Accelerated Kernel Stein Discrepancy with Rényi Landmark Selection for Stable and Efficient GAN Training

Michael Carlo, Giovanni Dettori, Ryan Gumsheimer

November 2025

1 Project Plan

Our project investigates the replacement of the classical adversarial discriminator in GAN training with a kernel-based distance metric, namely Kernel Stein Discrepancy (KSD). The goal is to assess whether a kernelized objective can provide an alternative that improves training stability and efficiency without compromising sample quality.

We will begin by reproducing the standard GAN setup using benchmark image datasets such as MNIST and CIFAR-10 to establish a reliable baseline and validate our implementation. Building on this, we will integrate KSD loss with Rényi landmark selection into the training pipeline, using it as a direct substitute for the adversarial loss. Finally, we will compare the two training paradigms—adversarial versus kernel-based—using both quantitative metrics (e.g., Fréchet Inception Distance, FID) and qualitative sample evaluation. All experiments will be implemented in Python, version-controlled via GitHub, and executed with GPU acceleration where appropriate.

2 Background and Motivation

Goodness-of-fit testing and model evaluation are central problems in modern machine learning and statistics. Given samples from an unknown distribution Q and a target model P with density $p(x)$, we often wish to determine how well Q approximates P. Classical divergence measures such as the Kullback–Leibler or Jensen–Shannon divergences are theoretically appealing but typically require access to the density $q(x)$ of the model distribution, which is unavailable in implicit generative models like GANs.

Kernel methods provide a powerful alternative by embedding probability measures into a *reproducing kernel Hilbert space* (RKHS), thereby allowing discrepancies between distributions to be expressed as distances between their embeddings. Formally, every positive-definite kernel $k(x, x')$ defines a unique Hilbert space of functions $\mathcal{H}_k$, where evaluation is continuous and satisfies

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k}, \qquad \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}_k} = k(x, x').$$

This property allows nonlinear relationships in the data space to appear as linear relations in the induced Hilbert space, so that expectations and variances can be expressed as inner products in $\mathcal{H}_k$.

In this framework, each probability distribution P on $\mathcal{X}$ can be represented by its *mean embedding* in the RKHS,

$$\mu_P := \mathbb{E}_{X\sim P}[\,k(\cdot, X)\,] \in \mathcal{H}_k,$$

which satisfies, for any $f \in \mathcal{H}_k$,

$$\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mu_P\rangle_{\mathcal{H}_k}.$$

Expectations under P thus become inner products between the test function f and the embedding μ_P. Similarly, comparing two distributions reduces to comparing their embeddings: the closer μ_P and μ_Q are in $\mathcal{H}_k$, the more similar the corresponding distributions are in the input space.

This observation provides a unifying language for defining statistical *discrepancies* between probability measures. A broad family of divergences can be expressed as *integral probability metrics* (IPMs) of the form

$$D_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \Big| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \Big|,$$

where $\mathcal{F}$ is a class of discriminating test functions. Choosing $\mathcal{F}$ as the unit ball of the RKHS $\mathcal{H}_k$ yields the *Maximum Mean Discrepancy* (MMD),

$$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k},$$

which admits a closed-form expression in terms of kernel evaluations and can be efficiently estimated from samples. When the kernel k is *characteristic* meaning that the mean embedding map $P \mapsto \mu_P$ is injective, the MMD uniquely identifies probability distributions, i.e. $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$.

However, MMD and related kernel-based divergences still require samples from both P and Q, or explicit knowledge of their normalized densities. This poses a serious limitation when the target distribution P is known only up to a normalizing constant and evaluating $p(x)$ or sampling from P may be computationally intractable. In such cases, standard IPM or embedding-based methods cannot be directly applied.

An elegant solution to this issue arises by exploiting *Stein's identity*, which allows one to construct kernelized discrepancies that depend only on the *score function* of the target distribution, $\nabla_x \log p(x)$. This observation leads directly to the development of the *Kernel Stein Discrepancy* (KSD), a measure of goodness-of-fit that preserves the nonparametric flexibility of kernel methods while not needing a fully normalized target density.

A particularly elegant instance of this framework is provided by *Stein's method.* Let P be a target distribution on $\mathbb{R}^d$ with smooth log-density $\log p(x)$ and score function $\nabla_x \log p(x)$. Stein's identity states that for any sufficiently regular test function $f : \mathbb{R}^d \to \mathbb{R}^d$,

$$\mathbb{E}_{X \sim P}\big[\, \nabla_x \log p(X)^\top f(X) + \nabla_x \cdot f(X) \,\big] = 0.$$

The operator inside the expectation,

$$(T_p f)(x) = \nabla_x \log p(x)^\top f(x) + \nabla_x \cdot f(x),$$

is known as the *Stein operator.* When comparing two distributions, say Q and P, we can evaluate the above expectations under Q: any deviation from zero quantifies a discrepancy between Q and P.

