On the Generative Utility of Cyclic Conditionals

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study whether and how can we model a joint distribution $p(x, z)$ using two conditional models $p(x|z)$ and $q(z|x)$ that form a cycle. This is motivated by the observation that deep generative models, in addition to a likelihood model $p(x|z)$, often also use an inference model $q(z|x)$ for data representation, but they rely on a usually uninformative prior $p(z)$ to define a joint distribution, which may render problems like posterior collapse and manifold mismatch. To explore the possibility to model a joint distribution using only $p(x|z)$ and $q(z|x)$, we study their *compatibility* and *determinacy*, corresponding to the existence and uniqueness of a joint distribution whose conditional distributions coincide with them. We develop a general theory for novel and operable equivalence criteria for compatibility, and sufficient conditions for determinacy. Based on the theory, we propose the CyGen framework for cyclic-conditional generative modeling, including methods to enforce compatibility and use the determined distribution to fit and generate data. With the prior constraint removed, CyGen better fits data and captures more representative features, supported by experiments showing better generation and downstream classification performance.

1 Introduction

Deep generative models have achieved a remarkable success in the past decade for generating realistic complex data x and extracting useful representations through their latent variable z. Variational auto-encoders (VAEs) [28, 42, 8, 9, 29, 47] follow the Bayesian framework and specify a prior $p(z)$ and a likelihood model $p(x|z)$, so that a joint distribution $p(x, z) = p(z)p(x|z)$ is defined for generative modeling (the joint induces a distribution $p(x)$ on data). An inference model $q(z|x)$ is also used, for extracting representations (and learning). Other frameworks like generative adversarial nets [17, 13, 14], flow-based models [12, 27, 18] and diffusion-based models [22, 45, 32] follow the same structure, with different choices of the conditional models $p(x|z)$ and $q(z|x)$ and training objectives. While for the prior $p(z)$, there is often not much knowledge for complex data (like images, text, audio), and these models widely adopt an uninformative prior such as a standard Gaussian. This however, introduces some side effects:

- *Posterior collapse* [9, 21, 40]: The standard Gaussian prior tends to squeeze $q(z|x)$ towards the origin for all x, which degrades the representativeness of the inferred z for x and hurts downstream tasks in the latent space like classification and clustering.
- *Manifold mismatch* [11, 15, 25]: Typically the likelihood model is continuous (keeps topology), so the standard Gaussian prior would restrict the modeled data distribution to a simply-connected support, which limits the capacity for fitting data from a non-(simply) connected support.

While there are works trying to mitigate the two problems, they either require a strong domain knowledge [25, 31, 35] or make inference harder [38].

One question then naturally emerges: *Can we model a joint distribution $p(x, z)$ only using the likelihood $p(x|z)$ and inference $q(z|x)$ models?* If we can, the limitations from the prior issue

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

are then removed from the root. Also, the inference model $q(z|x)$ is then no longer a struggling approximation to a predefined posterior but participate in defining the joint distribution. Modeling conditionals is also argued to be much easier than modeling marginals or the joint directly [1, 4, 5]. In some cases, one may even have better knowledge on the conditionals than on the prior, *e.g.* shift/rotation invariance of image representations (CNNs [34] / SphereNet [10]), and rules to extract frequency/energy features for audio [41]. It is then more natural and effective to incorporate this knowledge into the conditionals than using an uninformative prior.

In this paper, we explore such a possibility, and develop a systematic theory and a CyGen algorithmic framework for Cyclic-conditional Generative modeling. Theoretical analysis on the question breaks down into two sub-problems: when can two given cyclic conditionals correspond to a common joint, and when can they determine the joint. We term them *compatibility* and *determinacy* respectively, corresponding to the existence and uniqueness of a joint that induces the given conditionals. For this, we develop novel compatibility criteria and sufficient conditions for determinacy. In contrast to existing works [6, 4, 33], our results are operable (vs. existential) and self-contained (vs. requiring a marginal), and are general enough to cover both continuous and discrete cases. Our compatibility criteria are also equivalence conditions (vs. [2, 3]). The theory also makes practical algorithms possible for CyGen even without a prior model. We propose learning methods to enforce compatibility and to fit the determined distribution to data. A technique for efficient implementation is designed. We also propose data generation methods for CyGen. We show the practical utility of CyGen in both synthetic and real-world tasks, where the improved performance in downstream classification and data generation demonstrates the advantage to mitigate posterior collapse and manifold mismatch.

1.1 Related work

Denoising auto-encoders (DAEs). Auto-encoders consider extracting data features using a pair of encoder and decoder, where the decoder is for enforcing the encoder to preserve as representative information as possible. However, standard auto-encoders use deterministic encoder/decoder, which has insufficient determinacy for generative modeling (see Sec. 2.2.2). DAEs [49, 4, 5] use a probabilistic encoder and decoder for the robustness against random local perturbations. Its generative modeling effect is first noted through the equivalence to score matching (implying the utility for $p(x)$) in a Gaussian RBM [48] or infinitesimal noise [1] case. In more general cases, the utility to modeling the joint $p(x, z)$ is studied via the Gibbs chain, *i.e.* the Markov chain with transition kernel $p(x'|z')q(z'|x)$. Under a global [4, 33, 19] or local [5] shared support condition, the stationary distribution $\pi(x, z)$ of the Markov chain exists uniquely. But this approach does not give an explicit expression, and requires a large number of steps for the chain to converge for data generation and even for learning (Walkback [4], GibbsNet [33]). Moreover, the DAE loss is not proper for optimizing the inference model $q(z|x)$ as it promotes a mode-collapse behavior (see Sec. 3.2). As for compatibility, it is not really covered in DAEs. Existing results in DAEs only consider the statistical consistency (unbiasedness under infinite data) of the $p(x|z)$ estimator through fitting data from $p^*(x)q(z|x)$ [4, 5, 33, 19]. Particularly, they require a marginal $p^*(x)$ in advance, so that the joint is already defined by $p^*(x)q(z|x)$ regardless of $p(x|z)$, while compatibility (as well as determinacy) is a matter of only the two conditionals themselves. Moreover, in a non-asymptotic regime, compatibility is unlikely due to the support-covering bias of the forward KL loss they optimize [23, 46]. This makes $\pi(z|x)$ unnecessarily coincide with $q(z|x)$ [4], so one may not safely use $q(z|x)$ for inference and density ratio estimation in their methods. In contract, CyGen explicitly enforces compatibility.

