretic Evaluation of Generative Models in Learning Multi-modal Distributions

Mohammad Jalali
Department of Electrical and Computer Engineering
Isfahan University of Technology
mjalali@ec.iut.ac.ir

Cheuk Ting Li
Department of Information Engineering
The Chinese University of Hong Kong
ctli@ie.cuhk.edu.hk

Farzan Farnia
Department of Computer Science and Engineering
The Chinese University of Hong Kong
farnia@cse.cuhk.edu.hk

Abstract

The evaluation of generative models has received significant attention in the machine learning community. When applied to a multi-modal distribution which is common among image datasets, an intuitive evaluation criterion is the number of modes captured by the generative model. While several scores have been proposed to evaluate the quality and diversity of a model's generated data, the correspondence between existing scores and the number of modes in the distribution is unclear. In this work, we propose an information-theoretic diversity evaluation method for multi-modal underlying distributions. We utilize the *Rényi Kernel Entropy (RKE)* as an evaluation score based on quantum information theory to measure the number of modes in generated samples. To interpret the proposed evaluation method, we show that the RKE score can output the number of modes of a mixture of sub-Gaussian components. We also prove estimation error bounds for estimating the RKE score from limited data, suggesting a fast convergence of the empirical RKE score to the score for the underlying data distribution. Utilizing the RKE score, we conduct an extensive evaluation of state-of-the-art generative models over standard image datasets. The numerical results indicate that while the recent algorithms for training generative models manage to improve the mode-based diversity over the earlier architectures, they remain incapable of capturing the full diversity of real data. Our empirical results provide a ranking of widely-used generative models based on the RKE score of their generated samples[1].

1 Introduction

Deep generative models trained by generative adversarial networks (GANs) [1] and diffusion models [2] have achieved impressive results in various unsupervised learning settings [3, 4, 5, 6]. Due to their success in generating image samples with high visual quality, the analysis of large-scale generative models has received great attention in the machine learning community. In particular, the evaluation of generative models has been extensively studied in the recent literature to understand the benefits and drawbacks of existing approaches to training generative models.

To address the evaluation task for generative models, multiple assessment scores have been proposed in the literature. The existing evaluation metrics can be divided into two general categories: 1)

[1]The code repository is available at `https://github.com/mjalali/renyi-kernel-entropy`.

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

distance-based metrics including Fréchet Inception Distance (FID) [7] and Kernel Inception distance (KID) [8] scores, which measure a distance between the learned generative model and data distribution, 2) quality-based metrics including the Inception score [9], precision and recall scores [10, 11], and density and coverage scores [12], aiming to measure the quality and diversity of the generated samples based on the confidence and variety of labels assigned by a pre-trained neural net on ImageNet.

On the other hand, a different measure of diversity that is popular and standard in the information theory literature is entropy. The primary challenge with an entropy-based approach to the assessment of generative models is the statistical costs of entropy estimation in the typical high-dimensional spaces of image data, requiring a sample size exponentially growing with the dimension of data. Consequently, without further assumptions on the data distribution, it will be statistically and computationally infeasible to estimate the entropy value for high-dimensional image data.

In this work, we propose a novel information-theoretic approach for the diversity evaluation of generative models. Our proposed approach targets multi-modal data distributions comprised of several distinct modes, which is an applicable assumption to image datasets with a cluster-based structure due to their latent color and shape-based features. In this approach, we follow the entropy calculation in quantum information theory [13, 14] and utilize matrix-based Rényi entropy scores to evaluate the variety of samples produced by a generative model.

Considering the Rényi entropy in the Gaussian kernel space, we propose *Rényi kernel entropy (RKE)* and *relative Rényi kernel entropy (RRKE)* scores to measure the absolute and relative diversity of a multi-modal distribution with respect to the actual data distribution. We develop computationally efficient methods for estimating these entropy scores from empirical data with statistical convergence guarantees. We also prove that the RKE score will converge to standard differential entropy as the Gaussian kernel bandwidth approaches zero.

We provide an interpretation of the RKE score by deriving its closed-form expression for benchmark Gaussian mixture models (GMMs). In the GMM case, we show that the proposed RKE score reveals the number of Gaussian components in the underlying GMM, which motivates the method's application to general mixture distributions. We further extend this interpretation to mixture models with sub-Gaussian modes, e.g. modes with a bounded support set, to support the RKE score in more general settings. We discuss the numerical performance of the proposed score in synthetic mixture settings. Our numerical results demonstrate the fast convergence of the RKE score from a limited number of synthetic mixture data.

Next, we present the results of our numerical experiments on the evaluation of various state-of-the-art GAN and diffusion models using the RKE and RRKE metrics. Our numerical evaluation of the RKE score shows the lower diversity obtained by standard GAN models compared to real training data, which provides an information-theoretic numerical proof complementary to the birthday-paradox empirical proof for the same result in [15]. Furthermore, our empirical results suggest that the recent GAN and diffusion model architectures improve the mode-based variety of generated data over earlier GAN formulations. We can summarize the main contributions of this work as follows:

- Proposing an information theoretic approach to evaluate the diversity of generative models
- Developing computationally efficient methods to compute Rényi kernel entropy scores
- Providing theoretical and numerical support for the proposed evaluation methodology in the benchmark setting of Gaussian and sub-Gaussian mixture models
- Diversity evaluation of standard generative models using the information-theoretic method

2 Related Work

The evaluation of GAN-based generative models has been studied by a large body of related works. As surveyed in [16], several evaluation methods have been developed in the literature. The Inception score (IS) [9] uses the output of a pre-trained Inception-net model as features and proposes a score summing up the entropy-based diversity of assigned labels' distribution and confidence score averaged over the conditional labels' distribution. The modified IS (m-IS) in [17] substitutes the KL-divergence term in IS with a cross entropy term, which helps m-IS capture diversity within images from a certain class. Unlike these works, our proposed approach bases on matrix-based entropy scores which capture the number of clusters in a multi-modal distribution.

Also, several distance-based evaluation metrics have been proposed in the deep learning literature. The Wasserstein Critic [18] attempts to approximate the Wasserstein distance between the real and generated samples. The Fréchet Inception Distance (FID) [7] measures a distance based on the embedding of the last layer of the pre-trained Inception-net, where it fits multivariate Gaussian models to real and generated data and calculates their Fréchet distance [19]. [20] conduct a comprehensive comparison of several standard GAN architectures based on IS and FID scores, discussing the similarities and differences of the GANs' performance. As another variant of FID, [21] suggest a bias-free estimation of FID using quasi-Monte Carlo integration. In another related work, [8] propose Kernel Inception Distance (KID) as the squared maximum mean discrepancy (MMD) between two distributions. Adversarial accuracy and divergence scores in [22] utilize two classifiers to compute the closeness of distributions of real and fake data conditioned on category labels from the two classifiers. Unlike the above metrics, our proposed relative Rényi entropy score focuses on the number of common modes between real and fake data, and hence reduces the statistical complexity of estimating the entropy.

The diversity vs. quality tradeoff of GANs' generated samples has also been studied in multiple related works. [15] examine the diversity of GANs' data through a birthday paradox-based approach and suggest that the support set size of GANs' data could be smaller than real training data. The precision and recall evaluation by [10] assigns a two-dimensional score where precision is defined as the portion of fake data that can be generated by real distribution while recall is defined as the portion of real data that can be generated by the generative model. The improved precision and recall in [11] further address the sensitivity disadvantages of these scores by estimating the density function via the k-nearest neighbour method. Also, the density and coverage scores [12] provide a more robust version of precision and recall metrics to outliers. Additionally, [23] proposed a 3-dimensional metric, that measures the fidelity, diversity and generalization of models. We note that our work offers a complementary approach to the diversity evaluation for GANs, and since it directly estimates the matrix-based entropy from data, it requires a comparatively smaller sample size for proper estimation.

3 Preliminaries

3.1 Kernel-based Feature Maps and Representation

Throughout the paper, we use $\mathbf{X} \in \mathcal{X}$ to denote the data vector. Also, we denote the kernel feature map by $\phi : \mathbb{R}^t \to \mathbb{R}^d$ which gives us the kernel function $k : \mathbb{R}^t \times \mathbb{R}^t \to \mathbb{R}$ as the inner product of the corresponding feature vectors: $k(\mathbf{x}, \mathbf{y}) = \big\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \big\rangle$. Given n training data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ we use Φ to denote the normalized kernel feature map-based data matrix, i.e.,

$$\Phi = \frac{1}{\sqrt{n}} \begin{bmatrix} \phi(\mathbf{x}_1) \\ \vdots \\ \phi(\mathbf{x}_n) \end{bmatrix}.$$

Then, the kernel matrix $K \in \mathbb{R}^{n \times n}$ whose (i, j)th entry will be $\frac{1}{n} k(\mathbf{x}_i, \mathbf{x}_j)$ will be identical to $K = \Phi\Phi^\top$. Observe that, by definition, K will be a positive semi-definite (PSD) matrix, possessing positive eigenvalues $\lambda_1, \ldots, \lambda_n$ where the non-zero eigenvalues are shared with the empirical kernel covariance matrix $C = \Phi^\top\Phi$. We call a kernel function normalized if $k(\mathbf{x}, \mathbf{x}) = 1$ for every $\mathbf{x} \in \mathcal{X}$. A standard example of a normalized kernel function is the Gaussian kernel with bandwidth parameter σ defined as:

$$k_\sigma(\mathbf{x}, \mathbf{y}) := \exp\Big(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma^2}\Big).$$

For every normalized kernel function, the eigenvalues of kernel matrix K (and also empirical kernel covariance C) will be non-negative and add up to 1 as the trace of K will be 1. As a result, K's eigenvalues can be interpreted as a probability sequence.

3.2 Rényi Entropy for PSD Matrices

A standard extension of the entropy concept to PSD matrices is the matrix-based Rényi entropy [14]. The Rényi entropy of order $\alpha > 0$ for a PSD matrix $A \in \mathbb{R}^{d \times d}$ with eigenvalues $\lambda_1, \ldots, \lambda_d$ is defined as

$$\mathrm{RE}_\alpha(A) := \frac{1}{1-\alpha} \log\big(\mathrm{Tr}(A^\alpha)\big) = \frac{1}{1-\alpha} \log\Big(\sum_{i=1}^{d} \lambda_i^\alpha\Big),$$

where Tr denotes the trace operator. A commonly-used special case which we use throughout the paper is the Rényi entropy of order $\alpha = 2$ defined as $\mathrm{RE}_2(A) = \log\big(1/\sum_{i=1}^{d} \lambda_i^2\big)$.

Proposition 1. *For every PSD $A \in \mathbb{R}^d$, the following holds where $\|\cdot\|_F$ denotes the Frobenius norm:*

$$\mathrm{RE}_2(A) = \log\big(1/\mathrm{Tr}(AA^\top)\big) = \log\big(1/\|A\|_F^2\big).$$

3.3 Relative Rényi Entropy

To measure the relative diversity of a matrix $A \in \mathbb{R}^{d\times d}$ with respect to another matrix $B \in \mathbb{R}^{d\times d}$, one can use the sandwiched relative Rényi entropy of order α [24] defined as

$$\mathrm{RRE}_\alpha(A, B) = \frac{1}{\alpha - 1} \log\Big(\mathrm{Tr}\left(\big(B^{\frac{1-\alpha}{2\alpha}} A B^{\frac{1-\alpha}{2\alpha}}\big)^\alpha\right)\Big).$$

A widely-used special case is the relative entropy of order $\alpha = \frac{1}{2}$ which is commonly called the Fidelity score in quantum information theory. The definition of the Fidelity score is

$$\mathrm{RRE}_{1/2}(A, B) := -2\log\left(\mathrm{Tr}\big(\sqrt{B^{1/2} A B^{1/2}}\big)\right).$$

We note that the relative Rényi entropy of order $\alpha = 2$ requires an invertible matrix B which may not hold in the applications to multi-modal distributions with rank-deficient kernel covariance matrices as discussed in the next sections.

