On the Generative Utility of Cyclic Conditionals

Chang Liu[1,*], **Haoyue Tang**[2,†], **Tao Qin**[1], **Jintao Wang**[2], **Tie-Yan Liu**[1]
[1] Microsoft Research Asia, Beijing, 100080. [2] Tsinghua University, Beijing, 100084.

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 extracting representation, but they rely on a usually uninformative prior distribution $p(z)$ 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 operable equivalence criteria for compatibility, and sufficient conditions for determinacy. Based on the theory, we propose a novel generative modeling framework CyGen that only uses the two cyclic conditional models. We develop methods to achieve compatibility and determinacy, and to use the conditional models to fit and generate data. With the prior constraint removed, CyGen better fits data and captures more representative features, supported by both synthetic and real-world experiments.

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) [40, 62, 10, 11, 41, 75] follow the Bayesian framework and specify a prior distribution $p(z)$ 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 to approximate the posterior distribution $p(z|x)$ (derived from the joint $p(x, z)$), which serves for extracting representations. Other frameworks like generative adversarial nets [25, 20, 22], flow-based models [19, 55, 39, 26] and diffusion-based models [69, 33, 71, 44] 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** [11, 29, 59]: 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** [17, 23, 36]: 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 require either a strong domain knowledge [43, 36], or additional cost to learn a complicated prior model [47, 16, 74] sometimes even at the cost of inconvenient inference [54, 81].

[*]Correspondence to: Chang Liu <changliu@microsoft.com>.
[†]Work done during an internship at Microsoft Research Asia.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

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 specifying or learning a prior 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 participates in defining the joint distribution (avoid "inner approximation"). Modeling conditionals is also argued to be much easier than modeling marginal or joint distributions directly [1, 6, 7]. In many cases, one may even have better knowledge on the conditionals than on the prior, *e.g.* shift/rotation invariance of image representations (CNNs [46] / SphereNet [15]), and rules to extract frequency/energy features for audio [60]. 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 both a systematic theory and a novel generative modeling framework CyGen (**Cy**clic-conditional **Gen**erative model). **(1)** Theoretical analysis on the question amounts to two sub-problems: can two given cyclic conditionals correspond to a common joint, and if yes, can they determine the joint uniquely. We term them *compatibility* and *determinacy* of two conditionals, corresponding to the existence and uniqueness of a common joint. For this, we develop novel compatibility criteria and sufficient conditions for determinacy. Beyond existing results, ours are operable (vs. existential [8]) and self-contained (vs. need a marginal [6, 45]), and are general enough to cover both continuous and discrete cases. Our compatibility criteria are also equivalence (vs. unnecessary [2, 3]) conditions. **(2)** In addition to its independent contribution, the theory also enables generative modeling using only the two cyclic conditional models, *i.e.* the CyGen framework. We develop methods for achieving compatibility and determinacy to make an eligible generative model, and for fitting and generating data to serve as a generative model. Efficient implementation techniques are designed. Note CyGen also determines a prior implicitly; it just does not need an explicit model for it (vs. [47, 16, 74, 54]). We show the practical utility of CyGen in both synthetic and real-world tasks. The improved performance in downstream classification and data generation demonstrates the advantage to mitigate posterior collapse and manifold mismatch.

1.1 Related work

Dependency networks ([30]; similarly [34]) are perhaps the first to pursue the idea of modeling a joint by a set of conditionals. They use Gibbs sampling to determine the joint and are equivalent to undirected graphical models. They do not allow latent variables, so compatibility is not a key consideration as the data already specifies a joint as the common target of the conditionals. Beyond that, we introduce latent variables to better handle sensory data like images, for which we analyze the conditions for compatibility and determinacy and design novel methods to solve this different task.

Denoising auto-encoders (DAEs). AEs [66, 4] aim to extract data features by enforcing reconstruction through its encoder and decoder, which are deterministic hence insufficient determinacy (see Sec. 2.2.2). DAEs [77, 6, 7] use a probabilistic encoder and decoder for robust reconstruction against random data corruption. Their utility as a generative model is first noted through the equivalence to score matching (implies modeling $p(x)$) for a Gaussian RBM [76] or an infinitesimal Gaussian corruption [1]. 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 [6, 45, 27] or local [7] shared support condition, its stationary distribution $\pi(x, z)$ exists uniquely. But this is *not really determinacy*: even incompatible conditionals can have this unique existence, in which case $\pi(z|x) \neq q(z|x)$ [30, 6]. Moreover, the Gibbs chain does not give an explicit expression of $\pi(x, z)$ (thus intractable likelihood evaluation), and requires many iterations to converge for data generation and even for training (Walkback [6], GibbsNet [45]), making the algorithms costly and unstable.

