Explicit Density Approximation for Neural Implicit Samplers Using a Bernstein-Based Convex Divergence

Anonymous Author
Anonymous Institution

Abstract

Rank-based objectives such as the invariant statistical loss (ISL) are robust, likelihood-free tools for training implicit generative models. We propose *dual-ISL*, obtained by interchanging the roles of the target p and model density $\tilde{p}$ within ISL, which induces a *convex* optimization problem over model densities. We show that the associated rank-based discrepancy d_K is *continuous* under weak and L^1 convergence and *convex* in its first argument, properties not shared by classical divergences such as KL or Wasserstein distances. Additionally, we prove that d_K admits an L^2 interpretation: it is the projection of the density ratio $q = p/\tilde{p}$ onto a Bernstein polynomial basis. This yields explicit truncation-error bounds, sharp convergence rates, and a closed-form expression for the truncated density approximation. To handle multivariate data, we further introduce a sliced dual-ISL via random one-dimensional projections that preserves both continuity and convexity. Empirically, across several benchmarks, dual-ISL delivers faster and smoother convergence than standard ISL and offers competitive, often superior, mode coverage relative to state-of-the-art implicit models (modern GAN baselines, including multi-critic setups), while providing an explicit density approximation.

1 Introduction

Implicit generative models are a class of models that learn to generate data samples without explicitly modeling the underlying probability distribution Mohamed and Lakshminarayanan (2016), enabling flexible modeling of high-dimensional data across vision (e.g., implicit maximum likelihood estimation Li and Malik (2018), StyleGAN Sauer et al. (2022)), audio (e.g., NSF Wang et al. (2020), FA-GAN Shen et al. (2024)), and text domains (e.g., Residual EBM Deng et al. (2020)). Instead of directly estimating the data distribution, these models learn a mapping from a simple input distribution (such as a multivariate Gaussian) to the data space through a deterministic or stochastic function. A prominent example is the generator in a generative adversarial network (GAN) Goodfellow et al. (2014), which transforms random noise vectors into realistic data samples. The generator is trained in tandem with a discriminator that learns to distinguish real data from generated data, providing feedback that guides the generator to improve.

The invariant statistical loss (ISL) is a rank-based loss function recently proposed by de Frutos et al. (2024b) that compares the empirical order statistics of samples from the data and from the implicit generative model. In this work, we introduce dual-ISL, a novel likelihood-free objective obtained by swapping the roles of the data and model distributions within the ISL framework. Remarkably, the induced discrepancy, d_K, admits a fully *explicit closed-form density approximation*: it is exactly the L^2–projection of the density ratio

$$q(x) \;=\; \frac{p_{\text{target}}(x)}{p_{\text{model}}(x)}, \qquad x = F^{-1}_{\text{target}}(t),\; t \in [0,1],$$

onto the space of dual-Bernstein polynomials of degree K Lorentz (2012); Jiittler (1998). Writing

$$q_K(x) = \sum_{n=0}^{K} \mathbb{Q}_K(n)\, \tilde{b}_{n,K}\big(F_{\text{target}}(x)\big),$$

with computable coefficients $\{\mathbb{Q}_K(n)\}$ immediately yields $p_{\text{target}} \approx p_{\text{model}}(x) \cdot q_K(x)$. This explicit representation not only provides analytic error bounds via Bernstein approximation theory and ensures convexity over the space of densities, but also enables efficient density evaluation without auxiliary sampling

Preliminary work. Under review by AISTATS 2026. Do not distribute.

and provable convergence rates inherited from polynomial approximation. By marrying likelihood-free training with a tractable, closed-form density, dual-ISL bridges a critical gap, offering both rigorous theory and practical stability in implicit generative modeling.

Contributions and paper organisation. In this work, we endow implicit generative models with an explicit, tractable, and optimizable density approximation. By reinterpreting rank-based losses through the lens of Bernstein polynomial projections, dual-ISL provides a convex, likelihood-free training objective that approximates the true data distribution with closed-form densities. This projection-based framework mitigates fundamental issues such as mode collapse and non-convexity, enables rigorous analysis via polynomial approximation theory, and extends naturally to multivariate settings via slicing. The result is a principled and scalable method that bridges implicit modeling and explicit density estimation. In Section 2, we briefly review invariant statistical loss as defined in the literature. In Section 3, we introduce a new theoretical loss, dual to ISL, and prove convexity over the space of model densities. In Section 4, we analyse this dual loss as a projection onto the Bernstein polynomial basis, providing exact convergence rates, a new interpretation as polynomial interpolation, and an explicit density estimator. In Section 5, we extend dual-ISL to multivariate settings using slicing, retaining the Bernstein interpretation and establishing convexity and convergence results. Finally, Section 6 is devoted to the conclusions and a discussion of limitations.

Related work The implicit approach to generative modeling trains a sampler that maps noise to data, avoiding explicit likelihoods (Mohamed and Lakshminarayanan, 2016). The most widely explored family is GANs (Goodfellow et al., 2014), with numerous stabilizers, Wasserstein objectives (Arjovsky et al., 2017), gradient penalties (Gulrajani et al., 2017), and spectral normalization (Miyato et al., 2018), yet optimization remains nonconvex and prone to mode collapse. Beyond adversarial critics, integral probability metrics (IPMs) replace the discriminator with a distance between distributions. Kernel MMD offers a non-adversarial objective, though performance hinges on kernel choice and feature learning (Li et al., 2017); related IPMs include the energy distance and kernelized optimal transport variants. Other non-adversarial implicit objectives include, implicit maximum likelihood estimation, which matches each datum to its nearest generated sample (Li and Malik, 2018), and the sliced wasserstein generator, which directly minimizes the sliced Wasserstein distance between real and model samples (Deshpande et al., 2018).