4 A Diversity Metric for Multi-modal Distributions

4.1 Kernel-based Rényi Entropy Scores

Given the kernel matrix K computed using the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of random vector $\mathbf{X}$, the empirical *Rényi kernel entropy (RKE)* of the observed data can be defined as

$$\widehat{\mathrm{RKE}}_\alpha(\mathbf{X}) := \mathrm{RE}_\alpha(K),$$

which, as discussed in [14], is a diversity measure of the data. This diversity measurement approach will be interpreted and justified in the next subsection. Note that if the dimension d of the feature space is finite, since $K = \Phi\Phi^\top$ and empirical covariance matrix $\widehat{C} = \Phi^\top\Phi$ share the same eigenvalues, we have

$$\widehat{\mathrm{RKE}}_\alpha(\mathbf{X}) \;=\; \mathrm{RE}_\alpha(\widehat{C}) \;=\; \mathrm{RE}_\alpha\left(\frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right).$$

Therefore, we can see that the empirical Rényi kernel entropy $\widehat{\mathrm{RKE}}_\alpha(\mathbf{X})$ is an estimate of the following quantity about the underlying distribution P_X where $\mathbf{X}$ is sampled from, which we call the *Rényi kernel entropy* of P_X:

$$\mathrm{RKE}_\alpha(\mathbf{X}) := \mathrm{RE}_\alpha(C_X).$$

Here C_X denotes the kernel covariance matrix of distribution P_X defined as

$$C_X := \mathbb{E}_{P_X}\left[\phi(X)\phi(X)^\top\right] = \int P_X(x)\phi(x)\phi(x)^\top \mathrm{d}x.$$

Similarly, we can use the kernel-based relative Rényi entropy as a measure of joint diversity between distributions P_X, P_Y of random vectors $\mathbf{X}, \mathbf{Y}$. Here, for random vectors $\mathbf{X}$, $\mathbf{Y}$ distributed according to P_X, P_Y, we define the *relative Rényi kernel entropy (RRKE$_\alpha$)* score as the order-α relative kernel entropy between their kernel covariance matrices C_X, C_Y:

$$\mathrm{RRKE}_\alpha(\mathbf{X}, \mathbf{Y}) = \mathrm{RRE}_\alpha(C_X, C_Y).$$

In order to estimate the above relative entropy score, we can use the empirical RRKE score between the empirical kernel covariance matrices for samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from P_X and samples $\mathbf{y}_1, \ldots, \mathbf{y}_m$ from P_Y:

$$\widehat{\mathrm{RRKE}}_\alpha(\mathbf{X}, \mathbf{Y}) \;:=\; \mathrm{RRE}_\alpha\left(\frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top,\; \frac{1}{m}\sum_{j=1}^{m} \phi(\mathbf{y}_j)\phi(\mathbf{y}_j)^\top\right).$$

In the rest of this section, we show how the kernel-based Rényi entropy score evaluated under a Gaussian kernel can relate to the number of modes of multi-modal distributions with sub-Gaussian components, and subsequently how the relative Rényi entropy score counts the number of joint modes between two multi-modal distributions with sub-Gaussian modes.

4.2 RKE as a Measure of the Number of Modes

We consider a multi-modal underlying distribution P_X consisting of k distinct modes. Here, our goal is to show that the order-2 RKE score under the Gaussian kernel can count the number of present modes in the distribution. To this end, we analyze the order-2 Rényi kernel entropy score of the kernel covariance matrix C_X of mixtures of Gaussian and sub-Gaussian components and theoretically show that the RKE score reduces to the number of well-separated modes.

First, we derive the closed-form expression of the order-2 Rényi kernel entropy of a Gaussian mixture model under the Gaussian kernel. Here, we use the following notation to denote a k-component Gaussian mixture model where the ith component has frequency ω_i, mean vector $\boldsymbol{\mu}_i$ and Covariance matrix Σ_i: $P_{\mathrm{GMM}(\boldsymbol{\omega},\boldsymbol{\mu},\Sigma)} = \sum_{i=1}^{k} \omega_i \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$. Also, for a positive definite matrix $A \in \mathbb{R}^{d\times d}$, we use $\|\cdot\|_A$ to denote the A-norm defined as $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A \mathbf{x}}$.

Theorem 1. *Suppose that the distribution of $\mathbf{X}$ is given by the Gaussian mixture model $P_{\mathrm{GMM}(\boldsymbol{\omega},\boldsymbol{\mu},\Sigma)}$. Then, order-2 Rényi kernel score under the Gaussian kernel with bandwidth σ, denoted by G_σ, is*

$$\mathrm{RKE}_2^{\mathrm{G}_\sigma}(\mathbf{X}) = -\log\Bigg(\sum_{i=1}^{k}\sum_{j=1}^{k}\Bigg[\omega_i\omega_j e^{-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2_{A_{i,j}}}{\sigma^2}} \det\Big(I+\frac{2}{\sigma^2}(\Sigma_i+\Sigma_j)\Big)^{-\frac{1}{2}}\Bigg]\Bigg),$$

where $A_{i,j}$ is defined as follows given the $d\times d$ identity matrix I:

$$\begin{aligned} A_{i,j} \;:=\; & I - \big(I + 2\sigma^2\Sigma_i^{-1} - (I+2\sigma^2\Sigma_j^{-1})^{-1}\big)^{-1} - \big(I + 2\sigma^2\Sigma_j^{-1} - (I+2\sigma^2\Sigma_i^{-1})^{-1}\big)^{-1} \\ & + \big(2\sigma^2\Sigma_i^{-1} + 2\sigma^2\Sigma_j^{-1} + 4\sigma^4\Sigma_i^{-1}\Sigma_j^{-1}\big)^{-1} + \big(2\sigma^2\Sigma_i^{-1} + 2\sigma^2\Sigma_j^{-1} + 4\sigma^4\Sigma_j^{-1}\Sigma_i^{-1}\big)^{-1}. \end{aligned}$$

Proof. We defer the proof to the Appendix. □

Corollary 1. *Suppose $\mathbf{X} \sim \mathrm{GMM}(\boldsymbol{\omega}, \boldsymbol{\mu}, \Sigma)$ follows from a Gaussian mixture model with isotropic covariance matrices $\Sigma_i = \sigma_i^2 I$. Defining the coefficients $a_{i,j} := 1 + \frac{2(\sigma_i^2+\sigma_j^2)}{\sigma^2}$, the RKE score will be*

$$\mathrm{RKE}_2^{\mathrm{G}_\sigma}(\mathbf{X}) = -\log\Bigg(\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j a_{i,j}^{-\frac{d}{2}} e^{-\frac{a_{i,j}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2}}\Bigg).$$

Theorem 1 and Corollary 1 show that if $\sigma \gg \max_i \|\Sigma_i\|_{\mathrm{sp}}$, i.e. the kernel bandwidth dominates the spectral norm (maximum eigenvalue) of the component-wise covariance matrices, then $A_{i,j} \approx I$ and $a_{i,j} \approx 1$ and the RKE score will approximately be

$$\mathrm{RKE}_2^{\mathrm{G}_\sigma}(\mathbf{X}) \approx -\log\Big(\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j e^{-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2}}\Big).$$

The next theorem shows the above approximation generalizes to mixtures of sub-Gaussian components, e.g. modes with bounded support sets, and also provides an approximation error bound.

Theorem 2. *Suppose that $\mathbf{X} \in \mathbb{R}^d$ has a mixture distribution $P_{\mathrm{SGMM}} = \sum_{i=1}^{k}\omega_i P_i$ where the kth component occurs with probability ω_k and has a sub-Gaussian distribution with parameter $\boldsymbol{\sigma}_i$, i.e. its moment-generating function (MGF) M_{P_i} satisfies the following for every vector $\boldsymbol{\beta} \in \mathbb{R}^d$ given the mean vector $\boldsymbol{\mu}_i$ for P_i: $\mathbb{E}_{P_i}\big[\exp\big(\boldsymbol{\beta}^\top(\mathbf{X}-\boldsymbol{\mu}_i)\big)\big] \le \exp\big(\|\boldsymbol{\beta}\|_2^2\sigma_i^2/2\big)$. Then, the following approximation error bound holds where we define $\alpha_i^2 := 1 + 2\sigma_i^2/\sigma^2$:*

$$\Bigg|\exp\Big(-\mathrm{RKE}_2^{\mathrm{G}_\sigma}(\mathbf{X})\Big) - \sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j e^{-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2}}\Bigg| \;\le\; \sqrt{\sum_{i=1}^{k} 8\omega_i\big(1-\alpha_i^{-d}\big)}.$$

Proof. We defer the proof to the Appendix. □

More generally, we show that the RKE under a Gaussian kernel with a small bandwidth provides an estimate of the differential Rényi entropy, which connects the approach with a smoothed differential entropy estimator in high-dimensional settings.

Theorem 3. *Suppose that the underlying distribution of X has a continuous probability density function P_X. The order-2 Rényi kernel score under Gaussian kernel with bandwidth σ for this underlying distribution satisfies*

$$\lim_{\sigma \to 0} \Big(\mathrm{RKE}_2^{\mathrm{G}_\sigma}(\mathbf{X}) + \frac{d}{2} \log(\pi \sigma^2) \Big) = -\log \Big(\int P_X(\mathbf{x})^2 \mathrm{d}\mathbf{x} \Big),$$

where the right hand side is the order-2 Rényi differential entropy of P_X.

Proof. We defer the proof to the Appendix. □

4.3 RRKE as the Number of Common Modes

In order to measure the joint mode-based diversity, we propose the order-$\frac{1}{2}$ relative Rényi entropy score. We note that our choice of order $\frac{1}{2}$ is due to the existing inverse covariance matrix term in the relative entropies of orders greater than 1, which is not applicable to rank deficient matrices expected in the case of well-separated multi-modal distributions. Furthermore, the order $\frac{1}{2}$ relative entropy, well-known as the fidelity score, is a commonly-used and well-analyzed relative entropy case in quantum information theory.

To interpret the application of the order-$\frac{1}{2}$ relative entropy, we show the following theorem discussing how the negative RRKE score could approximate the joint diversity between two input multimodal distributions with sub-Gaussian components.

Theorem 4. *Suppose that $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$ are random vectors with mixture distributions $P_{\mathrm{MM}_1} = \sum_{i=1}^k \omega_i P_i$ and $P_{\mathrm{MM}_2} = \sum_{i=1}^k \eta_i Q_i$, respectively, where ω_i, η_i denote the frequency of the ith component. We also assume that P_i and Q_i have mean vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\zeta}_i$ respectively and are both σ_i-sub-Gaussian. Then, the following approximation error bound holds for order-$\frac{1}{2}$ relative Rényi entropy where we define $\alpha_i^2 = 1 + 2\sigma_i^2/\sigma^2$:*

$$\left| \exp\Big(-\mathrm{RRKE}_2^{\mathrm{G}_\sigma}(\mathbf{X}, \mathbf{Y})\Big) - \sum_{i=1}^k \sum_{j=1}^k \sqrt{\omega_i \eta_j} e^{-\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\zeta}_j\|_2^2}{\sigma^2}} \right| \leq \sqrt[4]{\sum_{i=1}^k 32\big(\omega_i + \eta_i\big)\big(1 - \alpha_i^{-d}\big)}$$

Proof. We defer the proof to the Appendix. □

The above theorem shows that if the ith mode of mixture distribution P_{mm_1} and the jth mode of mixture distribution P_{mm_2} are sufficiently close, then they add nearly $\sqrt{\omega_i \eta_j}$ to the RRKE score.

