Structure-preserving GANs

Jeremiah Birrell [1] **Markos A. Katsoulakis** [1] **Luc Rey-Bellet** [1] **Wei Zhu** [1]

Abstract

Generative adversarial networks (GANs), a class of distribution-learning methods based on a two-player game between a generator and a discriminator, can generally be formulated as a minmax problem based on the variational representation of a divergence between the unknown and the generated distributions. We introduce structure-preserving GANs as a data-efficient framework for learning distributions with additional structure such as group symmetry, by developing new variational representations for divergences. Our theory shows that we can reduce the discriminator space to its projection on the invariant discriminator space, using the conditional expectation with respect to the σ-algebra associated to the underlying structure. In addition, we prove that the discriminator space reduction must be accompanied by a careful design of structured generators, as flawed designs may easily lead to a catastrophic "mode collapse" of the learned distribution. We contextualize our framework by building symmetry-preserving GANs for distributions with intrinsic group symmetry, and demonstrate that both players, namely the equivariant generator and invariant discriminator, play important but distinct roles in the learning process. Empirical experiments and ablation studies across a broad range of data sets, including real-world medical imaging, validate our theory, and show our proposed methods achieve significantly improved sample fidelity and diversity—almost an order of magnitude measured in Fréchet Inception Distance—especially in the small data regime.

[1]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003, USA. Correspondence to: Jeremiah Birrell <birrell@math.umass.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

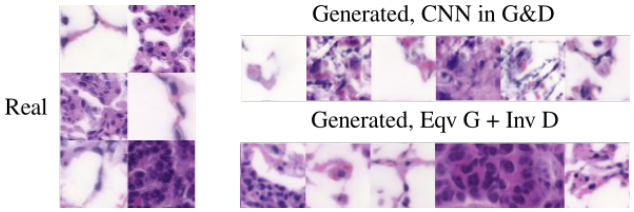


Figure 1. Real and GAN generated ANHIR images dyed with the H&E stain [cf. Section 5.5]. Left panel: real images. Right panels: randomly selected D_2^L-GAN generated samples after 40,000 generator iterations. Top right panel: `CNN G&D`, i.e., the baseline model. Bottom right panel: `Eqv G + Inv D`, i.e., our proposed framework contextualized in learning group-invariant distributions. More images are available in Appendix F.

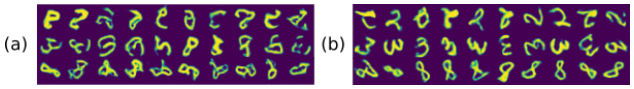


Figure 2. Randomly generated digits 2, 3 and 8 by GANs trained on the rotated MNIST images using **1% (600)** training samples. (a): the baseline CNN model. (b): our proposed framework for learning group-invariant distributions.

1. Introduction

Since their introduction by Goodfellow et al. (2014), generative adversarial networks (GANs) have become a burgeoning domain in distribution learning with a diverse range of innovative applications (Karras et al., 2019; Zhu et al., 2019; Mustafa et al., 2019; Yi et al., 2019). Mathematically, the minmax game between a generator and a discriminator in GAN can typically be formulated as minimizing a divergence— or other notions of "distance"—with a variational representation between the unknown and the generated distributions. Such formulations, however, do not make prior *structural* assumptions on the probability measures, making them sub-optimal in sample efficiency when learning distributions with intrinsic structures, such as the (rotation) group symmetry for medical images without preferred orientation; see Figure 1.

We introduce, in this work, the *structure-preserving GANs*, a data-efficient framework for learning probability measures

with embedded structures, by developing new variational representations for divergences between structured distributions. We demonstrate that efficient adversarial learning can be achieved by reducing the discriminator space to its projection onto its invariant subspace, using the conditional expectation with respect to the σ-algebra associated to the underlying structure; such practice, which is rigorously justified by our theory and generally applicable to a broad range of variational divergences, acts effectively as an unbiased regularization to prevent discriminator overfitting, a common challenge for GAN optimization in the limited data regime (Zhao et al., 2020). Furthermore, our theory suggests that the discriminator space reduction must be accompanied by *correctly* building generators sharing the same probabilistic structure, as the lack of which may easily lead to "mode collapse" in the trained model, i.e., the generated distribution samples only a subset of the support of the data source [cf. Figure 4a (2nd row)].

As an example, we contextualize our framework by building symmetry-preserving GANs for learning distributions with group symmetry. Unlike prior empirical work, our choice of equivariant generators and invariant discriminators is theoretically founded, and we show (theoretically and empirically) how flawed design of equivariant generators results easily in the aforementioned mode collapse [cf. Figure 4a (4th row)]. Experiments and ablation studies over synthetic and real-world data sets validate our theory, disentangle the contribution of the structural priors on generators and discriminators, and demonstrate the significant outperformance of our framework in terms of both sample quality and diversity—in some cases almost by an order of magnitude measured in Fréchet Inception Distance; see Figure 1 and 2 for a visual illustration.

2. Related Work

Neural generation of group-invariant distributions has mainly been proposed in a flow-based framework (Köhler et al., 2019; 2020; Rezende et al., 2019; Liu et al., 2019; Biloš & Günnemann, 2021; Boyda et al., 2021; Garcia Satorras et al., 2021). Such models typically use an equivariant normalizing-flow to push-forward a group-invariant prior distribution to a complex invariant target. In the context of GANs, Dey et al. (2021) intuitively replace the 2D convolutions with group convolutions (Cohen & Welling, 2016a) to build group-equivariant GANs; however, their empirical study has not been justified by theory, and their incomplete design of the equivariant generator may easily lead to a "mode collapse" of the learned model; see the discussion of Theorem 4.6. The existence of symmetry can often be deduced from prior or domain knowledge of the distribution, e.g., the rotation symmetry for medical images without preferred orientation. Symmetry detection from data has also been studied in recent works such as (Dehmamy et al., 2021). When extended from group symmetry to probability structures induced from other operators, our work is also related to GAN-assisted coarse-graining (CG) for molecular dynamics (Durumeric & Voth, 2019) and cosmology (Mustafa et al., 2019; Feder et al., 2020); see the end of Section 4.1 for a detailed discussion.

3. Background and Motivation

3.1. Generative adversarial networks

Generative adversarial networks are a class of methods in learning a probability distribution via a zero-sum game between a generator and a discriminator (Goodfellow et al., 2014; Arjovsky et al., 2017; Nowozin et al., 2016; Gulrajani et al., 2017). Specifically, let $(X, \mathcal{M})$ be a measurable space, and $\mathcal{P}(X)$ be the set of probability measures on X; given a target distribution $Q \in \mathcal{P}(X)$, the original GAN proposed by Goodfellow et al. (2014) learns Q by solving

$$\inf_{g\in G} D(Q\|P_g) = \inf_{g\in G}\sup_{\gamma\in\Gamma} H[\gamma; Q, P_g], \tag{1}$$

where $H[\gamma; Q, P_g] = E_Q[\log\gamma] + E_{P_g}[\log(1-\gamma)]$. The map $g : Z \to X$ in Eq. (1) is called a *generator*, which maps a random vector $z \in Z$ to a generated sample $g(z) \in X$, pushing forward the noise distribution $P \in \mathcal{P}(Z)$ (typically a Gaussian) to a probability measure $P_g \in \mathcal{P}(X)$, i.e., $P_g := g_*P := P \circ g^{-1}$; the test function $\gamma : X \to \mathbb{R}$ is called a *discriminator*, which aims to differentiate the source distribution Q and the generated probability measure P_g by maximizing $H[\gamma; Q, P_g]$. The spaces G and Γ, respectively, of generators and discriminators are both parametrized by neural networks (NNs), and the solution of model (1) is the best generator $g \in G$ that is able to "fool" all discriminators $\gamma \in \Gamma$ by achieving the smallest $D(Q\|P_g)$, which measures the "dissimilarity" between Q and P_g.

3.2. Variational representations for divergences

Mathematically, most GANs can be formulated as minimizing the "distance" between the probability measures Q and P_g according to some divergence or probability metric with a variational representation $\sup_{\gamma\in\Gamma} H(\gamma; Q, P_g)$ as in (1). We hereby recast these formulations in a unified but flexible mathematical framework that will prove essential in Section 4.1. Let $\mathcal{M}(X)$ be the space of measurable functions on X and $\mathcal{M}_b(X)$ be the subspace of bounded measurable functions. Given an objective functional $H : \mathcal{M}(X)^n \times \mathcal{P}(X) \times \mathcal{P}(X) \to [-\infty, \infty]$ and a test function space $\Gamma \subset \mathcal{M}(X)^n$, $n \in \mathbb{Z}^+$, we define

$$D_H^\Gamma(Q\|P) = \sup_{\gamma\in\Gamma} H(\gamma; Q, P)\,. \tag{2}$$

D_H^Γ is called a *divergence* if $D_H^\Gamma \geq 0$ and $D_H^\Gamma(Q\|P) = 0$ if and only if $Q = P$, hence providing a notion of "distance"

between probability measures. Variational representations of the form (2) have been widely used, including in GANs (Goodfellow et al., 2014; Nowozin et al., 2016; Arjovsky et al., 2017), divergence estimation (Nguyen et al., 2007; 2010; Ruderman et al., 2012; Birrell et al., 2021), determining independence through mutual information estimation (Belghazi et al., 2018), uncertainty quantification of stochastic processes (Chowdhary & Dupuis, 2013; Dupuis et al., 2016), bounding risk in probably approximately correct (PAC) learning (McAllester, 1999; Shawe-Taylor & Williamson, 1997; Catoni et al., 2008), parameter estimation (Broniatowski & Keziou, 2009), statistical mechanics and interacting particles (Kipnis & Landim, 1999), and large deviations (Dupuis & Ellis, 2011). It is known that formula (2) includes, through suitable choices of functional $H(\gamma; Q, P)$ and function space Γ, many divergences and probability metrics. Below we list several classes of examples.

(a) f-divergences. Let $f : [0, \infty) \to \mathbb{R}$ be convex and lower semi-continuous (LSC), with $f(1) = 0$ and f strictly convex at $x = 1$. The f-divergence between Q and P is

$$D_f(Q\|P) = \sup_{\gamma \in \mathcal{M}_b(X)} \{E_Q[\gamma] - E_P[f^*(\gamma)]\}, \quad (3)$$

where f^* denotes the Legendre transform of f. Some notable examples of the f-divergences include the Kullback-Leibler (KL) divergence and the family of α-divergences, which are constructed, respectively, from

$$f_{KL} = x \log x, \quad f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha - 1)}, \alpha > 0, \alpha \neq 1. \quad (4)$$

The flexibility of f allows one to tailor the divergence to the data source, e.g., for heavy tailed data. However, the formula (3) becomes $D_f(Q\|P) = \infty$ when Q is not absolutely continuous with respect to P, limiting its efficacy in comparing distributions with low-dimensional support.

(b) Γ-Integral Probability Metrics (IPMs). Given $\Gamma \subset \mathcal{M}_b(X)$, the Γ-IPM between Q and P is defined as

$$W^\Gamma(Q, P) = \sup_{\gamma \in \Gamma} \{E_Q[\gamma] - E_P[\gamma]\}. \quad (5)$$

Apart from the Wasserstein metric when $\Gamma = \mathrm{Lip}^1(X)$ (the space of 1-Lipschitz functions), examples of IPMs also include the total variation metric, the Dudley metric, and maximum mean discrepancy (MMD) (Müller, 1997; Sriperumbudur et al., 2012). With suitable choices of Γ, IPMs are able to meaningfully compare not-absolutely continuous distributions, but they could potentially fail at comparing distributions with heavy tails (Birrell et al., 2022).

(c) (f, Γ)-divergences. This class of divergences was introduced by Birrell et al. (2022) and they subsume both f-divergences and Γ-IPMs. Given a function f satisfying the same condition as in the definition of the f-divergence and $\Gamma \subset \mathcal{M}_b(X)$, the (f, Γ)-divergence is defined as

$$D_f^\Gamma(Q\|P) = \sup_{\gamma \in \Gamma} \left\{E_Q[\gamma] - \Lambda_f^P[\gamma]\right\}, \quad (6)$$

where $\Lambda_f^P[\gamma] = \inf_{\nu \in \mathbb{R}} \{\nu + E_P[f^*(\gamma - \nu)]\}$. One can verify that (6) includes as a special case the f-divergence (3) when $\Gamma = \mathcal{M}_b(X)$, and it is demonstrated in (Birrell et al., 2022) that under suitable assumptions on Γ we have

$$0 \leq D_f^\Gamma(Q\|P) \leq \min\{D_f(Q\|P), W^\Gamma(Q, P)\}, \quad (7)$$

making D_f^Γ suitable to compare not-absolutely continuous distributions with heavy tails. An example of the (f, Γ)-divergence is the Lipschitz α-divergence,

$$D_\alpha^L(Q\|P) = \sup_{\gamma \in \mathrm{Lip}_b^L(X)} \{E_Q[\gamma] - \Lambda_{f_\alpha}^P[\gamma]\}, \quad (8)$$

where $f = f_\alpha$ as in Eq. (4), and $\Gamma = \mathrm{Lip}_b^L(X)$ is the space of bounded L-Lipschitz functions.

(d) Sinkhorn divergences. The Wasserstein metric associated with a cost function $c : X^2 \to \mathbb{R}^+$ has the variational representation $W_c^\Gamma(Q, P) = \sup_{\gamma=(\gamma_1,\gamma_2)\in\Gamma}\{E_P[\gamma_1] + E_Q[\gamma_2]\}$, where $\Gamma = \{(\gamma_1, \gamma_2) \in C(X)^2 : \gamma_1(x) + \gamma_2(y) \leq c(x, y)\}$, and $C(X)$ is the space of continuous functions on X. The Sinkhorn divergence is given by

$$\mathcal{SD}_{c,\epsilon}^\Gamma(Q, P) = W_{c,\epsilon}^\Gamma(Q, P) - \frac{W_{c,\epsilon}^\Gamma(Q, Q) + W_{c,\epsilon}^\Gamma(P, P)}{2}, \quad (9)$$

where $W_{c,\epsilon}^\Gamma(Q, P)$ is the entropic regularization of the Wasserstein metrics [cf. Eq. (33)].

We refer to Appendix A for a detailed discussion of the variational divergences introduced above. In all the aforementioned examples, the choice of the discriminator space, Γ, is a defining characteristic of the divergence. We will explain, in Section 4.1, a general framework, i.e., the structure-preserving GANs, for incorporating added structural knowledge of the probability distributions or data sets into the choice of Γ, leading to enhanced performance and data efficiency in adversarial learning of structured distributions.

3.3. Group invariance and equivariance

We first introduce the structure-preserving GAN framework in the context of learning distributions with group symmetry. We emphasize that the focus of this work is not to discuss the group-invariance properties of probability measures (which can be found in, e.g., (Schindler, 2003)), but to understand how to incorporate such structural information into the generator/discriminator of GANs such that invariant probability distributions can be learned more efficiently. However,

we first require the following background and notations.

Groups and group actions. A *group* is a set Σ equipped with a binary operator, the group product, satisfying the axioms of associativity, identity, and invertibility. Given a group Σ and a set X, a map $T : \Sigma \times X \to X$ is called a *group action* if, for all $\sigma \in \Sigma$, $T_\sigma := T(\sigma, \cdot) : X \to X$ is an automorphism on X, and $T_{\sigma_1} \circ T_{\sigma_2} = T_{\sigma_1 \cdot \sigma_2}, \forall \sigma_1, \sigma_2 \in \Sigma$. In this paper, we will consider mainly the 2D rotation group $SO(2) = \{R_\theta \in \mathbb{R}^{2\times 2} : \theta \in \mathbb{R}\}$ and roto-reflection group $O(2) = \{R_{m,\theta} \in \mathbb{R}^{2\times 2} : m \in \mathbb{Z}, \theta \in \mathbb{R}\}$, where R_θ is the 2D rotation matrix of angle θ, and $R_{m,\theta}$ has a further reflection if $m \equiv 1 \pmod 2$. The natural actions of $SO(2)$ and $O(2)$ on $\mathbb{R}^2$ are matrix multiplications, which can be lifted to actions on the space of (k-channel) planar signals $L^2(\mathbb{R}^2, \mathbb{R}^k)$, e.g., RGB images. More specifically, when Σ is $SO(2)$ or $O(2)$ let $T_\sigma f(x) := f(\sigma^{-1}x),\ \forall \sigma \in \Sigma, \forall f \in L^2(\mathbb{R}^2, \mathbb{R}^k)$. We will also consider the finite subgroups C_n, D_n, respectively, of $SO(2)$ and $O(2)$, with the rotation angles θ restricted to integer multiples of $2\pi/n$.

Group equivariance and invariance. Let T^Z and T^X, respectively, be Σ-actions on the spaces Z and X. A map $g : Z \to X$ is called Σ-*equivariant* if $T^X_\sigma \circ g = g \circ T^Z_\sigma, \forall \sigma \in \Sigma$. A map $\gamma : X \to Y$ is called Σ-*invariant* if $\gamma \circ T^X_\sigma = \gamma, \forall \sigma \in \Sigma$. Invariance is thus a special case of equivariance after equipping Y with the action $T^Y_\sigma y \equiv y, \forall \sigma \in \Sigma$. In the context of NNs, achieving equivariance/invariance via group-equivariant CNNs (G-CNNs) has been well-studied, and we refer the reader to (Cohen et al., 2019; Weiler & Cesa, 2019) for a complete theory of G-CNNs.

Let G be a collection of measurable maps $g : Z \to X$. We denote its subset of Σ-equivariant maps as $G^{\text{eqv}}_\Sigma := \{g \in G : T^X_\sigma \circ g = g \circ T^Z_\sigma,\ \forall \sigma \in \Sigma\}$. Similarly, let Γ be a set of measurable functions $\gamma : X \to Y$; its subset, $\Gamma^{\text{inv}}_\Sigma$, of Σ-invariant functions is defined as

$$\Gamma^{\text{inv}}_\Sigma := \{\gamma \in \Gamma : \gamma \circ T^X_\sigma = \gamma,\ \forall \sigma \in \Sigma\}. \tag{10}$$

The function space Γ is called *closed under* Σ if

$$\gamma \circ T^X_\sigma \in \Gamma,\ \ \forall \sigma \in \Sigma,\ \ \forall \gamma \in \Gamma. \tag{11}$$

Finally, a probability measure $P \in \mathcal{P}(X)$ is called Σ-*invariant* if $P = P \circ (T^X_\sigma)^{-1}$ for all $\sigma \in \Sigma$. For instance, the distribution of medical images without orientation preference should be $SO(2)$-invariant; see Figure 1. The set of all Σ-invariant distributions on X is denoted as

$$\mathcal{P}_\Sigma(X) := \{P \in \mathcal{P}(X) : P \text{ is } \Sigma\text{-invariant}\}. \tag{12}$$

3.4. Definition of Haar measure on Σ and the symmetrization operators S_Σ and S^Σ

We will make frequent use of the symmetrization operators, on both functions and probability distributions, that are induced by a group action on X. These are constructed using the unique Haar probability measure, μ_Σ, of a compact Hausdorff topological group Σ (see, e.g., Chapter 11 in Folland (2013)). Intuitively the Haar measure is the uniform probability measure on Σ. Mathematically, this is expressed via the invariance of Haar measure under group multiplication, $\mu_\Sigma(\sigma \cdot E) = \mu_\Sigma(E \cdot \sigma) = \mu_\Sigma(E)$ for all $\sigma \in \Sigma$ and all Borel sets $E \subset \Sigma$. This is a generalization of the invariance of Lebesgue measure under translations and rotations. The Haar measure can be used to define symmetrization operators on both functions and probability measures as follows (going forward, we assume the group action is measurable).

Symmetrization of functions: $S_\Sigma : \mathcal{M}_b(X) \to \mathcal{M}_b(X)$,

$$S_\Sigma[\gamma](x) := \int_\Sigma \gamma(T_{\sigma'}(x))\mu_\Sigma(d\sigma') = E_{\mu_\Sigma}[\gamma \circ T_{\sigma'}(x)]. \tag{13}$$

Symmetrization of probability measures (dual operator): $S^\Sigma : \mathcal{P}(X) \to \mathcal{P}(X)$, defined for $\gamma \in \mathcal{M}_b(X)$ by

$$E_{S^\Sigma[P]}\gamma := \int_X S_\Sigma[\gamma](x)dP(x) = E_P S_\Sigma[\gamma]. \tag{14}$$

Remark 3.1. Sampling from $S^\Sigma[P]$: If $x_i, i = 1, ..., N$ are samples from P, and $\sigma_j, j = 1, ..., M$ are samples from the Haar probability measure μ_Σ (all independent) then $T_{\sigma_j}(x_i)$ are samples from $S^\Sigma[P]$. If P is Σ-invariant then the use of $T_{\sigma_j}(x_i)$ can be viewed as a form of data augmentation.

The following lemma provides several key properties of the symmetrization operators. Proofs and further details can be found in Appendix B, Lemma B.1.

Lemma 3.2. *(a) The symmetrization operator $S_\Sigma : \mathcal{M}_b(X) \to \mathcal{M}_b(X)$ is a projection onto the subspace of Σ-invariant bounded measurable functions, $\mathcal{M}^{inv}_{b,\Sigma}$ [cf. Eq. (10)].*

(b) The symmetrization operator $S^\Sigma : \mathcal{P}(X) \to \mathcal{P}(X)$ is a projection onto the subset of Σ-invariant probability measures, $\mathcal{P}_\Sigma(X)$ [cf. Eq. (12)].

(c) S_Σ is the conditional expectation with respect to the σ-algebra $\mathcal{M}_\Sigma$ of Σ-invariant sets, $\mathcal{M}_\Sigma := \{$Measurable sets $B \subset \mathcal{M} : T_\sigma(B) = B, \forall \sigma \in \Sigma\}$, i.e., $S_\Sigma[\gamma] = E_P[\gamma|\mathcal{M}_\Sigma]$ for all $\gamma \in \mathcal{M}_b(X)$, $P \in \mathcal{P}_\Sigma(X)$.

