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

José Manuel de Frutos Universidad Carlos III
Pablo M. Olmos Universidad Carlos III
Manuel A. Vázquez Universidad Carlos III
Joaquín Míguez Universidad Carlos III

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

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

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

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

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 $\{0, \ldots, K\}$, i.e. $\mathbb{Q}_K(n) = \frac{1}{K+1}$ for all $n \in \{0, \ldots, K\}$.*

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

$$\begin{aligned} d_K(p,\tilde{p}) &:= \frac{1}{K+1}\|\mathbb{Q}_K - \mathbb{U}_K\|_{\ell_1} \\ &= \frac{1}{K+1}\sum_{n=0}^{K}\Big|\tfrac{1}{K+1} \;-\; \mathbb{Q}_K(n)\Big| \\ &= \frac{2}{K+1}\mathrm{TV}(\mathbb{Q}_K,\mathbb{U}). \end{aligned} \tag{2}$$

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 $d_K\big(p,\tilde{p}_\theta\big)$ 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 $\{p_n\}_{n\geq 1}$ 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

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.

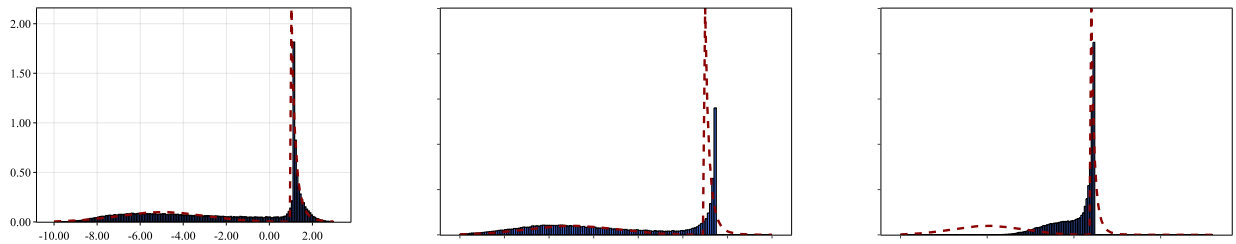
4.1 Projection interpretation

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*

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.

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 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

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

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).

Sketch of proof. The full proof is given in Appendix B.

Let $q = p/\tilde{p}$ and let $\tilde{F}$ be the cdf of $\tilde{p}$. Since $\tilde{p} > 0$ and $p, \tilde{p} \in C(\mathbb{R})$, $\tilde{F}$ is strictly increasing, so the change of variables $t = \tilde{F}(x)$ induces a continuous function $\tilde{q} : [0,1] \to \mathbb{R}$, $\tilde{q}(t) := q\Big(\tilde{F}^{-1}(t)\Big)$.

For fixed K, let

$$V_K = \operatorname{span}\{b_{0,K}, \dots, b_{K,K}\} \subset L^2([0,1]).$$

The degree-K Bernstein truncation $\tilde{q}_K$ is the $L^2([0,1])$-projection of $\tilde{q}$ onto V_K, and using the dual Bernstein basis $\{\tilde{b}_{m,K}\}_{m=0}^{K}$ it admits the representation

$$\tilde{q}_K(t) = \sum_{m=0}^{K} \big\langle \tilde{q}, b_{m,K} \big\rangle_{L^2([0,1])} \, \tilde{b}_{m,K}(t).$$

By Theorem 4.1, the coefficients satisfy

$$\mathbb{Q}_K(m) = \big\langle \tilde{b}_{m,K}, q \big\rangle_{L^2([0,1])}, \qquad m = 0, \dots, K.$$

Therefore,

$$\tilde{q}_K(t) = \sum_{m=0}^{K} \mathbb{Q}_K(m) \, \tilde{b}_{m,K}(t),$$

which is the claimed Bernstein-basis truncation formula. □

Remark 4.1. *With $\tilde{p}$ fixed, the map $\Phi_K : p \longmapsto \big(\mathbb{Q}_K(0), \dots, \mathbb{Q}_K(K)\big)$, 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,*

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

Thus Φ_K factors through the quotient of $L^2([0,1])$ by the subspace orthogonal to $\operatorname{span}\{b_{n,K}\}$, 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*

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

Sketch of proof. See Appendix B for the full proof.

Using the dual-basis representation, we add and subtract $1/(K+1)$ inside the sum,

$$\begin{aligned} \tilde{q}_K(t) &= \sum_{m=0}^{K} \mathbb{Q}_K(m) \, \tilde{b}_{m,K}(t) \\ &= \sum_{m=0}^{K} \Big[\frac{1}{K+1} - \Big(\frac{1}{K+1} - \mathbb{Q}_K(m) \Big) \Big] \tilde{b}_{m,K}(t). \end{aligned}$$

By Lemma B.1, we have $\sum_{m=0}^{K} \tilde{b}_{m,K}(t) = K+1$, and applying the triangle inequality yields

$$\begin{aligned} |\tilde{q}_K(t) - 1| &\le (K+1) \sum_{m=0}^{K} \left| \frac{1}{K+1} - \mathbb{Q}_K(m) \right| \\ &= (K+1)^2 d_K(p, \tilde{p}). \end{aligned}$$

Since $q_K(x) = \tilde{q}_K(\tilde{F}(x))$, taking suprema over $x \in \mathbb{R}$ and $t \in [0,1]$ is equivalent, yielding the claim. □

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).

Having quantified the approximation error of the Bernstein ratio approximation, we next translate the Bernstein projection view into an explicit density reconstruction. Using the coefficients $\mathbb{Q}_K(m)$ to approximate the density ratio and multiplying by the reference density $\tilde{p}$, we define a closed-form estimator p_K. Theorem 4.4 shows that this reconstruction is consistent, i.e., $p_K(x) \to p(x)$ pointwise as $K \to \infty$.

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

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

Figures 2(c–d) show that ISL's Bernstein-based approximation accurately recovers the true density in both 1D and 2D. In Figure 2c, we compare the ground-truth heavy-tailed Cauchy density with location 1 and scale 2 (red) against the dual-ISL estimate at $K = 10$ (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

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

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*

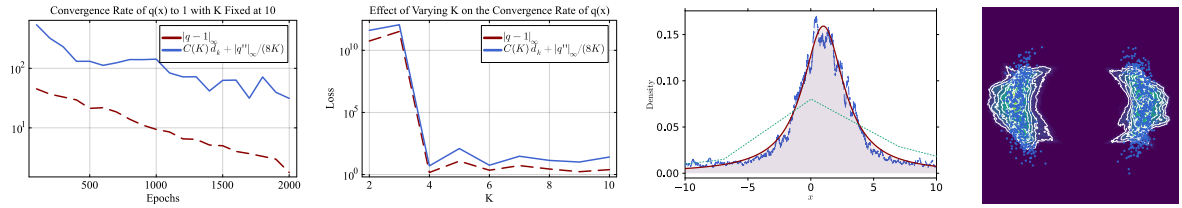
$$\textit{(i)} \quad \int_{\mathbb{S}^d} \|q^s - 1\|_\infty \, \mathrm{d}s \le (K+1)^2 \, d_K^{\mathbb{S}^d}(p, \tilde{p}) + C_d \, \frac{\|\nabla^2 q\|_\infty}{8K}$$

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

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,



(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.

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/FMNIST experiments. We report Fréchet Inception Distance (FID; lower is better) and the Precision/Recall metrics. Table 10 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 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.

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.

Acknowledgements

This work was partially supported by the Office of Naval Research (award N00014-22-1-2647) and by Spain's Agencia Estatal de Investigación (ref. PID2024-158181NB-I00 NISA, PID2024-157856NB-I00 CARTESIAN, and PID2021-123182OB-I00 EPiCENTER), funded by MCIN/AEI/10.13039/501100011033 and by the ERDF ("A way of making Europe"). Pablo M. Olmos was also supported by the Comunidad de Madrid (IND2024/TIC-34728), the IDEA-CM project (TEC-2024/COM-89), the ELLIS Unit Madrid (European Laboratory for Learning and Intelligent Systems), and the 2024 Leonardo Grant for Scientific Research and Cultural Creation from the BBVA Foundation, and has also received funding from the EU Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101226456 (MLCARE: Machine Learning Computational Advancements for peRsonalized mEdicine). The views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the EU or the European Commission (EC). Neither the EU nor the EC can be held responsible for them.

References

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 implicit 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 convolutional 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

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.**

SUPPLEMENTARY MATERIALS

A PROOFS OF THEOREMS SECTION 3

In this appendix, we establish three key analytic properties of the rank-based divergence d_K

- Continuity in its first argument under weak convergence (Theorem 3.1).
- Continuity in its second argument under L^1 norm (Theorem A.1).
- Convexity in its first argument (Theorem 3.2).

A.1 Proof of Theorem 3.1

Proof of Theorem 3.1.

(i) For each fixed $m \in \{0, 1, \ldots, K\}$, define

$$h_m(y) = \binom{K}{m} [\tilde{F}(y)]^m \, [1 - \tilde{F}(y)]^{K-m}.$$

Since $\tilde{F}(y)$ is the cdf of the fixed density $\tilde{p}$, the function $h_m(y)$ is continuous and bounded on $\mathbb{R}$. By the definition of weak convergence $p_n \xrightarrow{w} p$ we have that for every bounded continuous function h,

$$\lim_{n\to\infty} \int_{\mathbb{R}} h(y)\, p_n(y)\, dy = \int_{\mathbb{R}} h(y)\, p(y)\, dy.$$

Taking $h(y) = h_m(y)$ yields

$$\lim_{n\to\infty} \mathbb{Q}_K^{(n)}(m) = \lim_{n\to\infty} \int_{\mathbb{R}} h_m(y)\, p_n(y)\, dy = \int_{\mathbb{R}} h_m(y)\, p(y)\, dy = \mathbb{Q}_K(m).$$

(ii) The discrepancy between p and $\tilde{p}$ is

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

Since for each m we have shown that

$$\lim_{n\to\infty} \mathbb{Q}_K^{(n)}(m) = \mathbb{Q}_K(m),$$

and because the absolute value function is continuous, it follows that

$$\lim_{n\to\infty} \left| \frac{1}{K+1} - \mathbb{Q}_K^{(n)}(m) \right| = \left| \frac{1}{K+1} - \mathbb{Q}_K(m) \right|.$$

As the sum is finite (from $m = 0$ to K), we can exchange the limit and the summation to conclude that

$$\begin{aligned} \lim_{n\to\infty} d_K(p_n, \tilde{p}) &= \frac{1}{K+1} \sum_{m=0}^{K} \lim_{n\to\infty} \left| \frac{1}{K+1} - \mathbb{Q}_K^{(n)}(m) \right| \\ &= \frac{1}{K+1} \sum_{m=0}^{K} \left| \frac{1}{K+1} - \mathbb{Q}_K(m) \right| = d_K(p, \tilde{p}). \end{aligned}$$

□

A.2 Proof of Theorem A.1

We now state and prove that the divergence d_K is continuous with respect to its second argument in the L^1 norm.

Theorem A.1 (Continuity in the second argument)**.** *Let p and $\{\tilde{p}_n\}_{n\geq 1}$ be continuous densities on $\mathbb{R}$ such that $\tilde{p}_n \to \tilde{p}$ in the L^1 norm. Then, the discrepancy function d_K is continuous in its second argument, i.e.*

$$\lim_{n\to\infty} d_K\big(p, \tilde{p}_n\big) \;=\; d_K(p, \tilde{p}).$$

Proof. Recall that for any pair of densities $p, \tilde{p}$ we define (in Section 4)

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

and we write the mth component of $\Phi_K(p, \tilde{p})$ as

$$[\Phi_K(p, \tilde{p})]_m \;=\; \mathbb{Q}_K(m).$$

Thus, if $\tilde{p}_n$ has cdf $\tilde{F}_n(y)$, then by definition

$$[\Phi_K(p, \tilde{p}_n)]_m = \int_{\mathbb{R}} \binom{K}{m} \big[\tilde{F}_n(y)\big]^m \big[1 - \tilde{F}_n(y)\big]^{K-m} p(y)\,\mathrm{d}y.$$

Hence,

$$\begin{aligned}
&\|\Phi_K(p, \tilde{p}_n) - \Phi_K(p, \tilde{p})\|_{\ell^1} \\
&\qquad = \sum_{m=0}^{K} \left| \int_{\mathbb{R}} \binom{K}{m} \Big([\tilde{F}_n(y)]^m [1 - \tilde{F}_n(y)]^{K-m} \;-\; [\tilde{F}(y)]^m [1 - \tilde{F}(y)]^{K-m} \Big) p(y)\,\mathrm{d}y \right|
\end{aligned}$$

and by the triangle inequality we have

$$\begin{aligned}
&\|\Phi_K(p, \tilde{p}_n) - \Phi_K(p, \tilde{p})\|_{\ell^1} \\
&\qquad \leq \sum_{m=0}^{K} \binom{K}{m} \int_{\mathbb{R}} \Big| [\tilde{F}_n(y)]^m [1 - \tilde{F}_n(y)]^{K-m} - [\tilde{F}(y)]^m [1 - \tilde{F}(y)]^{K-m} \Big| p(y)\,\mathrm{d}y.
\end{aligned}$$

The function

$$f(a) \;=\; a^m\,(1-a)^{K-m},$$

is Lipschitz on $[0, 1]$ with a Lipschitz constant $C_{K,m} < \infty$. As a consequence,

$$\big| f(\tilde{F}_n(y)) - f(\tilde{F}(y)) \big| \leq C_K \,\big| \tilde{F}_n(y) - \tilde{F}(y) \big|,$$

where $C_K = \sup_{0\leq m\leq K} C_{K,m} < \infty$, and therefore,

$$\|\Phi_K(p, \tilde{p}_n) - \Phi_K(p, \tilde{p})\|_{\ell^1} \;\leq\; C'_K \int_{\mathbb{R}} \big| \tilde{F}_n(y) - \tilde{F}(y) \big|\, p(y)\,\mathrm{d}y, \tag{7}$$

where the constant $C'_K = C_K \sum_{m=0}^{K} \binom{K}{m} < \infty$ depends only on K.