5 Estimation Procedure and Guarantees

As discussed earlier, the RKE and RRKE scores for a proper kernel function provide measures of the absolute and relative mode-based diversity of multi-modal distributions. Here, we propose kernel-based estimators for these entropy scores. Our estimators suggest computationally and statistically feasible ways for approximating the entropy measures by exploiting the connections between the kernel similarity values and the Rényi entropy scores. In addition, we provide non-asymptotic estimation error bounds for the proposed estimators to analyze their sample complexity. Regarding the RKE score, the following results connect this score with the kernel function $k(\mathbf{x}, \mathbf{x}')$.

Theorem 5. *Given a random vector $\mathbf{X}$ distributed as P_X, the order-2 RKE score is the result of the following equation, where $\mathbf{X}, \mathbf{X}'$ are IID draws of P_X:*

$$\mathrm{RKE}_2(\mathbf{X}) = -\log\Big(\mathbb{E}_{X, X' \overset{\mathrm{iid}}{\sim} P_X}\big[k^2(\mathbf{X}, \mathbf{X}')\big]\Big).$$

Corollary 2. *For samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $K_{XX} = [\frac{1}{n} k(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$ being the normalized kernel matrix for the observed samples, the empirical order-2 Kernel entropy is given by*

$$\widehat{\mathrm{RKE}}_2(\mathbf{X}) \;=\; -\log\Big(\big\|K_{XX}\big\|_F^2\Big),$$

Proof. We defer the proof to the Appendix. We note that the result of Corollary 2 for the empirical RKE score has already been shown and discussed in [14]. □

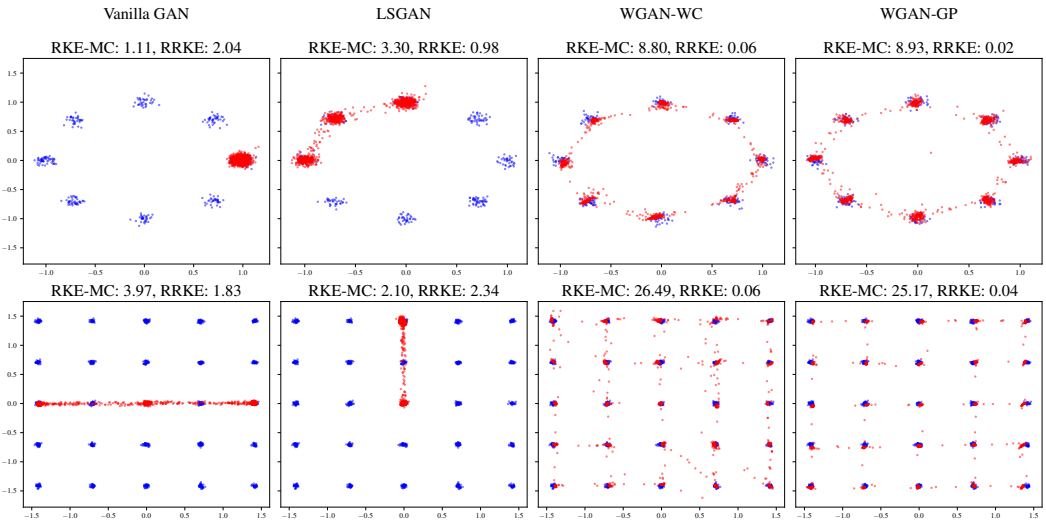


Figure 1: GANs' generated samples for Gaussian mixtures and the RKE-MC ↑ and RRKE ↓ scores.

As the above results suggest, we can use the normalized Frobenius norm of the kernel matrix to estimate the RKE score. Therefore, the computation of order-2 RKE only requires the Frobenius norm of the kernel matrix, which can be computed in $O(n^2)$ complexity. We note that for a general order-α Rényi entropy, one can apply the randomized algorithm in [25] for the computation of the order-α RKE score. Our next result bounds the estimation error for the empirical RKE score.

Theorem 6. *Consider a normalized kernel function satisfying $k(\mathbf{x},\mathbf{x}) = 1$ for every $\mathbf{x} \in \mathcal{X}$. Then, for every $\delta > 0$ the following bound will hold with probability at least $1-\delta$:*

$$\left|\exp\left(-\widehat{\mathrm{RKE}}_2(\mathbf{X})\right) - \exp\left(-\mathrm{RKE}_2(\mathbf{X})\right)\right| \leq O\left(\sqrt{\frac{\log\frac{n}{\delta}}{n}}\right)$$

Proof. We defer the proof to the Appendix. □

As implied by the above results, the order-2 Rényi kernel entropy can be efficiently estimated from training data and the probability of an ϵ-large error will exponentially diminish with the sample size n. Next, we discuss the computation approach for the order-$\frac{1}{2}$ RRKE score. The following result reveals the kernel-based representation of this score in the empirical case.

Theorem 7. *Consider empirical samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$ drawn from P_X and $\mathbf{y}_1,\ldots,\mathbf{y}_m$ drawn from P_Y. Then, the following identity holds for their empirical order-$\frac{1}{2}$ RRKE score:*

$$\widehat{\mathrm{RRKE}}_{\frac{1}{2}}(\mathbf{X},\mathbf{Y}) = -\log\left(\left\|K_{XY}\right\|_*^2\right)$$

where $K_{XY} = \left[\frac{1}{\sqrt{nm}}k(\mathbf{x}_i,\mathbf{y}_j)\right]_{n\times m}$ denotes the normalized cross kernel matrix and $\|\cdot\|_*$ denotes the nuclear norm, i.e. the sum of a matrix's singular values.

Proof. We defer the proof to the Appendix. □

As implied by the above theorem, the RRKE score can be computed using the singular value decomposition (SVD) for finding the singular values of K_{XY}. Note that the application of SVD requires $O(\min\{m^2n, mn^2\})$ computations given n, m samples from P_X and P_Y, respectively. Therefore, computing the RRKE score could be more expensive than the RKE score, since the nuclear norm is more costly to compute than the Frobenius norm.

6 Numerical Results

We tested the performance of the proposed entropy-based diversity evaluation approach on several combinations of standard GAN architectures and datasets. Specifically, we used the synthetic 8-component and 25-component Gaussian mixture datasets in [26] and the following image datasets:

Table 1: Evaluated scores for three image datasets. RKE-MC (Mode Count) denotes $\exp(\mathrm{RKE})$.

	Method	IS ↑	FID ↓	Precision ↑	Recall ↑	Density ↑	Coverage ↑	**RRKE ↓**	**RKE-MC ↑**
CIFAR-10	Dataset	11.57	-	-	-	-	-	-	39.58
	NVAE	5.85	51.67	0.36	0.50	0.28	0.60	2.01	17.65
	VDVAE	10.51	37.51	0.34	0.78	0.23	0.21	1.81	32.49
	DCGAN	5.75	54.30	0.59	0.25	0.49	0.23	0.98	10.19
	WGAN-WC	2.59	157.26	0.36	0.00	0.18	0.03	2.09	10.64
	WGAN-GP	7.51	21.66	0.62	0.56	0.57	0.51	0.74	19.07
	SAGAN	8.62	10.17	0.68	0.62	0.73	0.73	0.65	24.46
	SNGAN	8.81	9.23	0.70	0.62	0.77	0.74	0.62	25.83
	ContraGAN	9.69	4.02	0.75	0.62	0.99	0.86	0.52	29.80
Tiny-ImageNet	Dataset	33.99	-	-	-	-	-	-	155.86
	SAGAN	8.21	46.98	0.55	0.49	0.44	0.27	1.42	25.68
	SNGAN	8.12	48.96	0.55	0.46	0.40	0.26	1.46	27.18
	BigGAN	11.57	27.34	0.60	0.58	0.53	0.43	1.23	39.61
	ContraGAN	13.79	21.36	0.54	0.54	0.54	0.45	1.26	56.94
ImageNet	Dataset	357.35	-	-	-	-	-	-	1823.52
	SAGAN-256	29.67	44.66	0.57	0.58	0.42	0.35	2.34	105.57
	SNGAN-256	31.92	35.75	0.54	0.64	0.41	0.38	2.22	115.62
	ContraGAN-256	24.91	34.79	0.67	0.51	0.64	0.33	2.54	152.89
	BigGAN-256	28.33	33.48	0.58	0.61	0.49	0.37	2.28	106.07
	ReACGAN-256	52.53	15.65	0.74	0.42	0.79	0.41	2.15	119.76
	BigGAN-2048	96.42	4.49	0.71	0.58	0.80	0.65	1.83	606.18
	StyleGAN-XL	204.73	1.94	0.77	0.61	0.67	0.81	1.50	1375.17
	LDM-4-G	242.62	3.60	0.86	0.60	0.69	0.78	1.56	1321.24
	ADM-G	188.70	3.86	0.82	0.64	0.66	0.82	1.47	1407.75

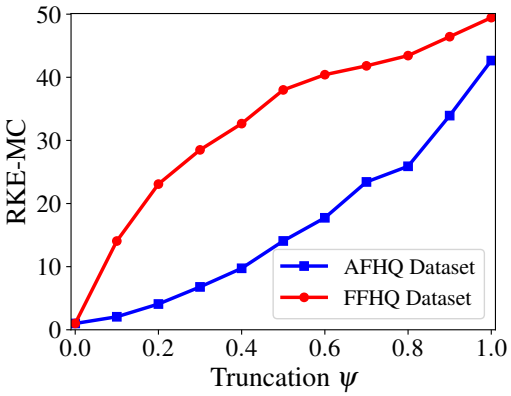
CIFAR-10 [27], Tiny-ImageNet [28], MS-COCO [29], AFHQ [30], FFHQ [31] and ImageNet [32]. We evaluated the performance of the following list of widely-used VAE, GAN and diffusion model architectures: NVAE [33], Very Deep VAE (VDVAE) [34], Vanilla GAN [1], LSGAN [35], Wassesrtein GAN with weight clipping (WGAN-WC) [18], Wassesrtein GAN with gradient penalty (WGAN-GP) [26], DCGAN [36], Self-Attention GAN (SAGAN) [37], Spectrally-Normalized GAN (SNGAN) [38], ContraGAN [39], ReACGAN [40], BigGAN [3], StyleGAN3 [41], StyleGAN-XL [42], GigaGAN [43], LDM [44] ADM-G [45] and BK-SDM [46]. To have a fair evaluation of the models, we downloaded the trained generative models from the StudioGAN repository [47].

In our evaluation of generative models, we compared the performance of order-2 Rényi Kernel Entropy Mode Count (RKE-MC), defined as $\exp(\mathrm{RKE}_2(\mathbf{X}))$, and order-$\frac{1}{2}$ Relative Rényi Kernel Entropy (RRKE) with the following standard baselines widely used in the evaluation of generative models: Inception Score (IS) [9], Fréchet Inception Distance (FID) [7], Kernel Inception Distance (KID) [8], precision and recall [11], density and coverage [12].

To compute the RKE and RRKE scores for the Gaussian mixture cases, we measured the scores based on the output of the trained generator. For image datasets, we followed the standard approach in the literature and evaluated the scores for the representation of the generator's output characterized by an Inception-net-V3 model pre-trained on ImageNet. We note that the Inception-net-based evaluation methodology is consistent with the baseline methods. Also, to select the bandwidth parameter σ for the Gaussian kernel in the RKE and RRKE scores, we performed cross-validation and chose the smallest bandwidth σ for which the reported score's standard deviation across 5,000 validation samples is below 0.01. Note that if the kernel bandwidth becomes overly small, the RKE score will grow almost logarithmically with the number of samples, and the standard deviation of its exponential, i.e. RKE mode count (RKE-MC), will increase almost linearly with the sample size and thus suffer from a large variance across disjoint sample sets. We provide a more detailed discussion of the bandwidth parameter's selection and the resulting variance in the Appendix.