Restricting f to lie in a vector-valued RKHS $\mathcal{H}_k^d$ induced by a kernel k, one therefore obtains the *Kernel Stein Discrepancy* (KSD),

$$\mathrm{KSD}^2(Q, P) = \Big\| \, \mathbb{E}_{X \sim Q}[T_p k(\cdot, X)] \, \Big\|_{\mathcal{H}_k}^2.$$

The KSD measures the squared RKHS norm of the Stein witness function, which equals zero if and only if $Q = P$ when k is characteristic and p satisfies mild smoothness and boundary conditions [1]. Unlike divergences that require evaluating $q(x)$, KSD depends only on samples from Q and on the score function of P, making it well suited for implicit models and high-dimensional data. When P is empirical, say P_{data}, its score function $\nabla_x \log p_{\text{data}}(x)$ can be approximated using a pretrained score network, or a kernel density estimate.

In practice, evaluating the population KSD requires replacing expectations by empirical averages. Given samples $\{x_i\}_{i=1}^n \sim Q$, empirical estimators are obtained by replacing expectations with sample averages. The two standard estimators are the *V-statistic* and the *U-statistic*, respectively defined as

[1] For Stein's identity to hold, the target density p must be continuously differentiable and satisfy $p(x)f(x) \to 0$ as $\|x\| \to \infty$ for all admissible test functions f. These assumptions ensure that integration by parts is valid and that the Stein operator yields zero expectation under the true distribution P.

$$\widehat{\mathrm{KSD}}^2_V(Q,P) = \frac{1}{n^2}\sum_{i,j=1}^{n} h_p(x_i,x_j), \qquad \widehat{\mathrm{KSD}}^2_U(Q,P) = \frac{1}{n(n-1)}\sum_{i\neq j} h_p(x_i,x_j),$$

where the *Stein kernel* $h_p(x,x')$ is given by

$$h_p(x,x') = \nabla_x \log p(x)^\top \nabla_{x'} \log p(x')\, k(x,x') + \nabla_x \log p(x)^\top \nabla_{x'} k(x,x') + \nabla_{x'} \log p(x')^\top \nabla_x k(x,x') + \mathrm{tr}\big(\nabla^2_{x,x'} k(x,x')\big)\,.$$

The V-statistic is a biased estimator of the population KSD, while the U-statistic is unbiased. Both estimators have quadratic computational complexity $\mathcal{O}(n^2)$ in the number of samples, which makes them prohibitively expensive for large-scale applications.

To address this limitation, [6] introduced the *Nyström-KSD estimator*, an accelerated approximation that projects the empirical mean embedding $\frac{1}{n}\sum_{i=1}^{n} h_p(\cdot, x_i)$ onto the subspace spanned by a small set of $m \ll n$ Nyström points $\{\tilde{x}_j\}_{j=1}^{m}$. The resulting estimator takes the form

$$\widehat{\mathrm{KSD}}^2_N(Q,P) = \beta^\top K^-_{m,m}\,\beta, \quad \beta = \frac{1}{n} K_{m,n}\mathbf{1}_n,$$

where $K_{m,m}$ and $K_{m,n}$ are Gram matrices built from the Stein kernel h_p, and $K^-_{m,m}$ denotes the Moore–Penrose pseudoinverse. This approach reduces the computational cost to $\mathcal{O}(mn + m^3)$, while preserving the $\mathcal{O}(n^{-1/2})$ statistical rate of convergence, achieving the same asymptotic efficiency as the $U-$ and $V-$ estimators (so called *quadratic*).

Recently it has been shown that this rate cannot be improved. [6] established the $\sqrt{n}$ consistency of both the quadratic and Nyström KSD estimators under a sub-Gaussian assumption on the Stein feature map, and [3] proved that the *minimax lower bound* for KSD estimation is $\mathcal{O}(n^{-1/2})$. Consequently, all known KSD estimators are *minimax optimal*, achieving the fastest possible convergence rate for KSD estimation under standard smoothness assumptions on p and the kernel k.

3 Proposed Research

While the Nyström-KSD estimator achieves the optimal $\mathcal{O}(n^{-1/2})$ convergence rate under *uniform* landmark selection, uniform sampling may not be the most informative strategy in practice. In high-dimensional or heterogeneous datasets, uniformly chosen landmarks often oversample dense regions and underrepresent informative or low-probability areas. To address this, we propose replacing the uniform selection of Nyström landmarks with a *Rényi sampling scheme*, designed to better capture information in the data distribution.

The proposed sampling criterion is inspired by the quadratic Rényi entropy defined in [4] for a probability density $p(x)$ as

$$H_R = -\log \int p(x)^2\,dx,$$

which measures the spread, or uncertainty, of the distribution. Given an empirical kernel density estimate $\hat{p}(x)$ constructed from data samples $\{x_i\}_{i=1}^{N}$, this entropy can be approximated as

$$\int \hat{p}(x)^2\,dx = \frac{1}{N^2}\mathbf{1}_v^\top \Omega \mathbf{1}_v,$$

where $\mathbf{1}_v = [1,\ldots,1]^\top$ and Ω is the kernel (Gram) matrix with entries $\Omega_{ij} = k(x_i,x_j)$. Minimizing this quantity corresponds to maximizing the entropy of the selected subset, hence encouraging a set of landmarks that are diverse and informative in the kernel-induced feature space.