As for *compatibility*, it is not really covered in DAEs. Existing results only consider the statistical consistency (unbiasedness under infinite data) of the $p(x|z)$ estimator by fitting (x, z) data from $p^*(x)q(z|x)$ [6, 7, 45, 27], where $p^*(x)$ denotes the true data distribution. 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 only of the two conditionals.

More crucially, the DAE loss is not proper for optimizing $q(z|x)$ as it promotes a mode-collapse behavior. This hinders both compatibility and determinacy (Sec. 3.2): one may not use $q(z|x)$ for inference, and data generation may depend on initialization. In contrast, CyGen explicitly enforces compatibility and guarantees determinacy, and enables likelihood evaluation and better generation.

Dual learning considers conversion tasks between two modalities in both directions, *e.g.*, machine translation [28, 80, 79] and image style transfer [37, 84, 82, 48]. 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). Technically, they did not consider determinacy:

they require an external marginal from either a model or data to determine a joint. In fact, determinacy is insufficient for deterministic conversions (see Sec. 2.2.2). Their cycle-consistency loss [37, 84, 82] is a version of our compatibility criterion in a specific Dirac case (see Sec. 2.2.1), and we extend the loss to the more general absolutely continuous case (allowing probabilistic conversion) (see Sec. 3.1).

2 Compatibility and Determinacy Theory

To be a generative model, a system needs to determine a distribution on the data variable x. With latent variable z, this amounts to determining a joint distribution on (x, z). In this section we build the general theory on the conditions for compatibility and determinacy. It also lays the foundation of our novel CyGen framework for generative modeling. We begin with formalizing the problems.

Setup. Denote the measure spaces of the two random variables x and z as $(\mathbb{X}, \mathscr{X}, \xi)$ and $(\mathbb{Z}, \mathscr{Z}, \zeta)$[3], 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. We use $\mathcal{X} \in \mathscr{X}$, $\mathcal{Z} \in \mathscr{Z}$ to denote measurable sets, and use "$\overset{\xi}{=}$", "$\subseteq^{\xi}$" as the extensions of "$=$", "$\subseteq$" up to a set of ξ-measure-zero (Def. A.1). Following the convention in machine learning, we call a "probability measure" as a "distribution". 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 $\mathbb{X}$, $\mathbb{Z}$ to have 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 [9, Thm. 18.2]. Define the *slice* of $\mathcal{W} \in \mathscr{X}\otimes\mathscr{Z}$ at z as $\mathcal{W}_z := \{x \mid (x, z) \in \mathcal{W}\} \in \mathscr{X}$ [9, Thm. 18.1(i)], 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). In a similar style, denote the *marginal* of a joint π on $\mathbb{Z}$ as $\pi^{\mathbb{Z}}(\mathcal{Z}) := \pi(\mathbb{X}\times\mathcal{Z})$. To keep the same level of generality, we follow the general definition of conditionals ([9, p.457]; see also Appx. A.4): the conditional $\pi(\mathcal{X}|z)$ of a joint π is the density function (R-N derivative) of $\pi(\mathcal{X}\times\cdot)$ w.r.t $\pi^{\mathbb{Z}}$. We highlight the key characteristic under this generality that $\pi(\cdot|z)$ 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. The goals of 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. The two cases correspond to different types of generative models, and lead to qualitatively different conclusions. Note the concepts only involve the two conditionals, so we desire self-contained conditions, meaning 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 $x \in \mathbb{X}$,[4] the conditionals $\mu(\cdot|z)$ and $\nu(\cdot|x)$ are either absolutely continuous (w.r.t ξ and ζ, resp.) [9, 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 conditional models are often specified in the form of density, this case covers many generative model frameworks, notably VAEs [40, 62, 61, 41, 75] and diffusion-based models [69, 33, 71].

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 formulae 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

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

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

$p(x|z)\frac{1}{Ab(z)} = q(z|x)\frac{a(x)}{A}$ 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].

However, it is more complicated than imagined. Berti et al. [8, 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.[5] To address the problem, Berti et al. [8] 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 $\mathcal{S}$.