Diversity Evaluation for Synthetic Mixture Datasets. We measured the RKE-MC and RRKE scores for the 8 and 25 component Gaussian mixture datasets in [26]. Figure 1 shows the real samples in blue and GANs' generated samples in red. As the evaluated scores suggest, the RKE-MC scores managed to count the number of captured modes and the RRKE relative distance increased under a

(a) AFHQ generated samples with different ψ.



(b) RKE Mode Count for StyleGAN generated samples on AFHQ and FFHQ varying with truncation factor ψ.

(c) FFHQ generated samples with different ψ.

Figure 2: RKE mode count's behavior under the truncation of the StyleGAN generated samples on AFHQ and FFHQ datasets.

worse coverage of the underlying Gaussian mixture. The rest of our numerical results on mixtures of Gaussians are discussed in the Appendix.

Diversity Evaluation for Real Image Datasets. We measured the proposed and baseline scores on the mentioned image datasets. Table 1 contains the evaluated results. Also, since the IS is a combination of both diversity and quality factors, we propose the following information-theoretic decomposition of IS to IS-quality and IS-diversity:

$$\mathrm{IS}(X) := \exp\big(I(X;\hat{Y})\big) = \exp\big(H(\hat{Y})\big)\exp\big(-H(\hat{Y}|X)\big),$$

where we call $\exp(H(\hat{Y}))$ IS-diversity and $\exp(-H(\hat{Y}|X))$ IS-quality. The decomposed Inception and KID scores are presented in the complete table of our numerical evaluations in the Appendix.

Our numerical results show that the RKE score consistently agrees with the majority of other diversity scores, and also for all GANs remains lower than the actual dataset's RKE. Therefore, due to the estimation guarantee in Theorem 6, our numerical results provide an information-theoretic numerical proof for the empirical result in [15] that applies the birthday paradox to show GAN models cannot capture the full diversity of training data. In addition, the recent generative models StyleGAN-XL and ADM-G achieved the highest RKE, showing their diversity improvement over other generative model baselines.

Also, while the coverage and IS-diversity scores were able to differentiate between the CIFAR-10-trained generative models, WGAN-WC had the lowest score for coverage and recall, despite the Inception score reporting it as the most diverse case. Meanwhile, RKE ranked the absolute diversity of WGAN-WC to be similar to DCGAN while RRKE score shows that the modes captured by WGAN-WC are not common with that of the dataset. In the ImageNet experiments, RKE scores suggested that ContraGAN's samples are more diverse than SAGAN and SNGAN, while its coverage and recall scores were lower than those baselines. However, we note that ContraGAN reached a worse RRKE but its better RKE indicates that it captures a diverse set of modes that could have a smaller intersection with the actual ImageNet modes. The above evaluation of absolute vs. relative diversity of generated samples of ContraGAN was not revealed by the other evaluation metrics.

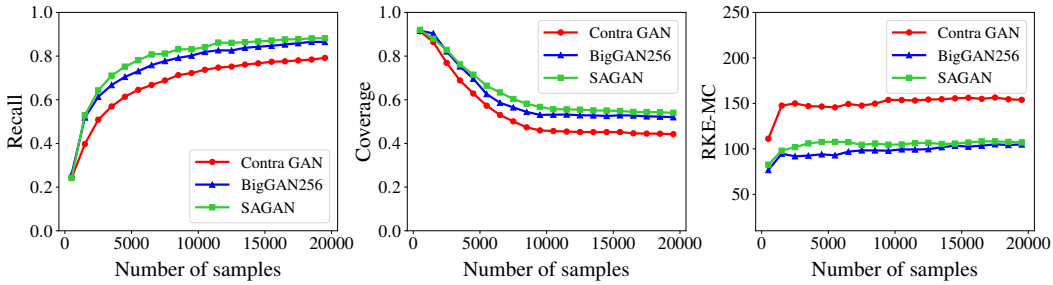


Figure 3: Comparing convergence of Recall, Coverage, and RKE scores on ImageNet dataset.

To assess the correlation between data diversity and RKE score, we repeated the dataset truncation experiment of [11, 12] for the RKE-MC measurement over AFHQ and FFHQ datasets. The numerical results in Figure 2 indicate a significant correlation between the truncation factor and the evaluated RKE. Also, we observed that the empirical RKE-MC scores manage to converge to the underlying RKE-MC using relatively few samples. Figure 3 plots the evaluated RKE-MC, recall, and coverage scores under different sample sizes, which shows RKE-MC can be estimated well with ≈ 2000 data.

Diversity Evaluation for text-to-image generative models. We used the proposed RKE and RRKE scores to evaluate text-to-image generative models. In our experiments, we evaluated the state-of-the-art text-to-image generative model GigaGAN and BK-SDM on the MS-COCO dataset. The numerical results in Table 2 indicate that the GigaGAN model achieves lower RRKE in comparison to BK-SDM. On the other hand, BK-SDM reaches higher RKE-based absolute mode diversity compared to GigaGAN.

Table 2: Zero-shot evaluation on 30K images from MSCOCO validation set for text-to-image generative models. RKE-MC (Mode Count) denotes $\exp(\mathrm{RKE})$.

	Method	IS ↑	FID ↓	**RRKE** ↓	**RKE-MC** ↑
COCO	GigaGAN	33.34	9.09	0.39	58.78
	BK-SDM (Base)	33.79	15.76	0.44	73.05

7 Conclusion

In this work, we proposed a diversity evaluation method for generative models based on entropy measures in quantum information theory. The proposed matrix-based Rényi entropy scores were shown to correlate with the number of modes in a mixture distribution with sub-Gaussian components and can be estimated from empirical data with theoretical guarantees. Our numerical results suggest that while state-of-the-art generative models reach a lower entropy-based diversity score than training data, the recent GAN and diffusion model architectures such as StyleGAN-XL and ADM manage to significantly improve the diversity factor over earlier geneartive models. A future direction for our work is to extend the diversity evaluation to non-GAN and non-diffusion models such as variational autoencoders (VAEs) and flow-based models. Also, studying the effects of the pre-trained Inception model on RKE and RRKE evaluations and comparing their robustness to the choice of pre-trained models vs. the baseline scores studied in [48] will be an interesting topic for future exploration.

Acknowledgments

The work of Farzan Farnia was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920 and by a CUHK Direct Research Grant [CUHK Project No. 4055164]. The work of Cheuk Ting Li was partially supported by an ECS grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.: CUHK 24205621]. The authors also thank the anonymous reviewers and metareviewer for their constructive feedback and suggestions.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc.

[8] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[9] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[10] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. NIPS'18, page 5234–5243, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. *Improved Precision and Recall Metric for Assessing Generative Models*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[12] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[13] Gregg Jaeger. *Quantum information*. Springer, 2007.

[14] 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.

[15] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018.

[16] Ali Borji. Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022.

[17] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 4941–4949, 2017.

[18] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 214–223. JMLR.org, 2017.

[19] L. N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5:64–72, 1969.

[20] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[21] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6069–6078, 2020.

[22] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. LR-GAN: Layered recursive generative adversarial networks for image generation. In *International Conference on Learning Representations*, 2017.

[23] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 17–23 Jul 2022.

[24] Martin Müller-Lennert, Frédéric Dupuis, Oleg Szehr, Serge Fehr, and Marco Tomamichel. On quantum rényi entropies: A new generalization and some properties. *Journal of Mathematical Physics*, 54(12):122203, 2013.

[25] Yuxin Dong, Tieliang Gong, Shujian Yu, and Chen Li. Optimal randomized approximations for matrix-based rényi's entropy. *IEEE Transactions on Information Theory*, 2023.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc.

[27] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009.

[28] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge, CS 231N, 2015.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[30] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[33] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[34] Rewon Child. Very deep {vae}s generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

[35] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[37] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7354–7363, 09–15 Jun 2019.

[38] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[39] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21357–21369. Curran Associates, Inc., 2020.

[40] Minguk Kang, Woohyeon Joseph Shim, Minsu Cho, and Jaesik Park. Rebooting ACGAN: Auxiliary classifier GANs with stable training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[41] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

[42] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.

[43] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[45] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[46] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: Architecturally compressed stable diffusion for efficient text-to-image generation. *ICML Workshop on Efficient Systems for Foundation Models (ES-FoMo)*, 2023.

[47] MinGuk Kang, Joonghyuk Shin, and Jaesik Park. StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[48] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

A Appendix

A.1 Proof of Proposition 1

The proposition is directly implied by the fact that the eigenvalues of $XX^\top = X^2$ are $\lambda_1^2, \ldots, \lambda_d^2$. Therefore, we have

$$\|X\|_F^2 = \mathrm{Tr}(XX^\top) = \mathrm{Tr}(X^2) = \sum_{i=1}^{d} \lambda_i^2.$$

The proof is therefore a direct consequence of the definition of order-2 Rényi entropy.

A.2 Proof of Theorem 1

We apply Theorem 5 which reveals that for a Gaussian kernel bandwidth of $\sqrt{2}\sigma$ the following holds. Note that, without loss of generality and for simplicity of theoretical derivations, we derive the equations for a bandwidth of $\sqrt{2}\sigma$:

$$\begin{aligned}
\mathrm{RKE}_2^{\mathrm{G}_{\sqrt{2}\sigma}}(\mathbf{X}) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P_X}\left[k^2_{\sqrt{2}\sigma}(\mathbf{X},\mathbf{X}')\right] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \mathbb{E}_{\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu}_i,\Sigma_i),X'\sim\mathcal{N}(\boldsymbol{\mu}_j,\Sigma_j)}\left[k^2_{\sqrt{2}\sigma}(\mathbf{X},\mathbf{X}')\right] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \int \frac{-1}{\sqrt{(2\pi)^{2d}\det(\Sigma_i)\det(\Sigma_j)}} \\
&\quad \times \exp\Big(\frac{-1}{2}\Big(\|\mathbf{x}-\boldsymbol{\mu}_i\|^2_{\Sigma_i^{-1}} + \|\mathbf{x}'-\boldsymbol{\mu}_j\|^2_{\Sigma_j^{-1}} + \sigma^{-2}\|\mathbf{x}-\mathbf{x}'\|_2^2\Big)\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \int \frac{-1}{\sqrt{(2\pi)^{2d}\det(\Sigma_i)\det(\Sigma_j)}} \times \exp\Big(\frac{-1}{2}\Big(\|\mathbf{x}-\boldsymbol{\mu}_i\|^2_{\Sigma_i^{-1}} \\
&\quad + \|\mathbf{x}'-\boldsymbol{\mu}_j\|^2_{\Sigma_j^{-1}} + \sigma^{-2}\big\|\big(\mathbf{x}-\boldsymbol{\mu}_i\big) - \big(\mathbf{x}'-\boldsymbol{\mu}_j\big) + \big(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\big)\big\|_2^2\Big)\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \int \frac{1}{\sqrt{(2\pi)^{2d}\det(\Sigma_i\Sigma_j)}} \\
&\quad \times \exp\Big(\frac{-1}{2}\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix}^\top \begin{bmatrix}\Sigma_i^{-1}+\sigma^{-2}I & -\sigma^{-2}I\\ -\sigma^{-2}I & \Sigma_j^{-1}+\sigma^{-2}I\end{bmatrix}\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix} \\
&\quad + \begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix}^\top \begin{bmatrix}\frac{1}{\sigma^2}\big(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i\big)\\ \frac{1}{\sigma^2}\big(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\big)\end{bmatrix} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&\overset{(a)}{=} \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \int \frac{1}{\sqrt{(2\pi)^{2d}\det(\Sigma_i\Sigma_j)}} \exp\Big(\frac{-1}{2}\Big(\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix} - \mathbf{b}_{i,j}\Big)^\top C_{i,j}\Big(\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix} - \mathbf{b}_{i,j}\Big) \\
&\quad + \frac{1}{2}\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \exp\Big(\frac{1}{2}\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2\Big) \int \frac{1}{\sqrt{(2\pi)^{2d}\det(\Sigma_i\Sigma_j)}} \\
&\quad \times \exp\Big(\frac{-1}{2}\Big(\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix} - \mathbf{b}_{i,j}\Big)^\top C_{i,j}\Big(\begin{bmatrix}\mathbf{x}-\boldsymbol{\mu}_i\\ \mathbf{x}'-\boldsymbol{\mu}_j\end{bmatrix} - \mathbf{b}_{i,j}\Big)\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \\
&\overset{(b)}{=} \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \exp\Big(\frac{1}{2}\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2\Big) \frac{\sqrt{(2\pi)^{2d}\det(C_{i,j}^{-1})}}{\sqrt{(2\pi)^{2d}\det(\Sigma_i\Sigma_j)}}
\end{aligned}$$