In this setting, we aim to optimize the selection of Nyström landmarks $\tilde{X} = \{\tilde{x}_1,\ldots,\tilde{x}_m\}$ by maximizing the entropy of their induced kernel density, or equivalently, by minimizing

$$\mathcal{E}(\tilde{X}) = \frac{1}{m^2}\mathbf{1}_m^\top \Omega(\tilde{X})\mathbf{1}_m,$$

subject to the constraint that Ω is normalized with respect to the kernel bandwidth. This criterion ensures that the induced subspace $\text{span}\{h_p(\cdot, \tilde{x}_i)\}_{i=1}^m$ better captures the geometry of the Stein feature map and the expressive regions of the data distribution. By promoting landmarks that are both representative and diverse, this strategy is expected to reduce the finite-sample bias of the Nyström-KSD estimator and stabilize its gradients when used as a loss in generative models.

Generative Adversarial Networks (GANs) provide a widely used framework for learning generative models that can synthesize data samples resembling those drawn from an unknown distribution P_{data}. A GAN consists of two components: a generator $G_\theta(z)$, which maps latent variables $z \sim p_z$ (typically standard Gaussian noise) to the data space, and a discriminator $D_\phi(x)$, which aims to distinguish between real samples $x \sim P_{\text{data}}$ and generated samples $x = G_\theta(z)$. The two networks are trained in a minimax game:

$$\min_\theta \max_\phi \mathcal{L}_{\text{GAN}}(\theta, \phi) = \mathbb{E}_{x \sim P_{\text{data}}}\big[\log D_\phi(x)\big] + \mathbb{E}_{z \sim p_z}\big[\log(1 - D_\phi(G_\theta(z)))\big].$$

At equilibrium, the generator distribution Q_θ ideally converges to the data distribution P_{data}, achieving a Nash equilibrium in which $D_\phi(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + q_\theta(x)}$.

However, the adversarial objective is known to be unstable and provides only an implicit measure of discrepancy between P_{data} and Q_θ. Moreover, the Jensen–Shannon divergence underlying $\mathcal{L}_{\text{GAN}}$ requires that both distributions admit well-defined densities, which is often restrictive in high-dimensional or implicit generative settings.

The *Kernel Stein Discrepancy* offers a non-adversarial alternative that directly quantifies the discrepancy between the model distribution Q_θ and the target P_{data} through their score functions, without needing to train an explicit discriminator. By using the accelerated Nyström-KSD estimator, this discrepancy can be efficiently approximated from mini-batches of generated samples, leading to a tractable training loss

$$\mathcal{L}_{\text{KSD}}(\theta) = \widehat{\text{KSD}}_N^2(Q_\theta, P_{\text{data}}),$$

whose gradients with respect to θ can be computed by automatic differentiation through the Nyström projection. This yields a generator trained to minimize the kernel-based discrepancy to the data distribution, sidestepping adversarial optimization while retaining theoretical convergence guarantees. Furthermore, the proposed *Rényi Nyström-KSD* estimator enhances finite-sample efficiency by adaptively selecting informative landmarks, thus combining the statistical rigor of Stein's method with the practical scalability of modern deep generative models.

4 Data and Code Implementation

As part of our research we propose the replication of the GAN architecture as done in [5] on the MNIST dataset used by [8] and a respective comparison with our proposed KSD approach. This classic benchmark consists of 70,000 grayscale images of handwritten digits 0–9 (split into 60,000 training and 10,000 test examples). Each image is 28×28 pixels, centered and normalized, depicting a single handwritten numeral.

This approach will allow us to compare performance of training a generator using the KSD as discriminator with the original adversarial generator-discriminator approach. Finally we will move to the CIFAR-10 dataset used in [7] to have a more realistic benchmark for real world applications. The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. Both datasets are available on Kaggle ([2] and [1]) and can be imported into code.

Code sharing and implementation will be done using Github, to facilitate version control, branching and as part of the deliverables of the project. In case computational power were to become a bottleneck we will make use of Google Colab to rent GPU time.

References

[1] Cifar-10 dataset. Dataset on Kaggle, 2025. Available at `https://www.kaggle.com/datasets/ayush1220/cifar10`.

[2] Mnist dataset. Dataset on Kaggle, 2025. Available at `https://www.kaggle.com/datasets/hojjatk/mnist-dataset`.

[3] José Cribeiro-Ramallo, Agnideep Aich, Florian Kalinke, Ashit Baran Aich, and Zoltán Szabó. The minimax lower bound of kernel stein discrepancy estimation. *arXiv preprint*, 2025.

[4] Mark Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14(3):669–688, 2002.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Florian Kalinke, Zoltán Szabó, and Bharath K. Sriperumbudur. Nyström kernel stein discrepancy. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*. PMLR, 2025.

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

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.