Lemma 3.2 implies that since S_Σ, S^Σ are projections onto $\mathcal{M}^{\text{inv}}_{b,\Sigma}, \mathcal{P}_\Sigma(X)$ respectively, i.e. $S_\Sigma \circ S_\Sigma = S_\Sigma$, $S^\Sigma \circ S^\Sigma = S^\Sigma$, they are necessarily *structure-preserving*, namely here symmetry-preserving. We discuss a general concept of structure-preserving operators at the end of Section 4.1.

4. Theory

We present in this section our theory for structure-preserving GANs. The results are first stated for the special case of

learning group-invariant distributions. We then extend the theory to a general class of structure-preserving operators.

4.1. Invariant discriminator theorem

We demonstrate under assumptions outlined below and for broad classes of divergences and probability metrics that for Σ-invariant probability measures P, Q we can restrict the test function space Γ (discriminator space in GANs) in (2) to the subset of Σ-invariant functions, $\Gamma_\Sigma^{\text{inv}}$ [cf. Eq. (10)], without changing the divergence/probability metric, i.e.,

$$D_H^\Gamma(Q\|P) = D_H^{\Gamma_\Sigma^{\text{inv}}}(Q\|P) \quad \text{for all } Q, P \in \mathcal{P}_\Sigma\,. \tag{15}$$

The space $\Gamma_\Sigma^{\text{inv}}$ is a much "smaller" and more efficient discriminator space to optimize over in the proposed GANs. We rigorously formulate our results in the following theorem, which first considers the (f,Γ) divergence (6), the Γ-IPM (5), and the Sinkhorn divergence (9). The proof is found in Appendix B.

Theorem 4.1. *If $S_\Sigma[\Gamma] \subset \Gamma$ and the probability measures P, Q are Σ-invariant then*

$$D^\Gamma(Q\|P) = D^{\Gamma_\Sigma^{\text{inv}}}(Q\|P)\,, \tag{16}$$

where D^Γ is an (f,Γ)-divergence or a Γ-IPM. Eq. (16) also holds for Sinkhorn divergences if the cost is Σ-invariant (i.e., $c(T_\sigma(x), T_\sigma(y)) = c(x,y)$ for all $\sigma \in \Sigma$, $x, y \in X$).

Remark 4.2. Eq. (16) can be generalized to a wider range of objective functionals satisfying appropriate convexity, continuity, and invariance conditions; see Theorem B.10.

For the Σ-invariant (f,Γ)-divergences, we also obtain a refined version of (7), given by the following infimal convolution formula (for appropriate Γ and f):

$$D_f^{\Gamma_\Sigma^{\text{inv}}}(Q\|P) = \inf_{\eta \in \mathcal{P}_\Sigma(X)} \{D_f(\eta\|P) + W^{\Gamma_\Sigma^{\text{inv}}}(Q,\eta)\} \tag{17}$$

for all $Q, P \in \mathcal{P}_\Sigma(X)$. See Appendix D for details on (17) and other results generalizing those in (Birrell et al., 2022).

Theorem 4.1 suggests that the discriminator space reduction effectively acts as an unbiased regularization to prevent discriminator overfitting, a common challenge for GAN optimization in the small data regime. Using invariant discriminators can thus improve the data-efficiency of the model; this will be empirically verified in Tables 1-3.

Examples satisfying the key condition $S_\Sigma[\Gamma] \subset \Gamma$ of Theorem 4.1

1. First we consider the standard f-divergence (3) between two Σ-invariant probability measures P and Q. The identity $S_\Sigma[\mathcal{M}_b(X)] = \mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$ from Lemma 3.2 implies that the functions space can be restricted to the Σ-invariant bounded functions $\mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$, giving rise to an (f,Γ)-divergence (6) with $\Gamma = \mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$, i.e., $D_f(Q\|P) = D_f^{\mathcal{M}_{b,\Sigma}^{\text{inv}}(X)}(Q\|P)$.
2. If the group Σ is finite and the function space $\Gamma \subset \mathcal{M}_b(X)$ is convex and closed under Σ in the sense of (11), then $S_\Sigma[\Gamma] \subset \Gamma$, as readily follows from the definition (13). Our implemented examples in Section 5 fall under this category.
3. The space of 1-Lipschitz functions on a metric space (X,d), assuming the action is 1-Lipschitz, i.e., $d(T_\sigma(x), T_\sigma(y)) \le d(x,y)$ for all $\sigma \in \Sigma$, $x, y \in X$.
4. The unit ball in an appropriate RKHS; see Lemma C.1.
5. More generally, if Γ is convex and closed in the weak topology on Γ induced by integration against finite signed measures; see Lemma C.3 for a proof.

Extension to other structure-preserving operators Let $K_x(dx')$ be a probability kernel from X to X and define $S_K : \mathcal{M}_b(X) \mapsto \mathcal{M}_b(X)$ by $S_K[f](x) := \int f(x')K_x(dx')$. K also defines a dual map $S^K : \mathcal{P}(X) \to \mathcal{P}(X)$, $S^K[P] := \int K_x(\cdot)P(dx)$. Let $\mathcal{P}_K(X)$ be the set of K-invariant probability measures, i.e., $\mathcal{P}_K(X) = \{P \in \mathcal{P}(X) : S^K[P] = P\}$. In this setting we have the following generalization of Theorem 4.1.

Theorem 4.3. *If $S_K[\Gamma] \subset \Gamma$ and $Q, P \in \mathcal{P}_K(X)$ then*

$$D^\Gamma(Q\|P) = D^{S_K[\Gamma]}(Q\|P)\,, \tag{18}$$

where D^Γ is an (f,Γ)-divergence or a Γ-IPM. It also holds when D^Γ is a Sinkhorn divergence if $S_K[c(\cdot,y)] = c(\cdot,y)$ and $S_K[c(x,\cdot)] = c(x,\cdot)$ for all $x, y \in X$.

In addition, if S_K is a projection (i.e., $S_K \circ S_K = S_K$) then $S_K[\Gamma] = \Gamma_K^{\text{inv}}$ where $\Gamma_K^{\text{inv}} := \{\gamma \in \Gamma : S_K[\gamma] = \gamma\}$.

Remark 4.4. Conditional expectations, $S_K[f] := E_P[f|\mathcal{A}]$, are a special case of Theorem 4.3 with kernel being a regular conditional probability, $K = P(\cdot|\mathcal{A})$. Here Γ_K^{inv} is the set of $\mathcal{A}$-measurable functions in Γ, which can be significantly "smaller" than Γ. The case where $\mathcal{A} = \sigma(\xi)$ for some random variable ξ has particular importance in coarse graining of molecular dynamics (Noid, 2013; Pak & Voth, 2018), see Appendix E. The result for Σ-invariant measures, Theorem 4.1, is also special case of Theorem 4.3, where the kernel is $K_x = \mu_\Sigma \circ R_x^{-1}$, $R_x(\sigma) := T_\sigma(x)$. Alternatively, Lemma 3.2 (c) shows S_Σ can be written as a conditional expectation.

Remark 4.5. Theorem 4.3 is an instance of the data processing inequality; see Theorem 2.21 in (Birrell et al., 2022).

4.2. Equivariant generator theorem

Theorem 4.1 provides the theoretical justification for reducing the discriminator space Γ to its Σ-invariant subset $\Gamma_\Sigma^{\text{inv}}$

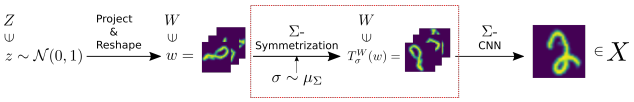


Figure 3. The Σ-symmetrization layer (enclosed in the red rectangle), which is missing in (Dey et al., 2021), ensures generator equivariance, which is critical in preventing GAN "mode collapse" [cf. Remark 4.11].

when the source Q and the generated measure P_g are **both** Σ-invariant. Our next theorem, however, shows that such practice could easily lead to "mode collapse" if one of the two distributions is **not** Σ-invariant, see Figure 4a; the proof is deferred to Appendix B.

Theorem 4.6. *Let $S_\Sigma[\Gamma] \subset \Gamma$ and $P, Q \in \mathcal{P}(X)$, i.e., not necessarily Σ-invariant. We have*

$$D^{\Gamma^{inv}_\Sigma}(Q\|P) = D^\Gamma(S^\Sigma[Q]\|S^\Sigma[P])\,, \tag{19}$$

where D^Γ is an (f,Γ)-divergence or a Γ-IPM.

Remark 4.7. The analogous result for the Sinkhorn divergences also holds if the cost is separately Σ-invariant in each variable, i.e., $c(T_\sigma(x), y) = c(x, y)$ and $c(x, T_\sigma(y)) = c(x, y)$ for all $\sigma \in \Sigma$, $x, y \in X$. However, this is a strong assumption that is not satisfied by most commonly used cost functions and actions.

Theorem 4.6 has the following implications: If one uses a Σ-invariant GAN (i.e., invariant discriminators and equivariant generators) to learn a non-invariant data source Q then one will in fact learn the symmetrized version $S^\Sigma[Q]$. On the other hand, if the data source Q is Σ-invariant (i.e., $S^\Sigma[Q] = Q$, cf. Lemma 3.2) but the GAN generated distribution P_g is not then discriminators from $\Gamma^{\text{inv}}_\Sigma$ alone can not differentiate Q and P_g, i.e., $D^{\Gamma^{\text{inv}}_\Sigma}(Q\|P_g) = 0$, as long as $Q = S^\Sigma[P_g]$. This suggests that P_g can easily suffer from "mode collapse", as it only needs to equal Q after Σ-symmetrization; we refer readers to Figure 4a (2nd and 4th rows) for a visual illustration, where a unimodal P_g can be erroneously selected as the "best" fitting model, even though its Σ-symmetrization $S^\Sigma[P_g]$ should be the "correct" one.

To prevent this from happening, one needs to ensure the generator produces a Σ-invariant distribution P_g; this is guaranteed by the following Theorem.

Theorem 4.8. *If $P_Z \in \mathcal{P}(Z)$ is Σ-invariant and $g : Z \to X$ is Σ-equivariant then the push-forward measure $P_g := P_Z \circ g^{-1}$ is Σ-invariant, i.e., $P_g \in \mathcal{P}_\Sigma(X)$.*

See Appendix B for a proof. We note that equivariant flow-based methods have also been proposed based on a similar strategy to Theorem 4.8. We refer readers to Section 2 for a discussion of related works.

Remark 4.9. Suppose $g = \gamma_2 \circ \gamma_1$ is a composition of two maps, $\gamma_1 : Z \to W$ and $\gamma_2 : W \to X$. Even if γ_1 is not Σ-equivariant (in fact, Z does not even need to be equipped with a Σ-action T^Z_σ), as long as $P_{\gamma_1} \in \mathcal{P}(W)$ is Σ-invariant and γ_2 is Σ-equivariant, the push-forward measure $P_g \in \mathcal{P}(X)$ is still Σ-invariant.

To construct the Σ-invariant noise source required in Theorem 4.8 (or Remark 4.9) one can begin with an arbitrary noise source and use a **Σ-symmetrization layer**, as described by the following theorem.

Theorem 4.10. *Let $W \sim \mu_\Sigma$ and N be a Z-valued random variable (i.e., an arbitrary noise source). If N and W are independent then the distribution of $T^Z(W, N)$ is Σ-invariant.*

Remark 4.11. Dey et al. (2021) also proposed to use G-CNNs to generate images with C_4/D_4-invariant distributions. However, the first step in their model, i.e., the "Project & Reshape" step [cf. Figure 3], uses a fully-connected layer which destroys the group symmetry in the noise source, leading to non-invariant final distribution P_g even if the subsequent layers are all Σ-equivariant. This easily leads to "mode collapse" [cf. Theorem 4.6], which we will empirically demonstrate in Section 5; see, e.g., Figure 4a (4th row). An easy remedy for this is to add a Σ-symmetrization layer: let w be the output of "Project & Reshape"; the Σ-symmetrization layer draws a random $\sigma \sim \mu_\Sigma$ and transforms w into $T^W_\sigma(w)$, producing a Σ-invariant distribution on the layer output (see Theorem 4.10). The final distribution P_g is thus Σ-invariant if subsequent layers are all Σ-equivariant by Remark 4.9. See Figure 3 for a visual illustration.

5. Experiments

We present experiments on both synthetic and real-world data sets with embedded group symmetry to empirically verify our theory for structure-preserving GANs in Section 4.

5.1. Algorithmic Feasibility

Theorems 4.1 and 4.8 imply that one can build invariant GANs by using Σ- invariant discriminators, Σ-equivariant generators, and a Σ-invariant noise source. Equivariant networks for arbitrary group symmetry (and gauge invariance) have been studied in recent works such as (Cohen & Welling, 2016b). Invariant noise sources can be constructed as shown in Theorem 4.10. We note that the symmetrization operators S^Σ, S_Σ are only used in the proofs of theoretical properties of the proposed GANs and are not needed in practical implementations. The necessary invariance/equivariance is built into the discriminator/generator via the structure of the layers; see Appendix G.4.

5.2. Data sets and common experimental setups

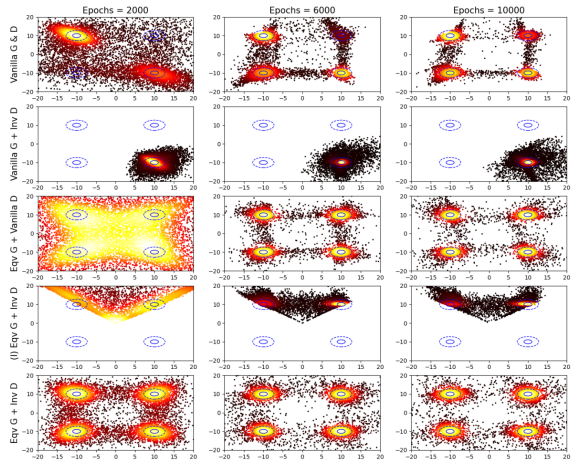
Toy example. Following (Birrell et al., 2022), this synthetic data source is a mixture of four 2D t-distributions with 0.5 degrees of freedom, embedded in a plane in $\mathbb{R}^{12}$. The four centers of the t-distributions are located (in the supporting plane) at coordinates $(\pm 10, \pm 10)$, exhibiting C_4-symmetry [cf. Figure 4a].

RotMNIST is built by randomly rotating the original 10-class 28×28 MNIST digits (LeCun et al., 1998), resulting in an $SO(2)$-invariant distribution. We use different portions of the 60,000 training images for experiments in Section 5.4.

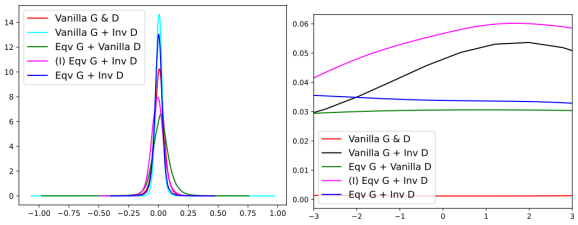
ANHIR consists of pathology slides stained with 5 distinct dyes for the study of cellular compositions (Borovec et al., 2020). Following (Dey et al., 2021), we extract from the original images 28,407 foreground patches of size 64×64. The staining dye is used as the class label for conditioned image synthesis. As the images have no preferred orientation/reflection, the distribution is $O(2)$-invariant.

LYSTO contains 20,000 patches extracted from whole-slide images of breast, colon and prostate cancer stained with immunohistochemical markers (Ciompi et al., 2019). The images are classified into 3 categories based on the organ source, and we downsize the images to 64×64. Similar to ANHIR, this data set is also $O(2)$-invariant.

Common experimental setups. To verify our theory in Section 4, and to quantify and disentangle the contributions of the structure-preserving discriminator (`D`) and generator (`G`) (Theorem 4.1 and Theorem 4.6), we replace the baseline `G` and/or `D` by their group-equivariant/invariant counterparts, `Eqv G` and `Inv D`, while adjusting the number of filters according to the group size to ensure a similar number of trainable parameters. We also consider the incomplete attempt by Dey et al. (2021) at building equivariant generators (`(I)Eqv G`), wherein the first fully-connected layer destroys the symmetry in the noise source, resulting in non-equivariant `G` even if subsequent layers are all equivariant [cf. Remark 4.11]. We use the Fréchet Inception Distance (FID) (Heusel et al., 2017) to evaluate the quality and diversity of the GAN generated samples after embedding them in the feature space of a pre-trained Inception-v3 network (Szegedy et al., 2016). Due to the simplicity of RotMNIST, we replace the inception-featurization by the encoding feature space of an autoencoder trained on the *rotated* digits. We note that, compared to classifiers, autoencoders are guaranteed to produce *different* features for rotated versions of the same digit; they are thus more suitable to measure sample diversity in rotation.



(a) 2D projection of the generated samples.



(b) D_2^L-GANs. (c) WGANs

Figure 4. This figure illustrates how our method can simultaneously handle heavy tails and low-dimensional support. Panel (a): 2D projection of the D_2^L-GAN generated samples onto the support plane of the source Q [cf. Section 5.3]. Each column shows the result after a given number of training epochs. The rows correspond to different settings for the generators (`G`) and discriminators (`D`); in particular, the 2nd and 4th rows use invariant `D` accompanied by, respectively, a baseline `G` and an incorrectly constructed equivariant `G`, leading to mode collapse [cf. Theorem 4.6]. The blue ovals mark the 25% and 50% probability regions of the data source Q, while the heat-map shows the generator samples. Panel (b) and (c): Generator distribution, projected onto components orthogonal to the support plane of Q. Values concentrated around zero indicate convergence to the sub-manifold. Models are trained on **200** training points.

5.3. Toy Example

We test the performance of different GANs (and their equivariant versions) based on 3 types of divergences, namely the Wasserstein-GAN (WGAN) based on the Γ-IPM Eq. (2), the D_{f_α}-GAN based on the classical f-divergence Eq. (3) and (4), and the D_α^L-GAN based on the (f, Γ)-divergence Eq. (8), in learning the C_4-invariant mixture Q. We use fully-connected networks with 3 hidden layers for the baseline `G` and `D` (`Vanilla G&D`). The generator pushes forward a 10D Gaussian noise source, which is itself C_4-invariant after prescribing a proper group action, e.g., $\pi/2$-rotations in the first two dimensions. Equivariant `G` (`Eqv G`) and invariant `D` (`Inv D`) are built by replacing fully-connected

layers with C_4-convolutional layers based on Theorem 4.8 due to the C_4-invariance of the noise source. We also mimic the incomplete attempt by Dey et al. (2021) in building equivariant generators (`(I) Eqv G`) by leaving the first fully-connected layer unchanged and replacing only the subsequent layers by C_4-convolutions.

Figure 4a displays the 2D projection of the generated samples learned by the $D^L_{\alpha=2}$-GAN (and its equivariant versions) on 200 training samples. It is clear that the baseline model without structural prior (`Vanilla G&D`) has difficulty in learning Q in such small data regime. Using an `Inv D` alone without an `Eqv G` (`Vanilla G + Inv D`) or with an incorrectly imposed `Eqv G` (`(I) Eqv G + Inv D`) leads easily to "mode collapse", validating Theorem 4.6. On the other hand, D^L_α-GAN with an `Eqv G` (even without an `Inv D`) is able to learn all 4 modes of Q. We omit the results of (equivariant) D_{f_α}-GANs and WGANs from Figure 4a, as both fail to learn the data source Q; this is unsurprising due to the lack of absolute continuity between Q and P_g (the former is supported on a plane, while the latter is the entire 12D space) and the fact that Q is heavy-tailed (as the mean does not exist.) This demonstrates the importance of our framework's broad applicability to a variety of variational divergences, as an improper choice of the divergence—even with structural prior—can fail to learn the source distribution.

Figure 4 (b) and (c) show the generated distribution projected onto components orthogonal to the support plane of Q. Values concentrated around zero indicate successful learning of the low-dimensional source distribution, i.e., generating high-fidelity samples. Figure 4b indicates that an `Inv D` in the D^L_α-GAN helps produce a distribution with sharper support, whereas `Eqv G` alone without `Inv D` tends to generate relatively low-quality samples away from the supporting plane. In contrast, Figure 4c indicates that WGAN (even with symmetry prior) fails to learn the support plane due to Q being heavy-tailed. Results with different numbers of training samples and α's are shown in Appendix F, and the conclusions are similar.

5.4. RotMNIST

We adopt a similar setup to Dey et al. (2021). Specifically, in the baseline `G`, a fully-connected layer first projects and reshapes the concatenated Gaussian noise and class embedding into a 2D feature map (see Figure 3); spectrally-normalized convolutions (Miyato et al., 2018), interspersed with pointwise-nonlinearities, class-conditional batch-normalizations, and upsamplings, are subsequently used to increase the spatial dimension. We note again that replacing 2D convolutions with C_n-convolutions does not simply lead to `Eqv G`, as the distribution after the "project and reshape" layer is no longer C_n-invariant. This can be fixed by adding a C_n-symmetrization layer after the first linear embedding; see Remark 4.11. We consider GANs with the relative average loss (RA-GANs) (Jolicoeur-Martineau, 2019) in addition to the D^L_α-GANs for this experiment. All configurations are trained with a batch size of 64 for 20,000 generator iterations. Implementation details are available in Appendix G.

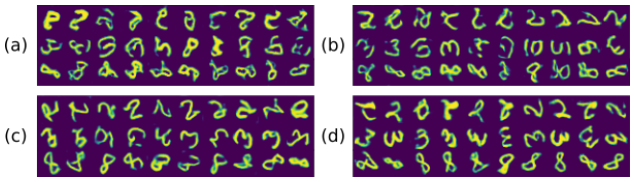


Figure 5. Randomly generated digits 2, 3 and 8 by the RA-GANs trained on RotMNIST after 20K generator iterations and using 1% (600) training data. (a): `CNN G&D`. (b): `(I) Eqv G + Inv D`, $\Sigma = C_4$. (c) & (d): `Eqv G + Inv D`, i.e., our models with correctly constructed equivariant generators. (c): $\Sigma = C_4$. (d): $\Sigma = C_8$. More images are available in Appendix F.