Dual learning considers tasks where the conversions between two modalities in both directions are desired, *e.g.* machine translation [20, 52, 51] and image style transfer [26, 55, 53, 36]. Although we also consider both directions (generation and inference), the fundamental distinction is that in generative modeling there is no data of the latent variable z (not even unpaired)[1]. Technically, they did not consider determinacy; they require an external marginal from either a model or data to determine a joint or the other-side marginal. Their cycle-consistency loss [26, 55, 53] is a version of our compatibility criterion in a specific Dirac case, and we extend the loss to the more general absolutely continuous case (allowing probabilistic conversion).

2 Compatibility and Determinacy Theory

To be a generative model, a system needs to determine a distribution on the data variable x, and with a latent variable z, this amounts to determining a joint distribution on (x, z). This leads to the goal

[1] Constructing z data as samples from a prior returns to the discussion above.

of the analysis on the compatibility and determinacy of cyclic conditionals. In this section we carry out theoretical analysis on the conditions for compatibility and determinacy, which also inspires our novel CyGen framework for generative modeling. We begin with formalizing the problems.

Setup In this work we focus on cyclic conditionals on two random variables x and z (extension to more variables follows a similar way). Following the convention in machine learning, we call a "probability measure" as a "distribution", and call a "probability density/mass function" as a "density function". Denote the measure spaces of x and z as $(\mathbb{X}, \mathscr{X}, \xi)$ and $(\mathbb{Z}, \mathscr{Z}, \zeta)$[2], where $\mathscr{X}$, $\mathscr{Z}$ are the respective sigma-fields, and the base measures ξ, ζ (*e.g.*, Lebesgue measure on Euclidean spaces, counting measure on finite/discrete spaces) are sigma-finite (unnecessarily finite). We use $\mathcal{X} \in \mathscr{X}$, $\mathcal{Z} \in \mathscr{Z}$ to denote measurable sets, and use "$\subseteq^{\xi}$", "$\subseteq^{\xi}$" as the extensions of "$=$", "$\subseteq$" up to a set of ξ-measure-zero (Def. A.1). Note that we do not require any further structures such as topology, metric, or linearity, for the interest of the most general conclusions that unify Euclidean/manifold and finite/discrete spaces and allow $\mathbb{X}$, $\mathbb{Z}$ with different dimensions or types.

Joint and conditional distributions are defined on the product measure space $(\mathbb{X}\times\mathbb{Z}, \mathscr{X}\otimes\mathscr{Z}, \xi\otimes\zeta)$, where "$\times$" is the usual Cartesian product, $\mathscr{X}\otimes\mathscr{Z} := \sigma(\mathscr{X}\times\mathscr{Z})$ is the sigma-field generated by measurable rectangles from $\mathscr{X}\times\mathscr{Z}$, and $\xi\otimes\zeta$ is the product measure [7, Thm. 18.2]. Define the *slice* of $\mathcal{W} \in \mathscr{X}\otimes\mathscr{Z}$ at z as $\mathcal{W}_z := \{x \mid (x, z) \in \mathcal{W}\}$, and its *projection* onto $\mathbb{Z}$ as $\mathcal{W}^{\mathbb{Z}} := \{z \mid \exists x \in \mathbb{X} \text{ s.t. } (x, z) \in \mathcal{W}\} \in \mathscr{Z}$ (Appx. A.3). Similarly denote the *marginal* of a joint π onto $\mathbb{Z}$ as $\pi^{\mathbb{Z}}(\mathcal{Z})$. To keep the same level of generality, we follow the general definition of conditionals [7, p.457]: the conditional $\mu(\mathcal{X}|z)$ of joint π is the density function of $\pi(\mathcal{X}\times\cdot)$ w.r.t $\pi^{\mathbb{Z}}$ (Appx. A.4). We highlight the key characteristic under this generality that $\mu(\cdot|z)$ can be arbitrary on a set of $\pi^{\mathbb{Z}}$-measure-zero, particularly, outside the support of $\pi^{\mathbb{Z}}$. Appx. A provides more background on measure theory. The goals of the analysis can be then formalized below.

Definition 2.1 (compatibility and determinacy)**.** We say two conditionals $\mu(\mathcal{X}|z)$, $\nu(\mathcal{Z}|x)$ are *compatible*, if there exists a joint distribution π on $(\mathbb{X}\times\mathbb{Z}, \mathscr{X}\otimes\mathscr{Z})$ such that $\mu(\mathcal{X}|z)$ and $\nu(\mathcal{Z}|x)$ are its conditional distributions. We say two compatible conditionals have *determinacy* on a set $\mathcal{S} \in \mathscr{X}\otimes\mathscr{Z}$, if there is only one joint distribution concentrated on $\mathcal{S}$ that makes them compatible.

We now analyze the conditions for compatibility and determinacy in the absolutely continuous and the Dirac cases. Distributions have qualitatively different behaviors in the two cases, and they correspond to different types of generative models. Note that the concepts only involve the two conditionals, so we desire conditions free of any marginal or joint distributions (vs. DAE results).

2.1 Absolutely Continuous Case

We first consider the case where for any $z \in \mathbb{Z}$ and any $x \in \mathbb{X}$,[3] the conditionals $\mu(\cdot|z)$ and $\nu(\cdot|x)$ are either absolutely continuous (w.r.t ξ and ζ, resp.) [7, p.448], or zero in the sense of a measure. This is equivalent to that they have density functions $p(x|z)$ and $q(z|x)$ (non-negative by definition; may integrate to zero), including "smooth" distributions on Euclidean spaces or manifolds, as well as *all* distributions on finite/discrete spaces. We also refer to such a conditional by its density. As conditionals are often specified in the form of density, this case covers many generative model frameworks, notably variational auto-encoders (VAEs) [28, 42] and diffusion-based models [22, 45].

2.1.1 Compatibility criterion in the absolutely continuous case

To put the concept into practical use, we need an operable criterion to tell the compatibility of two given conditionals. In the absolutely continuous case, one may expect that when $p(x|z)$ and $q(z|x)$ are compatible, the joint is also absolutely continuous (w.r.t $\xi\otimes\zeta$) with some density $p(x, z)$. This intuition is verified by Lem. C.1 in Appx. C.1. For a sloppy inspiration, one could then safely apply density-function calculus and get $\frac{p(x|z)}{q(z|x)} = \frac{p(x,z)}{p(z)} / \frac{p(x,z)}{p(x)} = \frac{p(x)}{p(z)}$ factorizes into a function of x and a function of z. Conversely, if the ratio factorizes as such $\frac{p(x|z)}{q(z|x)} = a(x)b(z)$, one could get $p(x|z)\frac{1}{Ab(z)} = q(z|x)\frac{a(x)}{A}$ where $A := \int_{\mathbb{X}} a(x)\xi(\mathrm{d}x)$, which defines a joint density and compatibility is achieved. This intuition leads to the classical compatibility criterion [2, Thm. 4.1; 3, Thm. 1].