Note that

$$\tilde{F}_n(y) \;-\; \tilde{F}(y) \;=\; \int_{-\infty}^{y} \big(\tilde{p}_n(t) - \tilde{p}(t)\big)\,\mathrm{d}t,$$

hence

$$\big| \tilde{F}_n(y) - \tilde{F}(y) \big| \;\leq\; \int_{\mathbb{R}} \big| \tilde{p}_n(t) - \tilde{p}(t) \big|\,\mathrm{d}t \;=\; \|\tilde{p}_n - \tilde{p}\|_{L^1},$$

and, as a consequence,

$$\int_{\mathbb{R}} \big|\tilde{F}_n(y) - \tilde{F}(y)\big| p(y)\,\mathrm{d}y \;\le\; \|\tilde{p}_n - \tilde{p}\|_{L^1} \int_{\mathbb{R}} p(y)\,\mathrm{d}y \;=\; \|\tilde{p}_n - \tilde{p}\|_{L^1}, \tag{8}$$

since $\int_{\mathbb{R}} p(y)\,\mathrm{d}y = 1$.

Combining Equation 7 and 8 we arrive at

$$\|\Phi_K(p,\tilde{p}_n) - \Phi_K(p,\tilde{p})\|_{\ell^1} \;\le\; C'_K\,\|\tilde{p}_n - \tilde{p}\|_{L^1} \;\longrightarrow\; 0 \quad \text{as } n \to \infty.$$

Finally, since

$$d_K\big(p,\tilde{p}_n\big) \;=\; \frac{1}{K+1}\,\Big\|\Phi_K(p,\tilde{p}_n) - \mathbb{U}_K\Big\|_{\ell^1},$$

we have

$$\big|d_K\big(p,\tilde{p}_n\big) \;-\; d_K\big(p,\tilde{p}\big)\big| \;=\; \frac{1}{K+1}\big|\|\Phi_K(p,\tilde{p}_n) - \mathbb{U}_K\|_{\ell^1} - \|\Phi_K(p,\tilde{p}) - \mathbb{U}_K\|_{\ell^1}\big|. \tag{9}$$

By the triangle inequality for ℓ^1,

$$|\|\Phi_K(p,\tilde{p}_n) - \mathbb{U}_K\|_{\ell^1} - \|\Phi_K(p,\tilde{p}) - \mathbb{U}_K\|_{\ell^1}| \;\le\; \|\Phi_K(p,\tilde{p}_n) - \Phi_K(p,\tilde{p})\|_{\ell^1}, \tag{10}$$

hence taking 9 and 10 together yields

$$\big|d_K\big(p,\tilde{p}_n\big) \;-\; d_K\big(p,\tilde{p}\big)\big| \;\le\; \frac{1}{K+1}\,\|\Phi_K(p,\tilde{p}_n) - \Phi_K(p,\tilde{p})\|_{\ell^1}.$$

Since $\|\Phi_K(p,\tilde{p}_n) - \Phi_K(p,\tilde{p})\|_{\ell^1} \to 0$, it follows that

$$\lim_{n\to\infty} d_K\big(p,\tilde{p}_n\big) \;=\; d_K\big(p,\tilde{p}\big),$$

□

A.3 Proof of Theorem 3.2

Finally we give the proof of the convexity of d_K w.r.t. the first argument.

Proof of Theorem 3.2. Let p_1, p_2 be two densities and $\lambda \in [0,1]$. Define

$$p(y) = \lambda p_1(y) + (1-\lambda) p_2(y).$$

Since

$$[\Phi_K(p,\tilde{p})]_n = \int_{\mathbb{R}} \binom{K}{n} [\tilde{F}(y)]^n [1-\tilde{F}(y)]^{K-n}\, p(y)\,\mathrm{d}y,$$

linearity of the integral gives

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

Hence, for each n,

$$\begin{aligned}\Big|[\Phi_K(p,\tilde{p})]_n - \tfrac{1}{K+1}\Big| \;&=\; \Big|\lambda\big([\Phi_K(p_1,\tilde{p})]_n - \tfrac{1}{K+1}\big) + (1-\lambda)\big([\Phi_K(p_2,\tilde{p})]_n - \tfrac{1}{K+1}\big)\Big| \\ &\le\; \lambda\,\Big|[\Phi_K(p_1,\tilde{p})]_n - \tfrac{1}{K+1}\Big| + (1-\lambda)\,\Big|[\Phi_K(p_2,\tilde{p})]_n - \tfrac{1}{K+1}\Big|.\end{aligned}$$

Summing over $n = 0,\dots,K$ and dividing by $K+1$ yields

$$d_K(p,\tilde{p}) = \frac{1}{K+1}\sum_{n=0}^{K}\big|\Phi_K(p,\tilde{p})_n - \tfrac{1}{K+1}\big| \le \lambda\, d_K(p_1,\tilde{p}) + (1-\lambda)\, d_K(p_2,\tilde{p}).$$

□

B PROOFS OF THEOREMS SECTION 4

In this appendix we give complete proofs for the theorems and claims in Sections 4.

- Theorem B.1. Characterizes the binomial mapping Φ_K, showing it is well-defined, linear in its first argument, bounded, and continuous under mild regularity conditions.
- Theorem 4.1. Establishes the Riesz representation of Φ_K, expressing each probability mass $\mathbb{Q}_K(n)$ as an L^2 inner product with a Bernstein basis function.
- Theorem 4.3. Provides a uniform bound on the deviation $\|q_K - 1\|_\infty$ in terms of the discrepancy $d_K(p, \tilde{p})$.
- Theorem 4.4. Derives an explicit Bernstein-based representation for a push-forward density p_K via a continuously differentiable map.

B.1 Properties of the map Φ_K

Theorem B.1. *Let p and $\tilde{p}$ be pdfs on $\mathbb{R}$, with cdfs F and $\tilde{F}$, respectively. For each integer $K \geq 1$, recall that*

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

where

$$\mathbb{Q}_K(n) \;=\; \int_{\mathbb{R}} \binom{K}{n} \big[\tilde{F}(y)\big]^n \big[1 - \tilde{F}(y)\big]^{K-n} p(y)\,\mathrm{d}y, \quad n = 0, \ldots, K. \tag{11}$$

Then the following properties hold

(i) Well-definedness. *For each fixed pair $(p, \tilde{p})$ and integer K, the integral on the right -hand side of Equation 11 uniquely determines a pmf $\mathbb{Q}_K$.*

(ii) Non-surjectivity. *Let Δ^K be the set of all pmfs on $\{0, 1, \ldots, K\}$. The image of Φ_K is strictly contained in Δ^K.*

(iii) Continuity. *Assume that $\|p_n - p\|_{L^1(\mathbb{R})} \to 0$ and $\|\tilde{F}_n - \tilde{F}\|_\infty \to 0$. Then*

$$\big\|\Phi_K\big(p_n, \tilde{p}_n\big) \;-\; \Phi_K\big(p, \tilde{p}\big)\big\|_{\ell^1} \;\longrightarrow\; 0.$$

(iv) Linearity. *The operator Φ_K is linear in its first argument. Thus, for any $\alpha \in [0, 1]$,*

$$\Phi_K\big(\alpha p_1 + (1-\alpha) p_2, \tilde{p}\big) = \alpha\, \Phi_K(p_1, \tilde{p}) + (1-\alpha)\, \Phi_K(p_2, \tilde{p}).$$

(v) Bounded operator. *Φ_K is a bounded operator from the space of continuous pdfs p on $\mathbb{R}$ to the space of pmfs $\mathbb{Q}_K$ on $\{0, 1, \ldots, K\}$. Specifically*

$$|||\Phi_K||| = \sup_{\|p\|_{L^1} \leq 1} \|\Phi_K(p, \tilde{p})\|_{TV} = \|\mathbb{Q}_K\|_{TV} = 1,$$

where $|||\cdot|||$ denotes the operator norm and $\|\cdot\|_{TV}$ the total variation norm.

Proof of Theorem B.1.

(i) Fix $K \geq 0$ and recall that

$$\mathbb{Q}_K(n) = \int_{\mathbb{R}} \binom{K}{n} [\tilde{F}(y)]^n\, [1 - \tilde{F}(y)]^{K-n}\, p(y)\,\mathrm{d}y, \qquad n = 0, \ldots, K.$$

Non-negativity. Since $p(y) \geq 0$ and each Bernstein-integrand $\binom{K}{n}[\tilde{F}(y)]^n[1 - \tilde{F}(y)]^{K-n} \geq 0$, it follows that $\mathbb{Q}_K(n) \geq 0$ for every n.

Normalization. By Fubini's theorem we may interchange sum and integral, i.e.,

$$\sum_{n=0}^{K} \mathbb{Q}_K(n) = \int_{\mathbb{R}} p(y) \sum_{n=0}^{K} \binom{K}{n} [\tilde{F}(y)]^n [1-\tilde{F}(y)]^{K-n} \, \mathrm{d}y.$$

But the inner sum is $(\tilde{F}(y) + (1-\tilde{F}(y)))^K = 1^K = 1$ by the binomial theorem, hence

$$\sum_{n=0}^{K} \mathbb{Q}_K(n) = \int_{\mathbb{R}} p(y) \, \mathrm{d}y = 1.$$

Uniqueness. Each $\mathbb{Q}_K(n)$ is defined by a single integral depending only on p and $\tilde{F}$. Thus the mapping $\Phi_K : (p, \tilde{p}) \mapsto (\mathbb{Q}_K(0), \ldots, \mathbb{Q}_K(K))$ is well-defined and unique for each choice of $(p, \tilde{p})$ and K.

(ii) To see that Φ_K is not surjective, it suffices to exhibit a pmf in Δ^K that cannot arise from any $(p, \tilde{p})$. We do this for $K = 2$.

Define

$$\mathbb{Q} = (0, 1, 0) \in \Delta^2,$$

so that $\mathbb{Q}(0) = 0$, $\mathbb{Q}(1) = 1$, and $\mathbb{Q}(2) = 0$. Suppose, for the sake of contradiction, that there exist densities $p, \tilde{p}$ on $\mathbb{R}$ with cdf $\tilde{F}$ such that $\Phi_2(p, \tilde{p}) = \mathbb{Q}$. Then by definition

$$\begin{aligned}
\mathbb{Q}(0) &= \int_{\mathbb{R}} [1-\tilde{F}(y)]^2 \, p(y) \, \mathrm{d}y = 0, \\
\mathbb{Q}(1) &= 2 \int_{\mathbb{R}} \tilde{F}(y) \, [1-\tilde{F}(y)] \, p(y) \, \mathrm{d}y = 1, \\
\mathbb{Q}(2) &= \int_{\mathbb{R}} [\tilde{F}(y)]^2 \, p(y) \, \mathrm{d}y = 0.
\end{aligned}$$

The first and third equations force

$$[1-\tilde{F}(y)]^2 = 0 \quad \text{and} \quad [\tilde{F}(y)]^2 = 0 \quad p\text{-almost everywhere},$$

hence $\tilde{F}(y) = 1$ and $\tilde{F}(y) = 0$ p-a.e. Since p is a probability density, its support has positive measure, so we cannot have $\tilde{F} \equiv 1$ and $\tilde{F} \equiv 0$ on that support. This contradiction shows no $(p, \tilde{p})$ can produce $\mathbb{Q} = (0, 1, 0)$. Therefore Φ_2, and hence Φ_K for general K, fails to be surjective.

(iii) Let $p_i \to p$ in $L^1(\mathbb{R})$ and let $\tilde{p}_j$ have cdfs $\tilde{F}_j \to \tilde{F}$ uniformly on $\mathbb{R}$. We show

$$\left\| \Phi_K(p_n, \tilde{p}_n) - \Phi_K(p, \tilde{p}) \right\|_1 \;\to\; 0.$$

By the triangle inequality,

$$\begin{aligned}
\left\| \Phi_K(p_i, \tilde{p}_n) - \Phi_K(p, \tilde{p}) \right\|_{\ell^1} \leq \; & \left\| \Phi_K(p_i, \tilde{p}_n) - \Phi_K(p_i, \tilde{p}) \right\|_{\ell^1} \\
& + \; \left\| \Phi_K(p_i, \tilde{p}) - \Phi_K(p, \tilde{p}) \right\|_{\ell^1},
\end{aligned} \tag{12}$$

and we handle each term separately.

Continuity in $\tilde{p}$. Fix p_i. For each $m = 0, \ldots, K$, set

$$g_j(y) = [\tilde{F}_j(y)]^m \, [1-\tilde{F}_j(y)]^{K-m}, \quad g(y) = [\tilde{F}(y)]^m \, [1-\tilde{F}(y)]^{K-m}.$$

Uniform convergence $\|\tilde{F}_j - \tilde{F}\|_\infty \to 0$ implies $\|g_j - g\|_\infty \to 0$. Hence for each coordinate m we have

$$\left| \Phi_K(p_i, \tilde{p}_n)_m - \Phi_K(p_i, \tilde{p})_m \right| \leq \binom{K}{m} \int_{\mathbb{R}} |g_i(y) - g(y)| \, p_i(y) \, \mathrm{d}y \leq \binom{K}{m} \|g_n - g\|_\infty,$$

and summing over m yields

$$\left\| \Phi_K(p_i, \tilde{p}_j) - \Phi_K(p_i, \tilde{p}) \right\|_{\ell^1} \;\leq\; \sum_{m=0}^{K} \binom{K}{m} \|g_j - g\|_\infty = 2^K \, \|g_j - g\|_\infty \;\longrightarrow\; 0. \tag{13}$$

Continuity in p. Fix $\tilde{p}$ (and write $\tilde{F}$ for its cdf). For each $m = 0, \dots, K$,

$$\big[\Phi_K(p_i, \tilde{p})\big]_m - \big[\Phi_K(p, \tilde{p})\big]_m = \binom{K}{m} \int_{\mathbb{R}} [\tilde{F}(y)]^m [1 - \tilde{F}(y)]^{K-m} \big(p_i(y) - p(y)\big)\, \mathrm{d}y.$$

Taking absolute values and using $\int |p_i - p| = \|p_i - p\|_{L^1}$ yields

$$\big|\Phi_K(p_i, \tilde{p})_m - \Phi_K(p, \tilde{p})_m\big| \;\le\; \binom{K}{m} \|p_i - p\|_{L^1}.$$

Summing over m gives

$$\begin{aligned}
\big\|\Phi_K(p_i, \tilde{p}) - \Phi_K(p, \tilde{p})\big\|_{\ell^1} &= \sum_{m=0}^{K} \big|\Phi_K(p_i, \tilde{p})_m - \Phi_K(p, \tilde{p})_m\big| \\
&\le \Big(\sum_{m=0}^{K} \binom{K}{m}\Big) \|p_i - p\|_{L^1} \\
&= 2^K \|p_i - p\|_{L^1} \longrightarrow 0. \qquad (14)
\end{aligned}$$

Since $p_i \to p$ in L^1, the right-hand side tends to zero. Hence $\Phi_K(\cdot, \tilde{p})$ is continuous in its first argument.