Table 1. The median of the FIDs (lower is better), calculated every 1,000 generator update for 20,000 iterations, averaged over three independent trials. The number of the training samples used for experiments varies from 1% (600) to 10% (6,000) of the entire training set. See Appendix F for further results.

	Architecture	1%	5%	10%
RA-GAN	`CNN G&D`	295	357	348
	`Eqv G + CNN D`, $\Sigma = C_4$	389	333	355
	`CNN G + Inv D`, $\Sigma = C_4$	223	181	188
	`(I) Eqv G+Inv D`, $\Sigma = C_4$	173	141	132
	`Eqv G + Inv D`, $\Sigma = C_4$	**98**	78	89
	`Eqv G + Inv D`, $\Sigma = C_8$	123	**52**	**51**
D^L_2-GAN	`CNN G&D`	280	261	283
	`Eqv G + CNN D`, $\Sigma = C_4$	253	271	251
	`CNN G + Inv D`, $\Sigma = C_4$	330	208	192
	`(I) Eqv G+Inv D`, $\Sigma = C_4$	273	147	133
	`Eqv G + Inv D`, $\Sigma = C_4$	149	99	88
	`Eqv G + Inv D`, $\Sigma = C_8$	**122**	**55**	**57**

Table 1 shows the median of the FIDs, calculated every 1,000 generator update, averaged over three independent trials. It is clear that our proposed models (`Eqv G + Inv D`) consistently achieve significantly improved results compared to the baseline `CNN G&D` and the prior approach (`(I) Eqv G + Inv D`); the out-performance is even more pronounced when increasing the group size from $\Sigma = C_4$ to C_8. We note that, similar to RotMNIST, one can also use a custom autoencoder featurization for FID evaluation, and the superiority of our model (`Eqv G + Inv D`) is even

Table 2. The (min, median) of the FIDs over the course of training, averaged over three independent trials on the medical images, where the plus sign "+" after the data set, e.g., ANHIR+, denotes the presence of data augmentation during training.

Loss	Architecture	ANHIR	ANHIR+
D_2^L	`CNN G&D`	(313, 485)	(347, 539)
	`(I)Eqv G + Inv D`	(120, 176)	(119, 177)
	`Eqv G + Inv D`	**(97, 157)**	**(90, 128)**
Loss	Architecture	LYSTO	LYSTO+
D_2^L	`CNN G&D`	(289, 410)	(265, 376)
	`(I)Eqv G + Inv D`	(253, 343)	(244, 329)
	`Eqv G + Inv D`	**(205, 259)**	**(192, 259)**

more prominent under such metric: for instance, on ANHIR, the median FIDs calculated through autoencoder featurization of the three comparing models are, respectively, 1221 (`CNN G&D`), 936 ((`(I)Eqv G + Inv D`)), and 329 (`Eqv G + Inv D`). See Figure 5 also for randomly generated samples by RA-GANs trained with 1% training data. More results are available in Appendix F.

5.5. ANHIR and LYSTO

Compared to RotMNIST, ResNet and its D_4-equivariant counterpart are used instead of CNNs for `G` and `D`. All models are trained for 40,000 generator iterations with a batch size of 32. Implementation details are available in Appendix G.

Table 2 displays the minimum and median of the FIDs, calculated every 2,000 generator update, averaged over three independent trials. The plus sign "+" after the data set, e.g., ANHIR+, denotes the presence of data augmentation (random 90° rotations and reflection) during training. It is clear that augmentation usually (but not always) has a positive effect on the results evaluated by the FID; however, our proposed model even without data augmentation still consistently and significantly outperforms the baseline model (`CNN G&D`) and the prior approach (`(I)Eqv G + Inv D`) (Dey et al., 2021) with augmentation. Figure 1 presents a random collection of real and generated ANHIR images, visually verifying the improved sample fidelity of our model over the baseline. More results are available in Appendix F.

5.6. Discussion of empirical findings

Consistently across all experiments, our proposed structure-preserving GAN outperforms prior approaches in generating high-fidelity and diverse samples by a significant margin, in some cases almost an order of magnitude measured in FID. The results also show that, compared to data-augmentation (a common strategy for learning from limited data), building theoretically-guided structural probabilistic priors directly into the two GAN players achieves substantially improved performance and data efficiency in adversarial learning.

Acknowledgements

The research of J.B., M.K. and L.R.-B. was partially supported by the Air Force Office of Scientific Research (AFOSR) under the grant FA9550-21-1-0354. The research of M. K. and L.R.-B. was partially supported by the National Science Foundation (NSF) under the grants DMS-2008970 and TRIPODS CISE-1934846. The research of W.Z. was partially supported by NSF under DMS-2052525 and DMS-2140982. We thank Neel Dey for sharing the pre-processed ANHIR data set. This work was performed in part using high performance computing equipment obtained under a grant from the Collaborative R&D Fund managed by the Massachusetts Technology Collaborative.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

Biloš, M. and Günnemann, S. Scalable normalizing flows for permutation invariant densities. In *International Conference on Machine Learning*, pp. 957–967. PMLR, 2021.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Rey-Bellet, L., and Wang, J. Variational representations and neural network estimation of Rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021. doi: 10.1137/20M1368926. URL https://doi.org/10.1137/20M1368926.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Rey-Bellet, L. (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *Journal of Machine Learning Research,* (to appear), 2022. URL https://arxiv.org/abs/2011.05953.

Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D. V., Bueno, G., Khvostikov, A. V., Bakas, S., Eric, I., Chang,

C., Heldmann, S., et al. Anhir: automatic non-rigid histological image registration challenge. *IEEE transactions on medical imaging*, 39(10):3042–3052, 2020.

Bot, R., Grad, S., and Wanka, G. *Duality in Vector Optimization*. Vector Optimization. Springer Berlin Heidelberg, 2009. ISBN 9783642028861.

Boyda, D., Kanwar, G., Racanière, S., Rezende, D. J., Albergo, M. S., Cranmer, K., Hackett, D. C., and Shanahan, P. E. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Broniatowski, M. and Keziou, A. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16 – 36, 2009. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2008.03.011. URL http://www.sciencedirect.com/science/article/pii/S0047259X08001036.

Catoni, O., Euclid, P., Library, C. U., and Press, D. U. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Lecture notes-monograph series. Cornell University Library, 2008. URL https://books.google.gr/books?id=-EtrnQAACAAJ.

Chowdhary, K. and Dupuis, P. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):635–662, 2013. doi: 10.1051/m2an/2012038.

Ciompi, F., Jiao, Y., and van der Laak, J. Lymphocyte assessment hackathon (LYSTO), October 2019. URL https://doi.org/10.5281/zenodo.3513571.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016a.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016b. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Cohen, T. S., Geiger, M., and Weiler, M. A general theory of equivariant CNNs on homogeneous spaces. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b9cfe8b6042cf759dc4c0cccb27a6737-Paper.pdf.

Cohn, D. *Measure Theory*. Birkhäuser Boston, 2013. ISBN 9781489903990. URL https://books.google.com/books?id=rgXyBwAAQBAJ.

Dehmamy, N., Walters, R., Liu, Y., Wang, D., and Yu, R. Automatic symmetry discovery with lie algebra convolutional network. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2503–2515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/148148d62be67e0916a833931bd32b26-Paper.pdf.

Dey, N., Chen, A., and Ghafurian, S. Group equivariant generative adversarial networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rgFNuJHHXv.

Dupuis, P. and Ellis, R. S. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

Dupuis, P., Katsoulakis, M. A., Pantazis, Y., and Plechac, P. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):80–111, 2016. doi: 10.1137/15M1025645.

Durumeric, A. E. and Voth, G. A. Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. *The Journal of chemical physics*, 151(12):124110, 2019.

Feder, R. M., Berger, P., and Stein, G. Nonlinear 3d cosmic web simulation with heavy-tailed generative adversarial networks. *Physical Review D*, 102(10):103504, 2020.

Folland, G. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 9781118626399. URL https://books.google.com/books?id=wI4fAwAAQBAJ.

Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I., and Welling, M. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems*, 34, 2021.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf.

Glaser, P., Arbel, M., and Gretton, A. KALE flow: A relaxed kl gradient flow for probabilities with disjoint support. *arXiv e-prints*, art. arXiv:2106.08929, June 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jolicoeur-Martineau, A. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1erHoR5t7.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kipnis, C. and Landim, C. *Scaling Limits of Interacting Particle Systems*. Springer-Verlag, 1999.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: sampling configurations for multi-body systems with symmetric energies. *arXiv preprint arXiv:1910.00753*, 2019.

Köhler, J., Klein, L., and Noé, F. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International Conference on Machine Learning*, pp. 5361–5370. PMLR, 2020.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, W., Burkhart, C., Polińska, P., Harmandaris, V., and Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *The Journal of Chemical Physics*, 153(4):041101, 2020.

Liu, J., Kumar, A., Ba, J., Kiros, J., and Swersky, K. Graph normalizing flows. *arXiv preprint arXiv:1905.13177*, 2019.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pp. 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674. doi: 10.1145/307400.307435. URL https://doi.org/10.1145/307400.307435.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Mustafa, M., Bard, D., Bhimji, W., Lukić, Z., Al-Rfou, R., and Kratochvil, J. M. CosmoGAN: creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks. *Computational Astrophysics and Cosmology*, 6(1):1, December 2019. ISSN 2197-7909. doi: 10.1186/s40668-019-0029-9. URL https://comp-astrophys-cosmol.springeropen.com/articles/10.1186/s40668-019-0029-9.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997. doi: 10.2307/1428011.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Nonparametric estimation of the likelihood ratio and divergence functionals. In *2007 IEEE International Symposium on Information Theory*, pp. 2016–2020, 2007.

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

Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of Chemical Physics*, 139(9):090901, 2013. doi: 10.1063/1.4818908. URL https://doi.org/10.1063/1.4818908.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 271–279, 2016.

Pak, A. J. and Voth, G. A. Advances in coarse-grained modeling of macromolecular complexes. *Current Opinion in Structural Biology*, 52:119–126, 2018. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2018.11.005. URL https://www.sciencedirect.com/science/article/pii/S0959440X18300939. Cryo electron microscopy: the impact of the cryo-EM revolution in biology • Biophysical and computational methods - Part A.

Rezende, D. J., Racanière, S., Higgins, I., and Toth, P. Equivariant hamiltonian flows. *arXiv preprint arXiv:1909.13739*, 2019.

Ruderman, A., Reid, M. D., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 1155–1162, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Rudin, W. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 2006. ISBN 9780070619883.

Schindler, W. *Measures with Symmetry Properties*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2003. ISBN 9783540362104. URL https://books.google.com/books?id=xyt8CwAAQBAJ.

Shawe-Taylor, J. and Williamson, R. C. A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, COLT '97, pp. 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916. doi: 10.1145/267460.267466. URL https://doi.org/10.1145/267460.267466.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722. URL https://doi.org/10.1214/12-EJS722.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008. ISBN 9780387772424. URL https://books.google.com/books?id=HUnqnrpYt4IC.

Stieffenhofer, M., Bereau, T., and Wand, M. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. *APL Materials*, 9(3):031107, 2021. doi: 10.1063/5.0039102. URL https://doi.org/10.1063/5.0039102.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Weiler, M. and Cesa, G. General E(2)-equivariant steerable CNNs. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/45d6637b718d0f24a237069fe41b0db4-Paper.pdf.

Yi, X., Walia, E., and Babyn, P. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

Zhao, S., Liu, Z., Lin, J., Zhu, J.-Y., and Han, S. Differentiable augmentation for data-efficient gan training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7559–7570. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/55479c55ebd1efd3ff125f1337100388-Paper.pdf.

Zhu, M., Pan, P., Chen, W., and Yang, Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

A. More details on variational representations of divergences and probability metrics

We provide, in this appendix, more details on variational representations of the divergences and probability metrics discussed in Section 3.2. Recall the notation introduced in the main paper: let $(X, \mathcal{M})$ be a measurable space, $\mathcal{M}(X)$ be the space of measurable functions on X, and $\mathcal{M}_b(X)$ be the subspace of bounded measurable functions. We denote $\mathcal{P}(X)$ as the set of probability measures on X. Given an objective functional $H : \mathcal{M}^n(X) \times \mathcal{P}(X) \times \mathcal{P}(X) \to [-\infty, \infty]$ and a test function space $\Gamma \subset \mathcal{M}(X)^n, n \in \mathbb{Z}^+$, we define

$$D_H^\Gamma(Q\|P) = \sup_{\gamma \in \Gamma} H(\gamma; Q, P)\,. \tag{20}$$

D_H^Γ is called a *divergence* if $D_H^\Gamma \geq 0$ and $D_H^\Gamma(Q\|P) = 0$ if and only if $Q = P$, hence providing a notion of "distance" between probability measures. D_H^Γ is further called a *probability metric* if it satisfies the triangle inequality (i.e., $D_H^\Gamma(Q\|P) \leq D_H^\Gamma(Q\|\nu) + D_H^\Gamma(\nu\|P)$ for all $Q, P, \nu \in \mathcal{P}(X)$) and is symmetric (i.e., $D_H^\Gamma(Q\|P) = D_H^\Gamma(P\|Q)$ for all $P, Q \in \mathcal{P}(X)$). It is well known that formula (20) includes, through suitable choices of objective functional $H(\gamma; Q, P)$ and function space Γ, many divergences and probability metrics. Below we further elaborate on the examples discussed in Section 3.2.

(a) f-divergences. Let $f : [0, \infty) \to \mathbb{R}$ be convex and lower semi-continuous (LSC), with $f(1) = 0$ and f strictly convex at $x = 1$. The f-divergence between Q and P can be defined based on two equivalent variational representations (Birrell et al., 2022), namely

$$D_f(Q\|P) = \sup_{\gamma \in \mathcal{M}_b(X)} \{E_Q[\gamma] - E_P[f^*(\gamma)]\} \tag{21}$$

$$= \sup_{\gamma \in \mathcal{M}_b(X)} \{E_Q[\gamma] - \Lambda_f^P[\gamma]\}\,, \tag{22}$$

where f^* in the first representation (21) denotes the Legendre transform (LT) of f,

$$f^*(y) = \sup_{x \in \mathbb{R}} \{yx - f(x)\}, \quad \forall y \in \mathbb{R}, \tag{23}$$

and $\Lambda_f^P[\gamma]$ in the second representation (22) is defined as

$$\Lambda_f^P[\gamma] \coloneqq \inf_{\nu \in \mathbb{R}} \{\nu + E_P[f^*(\gamma - \nu)]\}\,, \quad \gamma \in \mathcal{M}_b(\Omega)\,. \tag{24}$$

The two variational representations Eq. (21) and Eq. (22) share the same $\Gamma = \mathcal{M}_b(X)$, and their equivalence is due to $\mathcal{M}_b(\Omega)$ being closed under the shift map $\gamma \mapsto \gamma - \nu$ for $\nu \in \mathbb{R}$. Examples of the f-divergences include the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), the total variation distance, the χ^2-divergence, the Hellinger distance, the Jensen-Shannon divergence, and the family of α-divergences (Nowozin et al., 2016). For instance, the KL-divergence is constructed from

$$f_{KL} = x \log x, \quad \forall x \geq 0. \tag{25}$$

A key element in the second variational representation for D_f [Eq. (22)] is the functional $\Lambda_f^P[\gamma]$, which is a generalization of the cumulant generating function from the KL-divergence case to the f-divergence case. Indeed, for the KL-divergence where $f(x) = f_{KL}(x) = x \log x$, it is straightforward to show that Λ_f^P becomes the standard cumulant generating function, $\Lambda_{f_{KL}}^P[\gamma] = \log E_P[e^\gamma]$, and Eq. (22) becomes the Donsker-Varadhan variational formula; see Appendix C.2 in (Dupuis & Ellis, 2011). The flexibility of f allows one to tailor the divergence to the data source, e.g., for heavy tailed data. Moreover, the strict concavity of f in γ can result in improved statistical learning, estimation, and convergence performance. However, the variational representations (21) and (22) both result in $D_f(Q\|P) = \infty$ if Q is not absolutely continuous with respect to P, limiting their efficacy in comparing distributions with low-dimensional support.

(b) Γ-Integral Probability Metrics (IPMs). Given $\Gamma \subset \mathcal{M}_b(X)$, the Γ-IPM between Q and P is defined as

$$W^\Gamma(Q, P) = \sup_{\gamma \in \Gamma} \{E_Q[\gamma] - E_P[\gamma]\}. \tag{26}$$

We refer to (Müller, 1997; Sriperumbudur et al., 2012) for a complete theory and conditions on Γ ensuring that $W^\Gamma(Q,P)$ is a metric. Apart from the Wasserstein metric when $\Gamma = \mathrm{Lip}^1(X)$ is the space of 1-Lipschitz functions, examples of IPMs also include: the total variation metric, where Γ is the unit ball in $\mathcal{M}_b(X)$; the Dudley metric, where Γ is the unit ball in the space of bounded and Lipschitz continuous functions; and maximum mean discrepancy (MMD), where Γ is the unit ball in an RKHS (Müller, 1997; Sriperumbudur et al., 2012). With suitable choices of Γ, IPMs are able to meaningfully compare not-absolutely continuous distributions, but they could potentially fail at comparing distributions with heavy tails (Birrell et al., 2022).

(c) (f,Γ)-divergences. This class of divergences were introduced in (Birrell et al., 2022) and they subsume both f-divergences and Γ-IPMs. Given a function f satisfying the same condition as in the definition of the f-divergence and $\Gamma \subset \mathcal{M}_b(X)$, the (f,Γ)-divergence is defined as

$$D_f^\Gamma(Q\|P) = \sup_{\gamma\in\Gamma} \left\{ E_Q[\gamma] - \Lambda_f^P[\gamma] \right\}, \tag{27}$$

where $\Lambda_f^P[\gamma]$ is again given by Eq. (24), implying that Eq. (6) includes as a special case the f-divergence (3) when $\Gamma = \mathcal{M}_b(X)$ and the $\Gamma \subset \mathcal{M}_b(X)$ implies

$$D_f^\Gamma(Q\|P) \le D_f(Q\|P) \tag{28}$$

for any $\Gamma \subset \mathcal{M}_b(X)$. It is demonstrated in (Birrell et al., 2022) that one also has

$$D_f^\Gamma(Q\|P) \le W^\Gamma(Q,P)\,. \tag{29}$$

Some notable examples of such Γ's can be found in (Birrell et al., 2022), for instance the 1-Lipschitz functions $\mathrm{Lip}^1(X)$, the RKHS unit ball, ReLU neural networks, ReLU neural networks with spectral normalizations, etc. The property (29) readily implies that (f,Γ) divergences can be defined for non-absolutely continuous probability distributions. If X is further assumed to be a complete separable metric space then, under stronger assumptions on f and Γ, one has the following Infimal Convolution Formula:

$$D_f^\Gamma(Q\|P) = \inf_{\eta\in\mathcal{P}(X)} \left\{ D_f(\eta\|P) + W^\Gamma(Q,\eta) \right\}, \tag{30}$$

which implies, in particular, $0 \le D_f^\Gamma(Q\|P) \le \min\{D_f(Q\|P), W^\Gamma(Q,P)\}$, i.e., Eq. (28) and Eq. (29).

(d) Sinkhorn divergences. The Wasserstein (or "earth-mover") metric associated with a cost function $c : X \times X \to \mathbb{R}^+$ has the variational representation

$$W_c^\Gamma(Q,P) = \inf_{\pi\in\mathrm{Co}(Q,P)} E_\pi[c(x,y)] = \sup_{\gamma=(\gamma_1,\gamma_2)\in\Gamma} \{E_P[\gamma_1] + E_Q[\gamma_2]\}, \tag{31}$$

where $\mathrm{Co}(Q,P)$ is the set of all couplings of P and Q and $\Gamma = \{\gamma = (\gamma_1,\gamma_2) \in C(X)\times C(X) : \gamma_1(x) + \gamma_2(y) \le c(x,y)\,, x,y \in X\}$, with $C(X)$ being the space of continuous functions on X ($C_b(X)$ will denote the subspace of bounded continuous functions). The Sinkhorn divergence is given by

$$\mathcal{SD}_{c,\epsilon}^\Gamma(Q,P) = W_{c,\epsilon}^\Gamma(Q,P) - \frac{1}{2}W_{c,\epsilon}^\Gamma(Q,Q) - \frac{1}{2}W_{c,\epsilon}^\Gamma(P,P), \tag{32}$$

with $W_{c,\epsilon}^\Gamma(Q,P)$ being the entropic regularization of the Wasserstein metrics (Genevay et al., 2016),

$$W_{c,\epsilon}^\Gamma(Q,P) = \inf_{\pi\in\mathrm{Co}(Q,P)} \{E_\pi[c(x,y)] + \epsilon R(\pi\|P\times Q)\} \tag{33}$$

$$= \sup_{\gamma=(\gamma_1,\gamma_2)\in\Gamma} \left\{ E_P[\gamma_1] + E_Q[\gamma_2] - \epsilon E_{P\times Q}\left[\exp\left(\frac{\gamma_1\oplus\gamma_2 - c}{\epsilon}\right)\right] + \epsilon \right\}, \tag{34}$$

where now $\Gamma = C_b(X)\times C_b(X)$ and $\gamma_1\oplus\gamma_2(x,y) \coloneqq \gamma_1(x) + \gamma_2(y)$.

B. Proofs

In this appendix we provide proofs of results that were stated in the main text. First we prove the properties of the symmetrization operators from Lemma 3.2.