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*[6] 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 for almost all $z \in \mathcal{S}^{\mathbb{Z}}$ and $x \in \mathcal{S}^{\mathbb{X}}$ (Lem. B.16). This is critical for the normalizedness of distributions in our criterion. Appx. B.3 shows more facts. With this concept, our compatibility criterion is presented below.

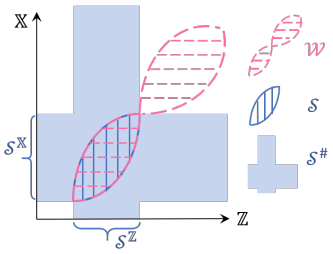

Figure 1: Illustration of a $\xi\otimes\zeta$-complete component $\mathcal{S}$ of $\mathcal{W}$.

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 $\mathcal{S}$,[7] 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.

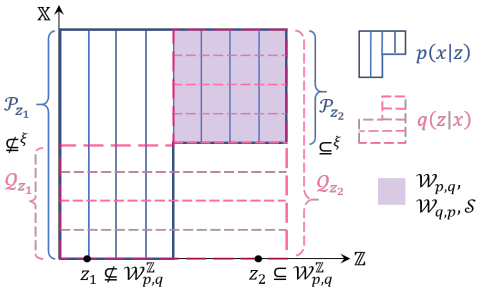

Figure 2: Illustration of our compatibility criterion in the absolutely continuous case (Thm. 2.3). The conditionals are uniform on the respective depicted slices. For condition **(i)**, $\mathcal{P}_z \subseteq^\xi \mathcal{Q}_z$ is *not* satisfied on the left half, *e.g.* z_1, so $\mathcal{W}_{p,q}$ 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.

[5] 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.

[6] 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.

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

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. 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 $(\mathcal{W}_{p,q})_z = \mathcal{P}_z$ which is 1 by construction; similarly for $q(z|x)$. In contrast, Appx. Example C.3 shows $\mathcal{S} = \mathcal{W}_{p,q} \cap \mathcal{W}_{q,p}$ is inappropriate. Conditions **(ii)** and **(iii)** cannot be guaranteed by condition **(i)** (Appx. Example B.13), while are needed to rule out special cases (Appx. Lem. B.14, Example B.15). Appx. C.2 gives a rigorous proof. Finally, 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 (necessarily exists 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 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) [25] and flow-based models [19, 55, 39, 26], 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.

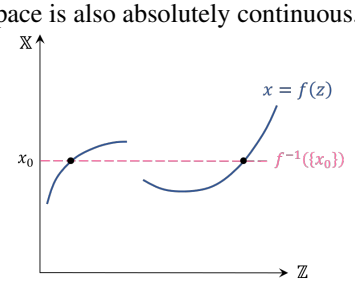

Figure 3: Illustration of our compatibility criterion in the Dirac case (Thm. 2.6).

Theorem 2.6 (compatibility criterion, Dirac)**.** *Suppose that $\mathscr{X}$ contains all the single-point sets: $\{x\} \in \mathscr{X}, \forall x \in \mathbb{X}$. 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*[8]*, if and only if there exists $x_0 \in \mathbb{X}$ such that $\nu(f^{-1}(\{x_0\})|x_0) = 1$.*

See Appx. C.6 for proof. Note such x_0 must be in the image set $f(\mathbb{Z})$, otherwise $\nu(f^{-1}(\{x_0\})|x_0) = \nu(\emptyset|x_0) = 0$. For a typical GAN generator, the preimage set $f^{-1}(\{x_0\})$ is discrete, so a compatible inference model must not be absolutely continuous. 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.

[8] 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.

Nevertheless, in practice one often desires the compatibility to hold over a set $\mathcal{X}$ to make a useful model (*i.e.*, allowing compatible joints supported on $\mathcal{X}$). When $\nu(\cdot|x)$ is also chosen in the Dirac form $\delta_{g(x)}$, this can be achieved by minimizing $\mathbb{E}_{p(x)}\ell\big(x, f(g(x))\big)$, where $p(x)$ is a distribution on $\mathcal{X}$ and ℓ is a metric on $\mathbb{X}$. This is the *cycle-consistency loss* used in dual learning [37, 84, 82, 48]. When f is invertible, minimizing the loss (*i.e.*, $g = f^{-1}$ a.e. on $\mathcal{X}$) is also necessary, as $f^{-1}(x)$ only has one element. Particularly, flow-based models are naturally compatible.