Combining in Equations 12, 13 and 14 yields

$$\big\|\Phi_K(p_i, \tilde{p}_j) - \Phi_K(p, \tilde{p})\big\|_{\ell^1} \;\longrightarrow\; 0,$$

i.e. Φ_K is jointly continuous in $(p, \tilde{p})$.

(iv) **Linearity in p.** Let p_1, p_2 be two probability densities on $\mathbb{R}$ and $\alpha \in [0, 1]$. Set

$$p = \alpha\, p_1 + (1 - \alpha)\, p_2.$$

Then for each $n = 0, \dots, K$,

$$\Phi_K(p, \tilde{p})_n = \binom{K}{n} \int_{\mathbb{R}} [\tilde{F}(y)]^n \, [1 - \tilde{F}(y)]^{K-n} \, p(y)\, \mathrm{d}y.$$

By linearity of the integral,

$$\begin{aligned}
\Phi_K(p, \tilde{p})_n &= \binom{K}{n} \int_{\mathbb{R}} [\tilde{F}]^n [1 - \tilde{F}]^{K-n} \big(\alpha p_1 + (1 - \alpha) p_2\big)\, \mathrm{d}y \\
&= \alpha\, \Phi_K(p_1, \tilde{p})_n + (1 - \alpha)\, \Phi_K(p_2, \tilde{p})_n.
\end{aligned}$$

Since this holds for every coordinate n, we conclude

$$\Phi_K\big(\alpha p_1 + (1 - \alpha) p_2,\ \tilde{p}\big) = \alpha\, \Phi_K(p_1, \tilde{p}) + (1 - \alpha)\, \Phi_K(p_2, \tilde{p}),$$

i.e. $\Phi_K(\cdot, \tilde{p})$ is linear.

(v) Boundedness and operator norm. Recall the total-variation norm on $\mathbb{R}^{K+1}$ is just the ℓ^1-norm. For any pair of densities $p, \tilde{p}$,

$$\|\Phi_K(p, \tilde{p})\|_{\mathrm{TV}} = \sum_{n=0}^{K} [\Phi_K(p, \tilde{p})]_n = \int_{\mathbb{R}} p(y) \sum_{n=0}^{K} \binom{K}{n} [\tilde{F}(y)]^n [1 - \tilde{F}(y)]^{K-n}\, \mathrm{d}y.$$

By the binomial theorem the inner sum is $\big(\tilde{F} + (1 - \tilde{F})\big)^K = 1$. Hence

$$\|\Phi_K(p, \tilde{p})\|_{\mathrm{TV}} = \int_{\mathbb{R}} p(y)\, \mathrm{d}y = \|p\|_{L^1}.$$

Taking the supremum over all p with $\|p\|_{L^1} \le 1$ shows

$$|||\Phi_K||| := \sup_{\|p\|_{L^1} \le 1} \|\Phi_K(p, \tilde{p})\|_{\mathrm{TV}} = \sum_{n=0}^{K} \mathbb{Q}_K(n) = 1.$$

Thus Φ_K is a bounded linear operator with $\|\Phi_K\| = 1$.

□

B.2 Proof of Theorem 4.1

Proof of Theorem 4.1. Recall that, by Theorem B.1, for each fixed $\tilde{p}$, the map

$$\Phi_K(\cdot, \tilde{p}) : C(\mathbb{R}) \longrightarrow \mathbb{R}^{K+1}$$

is linear. To exhibit its Riesz representation, it suffices to find, for each $n = 0, 1, \ldots, K$, a function $f_n(y) \in L^2(\mathbb{R})$ such that

$$[\Phi_K(p, \tilde{p})]_n \;=\; \mathbb{Q}_K(n) \;=\; \int_{\mathbb{R}} f_n(y)\, p(y)\, \mathrm{d}y \;=\; \langle f_n,\, p\rangle_{L^2(\mathbb{R})}.$$

However, by definition of Φ_K,

$$\mathbb{Q}_K(n) \;=\; \binom{K}{n} \int_{\mathbb{R}} \big[\tilde{F}(y)\big]^n \big[1 - \tilde{F}(y)\big]^{K-n} p(y)\, \mathrm{d}y = \int_{\mathbb{R}} \Big(b_{n,K} \circ \tilde{F}\Big)(y)\, p(y)\, \mathrm{d}y,$$

which is exactly $\langle\, b_{n,K} \circ \tilde{F},\; p\rangle$ for the Bernstein polynomial of degree K. Therefore

$$\Phi_K(p, \tilde{p}) \;=\; \big(\langle f_0, p\rangle,\, \langle f_1, p\rangle,\, \ldots,\, \langle f_K, p\rangle\big),$$

and the Riesz theorem yields the claimed representation, with $\{f_n\}_{n=0}^K$ playing the role of the dual elements.

Now assume $\tilde{p}(x) > 0$ for all x. Then $\tilde{F}'(x) = \tilde{p}(x) > 0$. So $\tilde{F}$ is continuous and strictly increasing. Hence $\tilde{F}$ is a bijection and admits the inverse $\tilde{F}^{-1}$.

Setting $q(x) = p(x)/\tilde{p}(x)$ and changing variables $t = \tilde{F}(x)$ (so $x = \tilde{F}^{-1}(t)$ and $\mathrm{d}t = \tilde{p}(\tilde{F}^{-1}(t))\, \mathrm{d}x$) yields

$$\mathbb{Q}_K(n) = \int_{\mathbb{R}} b_{n,K}\big(\tilde{F}(x)\big)\, p(x)\, \mathrm{d}x = \int_0^1 b_{n,K}(t)\, q\big(\tilde{F}^{-1}(t)\big)\, \mathrm{d}t = \big\langle b_{n,K},\, q \circ \tilde{F}^{-1}\big\rangle_{L^2([0,1])}.$$

This completes the proof. □

B.3 Proof of Theorem 4.2

Proof of Theorem 4.2. **Existence of the Bernstein expansion.** Since $p, \tilde{p} \in C(\mathbb{R})$ and $\tilde{p}(x) > 0$ for all x, the ratio $q(x) \;=\; \frac{p(x)}{\tilde{p}(x)}$ is well-defined and continuous on $\mathbb{R}$. Moreover, the cdf $\tilde{F}$ is C^1 with $\tilde{F}'(x) = \tilde{p}(x) > 0$; hence $\tilde{F}$ is strictly increasing and continuous. It follows that $\tilde{F}^{-1}$ exists and is continuous . Therefore, setting

$$\tilde{q}(t) \;=\; q\big(\tilde{F}^{-1}(t)\big),$$

we obtain $\tilde{q} \in C([0, 1])$.

Since $\{b_{n,K}\}_{n\geq 0}$ is a Schauder basis of $C([0, 1])$ (see (Lorentz, 2012, Theorem 1.1.1) and (Megginson, 2012, Chapter 4)), there are unique coefficients $\{\alpha_n\}_{n\geq 0}$ such that

$$\tilde{q}(t) = \sum_{n=0}^{\infty} \alpha_n\, b_{n,n}(t), \qquad t \in [0, 1].$$

Truncating at degree K yields

$$\tilde{q}_K(t) := \sum_{n=0}^{K} \alpha_n\, b_{n,K}(t).$$

Dual-basis (projection) representation. Let

$$V_K = \mathrm{span}\{b_{0,K}, \ldots, b_{K,K}\} \subset L^2([0, 1]),$$

and define the Gram matrix G with entries

$$G_{n,m} = \big\langle b_{n,K},\, b_{m,K}\big\rangle = \int_0^1 b_{n,K}(t)\, b_{m,K}(t)\, \mathrm{d}t, \qquad n, m \in \{0, \ldots, K\},$$

which is nonsingular (positive-definite). The dual (biorthogonal) basis $\{\tilde{b}_{n,K}\} \subset V_K$ is given by (see (Jiittler, 1998, Section 2))

$$\tilde{b}_{n,K}(t) = \sum_{m=0}^{K} (G^{-1})_{n,m}\, b_{m,K}(t),$$

so that $\langle \tilde{b}_{n,K}, b_{m,K} \rangle = \delta_{n,m}$.

Let $P_K : L^2([0,1]) \to V_K$ the orthogonal projection constructed as $P_K(f) \in V_K$ that satisfies $f - P_K(f) \perp V_K$. We write

$$P_K(f) = \sum_{n=0}^{K} c_n\, \tilde{b}_{n,K},$$

and impose for each m,

$$0 = \big\langle f - P_K(f),\, b_{m,K} \big\rangle = \big\langle f,\, b_{m,K} \big\rangle - \sum_{n=0}^{K} c_n \big\langle \tilde{b}_{n,K},\, b_{m,K} \big\rangle = \big\langle f,\, b_{m,K} \big\rangle - c_m,$$

where $c_m = \langle f, b_{m,K} \rangle$. Thus

$$P_K(f) = \sum_{n=0}^{K} \big\langle f, b_{n,K} \big\rangle_{L^2([0,1])}\, \tilde{b}_{n,K}.$$

Applying this to $f = \tilde{q}$ gives

$$\tilde{q}_K = P_K(\tilde{q}) = \sum_{n=0}^{K} \big\langle \tilde{q},\, b_{n,K} \big\rangle_{L^2([0,1])}\, \tilde{b}_{n,K}(t).$$

Identification of the coefficients. Since $\tilde{q}(t) = q(\tilde{F}^{-1}(t))$ and $q(x) = p(x)/\tilde{p}(x)$, the change of variables $t = \tilde{F}(x)$ yields

$$\begin{aligned} \big\langle \tilde{q},\, b_{n,K} \big\rangle = \int_0^1 b_{n,K}(t)\, \tilde{q}(t)\, \mathrm{d}t &= \int_{\mathbb{R}} b_{n,K}\big(\tilde{F}(x)\big)\, q(x)\, \tilde{p}(x)\, \mathrm{d}x \\ &= \int_{\mathbb{R}} b_{n,K}\big(\tilde{F}(x)\big)\, p(x)\, \mathrm{d}x = \mathbb{Q}_K(n). \end{aligned}$$

Hence

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

□

B.4 Proof Theorem 4.3

In this section we derive a uniform bound on the Bernstein–truncated ratio q_K. The main tool is the dual-basis expansion of $\tilde{q}_K$ and the fact (Lemma B.1) that the dual Bernstein functions sum to $K+1$.

Lemma B.1. *The dual Bernstein basis functions $\tilde{b}_{m,K}(x)$ satisfy*

$$\sum_{m=0}^{K} \tilde{b}_{m,K}(x) = K + 1.$$

Proof. We begin by showing that the constant vector

$$e = (1, 1, \ldots, 1)^{\top}$$

is an eigenvector of the Gram matrix G with eigenvalue $\lambda = 1/(K+1)$. Writing out the nth component of $G\,e$ yields

$$(G\,e)_n = \sum_{m=0}^{K} G_{nm}\, e_m = \sum_{m=0}^{K} \int_0^1 b_{n,K}(x)\, b_{m,K}(x)\, \mathrm{d}x = \int_0^1 b_{n,K}(x) \Big(\sum_{m=0}^{K} b_{m,K}(x) \Big) \mathrm{d}x.$$

Since the Bernstein polynomials form a partition of unity, $\sum_{m=0}^{K} b_{m,K}(x) = 1$ (by the Binomial theorem), and each integrates to $1/(K+1)$ (indeed, $\int_0^1 b_{n,K}(x)\,\mathrm{d}x = \binom{K}{n} B(n+1, K-n+1) = \frac{1}{K+1}$ by the Beta-function identity), we obtain

$$(G\,e)_n = \int_0^1 b_{n,K}(x)\,\mathrm{d}x = \frac{1}{K+1}.$$

Hence

$$G\,e = \frac{1}{K+1}\,e, \tag{15}$$

i.e. e is an eigenvector with eigenvalue $1/(K+1)$.