Lemma B.1. *(a) The symmetrization operator $S_\Sigma : \mathcal{M}_b(X) \to \mathcal{M}_b(X)$ is a projection operator onto the subspace of Σ-invariant bounded measurable functions*

$$\mathcal{M}_{b,\Sigma}^{inv}(X) \coloneqq \{\gamma \in \mathcal{M}_b(X) : \gamma \circ T_\sigma = \gamma \ \ \textit{for all} \ \ \sigma \in \Sigma\}\,, \tag{35}$$

in the sense that

1. $S_\Sigma[\mathcal{M}_b(X)] = \mathcal{M}_{b,\Sigma}^{inv}(X)$,
2. $S_\Sigma \circ S_\Sigma = S_\Sigma$.

Moreover,

$$S_\Sigma[\gamma \circ T_\sigma] = S_\Sigma[\gamma] \tag{36}$$

for all $\gamma \in \mathcal{M}_b(X)$, $\sigma \in \Sigma$.

(b) The symmetrization operator $S^\Sigma : \mathcal{P}(X) \to \mathcal{P}(X)$ is a projection operator onto the subset of Σ-invariant probability measures

$$\mathcal{P}_\Sigma(X) \coloneqq \{P \in \mathcal{P}(X) : P \circ T_\sigma^{-1} = P \ \ \textit{for all} \ \ \sigma \in \Sigma\}\,, \tag{37}$$

in the sense that

1. $S^\Sigma[\mathcal{P}(X)] = \mathcal{P}_\Sigma(X)$,
2. $S^\Sigma \circ S^\Sigma = S^\Sigma$.

(c) S_Σ is the conditional expectation operator with respect to the σ-algebra of Σ-invariant sets. More specifically, for all $\gamma \in \mathcal{M}_b(X)$, $P \in \mathcal{P}_\Sigma(X)$ we have

$$S_\Sigma[\gamma] = E_P[\gamma|\mathcal{M}_\Sigma]\,. \tag{38}$$

where $\mathcal{M}_\Sigma$ is the σ-algebra of Σ-invariant sets,

$$\mathcal{M}_\Sigma \coloneqq \{\textit{Measurable sets } B \subset X : T_\sigma(B) = B \ \ \textit{for all } \sigma \in \Sigma\}\,. \tag{39}$$

Proof. We will need the following invariance property of integrals with respect to Haar measure, which can be proven using the invariance of Haar measure under left and right group multiplication:

$$\int_\Sigma h(\sigma \cdot \sigma')d\mu_\Sigma(\sigma') = \int_\Sigma h(\sigma' \cdot \sigma)d\mu_\Sigma(\sigma') = \int_\Sigma h(\sigma')d\mu_\Sigma(\sigma')\,. \tag{40}$$

(a) If $\gamma \in \mathcal{M}_b(X)$ then $\gamma' = S_\Sigma[\gamma] \in \mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$ by applying (40) with $h(\sigma) \coloneqq \gamma \circ T_\sigma(x)$, $x \in X$. Indeed we have

$$\gamma' \circ T_\sigma(x) = \int \gamma(T_{\sigma'}(T_\sigma(x)))d\mu_\Sigma(\sigma') = \int h(\sigma' \cdot \sigma)\mu_\Sigma(d\sigma') = \int h(\sigma')\mu_\Sigma(d\sigma') = \gamma'(x)\,.$$

Furthermore any $\gamma \in \mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$ belongs to the range of S_Σ since $\gamma \circ T_\sigma = \gamma$ for all $\sigma \in \Sigma$ implies that $\gamma = S_\Sigma[\gamma]$. This also shows that $S_\Sigma \circ S_\Sigma = S_\Sigma$. Finally, for $\gamma \in \mathcal{M}_b(X)$, $\sigma \in \Sigma$, $x \in X$ we can compute

$$S_\Sigma[\gamma \circ T_\sigma](x) = \int \gamma(T_{\sigma\cdot\sigma'}(x))\mu_\Sigma(d\sigma') = \int \gamma(T'_\sigma(x))\mu_\Sigma(d\sigma') = S_\Sigma[\gamma](x)\,,$$

where we again used the invariance property of integrals with respect to Haar measure (40).

(b) For $P \in \mathcal{P}(X)$, $\gamma \in \mathcal{M}_b(X)$, and $\sigma \in \Sigma$ we can use (36) to compute

$$\int \gamma dS^\Sigma[P] \circ T_\sigma^{-1} = \int \gamma \circ T_\sigma dS^\Sigma[P] = \int S_\Sigma[\gamma \circ T_\sigma] dP = \int S_\Sigma[\gamma] dP = \int \gamma dS^\Sigma[P] \,.$$

This holds for all $\gamma \in \mathcal{M}_b(X)$, hence $S^\Sigma[P] \circ T_\sigma^{-1} = S^\Sigma[P]$ for all $\sigma \in \Sigma$. Therefore $S^\Sigma[P] \in \mathcal{P}_\Sigma(X)$. Conversely, if $P \in \mathcal{P}_\Sigma(X)$ then $E_P[\gamma \circ T_\sigma] = E_P[\gamma]$ for all $\sigma \in \Sigma$ and $\gamma \in \mathcal{M}_b(X)$ and thus, by Fubini's theorem, $E_P[S_\Sigma[\gamma]] = E_P[\gamma]$. Hence $S^\Sigma[P] = P$ and so $P \in S^\Sigma[\mathcal{P}]$. This completes the proof that $S^\Sigma[\mathcal{P}(X)] = \mathcal{P}_\Sigma(X)$. Combining these calculations it is also clear that $S^\Sigma \circ S^\Sigma = S^\Sigma$.

(c) Let $\gamma \in \mathcal{M}_b(X)$ and $P \in \mathcal{P}_\Sigma(X)$. From part (a) we know that $S_\Sigma[\gamma] \in \mathcal{M}_{b,\Sigma}^{\text{inv}}(X)$ and from this it is straightforward to show that $S_\Sigma[\gamma]$ is $\mathcal{M}_\Sigma$-measurable. Now fix $A \in \mathcal{M}_\Sigma$ and note that $1_A \circ T_\sigma = 1_A$ for all $\sigma \in \Sigma$ (where 1_A denotes the indicator function for A). Using this fact together with $S^\Sigma[P] = P$ (see part (b)) we can compute

$$\begin{aligned}\int S_\Sigma[\gamma] 1_A dP &= \int \int \gamma \circ T_{\sigma'} 1_A \mu_\Sigma(d\sigma') dP = \int \int (\gamma 1_A) \circ T_{\sigma'} \mu_\Sigma(d\sigma') dP = \int S_\Sigma[\gamma 1_A] dP = \int \gamma 1_A dS^\Sigma[P] \\ &= \int \gamma 1_A dP \,.\end{aligned}$$

This proves $S_\Sigma[\gamma] = E_P[\gamma|\mathcal{M}_\Sigma]$ by the definition of conditional expectation. □

Now we prove Theorem 4.1.

Theorem B.2. *If $S_\Sigma[\Gamma] \subset \Gamma$ and the probability measures P, Q are Σ-invariant then*

$$D^\Gamma(Q\|P) = D^{\Gamma_\Sigma^{inv}}(Q\|P) \,, \tag{41}$$

where D^Γ is an (f, Γ)-divergence or a Γ-IPM. Eq. (41) also holds for Sinkhorn divergences if the cost is Σ-invariant (i.e., $c(T_\sigma(x), T_\sigma(y)) = c(x, y)$ for all $\sigma \in \Sigma$, $x, y \in X$).

Remark B.3. Note that the classical Sinkhorn divergence is obtained when $\Gamma = C_b(X) \times C_b(X)$ but the proof of this theorem applies to any $\Gamma \subset \mathcal{M}_b(X)^2$ with $S_\Sigma[\Gamma] \subset \Gamma$.

Proof. We first prove the Theorem for (f, Γ)-divergences. Start by using Jensen's inequality and the convexity of the Legendre transform f^* to obtain

$$\begin{aligned}&f^*(S_\Sigma[\gamma](x) - \nu) = f^*\left(\int \big(\gamma(T_\sigma(x)) - \nu\big)\mu_\Sigma(d\sigma)\right) \\ \leq &\int f^*(\gamma(T_\sigma(x)) - \nu)\mu_\Sigma(d\sigma) = S_\Sigma[f^*(\gamma(x) - \nu)]\end{aligned}$$

for all $\gamma \in \mathcal{M}_b(X)$. Therefore

$$\begin{aligned}&D_f^{S_\Sigma[\Gamma]}(Q\|P) \\ = &\sup_{\gamma \in \Gamma, \nu \in \mathbb{R}} \{E_Q[S_\Sigma[\gamma]] - \nu - E_P[f^*(S_\Sigma[\gamma] - \nu)]\} \\ \geq &\sup_{\gamma \in \Gamma, \nu \in \mathbb{R}} \{E_Q[S_\Sigma[\gamma] - \nu] - E_P[S_\Sigma[f^*(\gamma - \nu)]]\} \\ = &\sup_{\gamma \in \Gamma, \nu \in \mathbb{R}} \{E_Q[\gamma] - \nu - E_P[f^*(\gamma - \nu)]\} = D_f^\Gamma(Q\|P) \,,\end{aligned}$$

where in the next to last equality we use Lemma 3.2(c) together with the assumptions $P, Q \in \mathcal{P}_\Sigma(X)$ to conclude $E_P[S_\Sigma[f^*(\gamma - \nu)]] = E_P[f^*(\gamma - \nu)]$ and $E_Q[S_\Sigma[\gamma]] = E_Q[\gamma]$. Hence we obtain $D_f^\Gamma(Q\|P) \leq D_f^{S_\Sigma[\Gamma]}(Q\|P)$. Furthermore, since $S_\Sigma[\Gamma] \subset \Gamma$, we have from (6) that $D_f^{S_\Sigma[\Gamma]}(Q\|P) = D_f^\Gamma(Q\|P)$. We conclude by showing that $S_\Sigma[\Gamma] \subset \Gamma$ implies $S_\Sigma[\Gamma] = \Gamma_\Sigma^{\text{inv}}$. First, if $\gamma \in \Gamma_\Sigma^{\text{inv}}$, then $S_\Sigma[\gamma] = \gamma$, therefore $\Gamma_\Sigma^{\text{inv}} \subset S_\Sigma[\Gamma]$. Conversely, since $\Gamma \subset \mathcal{M}_b(X)$, the functions in $S_\Sigma[\Gamma]$ are Σ-invariant (see Lemma 3.2). We assumed $S_\Sigma[\Gamma] \subset \Gamma$, hence $S_\Sigma[\Gamma] \subset \Gamma_\Sigma^{\text{inv}}$.

The proof for Γ-IPMs is similar, but does not require Jensen's inequality due to the linearity of the objective functional in γ. Hence the hypothesis $S_\Sigma[\Gamma] \subset \Gamma$ is not necessary to obtain $W^\Gamma(Q,P) = W^{S_\Sigma[\Gamma]}(Q,P)$.

Finally, we prove the result for Sinkhorn divergences. Equation (32) implies that it suffices to show $W^\Gamma_{c,\epsilon}(Q,P) = W^{\Gamma^{\text{inv}}_\Sigma}_{c,\epsilon}(Q,P)$: By the same reasoning used for (f,Γ)-divergences, our assumptions imply $\Gamma^{\text{inv}}_\Sigma = S_\Sigma[\Gamma]$ and therefore

$$\begin{aligned}
&W^{\Gamma^{\text{inv}}_\Sigma}_{c,\epsilon}(Q,P)\\
=&W^{S_\Sigma[\Gamma]}_{c,\epsilon}(Q,P) = \sup_{(\gamma_1,\gamma_2)\in\Gamma}\left\{E_P[S_\Sigma[\gamma_1]] + E_Q[S_\Sigma[\gamma_2]] - \epsilon E_{P\times Q}\left[\exp\left(\frac{S_\Sigma[\gamma_1]\oplus S_\Sigma[\gamma_2]-c}{\epsilon}\right)\right]+\epsilon\right\}\\
=&\sup_{(\gamma_1,\gamma_2)\in\Gamma}\left\{E_{S^\Sigma[P]}[\gamma_1] + E_{S^\Sigma[Q]}[\gamma_2] - \epsilon E_{P\times Q}\left[\exp\left(\frac{\int \gamma_1(T_\sigma(x))+\gamma_2(T_\sigma(y))-c(x,y)\mu_\Sigma(d\sigma)}{\epsilon}\right)\right]+\epsilon\right\}.
\end{aligned}$$

Using Jensen's inequality followed by Fubini's theorem on the third term we obtain

$$\begin{aligned}
&W^{\Gamma^{\text{inv}}_\Sigma}_{c,\epsilon}(Q,P)\\
\geq&\sup_{(\gamma_1,\gamma_2)\in\Gamma}\left\{E_{S^\Sigma[P]}[\gamma_1] + E_{S^\Sigma[Q]}[\gamma_2] - \epsilon\int E_{P\times Q}\left[\exp\left(\frac{\gamma_1(T_\sigma(x))+\gamma_2(T_\sigma(y))-c(x,y)}{\epsilon}\right)\right]\mu_\Sigma(d\sigma)+\epsilon\right\}.
\end{aligned}$$

Finally, the Σ-invariance of Q, P, and c imply $S^\Sigma[P]=P$, $S^\Sigma[Q]=Q$, and

$$\begin{aligned}
&\int E_{P\times Q}\left[\exp\left(\frac{\gamma_1(T_\sigma(x))+\gamma_2(T_\sigma(y))-c(x,y)}{\epsilon}\right)\right]\mu_\Sigma(d\sigma)\\
=&\int E_{P\times Q}\left[\exp\left(\frac{\gamma_1(T_\sigma(x))+\gamma_2(T_\sigma(y))-c(T_\sigma(x),T_\sigma(y))}{\epsilon}\right)\right]\mu_\Sigma(d\sigma)\\
=&\int\int\int \exp\left(\frac{\gamma_1(x)+\gamma_2(y)-c(x,y)}{\epsilon}\right) Q\circ T_\sigma^{-1}(dx)P\circ T_\sigma^{-1}(dy)\mu_\Sigma(d\sigma)\\
=&\int\int \exp\left(\frac{\gamma_1(x)+\gamma_2(y)-c(x,y)}{\epsilon}\right) Q(dx)P(dy).
\end{aligned}$$

Therefore

$$W^{\Gamma^{\text{inv}}_\Sigma}_{c,\epsilon}(Q,P) \geq \sup_{(\gamma_1,\gamma_2)\in\Gamma}\left\{E_P[\gamma_1]+E_Q[\gamma_2]-\epsilon E_{P\times Q}\left[\exp\left(\frac{\gamma_1\oplus\gamma_2-c}{\epsilon}\right)\right]+\epsilon\right\} = W^\Gamma_{c,\epsilon}(Q,P).$$

The reverse inequality follows from $\Gamma^{\text{inv}}_\Sigma \subset \Gamma$ and so the proof is complete. □

Next we prove Theorem 4.3, a generalization of Theorem 4.1.

Theorem B.4. *Let $K_x(dx')$ be a probability kernel from X to X and define $S_K : \mathcal{M}_b(X) \mapsto \mathcal{M}_b(X)$ by $S_K[f](x) = \int f(x')K_x(dx')$. K also defines a dual map $S^K : \mathcal{P}(X)\to\mathcal{P}(X)$, $S^K[P] \coloneqq \int K_x(\cdot)P(dx)$. Let $\mathcal{P}_K(X)$ be the set of K-invariant probability measures, i.e., $\mathcal{P}_K(X) = \{P\in\mathcal{P}(X) : S^K[P]=P\}$.*

If $\Gamma\subset\mathcal{M}_b(X)$ such that $S_K[\Gamma]\subset\Gamma$ and $Q,P\in\mathcal{P}_K(X)$ then

$$D^\Gamma(Q\|P) = D^{S_K[\Gamma]}(Q\|P), \tag{42}$$

where D^Γ is an (f,Γ)-divergence or a Γ-IPM. It also holds for the Sinkhorn divergence if $S_K[c(\cdot,y)] = c(\cdot,y)$ and $S_K[c(x,\cdot)] = c(x,\cdot)$ for all $x,y\in X$.

In addition, if S_K is a projection (i.e., $S_K\circ S_K = S_K$) then $S_K[\Gamma] = \Gamma^{inv}_K$ where where $\Gamma^{inv}_K \coloneqq \{\gamma\in\Gamma : S_K[\gamma]=\gamma\}$.

Proof. We prove (42) for (f,Γ)-divergences. The proof for Γ-IPMs and Sinkhorn divergences are similar. We note that for Γ-IPMs, (42) does not require the assumption $S_K[\Gamma]\subset\Gamma$.

Fix $Q, P \in \mathcal{P}_K(X)$ and use Jensen's inequality along with the K-invariance of Q and P to compute

$$\begin{aligned}
D_f^{S_K[\Gamma]}(Q\|P) &= \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_Q[S_K[\gamma]-\nu] - E_P[f^*(S_K[\gamma]-\nu)]\} \\
&= \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_Q[S_K[\gamma-\nu]] - E_P[f^*(\int(\gamma(x')-\nu)K_x(dx'))]\} \\
&\geq \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_Q[S_K[\gamma-\nu]] - E_P[\int f^*(\gamma(x')-\nu)K_x(dx'))]\} \\
&= \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_{S^K[Q]}[\gamma-\nu] - E_{S^K[P]}[f^*(\gamma-\nu)]\} \\
&= \sup_{\gamma\in\Gamma,\nu\in\mathbb{R}} \{E_Q[\gamma-\nu] - E_P[f^*(\gamma-\nu)]\} = D_f^{\Gamma}(Q\|P)\,.
\end{aligned}$$

Therefore $D_f^{S_K[\Gamma]}(Q\|P) \geq D_f^{\Gamma}(Q\|P)$. Note that this computation is the same as the proof of the data processing inequality for (f,Γ)-divergences; see Theorem 2.21 in (Birrell et al., 2022). The assumption $S_K[\Gamma] \subset \Gamma$ implies the reverse inequality, hence we conclude $D_f^{S_K[\Gamma]}(Q\|P) = D_f^{\Gamma}(Q\|P)$.

Now suppose $S_K \circ S_K = S_K$. If $\gamma = S_K[\gamma'] \in S_K[\Gamma]$ then $S_K[\gamma] = S_K[S_K[\gamma']] = S_K[\gamma'] = \gamma$. This, together with the assumption that $S_K[\Gamma] \subset \Gamma$ implies $\gamma \in \Gamma_K^{\text{inv}}$. Conversely, if $\gamma \in \Gamma_K^{\text{inv}}$ then $\gamma = S_K[\gamma] \in S_K[\Gamma]$ by the definition of Γ_K^{inv}. This completes the proof. □

We now prove Theorem 4.6, which explains the potential "mode collapse" in GANs when restricting the test function space from Γ to $\Gamma_\Sigma^{\text{inv}}$ if at least one of the distributions Q and P is not Σ-invariant.

Theorem B.5. *Suppose $S_\Sigma[\Gamma] \subset \Gamma$ and $P, Q \in \mathcal{P}(X)$ (i.e., not necessarily Σ-invariant). Then*

$$D_f^{\Gamma_\Sigma^{inv}}(Q\|P) = D_f^{\Gamma}(S^\Sigma[Q]\|S^\Sigma[P])\,, \tag{43}$$

$$W^{\Gamma_\Sigma^{inv}}(Q,P) = W^{\Gamma}(S^\Sigma[Q], S^\Sigma[P])\,. \tag{44}$$

Remark B.6. The analogous result for the Sinkhorn divergences also holds if the cost is separately Σ-invariant in each variable, i.e., $c(T_\sigma(x), y) = c(x,y)$ and $c(x, T_\sigma(y)) = c(x,y)$ for all $\sigma \in \Sigma$, $x, y \in X$. Though this is not satisfied by most commonly used cost functions and actions one can always enforce it by replacing the cost function c with the symmetrized cost

$$c_\Sigma(x,y) \coloneqq \int\int c(T_\sigma(x), T_{\sigma'}(y))\mu_\Sigma(d\sigma)\mu_\Sigma(d\sigma')\,. \tag{45}$$

Proof. We prove only the validity of (43); the proof of (44) is similar.

$$\begin{aligned}
D_f^{\Gamma}(S^\Sigma[Q]\|S^\Sigma[P]) &= D_f^{\Gamma_\Sigma^{\text{inv}}}(S^\Sigma[Q]\|S^\Sigma[P]) \\
&= \sup_{\gamma\in\Gamma_\Sigma^{\text{inv}},\nu\in\mathbb{R}} \left\{E_{S^\Sigma[Q]}[\gamma-\nu] - E_{S^\Sigma[P]}[f^*(\gamma-\nu)]\right\} \\
&= \sup_{\gamma\in\Gamma_\Sigma^{\text{inv}},\nu\in\mathbb{R}} \{E_Q[\gamma-\nu] - E_P[f^*(\gamma-\nu)]\} \\
&= D_f^{\Gamma_\Sigma^{\text{inv}}}(Q\|P)\,,
\end{aligned}$$

where the first equality is due to Theorem 4.1, and the third equality holds as $\gamma - \nu$ and $f^*(\gamma-\nu)$ are both Σ-invariant when $\gamma \in \Gamma_\Sigma^{\text{inv}}$. □

Next we prove Theorem 4.8, which explains how to ensure the generator produces a Σ-invariant distribution P_g

Theorem B.7. *If $P_Z \in \mathcal{P}(Z)$ is Σ-invariant and $g : Z \to X$ is Σ-equivariant then the push-forward measure $P_g \coloneqq P_Z \circ g^{-1}$ is Σ-invariant, i.e., $P_g \in \mathcal{P}_\Sigma(X)$.*

Proof. The proof is based on the equivalence of the following commutative diagrams:

$$\begin{array}{ccc} Z & \xrightarrow{g} & X \\ {\scriptstyle T_\sigma^Z}\downarrow & & \downarrow{\scriptstyle T_\sigma^X} \\ Z & \xrightarrow[g]{} & X \end{array} \iff \begin{array}{ccc} \mathcal{P}(Z) & \xrightarrow{\circ g^{-1}} & \mathcal{P}(X) \\ {\scriptstyle \circ (T_\sigma^Z)^{-1}}\downarrow & & \downarrow{\scriptstyle \circ (T_\sigma^X)^{-1}} \\ \mathcal{P}(Z) & \xrightarrow[\circ g^{-1}]{} & \mathcal{P}(X) \end{array} \tag{46}$$

More specifically,

$$\begin{aligned} &P_g \circ (T_\sigma^X)^{-1} = P_Z \circ g^{-1} \circ (T_\sigma^X)^{-1} = P_Z \circ (T_\sigma^X \circ g)^{-1} \\ =&P_Z \circ (g \circ T_\sigma^Z)^{-1} = P_Z \circ (T_\sigma^Z)^{-1} \circ g^{-1} = P_Z \circ g^{-1} \\ =&P_g \,, \end{aligned}$$

where the third and fifth equalities are due to the equivariance and invariance, respectively, of g and P_Z. □

Next we prove Theorem 4.10, which provides a method for constructing Σ-invariant noise sources.