Rank- and order-statistic methods offer another likelihood-free route: classical goodness-of-fit tests, Kolmogorov–Smirnov and Cramér–von Mises, compare distributions via the probability-integral transform (Massey Jr, 1951). The invariant statistical loss (ISL) adapts this idea to implicit training by enforcing uniformity of a discrete rank statistic built from model/data samples (de Frutos et al., 2024b). Closely related *slicing/projection* approaches are likewise *order-statistic based*: they reduce multivariate comparisons to 1D problems by projecting and then *sorting* samples to form empirical cdfs/ranks, yielding scalable, analysable objectives (Kolouri et al., 2019). In particular, max-sliced and subspace-robust variants focus comparisons on informative directions (Deshpande et al., 2019).

2 The invariant statistical loss (ISL)

We briefly review the invariant statistical loss (ISL) from de Frutos et al. (2024b). ISL is built on a simple rank statistic whose distribution is exactly uniform when two samples come from the same probability density function, and which varies continuously under L^1-perturbations of the underlying densities.

2.1 Rank statistic and uniformity

Let $\tilde{y}_1, \dots, \tilde{y}_K$ be i.i.d. samples from a univariate real distribution with pdf $\tilde{p}$, and let y be a single sample independently drawn from another distribution with pdf p. Define the subset

$$\mathcal{A}_K \;:=\; \Big\{\tilde{y} \in \{\tilde{y}_k\}_{k=1}^K : \tilde{y} \le y\Big\},$$

and the *rank statistic*

$$A_K \;:=\; \big|\mathcal{A}_K\big|, \tag{1}$$

i.e., A_K counts how many samples in $\{\tilde{y}_1, \dots, \tilde{y}_K\}$ lie at or below y. Then A_K is a discrete random variable (r.v.) taking values in $\{0, 1, \dots, K\}$, and we denote its pmf by

$$\mathbb{Q}_K : \{0, \dots, K\} \to [0, 1].$$

When the two pdfs p and $\tilde{p}$ coincide, this pmf is exactly uniform (de Frutos et al., 2024b).

Theorem 2.1. *If $p = \tilde{p}$, then A_K is uniformly distributed on $\{0, \dots, K\}$, i.e. $\mathbb{Q}_K(n) = \frac{1}{K+1}$ for all $n \in \{0, \dots, K\}$.*

2.2 ISL discrepancy

The ISL discrepancy quantifies the deviation of the pmf $\mathbb{Q}_K$ from the uniform law on $\{0, \dots, K\}$. To be

specific, we define the discrepancy function

$$d_K(p,\tilde{p}) := \frac{1}{K+1}\|\mathbb{Q}_K - \mathbb{U}_K\|_{\ell_1} = \frac{1}{K+1}\sum_{n=0}^{K}\left|\tfrac{1}{K+1} - \mathbb{Q}_K(n)\right| = \frac{2}{K+1}\mathrm{TV}(\mathbb{Q}_K,\mathbb{U}). \quad (2)$$

where $\mathbb{U}_K$ is the uniform pmf on $\{0,\dots,K\}$ and $\mathrm{TV}(\cdot,\cdot)$ denotes total variation distance. By Theorem 2.1, $d_K(p,p) = 0$ for all K. Moreover, Theorem 2.2 below, ensures that $d_K(p,\tilde{p})$ depends continuously on $\tilde{p}$ in the L^1 sense, while Theorem 2.3 guarantees that if $d_K(p,\tilde{p}) = 0$ for all K, then $\tilde{p} = p$ almost everywhere. Hence, in the large-K limit, d_K behaves as a proper divergence, vanishing precisely when the two densities coincide.

Theorem 2.2 (Continuity). *If $\|p-\tilde{p}\|_{L^1(\mathbb{R})} \le \epsilon$, then for all $n \in \{0,\dots,K\}$,*

$$\frac{1}{K+1} - \epsilon \;\le\; \mathbb{Q}_K(n) \;\le\; \frac{1}{K+1} + \epsilon.$$

Theorem 2.3 (Identifiability). *Let $p,\tilde{p}$ be pdfs of univariate real r.v.s. If the rank statistic A_K in (1) is uniformly distributed on $\{0,\dots,K\}$ for* every *$K \in \mathbb{N}$, then $p = \tilde{p}$ almost everywhere.*

Finally, when $\tilde{p} = \tilde{p}_\theta$ depends smoothly on a parameter vector θ, one can show (under mild regularity assumptions) that $\theta \mapsto d_K(p,\tilde{p}_\theta)$ is continuous and differentiable, making it suitable for gradient-based optimization (see Theorem 4 in de Frutos et al. (2024a)). For full proofs and additional remarks, see de Frutos et al. (2024b,a).

2.3 A surrogate for ISL optimization

Directly minimizing the divergence $d_K(p,\tilde{p}_\theta)$ with respect to the generator parameters θ is normally not feasible: the pmf $\mathbb{Q}_K$ has to be approximated empirically and its dependence on θ is unknown. To overcome this difficulty, de Frutos et al. (2024b) introduced a carefully designed surrogate loss that (i) closely tracks d_K, and (ii) admits gradient optimization via standard backpropagation. This surrogate is constructing by approximating the pmf of $\mathbb{Q}_K$ using sigmoidal functions and a Gaussian kernel density estimator. For full details of the surrogate derivation, implementation, and bias-variance trade-offs, see (de Frutos et al., 2024a, Section 2.3).

3 The dual-invariant statistical loss

By interchanging the roles of the data distribution p and the model distribution $\tilde{p}$ in the ISL framework, we obtain a *dual* objective that remains likelihood-free, but crucially becomes convex in the model pdf $\tilde{p}$.

3.1 Continuity and convexity of $d_K(p,\tilde{p})$

Unlike most classical discrepancies, this rank-based measure is *weakly continuous*: if $p_n \xrightarrow{w} p$ weakly, then $\lim_{n\to\infty} d_K(p_n,\tilde{p}) = d_K(p,\tilde{p})$ (Theorem 3.1 below). In contrast, the Kullback–Leibler divergence does not enjoy weak continuity, and the Wasserstein and Energy distances require uniformly bounded moments to guarantee even this level of stability (Huster et al., 2021, Section 5). Finally, we show that d_K is convex in its first argument (Theorem 3.2), yielding a tractable convex optimization problem in the space of densities.