2.2.2 Determinacy in the Dirac case

As mentioned, for any x_0 satisfying the condition, two compatible conditionals have the determinacy on this point $\{x_0\}$ with the unique joint $\delta_{(x_0, f(x_0))}$. But when such x_0 is not unique, the distribution over these x_0 values is not determined, so the two conditionals do not have the determinacy globally on $\mathbb{X}\times\mathbb{Z}$. This is similar to the absolutely continuous case with multiple complete supports; particularly, each $\{(x_0, f(x_0))\}$ is a complete support for discrete $\mathbb{X}$ and $\mathbb{Z}$. This meets one's intuition: compatible Dirac conditionals can only determine a curve in $\mathbb{X}\times\mathbb{Z}$, but cannot determine a distribution on the curve. One exception is when $f(z) \equiv x_0$ is constant, so this x_0 is the only candidate. 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 cyclic conditionals achieve compatibility and a sufficient determinacy, so that they can determine a useful joint without specifying a prior. Note a certain prior is implicitly determined by the conditionals; we find we just do not need an explicit model for it. This inspires CyGen, a novel framework that only uses **Cy**clic conditionals for **Gen**erative modeling.

For the eligibility as a generative model, compatibility and a sufficient determinacy are required. For the latter, we just shown a deterministic likelihood or inference model is not suitable, so we use absolutely continuous conditionals as the theory suggests. The conditionals can then be modeled by parameterized densities $p_\theta(x|z)$, $q_\phi(z|x)$. We consider the common case where $\mathbb{X} = \mathbb{R}^{d_\mathbb{X}}$, $\mathbb{Z} = \mathbb{R}^{d_\mathbb{Z}}$, and $p_\theta(x|z)$, $q_\phi(z|x)$ have a.e.-full supports and are differentiable. Determinacy is then exactly guaranteed by Cor. 2.5. For compatibility, we develop an effective loss in Sec. 3.1 to enforce it.

For the usage as a generative model, we develop methods to fit the model-determined data distribution $p_{\theta,\phi}(x)$ to the true data distribution $p^*(x)$ in Sec. 3.2, and to generate data from $p_{\theta,\phi}(x)$ in Sec. 3.3.

3.1 Enforcing Compatibility

In this a.e.-full support case, the entire product 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:

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

Here, 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. [9, Thm. 15.2(ii)]. By integration, this means $\nabla_z r_{\theta,\phi}(x,z) = V(z)$ hence $r_{\theta,\phi}(x,z) = v(z) + u(x)$, $\xi\otimes\zeta$-a.e., for some functions $V(z)$, $v(z)$, $u(x)$ s.t. $V(z) = \nabla v(z)$. So the ratio $p_\theta(x|z)/q_\phi(z|x) = \exp\{r_{\theta,\phi}(x,z)\} = \exp\{u(x)\}\exp\{v(z)\}$ factorizes, $\xi\otimes\zeta$-a.s.

In the sense of enforcing compatibility, this loss generalizes the cycle-consistency loss to probabilistic conditionals. Also, the loss is different from the Jacobian-norm regularizers in contractive AE [63] and DAE [63, 1], and explains the "tied weights" trick for AEs [58, 77, 76, 63, 1] (see Appx. D.1).

Implication on Gaussian VAE which uses additive Gaussian conditional models, $p_\theta(x|z) := \mathcal{N}(x|f_\theta(z), \sigma_d^2 I_{d_\mathbb{X}})$ and $q_\phi(z|x) := \mathcal{N}(z|g_\phi(x), \sigma_e^2 I_{d_\mathbb{Z}})$. It is the vanilla and the most common form of VAE [40]. As its ELBO objective drives $q_\phi(z|x)$ to meet the joint $p(z)p_\theta(x|z)$, compatibility is enforced. Under our view, this amounts to minimizing the compatibility loss Eq. (2), which then enforces the match of Jacobians: $(\nabla_z f_\theta^\top(z))^\top = (\sigma_d^2/\sigma_e^2)\nabla_x g_\phi^\top(x)$. As the two sides indicate the equation is constant of both x and z, 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 in causality [83, 57], and explains the empirical observation that the latent space of such VAEs is quite linear [68]. It is also the root of

recent analyses on Gaussian VAE that the latent space coordinates the data manifold [16], and the inference model learns an isometric embedding after a proper rescaling [53].