[2]The symbol $\mathbb{Z}$ overwrites the symbol for the set of integers, which is not involved in this paper.

[3]There may be problems if absolute continuity holds only for ζ-a.e. z and ξ-a.e. x; see Example C.2.

However, it is more complicated than imagined. Berti et al. [6, Example 9] point out that the classical criterion is only sufficient but *not necessary*. The subtlety is about on which region does this factorization have to hold. The classical criterion requires it to be the positive region of $p(x|z)$ and also coincide with that of $q(z|x)$. But as mentioned, conditional $\mu(\cdot|z)$ can be arbitrary outside the support of the marginal $\pi^{\mathbb{Z}}$ (similarly for $\nu(\cdot|x)$), which may lead to additional positive regions that violate the requirement.[4] To address the problem, Berti et al. [6] give an equivalence criterion (Thm. 8), but it is *existential* thus less useful as the definition of compatibility is itself existential. Moreover, these criteria are restricted to either Euclidean or discrete spaces.

Next we give our *equivalence* criterion that is *operable*. In addressing the subtlety with regions, we first introduce a related concept that helps identify appropriate regions.

Definition 2.2 ($\xi\otimes\zeta$-complete component)**.** For a set $\mathcal{W} \in \mathscr{X}\otimes\mathscr{Z}$, we say that a set $\mathcal{S} \in \mathscr{X}\otimes\mathscr{Z}$ is a *$\xi\otimes\zeta$-complete component* of $\mathcal{W}$, if $\mathcal{S}^\sharp \cap \mathcal{W} \overset{\xi\otimes\zeta}{=} \mathcal{S}$, where $\mathcal{S}^\sharp := \mathcal{S}^{\mathbb{X}}\times\mathbb{Z} \cup \mathbb{X}\times\mathcal{S}^{\mathbb{Z}}$ is the *stretch* of $\mathcal{S}$.

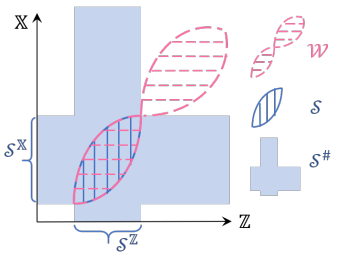

Figure 1: Illustration of a $\xi\otimes\zeta$-complete component $\mathcal{S}$ of $\mathcal{W}$.

Fig. 1 illustrates the concept. Roughly, the stretch $\mathcal{S}^\sharp$ of $\mathcal{S}$ represents the region where the conditionals are a.s. determined if $\mathcal{S}$ is the *support*[5] of the joint. If $\mathcal{S}$ is a complete component of $\mathcal{W}$, it is complete under stretching and intersecting with $\mathcal{W}$. Such a set $\mathcal{S}$ is an a.s. subset of $\mathcal{W}$ (Lem. B.12), while has a.s. the same slice as $\mathcal{W}$ does at any $z \in \mathcal{S}^{\mathbb{Z}}$ and $x \in \mathcal{S}^{\mathbb{X}}$. This is critical for the normalizedness of distributions in our criterion. Appx. B.3 shows more facts about the concept. With this concept, our compatibility criterion is presented below.

Theorem 2.3 (compatibility criterion, absolutely continuous)**.** *Let $p(x|z)$ and $q(z|x)$ be the density functions of two everywhere absolutely continuous (or zero) conditional distributions, and define:*

$$\mathcal{P}_z := \{x \mid p(x|z) > 0\}, \mathcal{P}_x := \{z \mid p(x|z) > 0\},$$
$$\mathcal{Q}_z := \{x \mid q(z|x) > 0\}, \mathcal{Q}_x := \{z \mid q(z|x) > 0\}.$$

Then they are compatible, if and only if they have a complete support $\mathcal{S}$*, defined as a* ***(i)*** *$\xi\otimes\zeta$-complete component of both*

$$\mathcal{W}_{p,q} := \bigcup_{z:\mathcal{P}_z \subseteq^\xi \mathcal{Q}_z} \mathcal{P}_z \times \{z\},\ \mathcal{W}_{q,p} := \bigcup_{x:\mathcal{Q}_x \subseteq^\zeta \mathcal{P}_x} \{x\}\times\mathcal{Q}_x,$$

such that: ***(ii)*** *$\mathcal{S}^{\mathbb{X}} \subseteq^\xi \mathcal{W}^{\mathbb{X}}_{q,p}$, $\mathcal{S}^{\mathbb{Z}} \subseteq^\zeta \mathcal{W}^{\mathbb{Z}}_{p,q}$,* ***(iii)*** *$(\xi\otimes\zeta)(\mathcal{S}) > 0$, and* ***(iv)*** *$\frac{p(x|z)}{q(z|x)}$ factorizes as $a(x)b(z)$, $\xi\otimes\zeta$-a.e. on $\mathcal{S}$,*[6] *where* ***(v)*** *$a(x)$ is ξ-integrable on $\mathcal{S}^{\mathbb{X}}$. For sufficiency,*

$$\pi(\mathcal{W}) := \frac{\int_{\mathcal{W}\cap\mathcal{S}} q(z|x)|a(x)|(\xi\otimes\zeta)(\mathrm{d}x\mathrm{d}z)}{\int_{\mathcal{S}^{\mathbb{X}}}|a(x)|\xi(\mathrm{d}x)}, \quad (1)$$

$\forall \mathcal{W} \in \mathscr{X}\otimes\mathscr{Z}$, is a compatible joint of them.

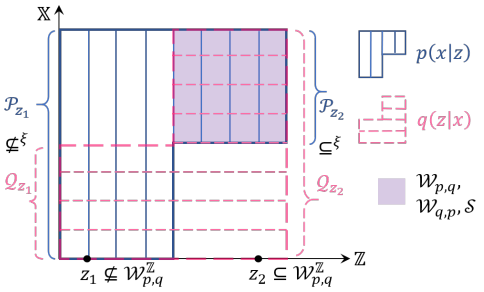

Figure 2: Illustration of our compatibility criterion in the absolutely continuous case (Thm. 2.3). The conditionals are uniform on the respective regions. For condition **(i)**, that $\mathcal{P}_z \subseteq^\xi \mathcal{Q}_z$ is *not* satisfied on the left half, *e.g.* z_1, so $\mathcal{W}_{p,q}$ does not cover the left half; it is satisfied on the right half, *e.g.* z_2, so $\mathcal{W}_{p,q}$ is composed of slices $\mathcal{P}_z$ on the right half, making the top-right quadrant (shaded). Similarly, $\mathcal{W}_{q,p}$ is the same region, and it is a $\xi\otimes\zeta$-complete component of itself. It also satisfies other conditions thus is a complete support $\mathcal{S}$.