A key insight in the continuity proof is that each probability mass $\mathbb{Q}_K(n)$, for $n = 0,\dots,K$, can be written as a continuous mixture of binomial pmf's. Indeed, drawing K i.i.d. samples $\tilde{y}_i \sim \tilde{p}$ and counting how many of them fall below y yields a Binomial$(K,\tilde{F}(y))$ distribution. Since y itself is drawn from p, one obtains the vector $\mathbb{Q}_K = (\mathbb{Q}_K(n))_{n=0}^K$, with,

$$\mathbb{Q}_K(n) := \int_{\mathbb{R}} h_n(y)\,p(y)\,\mathrm{d}y,$$

where,

$$h_n(y) := \binom{K}{n}\tilde{F}(y)^n\big(1-\tilde{F}(y)\big)^{K-n},$$

and the bounded, continuous functions h_n then ensures weak continuity of $\mathbb{Q}_K$ and hence of d_K. We formalize this argument in the following theorem.

Theorem 3.1 (Continuity under weak convergence). *Let $(p_n)_{n\ge1}$ be a sequence of pdfs on $\mathbb{R}$ converging weakly to a density p, and let $\tilde{p}$ be a fixed reference density with cdf $\tilde{F}$. For each $K \in \mathbb{N}$, define*

$$\mathbb{Q}_K^{(n)}(m) := \int_{\mathbb{R}}\binom{K}{m}\tilde{F}(y)^m\big(1-\tilde{F}(y)\big)^{K-m}\,p_n(y)\,\mathrm{d}y,$$

for $m = 0,\dots,K$. Then

(i) *(Pointwise convergence)* $\lim_{n\to\infty}\mathbb{Q}_K^{(n)}(m) = \mathbb{Q}_K(m)$ *for each* $m = 0,\dots,K$.

(ii) *(Continuity of d_K)* $\lim_{n\to\infty} d_K(p_n,\tilde{p}) = d_K(p,\tilde{p})$.

Proof. See Appendix A. □

Since strong convergence implies weak convergence, the previous theorem remains applicable when the sequence $\{p_n\}_{n\ge1}$ converges to p in the L^1 norm. We can also establish that the ISL divergence is continuous with respect to its second argument $\tilde{p}$; a detailed proof can be found in Appendix A, Theorem A.1.

We now see that the discrepancy $d_K(p, \tilde{p})$ is indeed convex in its first argument.

Theorem 3.2 (Convexity)**.** *For any probability distributions p_1, p_2 and $\tilde{p}$ on $\mathbb{R}$, and for any $\lambda \in [0,1]$, the discrepancy d_K satisfies*

$$d_K(\lambda p_1 + (1-\lambda)p_2, \tilde{p}) \leq \lambda\, d_K(p_1, \tilde{p}) + (1-\lambda)\, d_K(p_2, \tilde{p}).$$

Proof. See Appendix A. □

3.2 A dual loss function

Because d_K is convex in its first argument, we can obtain a new training criterion by swapping the data and model distributions. Specifically, let $\tilde{y} \sim \tilde{p}$ be a simulated sample from our generator and let $y_{1:K} \overset{\text{i.i.d.}}{\sim} p$ be K independent real data points. We then form the rank statistic as

$$\tilde{A}_K := \left|\left\{y \in \{y_k\}_{k=1}^K : y \leq \tilde{y}\right\}\right|$$

whose pmf $\tilde{\mathbb{Q}}_K(n) = \mathbb{P}\left(\tilde{A}_K = n\right)$ remains uniform if and only if $\tilde{p} = p$. All of our previous guarantees, continuity under small L^1 perturbations (Theorem 2.2) and identifiability when $\mathbb{Q}_K$ is exactly uniform for every K (Theorem 2.3), carry over unchanged. The dual-ISL discrepancy

$$d_K(\tilde{p}, p) = \frac{1}{K+1} \sum_{n=0}^{K} \left| \tilde{\mathbb{Q}}_K(n) - \frac{1}{K+1} \right|,$$

therefore yields a convex, likelihood-free training objective in the space of generator densities. The pseudocode for this method is provided in the supplementary material (see Algorithm 1).

3.3 Dual-ISL vs. ISL, GANs & Diffusion on 1D distributions

We start considering the same experimental setup as Zaheer et al. (2017); de Frutos et al. (2024b). We evaluate dual-ISL on six benchmark targets using $N = 1000$ i.i.d. samples drawn from each distribution. The first three are standard univariate pdfs, and the latter three are mixtures with equal mixing weights. Model_1 combines Gaussians $\mathcal{N}(5,2)$ and $\mathcal{N}(-1,1)$; Model_2 combines Gaussians $\mathcal{N}(5,2)$, $\mathcal{N}(-1,1)$, and $\mathcal{N}(-10,3)$; and Model_3 combines a Gaussian $\mathcal{N}(-5,2)$ with a Pareto$(5,1)$ distribution.

We train a 4-layer MLP generator (1-7–13-7–1 units, ELU activations) with $\epsilon \sim \mathcal{N}(0,1)$ input noise for 10^4 epochs with Adam (learning rate 10^{-2}), and compare Dual-ISL, ISL, GAN Goodfellow et al. (2014), WGAN Arjovsky et al. (2017), MMD-GAN Li et al. (2017), and a diffusion baseline using the Kolmogorov–Smirnov distance (KSD) metric (Table 1). Experimental details are provided in Supplementary Material Section D.1.

The convexity of the dual-ISL objective not only accelerates convergence, yielding faster, smoother, and more stable training curves compared to standard ISL (see Figure 6 in Appendix), but also enhances mode coverage on challenging mixtures. As shown in Figure 1, both dual-ISL and standard ISL successfully avoid the mode collapse exhibited by MMD-GAN, with dual-ISL most accurately capturing the heavy tail of the Pareto component.

The Appendix presents runtime benchmarks demonstrating dual-ISL's computational advantages over standard ISL. In Appendix D.3, we also propose a new ISL-based method with a monotonicity penalty that guarantees recovery of the optimal-transport map even for distributions without finite moments (e.g., heavy-tailed), an advantage over the p-Wasserstein distance which requires finite pth moments.