This finding reveals that the expectation to use deep neural networks for learning a flexible nonlinear representation will be disappointed in Gaussian VAE. So we use a non-additive-Gaussian model, *e.g.* a flow-based model [61, 41, 75, 26], for at least one of $p_\theta(x|z)$ and $q_\phi(z|x)$ (often the latter).

Efficient 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 unbiased and much cheaper method based on Hutchinson's trace estimator [35]: $\mathrm{tr}(A) = \mathbb{E}_{p(\eta)}[\eta^\top A\eta]$, where η is any random vector with zero mean and identity covariance (*e.g.*, a standard Gaussian). As the function within expectation is $\left\|\nabla_x \nabla_z^\top r\right\|_F^2 = \left\|\nabla_z \nabla_x^\top r\right\|_F^2 = \mathrm{tr}\left((\nabla_z \nabla_x^\top r)^\top \nabla_z \nabla_x^\top r\right)$, applying the estimator yields a formulation that reduces gradient evaluation complexity to $O(d_\mathbb{X} + d_\mathbb{Z})$:

$$(\min_{\theta,\phi})\ \ C(\theta,\phi) = \mathbb{E}_{\rho(x,z)}\mathbb{E}_{p(\eta_x)}\left\|\nabla_z\left(\eta_x^\top \nabla_x r_{\theta,\phi}(x,z)\right)\right\|_2^2, \text{where } \mathbb{E}[\eta_x] = 0, \mathrm{Var}[\eta_x] = I_{d_\mathbb{X}}. \quad (3)$$

As concluded from the above analysis on Gaussian VAE, we use a flow-based model for the inference model $q_\phi(z|x)$. But in common instances evaluating the inverse of the flow is intractable [61, 41, 75] or costly [26]. This however, disables the use of automatic differentiation tools for estimating the gradients in the compatibility loss. Appx. D.2 explains this problem in detail and shows our solution.

3.2 Fitting Data

After enforcing compatibility, Cor. 2.5 guarantees the a.e.-fully supported conditional models uniquely determine a joint, hence a data distribution $p_{\theta,\phi}(x)$. To fit $p_{\theta,\phi}(x)$ to the true data distribution $p^*(x)$, an explicit expression is required. For this, Eq. (1) is not helpful as we do not have explicit expressions of $a(x)$, $b(z)$. But when compatibility is given, we can safely use density function formulae:

$$p_{\theta,\phi}(x) = 1/\tfrac{1}{p_{\theta,\phi}(x)} = 1/\textstyle\int_\mathbb{Z} \tfrac{p_{\theta,\phi}(z')}{p_{\theta,\phi}(x)}\zeta(\mathrm{d}z') = 1/\int_\mathbb{Z} \tfrac{q_\phi(z'|x)}{p_\theta(x|z')}\zeta(\mathrm{d}z') = 1/\mathbb{E}_{q_\phi(z'|x)}[1/p_\theta(x|z')],$$

which is an explicit expression in terms of the two conditionals. Although other expressions are possible, this one has a simple form, and the Monte-Carlo expectation estimation in $\mathbb{Z}$ has a lower variance than in $\mathbb{X}$ since usually $d_\mathbb{Z} \ll d_\mathbb{X}$. We can thus fit data by maximum likelihood estimation:

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

The loss function can be estimated using the reparameterization trick [40] to reduce variance, and the `logsumexp` trick is adopted for numerical stability. This expression can also serve for data likelihood evaluation. The final training process of CyGen is the joint optimization with the compatibility loss.