Fig. 2 shows an illustration of the conditions. To understand the criterion, conditions **(iv)** and **(v)** stem from the starting inspiration, which also shows a hint for Eq. (1). Other conditions handle the subtlety to find a region $\mathcal{S}$ where **(iv)** and **(v)** must hold. This is essentially the support of a compatible joint π as there is no need and no way to control conditionals outside the support.

For necessity, informally, if z is in the support of $\pi^{\mathbb{Z}}$, then $p(x|z)$ determines the distribution on $\mathbb{X}\times\{z\}$; particularly, the joint π should be a.e. positive on $\mathcal{P}_z$, which in turn asks $q(z|x)$ to be so.

[4] The flexibility of $p(x|z)$ on a ξ-measure-zero set for a given z (similarly for $q(z|x)$) is not a vital problem, as one can adjust the conditions to hold only a.e.

[5] While the typical definition of support requires a topological structure which is absent under our generality, Def. B.8 in Appx. B.1 defines such a concept for absolutely continuous distributions.

[6] Formally, there exist functions a on $\mathcal{S}^{\mathbb{X}}$ and b on $\mathcal{S}^{\mathbb{Z}}$ s.t. $(\xi\otimes\zeta)\{(x,z) \in \mathcal{S} \mid \frac{p(x|z)}{q(z|x)} \neq a(x)b(z)\} = 0$.

This means $\mathcal{P}_z \subseteq^{\xi} \mathcal{Q}_z$ (unnecessary equal, since $q(z|x)$ is "out of control" outside the joint support), which leads to the definitions of $\mathcal{W}_{p,q}$ and $\mathcal{W}_{q,p}$. The joint support should be contained within the two sets in order to avoid *support conflict* (*e.g.*, although the bottom-left quadrant in Fig. 2 is part of the intersection of positive regions of the conditionals, a joint on it is required by $p(x|z)$ to also cover the top-left, on which $q(z|x)$ does not agree). Condition **(i)** indicates $\mathcal{S} \subseteq^{\xi\otimes\zeta} \mathcal{W}_{p,q}$ and $\mathcal{W}_{q,p}$ so $\mathcal{S}$ satisfies this requirement and also makes the ratio in **(iv)** a.e. well-defined. The complete-component condition in **(i)** also makes the conditionals *normalized* on $\mathcal{S}$: as mentioned, such an $\mathcal{S}$ has a.s. the same slice as $\mathcal{W}_{p,q}$ does for a given z in support $\mathcal{S}^{\mathbb{Z}}$, so the integral of $p(x|z)$ on $\mathcal{S}_z$ is the same as that on $(\mathcal{W}_{p,q})_z$ which is 1 by construction (similarly for $q(z|x)$). Example C.3 in Appx. C.2 shows $\mathcal{W}_{p,q} \cap \mathcal{W}_{q,p}$ is unnecessarily appropriate. Conditions **(ii)** and **(iii)** cannot be guaranteed by condition **(i)** (Example B.13), while are needed to rule out special cases (Lem. B.14; Example B.15). Appx. C.2 gives a rigorous proof following this intuition. Although the criterion relies on the *existence* of such a complete support, candidates are few (if any), so it is *operable*.

2.1.2 Determinacy in the absolutely continuous case

When compatible, absolutely continuous cyclic conditionals are very likely to have determinacy.

Theorem 2.4 (determinacy, absolutely continuous)**.** *Let $p(x|z)$ and $q(z|x)$ be two compatible conditional densities and $\mathcal{S}$ be a complete support that makes them compatible (also necessary due to Thm. 2.3). Suppose that $\mathcal{S}_z \overset{\xi}{=} \mathcal{S}^{\mathbb{X}}$, for ζ-a.e. z on $\mathcal{S}^{\mathbb{Z}}$, or $\mathcal{S}_x \overset{\zeta}{=} \mathcal{S}^{\mathbb{Z}}$, for ξ-a.e. x on $\mathcal{S}^{\mathbb{X}}$. Then their compatible joint supported on $\mathcal{S}$ is unique, which is given by Eq.* (1).

Proof is given in Appx. C.4. The condition in the theorem roughly means that the complete support $\mathcal{S}$ is "rectangular". From the perspective of Markov chain, this corresponds to the irreducibility of the Gibbs chain ($z_1 \sim q(z|x_0)$, $x_1 \sim p(x|z_1)$, ...) for the unique existence of a stationary distribution. When the conditionals have multiple such complete supports, on each of which the compatible joint is unique, while globally on $\mathbb{X}\times\mathbb{Z}$, they may have multiple compatible joints. In general, determinacy in the absolutely continuous case is sufficient, including the following common case (*e.g.*, in VAEs).

Corollary 2.5. *We call two conditional densities* have a.e.-full supports*, if $p(x|z) > 0, q(z|x) > 0$ for $\xi\otimes\zeta$-a.e. (x, z). If they are compatible, then their compatible joint is unique, since $\mathbb{X}\times\mathbb{Z}$ is the $\xi\otimes\zeta$-unique complete support (Prop. C.4 in Appx. C.3), which satisfies the condition in Thm. 2.4.*

2.2 Dirac Case

Many other prevailing generative models, including generative adversarial networks (GANs) [17] and flow-based models [12, 27], use a deterministic function $x = f(z)$ as the likelihood model. In such cases, the conditional $\mu(\mathcal{X}|z) = \delta_{f(z)}(\mathcal{X}) := \mathbb{I}[f(z) \in \mathcal{X}], \forall \mathcal{X} \in \mathscr{X}$ is a Dirac measure. Note it may not have a density function *e.g.* when ξ assigns zero to all single-point sets, like the Lebesgue measure on Euclidean spaces, so we keep the measure notion. This case is not exclusive to the absolutely continuous case: a Dirac conditional on a discrete space is also absolutely continuous.

2.2.1 Compatibility criterion in the Dirac case

Compatibility criterion is easier to imagine in this case. As illustrated in Fig. 3, it is roughly that the other-way conditional $\nu(\cdot|x)$ could find a way to put its mass only on the curve; otherwise support conflict is rendered.