4 An L^2-projection view of d_K

We adopt a projection-based view of ISL. From this point on, we treat p and $\tilde{p}$ interchangeably, so that, with a slight abuse of notation, our framework covers both standard ISL ($q = p/\tilde{p}$) and dual-ISL ($q = \tilde{p}/p$). Specifically, we show that the discrete pmf $\mathbb{Q}_K(n)$ coincides with the L^2-projection coefficients of the density ratio $q = p(x)/\tilde{p}(x)$ onto the degree-K Bernstein basis $\{b_{n,K}\}_{n=0}^K$. In this light, ISL becomes a purely likelihood-free density-ratio divergence, comparing projection coefficients rather than intractable likelihoods, and we conclude by deriving sharp convergence rate bounds.

4.1 Projection interpretation

To reveal the underlying geometry of $d_K(\cdot,\cdot)$, we define a linear operator that collects the $K+1$ probabilities $\mathbb{Q}_K(0), \ldots, \mathbb{Q}_K(K)$, into a single vector. We then show that each entry $\mathbb{Q}_K(n)$ is precisely the L^2 inner product between a density ratio and its corresponding Bernstein basis function.

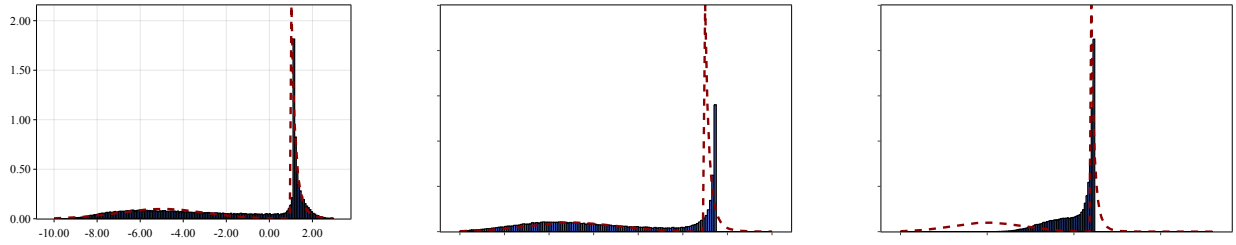
Definition 4.1 (Binomial mapping)**.** *Let $p, \tilde{p} \in C(\mathbb{R})$ be two continuous pdfs with cdfs F and $\tilde{F}$. For any integer $K \geq 1$, define the operator,*

$$\Phi_K(p, \tilde{p}) := \big(\mathbb{Q}_K(0), \mathbb{Q}_K(1), \ldots, \mathbb{Q}_K(K)\big) \in \mathbb{R}^{K+1}.$$

It is straightforward from the integral representation that, for each fixed $\tilde{p}$, the map $p \mapsto \Phi_K(p, \tilde{p})$ is linear

Target	Dual-ISL	ISL	GAN	WGAN	MMD-GAN	Diffusion
$\mathcal{N}(4,2)$	$\mathbf{0.018 \pm 0.005}$	0.020 ± 0.003	0.018 ± 0.003	0.024 ± 0.017	0.042 ± 0.026	0.020 ± 0.002
$\mathcal{U}(-2,2)$	0.034 ± 0.015	0.021 ± 0.004	0.049 ± 0.032	0.064 ± 0.062	0.104 ± 0.060	$\mathbf{0.013 \pm 0.002}$
Cauchy$(1,2)$	0.016 ± 0.003	$\mathbf{0.013 \pm 0.002}$	0.013 ± 0.002	0.052 ± 0.055	0.031 ± 0.008	0.114 ± 0.034
Pareto$(1,1)$	$\mathbf{0.090 \pm 0.080}$	0.198 ± 0.148	0.117 ± 0.041	0.106 ± 0.043	0.158 ± 0.168	0.209 ± 0.011
Mixture$_1$	0.016 ± 0.004	$\mathbf{0.016 \pm 0.002}$	0.017 ± 0.004	0.080 ± 0.069	0.054 ± 0.033	0.031 ± 0.031
Mixture$_2$	$\mathbf{0.016 \pm 0.002}$	0.017 ± 0.003	0.026 ± 0.014	0.031 ± 0.023	0.042 ± 0.061	0.050 ± 0.005
Mixture$_3$	$\mathbf{0.170 \pm 0.019}$	0.171 ± 0.012	0.190 ± 0.094	0.216 ± 0.040	0.187 ± 0.108	0.173 ± 0.024

Table 1: KSD over 10 runs for Dual-ISL and baselines. Setup: $K = 10$, 1000 epochs, $N = 1000$.



(a) dual-ISL (b) ISL (c) MMD-GAN

Figure 1: Comparison of dual-ISL, standard ISL, and MMD-GAN for modeling a mixture of Pareto and Normal distributions. Subfigure 1a displays the dual-ISL results, Subfigure 1b illustrates the performance of the standard ISL approach, and Subfigure 1c showcases the outcomes obtained via MMD-GAN.

and continuous under mild regularity conditions on p and $\tilde{p}$. A full statement of these and related properties appears in Theorem B.1.

The next result shows that Φ_K admits a Riesz representation see (Brézis, 2011, Theorem 4.11), expressing each probability mass $\mathbb{Q}_K$ as an L^2 inner product with a Bernstein basis function. Let the nth Bernstein polynomial of degree K be defined as

$$b_{n,K}(t) \;:=\; \binom{K}{n} t^n (1-t)^{K-n}, \qquad t \in [0,1].$$

Theorem 4.1. *(Riesz representation of Φ_K) Let $\tilde{p}$ be a fixed continuous density on $\mathbb{R}$ with cdf $\tilde{F}$. Then for any $K \geq 0$, the operator $\Phi_K(\cdot, \tilde{p})$ mapping $p \mapsto (\mathbb{Q}_K(0), \ldots, \mathbb{Q}_K(K))$ satisfies*

$$\mathbb{Q}_K(n) = \int_{\mathbb{R}} b_{n,K}\big(\tilde{F}(x)\big)\, p(x)\, \mathrm{d}x = \big\langle\, b_{n,K} \circ \tilde{F},\; p \big\rangle_{L^2(\mathbb{R})}.$$

for $n = 0, \ldots, K$. Moreover, if $\tilde{p}(x) > 0$ for all $x \in \mathbb{R}$, then defining the density ratio $q(x) = p(x)/\tilde{p}(x)$, we get

$$\begin{aligned}\mathbb{Q}_K(n) &= \int_0^1 b_{n,K}(t)\, q\big(\tilde{F}^{-1}(t)\big)\, \mathrm{d}t \\ &= \big\langle b_{n,K},\, q \circ \tilde{F}^{-1} \big\rangle_{L^2([0,1])}, \quad \text{for } n = 0, \ldots, K.\end{aligned}$$