Left-multiplying by G^{-1} (positive-definite matrix) on both sides of Equation 15 yields

$$G^{-1}e = (K+1)\,e.$$

Which implies that

$$(G^{-1}e)_n = \sum_{m=0}^{K} (G^{-1})_{n,m} = K+1 \quad \text{for each } n.$$

Finally, by the definition of the dual basis,

$$\begin{aligned}
\sum_{m=0}^{K} \tilde{b}_{m,K}(x) &= \sum_{m=0}^{K}\sum_{n=0}^{K} (G^{-1})_{n,m}\, b_{n,K}(x) \\
&= \sum_{n=0}^{K} b_{n,K}(x) \left(\sum_{m=0}^{K} (G^{-1})_{nm}\right) \\
&= (K+1)\sum_{n=0}^{K} b_{n,K}(x) \\
&= K+1.
\end{aligned}$$

where we again are used the partition-of-unity property $\sum_{n=0}^{K} b_{n,K}(x) = 1$. □

Proof of Theorem 4.3. We have

$$\tilde{q}_K(t) = \sum_{m=0}^{K} \mathbb{Q}_K(m)\tilde{b}_{m,K}(t) = \sum_{m=0}^{K} \left[\frac{1}{K+1} - \left(\frac{1}{K+1} - \mathbb{Q}_K(m)\right)\right] \tilde{b}_{m,K}(t).$$

Since $\sum_{m=0}^{K} \tilde{b}_{m,K}(t) = K+1$ (see Lemma B.1), we rewrite this explicitly as

$$\tilde{q}_K(t) = 1 - \sum_{m=0}^{K} \left(\frac{1}{K+1} - \mathbb{Q}_K(m)\right) \tilde{b}_{m,K}(t).$$

By the triangle inequality, we have

$$|\tilde{q}_K(t) - 1| \le \sum_{m=0}^{K} \left|\frac{1}{K+1} - \mathbb{Q}_K(m)\right| \; |\tilde{b}_{m,K}(t)| \le (K+1) \sum_{m=0}^{K} \left|\frac{1}{K+1} - \mathbb{Q}_K(m)\right|,$$

where the inequality $|\tilde{b}_{m,K}(t)| \le K+1$ follows from Lemma B.1. Recognizing the definition of $d_K(p, \tilde{p})$, we have

$$|\tilde{q}_K(t) - 1| \le (K+1)^2 d_K(p, \tilde{p}).$$

Since $q_K(x) = \tilde{q}_K\big(\tilde{F}(t)\big)$, the supremum over $x \in \mathbb{R}$ coincides with the supremum over $t \in [0,1]$. Hence

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

□

B.5 Proof of Theorem 4.4

Proof of Theorem 4.4. Recall from Theorem 4.2 that

$$\tilde{q}_K(t) = \sum_{m=0}^{K} \mathbb{Q}_K(m)\, \tilde{b}_{m,K}(t) \tag{16}$$

is precisely the degree-K truncation of the Bernstein expansion of $\tilde{q}$. Since $\{b_{n,K}\}_{n\geq 0}$ is a Schauder basis of $\big(C([0,1]), \|\cdot\|_\infty\big)$, these truncations satisfy

$$\|\tilde{q}_K - \tilde{q}\|_\infty = \sup_{t\in[0,1]} |\tilde{q}_K(t) - \tilde{q}(t)| \;\longrightarrow\; 0 \qquad \text{when } \; K \to \infty,$$

i.e. $\tilde{q}_K \to \tilde{q}$ uniformly on $[0,1]$ (see (Lorentz, 2012, Section 1.1)). From Equation. (4) and 16,

$$p_K(x) = \tilde{p}(x)\, \tilde{q}_K\big(\tilde{F}(x)\big).$$

Since $\tilde{p}$ and $\tilde{F}$ are continuous, for each fixed x the argument $t_x = \tilde{F}(x) \in [0,1]$ is constant, and uniform convergence of $\tilde{q}_K$ gives

$$\lim_{K\to\infty} \tilde{q}_K\big(t_x\big) = \tilde{q}\big(t_x\big).$$

Hence

$$\lim_{K\to\infty} p_K(x) = \tilde{p}(x) \lim_{K\to\infty} \tilde{q}_K(\tilde{F}(x)) = \tilde{p}(x)\, \tilde{q}\big(\tilde{F}(x)\big) = \tilde{p}(x)\, \frac{p(x)}{\tilde{p}(x)} = p(x).$$

□

C PROOFS OF THEOREMS SECTION 5

In this subsection we prove Theorem 5.1, which shows that for each projection direction s, the Bernstein-projected density ratio q^s converges uniformly to 1 at the rate controlled by $d_K(p,\tilde{p})$ and the Hessian of q.

Proof of Theorem 5.1. Integrating over $\mathbb{S}^d$ on both the left-hand and right hand sides of expression (3) yields

$$\int_{\mathbb{S}^d} \|q^s - 1\|_\infty \, ds \leq (K+1)^2 \int_{\mathbb{S}^d} d_K(s\#p, s\#\tilde{p})\, ds + \frac{1}{8K} \int_{\mathbb{S}^d} \|(q^s)''\|_\infty \, ds.$$

Since $\tilde{p}(x) > 0$ for any $x \in \mathbb{R}$ then $q \in C^2(\mathbb{R}^d)$. Note that for any unit vector v, the second directional derivative is $D^2_{v,v} q(x) = v^\top \nabla^2 q(x) v$, and hence $|D^2_{v,v} q(x)| \leq \|\nabla^2 q(x)\|$. Taking the supremum over x yields the uniform bound $\|\nabla^2 q\|_\infty$. Therefore, we can further write

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

where $\mathcal{L}(\mathbb{S}^d)$ denotes the Lebesgue measure of $\mathbb{S}^d$ and $\|\nabla^2 q\|_\infty$ denotes the supremum over $[0,1]^d$ of the operator norm of the Hessian of q. Similarly, we obtain

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

□

D SUPPLEMENTARY EXPERIMENTS

D.1 Evaluating dual-ISL on 1D Target Distributions

Following the evaluation setup of Zaheer et al. (2017); de Frutos et al. (2024b), we draw $N = 1000$ i.i.d. samples from each of six one-dimensional benchmark targets. The first three targets are classical pdf (e.g. $\mathcal{N}(0,1)$, Cauchy, and Pareto), while the remaining three are equally-weighted mixtures:

- **Model$_1$:** $\frac{1}{2}\mathcal{N}(5,2) + \frac{1}{2}\mathcal{N}(-1,1)$.
- **Model$_2$:** $\frac{1}{3}\mathcal{N}(5,2) + \frac{1}{3}\mathcal{N}(-1,1) + \frac{1}{3}\mathcal{N}(-10,3)$.
- **Model$_3$:** $\frac{1}{2}\mathcal{N}(-5,2) + \frac{1}{2}\text{Pareto}(5,1)$.

All non-diffusion methods (Dual-ISL, ISL, GAN, WGAN and MMD-GAN) use the same generator architecture: a four-layer MLP with ELU activations and layer widths [1, 7, 13, 13, 7, 1]. Each is trained for 10^4 epochs using Adam with a fixed learning rate of 10^{-2}. By contrast, the DDPM baseline employs a four-layer ELU-MLP score network with identical widths, augmented by a 16-dimensional sinusoidal time embedding. It is also trained for 10^4 epochs with Adam (lr $= 10^{-2}$) across $T = 200$ diffusion steps, where the noise schedule $\{\beta_t\}$ is linearly spaced from 10^{-4} to 2×10^{-2}. Table 1 summarizes the quantitative results, and the corresponding visualizations are shown in Figure 3.

Referring to (Zaheer et al., 2017, Theorem 1), in the one-dimensional case any generator that perfectly pushes forward a simple base distribution p_z (e.g. uniform or Gaussian) to a target distribution p must implement one of at most two continuous maps. Concretely, if F_z and F are the cdfs of p_z and p, then the two solutions are

$$f_+(z) \;=\; F^{-1}\big(F_z(z)\big) \qquad \text{and} \qquad f_-(z) \;=\; F^{-1}\big(1 - F_z(z)\big).$$

In practice, a learning algorithm that truly captures the underlying structure of a multimodal or heavy-tailed distribution should recover one of these two "probability-integral-transform" maps.

In Figure 4, we therefore plot, for each method, the learned generator $f_\theta(z)$ against the theoretical target map $f_+(z)$ for the challenging mixture $\frac{1}{2}\mathcal{N}(-5,2) + \frac{1}{2}\text{Pareto}(5,1)$ ("Model$_3$"). A perfect fit would lie exactly on the diagonal. As seen, dual-ISL (Subfigure 4a) closely tracks the true transformation across the full support of z, whereas classical ISL (Subfigure 4b) suffers from local distortions around the Pareto tail, and MMD-GAN (Subfigure 4c) exhibits even larger deviations, especially where the two modes meet. This visualization makes explicit how dual-ISL more faithfully learns the correct mapping, rather than merely matching moments or densities.

D.2 Computational Benchmarking of Dual-ISL vs (Classical) ISL

Next, we measure execution time using dedicated benchmarking tools Chen and Revels (2016). These tools first warm up and calibrate the code to determine the optimal number of iterations per measurement, then execute the code in bundled loops to collect multiple independent samples. We compute statistics, including the minimum, median, mean, and standard deviation, while tracking garbage collection time separately. Table 4 summarizes the results for various target distributions at a fixed K. Figure 5 combines two perspectives: (a) total runtime as a function of K, and (b) estimation accuracy versus runtime at a fixed K. These findings show that Dual-ISL not only runs faster than classical ISL, but also achieves a superior accuracy–runtime balance, and this advantage grows even larger as K increases.

Target	**ISL**			**Dual-ISL**		
	Median	**Mean ± σ**	**Memory**	**Median**	**Mean ± σ**	**Memory**
$\mathcal{N}(4,2)$	239.374	241.280(434)	17.42 GiB	22.281	22.502(118)	9.27 GiB
Pareto(1, 1)	238.043	239.007(1607)	17.42 GiB	22.207	22.109(518)	9.27 GiB
Model$_1$	241.683	241.905(1722)	17.42 GiB	21.685	21.709(44)	9.27 GiB
Model$_2$	237.825	239.805(1832)	17.42 GiB	22.921	22.886(453)	9.27 GiB

Table 4: Runtime (in seconds) and memory benchmarks for ISL vs. Dual-ISL at fixed $K = 10$. Each cell reports median runtime, mean ± standard deviation, and peak memory usage.

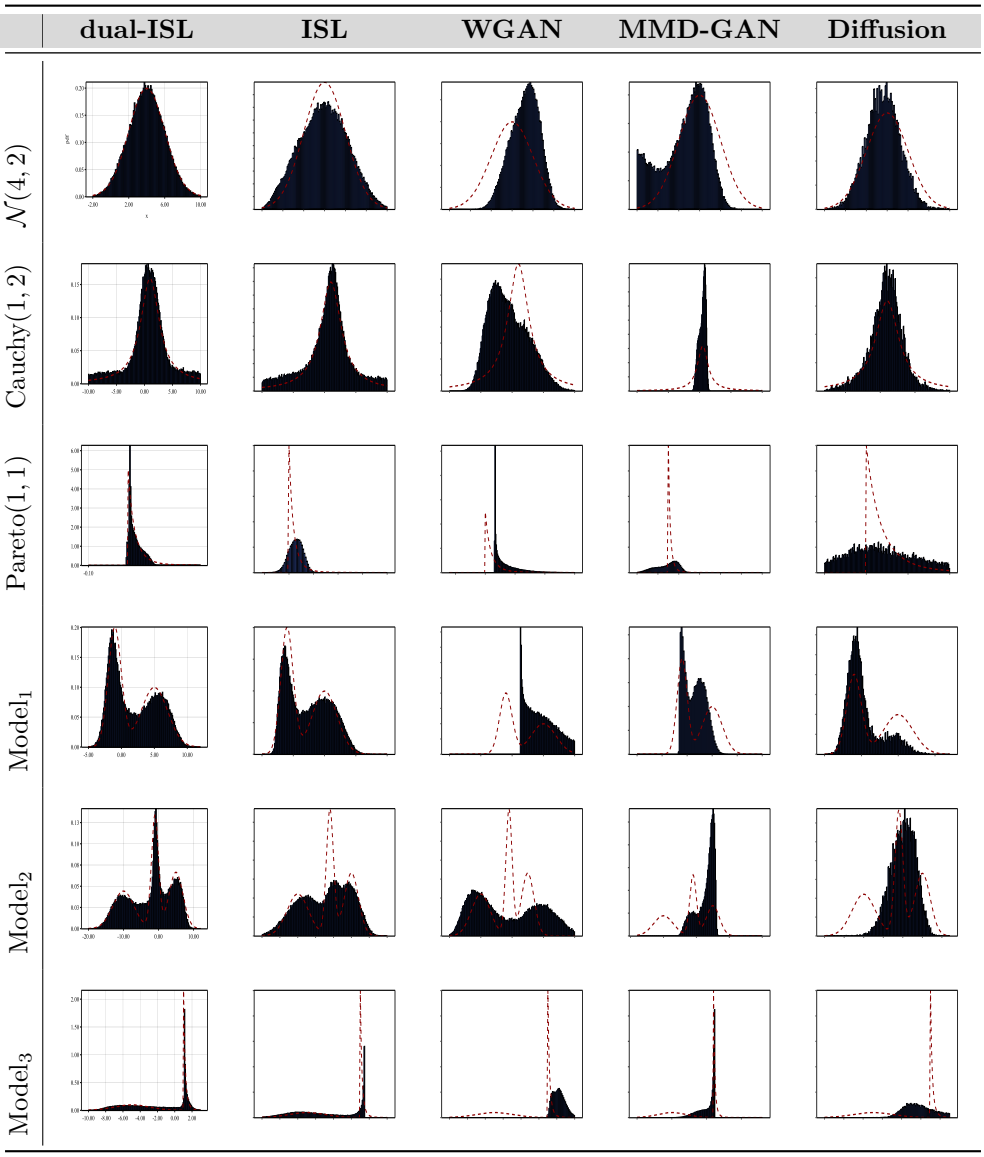
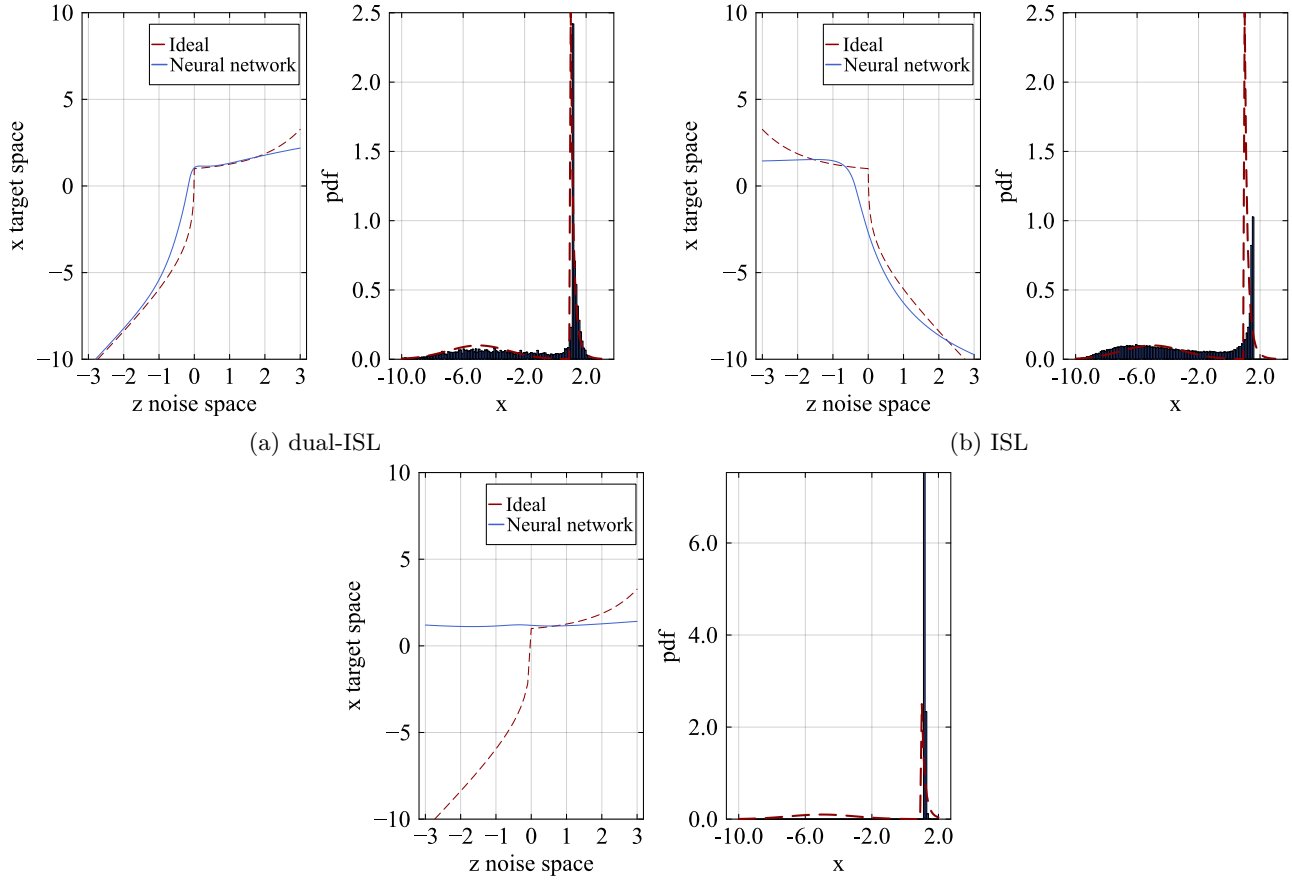