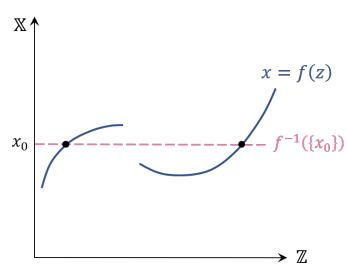

Figure 3: Illustration of our compatibility criterion in the Dirac case (Thm. 2.6).

Theorem 2.6 (compatibility criterion, Dirac)**.** *Suppose that $\mathscr{X}$ contains every single-point set: $\{x\} \in \mathscr{X}, \forall x \in \mathbb{X}$. Conditional distribution $\nu(\mathcal{Z}|x)$ is compatible with $\mu(\mathcal{X}|z) := \delta_{f(z)}(\mathcal{X})$ where function $f : \mathbb{Z} \to \mathbb{X}$ is $\mathscr{X}/\mathscr{Z}$-measurable*[7]*, if and only if there exists $x_0 \in \mathbb{X}$ such that $\nu(f^{-1}(\{x_0\})|x_0) = 1$.*

Proof is given in Appx. C.6. Note that since the empty set always gets measure zero, such an x_0 must be in the image space $f(\mathbb{Z})$. What may be counter-intuitive is that $\nu(\cdot|x)$ is not required to concentrate on the curve for *any* x; one x_0 is sufficient as $\delta_{(x_0, f(x_0))}$ is a compatible joint. Nevertheless, in practice one often desires

[7]For meaningful discussion, we require f to be $\mathscr{X}/\mathscr{Z}$-measurable, which includes any function between discrete sets and continuous functions when $\mathscr{X}$ and $\mathscr{Z}$ are the Borel sigma-fields.

the compatibility to hold over a set $\mathcal{X}$ to be useful. When $\nu(\cdot|x)$ is also chosen to be in the Dirac form $\delta_{g(x)}$, this can be achieved by minimizing $\mathbb{E}_{p(x)} D\big(x, f(g(x))\big)$, where $p(x)$ is a distribution on $\mathcal{X}$ and D is a metric on $\mathbb{X}$. This is the cycle-consistency loss used in dual learning [26, 55, 53, 36]. When f is invertible, achieving zero of this loss is also necessary, as there is only one element in $f^{-1}(x)$ for any $x \in \mathcal{X}$.

2.2.2 Determinacy in the Dirac case

As explained above, for any x_0 satisfying the condition, the two conditionals have determinacy on this point $\{x_0\}$ with compatible joint $\delta_{(x_0, f(x_0))}$. Similar to the absolutely continuous case, when such x_0 is not unique, there are multiple compatible joints globally on $\mathbb{X}\times\mathbb{Z}$. This is however often the case, as compatibility is usually desired to hold on a set $\mathcal{X}$ (with multiple elements) to be useful, but a joint distribution on $\mathcal{X}$ cannot be determined uniquely. This meets one's intuition: compatible Dirac conditionals can only determine a curve in $\mathbb{X}\times\mathbb{Z}$ from two sides, but the distribution on the curve is undetermined. One exception is when $f(z) \equiv x_0$ is constant, so this x_0 is the only candidate in the theorem. The joint then degenerates to a distribution on $\mathbb{Z}$, which is fully determined by $\nu(\cdot|x_0)$.

In general, determinacy in the Dirac case is insufficient, and this type of generative models (GANs, flow-based models) have to specify a prior to define a joint.

3 Generative Modeling using Cyclic Conditionals

The theory suggests it is possible that two conditionals achieve compatibility and particularly a sufficient determinacy, so that they can define a useful joint distribution without specifying a prior for generative modeling. For a sufficient determinacy, a deterministic likelihood or inference model is undesired due to the analysis on the Dirac case, so we consider absolutely continuous conditionals and use probabilistic likelihood and inference models, represented by parameterized density models $p_\theta(x|z)$, $q_\phi(z|x)$. We take $\mathbb{X}$ and $\mathbb{Z}$ as Euclidean spaces which is common in generative modeling, and assume the conditional densities have a.e.-full supports and are differentiable, which hold for common architectures.

As a generative model, the conditionals are first expected to be compatible as the prerequisite for determinacy. They then need to fit their determined distribution $p_{\theta,\phi}(x)$ to the data distribution $p^*(x)$. After learning (achieving the above two goals), they are also expected to generate novel data samples. We now develop methods for the three tasks, and call the resulting framework CyGen for "Cyclic Generative" modeling.

3.1 Enforcing Compatibility

In the a.e.-full support case, the entire space $\mathbb{X}\times\mathbb{Z}$ is the only possible complete support (Prop. C.4 in Appx. C.3), so for compatibility, condition **(iv)** in Thm. 2.3 is the most critical one. For this, we do not have to find functions $a(x)$, $b(z)$ in Thm. 2.3, but only need to enforce such a factorization. So we propose the following loss function to enforce compatibility:

$$(\min_{\theta,\phi})\ \ C(\theta,\phi) := \mathbb{E}_{\rho(x,z)} \big\| \nabla_x \nabla_z^\top r_{\theta,\phi}(x,z) \big\|_F^2, \quad r_{\theta,\phi}(x,z) := \log\big(p_\theta(x|z)/q_\phi(z|x)\big), \tag{2}$$

where ρ is some absolutely continuous reference distribution on $\mathbb{X}\times\mathbb{Z}$, which can be taken as $p^*(x)q_\phi(z|x)$ in practice as it gives samples to estimate the expectation. When $C(\theta,\phi) = 0$, we have $\nabla_x \nabla_z^\top r_{\theta,\phi}(x,z) = 0$, $\xi\otimes\zeta$-a.e. [7, Thm. 15.2(ii)], which means $\nabla_z^\top r_{\theta,\phi}(x,z) = \tilde{b}(z)$ hence $r_{\theta,\phi}(x,z) = \tilde{B}(z) + \tilde{a}(x)$, $\xi\otimes\zeta$-a.e., for some functions $\tilde{b}(z)$, $\tilde{B}(z)$, $\tilde{a}(x)$ s.t. $\tilde{b}(z) = \nabla_z \tilde{B}(z)$. So the ratio $r_{\theta,\phi}(x,z)$ factorizes $\xi\otimes\zeta$-a.s.