$$\overset{(c)}{=} \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \exp\Big(\frac{1}{2}\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2\Big) \frac{1}{\sqrt{\det(\begin{bmatrix}\Sigma_i & 0\\ 0 & \Sigma_j\end{bmatrix} C_{i,j})}}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} \omega_i\omega_j \exp\Big(\frac{1}{2}\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} - \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2\Big) \frac{1}{\sqrt{\det(\begin{bmatrix}I+\sigma^{-2}\Sigma_i & -\sigma^{-2}\Sigma_i\\ -\sigma^{-2}\Sigma_j & I+\sigma^{-2}\Sigma_j\end{bmatrix})}}$$

$$\overset{(d)}{=} \sum_{i=1}^{k}\sum_{j=1}^{k}\Bigg[\omega_i\omega_j \exp\Big(\frac{1}{2\sigma^2}\big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big)^\top\big((I - A_{i,j}) - I\big)\big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big)\Big) \times \det\Big(I + \frac{1}{\sigma^2}(\Sigma_i + \Sigma_j)\Big)^{-1/2}\Bigg]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k}\Bigg[\omega_i\omega_j \exp\Big(-\frac{1}{2\sigma^2}\big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big)^\top A_{i,j}\big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\big)\Big) \det\Big(I + \frac{1}{\sigma^2}(\Sigma_i + \Sigma_j)\Big)^{-1/2}\Bigg]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k}\Bigg[\omega_i\omega_j e^{-\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2_{A_{i,j}}}{2\sigma^2}} \det\Big(I + \frac{1}{\sigma^2}(\Sigma_i + \Sigma_j)\Big)^{-1/2}\Bigg],$$

Here (a) comes from defining $C_{i,j} = \begin{bmatrix}\Sigma_i^{-1} + \sigma^{-2}I & -\sigma^{-2}I\\ -\sigma^{-2}I & \Sigma_j^{-1} + \sigma^{-2}I\end{bmatrix}$ and $\mathbf{b}_{i,j} = \frac{1}{\sigma^2}C_{i,j}^{-1}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}$. (b) follows from the unit integral of a multivariate Gaussian probability density function with mean vector $\mathbf{b}_{i,j}$ and covariance matrix $C_{i,j}^{-1}$. (c) uses the determinant of Block diagonal matrices as $\det(\begin{bmatrix}\Sigma_i & 0\\ 0 & \Sigma_j\end{bmatrix}) = \det(\Sigma_i)\det(\Sigma_j) = \det(\Sigma_i\Sigma_j)$. (d) is based on the determinant of block matrices $\begin{bmatrix}A & B\\ C & D\end{bmatrix}$ that if $CD = DC$ (which holds in our case as $-\sigma^{-2}\Sigma_j$ and $I + \sigma^{-2}\Sigma_j$ commute) then $\det(\begin{bmatrix}A & B\\ C & D\end{bmatrix}) = \det(AD - BC)$, and furthermore the fact that by defining $F_i = \sigma^2\Sigma_i^{-1}$ and $F_j = \sigma^2\Sigma_j^{-1}$ we will have

$$\begin{aligned}
\mathbf{b}_{i,j}^\top C_{i,j}\mathbf{b}_{i,j} &= \frac{1}{\sigma^4}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}^\top C_{i,j}^{-1}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}\\
&= \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}^\top \big(\sigma^2 C_{i,j}\big)^{-1}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}\\
&= \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}^\top \begin{bmatrix}I + \sigma^2\Sigma_i^{-1} & -I\\ -I & I + \sigma^2\Sigma_j^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}\\
&= \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}^\top \begin{bmatrix}I + F_i & -I\\ -I & I + F_j\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}\\
&\overset{(e)}{=} \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}^\top \begin{bmatrix}\big(I + F_i - (I + F_j)^{-1}\big)^{-1} & (F_i + F_j + F_iF_j)^{-1}\\ (F_i + F_j + F_jF_i)^{-1} & \big(I + F_j - (I + F_i)^{-1}\big)^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\end{bmatrix}\\
&= \frac{1}{\sigma^2}\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)^\top\Bigg(\big(I + F_i - (I + F_j)^{-1}\big)^{-1}\\
&\quad + \big(I + F_j - (I + F_i)^{-1}\big)^{-1} - (F_i + F_j + F_iF_j)^{-1} - (F_i + F_j + F_jF_i)^{-1}\Bigg)\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)\\
&= \frac{1}{\sigma^2}\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)^\top\Bigg(\big(I + \sigma^2\Sigma_i^{-1} - (I + \sigma^2\Sigma_j^{-1})^{-1}\big)^{-1} + \big(I + \sigma^2\Sigma_j^{-1} - (I + \sigma^2\Sigma_i^{-1})^{-1}\big)^{-1}
\end{aligned}$$

$$- (\sigma^2\Sigma_i^{-1} + \sigma^2\Sigma_j^{-1} + \sigma^4\Sigma_i^{-1}\Sigma_j^{-1})^{-1} - (\sigma^2\Sigma_i^{-1} + \sigma^2\Sigma_j^{-1} + \sigma^4\Sigma_j^{-1}\Sigma_i^{-1})^{-1}\Big)\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)$$
$$= \frac{1}{\sigma^2}\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)^\top\Big(I - A_{i,j}\Big)\Big(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\Big)$$

In the above, (e) holds because F_i and F_j are positive definite matrices for the supposed invertible and thus positive definite Σ_i and Σ_j. Hence, the matrices $I + F_i$ and $I + F_j$ are both positive definite and invertible. Also, the Schur complement $(I + F_j) - (-I)(I + F_i)^{-1}(-I) = I + F_j - (I + F_i)^{-1}$ will be a positive definite and invertible matrix because $(I + F_i)^{-1} \prec I \prec I + F_j$. Therefore, for the inverse of the block matrix, we will have

$$\begin{bmatrix} I + F_i & -I \\ -I & I + F_j \end{bmatrix}^{-1} = \begin{bmatrix} \big(I + F_i - (I + F_j)^{-1}\big)^{-1} & (F_i + F_j + F_iF_j)^{-1} \\ (F_i + F_j + F_jF_i)^{-1} & \big(I + F_j - (I + F_i)^{-1}\big)^{-1} \end{bmatrix}.$$

The above discussion completes the proof.

A.3 Proof of Theorem 2

Note that according to the definition of the kernel covariance matrix we have

$$C_X - \sum_{i=1}^{k} \omega_i \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top = \sum_{i=1}^{k}\Big[\omega_i\Big(\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top\big] - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big)\Big].$$

Therefore, applying Jensen's inequality for the convex Frobenius norm-squared function $\|\cdot\|_F^2$ shows that

$$\begin{aligned}
&\Big\|C_X - \sum_{i=1}^{k} \omega_i \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_F^2 \\
&= \Big\|\sum_{i=1}^{k}\Big[\omega_i\Big(\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top\big] - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big)\Big]\Big\|_F^2 \\
&\le \sum_{i=1}^{k}\Big[\omega_i\Big\|\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top\big] - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_F^2\Big] \\
&= \sum_{i=1}^{k}\Big[\omega_i\Big\|\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big]\Big\|_F^2\Big] \\
&\le \sum_{i=1}^{k}\omega_i\mathbb{E}\Big[\Big\|\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_F^2\Big] \\
&= \sum_{i=1}^{k}\omega_i\mathbb{E}\Big[2 - 2\big(\phi(\boldsymbol{\mu}_i)^\top\phi(\boldsymbol{\mu}_i + Z_i)\big)^2\Big] \\
&= \sum_{i=1}^{k} 2\omega_i\mathbb{E}\Big[1 - \exp\big(\frac{\|\mathbf{Z}_i\|_2^2}{\sigma^2}\big)\Big] \\
&\le \sum_{i=1}^{k} 2\omega_i\big(1 - \frac{1}{\alpha_i^d}\big).
\end{aligned}$$

Note that the inequality before the last holds, since $\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top$ is a rank two matrix where $\mathbf{a} = \phi(\boldsymbol{\mu}_i + Z_i)$ and $\mathbf{b} = \phi(\boldsymbol{\mu}_i)$ have both unit norms. Therefore, for the Frobenius norm-squared of $\mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top$ will be equal to $2 - 2(\mathbf{a}^\top\mathbf{b})^2$. Therefore, we have

$$\Big|\big\|C_X\big\|_F - \big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F\Big| \le \Big\|C_X - \sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_F \le \sqrt{\sum_{i=1}^{k} 2\omega_i\big(1 - \frac{1}{\alpha_i^d}\big)}.$$

On the other hand, the special zero-covariance case of Theorem 1 shows that

$$\big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F^2=\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j\exp(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2})$$

Also, we know that the Gaussian kernel is always upper-bounded by 1 and thus $\big\|C_X\big\|_F+\big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F\leq 2$. Therefore, knowing that $\mathrm{RKE}_2^{G_\sigma}(\mathbf{X})=-\log\big(\|C_X\|_F^2\big)$ shows

$$\begin{aligned}
&\Big|\exp\big(-\mathrm{RKE}_2^{G_\sigma}(\mathbf{X})\big)-\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j\exp\big(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2}\big)\Big|\\
=&\Big|\big\|C_X\big\|_F^2-\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_i\omega_j\exp\big(-\frac{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|_2^2}{\sigma^2}\big)\Big|\\
=&\Big|\big\|C_X\big\|_F^2-\big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F^2\Big|\\
=&\Big|\big\|C_X\big\|_F+\big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F\Big|\times\Big|\big\|C_X\big\|_F-\big\|\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big\|_F\Big|\\
\leq& 2\Big\|C_X-\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_F\\
\leq&\sqrt{\sum_{i=1}^{k}8\omega_i\big(1-\frac{1}{\alpha_i^d}\big)}.
\end{aligned}$$

The theorem's proof is hence complete.