Comparison with DAE. We note that the DAE loss [77, 6] $\mathbb{E}_{p^*(x)q_\phi(z'|x)}[-\log p_\theta(x|z')]$ is a *lower bound* of Eq. (4) due to Jensen's inequality, so it is not suitable for maximizing likelihood. In fact, the DAE loss minimizes $\mathbb{E}_{q_\phi(z)}\mathrm{KL}(q_\phi(x|z)\|p_\theta(x|z))$ for $p_\theta(x|z)$ to match $q_\phi(x|z)$, where $q_\phi(z)$ and $q_\phi(x|z)$ are induced from the joint $p^*(x)q_\phi(z|x)$, but it is not a proper loss for $q_\phi(z|x)$ as a *mode-collapse* behavior is promoted: the optimal $q_\phi(z|x)$ only concentrates on the point(s) of $\mathrm{argmin}_{z'}\, p_\theta(x|z')$, and an additional entropy term $-\mathbb{E}_{q_\phi(z)}\mathbb{H}[q_\phi(x|z)]$ is required to optimize the same KL loss. This behavior *hurts determinacy*, as $q_\phi(z|x)$ tends to be a (mixture of) Dirac measure (Sec. 2.2.2). The resulting Gibbs chain may also converge differently depending on initialization, as ergodicity is broken. This behavior also hurts compatibility, as $q_\phi(x|z)$ deviates from $p_\theta(x|z)$ (not Dirac) and does not match the Gibbs stationary distribution [30, 6]. In contrast, CyGen follows a more fundamental logic: enforce compatibility explicitly and follow the maximum likelihood principle faithfully. It leads to a proper loss for both conditionals that does not hinder determinacy.

3.3 Data Generation

Generating samples from the learned data distribution $p_{\theta,\phi}(x)$ is not as straightforward as typical models that specify a prior, since ancestral sampling is not available. But it is still tractable via Markov chain Monte Carlo methods (MCMCs). We propose using *dynamics-based MCMCs*, which are often more efficient than Gibbs sampling (used in DAE [6] and GibbsNet [45]). They only require an *unnormalized* density function of the target distribution, which is readily available in CyGen when compatible: $p_{\theta,\phi}(x) = \frac{p_{\theta,\phi}(x)}{p_{\theta,\phi}(z)}p_{\theta,\phi}(z) = \frac{p_\theta(x|z)}{q_\phi(z|x)}p_{\theta,\phi}(z) \propto \frac{p_\theta(x|z)}{q_\phi(z|x)}$ for any $z \in \mathbb{Z}$. In practice, this z can be taken as a sample from $q_\phi(z|x)$ to lie in a high probability region for a confident estimate.

Stochastic gradient Langevin dynamics (SGLD) [78] is a representative instance, which has been shown to produce complicated realistic samples in energy-based [21], score-based [70] and diffusion-

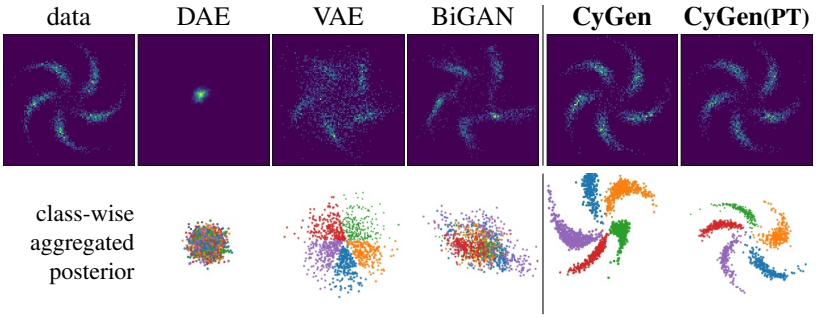

Figure 4: Generated data (DAE and CyGen use SGLD) and class-wise aggregated posteriors of DAE, VAE, BiGAN and CyGen. Also shows results of CyGen(PT) that is PreTrained as a VAE. (Best view in color.)

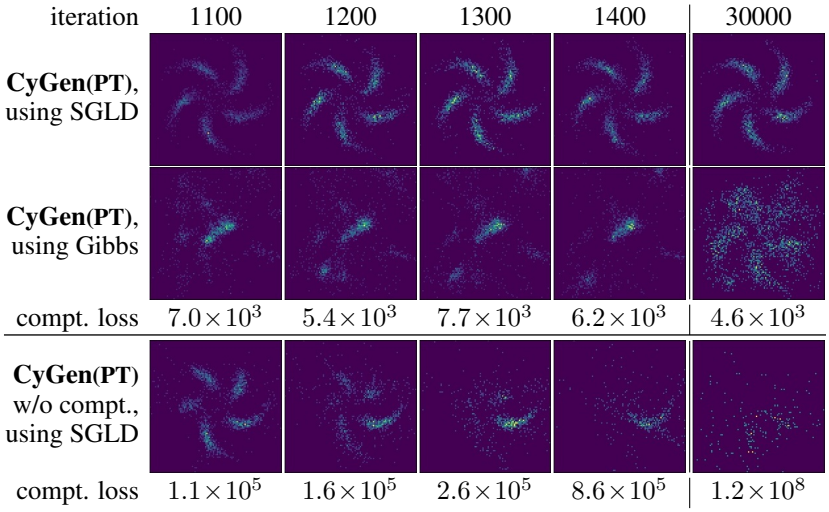

Figure 5: Generated data along the training process of CyGen after VAE pretraining (iteration 1000), using SGLD (rows 1,3) and Gibbs sampling (row 2) for generation, and with (rows 1,2) and without the compatibility loss (row 3) for training. See also Appx. Fig. 13.