Theorem B.8. *Let $W \sim \mu_\Sigma$ and N be a Z-valued random variable (i.e., an arbitrary noise source). If W and N are independent then the distribution of $T^Z(W, N)$ is Σ-invariant.*

Proof. Let P_Z denote the distribution of N. Independence of W and N implies $(W, N) \sim \mu_\Sigma \times P_Z$. Therefore $T^Z(W, N) \sim (\mu_\Sigma \times P_Z) \circ (T^Z)^{-1} \coloneqq P_Z^\Sigma$. We need to show that P_Z^Σ is Σ-invariant: For $\sigma \in \Sigma$ we can compute

$$\begin{aligned} P_Z^\Sigma \circ (T_\sigma^Z)^{-1} =&(\mu_\Sigma \times P_Z) \circ (T^Z)^{-1} \circ (T_\sigma^Z)^{-1} \\ =&(\mu_\Sigma \times P_Z) \circ (T_\sigma^Z \circ T^Z)^{-1} \\ =&(\mu_\Sigma \times P_Z) \circ (T^Z \circ (T_\sigma^\Sigma \times id))^{-1} \\ =&(\mu_\Sigma \times P_Z) \circ (T_\sigma^\Sigma \times id)^{-1} \circ (T^Z)^{-1} \,, \end{aligned} \tag{47}$$

where T^Σ is the left-multiplication action of Σ on itself. Invariance of μ_Σ implies

$$(\mu_\Sigma \times P_Z) \circ (T_\sigma^\Sigma \times id)^{-1} =(\mu_\Sigma \circ (T_\sigma^\Sigma)^{-1}) \times P_Z = \mu_\Sigma \times P_Z \,. \tag{48}$$

Therefore

$$P_Z^\Sigma \circ T_\sigma^{-1} = (\mu_\Sigma \times P_Z) \circ (T^Z)^{-1} = P_Z^\Sigma \,. \tag{49}$$

This proves P_Z^Σ is Σ-invariant as claimed. □

Next we show how the proof of Theorem 4.1 can be generalized to a wider variety of objective functionals. This result will utilize a certain topology on the space of bounded measurable functions which we describe in the following definition.

Definition B.9. Let V be a subspace of $\mathcal{M}_b(X)^n$, $n \in \mathbb{Z}^+$, and $M(X)$ be the set of finite signed measures on X. For $\nu \in M(X)^n$ we define $\tau_\nu : V \to \mathbb{R}$ by $\tau_\nu(\gamma) \coloneqq \sum_{i=1}^n \int \gamma^i d\nu_i$ and we let $\mathcal{T} = \{\tau_\nu : \nu \in M(X)^n\}$. $\mathcal{T}$ is a separating vector space of linear functionals on V and we equip V with the weak topology from $\mathcal{T}$ (i.e., the weakest topology on V for which every $\tau \in \mathcal{T}$ is continuous). This makes V a locally convex topological vector space with dual space $V^* = \mathcal{T}$; see Theorem 3.10 in (Rudin, 2006). In the following we will abbreviate this by saying that V *has the $M(X)$-topology*.

Theorem B.10. *Let V be a subspace of $\mathcal{M}_b(X)^n$, $n \in \mathbb{Z}^+$, that is closed under Σ in the sense of (11) and satisfies $S_\Sigma[V] \subset V$. Given an objective functional $H : V \times \mathcal{P}(X) \times \mathcal{P}(X) \to [-\infty, \infty)$ and a test function space $\Gamma \subset V$ we define*

$$D_H^\Gamma(Q \| P) \coloneqq \sup_{\gamma \in \Gamma} H(\gamma; Q, P) \,. \tag{50}$$

If $H(\cdot; Q, P)$ is concave and upper semi-continuous (USC) in the $M(X)$-topology on V (see Definition B.9) and

$$H(\gamma \circ T_\sigma; Q, P) = H(\gamma; Q \circ T_\sigma^{-1}, P \circ T_\sigma^{-1}) \tag{51}$$

for all $\sigma \in \Sigma$, $\gamma \in V$, and $Q, P \in \mathcal{P}(X)$ then for all Σ-invariant Q, P we have

$$D_H^{\Gamma}(Q\|P) \leq D_H^{S_\Sigma[\Gamma]}(Q\|P)\,. \tag{52}$$

If, in addition, $S_\Sigma[\Gamma] \subset \Gamma$ then $S_\Sigma[\Gamma] = \Gamma_\Sigma^{inv}$ and

$$D_H^{\Gamma}(Q\|P) = D_H^{\Gamma_\Sigma^{inv}}(Q\|P)\,. \tag{53}$$

Remark B.11. See Appendix C for conditions implying $S_\Sigma[\Gamma] \subset \Gamma$.

Proof. Fix $\gamma \in \Gamma$ and Σ-invariant Q, P. Define $G := -H(\cdot; Q, P)$ and note that $G : V \to (-\infty, \infty]$ is LSC and convex. Convex conjugate duality (see the Fenchel-Moreau Theorem, e.g., Theorem 2.3.6 in Bot et al. (2009)) and Fubini's theorem then imply

$$\begin{aligned}
G(S_\Sigma[\gamma]) &= \sup_{\nu \in M(X)^n} \{\tau_\nu(S_\Sigma[\gamma]) - G^*(\tau_\nu)\} \\
&= \sup_{\nu \in M(X)^n} \{\sum_i \int S_\Sigma[\gamma^i] d\nu_i - G^*(\tau_\nu)\} \\
&= \sup_{\nu \in M(X)^n} \{\int \sum_i \int \gamma^i \circ T_\sigma d\nu_i - G^*(\tau_\nu)\mu_\Sigma(d\sigma)\} \\
&= \sup_{\nu \in M(X)^n} \{\int \tau_\nu(\gamma \circ T_\sigma) - G^*(\tau_\nu)\mu_\Sigma(d\sigma)\} \\
&\leq \int G(\gamma \circ T_\sigma)\mu_\Sigma(d\sigma)\,.
\end{aligned}$$

We can use our assumptions to compute

$$\begin{aligned}
G(\gamma \circ T_\sigma) &= -H(\gamma \circ T_\sigma; Q, P) \\
&= -H(\gamma; Q \circ T_\sigma^{-1}, P \circ T_\sigma^{-1}) \\
&= -H(\gamma; Q, P)
\end{aligned}$$

and hence we obtain

$$H(S_\Sigma[\gamma]; Q, P) \geq H(\gamma; Q, P)\,.$$

Taking the supremum over $\gamma \in \Gamma$ gives (52). If $S_\Sigma[\Gamma] \subset \Gamma$ then we clearly have the bound $D_H^{S_\Sigma[\Gamma]} \leq D_H^{\Gamma}$ and hence $D_H^{S_\Sigma[\Gamma]} = D_H^{\Gamma}$. The equality $S_\Sigma[\Gamma] = \Gamma_\Sigma^{\text{inv}}$ was shown in the proof of Theorem 4.1 and so we are done. □

Theorem B.10 applies to many classes of divergences, some of which have not been discussed in the main text. For example:

1. Integral probability metrics and MMD (5); see (Müller, 1997; Sriperumbudur et al., 2012).
2. (f, Γ) divergences (6); concavity and USC of the objective functional follows Proposition B.8 in (Birrell et al., 2022).
3. Sinkhorn divergences (9); concavity and USC of the objective functional follows Lemma B.7 in (Birrell et al., 2022).
4. Rényi divergence for $\alpha \in (0, 1)$; see Theorem 3.1 in (Birrell et al., 2021).
5. The Kullback-Leibler Approximate Lower bound Estimator (KALE); see Definition 1 in (Glaser et al., 2021).

C. Conditions Ensuring $S_\Sigma[\Gamma] \subset \Gamma$

In this appendix we provide conditions under which the test function space Γ is closed under symmetrization, that being a key assumption in our main results in Section 4. First we show that $S_\Sigma[\Gamma] \subset \Gamma$ when Γ is the unit ball in an appropriate RKHS.

Lemma C.1. *Let $V \subset \mathcal{M}_b(X)$ be a separable RKHS with reproducing-kernel $k : X \times X \to \mathbb{R}$. Let $\Gamma = \{\gamma \in V : \|\gamma\|_V \leq 1\}$ be the unit ball in V. Suppose we have a measurable group action $T : \Sigma \times X \to X$ and k is Σ-invariant under this action (i.e., $k(T_\sigma(x), T_\sigma(y)) = k(x, y)$ for all $\sigma \in \Sigma$, $x, y \in X$). Then $S_\Sigma[\Gamma] \subset \Gamma$.*

Remark C.2. The proof will use many standard properties of a RKHS. In particular, recall that the assumption $X \subset \mathcal{M}_b(X)$ implies k is bounded and jointly measurable. See Chapter 4 in (Steinwart & Christmann, 2008) for this and further background. See (Sriperumbudur et al., 2011) and references therein for more discussion of characteristic kernels as well as the related topic of universal kernels.

Proof. The Σ-invariance of k implies

$$k(T_\sigma(x), y) = k(T_\sigma(x), T_\sigma(T_{\sigma^{-1}}(y))) = k(x, T_{\sigma^{-1}}(y)) \tag{54}$$

and

$$\langle k(\cdot, T_\sigma(x)), k(\cdot, T_\sigma(y))\rangle_V = k(T_\sigma(x), T_\sigma(y)) = k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_V \tag{55}$$

for all $\sigma \in \Sigma$ and $x, y \in X$. Next we will show that the map $U_\sigma : \gamma \mapsto \gamma \circ T_\sigma$ is an isometry on V for all $\sigma \in \Sigma$, $\gamma \in V$: It is clearly a linear map. To show its range is contained in V, first recall that the span of $\{k(\cdot, x)\}_{x\in X}$ is dense in V. Therefore, given $\gamma \in V$ there is a sequence $\gamma_n \to \gamma$ having the form

$$\gamma_n = \sum_{i=1}^{N_n} a_{n,i} k(\cdot, x_{n,i})$$

for some $a_{n,i} \in \mathbb{R}$, $x_{n,i} \in X$. Equation (54) implies

$$\gamma_n \circ T_\sigma = \sum_{i=1}^{N_n} a_{n,i} k(T_\sigma(\cdot), x_{n,i}) = \sum_{i=1}^{N_n} a_{n,i} k(\cdot, T_{\sigma^{-1}}(x_{n,i}))\,.$$

Combining Eq. (56) with Eq. (55) we can conclude that $\|\gamma_n \circ T_\sigma\|_V = \|\gamma_n\|_V$ and $\|\gamma_n \circ T_\sigma - \gamma_m \circ T_\sigma\|_V = \|\gamma_n - \gamma_m\|_V$. γ_n converges in V, hence is Cauchy, therefore $\gamma_n \circ T_\sigma$ is Cauchy as well. We have assumed V is complete, therefore $\gamma_n \circ T_\sigma \to \tilde{\gamma}$ for some $\tilde{\gamma} \in V$. V is a RKHS, hence the evaluation maps are continuous and we find $\tilde{\gamma}(x) = \lim_n \gamma_n(T_\sigma(x)) = \gamma(T_\sigma(x))$ for all x. Therefore $\gamma \circ T_\sigma = \tilde{\gamma} \in V$ and

$$\|\gamma \circ T_\sigma\|_V = \lim_n \|\gamma_n \circ T_\sigma\|_V = \lim_n \|\gamma_n\|_V = \|\gamma\|_V\,.$$

This proves U_σ is an isometry on V.

Now fix $\gamma \in \Gamma$. We will show that the map $\sigma \to U_\sigma[\gamma]$ is Bochner integrable (see, e.g., Appendix E in Cohn (2013)): It clearly has has separable range since V was assumed to be separable. By the same reasoning as above, given $\tilde{\gamma} \in V$ we have a sequence $\tilde{\gamma}_n \to \tilde{\gamma}$ where

$$\tilde{\gamma}_n = \sum_{i=1}^{N_n} a_{n,i} k(\cdot, x_{n,i})\,.$$

Hence

$$\begin{aligned}\langle \tilde{\gamma}, U_\sigma[\gamma]\rangle_V &= \lim_n \sum_{i=1}^{N_n} a_{n,i} \langle k(\cdot, x_{n,i}), U_\sigma[\gamma]\rangle_V = \lim_n \sum_{i=1}^{N_n} a_{n,i}, U_\sigma[\gamma](x_{n,i}) \\ &= \lim_n \sum_{i=1}^{N_n} a_{n,i}, \gamma(T_\sigma(x_{n,i}))\,,\end{aligned}$$

which is now clearly measurable in σ due to the measurability of the action. Therefore $\sigma \mapsto U_\sigma[\gamma]$ is strongly measurable. $\|U_\sigma[\gamma]\|_V = \|\gamma\|_V \leq 1$, therefore the Bochner integral $\int U_\sigma[\gamma]\mu_\Sigma(d\sigma)$ exists in V and satisfies

$$\|\int U_\sigma[\gamma]\mu_\Sigma(d\sigma)\|_V \leq \int \|U_\sigma[\gamma]\|_V \mu_\Sigma(d\sigma) \leq 1\,.$$

This proves $\int U_\sigma[\gamma]\mu_\Sigma(d\sigma) \in \Gamma$. Finally, V is a RKHS and so the evaluation maps are in V^*. Therefore evaluation commutes with the Bochner integral and we find

$$(\int U_\sigma[\gamma]\mu_\Sigma(d\sigma))(x) = \int U_\sigma[\gamma](x)\mu_\Sigma(d\sigma) = \int \gamma(T_\sigma(x))\mu_\Sigma(d\sigma) = S_\Sigma[\gamma](x)\,.$$

Hence we can conclude $S_\Sigma[\gamma] \in \Gamma$ for all $\gamma \in \Gamma$ as claimed. □

The next result provides a general framework for proving $S_\Sigma[\Gamma] \subset \Gamma$.

Lemma C.3. *Let $V \subset \mathcal{M}_b(X)^n$, $n \in \mathbb{Z}^+$, be a subspace equipped with the $M(X)$-topology (see Definition B.9) and $\Gamma \subset V$. If Γ is convex and closed, the group action $T : \Sigma \times X \to X$ is measurable, $S_\Sigma[V] \subset V$, and Γ is closed under Σ (i.e., $\gamma \circ T_\sigma \in \Gamma$ for all $\gamma \in \Gamma$, $\sigma \in \Sigma$) then $S_\Sigma[\Gamma] \subset \Gamma$.*

Proof. Suppose we have $\gamma \in \Gamma$ with $S_\Sigma[\gamma] \notin \Gamma$. As noted in Definition B.9, V is a locally convex topological vector space with $V^* = \{\tau_\nu : \nu \in M(X)^n\}$, $\tau_\nu(\gamma) \coloneqq \sum_{i=1}^n \int \gamma^i d\nu_i$. The separating hyperplane theorem (see Theorem 3.4(b) in Rudin (2006)) applied to $A = \{S_\Sigma[\gamma]\}$ and $B = \Gamma$ therefore implies the existence of $\nu \in M(X)^n$ such that

$$\tau_\nu(\tilde{\gamma}) > \tau_\nu(S_\Sigma[\gamma]) \tag{56}$$

for all $\tilde{\gamma} \in \Gamma$. We have assumed Γ is closed under Σ and so we can let $\tilde{\gamma} = \gamma \circ T_\sigma$ to get

$$\sum_{i=1}^n \int \gamma^i \circ T_\sigma d\nu_i - \sum_{i=1}^n \int S_\Sigma[\gamma^i] d\nu_i > 0 \tag{57}$$

for all $\sigma \in \Sigma$. Integrating with respect to $\mu_\Sigma(d\sigma)$ and using Fubini's theorem to change the order of integration we obtain a contradiction. Therefore $S_\Sigma[\gamma] \in \Gamma$ as claimed. □

We end this section with several examples of function spaces, V, that are useful in conjunction with Lemma C.3:

1. $V = \mathcal{M}_b(X)^n$, $n \in \mathbb{Z}^+$, in which case $S_\Sigma[V] \subset V$ follows from measurability of the action.
2. X is a metric space, the action $T : \Sigma \times X \to X$ is continuous, and $V = C_b(X)^n$, $n \in \mathbb{Z}^+$. In this case, $S_\Sigma[V] \subset V$ follows from the dominated convergence theorem.
3. X is a metric space, the action $T : \Sigma \times X \to X$ is continuous, T_σ is 1-Lipschitz for all $\sigma \in \Sigma$, and $V = \mathrm{Lip}_b^1(X)^n$, $n \in \mathbb{Z}^+$. In this case, $S_\Sigma[V] \subset V$ follows from the following calculation:

$$\begin{aligned}|S_\Sigma[\gamma](x) - S_\Sigma[\gamma](y)| &\leq \int |\gamma(T_\sigma(x)) - \gamma(T_\sigma(y))|\mu_\Sigma(d\sigma) \leq \int d(T_\sigma(x), T_\sigma(y))\mu_\Sigma(d\sigma) \\ &\leq \int d(x,y)\mu_\Sigma(d\sigma) = d(x,y)\end{aligned}$$

   for all $\gamma \in \mathrm{Lip}_b^1(X)$.

D. Additional Properties of Σ-Invariant (f, Γ)-Divergences

In this appendix we derive further properties of (f, Γ)-divergences between Σ-invariant distributions. Here we will assume that X is a complete separable metric space (with metric d). Our analysis will require the following notion of a determining set of functions.

Definition D.1. Given $\mathcal{Q} \subset \mathcal{P}(X)$, a subset $\Psi \subset \mathcal{M}_b(X)$ will be called $\mathcal{Q}$**-determining** if for all $Q, P \in \mathcal{Q}$, $E_Q[\psi] = E_P[\psi]$ for all $\psi \in \Psi$ implies $Q = P$.

We will also need f and Γ to satisfy one of the following admissibility criteria, as introduced in (Birrell et al., 2022).

Definition D.2. For a, b with $-\infty \leq a < 1 < b \leq \infty$ we define $\mathcal{F}_1(a, b)$ to be the set of convex functions $f : (a, b) \to \mathbb{R}$ with $f(1) = 0$. For $f \in \mathcal{F}_1(a, b)$, if b is finite we extend the definition of f by $f(b) \coloneqq \lim_{x \nearrow b} f(x)$. Similarly, if a is finite we define $f(a) \coloneqq \lim_{x \searrow a} f(x)$ (convexity implies these limits exist in $(-\infty, \infty]$). Finally, extend f to $x \notin [a, b]$ by $f(x) = \infty$. The resulting function $f : \mathbb{R} \to (-\infty, \infty]$ is convex and LSC.

We will call $f \in \mathcal{F}_1(a, b)$ **admissible** if $\{f^* < \infty\} = \mathbb{R}$ and $\lim_{y \to -\infty} f^*(y) < \infty$ (note that this limit always exists by convexity). If f is also strictly convex at 1 then we will call f **strictly admissible**. We will call $\Gamma \subset C_b(X)$ **admissible** if $0 \in \Gamma$, Γ is convex, and Γ is closed in the $M(X)$-topology on $C_b(X)$ (see Definition B.9). Γ will be called **strictly admissible** if it also satisfies the following property: There exists a $\mathcal{P}(X)$-determining set $\Psi \subset C_b(X)$ such that for all $\psi \in \Psi$ there exists $c \in \mathbb{R}$, $\epsilon > 0$ such that $c \pm \epsilon\psi \in \Gamma$. Finally, an admissible $\Gamma \subset C^{\text{inv}}_{b,\Sigma}(X)$ (the set of Σ-invariant bounded continuous functions) will be called Σ**–strictly admissible** if there exists a $\mathcal{P}_\Sigma(X)$-determining set $\Psi \subset C_b(X)$ such that for all $\psi \in \Psi$ there exists $c \in \mathbb{R}$, $\epsilon > 0$ such that $c \pm \epsilon\psi \in \Gamma$.

One way to construct a Σ-strictly admissible set is to start with an appropriate strictly admissible set and then restrict to the subset of Σ-invariant functions; see Appendix D.1 for a proof.

Lemma D.3. *Let $\Gamma \subset C_b(X)$.*

1. *If Γ is admissible then Γ^{inv}_Σ is admissible.*
2. *If Γ is strictly admissible and $S_\Sigma[\Gamma] \subset \Gamma$ then Γ^{inv}_Σ is Σ-strictly admissible.*

Below are several useful examples of strictly admissible Γ that satisfy $S_\Sigma[\Gamma] \subset \Gamma$.

1. $\Gamma \coloneqq C_b(X)$, if the action is continuous in x, i.e., if $T_\sigma : X \to X$ is continuous for all $\sigma \in \Sigma$.
2. $\Gamma \coloneqq \{g \in C_b(X) : |g| \leq C\}$ for any $C > 0$ and assuming the action is continuous in x,
3. $\Gamma \coloneqq \text{Lip}^L_b(X)$ for any $L > 0$ and assuming the action is 1-Lipschitz, i.e., $d(T_\sigma(x), T_\sigma(y)) \leq d(x, y)$ for all $\sigma \in \Sigma$, $x, y \in X$.
4. $\Gamma \coloneqq \{g \in \text{Lip}^L_b(X) : |g| \leq C\}$ for any $C, L > 0$ and assuming the action is 1-Lipschitz.
5. The unit ball in an appropriate RKHS V, $\Gamma \coloneqq \{g \in V : \|g\|_V \leq 1\}$, assuming the kernel is Σ-invariant; see Lemma D.6 for details.