Figure 3: One-dimensional density estimation across six benchmark targets. Each row corresponds to a different true distribution (top to bottom: $\mathcal{N}(4,2)$, Cauchy$(1,2)$, Pareto$(1,1)$, Model$_1$, Model$_2$, Model$_3$). In each subplot, the red curve shows the ground-truth density and the blue curve shows the model's estimated density. Columns (left to right) compare dual-ISL, classical ISL, WGAN, MMD-GAN, and a DDPM diffusion baseline, respectively. Dual-ISL more accurately captures multi-modal and heavy-tailed shapes, with reduced mode-collapse and smoother estimates.

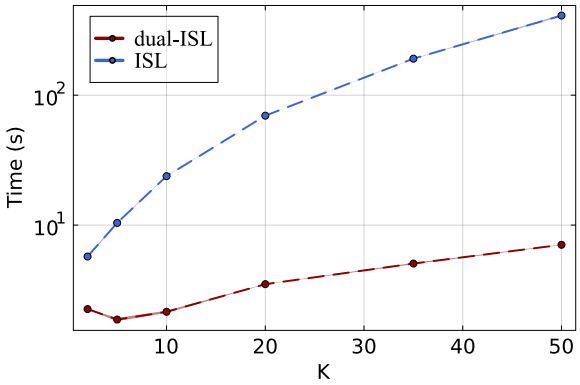


(a) dual-ISL

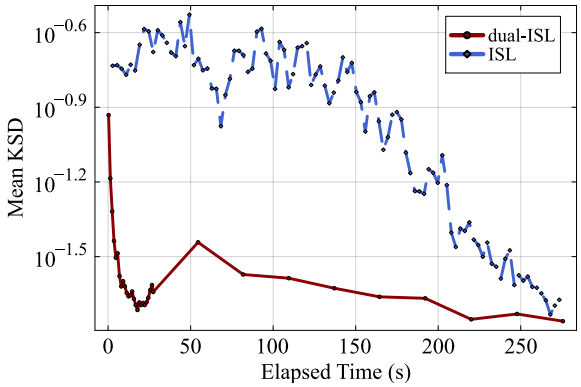
(b) ISL

(c) MMD-GAN

Figure 4: Comparison of learned generator mappings $f_\theta(z)$ against the true probability-integral-transform $f_+(z)$ or $f_-(z)$ for Model$_3$. Dual-ISL closely follows the ideal map even in the heavy-tailed region, while ISL and MMD-GAN display growing errors, particularly near the mode boundaries and in the Pareto tail.

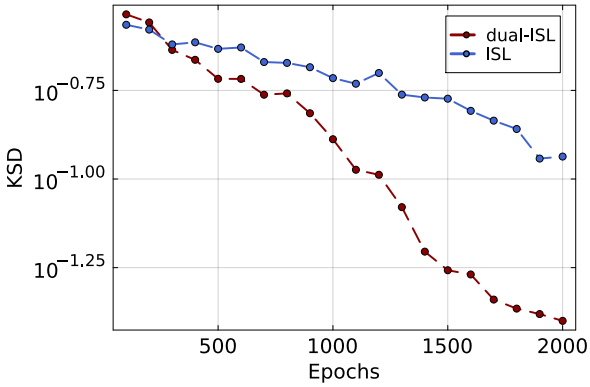


(a) Target density Model$_2$

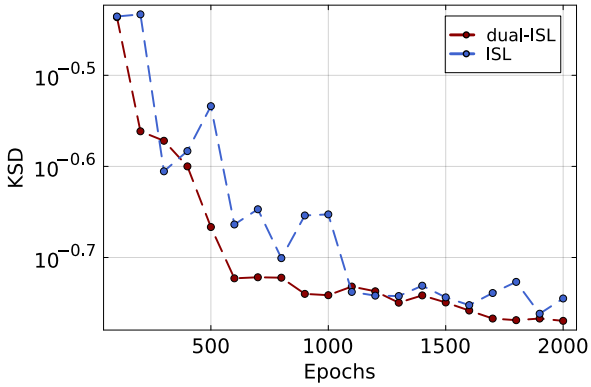


(b) Target density Model$_2$

Figure 5: Computation time and accuracy trade-off for classical ISL versus Dual-ISL as K increases. (a) Total runtime (in seconds) for 1000 training epochs with batch size $N = 1000$ on the Model$_2$ target. (b) Runtime versus mean KSD for both methods with $K = 10$ and batch size $N = 1000$, illustrating how Dual-ISL maintains lower runtimes improving also accuracy. Dual-ISL consistently outperforms classical ISL in speed, with the gap widening at larger K.



(a) $\mathcal{N}(4, 2)$



(b) Model$_3$

Figure 6: Training curves for dual-ISL versus classical ISL (dashed lines indicate mean over 10 runs). Left: target $\mathcal{N}(4, 2)$. Right: target mixture Model$_3$ (mixture Pareto–Normal).

D.3 Moment-Agnostic Optimal Transport via Monotonicity-Penalized ISL

Our ISL framework applies to *any* probability law, including heavy-tailed distributions lacking finite higher-order moments. Indeed, the rank statistic in Eq. 1 is always well-defined, whereas the classical p–Wasserstein distance is only finite when both distributions possess finite p-th order moments.

Unique transport in one dimension. In one dimension, any continuous map that pushes a simple base law p_z onto a target p must coincide with one of two inverses of the base cdf. If F_z and F denote the cdfs of p_z and p, then

$$f_+(z) = F^{-1}\big(F_z(z)\big), \quad f_-(z) = F^{-1}\big(1 - F_z(z)\big).$$

The monotone map f_+ is in fact the unique optimal transport in $\mathbb{R}$. To recover this map, we augment our ISL loss with a *monotonicity penalty*.

Monotonicity-constrained training. Given a batch of inputs $\{x_i\}_{i=1}^N$ sorted as

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(N)}, \qquad f_\theta(x_{(i)}) \;=\; \text{model output at } x_{(i)},$$

we define

$$\text{Penalty} \;=\; \frac{1}{N}\sum_{i=1}^{N-1} \max\big\{0,\; f_\theta(x_{(i)}) - f_\theta(x_{(i+1)})\big\},$$

which is zero if and only if f_θ is non-decreasing. The overall training objective becomes

$$\mathcal{L}(p,\tilde{p}) \;=\; d_K(p,\tilde{p}) \;+\; \lambda\,\frac{1}{N}\sum_{i=1}^{N-1}\max\{0,\; f_\theta(x_{(i)}) - f_\theta(x_{(i+1)})\},$$

where $d_K(p,\tilde{p})$ is our rank-based discrepancy and $\lambda > 0$ weights the monotonicity constraint. As $\lambda \to \infty$, any violation of monotonicity incurs infinite cost, forcing f_θ to converge to the unique optimal transport map f_+.

This loss not only recovers the optimal transport in one dimension under minimal smoothness, but also extends beyond the Wasserstein framework to handle distributions with heavy tails.

We evaluate three training objectives, Dual-ISL with monotonicity penalty, 1-Wasserstein, and 2-Wasserstein, on a suite of heavy-tailed target distributions using a five-layer MLP with ELU activations and layer widths $[16, 16, 32, 32, 16, 1]$. Each model is trained for 10^4 epochs via vanilla gradient descent with a fixed learning rate of 10^{-2}; we avoid adaptive optimizers to ensure that observed differences stem solely from the loss functions.

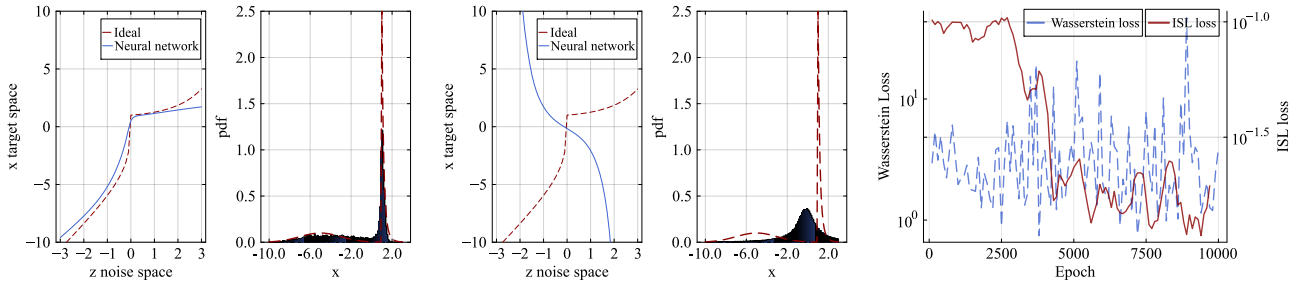
Performance is measured by two complementary metrics.

1. **Kolmogorov–Smirnov distance (KSD):** the maximum absolute deviation between the empirical cdfs of the real pdf p and that of the generated distribution $\tilde{p}$.
2. **Tail-fit error** A_{CCDF}: the area between the log–log complementary cdfs of real and generated samples, defined for n data points by

$$A_{\mathrm{CCDF}} = \sum_{i=1}^{n}\Big[\log\big(F_p^{-1}(i/n)\big) - \log\big(\tilde{F}_{\tilde{p}}^{-1}(i/n)\big)\Big]\,\log\big(\tfrac{i+1}{i}\big),$$

where F_p^{-1} and $\tilde{F}_{\tilde{p}}^{-1}$ are the inverse empirical ccdfs of p and $\tilde{p}$, respectively.

Figure 7 compares the Dual-ISL loss (with monotonicity penalty) against the 1-Wasserstein loss on a Pareto–Normal mixture. Figure (a) shows that Dual-ISL accurately recovers the true transport map despite the heavy Pareto tails. Figure (b) demonstrates that the 1-Wasserstein loss fails to learn the correct mapping under heavy-tailed behavior. Figure (c) plots both losses over 10 000 training epochs: the Wasserstein loss oscillates and does not converge, whereas the Dual-ISL loss decreases smoothly and reliably, highlighting its stability and robustness.



(a) Dual-ISL **(b)** Wasserstein **(c)** Optimal-Transport

Figure 7: Comparison of transport objectives on a Pareto–Normal mixture: (a) Dual-ISL with monotonicity penalty, (b) 1-Wasserstein, and (c) Training dynamics over 10000 epochs, plotting Dual-ISL loss (solid) and 1-Wasserstein loss (dashed). The Dual-ISL curve decreases smoothly and converges reliably, whereas the Wasserstein loss oscillates and does not settle.

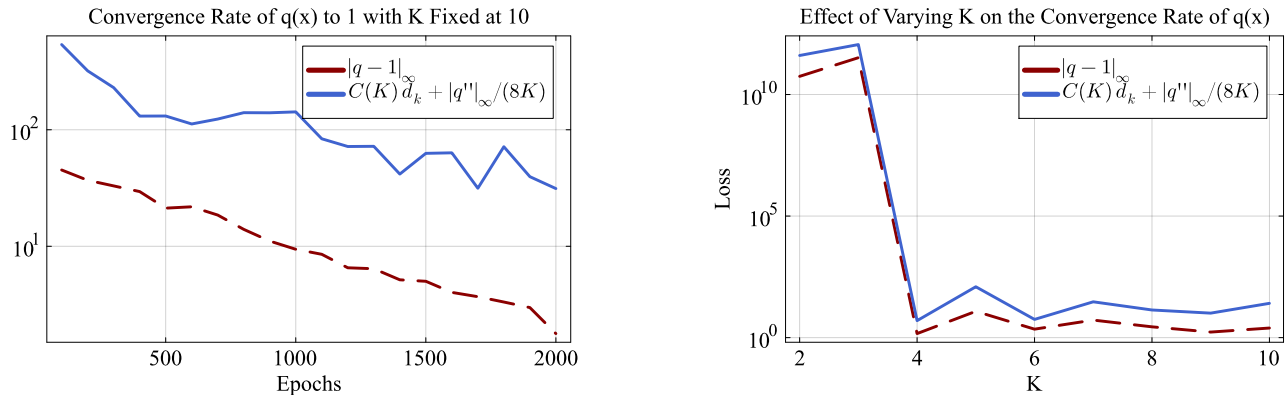
Table 5 shows that Dual-ISL (with monotonicity penalty) not only matches but frequently outperforms the classical OT baselines (1-Wasserstein and 2-Wasserstein) for $K = 10$. For the moderately heavy-tailed Cauchy$(5, 10)$, 1-Wasserstein attains the lowest KSD, yet Dual-ISL achieves a smaller A_{CCDF}, indicating superior tail alignment. As the tail heaviness grows in Cauchy$(5, 20)$ and Pareto$(1, 1)$, Dual-ISL outperforms both OT metrics on both measures, note that 2-Wasserstein is undefined for Pareto due to its infinite second moment, underscoring Dual-ISL's robustness where classical OT either diverges or loses precision. Finally, on the multimodal Model$_3$, Dual-ISL yields the lowest KSD and A_{CCDF}, demonstrating that its rank-statistic formulation reliably recovers the unique monotone optimal-transport map in one dimension while avoiding the instability and non-differentiability of conventional OT losses.