based [33, 71] models. It gives the following transition:

$$x^{(t+1)} = x^{(t)} + \varepsilon \nabla_{x^{(t)}} \log \frac{p_\theta(x^{(t)}|z^{(t)})}{q_\phi(z^{(t)}|x^{(t)})} + \sqrt{2\varepsilon}\,\eta_x^{(t)}, \text{where } z^{(t)} \sim q_\phi(z|x^{(t)}), \eta_x^{(t)} \sim \mathcal{N}(0, I_{d_\mathbb{X}}), \quad (5)$$

and ε is a step size parameter. Method to draw $z \sim p_{\theta,\phi}(z)$ can be developed symmetrically (see Appx. Eq. (25)). Also applicable are other dynamics-based MCMCs [14, 18, 52], and particle-based variational inference methods [51, 12, 49, 50, 72] which are more sample-efficient.

4 Experiments

We demonstrate the power of CyGen for data generation and representation learning. Baselines include DAE, and generative models using Gaussian prior *e.g.* VAE and BiGAN (Appx. E.1). For a fair comparison, all methods use the same architecture, which is an additive Gaussian $p_\theta(x|z)$ and a Sylvester flow (Householder version) [75] for $q_\phi(z|x)$ (Appx. E.2), as required by CyGen (Sec. 3.1). It is necessarily probabilistic for determinacy, so we exclude flow-based generative models and common BiGAN/GibbsNet architectures, which are deterministic. We also considered GibbsNet [45] which also aims at the prior issue, but it does not produce reasonable results using the same architecture, due to its unstable training process (see Appx. E.1). Codes: `https://github.com/changliu00/cygen`.

4.1 Synthetic Experiments

For visual verification of the claims, we first consider a 2D toy dataset (Fig. 4 top-left). Appx. E.3 shows more details and results, including the investigation on another similar dataset.

Data generation. The learned data distributions (as the histogram of generated data samples) are shown in Fig. 4 (row 1). We see the five clusters are blurred to overlap in VAE's distribution and are still connected in BiGAN's, due to the specified prior. In contrast, our CyGen fits this distribution much better; particularly it clearly separates the five non-connected clusters. This verifies the advantage to overcome the *manifold mismatch* problem. As for DAE, it cannot capture the data distribution due to collapsed inference model and insufficient determinacy (Sec. 3.2).

Representation. Class-wise aggregated posteriors (as the scatter plot of z samples from $q_\phi(z|x)p^*(x|y)$ for each class/cluster y) in Fig. 4 (row 2) show that CyGen mitigates the *poste-*

rior collapse problem, as the learned inference model $q_\phi(z|x)$ better separates the classes with a margin in the latent space. This more informative and representative feature would benefit downstream tasks like classification or clustering in the latent space. In contrast, the specified Gaussian prior squeezes the VAE latent clusters to touch, and the BiGAN latent clusters even to mix. The mode-collapsed inference model of DAE locates all latent clusters in the same place.

Incorporating knowledge into conditionals. CyGen alone (without pretraining) already performs well. When knowledge is available, we can further incorporate it into the conditional models. Fig. 4 shows pretraining CyGen's likelihood model as in a VAE (CyGen(PT)) embodies VAE's knowledge that the prior is centered and centrosymmetric, as the (all-class) aggregated posterior ($\approx$ prior) is such. Note its data generation quality is not sacrificed. Appx. Fig. 14 verifies this directly via the priors.

Comparison of data generation methods. We then make more analysis on CyGen. Fig. 5 (rows 1,2) shows generated data of CyGen using SGLD and Gibbs sampling. We see SGLD better recovers the true distribution, and is more robust to slight incompatibility.