This loss also reveals some interesting facts about the vanilla VAE [28], which adopts Gaussian models $p_\theta(x|z) := \mathcal{N}(x|f_\theta(z), \sigma^2_{x|z} I_{d_\mathbb{X}})$ and $q_\phi(z|x) := \mathcal{N}(z|g_\phi(x), \sigma^2_{z|x} I_{d_\mathbb{Z}})$. Minimizing the compatibility loss Eq. (2) then amounts to enforcing $(\nabla_z f_\theta^\top(z))^\top = (\sigma^2_{x|z}/\sigma^2_{z|x})\nabla_x g_\phi^\top(x)$. As the two sides indicate the equation is constant of x and z respectively, it must be a constant, so $f_\theta(z)$ and $g_\phi(x)$ must be affine. This conclusion coincides with the theory on additive noise models [54, 39], and also explains the empirical observation that the latent space of such VAEs is quite linear [44]. This finding reveals that the expectation to use deep neural networks for nonlinearity will be disappointed in the vanilla VAE. So we use a non-Gaussian model, *e.g.* a flow-based model, for at least one of $p_\theta(x|z)$ and $q_\phi(z|x)$ (often the latter).

For implementation, direct Jacobian evaluation for Eq. (2) is of complexity $O(d_{\mathbb{X}} d_{\mathbb{Z}})$, which is often prohibitively large. We thus propose a stochastic but much cheaper and unbiased method based on Hutchinson's trace estimator [24]: $\operatorname{tr}(A) = \mathbb{E}_{p(\eta)}[\eta^\top A \eta]$, where η is any random vector with zero mean and identity covariance (*e.g.*, a standard Gaussian). As the function within expectation is $\left\|\nabla_x \nabla_z^\top r\right\|_F^2 = \left\|\nabla_z \nabla_x^\top r\right\|_F^2 = \operatorname{tr}\left((\nabla_z \nabla_x^\top r)^\top \nabla_z \nabla_x^\top r\right)$, applying the estimator yields:

$$\left(\min_{\theta,\phi}\right) \; C(\theta,\phi) = \mathbb{E}_{\rho(x,z)} \mathbb{E}_{p(\eta_x)} \left\|\nabla_z \left(\eta_x^\top \nabla_x r_{\theta,\phi}(x,z)\right)\right\|_2^2, \tag{3}$$

which reduces the complexity of gradient evaluation to $O(d_{\mathbb{X}} + d_{\mathbb{Z}})$.

As concluded from the above analysis, it is preferred to use a flow-based model for the inference model $q_\phi(z|x)$. But in common choices [29, 47] the evaluation of the inverse of the flow is not tractable, so density evaluation is only possible at samples the flow model draws but not at an arbitrary point. This disables the use of automatic gradient estimation tools for the gradients in the compatibility loss. Appx. D.1 explains this problem in detail and shows our solution.

3.2 Fitting Data

In the a.e.-full support case, compatible conditionals enjoy a quite sufficient determinacy that the compatible joint is unique on $\mathbb{X} \times \mathbb{Z}$ (Cor. 2.5). To fit data distribution $p^*(x)$, an explicit expression of the model-determined distribution $p_{\theta,\phi}(x)$ is required. For this, Eq. (1) is not very helpful as we do not have explicit expressions of $a(x)$, $b(z)$ in Thm. 2.3. But when compatibility is given, we can safely use density-function calculus for deduction. We can thus write $p_{\theta,\phi}(x) = 1/\frac{1}{p_{\theta,\phi}(x)} = 1/\int_{\mathbb{Z}} \frac{p_{\theta,\phi}(z')}{p_{\theta,\phi}(x)} \zeta(\mathrm{d}z') = 1/\int_{\mathbb{Z}} \frac{q_\phi(z'|x)}{p_\theta(x|z')} \zeta(\mathrm{d}z') = 1/\mathbb{E}_{q_\phi(z'|x)}[1/p_\theta(x|z')]$, which is an explicit expression in terms of the two conditionals. Although other expressions are possible, this one is preferred as it is simple, and as $d_{\mathbb{Z}}$ is often much smaller than $d_{\mathbb{X}}$ in generative modeling, Monte-Carlo expectation estimation in $\mathbb{Z}$ has lower variance than in $\mathbb{X}$. We can thus fit data through maximum likelihood estimate:

$$\left(\min_{\theta,\phi}\right) \quad -\mathbb{E}_{p^*(x)}[\log p_{\theta,\phi}(x)] = \mathbb{E}_{p^*(x)}[\log \mathbb{E}_{q_\phi(z'|x)}[1/p_\theta(x|z')]]. \tag{4}$$

The loss function can be estimated using the reparameterization trick [28] to reduce variance, and the `logsumexp` trick is adopted for numerical stability. This expression can also serve for data likelihood evaluation. The final training procedure for CyGen is the joint optimization of Eqs. (3, 4).

An interesting observation is that, this loss function is an upper bound of the DAE [49, 4] loss function $-\mathbb{E}_{p^*(x) q_\phi(z'|x)}[\log p_\theta(x|z')]$, so minimizing the DAE loss may not serve for maximizing data likelihood $\mathbb{E}_{p^*(x)}[\log p_{\theta,\phi}(x)]$, but would instead collapse $q_\phi(z|x)$ to be peaked at $\operatorname{argmin}_{z'} p_\theta(x|z')$ for every x in the support of $p^*(x)$. In fact, the DAE loss is proper for optimizing $p_\theta(x|z)$ to match the joint $p^*(x) q_\phi(z|x)$, but not proper for optimizing $q_\phi(z|x)$.

3.3 Data Generation

Generating novel data points from $p_{\theta,\phi}(x)$ modeled by the two conditionals $p_\theta(x|z)$ and $q_\phi(z|x)$ is not as straightforward as the typical prior-likelihood generative models. But it is still practical.

Gibbs Sampling has been considered [4, 33] to draw samples from a joint distribution represented by two compatible conditionals. The sampling procedure is to alternately draw from $p_\theta(x|z)$ and $q_\phi(z|x)$, *i.e.*, $z_t \sim q_\phi(z|x_{t-1})$, $x_t \sim p_\theta(x|z_t)$, $t = 1, 2, \cdots$. In the a.e.-full support case, the chain has a unique stationary distribution since ergodicity is satisfied, and given compatibility, this is exactly the joint that they determine. Gibbs sampling is easy to implement, but often converges insufficiently fast.