Target	Dual-ISL		1-Wasserstein		2-Wasserstein	
	KSD	ACCDF	KSD	ACCDF	KSD	ACCDF
Cauchy$(5, 10)$	0.069 ± 0.059	$\mathbf{19.401 \pm 0.539}$	$\mathbf{0.037 \pm 0.024}$	21.437 ± 0.429	0.504 ± 0.267	28.678 ± 7.018
Cauchy$(5, 20)$	$\mathbf{0.045 \pm 0.010}$	$\mathbf{17.201 \pm 0.423}$	0.046 ± 0.014	21.123 ± 1.290	0.668 ± 0.168	42.398 ± 11.578
Pareto$(1, 1)$	$\mathbf{0.120 \pm 0.053}$	$\mathbf{21.769 \pm 0.072}$	0.240 ± 0.064	23.676 ± 0.072	0.950 ± 0.158	–
Model$_3$	$\mathbf{0.131 \pm 0.017}$	$\mathbf{19.406 \pm 1.303}$	0.148 ± 0.097	31.561 ± 2.769	0.219 ± 0.046	22.031 ± 1.387

Table 5: KSD and ACCDF (mean ± std) for Dual-ISL (with monotonicity penalty) vs. Wasserstein baselines at $K = 10$.

D.4 Empirical proof of the convergence rate

Our aim is to empirically validate Equation 3. To do this, we train a neural network using the same architecture as in our previous experiment. The network is fed an input pdf $\mathcal{N}(0, 1)$ and is tasked with approximating a target pdf defined as a mixture of Cauchy distributions. We estimate $\tilde{p}$ via a kernel density estimator and compute the second derivative of q using central finite differences with a sixth-order expansion. Each experiment is repeated 10 times, and the mean results are shown in Figures 8.



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

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

Figure 8: Empirical convergence of dual-ISL's Bernstein approximation (cf. Eq. 3). The solid blue curve shows the mean theoretical upper bound $\|q_K - 1\|_\infty \leq (K + 1)^3 d_K$, and the dashed red curve shows the observed $\|q - 1\|_\infty$.

D.5 Density estimation

D.5.1 1D density estimation

We employ the same fully-connected NN and training hyperparameters as in Appendix D.1. Once training converges, we approximate the implicit density with Equation 5. We restate the latter here for convenience

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

where

$$\mathbb{Q}_K(m) \;=\; \mathbb{P}\big(A_K = m\big), \quad b_{m,K}(u) = \binom{K}{m} u^m (1-u)^{K-m}\,,$$

and $\{\tilde{b}_{m,K}\}$ is the dual of the Bernstein polynomial basis.

The computation proceeds in three steps

1. **Monte Carlo estimation of weights**
   For each $m = 0, \ldots, K$, estimate
   $$\mathbb{Q}_K(m) = \mathbb{P}\big(\#\{\, x_i \leq x\} = m\big),$$

by sampling K independent latent vectors $z_i \sim \mathcal{N}(0, I)$, computing $x_i = f_\theta(z_i)$, and counting how many satisfy $x_i \leq x$. Repeat M times and take empirical frequencies:

$$\widehat{\mathbb{Q}}_K(m) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\Big(\#\{x_i^{(j)} \leq x\} = m\Big).$$

2. **Empirical cdf estimation**
   Draw N samples $\{x_i\}_{i=1}^N$ from the trained generator, sort them in ascending order, and form the empirical cdf

$$\widehat{\tilde{F}}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{x_i \leq x\}.$$

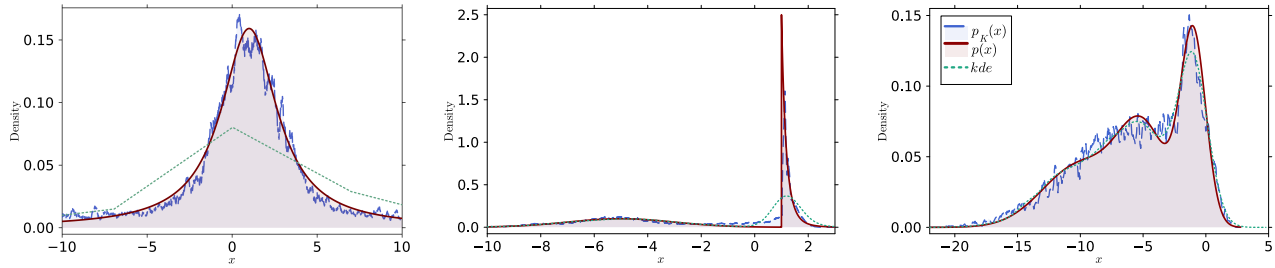
3. **Finite-difference density**
   Approximate the density of the push-forward distribution by a first-order finite difference:

$$\widehat{\tilde{p}}(x) \approx \frac{\widehat{\tilde{F}}(x + \Delta) - \widehat{\tilde{F}}(x)}{\Delta}, \qquad \Delta \ll 1.$$

We evaluate Dual-ISL on six univariate target distributions. For each target, we estimate the mixture weights $\widehat{\mathbb{Q}}_K$ via 10^5 Monte Carlo trials, compute the empirical cdf $\widehat{\tilde{F}}(x)$ from 10^5 samples drawn from the trained generator, and form the density estimate using a first-order finite difference with $\Delta = 0.1$. Table 6 reports the average Kolmogorov–Smirnov distance (over 10 independent runs) for the estimated density using Equation 5 with $K \in \{2, 5, 10\}$ versus a Gaussian kernel density estimator using the same number of samples and Silverman's rule for bandwidth selection. Figure 9 illustrates density estimates for three representative targets, true density (solid red), ISL estimate (dashed blue), and KDE (dotted green), all plotted on common axes to facilitate direct comparison of bias and tail behavior. As shown, Dual-ISL achieves the best results for every target except the Gaussian case, in which KDE with a Gaussian kernel performs marginally better.

Target	**Dual-ISL (K=2)**	**Dual-ISL (K=5)**	**Dual-ISL (K=10)**	**KDE**
$\mathcal{N}(4, 2)$	0.0202	0.0178	0.0167	**0.0110**
Cauchy$(1, 2)$	0.0237	0.0253	**0.0184**	0.2013
Pareto$(1, 1)$	0.0302	**0.0203**	0.0252	0.3872
Mixture$_1$	0.0395	**0.0095**	0.0120	0.0156
Mixture$_2$	0.0171	0.0167	**0.0070**	0.0145
Mixture$_3$	0.1853	0.1786	**0.0741**	0.1644

Table 6: Mean Kolmogorov–Smirnov distance (over 10 runs) for Dual-ISL versus Gaussian KDE. We used $\widehat{\mathbb{Q}}_K$ estimated with 10,000 trials, the empirical CDF $\widehat{\tilde{F}}$ from 10,000 samples, and a finite-difference step $\Delta = 0.1$.



(a) Cauchy **(b)** Model$_2$ **(c)** Model$_3$

Figure 9: Density estimates for three univariate distributions using the ISL estimator (blue dashed line) and Gaussian kernel density estimator (KDE; green dotted line). Figure (a) shows the true Cauchy target density, while Figures (b) and (c) display the corresponding ISL and KDE estimates for Model$_2$ and Model$_3$, respectively. All panels share identical axes to facilitate direct comparison of estimator bias and tail behavior.

D.5.2 1D density correction for normalizing flows

As an additional illustrative experiment, using the same benchmark targets and training setup as in Appendix D.1, we consider a *normalizing flow* trained with the Dual-ISL objective. In particular, we use a flow with four planar-flow layers. Unlike the implicit generator considered in Appendix D.5.1, the flow provides direct access to its density $\tilde{p}_\theta(x)$ at any evaluation point x.

This setting allows us to isolate the effect of the Bernstein correction from the numerical finite-difference approximation of $\widehat{\tilde{p}}(x)$ used in Equation (5). Specifically, we define the corrected density by replacing the Monte Carlo / finite-difference estimate of the push-forward density in Equation (5) with the exact flow density $\tilde{p}_\theta(x)$:

$$p_K^{\mathrm{corr}}(x) \;=\; \tilde{p}_\theta(x) \sum_{m=0}^{K} \widehat{\mathbb{Q}}_K(m)\, \tilde{b}_{m,K}\big(\widehat{\tilde{F}}(x)\big),$$

where $\widehat{\mathbb{Q}}_K(m)$ and $\widehat{\tilde{F}}(x)$ are computed as in Appendix D.5.1.

We compare the *raw* flow density $\tilde{p}_\theta$ and the *Bernstein-corrected* density p_K^{corr} using the kernel Stein discrepancy (KSD) with respect to the true target distribution. Table 7 reports the resulting KSD values on five representative targets.

Model	KSD (raw flow density $\tilde{p}_\theta$)	KSD (Bernstein + flow density)
$\mathcal{N}(4,2)$	1.20×10^{-2}	1.19×10^{-2}
Cauchy$(1,2)$	4.24×10^{-2}	3.43×10^{-2}
Model$_1$	3.52×10^{-2}	3.26×10^{-2}
Model$_2$	4.71×10^{-2}	3.17×10^{-2}
Model$_3$	2.15×10^{-1}	1.91×10^{-1}

Table 7: Kernel Stein discrepancy (KSD) of the raw normalizing-flow density and the Bernstein-corrected density obtained by multiplying the exact flow density by the Dual-ISL Bernstein correction factor.

The Bernstein-corrected density improves upon the raw flow density on all targets, with the largest gains observed for the heavier-tailed and multimodal cases. In addition, the corrected densities are noticeably smoother than the Monte Carlo-based density reconstructions reported in Appendix D.5.1, since the base density factor $\tilde{p}_\theta(x)$ is evaluated exactly in the normalizing-flow setting rather than approximated via finite differences.

D.5.3 2D density estimation

To generalize our implicit estimator to data in $\mathbb{R}^d$, let $\{s_\ell\}_{\ell=1}^m \subset \mathbb{S}^d$ be random unit vectors. Denote by

$$(s\#\hat{p})(y) \;=\; \text{ISL-estimate of the 1D pdf at } y = s^\top x,$$

the push-forward density along s. Then for any query point $x \in \mathbb{R}^d$,

$$\hat{p}(x) = \frac{1}{m} \sum_{\ell=1}^{m} \big(s_\ell\#\hat{p}\big)\big(s_\ell^\top x\big).$$

This Monte Carlo slicing, averaging one-dimensional ISL estimates, yields a consistent multivariate density approximation without ever constructing a full d-dimensional kernel.

In Figure 10, we compare the two-dimensional density estimates produced by our sliced Dual-ISL method against a Gaussian kernel density estimator with Silverman's rule for bandwidth selection. Across all experiments, the generator is a four-layer MLP that maps 2D standard normal noise to the data space, with each hidden layer comprising 32 tanh units. We train for 100 epochs using 1000 samples per epoch, and approximate the sliced discrepancy by averaging over $m = 10$ random projections with $K = 10$. For the KDE baseline, we use 10000 samples to construct each density estimate.

D.6 Experiments on 2D distributions

We begin by evaluating three synthetic 2D benchmarks, each with a distinct topology:

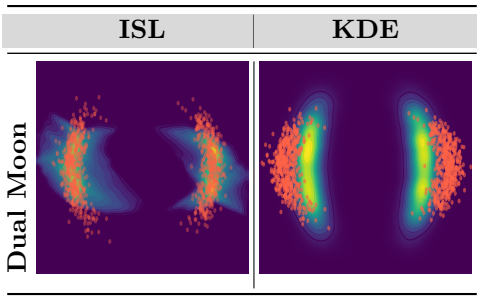


Figure 10: Two-dimensional density estimates on synthetic Dual Moon dataset. Left column: sliced Dual-ISL (averaged over $m = 10$ random projections with $K = 10$); right column: Gaussian KDE using Silverman's bandwidth rule. The generator is a four-layer MLP with 32 units per hidden layer, trained for 100 epochs with 1000 samples per epoch; the KDE baseline uses 10000 samples.

1. **Dual Moon**: A bimodal Gaussian mixture (two disconnected modes).
2. **Circle Gaussian**: An eight-component Gaussian mixture arranged in a circle.
3. **Two Ring**: A "double-ring" distribution consisting of two concentric circular supports.

Our aim is to test whether the sliced dual-ISL method can fully recover these supports, including disconnected or non-convex regions. We compare against both normalizing flows and GANs, measuring performance by

- **KL divergence** between the learned and true densities, and
- **Visual coverage** of the true support via sample plots.

In all experiments—GAN, WGAN, and dual-ISL- the generator was a four-layer MLP that maps two-dimensional standard normal noise into data-space samples, with each hidden layer comprising 32 units and tanh activations. For the GAN and WGAN variants, the discriminator (or critic) adopted a similar four-layer MLP but with 128-unit hidden layers using ReLU activations and a final sigmoid output. Every model was trained for 1 000 epochs with a batch size of 1000 under the Adam optimizer. In the adversarial setups, we swept the critic-to-generator update ratio from 1:1 to 5:1 and chose the learning rate from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. For ISL, we fix $K = 10$, drew $N = 1000$ samples per projection, averaged over $L = 10$ random projections, and used a constant learning rate of 10^{-3}.

For the normalizing flow baseline we adopted the RealNVP architecture of Dinh et al. (2016), consisting of four affine-coupling layers whose scale and translation nets are two-hidden-layer MLPs with 32 ReLU units each. RealNVP was trained under the same 1 000-epoch, batch-size-1 000 protocol, but with a fixed learning rate of 5×10^{-5} as in Stimper et al. (2023).