A.4 Proof of Theorem 3

By the continuity of P_X, we have

$$\begin{aligned}
&\lim_{\sigma\to 0}\left(\pi\sigma^2\right)^{-d/2}\mathbb{E}\left[k_\sigma^2(X,X')\right]\\
=&\lim_{\sigma\to 0}\left(\pi\sigma^2\right)^{-d/2}\mathbb{E}\left[\int k_\sigma^2(X,x')P_X(x')\mathrm{d}x'\right]\\
=&\lim_{\sigma\to 0}\mathbb{E}\left[\int\left(\pi\sigma^2\right)^{-d/2}\exp\left(-\frac{\|X-x'\|_2^2}{\sigma^2}\right)P_X(x')\mathrm{d}x'\right]\\
=&\,\mathbb{E}\left[\lim_{\sigma\to 0}\int\left(\pi\sigma^2\right)^{-d/2}\exp\left(-\frac{\|X-x'\|_2^2}{\sigma^2}\right)P_X(x')\mathrm{d}x'\right]\\
=&\,\mathbb{E}\left[P_X(X)\right]\\
=&\int P_X(x)^2\mathrm{d}x.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\lim_{\sigma\to 0}\left(\mathrm{RKE}_2(P_X)+\frac{d}{2}\log(\pi\sigma^2)\right)\\
=&\lim_{\sigma\to 0}\left(-\log\left(\left(\pi\sigma^2\right)^{-d/2}\mathbb{E}\left[k_\sigma^2(X,X')\right]\right)\right)\\
=&-\log\big(\int P_X(x)^2\mathrm{d}x\big).
\end{aligned}$$

A.5 Proof of Theorem 4

Using $\|\cdot\|_*$ to denote the nuclear norm, the following holds for every two PSD matrices $A, B \in \mathbb{R}^{d\times d}$ [24]:

$$\mathrm{Tr}\big(\sqrt{A^{1/2}BA^{1/2}}\big) \;=\; \mathrm{Tr}\big(\sqrt{A^{1/2}B^{1/2}\big(A^{1/2}B^{1/2}\big)^\top}\big) \;=\; \big\|\sqrt{A}\sqrt{B}\big\|_*. \tag{1}$$

Therefore, we can write

$$\begin{aligned}
&\Big|\mathrm{Tr}\Big(\sqrt{\sqrt{C_X}C_Y\sqrt{C_X}}\Big) - \mathrm{Tr}\Big(\sqrt{\sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\Big(\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top\Big)\sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}}\Big)\Big| \\
&\overset{(a)}{=} \Big|\Big\|\sqrt{C_X}\sqrt{C_Y}\Big\|_* - \Big\|\sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_*\Big| \\
&\overset{(b)}{\le} \Big|\Big\|\sqrt{C_X}\sqrt{C_Y}\Big\|_* - \Big\|\sqrt{C_X}\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_*\Big| \\
&\qquad + \Big|\Big\|\sqrt{C_X}\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_* - \Big\|\sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_*\Big| \\
&\overset{(c)}{\le} \Big\|\sqrt{C_X}\Big(\sqrt{C_Y} - \sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big)\Big\|_* \\
&\qquad + \Big\|\Big(\sqrt{C_X} - \sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\Big)\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_* \\
&\overset{(d)}{\le} \big\|\sqrt{C_X}\big\|_F\Big\|\sqrt{C_Y} - \sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_F \\
&\qquad + \Big\|\sqrt{C_X} - \sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\Big\|_F\Big\|\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_F \\
&\overset{(e)}{=} \Big\|\sqrt{C_Y} - \sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}\Big\|_F + \Big\|\sqrt{C_X} - \sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\Big\|_F \\
&\overset{(f)}{\le} \sqrt{\Big\|C_Y - \sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top\Big)\Big\|_*} + \sqrt{\Big\|C_X - \sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_*}
\end{aligned}$$

In the above, (a) holds due to the identity discussed in (1). (b) and (c) follow from the application of triangle inequality for the absolute value and nuclear norm functions, respectively. (d) is an application of Holder's inequality for Schatten norms where $\|AB\|_* \le \|A\|_F\|B\|_F$ for every pair of matrices A, B. (e) holds because all the four matrices $\sqrt{C_X}$, $\sqrt{C_Y}$, $\sqrt{\sum_{i=1}^k \omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}$, $\sqrt{\sum_{i=1}^k \eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top}$ have a unit Frobenius norm since the square of their eigenvalues will be the eigenvalues of a normalized kernel covariance matrix. Finally, (f) is the result of the matrix norm inequality that $\|\sqrt{A} - \sqrt{B}\|_F \le \sqrt{\|A-B\|_*}$ for every pair of PSD matrices A, B.

To further simplify the above upper-bound Next, we apply the Jensen's inequality for the convex nuclear norm-squared function which shows that

$$
\begin{aligned}
&\Big\|C_X - \sum_{i=1}^{k} \omega_i \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_*^2 \\
&= \Big\|\sum_{i=1}^{k}\Big[\omega_i\Big(\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top\big] - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big)\Big]\Big\|_*^2 \\
&\leq \sum_{i=1}^{k}\Big[\omega_i\Big\|\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top\big] - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_*^2\Big] \\
&= \sum_{i=1}^{k}\Big[\omega_i\Big\|\mathbb{E}\big[\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\big]\Big\|_*^2\Big] \\
&\leq \sum_{i=1}^{k}\Big[\omega_i\mathbb{E}\Big[\Big\|\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_*^2\Big]\Big] \\
&\leq \sum_{i=1}^{k}\Big[\omega_i\mathbb{E}\Big[4 - 4\big(\phi(\boldsymbol{\mu}_i)^\top\phi(\boldsymbol{\mu}_i + Z_i)\big)^2\Big] \\
&= \sum_{i=1}^{k} 4\omega_i\mathbb{E}\Big[1 - \exp\big(\frac{\|\mathbf{Z}_i\|_2^2}{\sigma^2}\big)\Big] \\
&\leq \sum_{i=1}^{k} 4\omega_i\big(1 - \frac{1}{\alpha_i^d}\big).
\end{aligned}
$$

The inequality before the last holds because $\phi(\boldsymbol{\mu}_i + Z_i)\phi(\boldsymbol{\mu}_i + Z_i)^\top - \phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top$ is a rank two matrix with Frobenius norm-squared of $2 - 2(\phi(\boldsymbol{\mu}_i + Z_i)^\top\phi(\boldsymbol{\mu}_i))^2$, and since the matrix's rank is bounded by 2, its nuclear norm will be upper-bounded by $\sqrt{2}$ times its Frobenius norm. As a result of the above inequality, we can write

$$
\begin{aligned}
&\Big|\mathrm{Tr}\Big(\sqrt{\sqrt{C_X}C_Y\sqrt{C_X}}\Big) - \mathrm{Tr}\Big(\sqrt{\sqrt{\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}\Big(\sum_{i=1}^{k}\eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top\Big)\sqrt{\sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top}}\Big)\Big| \\
&\leq \sqrt{\Big\|C_Y - \sum_{i=1}^{k}\eta_i\phi(\boldsymbol{\zeta}_i)\phi(\boldsymbol{\zeta}_i)^\top\Big)\Big\|_*} + \sqrt{\Big\|C_X - \sum_{i=1}^{k}\omega_i\phi(\boldsymbol{\mu}_i)\phi(\boldsymbol{\mu}_i)^\top\Big\|_*} \\
&\leq \sqrt[4]{\sum_{i=1}^{k} 4\omega_i\big(1 - \frac{1}{\alpha_i^d}\big)} + \sqrt[4]{\sum_{i=1}^{k} 4\eta_i\big(1 - \frac{1}{\alpha_i^d}\big)} \\
&\leq \sqrt[4]{\sum_{i=1}^{k} 32(\omega_i + \eta_i)\big(1 - \frac{1}{\alpha_i^d}\big)}
\end{aligned}
$$

where the last inequality holds as $\sqrt[4]{a} + \sqrt[4]{b} \leq \sqrt[4]{8(a+b)}$ holds for every $a, b \geq 0$, which is a consequence of $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ for every $a, b \geq 0$. The proof is therefore complete.

A.6 Proof of Theorem 5

Note that according to the definition of kernel covariance matrix we have

$$
\mathrm{Tr}\big(C_X C_X^\top\big) \;=\; \mathrm{Tr}\left(\int p(\mathbf{x})\phi(\mathbf{x})\phi(\mathbf{x})^\top \mathrm{d}\mathbf{x} \int p(\mathbf{x}')\phi(\mathbf{x}')\phi(\mathbf{x}')^\top \mathrm{d}\mathbf{x}'\right)
$$

$$
\begin{aligned}
&= \operatorname{Tr}\left(\int p(\mathbf{x})p(\mathbf{x}')\phi(\mathbf{x})\phi(\mathbf{x})^\top\phi(\mathbf{x}')\phi(\mathbf{x}')^\top \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\right)\\
&= \int p(\mathbf{x})p(\mathbf{x}')\operatorname{Tr}\Big(\phi(\mathbf{x})\phi(\mathbf{x})^\top\phi(\mathbf{x}')\phi(\mathbf{x}')^\top\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\\
&= \int p(\mathbf{x})p(\mathbf{x}')\operatorname{Tr}\Big(\phi(\mathbf{x}')^\top\phi(\mathbf{x})\phi(\mathbf{x})^\top\phi(\mathbf{x}')\Big)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\\
&= \int p(\mathbf{x})p(\mathbf{x}')k(\mathbf{x}',\mathbf{x})k(\mathbf{x},\mathbf{x}')\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\\
&= \int p(\mathbf{x})p(\mathbf{x}')k(\mathbf{x},\mathbf{x}')^2\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'\\
&= \mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P_X}\big[k(\mathbf{X},\mathbf{X}')^2\big],
\end{aligned}
$$

where the last line holds since the joint density function of independent $\mathbf{X}, \mathbf{X}'$ will be the product of the marginal density functions. Given the above result, the theorem is a direct consequence of Proposition 1.

A.7 Proof of Theorem 6

As shown in Theorem 5, we have

$$\exp\big(-\mathrm{RKE}_2(\mathbf{X})\big) = \mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P_X}\big[k^2(\mathbf{X},\mathbf{X}')\big].$$

As a result, we obtain the following:

$$\exp\big(-\widehat{\mathrm{RKE}}_2(\mathbf{X})\big) - \exp\big(-\mathrm{RKE}_2(\mathbf{X})\big) \;=\; \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\mathbf{x}_i,\mathbf{x}_j)^2 - \mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P_X}\big[k^2(\mathbf{X},\mathbf{X}')\big]$$

Note that according to the Cauchy-Schwarz inequality, for every $\mathbf{x},\mathbf{y}\in\mathcal{X}$ we have $k(\mathbf{x},\mathbf{y})^2 \le k(\mathbf{x},\mathbf{x})k(\mathbf{y},\mathbf{y}) = 1$, and therefore for a normalized kernel we will have $0 \le k(\mathbf{x},\mathbf{y}) \le 1$ at every $\mathbf{x},\mathbf{y}\in\mathcal{X}$. Therefore, if $\mathbf{z}_1,\ldots,\mathbf{z}_n$ are n IID samples from P_X which are also independent from $\mathbf{x}_1,\ldots,\mathbf{x}_n$, we will have:

$$\left|\frac{1}{n^2}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{x}_i)^2 - \frac{1}{n^2}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{z}_i)^2\right| \le \frac{n}{n^2} - \frac{0}{n^2} = \frac{1}{n}.$$

On the other hand, we know a complete graph of size n can be decomposed to $r = \lceil\frac{n}{2}\rceil$ matchings $m_1,\ldots,m_r$ with $[\frac{n}{2}]$ edges. We will have the following identity given these matchings

$$\frac{1}{n^2}\sum_{i\neq j} k(\mathbf{x}_i,\mathbf{x}_j)^2 \;=\; \frac{1}{n^2}\sum_{i=1}^r\sum_{t=1}^{[\frac{n}{2}]} k(\mathbf{x}_{m_i(t,0)},\mathbf{x}_{m_i(t,1)}).$$