The following result extends the infimal convolution formula and divergence properties from (Birrell et al., 2022) to the case where the models and test-function space are Σ-invariant.

Theorem D.4. *Suppose f and Γ are admissible and $\Gamma \subset C^{inv}_{b,\Sigma}(X)$. For $Q, P \in \mathcal{P}_\Sigma(X)$ we have the following properties:*

1. *Infimal Convolution Formula on $\mathcal{P}_\Sigma(X)$:*

$$D^\Gamma_f(Q\|P) = \inf_{\eta \in \mathcal{P}_\Sigma(X)} \{D_f(\eta\|P) + W^\Gamma(Q, \eta)\}. \tag{58}$$

2. *Existence of an Optimizer: If $D^\Gamma_f(Q\|P) < \infty$ then there exists $\eta_* \in \mathcal{P}_\Sigma(X)$ such that*

$$D^\Gamma_f(Q\|P) = D_f(\eta_*\|P) + W^\Gamma(Q, \eta_*). \tag{59}$$

   If f is strictly convex then there is a unique such η_*.

3. *$\mathcal{P}_\Sigma(X)$-Divergence Property for W^Γ: $W^\Gamma(Q, P) \geq 0$ and $W^\Gamma(Q, P) = 0$ if $Q = P$. If Γ is Σ-strictly admissible then $W^\Gamma(Q, P) = 0$ implies $Q = P$.*

4. *$\mathcal{P}_\Sigma(X)$-Divergence Property for D_f^Γ: $D_f^\Gamma(Q\|P) \geq 0$ and $D_f^\Gamma(Q\|P) = 0$ if $Q = P$. If f is strictly admissible and Γ is Σ-strictly admissible then $D_f^\Gamma(Q\|P) = 0$ implies $Q = P$.*

Proof. 1. Part 1 of Theorem 2.15 from (Birrell et al., 2022) implies an infimal convolution formula on $\mathcal{P}(X)$, hence

$$D_f^\Gamma(Q\|P) = \inf_{\eta\in\mathcal{P}(X)}\{D_f(\eta\|P) + W^\Gamma(Q,\eta)\} \leq \inf_{\eta\in\mathcal{P}_\Sigma(X)}\{D_f(\eta\|P) + W^\Gamma(Q,\eta)\}\,. \tag{60}$$

To prove the reverse inequality, we use the bound $D_f \geq D_f^{S_\Sigma[\mathcal{M}_b(X)]}$, the equality $S_\Sigma[\Gamma] = \Gamma$, and then Theorem B.5 to compute

$$\begin{aligned} D_f^\Gamma(Q\|P) &\geq \inf_{\eta\in\mathcal{P}(X)}\{D_f^{S_\Sigma[\mathcal{M}_b(X)]}(\eta\|P) + W^{S_\Sigma[\Gamma]}(Q,\eta)\} \\ &= \inf_{\eta\in\mathcal{P}(X)}\{D_f(S^\Sigma[\eta]\|P) + W^\Gamma(Q,S^\Sigma[\eta])\} \\ &= \inf_{\eta\in\mathcal{P}_\Sigma(X)}\{D_f(\eta\|P) + W^\Gamma(Q,\eta)\}\,. \end{aligned} \tag{61}$$

This proves the infimal convolution formula on $\mathcal{P}_\Sigma(X)$.

2. Now suppose $D_f^\Gamma(Q\|P) < \infty$. Part 2 of Theorem 2.15 from (Birrell et al., 2022) implies there exists $\eta_* \in \mathcal{P}(X)$ such that

$$D_f^\Gamma(Q\|P) = D_f(\eta_*\|P) + W^\Gamma(Q,\eta_*)\,. \tag{62}$$

We need to show that η_* can be taken to be Σ-invariant. To do this, first use the infimal convolution formula to bound

$$D_f^\Gamma(Q\|P) \leq D_f(S^\Sigma[\eta_*]\|P) + W^\Gamma(Q,S^\Sigma[\eta_*])\,. \tag{63}$$

The Σ-invariance of Q and P together with Theorem B.5 imply

$$W^\Gamma(Q,S^\Sigma[\eta_*]) = W^\Gamma(Q,\eta_*)\,. \tag{64}$$

and

$$D_f(S^\Sigma[\eta_*]\|P) = D_f^{\mathcal{M}_{b,\Sigma}^{\text{inv}}(X)}(\eta_*\|P) \leq D_f(\eta_*\|P)\,. \tag{65}$$

Therefore

$$D_f^\Gamma(Q\|P) \leq D_f(S^\Sigma[\eta_*]\|P) + W^\Gamma(Q,S^\Sigma[\eta_*]) \leq D_f(\eta_*\|P) + W^\Gamma(Q,\eta_*) = D_f^\Gamma(Q\|P)\,. \tag{66}$$

Hence

$$D_f^\Gamma(Q\|P) = D_f(S^\Sigma[\eta_*]\|P) + W^\Gamma(Q,S^\Sigma[\eta_*]) \tag{67}$$

with $S^\Sigma[\eta_*] \in \mathcal{P}_\Sigma(X)$ as claimed.

If f is strictly convex then uniqueness is a corollary of the corresponding uniqueness result from Part 2 of Theorem 2.15 in (Birrell et al., 2022).

3. Admissibility of Γ implies $0 \in \Gamma$, hence $W^\Gamma(Q\|P) \geq E_Q[0] - E_P[0] = 0$. If $Q = P$ then the definition clearly implies $W^\Gamma(Q,P) = 0$. If Γ is Σ-strictly admissible and $W^\Gamma(Q,P) = 0$ then $0 \geq E_Q[g] - E_P[g]$ for all $g \in \Gamma$. Letting $g = c \pm \epsilon\psi$ as in the definition of Σ-strict admissiblity we see that $0 \geq \pm(E_Q[\psi] - E_P[\psi])$. Hence $E_Q[\psi] = E_P[\psi]$ for all $\psi \in \Psi$. Ψ is a $\mathcal{P}_\Sigma(X)$-determining set and $Q, P \in \mathcal{P}_\Sigma(X)$, hence we can conclude that $Q = P$.

4. We know that $D_f \geq 0$ and $W^\Gamma \geq 0$, therefore the infimal convolution formula implies $D_f^\Gamma \geq 0$. If $Q = P$ we can bound

$$0 \leq D_f^\Gamma(Q\|P) \leq D_f(Q\|P) = 0\,, \tag{68}$$

hence $D_f^\Gamma(Q\|P) = 0$. Finally, suppose f is strictly admissible, Γ is Σ-strictly admissible, and $D_f^\Gamma(Q\|P) = 0$. Then Part 2 of this theorem implies

$$0 = D_f^\Gamma(Q\|P) = D_f(\eta_*\|P) + W^\Gamma(Q, \eta_*) \tag{69}$$

for some $\eta_* \in \mathcal{P}_\Sigma(X)$. Both terms are non-negative, hence

$$D_f(\eta_*\|P) = W^\Gamma(Q, \eta_*) = 0\,. \tag{70}$$

The $\mathcal{P}_\Sigma(X)$-divergence property for W^Γ then implies $Q = \eta_*$. f being strictly admissible implies that D_f has the divergence property, hence $\eta_* = P$. Therefore $Q = P$ as claimed.

□

D.1. Admissibility Lemmas

In this appendix we prove several lemmas regarding admissible test function spaces. First we prove the admissibility properties of $\Gamma_\Sigma^{\text{inv}}$ from Lemma D.3.

Lemma D.5. *Let $\Gamma \subset C_b(X)$.*

1. *If Γ is admissible then Γ_Σ^{inv} is admissible.*
2. *If Γ is strictly admissible and $S_\Sigma[\Gamma] \subset \Gamma$ then Γ_Σ^{inv} is Σ-strictly admissible.*

Proof. 1. The zero function is Σ-invariant, hence is in $\Gamma_\Sigma^{\text{inv}}$. If $\gamma_1, \gamma_2 \in \Gamma_\Sigma^{\text{inv}}$ and $t \in [0,1]$ then convexity of Γ implies $t\gamma_1 + (1-t)\gamma_2 \in \Gamma$. We have $(t\gamma_1 + (1-t)\gamma_2) \circ T_\sigma = t\gamma_1 \circ T_\sigma + (1-t)\gamma_2 \circ T_\sigma = t\gamma_1 + (1-t)\gamma_2$, hence we conclude that $\Gamma_\Sigma^{\text{inv}}$ is convex. Finally, we can write

$$\begin{aligned}\Gamma_\Sigma^{\text{inv}} =& \Gamma \bigcap_{\sigma\in\Sigma, x\in X} \{\gamma \in C_b(X) : \gamma(T_\sigma(x)) = \gamma(x)\} \\ =& \Gamma \bigcap_{\sigma\in\Sigma, x\in X} \{\gamma \in C_b(X) : \tau_{\delta_{T_\sigma(x)}}[\gamma] = \tau_{\delta_x}[\gamma]\}\,.\end{aligned}$$

We have assumed Γ is admissible, hence it is closed. The maps τ_ν, $\nu \in M(X)$ are continuous on $C_b(X)$, hence the sets $\{\gamma \in C_b(X) : \tau_{\delta_{T_\sigma(x)}}[\gamma] = \tau_{\delta_x}[\gamma]\}$ are also closed. Therefore $\Gamma_\Sigma^{\text{inv}}$ is closed. This proves $\Gamma_\Sigma^{\text{inv}}$ is admissible.

2. Now suppose Γ is strictly admissible and $S_\Sigma[\Gamma] \subset \Gamma$. In particular, Γ is admissible and so Part 1 implies $\Gamma_\Sigma^{\text{inv}}$ is admissible. Let Ψ be as in the definition of strict admissibility. For every $\psi \in \Psi$ there exists $c \in \mathbb{R}$, $\epsilon > 0$ such that $c \pm \epsilon\psi \in \Gamma$. Hence $c \pm \epsilon S_\Sigma[\psi] = S_\Sigma[c \pm \epsilon\psi] = S_\Sigma[\Gamma] = \Gamma_\Sigma^{\text{inv}}$ (see the proof of Theorem 4.1) and $S_\Sigma[\Psi] \subset C_b(X)$. Finally, suppose $Q, P \in \mathcal{P}_\Sigma(X)$ such that $E_Q[S_\Sigma[\psi]] = E_P[S_\Sigma[\psi]]$ for all $\psi \in \Psi$. Part (b) of Lemma 3.2 then implies $E_Q[\psi] = E_P[\psi]$ for all $\psi \in \Psi$. Ψ is $\mathcal{P}(X)$-determining, hence $Q = P$. Therefore $S_\Sigma[\Psi]$ is a $\mathcal{P}_\Sigma(X)$-determining set and we conclude that $\Gamma_\Sigma^{\text{inv}}$ is Σ-strictly admissible.

□

Next we provide assumptions under which the unit ball in a RKHS is closed under S_Σ and is (strictly) admissible.

Lemma D.6. *Let $V \subset C_b(X)$ be a separable RKHS with reproducing-kernel $k : X \times X \to \mathbb{R}$. Let $\Gamma = \{\gamma \in V : \|\gamma\|_V \leq 1\}$ be the unit ball in V. Then:*

1. *Γ is admissible.*
2. *If the kernel is characteristic (i.e., the map $P \in \mathcal{P}(X) \mapsto \int k(\cdot, x)P(dx) \in V$ is one-to-one) then Γ is strictly admissible.*
3. *If k is Σ-invariant the $S_\Sigma[\Gamma] \subset \Gamma$.*

Proof. 1. Admissibility was shown in Lemma C.9 in (Birrell et al., 2022).

2. Now suppose the kernel is characteristic. Let $P, Q \in \mathcal{P}(X)$ with $\int \gamma dP = \int \gamma dQ$ for all $\gamma \in \Gamma$ (and hence for all $\gamma \in V$). Therefore

$$0 = \int \gamma dQ - \int \gamma dP = \langle \gamma, \int k(\cdot, x)Q(dx) - \int k(\cdot, x)P(dx) \rangle_V \tag{71}$$

for all $\gamma \in V$. Therefore $\int k(\cdot, x)Q(dx) = \int k(\cdot, x)P(dx)$. We have assumed the kernel is characteristic, hence we conclude that $Q = P$. This proves Γ is $\mathcal{P}(X)$-determining. We also have $-\Gamma \subset \Gamma$, hence Γ is strictly admissible.

3. This was shown in Lemma C.1 above.

□

E. Coarse-graining and structure-preserving operators

We show in this section how to apply our structure preserving formalism, Theorem 4.3 in particular, in the context of coarse-graining. We refer to the reviews (Noid, 2013; Pak & Voth, 2018) for fundamental concepts in the coarse-graining of molecular systems. Mathematically, a coarse-graining of the state space X is given by a measurable (non-invertible) map

$$\xi : X \to Y$$

where $y = \xi(x)$ are thought of as the coarse variables and Y as a space of significantly less complexity than X. If $\mathcal{A} = \sigma(\xi)$ is the σ-algebra generated by the coarse-graining map ξ then a function is measurable with respect to $\mathcal{A}$ if it is constant on every level set $\xi^{-1}(y)$.

To complete the description of the coarse-graining one selects a kernel $K_y(dx)$, which in the coarse-graining literature is called the back-mapping. The kernel $K_y(dx)$ describes the conditional distribution of the fully resolved state $x \in \xi^{-1}(y)$, conditioned on the coarse-grained state $y = \xi(x)$, namely $K_y(dx) = P(dx|y)$; in particular $K_y(dx)$ is supported on the set $\xi^{-1}(y)$. The kernel induces naturally a projection $S_K : \mathcal{M}_b(X) \to \mathcal{M}_b(X)$ given by

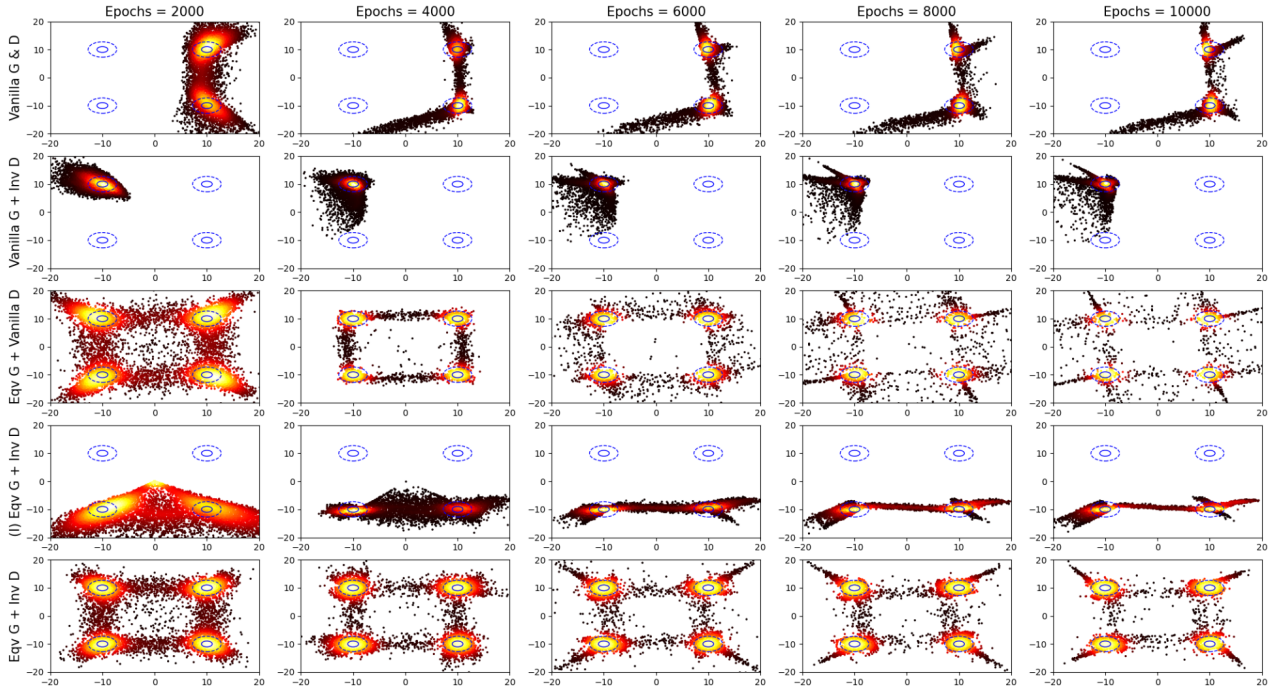
$$S_K[f](x) \; = \int_{\xi^{-1}(y)} f(x')K_y(dx') \qquad \text{for any } x \in \xi^{-1}(y)$$

and, by construction, $S_K[f](x)$ is $\mathcal{A}$-measurable. If a measure is S^K-invariant, i.e., $S^K[P] = P$, then it is uniquely determined by its value on $\mathcal{A}$, in other words it is completely specified by a probability measure $Q \in \mathcal{P}(Y)$ on the coarse variable $y = \xi(x)$. We refer to such a Q as a "coarse-grained" probability measure. Once a coarse-grained measure is constructed on Y, see (Noid, 2013; Pak & Voth, 2018) for a rich array of such methods, it can be then "reconstructed" as a measure on X by the kernel $K_y(dx)$ as $P(dx) = K_y(dx)Q(dy)$. For example, if we take X and Y to be discrete sets we can chose the trivial (uniform) reconstruction kernel with density $k_y(x) = \delta_x(\xi^{-1}(y))\frac{1}{|\xi^{-1}(y)|}$ and any coarse-grained measure with density $q(y)$ on the coarse variables y is reconstructed on X as a probability density on X:

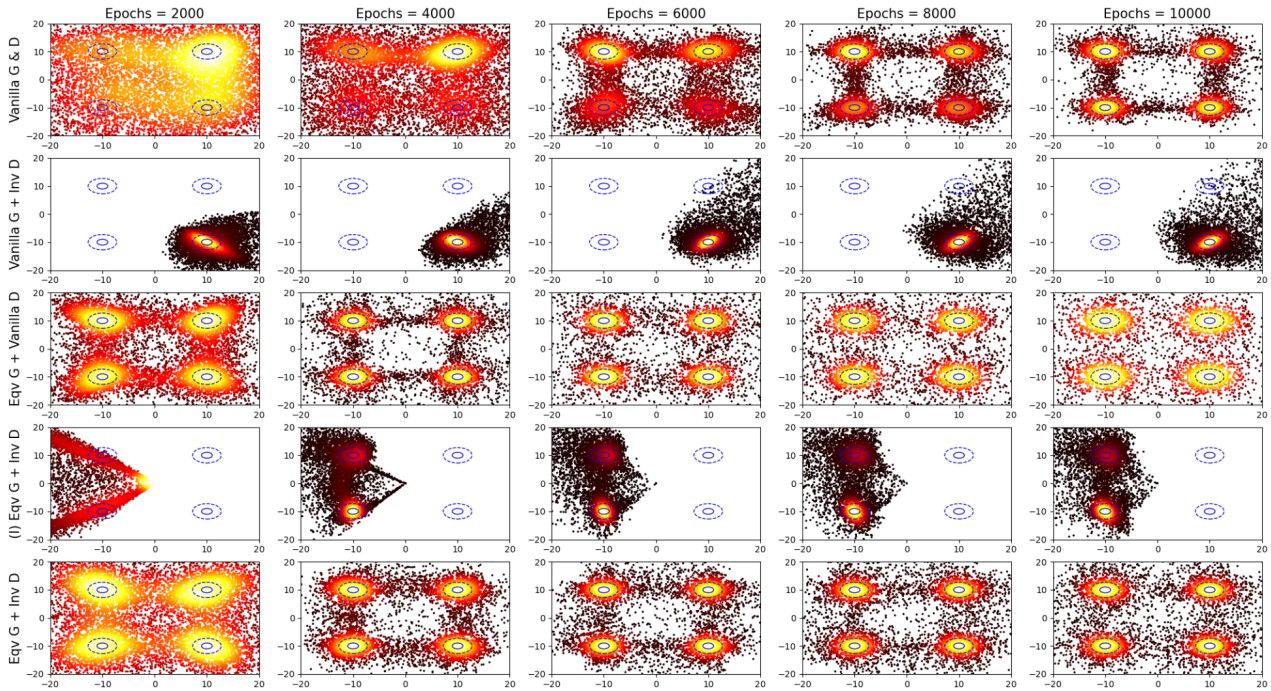
$$p(x) = \delta_x(\xi^{-1}(y))\frac{1}{|\xi^{-1}(y)|}q(y)\,, \quad \text{where} \quad y = \xi(x)\,, x \in X\,.$$

Finally, we note that back-mappings $K_y(dx) = P(dx|y)$ in coarse-graining—being probabilities conditioned on the coarse variables—can be constructed, to great accuracy, as generative models using *conditional GANs*, see (Li et al., 2020; Stieffenhofer et al., 2021).