Proof. See Appendix B. □

Theorem 4.1 implies that, if we define $\tilde{q}(t) = q\big(\tilde{F}^{-1}(t)\big)$ for $t \in [0,1]$ then each coefficient $\mathbb{Q}_K(n)$ is exactly the $L^2([0,1])$-projection of $\tilde{q}$ onto the Bernstein polynomial $b_{n,K}$. Equivalently, the vector $\{\mathbb{Q}_K(n)\}_{n=0}^K$ collects the best mean-square approximation coefficients of the *push-forward density ratio* $q \circ \tilde{F}^{-1}$ in the degree-K Bernstein basis.

Theorem 4.2 (Bernstein-basis truncation for the density-ratio)**.** *Let $p, \tilde{p} \in C(\mathbb{R})$ with $\tilde{p}(x) > 0$ for all $x \in \mathbb{R}$ and $K \in \mathbb{N}$. Then $\tilde{q} \in C([0,1])$ admits the Bernstein-polynomial expansion*

$$\tilde{q}(t) \;=\; \lim_{K \to \infty} \sum_{n=0}^{K} \alpha_{n,K}\, b_{n,K}(t),$$

where $\alpha = (\alpha_{n,K})_{n \geq 0}$ are the unique Bernstein-basis coordinates of $\tilde{q}$. Its degree-K truncation can be expressed as

$$\tilde{q}_K(t) \;:=\; \sum_{n=0}^{K} \alpha_{n,K}\, b_{n,K}(t) = \sum_{n=0}^{K} \mathbb{Q}_K(n)\, \tilde{b}_{n,K}(t),$$

where $\{\tilde{b}_{n,K}\}_{n=0}^K$ is the dual Bernstein basis (Jiittler, 1998).

Proof. See Proof in Appendix B. □

Remark 4.1. *With $\tilde{p}$ fixed, the map $\Phi_K : p \longmapsto \big(\mathbb{Q}_K(0), \ldots, \mathbb{Q}_K(K)\big)$, cannot distinguish between any*

two target densities p_1, p_2 whose pushed-through ratios $q_i \circ \tilde{F}^{-1}$ have the same degree-K Bernstein projections. Equivalently,

$$\Phi_K(p_1) = \Phi_K(p_2) \iff \langle b_{n,K},\, (q_1 - q_2) \circ \tilde{F}^{-1} \rangle_{L^2([0,1])} = 0 \quad \forall\, n = 0, \dots, K.$$

Thus Φ_K factors through the quotient of $L^2([0,1])$ by the subspace orthogonal to $\mathrm{span}\{b_{n,K}\}$, inducing a bijection onto its image.

4.2 Approximation error

We now quantify the truncation error in the approximation ratio $\tilde{q}_K(t)$ is an estimate of $q(x)$ and it remains uniformly close to 1, with its sup-norm deviation bounded by the discrepancy $d_K(p, \tilde{p})$.

Theorem 4.3. *Let $p, \tilde{p} \in C(\mathbb{R})$ be pdfs. Then $q_K(x)$ satisfies*

$$\|q_K - 1\|_\infty \le (K+1)^2\, d_K(p, \tilde{p}).$$

Proof. See Proof in Appendix B □

By standard Bernstein-approximation theory, one can bound the truncation error via the modulus of continuity of q. In particular, if $q \in C^2(\mathbb{R})$, then

$$\|q - q_K\|_\infty = O(K^{-1}),$$

and more generally, if q is α–Hölder continuous on $\mathbb{R}$, then

$$\|q - q_K\|_\infty = O\big(K^{-\alpha/2}\big).$$

See Gzyl and Palacios (1997) for the C^2 case and Mathé (1999) for the Hölder regime.

Remark 4.2. *If we assume that $q \in C^2(\mathbb{R})$, by the triangle inequality, we have*

$$\begin{aligned}\|q(x) - 1\|_\infty &\le \|1 - q_K(x)\|_\infty + \|q_K(x) - q(x)\|_\infty \\ &\le (K+1)^2 d_K(p, \tilde{p}) + \frac{\|q(x)''\|_\infty}{8K}. \end{aligned} \tag{3}$$

To empirically validate Equation 3, we train the same NN architecture under identical hyperparameters as in our earlier experiments. The model input is $z \sim \mathcal{N}(0,1)$ and approximates a mixture of Cauchy distributions. We recover the estimated density $\tilde{p}$ via kernel density estimation and compute the second derivative of the quotient q with sixth-order central finite differences. We average over 10 runs, means are shown in Figure 2 (a-b).

Theorem 4.4 (Explicit density approximation)**.** *Let $p, \tilde{p} \in C(\mathbb{R})$ with $\tilde{p}(x) > 0$ for all $x \in \mathbb{R}$, and let $\tilde{F}(x)$ be the cdf of $\tilde{p}$. Define*

$$p_K(x) := \tilde{p}(x) \sum_{m=0}^{K} \mathbb{Q}_K(m)\, \tilde{b}_{m,K}\big(\tilde{F}(x)\big). \tag{4}$$

Then for every $x \in \mathbb{R}$, $\lim_{K\to\infty} p_K(x) = p(x)$.