As a result, we have

$$\left|\big[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\mathbf{x}_i,\mathbf{x}_j)^2\big] - \big[\frac{1}{n^2}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{z}_i)^2\big] - \big[\frac{1}{n^2}\sum_{i=1}^r\sum_{t=1}^{[\frac{n}{2}]} k(\mathbf{x}_{m_i(t,0)},\mathbf{x}_{m_i(t,1)})\big]\right| \le \frac{1}{n}$$

Now, we note that $\{(\mathbf{x}_i,\mathbf{z}_i)_{i=1}^n\}$ are n independent samples of $(\mathbf{X},\mathbf{X}')$, and an application of the Hoeffding's inequality implies that with probability $1-\delta/n$ we have the following

$$\left|\frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\mathbf{z}_i)^2 - \mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P_X}\big[k^2(\mathbf{X},\mathbf{X}')\big]\right| \le \sqrt{\frac{2\log(n/\delta)}{n}}.$$

Similarly, every matching m_t provides $[\frac{n}{2}]$ independent sample pairs of $(\mathbf{X},\mathbf{X}')$, which implies that with probability $1-\delta/n$

$$\left|\frac{1}{[n/2]}\sum_{i=1}^{[n/2]} k(\mathbf{x}_{m_t(i,0)},\mathbf{x}_{m_t(i,1)})^2 - \mathbb{E}_{X,X'\overset{\text{iid}}{\sim}P_X}\big[k^2(\mathbf{X},\mathbf{X}')\big]\right| \le \sqrt{\frac{4\log(n/\delta)}{n}}.$$

Given the above bounds, an application of the union bound shows that with probability at least $1 - n \times \frac{\delta}{n} = 1 - \delta$, we will have the following

$$\left| \mathbb{E}_{X,X' \overset{\text{iid}}{\sim} P_X} \left[k^2(\mathbf{X}, \mathbf{X}') \right] - \left[\frac{1}{n^2} \sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{z}_i)^2 \right] - \left[\frac{1}{n^2} \sum_{i=1}^{r} \sum_{t=1}^{\lceil \frac{n}{2} \rceil} k(\mathbf{x}_{e_i(t,0)}, \mathbf{x}_{e_i(t,1)}) \right] \right| \leq O\Big(\sqrt{\frac{\log(n/\delta)}{n}} \Big),$$

which can be combined with the mentioned upper-bound to show that with probability at least $1 - \delta$ the following holds

$$\begin{aligned} \Big| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j)^2 - \mathbb{E}_{X,X' \overset{\text{iid}}{\sim} P_X} \left[k^2(\mathbf{X}, \mathbf{X}') \right] \Big| &\leq O\Big(\sqrt{\frac{\log(n/\delta)}{n}} + \frac{1}{n} \Big) \\ &= O\Big(\sqrt{\frac{\log(n/\delta)}{n}} \Big). \end{aligned}$$

Therefore, the proof is complete.

A.8 Proof of Theorem 7

Note that according to our definition, we will have

$$K_{XY} = \Phi_X \Phi_Y^\top.$$

We consider the SVD decomposition of matrices $\Phi_X = U_X S_X V_X^\top$ and $\Phi_Y = U_Y S_Y V_Y^\top$ where U_X, U_Y are unitary matrices in $\mathbb{R}^{n \times n}$, V_X, V_Y are unitary matrices in $\mathbb{R}^{d \times d}$, and S_X, S_Y are semi-diagonal matrices in $\mathbb{R}^{n \times d}$. Then, we will have

$$K_{XY} = U_X S_X V_X^\top V_Y S_Y U_Y^\top.$$

Also, we can obtain that

$$C_X = \Phi_X^\top \Phi_X = V_X S_X^\top S_X V_X^\top, \quad C_Y = \Phi_Y^\top \Phi_Y = V_Y S_Y^\top S_Y V_Y^\top.$$

As a result, we will have the following

$$\sqrt{C_X}\sqrt{C_Y} = V_X \sqrt{S_X^\top S_X} V_X^\top V_Y \sqrt{S_Y^\top S_Y} V_Y^\top$$

However, since S_X, S_Y are semi-diagonal and U_X, U_Y are unitary matrices, we will have the same set of non-zero singular-values for the following two matrices

$$\text{singular.values}\big(V_X \sqrt{S_X^\top S_X} V_X^\top V_Y \sqrt{S_Y^\top S_Y} V_Y^\top \big) = \text{singular.values}\big(U_X S_X V_X^\top V_Y S_Y U_Y^\top \big)$$

Therefore, K_{XY} shares the same singular values with $\sqrt{C_X}\sqrt{C_Y}$. Therefore, we will have

$$\text{Tr}(\sqrt{\sqrt{C_X} C_Y \sqrt{C_X}}) = \sum_{i=1}^{d} s_i(\sqrt{C_X}\sqrt{C_Y}) = \sum_{i=1}^{d} s_i(K_{XY}) = \|K_{XY}\|_*,$$

which due to the definition of order-$\frac{1}{2}$ RRKE score completes the proof.

A.9 Additional Experimental Results

A.9.1 Effect of bandwidth on RKE

We show the effect of different bandwidths on CIFAR10, Tiny-ImageNet, and ImageNet datasets in Figure 4. This plot indicates that the ranking of the models remains consistent for different bandwidth parameters in the range $\sigma \in [0.1, 0.5]$. It is important to note that for bandwidth values $\sigma > 0.5$, the Gaussian kernel assigns near-zero values to almost every pair of input samples, and therefore all the RKE mode count values are close to 1. On the other hand, for smaller $\sigma \approx 0$ bandwidth values, every data point would be counted as a separate mode (high sensitivity to between samples distances). We also experimented the effect of different bandwidths on StyleGAN3 with different truncation factors in Figure 5.

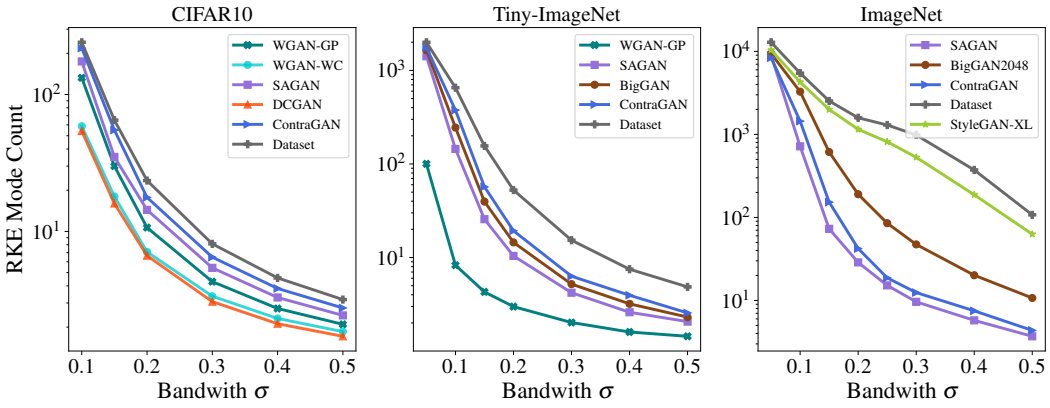
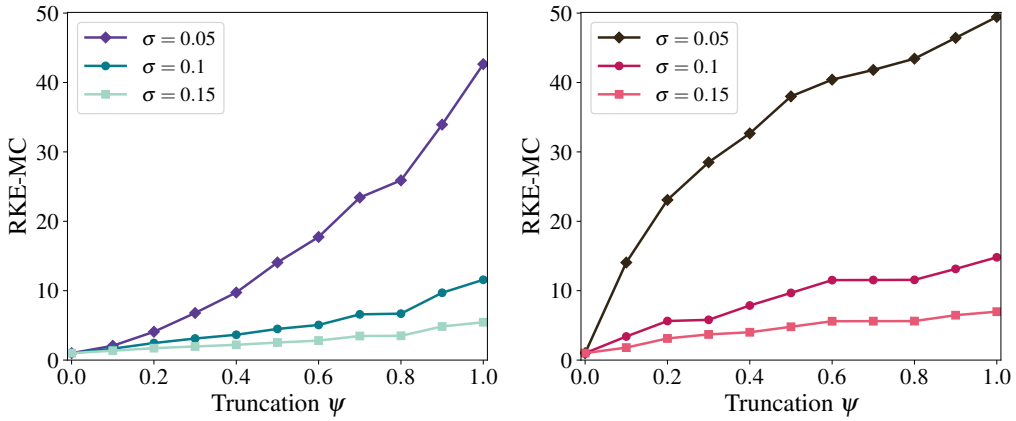


Figure 4: Effect of the bandwidth σ on the numerical evaluation



(a) AFHQ generated samples with different ψ.

(b) FFHQ generated samples with different ψ.

Figure 5: Effect of the bandwidth σ on StyleGAN3 with different truncation factor on AFHQ and FFHQ dataset.

A.9.2 Can existing metrics count the number of modes?

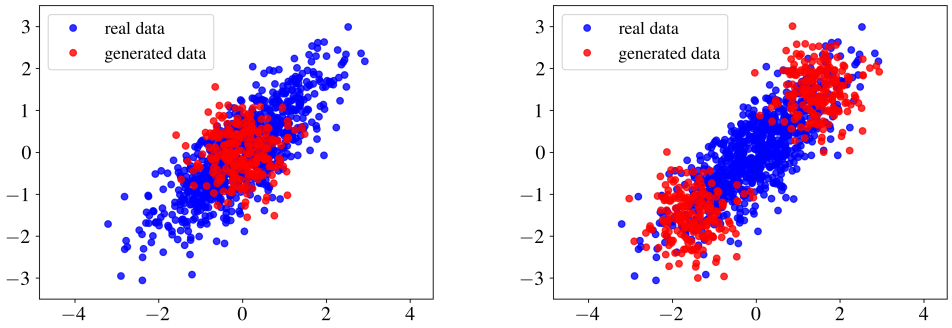
In our experiment on Gaussian distributions, we observed that existing metrics like Recall and Coverage are unable to quantify the number of modes. As shown in Figure 6, we have two generated data, one with a single mode and the other with two Gaussian modes. Recall and Coverage yielded the same results for both datasets. However, RKE-MC demonstrated the capability to count the number of modes in this experiment accurately.

A.9.3 Different orders of Rényi entropy scores in applications of the RKE evaluation

We evaluated the matrix-based Rényi entropy score of different orders and summarized the numerical results in Figure 7. As shown in this figure, order-2 Renyi entropy can successfully distinguish the diversity performance of BigGAN-2048 and SAGAN.

A.9.4 Comparison between our proposed algorithms for computing the RKE score and other algorithms

In our numerical evaluation of the RKE score, we focused on order-2 matrix-based Renyi entropy which reduces to the Frobenius norm of the kernel matrix. This algorithm will require computation for samples of dimension. In addition, Theorem 5 implies a randomized algorithm estimating the expected value using empirical samples which requires computation for pairs of fresh empirical samples. On the other hand, Dong et al. [25]'s computation method applies to a general order-α matrix-based Renyi entropy.



(a) Recall: 0.77, Coverage: 0.42 (b) Recall: 0.74, Coverage: 0.42

Figure 6: Recall and Coverage can not count the number of modes.