Figure 11 exposes the fundamental limitations of popular generative models on complex, multimodal data. GANs frequently collapse to a subset of modes, omitting entire regions of the true support. Normalizing flows, by enforcing invertibility, preserve the topology of the base distribution but struggle to represent disconnected clusters, instead "bridging" them with thin density filaments. In contrast, sliced Dual-ISL accurately recovers each connected component of the support, exemplified by the clean separation of the two moons, while still covering the full data manifold. On the Circle-of-Gaussians task, however, a low projection order K can allow leakage between rings; raising K sharpens the estimate and eliminates this spillover. A hybrid strategy ("Dual-ISL + GAN"), in which a model is first trained with Dual-ISL for 100 epochs and then fine-tuned adversarially, combines the stability and full-support coverage of Dual-ISL with the precision of a GAN, yielding the most faithful reconstructions. Table 8 reports KL divergences between each model and the true distribution, confirming that both Dual-ISL and especially the Dual-ISL + GAN variant outperform all baselines.

Figure 12 overlays the true data samples (scatter points) with Gaussian-KDE contours computed from 10^5 points generated by sliced dual-ISL. These contours closely match the underlying support of both the Dual Moon and Two Rings datasets, accurately delineating disconnected modes and concentric rings. This visualization confirms that sliced dual-ISL not only recovers the global topology of complex 2D manifolds but also captures fine geometric details in the learned density.

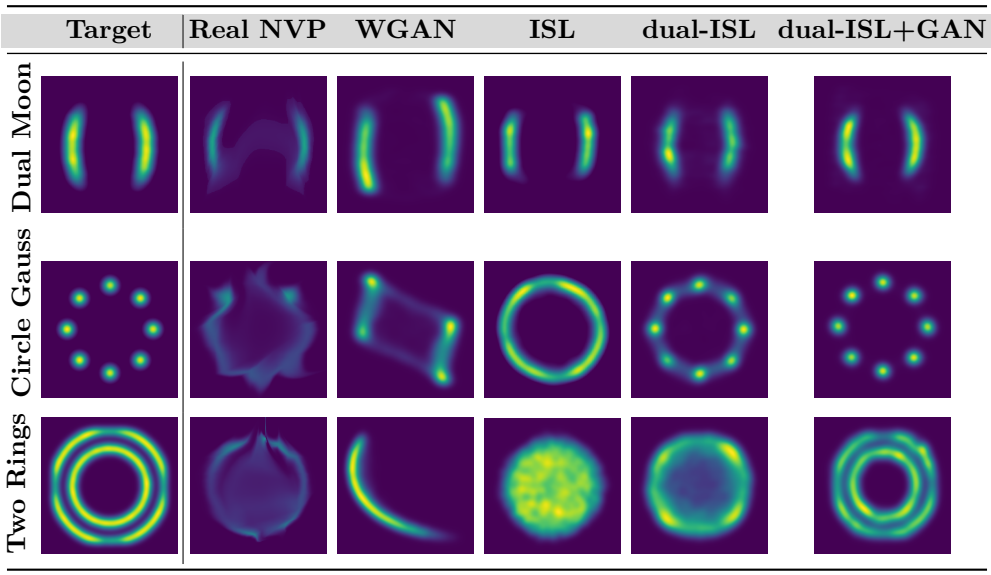
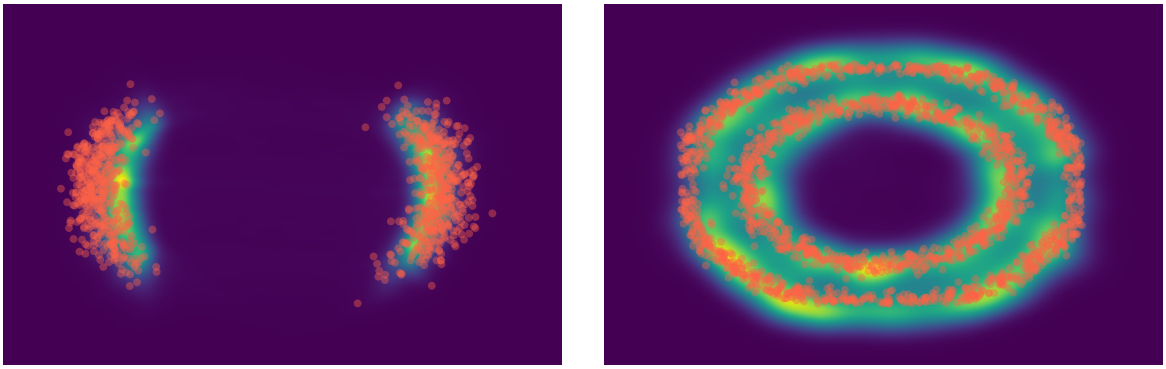


Figure 11: Comparison of various generative methods on three 2D toy distributions. Each row (left) shows the target, and each column the outputs of Real NVP, WGAN, ISL-slicing, Dual-ISL, and dual-ISL(pretrained)+GAN, respectively.

Dataset	Real NVP	GAN	WGAN	ISL	dual-ISL	dual-ISL+GAN
Dual Moon	1.77	1.23	1.02	0.43	0.35	**0.21**
Circle of Gaussians	2.59	2.24	2.38	1.61	0.44	**0.35**
Two Rings	2.69	1.46	2.74	0.56	0.43	**0.29**

Table 8: KL-divergence (lower is better) of different generative models on 2D toy benchmarks.



(a) Dual Moon (b) Two Rings

Figure 12: Scatter plots of the true Dual Moon (left) and Two Rings (right) datasets, overlaid with density contours learned by Sliced dual-ISL and approximated via Gaussian KDE on 10^5 generated samples. Models were trained for 1000 epochs with $L = 20$ random projections and a batch size of $N = 2000$.

D.7 High dimensional experiments

MNIST/Fashion-MNIST In this section, we evaluate the slicing dual-ISL method on high-dimensional image generation. In particular, we incorporate the dual-ISL objective into the training of a DCGAN generator Radford et al. (2015) and benchmark the resulting architecture on the MNIST and Fashion-MNIST datasets. We report precision (a proxy for sample fidelity) and recall (a proxy for sample diversity) to assess the quality and diversity of the generated images Sajjadi et al. (2018). Our models were trained for 40 epochs with a batch size of 128, except for the pretrained models, which were first pretrained with sliced dual-ISL for 20 epochs (using 20 random projections) and then trained as a DCGAN for 40 epochs.

In Table 9, we report our results alongside those of other implicit generative models. On MNIST, our simple ISL-based model achieves recall rates comparable to much more complex, multi-discriminator GANs Durugkar et al. (2016); Choi and Han (2022) (5 discriminator each), despite using far fewer parameters. Furthermore, by pretraining the DCGAN generator with sliced dual-ISL and then fine-tuning under the standard adversarial loss, we attain state-of-the-art precision scores.

On Fashion MNIST, our model matches MCL-GAN in recall and, while we do not quite reach GMAN's recall performance, our precision scores exceed theirs. This demonstrates that even with a simpler architecture, ISL can deliver competitive recall and precision results across diverse image-generation benchmarks.

Dataset	**Method**	$F_{1/8}\uparrow$	$F_8\uparrow$	**Precision**$\uparrow$	**Recall**$\uparrow$
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	93.58 ± 0.84	95.82 ± 1.61	94.03 ± 1.82	96.68 ± 2.42
	GMAN	97.60 ± 0.70	96.81 ± 1.71	97.60 ± 1.82	96.80 ± 2.42
	MCL-GAN	$\mathbf{97.71 \pm 0.19}$	$\mathbf{98.49 \pm 1.57}$	$\mathbf{97.70 \pm 1.33}$	$\mathbf{98.50 \pm 2.15}$
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	92.92 ± 1.23
	DCGAN	86.14 ± 0.11	88.92 ± 1.51	86.60 ± 1.58	88.97 ± 1.33
	dual-ISL with DCGAN	91.43 ± 0.19	91.87 ± 1.57	91.88 ± 1.35	92.42 ± 1.47
	GMAN	90.97 ± 0.09	$\mathbf{95.43 \pm 1.12}$	90.90 ± 1.33	$\mathbf{95.50 \pm 2.25}$
	MCL-GAN	$\mathbf{97.62 \pm 0.09}$	92.97 ± 1.28	$\mathbf{97.70 \pm 1.33}$	92.90 ± 2.31

Table 9: Quantitative comparison of generative models on MNIST and Fashion-MNIST (28×28) using $F_{1/8}$, F_8 (weighted harmonic mean of Precision and Recall, with $\beta = 1/8$ and $\beta = 8$ respectively), Precision, and Recall (mean $\pm$ std) in %. Results are shown for dual-ISL with $m = 20$ and $m = 50$ random projections, standard and Wasserstein DCGAN variants (with and without ISL pretraining), GMAN, and MCL-GAN. Boldface highlights the best score in each column per dataset. Higher is better.

Figure 13 shows the class-frequency distributions obtained by our sliced dual-ISL model (40 epochs, $m = 50$ random projections) versus a standard DCGAN. Our model produces all ten digit classes in nearly uniform proportions, closely matching the true uniform distribution, while the DCGAN exhibits pronounced class imbalance. To compute these frequencies, we generated 10 000 samples from each model and classified them with a pretrained digit recognizer. In a Kolmogorov–Smirnov test for uniformity (on 10 000 samples), the sliced dual-ISL model achieved $p = 0.070$, compared to $p = 0.642$ for DCGAN, indicating a significantly better match to the ideal uniform distribution.

Finally, Figure 14 presents an ablation study of precision and recall in dual-ISL on MNIST, comparing the effects of using 20 versus 100 random projections. Even with only 20 projections, the model already achieves strong performance, and increasing the number of projections yields only marginal gains in these quality metrics.

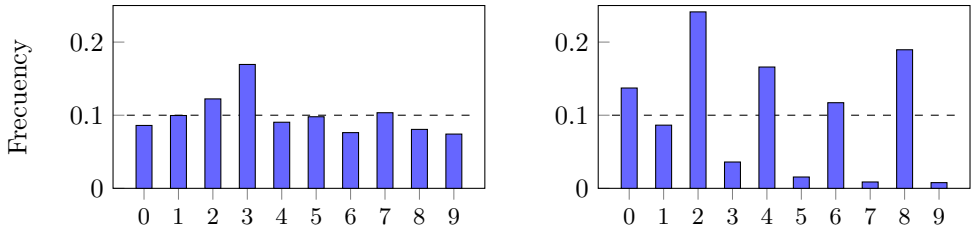


Figure 13: Class-frequency distributions of MNIST samples generated by two different models. Left: dual-ISL; Right: standard DCGAN. Each bar represents the proportion of generated images assigned to each digit class (0–9), illustrating that dual-ISL produces a more uniform coverage across all classes, whereas the DCGAN exhibits notable biases toward certain digits. The dashed line indicates the ideal uniform distribution across all classes.

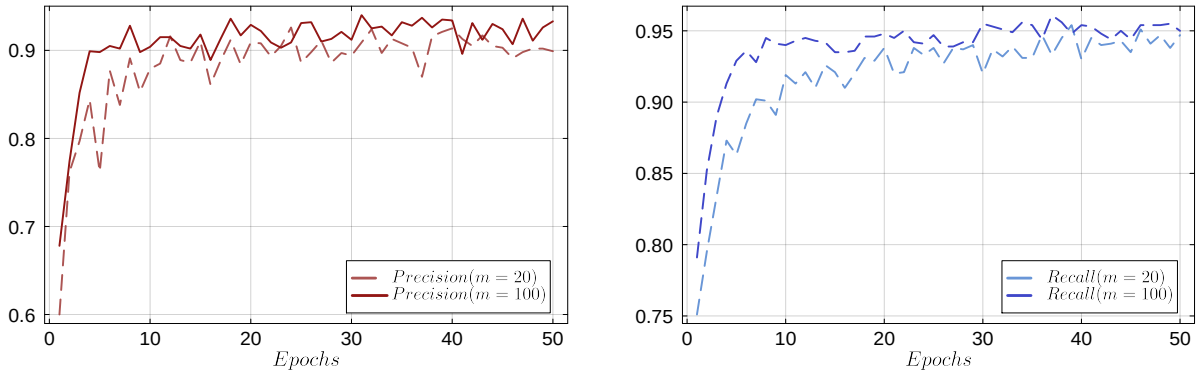
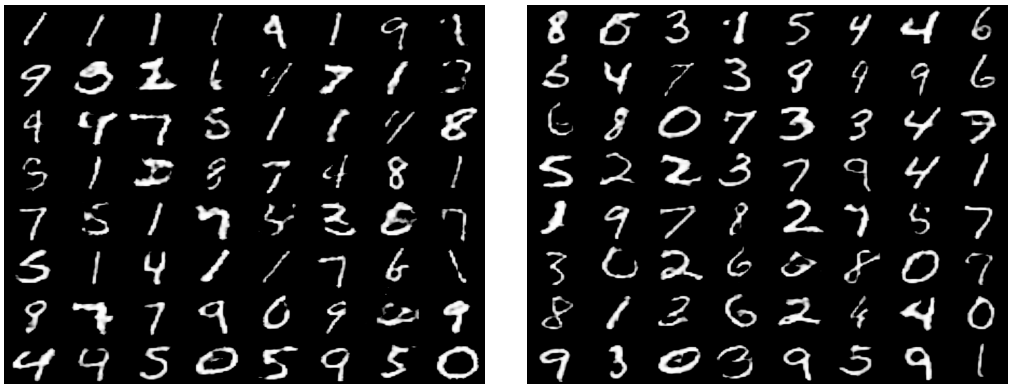


Figure 14: Precision (left) and recall (right) on the MNIST test set after 50 training epochs, using $m = 20$ and $m = 100$ random projections.