F. Additional Experiments

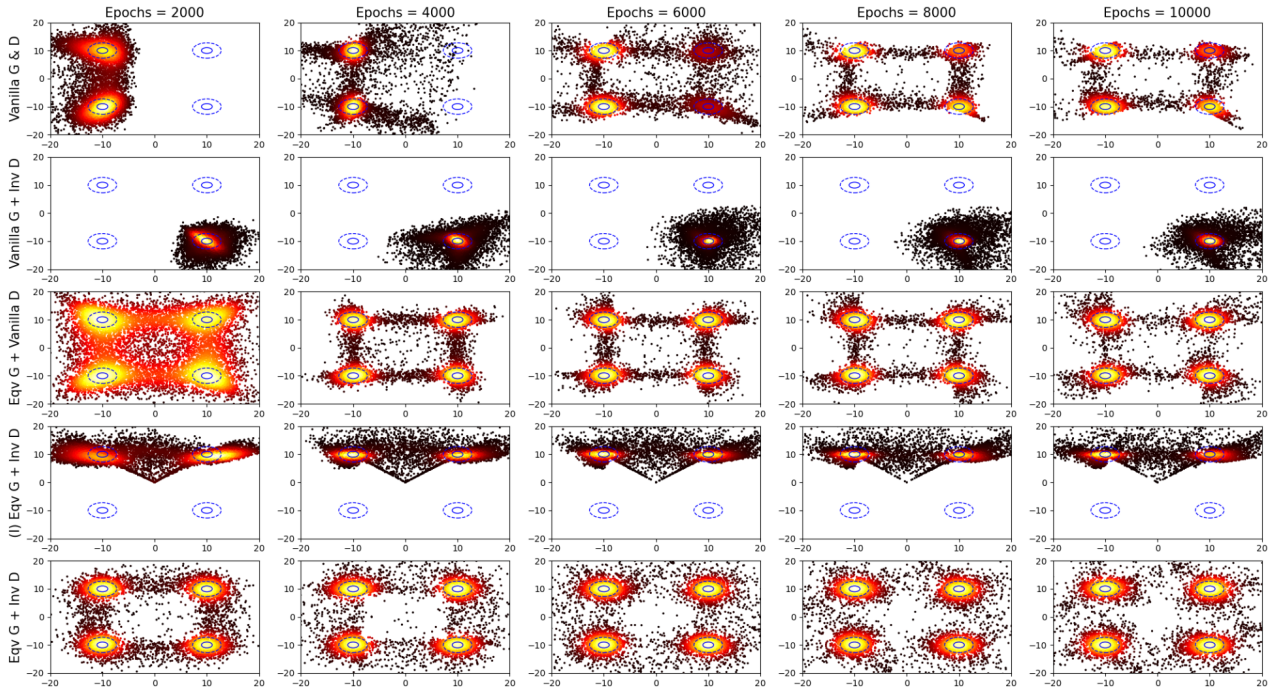


(a) Models trained with 50 training samples.

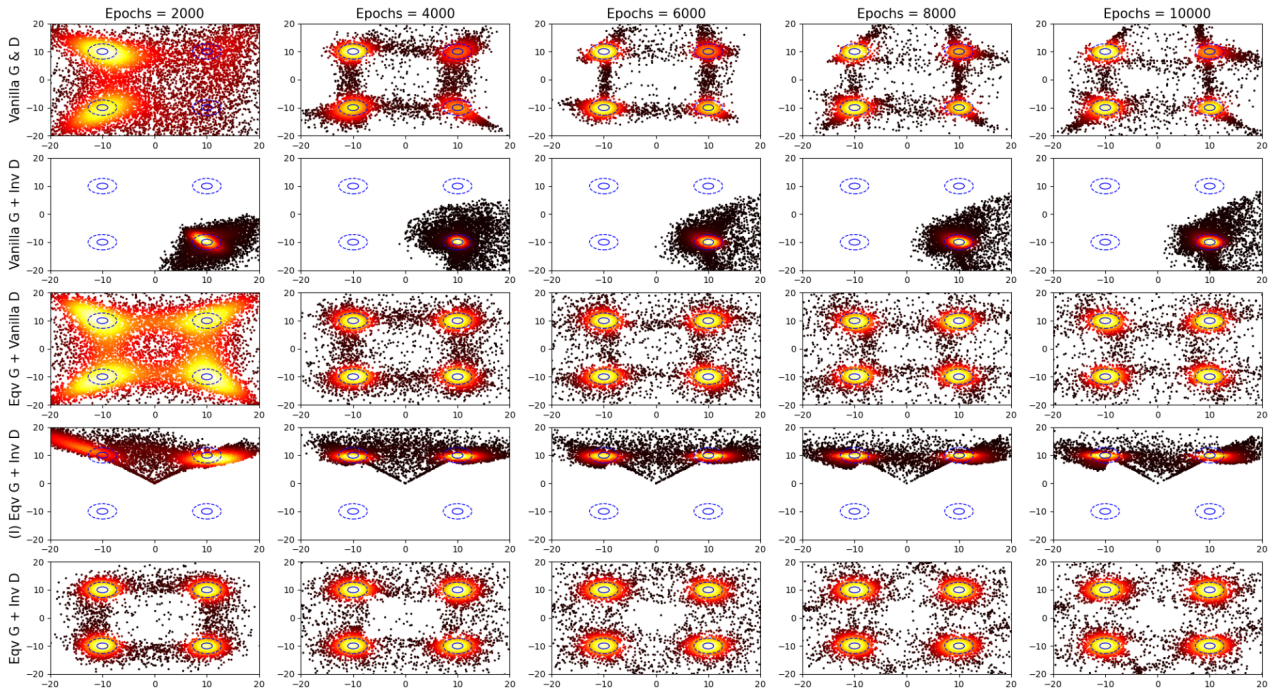


(b) Models trained with 5000 training samples.

Figure 6. 2D projection of the D_2^L-GAN generated samples onto the support plane of the source distribution Q [cf. Section 5.3]. Each column shows the result after a given number of training epochs. The rows correspond to different settings for the generators and discriminators. The solid and dashed blue ovals mark the 25% and 50% probability regions, respectively, of the data source Q, while the heat-map shows the generator samples. Panel (a): models are trained with **50** training samples. Panel (b): models are trained with **5000** training samples.



(a) D_α^L-GANs, $\alpha = 5$.



(b) D_α^L-GANs, $\alpha = 10$.

Figure 7. 2D projection of the D_α^L-GAN generated samples onto the support plane of the source distribution Q [cf. Section 5.3]. Each column shows the result after a given number of training epochs. The rows correspond to different settings for the generators and discriminators. The solid and dashed blue ovals mark the 25% and 50% probability regions, respectively, of the data source Q, while the heat-map shows the generator samples. Models are trained on **200** training points. Panel (a): $\alpha = 5$. Panel (b): $\alpha = 10$.

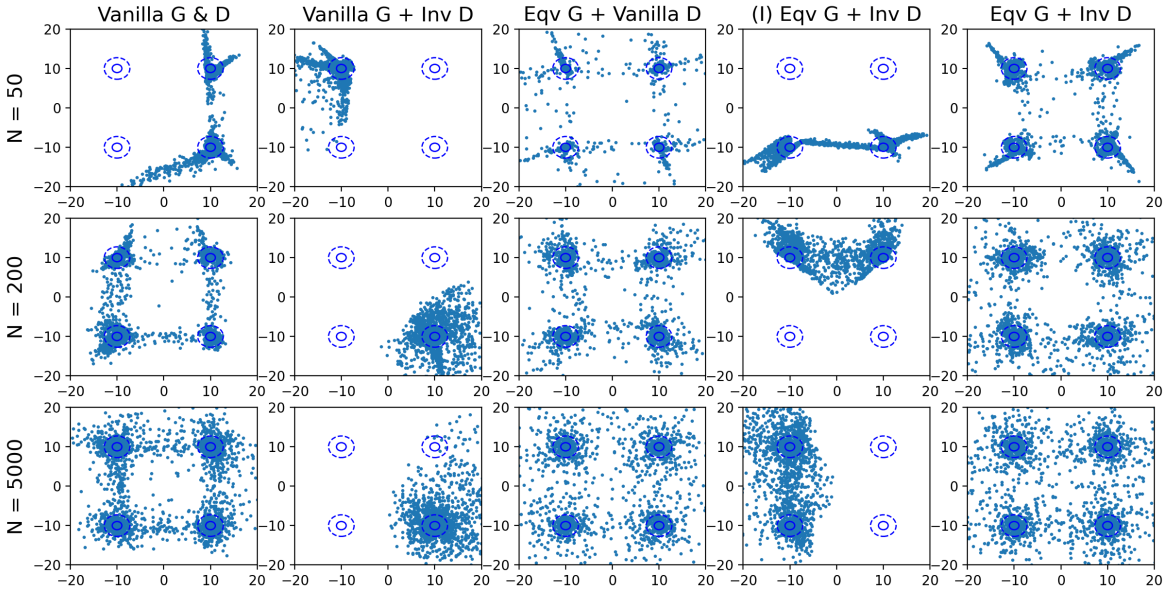
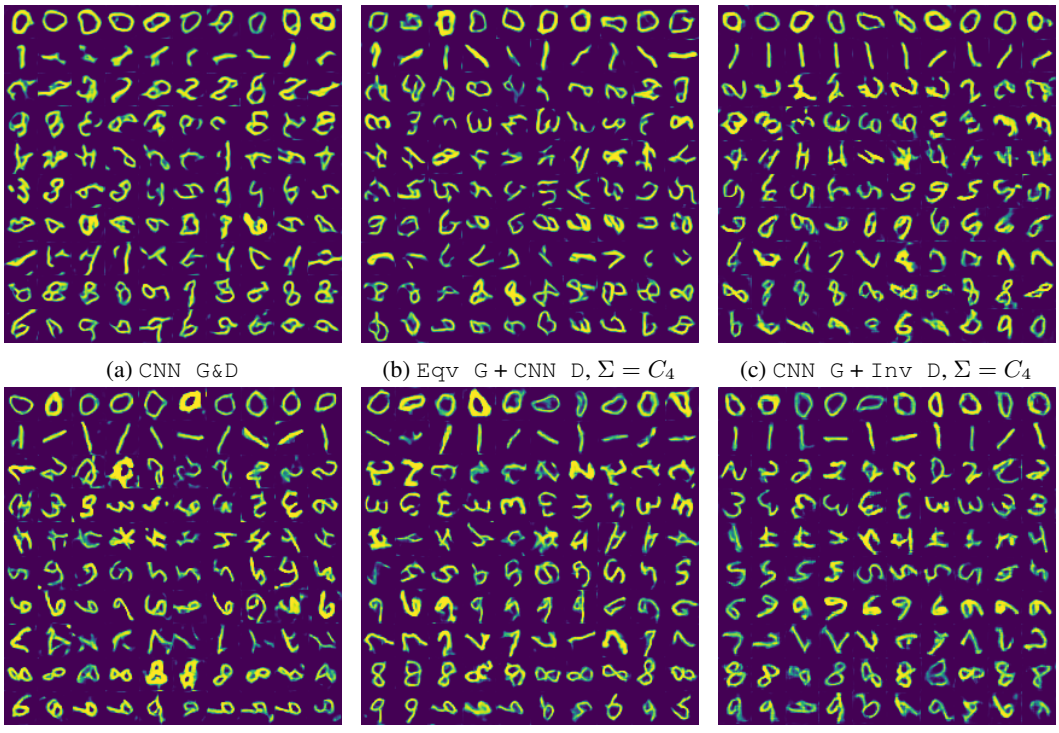


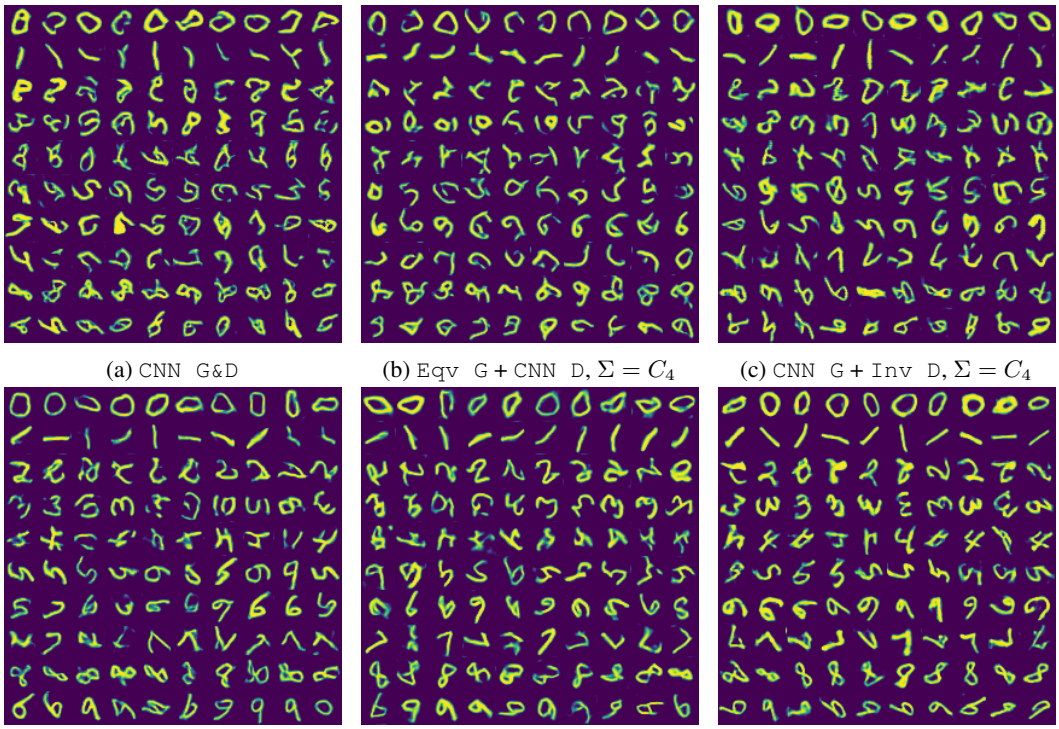
Figure 8. 2D projection of the D_2^L-GAN generated samples (3000 for each setting) onto the support plane of the source distribution Q [cf. Section 5.3]. Each GAN is trained for 10000 epochs. The rows correspond to the number of training points $N = 50$, 200, or 5000. The columns correspond to different settings for the generators and discriminators. The solid and dashed blue ovals mark the 25% and 50% probability regions, respectively, of the data source Q. Compared to Figure 6, heat maps are suppressed in this figure for easier examination of the sample quality.



(a) `CNN G&D` (b) `Eqv G + CNN D`, $\Sigma = C_4$ (c) `CNN G + Inv D`, $\Sigma = C_4$

(d) `(I) Eqv G + Inv D`, $\Sigma = C_4$ (e) `Eqv G + Inv D`, $\Sigma = C_4$ (f) `Eqv G + Inv D`, $\Sigma = C_8$

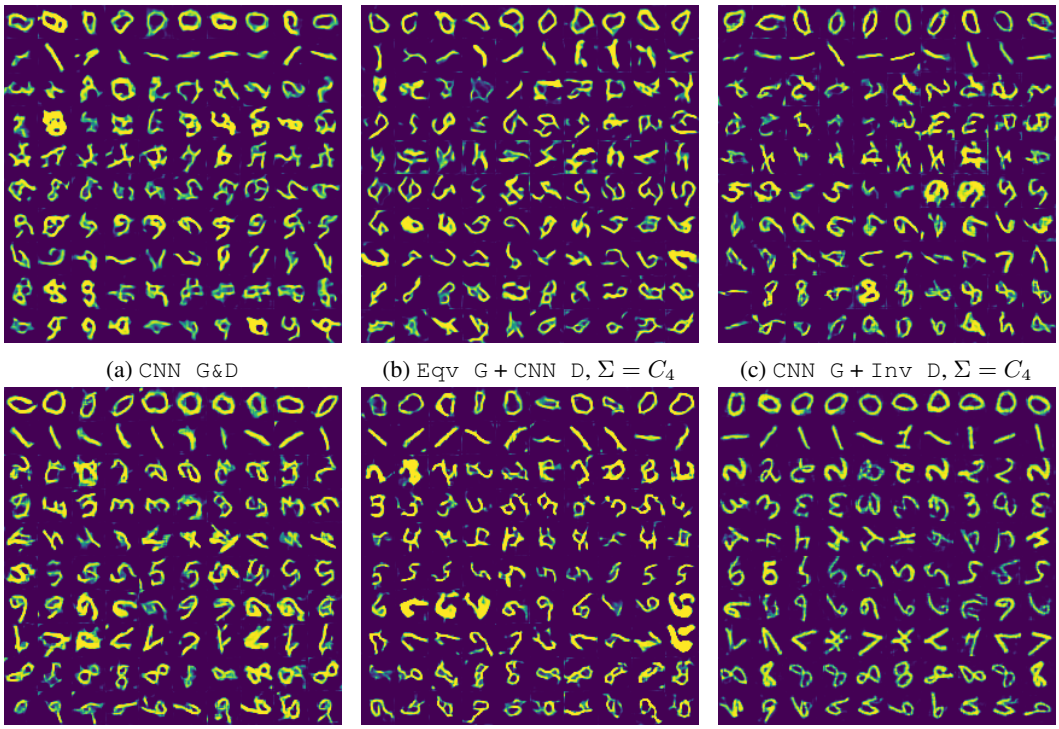
Figure 9. Randomly generated digits by the D_2^L-GANs trained on RotMNIST after 20K generator iterations with 1% (600) training data.



(a) `CNN G&D` (b) `Eqv G` + `CNN D`, $\Sigma = C_4$ (c) `CNN G` + `Inv D`, $\Sigma = C_4$

(d) `(I) Eqv G` + `Inv D`, $\Sigma = C_4$ (e) `Eqv G` + `Inv D`, $\Sigma = C_4$ (f) `Eqv G` + `Inv D`, $\Sigma = C_8$

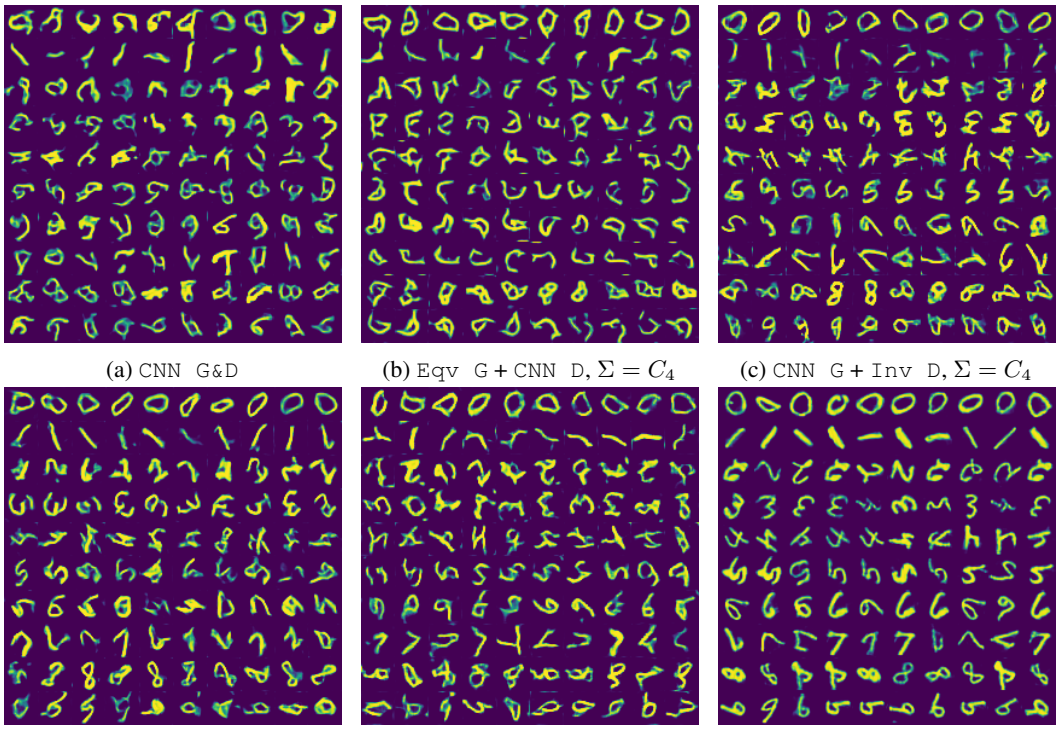
Figure 10. Randomly generated digits by the RA-GANs trained on RotMNIST after 20K generator iterations with 1% (600) training data.



(a) `CNN G&D` (b) `Eqv G + CNN D`, $\Sigma = C_4$ (c) `CNN G + Inv D`, $\Sigma = C_4$

(d) `(I) Eqv G + Inv D`, $\Sigma = C_4$ (e) `Eqv G + Inv D`, $\Sigma = C_4$ (f) `Eqv G + Inv D`, $\Sigma = C_8$

Figure 11. Randomly generated digits by the D_2^L-GANs trained on RotMNIST after 20K generator iterations with 0.33% (200) training data. Our model `Eqv G + Inv D`, $\Sigma = 8$ is the only one that can generate high-fidelity images in this setting. We note that the repetitively generated digits are inevitable in such a small data regime, as the models are forced to learn the empirical distribution of the limited training data (20 images per class).



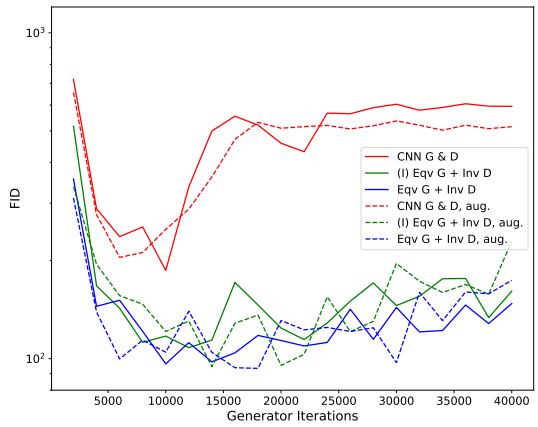
(a) CNN G&D (b) Eqv G + CNN D, $\Sigma = C_4$ (c) CNN G + Inv D, $\Sigma = C_4$

(d) (I) Eqv G + Inv D, $\Sigma = C_4$ (e) Eqv G + Inv D, $\Sigma = C_4$ (f) Eqv G + Inv D, $\Sigma = C_8$

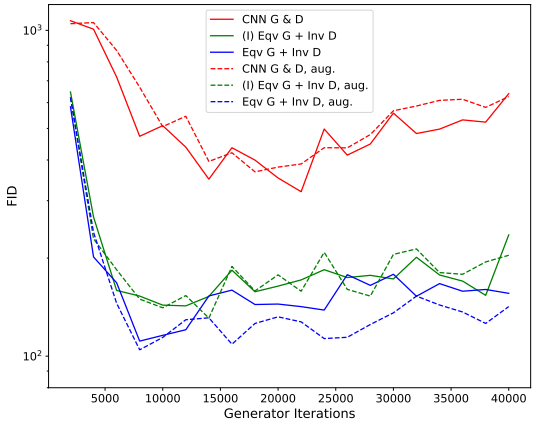
Figure 12. Randomly generated digits by the RA-GANs trained on RotMNIST after 20K generator iterations with 0.33% (200) training data. Our model Eqv G + Inv D, $\Sigma = 8$ is the only one that can generate high-fidelity images in this setting. We note that the repetitively generated digits are inevitable in such a small data regime, as the models are forced to learn the empirical distribution of the limited training data (20 images per class).