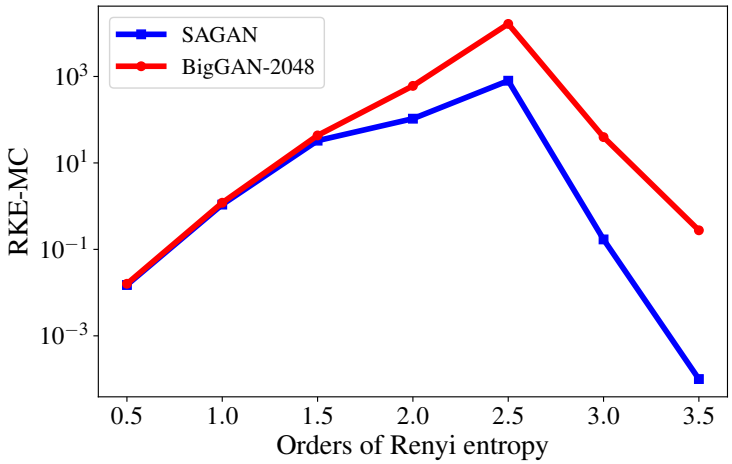


Figure 7: Effect of different Renyi entropy orders in RKE-MC in application to ImageNet data

The table 3 shows the time (in seconds) taken by the three algorithms in performing the computation of order-2 Renyi entropy. The results in the table indicate that Dong et al.'s method and the Frobenius norm-based approach result in similar time complexity, while the randomized algorithm based on empirical expected value can significantly reduce the computation time.

A.9.5 RKE-MC During Training

We trained the ContraGAN and SNGAN on CIFAR-10 data and recorded the evaluation scores every 2,000 generator iterations. As shown in Figure 8, RKE increased during the training. In Figure 9, we can see the diversity of generated samples for 10 classes of CIFAR10 during the training of ContraGAN.

Table 3: Comparison between algorithms for computing the RKE score

Algorithms	1000 samples	2000 samples	3000 samples
Frobenius norm (Ours)	9.12	36.35	81.85
Empirical expected value (Ours)	2.19	3.25	5.10
Dong et al. (Hutch++ based)	8.97	35.9	82.08

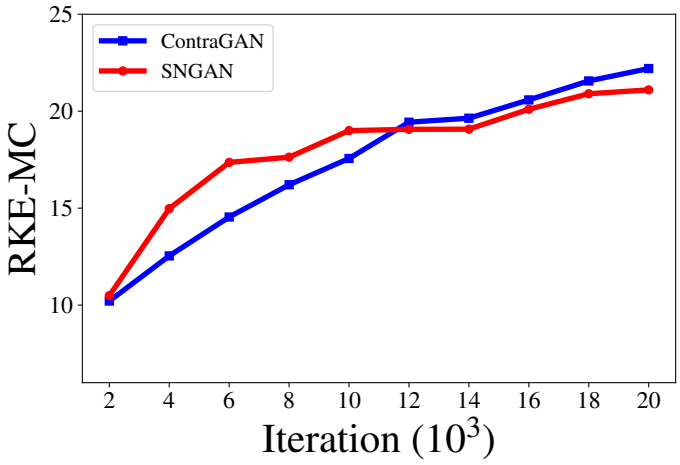


Figure 8: RKE Mode Count during training of ContraGAN and SNGAN on CIFAR10 dataset.

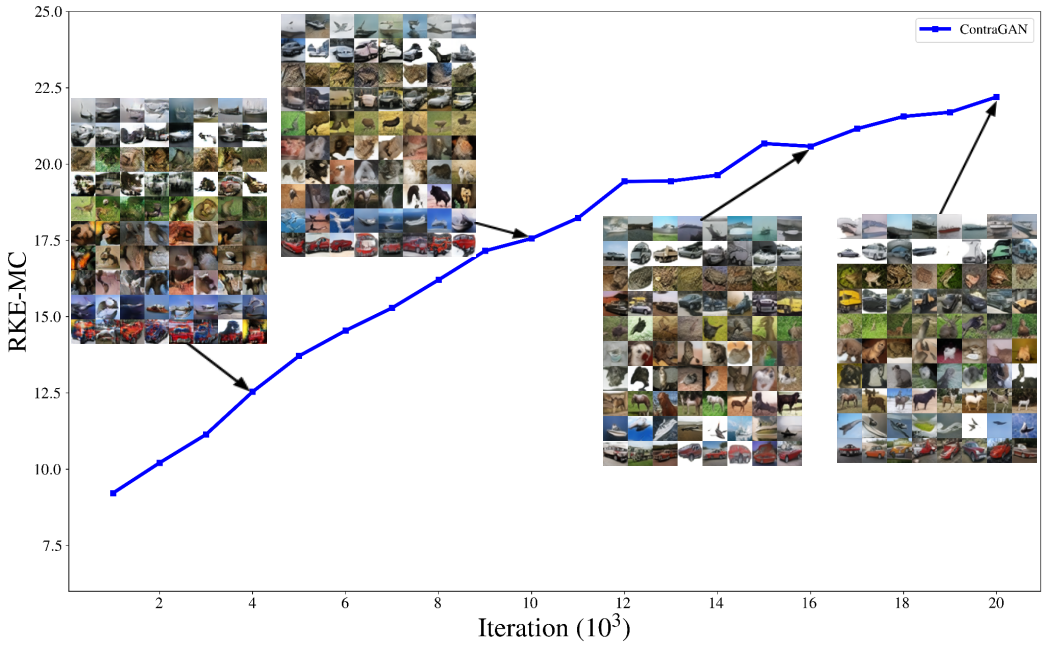


Figure 9: RKE Mode Count during training of Contra-GAN on CIFAR10 dataset with generated samples from 10 classes.

Table 4: Absolute evaluation for three image datasets. IS-diversity & IS-quality report $10^3 \exp(H(Y))$ and $10 \exp(-H(Y|X))$ and RKE-MC (Mode Count) denotes $\exp(\mathrm{RKE})$.

			Separated Inception Score				
	Method	IS ↑	IS-diversity ↑	IS-quality ↑	FID ↓	KID ↓	**RKE-MC** ↑
CIFAR-10	Dataset	11.57	39.87	29.01	-	-	39.58
	NVAE	5.85	25.05	23.36	51.67	0.0502	17.65
	VDVAE	10.51	38.40	27.37	37.51	0.0247	32.49
	DCGAN	5.75	16.41	34.89	54.30	0.0536	10.19
	WGAN-WC	2.59	35.25	7.37	157.26	0.0914	10.64
	WGAN-GP	7.51	25.82	28.85	21.66	0.0106	19.07
	SAGAN	8.62	28.33	30.13	10.17	0.0077	24.46
	SNGAN	8.81	30.08	29.31	9.23	0.0051	25.83
	ContraGAN	9.69	35.70	27.14	4.02	0.0023	29.80
Tiny-ImageNet	Dataset	33.99	60.56	56.12	-	-	155.86
	SAGAN	8.21	18.34	44.75	46.98	0.0531	25.68
	SNGAN	8.12	18.74	43.34	48.96	0.0567	27.18
	BigGAN	11.57	25.08	46.17	27.34	0.0262	39.61
	ContraGAN	13.79	28.96	47.62	21.36	0.0175	56.94
ImageNet	Dataset	357.35	370.78	96.37	-	-	1823.52
	SAGAN-256	29.67	35.66	83.22	44.66	0.0372	105.57
	SNGAN-256	31.92	35.70	89.41	35.75	0.0391	115.62
	ContraGAN-256	24.91	30.21	82.46	34.79	0.0403	152.89
	BigGAN-256	28.33	31.68	89.43	33.48	0.0440	106.07
	ReACGAN-256	52.53	62.24	75.05	15.65	0.0382	119.76
	BigGAN-2048	96.42	104.24	92.49	0.89	0.0038	606.18
	StyleGAN-XL	204.73	292.50	69.99	1.94	0.0035	1375.17
	LDM-4-G	242.62	252.43	94.65	3.60	0.0036	1321.24
	ADM-G	188.70	216.23	92.34	3.86	0.0036	1407.75

A.9.6 Datasets Results

We have divided the scores into absolute (Table 4) and relative scores (Table 5) where absolutes scores reports only based on the input samples but relative scores are based on the dataset and the generated samples.

A.9.7 The effect of standard deviation in RKE

In this experiment, we investigate the impact of standard deviation over σ. The real distribution is two Gaussian mixtures with $(-5, 0), (5, 0)$ as their centers with varying component std. The data and the first 10 eigenvalues of the kernel are shown in Figure 10. RKEs with different hyperparameter p's are shown in Table 6.

Table 5: Relative evaluation scores for CIFAR-10. A lower RRKE implies higher joint diversity.

	Method	Precision ↑	Recall ↑	Density ↑	Coverage ↑	**RRKE** ↓
CIFAR-10	NVAE	0.36	0.50	0.28	0.60	2.01
	VDVAE	0.34	0.78	0.23	0.21	1.81
	DCGAN	0.59	0.25	0.49	0.23	0.98
	WGAN-WC	0.36	0.00	0.18	0.03	2.09
	WGAN-GP	0.62	0.56	0.57	0.51	0.74
	SAGAN	0.68	0.62	0.73	0.73	0.65
	SNGAN	0.70	0.62	0.77	0.74	0.62
	ContraGAN	0.75	0.62	0.99	0.86	0.52
Tiny-ImageNet	SAGAN	0.55	0.49	0.44	0.27	1.42
	SNGAN	0.55	0.46	0.40	0.26	1.46
	BigGAN	0.60	0.58	0.53	0.43	1.23
	ContraGAN	0.62	0.54	0.54	0.45	1.26
ImageNet	SAGAN	0.57	0.58	0.42	0.35	2.34
	SNGAN	0.54	0.64	0.41	0.38	2.22
	ContraGAN	0.67	0.51	0.64	0.33	2.54
	BigGAN256	0.58	0.61	0.49	0.37	2.28
	ReACGAN-256	0.74	0.42	0.79	0.73	2.20
	BigGAN2048	0.71	0.58	0.80	0.65	1.83
	StyleGAN-XL	0.77	0.61	0.67	0.81	1.50
	LDM-4-G	0.86	0.60	0.69	0.78	1.56
	ADM-G	0.82	0.64	0.66	0.82	1.47

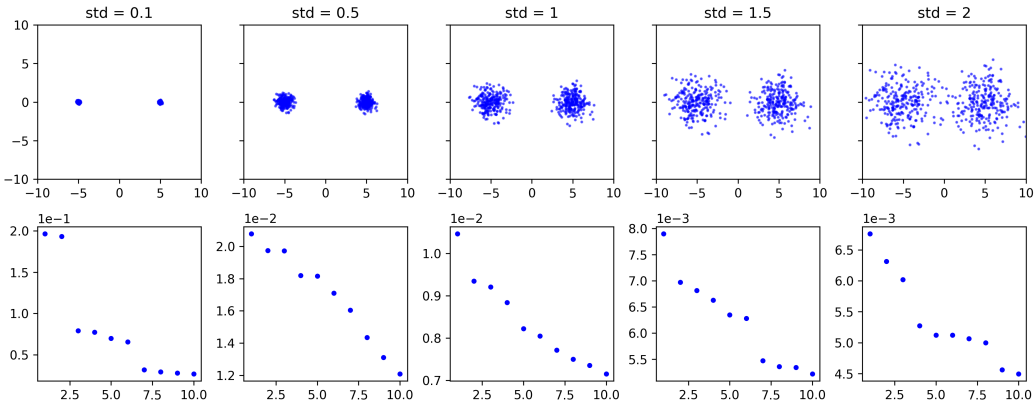


Figure 10: 500 samples from 2D Gaussian distribution with (-5, 0) and (5, 0) as their centers and their eigenvalue in the second row.

Table 6: RKE-MC result with different hyperparameter σ for Figure 10 samples.

σ	std = 0.1	std = 0.5	std = 1	std = 1.5	std = 2
0.1	9.69	139.95	295.92	380.01	423.77
0.5	2.31	9.69	31.26	63.07	100.57
1	2.07	3.94	9.69	18.95	31.19
2	**2.01**	2.48	3.94	6.35	9.64
5	1.96	**2.03**	**2.24**	**2.56**	**2.99**
10	1.46	1.47	1.50	1.55	1.62