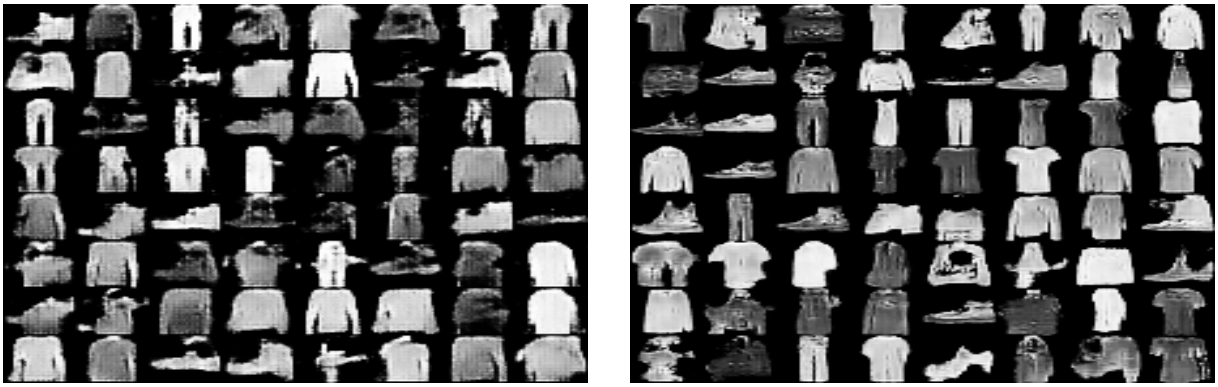


(a) DCGAN (b) DCGAN pretrained with ISL

Figure 15: Figure 15a shows digits generated by a standard DCGAN, while Figure 15b shows samples from a DCGAN pretrained with ISL. The ISL-pretrained model exhibits significantly greater diversity across all classes, whereas the vanilla DCGAN produces an overabundance of '1's.

CelebA We further evaluate our approach on the CelebA dataset, adopting the same training and evaluation setup used for the MNIST and FMNIST experiments. As quantitative measures, we report the Fréchet Inception Distance (FID; lower is better) alongside the standard Precision and Recall metrics that jointly capture sample fidelity and diversity.

Table 10 summarizes the results. The *dual-ISL+DCGAN* model achieves the highest Recall among all compared



(a) DCGAN (b) DCGAN pretrained with ISL

Figure 16: Comparison of Fashion-MNIST samples: 16a generated by DCGAN trained 40 epochs, and 16b generated by a DCGAN pretrained with dual-ISL. The ISL-pretrained model demonstrates greater class diversity and improved precision.

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 10: CelebA (64×64×3). 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.

methods, reflecting its superior ability to cover a broader range of modes in the data distribution. In contrast, variants emphasizing adversarial regularization, such as SN-DCGAN, exhibit higher Precision but noticeably lower Recall, suggesting a trade-off toward sample sharpness at the expense of diversity.

When the LS-DCGAN generator is pretrained with the sliced dual-ISL objective (*dual-ISL+LS-DCGAN*), we observe a clear improvement in Recall (0.824 vs. 0.524), while FID remains virtually unchanged (23.56 vs. 22.99) and Precision stays comparable (0.928 vs. 0.932). This trend indicates that dual-ISL enhances coverage of the true data manifold without compromising sample quality.

Taken together, these findings reinforce a consistent pattern across datasets: incorporating dual-ISL as a pretraining step yields models that generate more diverse samples while maintaining competitive fidelity, thus improving the balance between mode coverage and visual realism.

Training efficiency. Finally, we compare the computational cost of the proposed method across datasets. Table 11 reports the average training time per model on MNIST and CelebA under identical hardware and hyperparameter settings. On MNIST, dual-ISL and its hybrid variants train in roughly the same time as standard DCGAN, confirming that the pretraining stage adds negligible overhead. Although WDCGAN-GP is substantially slower due to gradient penalties, GMAN remains competitive despite using multiple discriminators. On CelebA, all models complete training in about one hour, with dual-ISL and its combinations remaining as efficient as DCGAN and LS-GAN. These results indicate that dual-ISL achieves improved stability and diversity without compromising runtime efficiency.

Dataset	Method	Time
MNIST	dual-ISL (m=20)	5m 34s
	dual-ISL (m=50)	8m 18s
	dual-ISL + DCGAN	7m 53s
	DCGAN	8m 09s
	WDCGAN-GP	14m 03s
	GMAN (5 discr.)	13m 50s

MNIST (28×28)

Dataset	Method	Time
CelebA	dual-ISL (m=100)	56m 54s
	DCGAN	59m 32s
	dual-ISL + DCGAN	59m 21s
	W-DCGAN-GP	61m 01s
	LS-GAN	59m 47s

CelebA ($64 \times 64 \times 3$)

Table 11: Average training time per model on MNIST and CelebA. dual-ISL pretraining introduces minimal computational cost while improving model stability and diversity. Time is reported in minutes (m) and seconds (s)

E PSEUDOCODES

Below we summarize two variants of the dual-ISL training procedure. Algorithm 1 presents the basic likelihood-free, rank-based update in the one-dimensional case. Algorithm 2 extends this to multi-dimensional data by drawing random one-dimensional projections and averaging the resulting rank losses.

Algorithm 1 dual-ISL Training

Require: Generator network f_θ, real samples $\{y_i\}_{i=1}^N$, batch size M, rank draws K, epochs T, learning rate η
Ensure: Trained parameters θ

```
for epoch = 1 to T do
  for each minibatch B ⊂ {y_i}_{i=1}^N of size M do
    Sample noise z ~ p_z
    Generate fictious sample ỹ = f_θ(z)
    Initialize histogram q ← 0
    for t = 1, ..., ⌊M/K⌋ do
      {y_i}_{i=1}^K ← RandomSubset(B, K) {draw K real samples from the minibatch}
      a_K ← Σ_{i=1}^K SoftIndicator[ỹ ≤ y_i] {differentiable count of the A_K statistic}
      q ← q + SoftHotEncoding(a_K, length = K + 1) {accumulate a differentiable one-hot into the histogram}
    end for
    q ← normalize(q)
    Compute loss: loss ← || (1/(K+1)) 1_{K+1} − q ||_{ℓ^1}
    Update: θ ← θ − η ∇_θ loss
  end for
end for
return θ
```

Algorithm 2 Dual-ISL with Random Projections

Require: Generator f_θ, real data $\{y_i\}_{i=1}^N \subset \mathbb{R}^d$, batch size M, rank draws K, projection draws L, epochs T, learning rate η

Ensure: Learned parameters θ

for epoch $= 1$ **to** T **do**
  for each minibatch $B \subset \{y_i\}$ of size M **do**
    Sample noise $z \sim p_z$
    Generate fictious sample $\tilde{y} = f_\theta(z)$
    Initialize histogram $\mathbf{q} \leftarrow \mathbf{0}$
    for $\ell = 1$ **to** L **do**
      Sample random unit vector $v_\ell \sim \text{Uniform}(S^{d-1})$
      for $t = 1, \ldots, \lfloor M/K \rfloor$ **do**
        $\{y_i\}_{i=1}^K \leftarrow \text{RandomSubset}(B, K)$ {draw K real samples from the minibatch}
        Compute projections $\tilde{u} = v_\ell^\top \tilde{y}$ and $u_i = v_\ell^\top y_i$ for all i
        $a_K \leftarrow \sum_{i=1}^K \text{SoftIndicator}[\tilde{u} \leq u_i]$ {differentiable count of the A_K statistic}
        $\mathbf{q} \leftarrow \mathbf{q} + \text{SoftHotEncoding}(a_K,\ \text{length} = K+1)$ {accumulate a differentiable one-hot into the histogram}
      end for
      $\mathbf{q} \leftarrow \text{normalize}(\mathbf{q})$
      $\text{loss} \leftarrow \| \mathbf{q} - \frac{1}{K+1}\mathbf{1}_{K+1} \|_1$
    end for
    Compute loss: $\text{projection_loss} = \text{mean}(\text{loss})$
    Update: $\theta \leftarrow \theta - \eta \nabla_\theta \text{projection_loss}$
  end for
end for
return θ

F EXPERIMENTAL SETUP

All experiments were performed on a MacBook Pro running macOS 13.2.1, equipped with an Apple M1 Pro CPU and 16 GB of RAM. When GPU acceleration was required, we used a single NVIDIA TITAN Xp with 12 GB of VRAM. Detailed hyperparameter settings for each experiment are provided in the corresponding sections. A repository containing all code and data is available at https://github.com/josemanuel22/dual-isl. The code will also be included in the supplementary materials in a folder.

G LIMITATIONS

Dual-ISL removes adversarial critics and yields stable, closed-form density surrogates, but scaling it beyond low dimensions exposes practical trade-offs. In what follows we examine the two main control knobs, (i) the number of projection slices m and (ii) the Bernstein truncation degree K—and how they affect accuracy, compute, and numerical stability. We also outline how structured/learned projections and an emerging link to optimal transport (OT) suggest paths to adaptive slicing, sharper guarantees, and faster solvers.

Extending dual-ISL to high-dimensional data via random projections introduces an inherent accuracy–compute trade-off. Using m slices improves capture of direction-dependent structure and, by standard Monte Carlo averaging, reduces estimator variance roughly as $1/m$; however, cost grows about linearly in m (and in the Bernstein order K), which can be prohibitive at image scale. Too few slices, in contrast, risk missing informative directions and biasing both density-ratio and quality estimates. Beyond purely random directions, structured or data-dependent projections (e.g., PCA/ICA axes, patch- or wavelet-based views, Fisher/score-aligned directions, or projections learned jointly with the generator) can raise information per slice; see, for example, generalized

or learned slicing (Kolouri et al., 2019), max-sliced and subspace-robust formulations that actively search for informative directions (Deshpande et al., 2019; Paty and Cuturi, 2019), and practical sampling and guarantee results for sliced distances (Nietert et al., 2022; Chen et al., 2020). Low-discrepancy schemes on the sphere (e.g., spherical Fibonacci grids) further reduce projection variance at fixed m (Keinert et al., 2015). A key open problem is to design adaptive policies that select or weight slices on the fly (bandit-style acquisition, leverage-score sampling, or gradient-aligned sampling) with finite-sample guarantees that link (m, K), ambient dimension, and approximation error.

Next, the truncation degree K governs the resolution of the Bernstein approximation used in dual-ISL. Small K is computationally stable and inexpensive but introduces bias by oversmoothing narrow modes and sharp transitions. Larger K can represent finer structure (e.g., high–curvature regions, thin modes), yet it also amplifies variance and can trigger numerical issues (ill-conditioned Gram systems; under/overflow in binomial terms) (Lorentz, 2012).

A classical Jackson-type estimate for Bernstein polynomials provides a rule of thumb: if f is α–Hölder on $[0, 1]$, then

$$\|f - B_K f\|_\infty = \mathcal{O}\big(K^{-\alpha/2}\big),$$

so bias decays at a predictable (but diminishing) rate as K grows. Equivalently, resolving features at spatial scale δ typically requires $K \asymp \delta^{-2}$. In practice, moderate K often balances detail and stability, but an automatic schedule is desirable. Promising directions include: (i) *adaptive* K per slice or training stage (increase K where residuals remain large); (ii) Recast the rank mapping in an orthonormal polynomial basis on $[0, 1]$ (e.g., Legendre/Jacobi); the corresponding dual basis is well-conditioned, so density estimation remains numerically stable as K grows. These strategies aim to reduce variance and conditioning issues while retaining the bias benefits of higher K.

Finally, dual-ISL naturally induces an *empirical coupling*: given two i.i.d. batches $\{x_i\}_{i=1}^n \sim p$ and $\{\tilde{x}_j\}_{j=1}^n \sim \tilde{p}_\theta$, the permutation $\pi \in S_n$ that aligns generated to real samples under ISL's rank criterion defines the discrete plan $\gamma_\pi = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, \tilde{x}_{\pi(i)})}$. In 1D, if the alignment enforces monotonicity (e.g., sorting both batches), γ_π matches the (unique) OT Monge map $T = F_p^{-1} \circ F_{\tilde{p}_\theta}$ (Villani, 2008; Santambrogio, 2015). Beyond 1D, a key question is when this coupling approaches an OT plan and how deviations depend on the number of slices m, the Bernstein degree K, dimension d, and sample size n.

Formally linking dual-ISL to OT could yield three benefits. *(i) Sharper bounds.* One can combine finite-sample Wasserstein rates with slicing complexity to bound $\mathbb{E}\, W_1(\gamma_\pi, \gamma^\star)$, separating slicing/truncation bias from statistical error (Fournier and Guillin, 2015). *(ii) Hybrid objectives.* Add a light OT term to ISL to enforce geometry while retaining stability, e.g., $\mathcal{L}(\theta) = d_K^{\text{slice}}(p, \tilde{p}_\theta) + \lambda \frac{1}{n}\sum_i \|x_i - \tilde{x}_{\pi(i)}\|$.

Our sliced Dual-ISL formulation introduces two main limitations that also motivate concrete research directions. *(i) Projection budget.* Performance depends on the number of slices m: too few miss anisotropic structure (bias), while many increase cost roughly linearly and can raise estimator variance for finite data. Future work should develop *adaptive*, data–driven policies to select/weight slices on the fly, and leverage *modality–aware* or learned directions to increase information per slice at fixed m. *(ii) Bernstein order.* The truncation degree K controls resolution: small K oversmooths (bias), large K captures fine detail but may amplify variance and cause numerical issues (ill–conditioning, over/underflow). Promising avenues include *automatic schedules* for K (per slice or stagewise), and *stable bases/solvers* (orthogonal polynomial). Finally, ISL implicitly induces a coupling between data and model samples, hinting at a tight link to optimal transport. Formalizing this connection could yield *sharper statistical guarantees* and *hybrid ISL–OT objectives/solvers* that preserve ISL's stability while improving geometric faithfulness. Addressing these points is key to scaling ISL to large images and extending it to non–Euclidean or discrete modalities.

H POTENTIAL SOCIETAL IMPACT

Implicit generative models offer strong data-synthesis capabilities but also carry dual-use risks. The ISL framework can produce realistic samples (e.g., faces or voices) and could likewise be misused to create deceptive content. Our results indicate that ISL models heavy-tailed distributions effectively, which may help represent rare or minority subpopulations more faithfully, subject to data quality and training practice. In addition, the closed-form density estimate provided by dual-ISL can facilitate inspection and auditing of model behaviour.

These properties do not by themselves ensure fairness or safety, but they can support responsible evaluation when combined with appropriate governance and safeguards.