Proof. See Proof in Appendix B. □

Remark 4.3. *In practice, one draws latent samples $z_1, \dots, z_N \overset{\text{i.i.d.}}{\sim} p_z$ and computes $x_i = f(z_i)$, where f is the neural network pushing p_z forward to $\tilde{p}$. One then forms the empirical cdf and density estimates*

$$\widehat{\tilde{F}}(x) = \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{x_i \le x\},$$

$$\widehat{\tilde{p}}(x) = \frac{\widehat{\tilde{F}}(x+\delta) - \widehat{\tilde{F}}(x-\delta)}{2\delta}.$$

Substituting these into Equation (4) yields the Monte Carlo approximation

$$\widehat{p}_K(x) = \widehat{\tilde{p}}(x) \sum_{m=0}^{K} \mathbb{Q}_K(m)\, \tilde{b}_{m,K}\big(\widehat{\tilde{F}}(x)\big). \tag{5}$$

Figure 2(c–d) shows that ISL's Bernstein approximation recovers the true density in 1D and 2D. In Figure 2c, we compare a Gaussian-mixture ground truth (red) with dual-ISL at $K = 2$ (light blue) and $K = 15$ (dark blue); Figure 2d overlays learned contours on two-moons. See Appendix D.5 for additional experiments and implementation details.

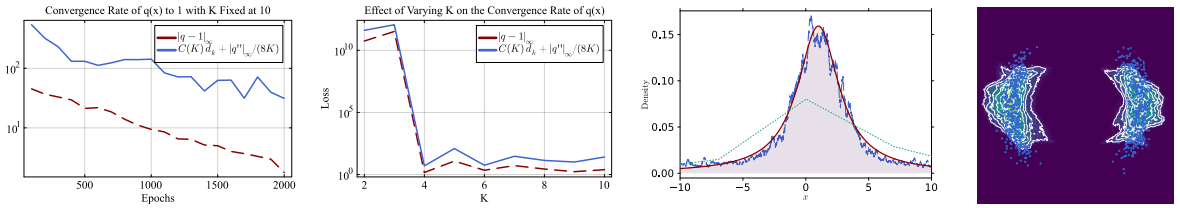
5 Sliced multivariate ISL via Bernstein polynomial approximation

When the data are multidimensional, the target $p(x)$ is a pdf on $\mathbb{R}^d$, with $d > 1$, and there is no finite set of univariate statistics that uniquely characterizes an arbitrary density (cf. Theorem 2.3 in 1D). Instead, we employ a sliced strategy: we assess a d-dimensional distribution by projecting it onto many random directions, computing the one-dimensional ISL discrepancy along each slice, and then averaging these values Kolouri et al. (2019).

One–dimensional projected statistic

For any unit vector in the d-dimensional sphere $s \in \mathbb{S}^d \subset \mathbb{R}^{d+1}$, denote by $s\#p$ the pdf of the one-dimensional projection $y = s^\top x$ with associated cdf denoted by $\tilde{F}_s$. Then the pmf of order K in direction s is

$$\begin{aligned}\mathbb{Q}_K^s(n) &= \int_{\mathbb{R}} \binom{K}{n} [\tilde{F}_s(y)]^n\, [1 - \tilde{F}_s(y)]^{K-n}\, (s\#p)(y)\, \mathrm{d}y \\ &= \int_0^1 \binom{K}{n} t^n (1-t)^{K-n} \frac{s\#p(\tilde{F}_s^{-1}(t))}{s\#\tilde{p}(\tilde{F}_s^{-1}(t))} \mathrm{d}t \\ &= \langle b_{n,K}, q^s \circ \tilde{F}_s^{-1} \rangle_{L^2([0,1])}, \qquad n = 0, \dots, K.\end{aligned}$$



(a) Fixed $K = 10$. $N = 1000$, lr 10^{-3}.

(b) Varying K. 1000 epochs, $N = 1000$, lr 10^{-3}.

(c) 1D Cauchy: true (red), dual-ISL K=10 (blue), KDE (green).

(d) Two-moons: dual-ISL density.

Figure 2: Four complementary views of dual-ISL. (a) Convergence at fixed K=10: the dashed red curve shows the empirical error $\|q-1\|_\infty$, while the solid blue curve shows the Bernstein bound from Eq. 3, $\|q_K - 1\|_\infty \le (K+1)^2 d_K$. (b) Effect of K: both the empirical error (red) and the theoretical bound (blue) decrease as K grows. (c) 1D Cauchy density estimation: dual-ISL (K=10, blue) closely tracks the ground-truth density (red) and improves over a KDE baseline (green). (d) 2D two-moons: dual-ISL's learned density (contours) aligns with the sample cloud, capturing the manifold structure.

where we have denoted by $q^s(x) = s\#q(x) = \dfrac{s\#p}{s\#\tilde{p}}(x)$ the push-forward of the quotient q by the linear transformation s.

Sliced ISL divergence

We then define the *sliced* ISL discrepancy by integrating over the unit sphere,

$$d_K^{\mathbb{S}^d}(p,\tilde{p}) \;=\; \int_{\mathbb{S}^d} d_K\big(s\#p,\; s\#\tilde{p}\big)\,\mathrm{d}s, \tag{6}$$

where d_K is the discrepancy in Definition 2. In practice, to approximate the integral in Equation 6, one randomly samples a finite set of directions $\{s_\ell\}_{\ell=1}^L$ and averages the resulting evaluations.

The following Theorem is derived using the bounds of Equation 3, and shows that, under mild smoothness assumptions on $q(x) = p(x)/\tilde{p}(x)$, if $\lim_{K\to\infty} d_K^{\mathbb{S}^d}(p,\tilde{p}) = 0$ then every one-dimensional projected ratio q^s converges uniformly to 1. By the Cramér–Wold theorem (Billingsley, 2017, Thm. 29.4), this ensures that $p = \tilde{p}$ on $\mathbb{R}^d$, and hence $d_{\mathbb{S}^d}^K$ becomes a proper divergence as $K \to \infty$.