Dynamics-Based MCMC are often more efficient, due to the leverage of gradient information. Generating samples from the marginal distribution $p_{\theta,\phi}(x)$ only requires an unnormalized density function of $p_{\theta,\phi}(x)$ for which CyGen could well serve when compatible: $p_{\theta,\phi}(x) = \frac{p_{\theta,\phi}(x)}{p_{\theta,\phi}(z)} p_{\theta,\phi}(z) = \frac{p_\theta(x|z)}{q_\phi(z|x)} p_{\theta,\phi}(z) \propto \frac{p_\theta(x|z)}{q_\phi(z|x)}, \forall z \in \mathbb{Z}$. In practice, this z can be taken as a sample from $q_\phi(z|x)$ to lie in a high probability region for a more confident estimate. For a representative method, stochastic gradient Langevin dynamics (SGLD) [50] gives the following transition:

$$x^{(t+1)} = x^{(t)} + \varepsilon \nabla_{x^{(t)}} \log \frac{p_\theta(x^{(t)}|z^{(t)})}{q_\phi(z^{(t)}|x^{(t)})} + \sqrt{2\varepsilon}\eta_x^{(t)}, \text{where } z^{(t)} \sim q_\phi(z|x^{(t)}), \eta_x^{(t)} \sim \mathcal{N}(0, I_{d_{\mathbb{X}}}),$$

and ε is a step size parameter. Method to draw from the prior $p_{\theta,\phi}(z)$ can be developed symmetrically.

4 Experiments

We demonstrate the power of CyGen on a variety of generation and inference tasks on synthetic and real-world datasets. We compare the representation learning and data generation ability of CyGen against other bi-directional methods that use a specified prior, including VAE and BiGAN.

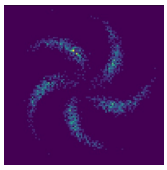

Figure 4: Synthetic Data

4.1 Synthetic Experiments

4.1.1 Generation and Inference

We first fit the model on a 2D toy dataset as illustrated in Fig. 4. Experiments are carried out by assuming $p_\theta(x|z)$ is a unimodal Gaussian distribution and the inference model $q_\phi(z|x)$ is a Sylvester flow [47]. All methods use the same encoder and decoder structure detailed in Appendix.

The learned distributions are visualized in Table 1. We use isotropic Gaussian as the prior for VAE and BiGAN. We found that with the same encoder/decoder structure, CyGen can fit multi-modal distribution well, showing that we are not restricted to the manifold mismatch problem. BiGAN and VAE have problems in fitting the regions between two disconnected clusters.

We then show the latent space learned by CyGen is more expressive by visualizing the aggregated posterior of the different clusters of the data samples in Table 1. From the figures, we see that the inference model learned by CyGen better separates data clusters in the latent space, so does not suffer from the posterior collapse problem.

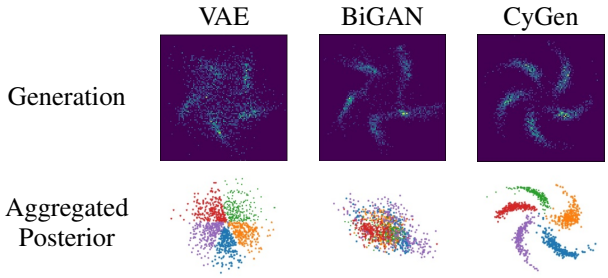

Table 1: Generation and Aggregated Posterior Comparisons of VAE, BiGAN and CyGen.

4.1.2 Impact of Compatible Loss

Next we compare different generation methods proposed in Section 3.3 and validate the necessity of the compatible loss. We plot the generation results of models with different compatible losses using Gibbs sampling and latent space Langevin dynamics.

From Table 2, we see that a smaller compatible loss indicates better generation qualities of both Gibbs and Langevin generation mechanisms. Moreover, Gibbs sampling suffers from mode collapse heavily when the compatible loss is large, while the Langevin dynamics can guarantee good generation qualities even if the conditionals are not exactly compatible. Therefore, data generation by running Langevin dynamics is more robust when two conditional are not compatible.

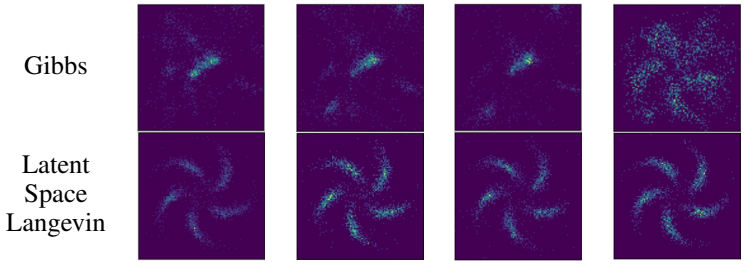

Table 2: Generation and Aggregated Posterior Comparisons of VAE, BiGAN and CyGen. Compatible loss from left to right: 7.0e3, 7.7e3, 6.2e3, 4.6e3.

4.2 Real-World Data

We compare the generation ability of CyGen on the MNIST and SVHN datasets against VAE and Denoising Auto-Encoder (DAE) using same encoder/decoder structure structure. Notice that two unimodal Gaussian conditionals cannot be compatible to a multi-modal joint Gaussian distributions. To enforce compatibility, our inference model uses the Householder flow structure. VAE models are

trained for 100 epochs. Data samples generated by VAE are obtained by pushing a Gaussian prior z through the likelihood model $p_\theta(x|z)$.

Based on the pretrained model from VAE, we then train CyGen as described in Section 3.2. For CyGen generation, we start from a random Gaussian noise z_0 and draw x_0 from the inference model $q_\phi(\cdot|z_0)$. Then we run Langevin dynamics on the data space x as described in Section 3.3. Generations from VAE are obtained by pushing isotropic Gaussian through the generation $p_\theta(x|z)$ and generations of DAE are obtained by Gibbs sampling starting from an isotropic Gaussian distribution on the latent space. As illustrated in the Fig 5, CyGen improves generation quality on MNIST dataset compared with the pretrained VAE model. DAE models has poor generation quality due to un-compatible conditionals $p_\theta(x|z)$ and $q_\phi(z|x)$.

We then show that the latent features sampled from the inference model $q_\phi(z|x)$ learned by CyGen is more expressive for downstream classification task. After training the generation and inference model, we sample the latent features directly from the inference model $q_\phi(z|x)$, and then train a 2-layer MLP classifier with 10 hidden units on top of the sampled latent features. We compare the classification result on test dataset against bi-generation mechanisms including VAE, DAE, BiGAN and GibbsNet. GANs with flow structured encoder is notoriously difficult to train and we show the results of BiGAN/GibbsNet with the same latent space dimension from [33], i.e., inference model without a flow structure. All the inference models have the same latent space dimension CyGen, and we train the classifier for 100 epochs using the same learning rate and optimizer. The downstream classification accuracy is shown in Table 3.