Table 3. The median of the FIDs (lower is better), calculated every 1,000 generator update for 20,000 iterations, averaged over three independent trials. The number of the training samples used for experiments varies from 0.33% (200) to 100% (60,000) of the entire training set.

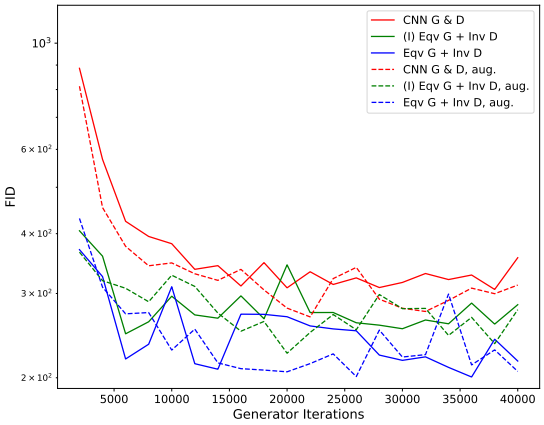
Loss	Architecture	0.33%	1%	5%	10%	25%	50%	100%
RA-GAN	CNN G&D	431	295	357	348	407	403	392
	Eqv G + CNN D, $\Sigma = C_4$	865	389	333	355	325	380	393
	CNN G + Inv D, $\Sigma = C_4$	382	223	181	188	185	177	176
	(I) Eqv G + Inv D, $\Sigma = C_4$	360	173	141	132	124	135	130
	Eqv G + Inv D, $\Sigma = C_4$	**190**	**98**	78	89	80	84	82
	Eqv G + Inv D, $\Sigma = C_8$	313	123	**52**	**51**	**59**	**52**	**57**
$D^{\Gamma}_{\alpha=2}$-GAN	CNN G&D	423	280	261	283	290	297	293
	Eqv G + CNN D, $\Sigma = C_4$	409	253	271	251	263	274	275
	CNN G + Inv D, $\Sigma = C_4$	511	330	208	192	190	183	173
	(I) Eqv G + Inv D, $\Sigma = C_4$	484	273	147	133	141	124	126
	Eqv G + Inv D, $\Sigma = C_4$	352	149	99	88	80	80	81
	Eqv G + Inv D, $\Sigma = C_8$	**293**	**122**	**55**	**57**	**53**	**53**	**51**



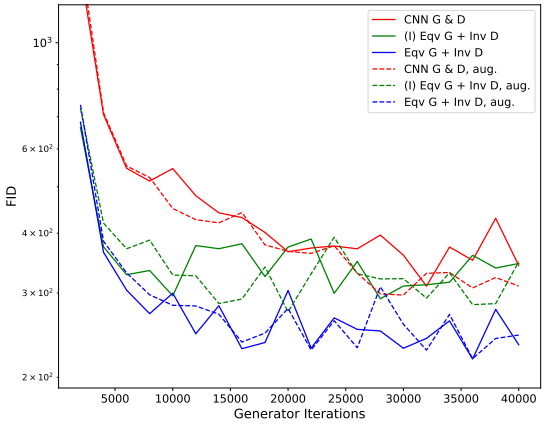
(a) ANHIR, RA-GAN



(b) ANHIR, D_2^L-GAN



(c) LYSTO, RA-GAN



(d) LYSTO, D_2^L-GAN

Figure 13. The curves of the Fréchet Inception Scores (FID), calculated after every 2,000 generator updates up to 40,000 iterations, averaged over three random trials on the medical data sets, ANHIR (top row) and LYSTO (bottom row). The symbol "aug." in the legend denotes the presence of data augmentation during GAN training.

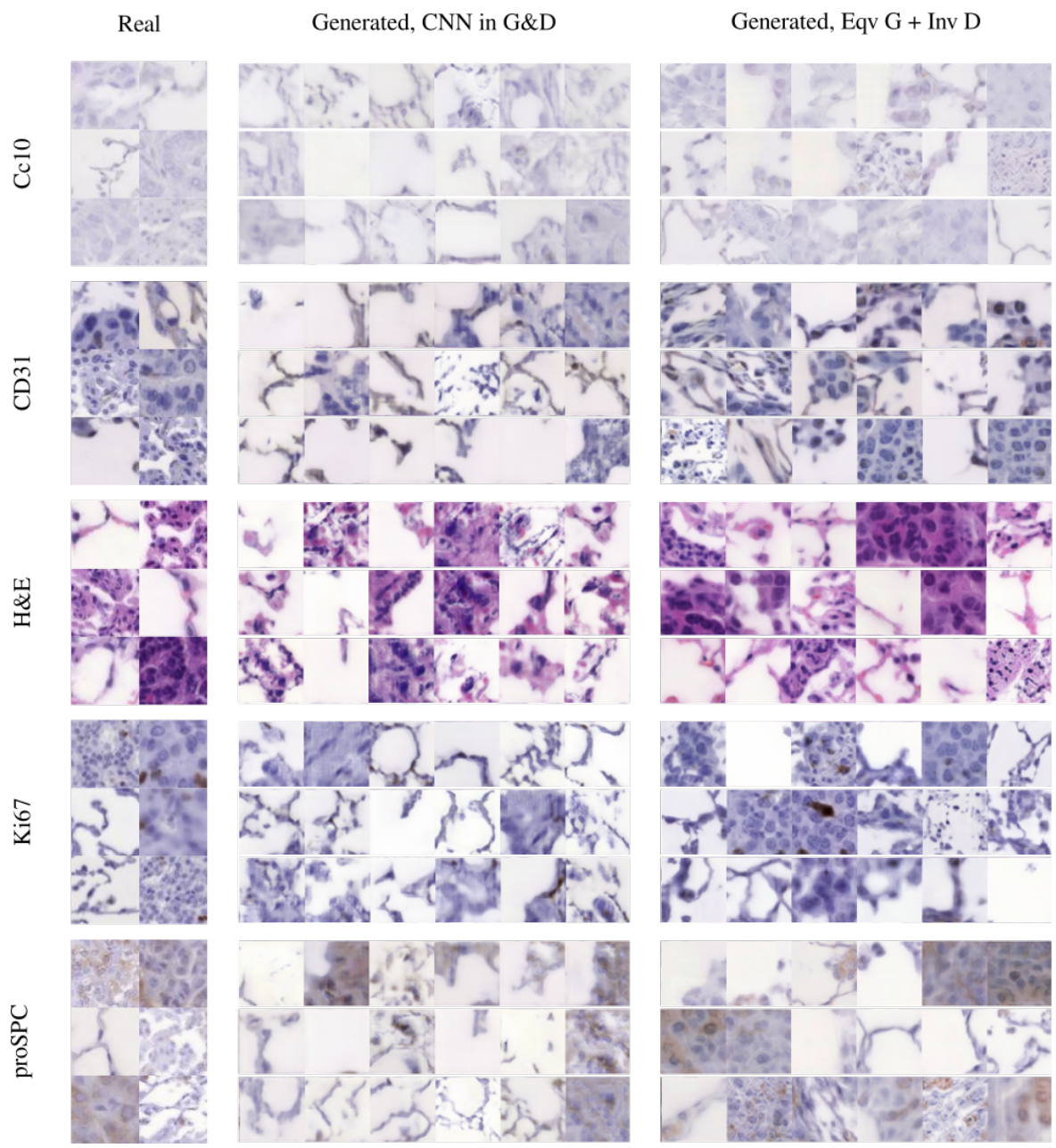


Figure 14. Real and GAN generated ANHIR images dyed with different stains. Left panel: real images. Middle and right panels: randomly selected D_2^L-GANs' generated samples after 40,000 generator iterations. Middle panel: `CNN G&D`. Right panel: `Eqv G` + `Inv D`.

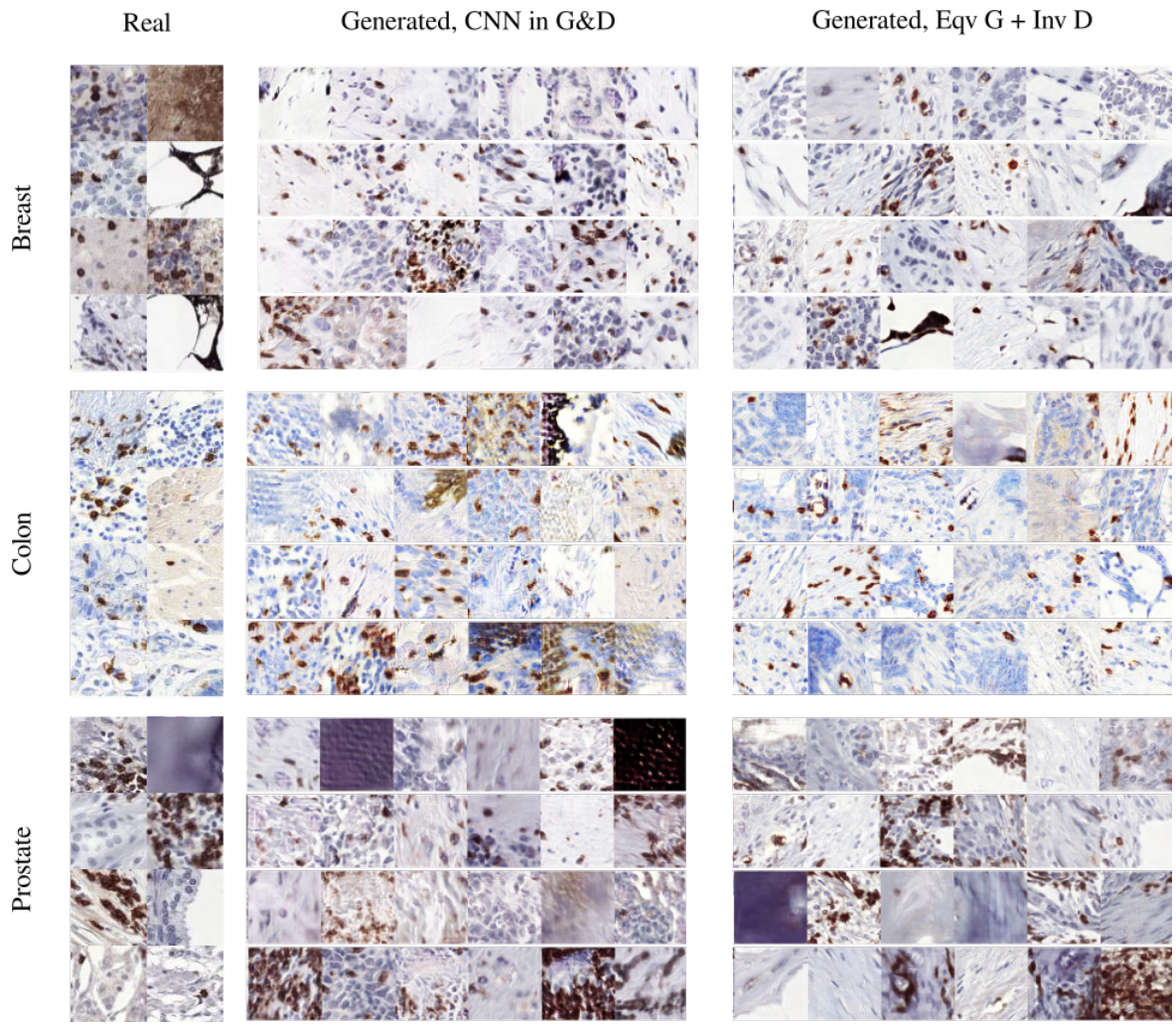


Figure 15. Real and GAN generated LYSTO images of breast, colon, and prostate cancer. Left panel: real images. Middle and right panels: randomly selected D_2^L-GANs' generated samples after 40,000 generator iterations. Middle panel: `CNN G&D`. Right panel: `Eqv G + Inv D`.

Table 4. The (min, median) of the FIDs over the course of training, averaged over three independent trials on the medical images, where the plus sign "+" after the data set, e.g., ANHIR+, denotes the presence of data augmentation during training.

Loss	Architecture	ANHIR	ANHIR+
RA	CNN G&D	(186, 523)	(184, 503)
	(I)Eqv G + Inv D	(100, 142)	(88, 140)
	Eqv G + Inv D	**(78, 125)**	**(84, 118)**
D_2^L	CNN G&D	(313, 485)	(347, 539)
	(I)Eqv G + Inv D	(120, 176)	(119, 177)
	Eqv G + Inv D	**(97, 157)**	**(90, 128)**

Loss	Architecture	LYSTO	LYSTO+
RA	CNN G&D	(281, 340)	(250, 312)
	(I)Eqv G + Inv D	(218, 272)	(212, 271)
	Eqv G + Inv D	**(175, 238)**	**(181, 227)**
D_2^L	CNN G&D	(289, 410)	(265, 376)
	(I)Eqv G + Inv D	(253, 343)	(244, 329)
	Eqv G + Inv D	**(205, 259)**	**(192, 259)**

G. Implementation Details

G.1. Common experimental setup

All models are trained using the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.0$ and $\beta_2 = 0.9$ (Zhang et al., 2019). Discriminators are updated twice after each generator update. An exponential moving average across iterations of the generator weights with $\alpha = 0.9999$ is used when sampling images (Brock et al., 2018).

G.2. RotMNIST

For RA-GAN, the training is stabilized by regularizing the discriminator $\gamma \in \Gamma$ with a zero-centered gradient panelty (GP) on the real distribution Q in the following form

$$R_1 = \frac{\lambda_1}{2} E_{x\sim Q} \|\nabla\gamma(x)\|_2^2. \tag{72}$$

We set the GP weight $\lambda_1 = 0.1$ according to (Dey et al., 2021). For the D_α^L-GAN, we use the one-sided GP as a soft constraint on the Lipschitz constant

$$R_2 = \lambda_2 E_{x\sim\rho_g} \max\{0, \|\nabla\gamma(x)\|^2 - 1\}, \tag{73}$$

where $\rho_g \sim TX + (1-T)Y$ (with $X \sim P_g$, $Y \sim Q$, and $T \sim \text{Unif}([0,1])$ all being independent.) The one-sided GP weight is set to $\lambda_2 = 10$ according to (Birrell et al., 2022). Unequal learning rates were set to $\eta_G = 0.0001$ and $\eta_D = 0.0004$ respectively. The neural architectures for the generators and discriminators are displayed in Table 5 and Table 6.

G.3. ANHIR and LYSTO

Similar to RotMNIST, the GP weights are set to $\lambda_1 = 0.1$ for the RA-GAN in (72) and $\lambda_2 = 10$ for the D_α^L-GAN in (73), and we consider only the case $\alpha = 2$. The learning rates were set to $\eta_G = 0.0001$ and $\eta_D = 0.0004$ respectively. ResNets instead of CNNs are used as baseline generators and discriminators, and the detailed architectural designs are specified in Table 7 and Table 8.

G.4. Architectures

Table 5. Generator architectures used in the RotMNIST experiment. ConvSN and C_4-ConvSN stand for spectrally-normalized 2D convolution and its C_4-equivariant counterpart. The incomplete attempt at building equivariant generators (`(I) Eqv G`) does not have the "C_4-symmetrization" layer. The C_8-equivariant generator (`Eqv G`, $\Sigma = C_8$) is built by replacing "3×3 C_4-ConvSN" with "5×5 C_8-ConvSN" while adjusting the number of filters to maintain a similar number of trainable parameters.

CNN Generator (`CNN G`)
Sample noise $z \in \mathbb{R}^{64} \sim \mathcal{N}(0, I)$ Embed label class y into $\hat{y} \in \mathbb{R}^{64}$ Concatenate z and $\hat{y}$ into $h \in \mathbb{R}^{128}$
Project and reshape h to $7 \times 7 \times 128$
3×3 ConvSN, $128 \to 512$
ReLU; Up $2\times$
3×3 ConvSN, $512 \to 256$
CCBN; ReLU; Up $2\times$
3×3 ConvSN, $256 \to 128$
CCBN; ReLU
3×3 ConvSN, $128 \to 1$
$\tanh()$

C_4-Equivariant Generator (`Eqv G`, $\Sigma = C_4$)
Sample noise $z \in \mathbb{R}^{64} \sim \mathcal{N}(0, I)$ Embed label class y into $\hat{y} \in \mathbb{R}^{64}$ Concatenate z and $\hat{y}$ into $h \in \mathbb{R}^{128}$
Project and reshape h to $7 \times 7 \times 128$
C_4-symmetrization of h
3×3 C_4-ConvSN, $128 \to 256$
ReLU; Up $2\times$
3×3 C_4-ConvSN, $256 \to 128$
CCBN; ReLU; Up $2\times$
3×3 C_4-ConvSN, $128 \to 64$
CCBN; ReLU
3×3 C_4-ConvSN, $64 \to 1$
C_4-Max Pool
$\tanh()$

Table 6. Discriminator architectures used in the RotMNIST experiment. The C_8-invariant discriminator (`Inv D`, $\Sigma = C_8$) is built by replacing "3×3 C_4-ConvSN" with "5×5 C_8-ConvSN" while adjusting the number of filters to maintain a similar number of trainable parameters.

CNN Discriminator (`CNN D`)
Input image $x \in \mathbb{R}^{28 \times 28 \times 1}$
3×3 ConvSN, $1 \to 128$
LeakyReLU; Avg. Pool
3×3 ConvSN, $128 \to 256$
LeakyReLU; Avg. Pool
3×3 ConvSN, $256 \to 512$
LeakyReLU; Avg. Pool
Global Avg. Pool into f
Embed label class y into $\hat{y}'$
Project $(\hat{y}', f)$ into a scalar

C_4-Invariant Discriminator (`Inv D`, $\Sigma = C_4$)
Input image $x \in \mathbb{R}^{28 \times 28 \times 1}$
3×3 C_4-ConvSN, $1 \to 64$
LeakyReLU; Avg. Pool
3×3 C_4-ConvSN, $64 \to 128$
LeakyReLU; Avg. Pool
3×3 C_4-ConvSN, $128 \to 256$
LeakyReLU; Avg. Pool
C_4-Max Pool
Global Avg. Pool into f
Embed label class y into $\hat{y}'$
Project $(\hat{y}', f)$ into a scalar

Table 7. Generator architectures used in the ANHIR and LYSTO experiments. The generator residual block (ResBlockG) is a cascade of [CCBN, ReLU, Up $2\times$, 3×3 ConvSN, CCBN, ReLU, 3×3 ConvSN] with a short connection consisting of [Up $2\times$, 1×1 ConvSN]. The equivariant residual block (D_4-ResBlockG) is built by replacing each component with its equivariant counterpart. The incomplete attempt at building equivariant generators (`(I)Eqv G`) does not have the "D_4-symmetrization" layer.

CNN Generator (`CNN G`)
Sample noise $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$ Embed label class y into $\hat{y} \in \mathbb{R}^{128}$ Concatenate z and $\hat{y}$ into $h \in \mathbb{R}^{256}$
Project and reshape h to $4 \times 4 \times 128$
ResBlockG, $128 \to 256$
ResBlockG, $256 \to 128$
ResBlockG, $128 \to 64$
ResBlockG, $64 \to 32$
ResBlockG, $32 \to 16$
BN; ReLU
3×3 ConvSN, $16 \to 3$
$\tanh()$

Equivariant Generator (`Eqv G`)
Sample noise $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$ Embed label class y into $\hat{y} \in \mathbb{R}^{128}$ Concatenate z and $\hat{y}$ into $h \in \mathbb{R}^{256}$
Project and reshape h to $4 \times 4 \times 128$
D_4-symmetrization of h
D_4-ResBlockG, $128 \to 90$
D_4-ResBlockG, $90 \to 45$
D_4-ResBlockG, $45 \to 22$
D_4-ResBlockG, $22 \to 11$
D_4-ResBlockG, $11 \to 5$
D_4-BN; ReLU
3×3 D_4-ConvSN, $5 \to 3$
D_4-Max Pool
$\tanh()$

Table 8. Discriminator architectures used in the ANHIR and LYSTO experiments. The discriminator residual block (ResBlockD) is a cascade of [ReLU, 3×3 ConvSN, ReLU, 3×3 ConvSN, Max Pool] with a short connection consisting of [1×1 ConvSN, Max Pool]. The equivariant residual block (D_4-ResBlockD) is built by replacing each component with its equivariant counterpart.

CNN Discriminator (`CNN D`)
Input image $x \in \mathbb{R}^{64\times64\times3}$
ResBlockD, $3 \to 16$
ResBlockD, $16 \to 32$
ResBlockD, $32 \to 64$
ResBlockD, $64 \to 128$
ResBlockD, $128 \to 256$
ReLU
Global Avg. Pool into f
Embed label class y into $\hat{y}'$
Project $(\hat{y}', f)$ into a scalar

Invariant Discriminator (`Inv D`)
Input image $x \in \mathbb{R}^{64\times64\times3}$
D_4-ResBlockD, $3 \to 5$
D_4-ResBlockD, $5 \to 11$
D_4-ResBlockD, $11 \to 22$
D_4-ResBlockD, $22 \to 45$
D_4-ResBlockD, $45 \to 90$
ReLU
D_4-Max Pool
Global Avg. Pool into f
Embed label class y into $\hat{y}'$
Project $(\hat{y}', f)$ into a scalar