Theorem 5.1 (Uniform convergence under slicing)**.** *Let $p, \tilde{p} \in C^2\left(\mathbb{R}^d\right)$. Then there is a constant $C_d = \mathcal{L}(\mathbb{S}^d)$ such that*

$$(i)\quad \int_{\mathbb{S}^d} \|q^s - 1\|_\infty \,\mathrm{d}s \le (K+1)^2\, d_K^{\mathbb{S}^d}(p,\tilde{p}) + C_d\,\frac{\|\nabla^2 q\|_\infty}{8K}$$

$$(ii)\quad \sup_{s\in\mathbb{S}^d} \|q^s - 1\|_\infty \le (K+1)^2 \sup_{s\in\mathbb{S}^d} d_K\big(s\#p,\; s\#\tilde{p}\big) + \frac{\|\nabla^2 q\|_\infty}{8K}$$

Here $\|\nabla^2 q\|_\infty = \sup_{x\in[0,1]^d} \|\nabla^2 q(x)\|$ and $\mathcal{L}(\mathbb{S}^d)$ is the surface measure of the sphere.

Proof. See Appendix C. □

Since $s \mapsto s\#(\cdot)$ is linear, compactness of $\mathbb{S}^d$ plus Theorems 3.1–3.2 imply that $(p,\tilde{p}) \mapsto d_K^{\mathbb{S}^d}(p,\tilde{p})$ is continuous and convex in its first argument. Consequently, by interchanging the roles of the model and target distributions in the slicing framework we obtain a *sliced dual-ISL* method that retains both convexity and differentiability (almost everywhere) under mild smoothness of the network parameters. Pseudocode for its implementation is given in Appendix E.

5.1 High-dimensional image generation

MNIST and Fashion-MNIST We evaluate sliced dual-ISL on MNIST and Fashion-MNIST by first pretraining the generator with the sliced dual-ISL objective (no discriminator), then fine-tuning adversarially. Following Sajjadi et al. (2018), we report Precision (sample fidelity) and Recall (sample diversity), together with F_β scores ($\beta \in \{1/8, 8\}$). All baselines train for 40 epochs with batch size 128; for the DCGAN pretrained with dual-ISL variant (dual-ISL+DCGAN), we run 20 epochs of dual-ISL pretraining followed by 40 epochs of DCGAN fine–tuning.

Table 2 shows that this GAN pretraining with dual-ISL, consistently improves the Precision–Recall trade–off over vanilla DCGAN. See Appendix D.7 for additional results on mode coverage (digit distributions), ablations, and runtime comparisons.

CelebA We further evaluate on CelebA and follow the same training protocol as in the MNIST/FM-NIST experiments. We report Fréchet Inception Distance (FID; lower is better) and the Precision/Recall metrics. Table 3 shows that *dual-ISL+DCGAN* attains the highest Recall among all compared methods, indicating the best mode coverage. In contrast, precision-oriented baselines (e.g., SN-DCGAN) achieve very

Dataset	Method	$F_{1/8}\uparrow$	$F_8\uparrow$	Precision↑	Recall↑
MNIST	dual-ISL (m=20)	85.00 ± 0.32	95.17 ± 1.76	84.85 ± 1.20	95.35 ± 1.39
	dual-ISL (m=50)	85.69 ± 0.29	95.81 ± 1.24	85.55 ± 1.11	96.23 ± 1.98
	DCGAN	93.58 ± 0.64	75.66 ± 1.46	93.85 ± 1.45	75.43 ± 2.56
	dual-ISL with DCGAN	$\mathbf{93.58 \pm 0.84}$	$\mathbf{95.82 \pm 1.61}$	$\mathbf{94.03 \pm 1.82}$	$\mathbf{96.68 \pm 2.42}$
FMNIST	dual-ISL (m=20)	81.84 ± 0.11	91.08 ± 1.83	81.48 ± 1.43	91.49 ± 2.15
	dual-ISL (m=50)	83.90 ± 0.09	91.18 ± 1.57	84.08 ± 1.31	$\mathbf{92.92 \pm 1.23}$
	DCGAN	86.14 ± 0.11	88.92 ± 1.51	86.60 ± 1.58	88.97 ± 1.33
	dual-ISL with DCGAN	$\mathbf{91.43 \pm 0.19}$	$\mathbf{91.87 \pm 1.57}$	$\mathbf{91.88 \pm 1.35}$	92.42 ± 1.47

Table 2: Quantitative comparison on MNIST/FMNIST. We report $F_{1/8}$ and F_8 (β-weighted F-scores of Precision/Recall with $\beta = 1/8$ and 8), Precision, and Recall (mean±std, %). Methods: dual-ISL with $m \in \{20, 50\}$ random projections, DCGAN, and DCGAN pretrained with dual-ISL.

Method	FID↓	Precision↑	Recall↑
DCGAN (2015)	30.93	0.839	0.834
dual-ISL with DCGAN	30.54	0.893	**0.952**
LS-DCGAN (2017)	**22.99**	0.932	0.524
dual-ISL with LS-DCGAN	23.56	0.928	0.824
W-DCGAN-GP (2017)	40.32	0.847	0.884
SN-DCGAN (2018)	32.94	**0.974**	0.887
DynGAN (2024)	48.06	0.955	0.718
GMAN (2016)	31.66	0.873	0.888

Table 3: CelebA (64×64). Best values per column in **bold**. dual-ISL with DCGAN gives the highest recall (diversity); LS-DCGAN achieves the lowest FID; SN-DCGAN attains the highest precision.

strong Precision but lower Recall. When we pretrain an LS-DCGAN generator with sliced dual-ISL (*dual-ISL+LS-DCGAN*), Recall improves markedly relative to vanilla LS-DCGAN (0.824 vs. 0.524) with only a small change in FID (23.56 vs. 22.99) and negligible Precision difference (0.928 vs. 0.932). Overall, the results underscore a consistent pattern we also observe on MNIST/FMNIST: dual-ISL pretraining substantially improves diversity (Recall) while keeping fidelity competitive. Further runtime comparison results appear in Appendix D.7.

6 Summary and limitations

We introduced *dual-ISL*, a likelihood-free, convex objective for training implicit generators. By swapping the roles of data and model in ISL, the resulting discrepancy is continuous and convex in the model density. An L^2 projection view shows that dual-ISL estimates the push-forward density ratio via a Bernstein basis, yielding a closed-form surrogate density with truncation and error bounds. A sliced extension preserves these properties in $\mathbb{R}^d$. Empirically, dual-ISL attains smoother, faster convergence and stronger mode coverage than standard ISL and competitive GAN baselines on MNIST/FMNIST and CelebA.