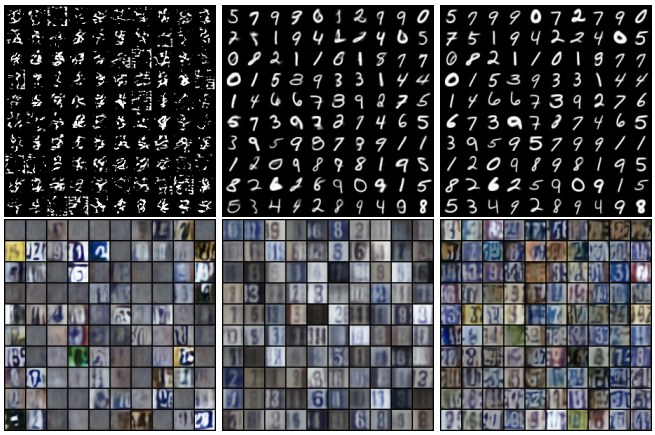

Figure 5: Generated samples on MNIST/SVHN dataset by DAE (left), VAE (middle) and CyGen running Langevin dynamics on data space x (right).

The empirical results demonstrate that the inference mapping $q_\phi(z|x)$ of CyGen preserves more information than dual generation mechanisms like BiGAN/VAE and GibbsNet. It can be concluded that our model achieves both good generation and inference performances compared with bi-directional models such as VAE and BiGAN.

Table 3: Downstream Classification Accuracy (%) Comparisons of different bi-directional generation mechanism

	VAE	DAE	BiGAN	GibbsNet	CyGen
MNIST	94.5±0.3	98.0±0.1	91.0	97.7	98.3± 0.1
SVHN	30.8±0.2	74.5±1.0	67.7	77.0	75.8±0.5

5 Conclusions

In this work we investigate the possibility of defining a joint distribution using two conditionals for generative modeling without an explicit prior distribution. We develop a systematic theory with novel and operable equivalence criteria for compatibility and sufficient conditions for determinacy, and exploit the theory for practical generative modeling methods using cyclic conditionals (CyGen), including enforcing compatibility, fitting data and data generation. Experimental results illustrate the benefits of CyGen over generative models that use a specified uninformative prior.

As for the limitation of CyGen, although it provides a way to remove the prior constraint, this may come at the cost of lack of identifiability and disentanglement of the learned representation, which is hard to guarantee even with a specified prior [37]. This may be mitigated by incorporating domain knowledge into either conditional models, which leads to an interesting future work.

References

[1] G. Alain and Y. Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

[2] B. C. Arnold and S. J. Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

[3] B. C. Arnold, E. Castillo, J. M. Sarabia, et al. Conditionally specified distributions: an introduction. *Statistical Science*, 16(3):249–274, 2001.

[4] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, 2013.

[5] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning*, pages 226–234, 2014.

[6] P. Berti, E. Dreassi, and P. Rigo. Compatibility results for conditional distributions. *Journal of Multivariate Analysis*, 125:190–203, 2014.

[7] P. Billingsley. *Probability and Measure*. John Wiley & Sons, New Jersey, 2012. ISBN 978-1-118-12237-2.

[8] J. Bornschein, S. Shabanian, A. Fischer, and Y. Bengio. Bidirectional Helmholtz machines. In *International Conference on Machine Learning*, pages 2511–2519. PMLR, 2016.

[9] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

[10] B. Coors, A. P. Condurache, and A. Geiger. SphereNet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018.

[11] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, 2018.

[12] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear independent components estimation. In *Workshop on the International Conference on Learning Representations*, 2015.

[13] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

[14] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

[15] L. Falorsi, P. de Haan, T. R. Davidson, N. De Cao, M. Weiler, P. Forré, and T. S. Cohen. Explorations in homeomorphic variational auto-encoding. *arXiv preprint arXiv:1807.04689*, 2018.

[16] J. Galambos. *Advanced probability theory*, volume 10. CRC Press, 1995.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, Montréal, Canada, 2014. NIPS Foundation.

[18] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[19] A. Grover and S. Ermon. Uncertainty autoencoders: Learning compressed representations via variational information maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2514–2524. PMLR, 2019.

[20] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.

[21] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[23] F. Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015.

[24] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[25] D. Kalatzis, D. Eklund, G. Arvanitidis, and S. Hauberg. Variational autoencoders with riemannian brownian motion priors. In *International Conference on Machine Learning*, pages 5053–5066. PMLR, 2020.

[26] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR.org, 2017.

[27] D. P. Kingma and P. Dhariwal. Glow: generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[28] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR 2014)*, Banff, Canada, 2014. ICLR Committee.

[29] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, Barcelona, Spain, 2016. NIPS Foundation.

[30] A. Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

[31] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *Proceedings of the International Conference on Learning Representations*, 2018.

[32] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[33] A. M. Lamb, D. Hjelm, Y. Ganin, J. P. Cohen, A. C. Courville, and Y. Bengio. GibbsNet: Iterative adversarial inference for deep graphical models. In *Advances in Neural Information Processing Systems*, pages 5089–5098, 2017.

[34] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551, 1989.

[35] C. Li, M. Welling, J. Zhu, and B. Zhang. Graphical generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2018.

[36] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[37] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[38] B. Pang, T. Han, E. Nijkamp, S.-C. Zhu, and Y. N. Wu. Learning latent space energy-based prior model. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[39] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[40] A. Razavi, A. v. d. Oord, B. Poole, and O. Vinyals. Preventing posterior collapse with delta-VAEs. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[41] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[42] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[43] A. Rinaldo. Advanced probability, February 2018.

[44] H. Shao, A. Kumar, and P. Thomas Fletcher. The Riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

[45] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR 2021)*, 2021.

[46] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–10, 2016.

[47] R. Van Den Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. Sylvester normalizing flows for variational inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 393–402. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

[48] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[49] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning*, pages 1096–1103, 2008.

[50] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 681–688, Bellevue, Washington USA, 2011. IMLS.

[51] Y. Xia, J. Bian, T. Qin, N. Yu, and T.-Y. Liu. Dual inference for machine learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 3112–3118, 2017.

[52] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T.-Y. Liu. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3789–3798. JMLR.org, 2017.

[53] Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2849–2857, 2017.

[54] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Covered in Conclusion 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We did not see possible negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appx. C for proof and Appx. A and Appx. B for technical backgrounds and our Lemmas.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] They are common public datasets and code bases.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] They are common public datasets and code bases and no such problems are found.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]