Relying on 1D slicing introduces an accuracy–compute trade-off: capturing complex cross-dimensional structure may require many directions m, while cost grows roughly linearly in m (and in K). Performance is also sensitive to the Bernstein order K: larger K better resolves sharp features but can increase variance and instability, whereas smaller K may underfit. Beyond random slicing, structured or learned projections (e.g., gradient-aligned directions or low-discrepancy spherical designs) can increase information per slice. A natural next step is to adapt m, the slice directions, and *per-slice* orders K via principled selection/weighting schemes, together with tighter finite-sample guarantees. See Appendix G for an extended discussion of limitations and future directions.

Bibliography

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, 2011.

Jiahao Chen and Jarrett Revels. Robust benchmarking in noisy environments. *arXiv preprint arXiv:1608.04295*, 2016.

Xiongjie Chen, Yongxin Yang, and Yunpeng Li. Augmented sliced Wasserstein distances. *arXiv preprint arXiv:2006.08812*, 2020.

Jinyoung Choi and Bohyung Han. Mcl-gan: Generative adversarial networks with multiple specialized discriminators. In *Advances in Neural Information Processing Systems*, volume 35, pages 29597–29609, 2022.

José Manuel de Frutos, Manuel A Vázquez, Pablo Olmos, and Joaquín Míguez. Robust training of im-

plicit generative models for multivariate and heavy-tailed distributions with an invariant statistical loss. *arXiv preprint arXiv:2410.22381*, 2024a.

José Manuel de Frutos, Pablo Olmos, Manuel A Vázquez, and Joaquín Míguez. Training implicit generative models via an invariant statistical loss. In *International Conference on Artificial Intelligence and Statistics*, pages 2026–2034. PMLR, 2024b.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. In *International Conference on Learning Representations (ICLR)*, 2020.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3483–3491, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10648–10656, 2019.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162 (3):707–738, 2015.

Ian J 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*, volume 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Henryk Gzyl and Jose Luis Palacios. The Weierstrass approximation theorem and large deviations. *The American mathematical monthly*, 104(7):650–653, 1997.

Todd Huster, Jeremy Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi O Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto gan: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pages 4523–4532. PMLR, 2021.

Bert Jiittler. The dual basis functions for the bernstein polynomials. *Advances in Computational Mathematics*, 8(4):345–352, 1998.

Benjamin Keinert, Matthias Innmann, Michael Sänger, and Marc Stamminger. Spherical fibonacci mapping. *ACM Transactions on Graphics (TOG)*, 34(6):1–7, 2015.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018.

George G Lorentz. *Bernstein polynomials*. American Mathematical Soc., 2012.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

Peter Mathé. Approximation of hölder continuous functions by bernstein polynomials. *The American mathematical monthly*, 106(6):568–574, 1999.

Robert E Megginson. *An introduction to Banach space theory*, volume 183. Springer Science & Business Media, 2012.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 35, pages 28179–28193, 2022.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep con-

volutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

Rubing Shen, Yanzhen Ren, and Zongkun Sun. Fagan: Artifacts-free and phase-aware high-fidelity gan-based vocoder. In *Proc. Interspeech 2024*, pages 3884–3888, 2024.

Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. normflows: A pytorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361, 2023. doi: 10.21105/joss.05361. URL https://doi.org/10.21105/joss.05361.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.

Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

Manzil Zaheer, C-l Li, Barnabás Póczos, and Ruslan Salakhutdinov. Gan connoisseur: Can gans learn simple 1D parametric distributions. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 1–6, 2017.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.** See sections *The Invariant Statistical Loss (ISL)*, *The dual-invariant statistical loss*, *An L^2-projection view of d_K*, and *Sliced multivariate ISL via Bernstein polynomial approximation*. Pseudocode is in *Appendix: Pseudocode*.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes.** We analyze continuity/weak continuity and convexity in *The dual-invariant statistical loss*; truncation/error bounds and convergence rates in *An L^2-projection view of d_K*; and provide runtime notes in the *Supplementary Materials*.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes.** An anonymized repository URL with dependencies and reproduction scripts is provided in the *Supplementary Materials*.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes.** Assumptions are stated with each theorem in *The dual-invariant statistical loss*, *An L^2-projection view of d_K*, and *Sliced multivariate ISL via Bernstein polynomial approximation*.
   (b) Complete proofs of all theoretical results. **Yes.** Proofs appear in *Appendix: Proofs for "The dual-invariant statistical loss"*, *Appendix: Proofs for "An L^2-projection view of d_K"*, and *Appendix: Multidimensional Extension of the Projection Schema*.
   (c) Clear explanations of any assumptions. **Yes.** Each result includes a brief explanation of its assumptions and scope, with clarifying remarks where relevant.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a

URL). **Yes.** An anonymized repo, scripts, and step-by-step commands are provided in the *Supplementary Materials*.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.** Architectures, data splits, and hyperparameters (including K and the number of slices m) are detailed in *The dual-invariant statistical loss* and *Sliced multivariate ISL via Bernstein polynomial approximation*, with full configs in the *Supplementary Materials*.

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.** We define KSD and Precision/Recall; tables and plots report mean $\pm$ std across runs (see the table titled "KSD over 10 runs for Dual-ISL and baselines" and "Quantitative comparison on MNIST/FMNIST").

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes.** Hardware and runtime details are provided in the *Supplementary Materials*.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. **Yes.** We cite MNIST, Fashion-MNIST, CelebA, and compared methods.

(b) The license information of the assets, if applicable. **Not Applicable.** Public datasets were used without redistribution; licenses are respected via citation.

(c) New assets either in the supplemental material or as a URL, if applicable. **Yes.** We release anonymized code and experiment configs; figures/tables are reproducible from provided scripts.

(d) Information about consent from data providers/curators. **Not Applicable.**

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. **Not Applicable.**

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**