Impact of the compatibility loss. Fig. 5 (rows 1,3) also shows the comparison with training CyGen without the compatibility loss. We see the compatibility is then indeed out of control, which invalidates the likelihood estimation Eq. (4) for fitting data and the gradient estimation in Eq. (5) for data generation, leading to the failure in row 3. Along the training process of the normal CyGen, we also find a smaller compatibility loss makes better generation (esp. using Gibbs sampling).

4.2 Real-World Datasets

We test the performance of CyGen on real-world image datasets MNIST and SVHN. We consider the VAE-pretrained version, CyGen(PT), for more stable training. Appx. E.4 shows more details. On these datasets, even BiGAN cannot produce reasonable results using the same architecture, similar to GibbsNet.

Data generation. From Fig. 6, We see that CyGen(PT) generates both sharp and diverse samples, as a sign to mitigate *manifold mismatch*. DAE samples are mostly imperceptible, due to the mode-collapsed $q_\phi(z|x)$ and the subsequent lack of determinacy (Sec. 3.2). VAE samples are a little blurry as a typical behavior due to the simply-connected prior. This observation is also quantitatively supported by the FID score [31, 67] on SVHN: CyGen achieves 102, while DAE 157 and VAE 128 (lower is better).

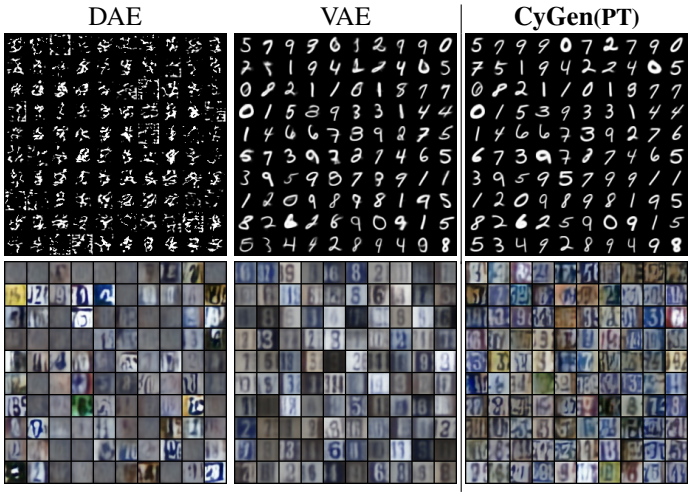

Figure 6: Generated data on the MNIST and SVHN datasets.

Table 7: Downstream classification accuracy (%) using learned representation by various models.
†: Results from [45] using a different, deterministic architecture (not suitable for CyGen).

	DAE	VAE	BiGAN†	GibbsNet†	**CyGen(PT)**
MNIST	98.0±0.1	94.5±0.3	91.0	97.7	**98.3±0.1**
SVHN	74.5±1.0	30.8±0.2	66.7	**79.6**	**75.8±0.5**

Representation. We then show in Table 7 that CyGen(PT)'s latent representation is more informative for the downstream classification task, as an indicator to avoid *posterior collapse*. BiGAN and GibbsNet make random guess using the same probabilistic flow architecture, and their reported results in [45] using a different, deterministic architecture (not suitable for CyGen due to insufficient determinacy) are still not always better, due to the prior constraint. We conclude that CyGen achieves both superior generation and representation learning performance.

5 Conclusions and Discussions

In this work we investigate the possibility of defining a joint distribution using two conditional distributions, under the motivation for generative modeling without an explicit prior. We develop a systematic theory with novel and operable equivalence criteria for compatibility and sufficient conditions for determinacy, and propose a novel generative modeling framework CyGen that only

uses cyclic conditional models. Methods for achieving compatibility and determinacy, fitting data and data generation are developed. Experiments show the benefits of CyGen over DAE and prevailing generative models that specify a prior in overcoming manifold mismatch and posterior collapse.

The novel CyGen framework broadens the starting point to build a generative model, and the general theory could also foster a deeper understanding of other machine learning paradigms, *e.g.*, dual learning and self-supervised learning, and inspire more efficient algorithms.

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] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

[5] J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

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

[7] 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.

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

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

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

[11] 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.

[12] C. Chen, R. Zhang, W. Wang, B. Li, and L. Chen. A unified particle-optimization framework for scalable Bayesian sampling. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, Monterey, California USA, 2018. Association for Uncertainty in Artificial Intelligence.

[13] R. T. Chen, J. Behrmann, D. Duvenaud, and J.-H. Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, 2019.

[14] T. Chen, E. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1683–1691, Beijing, China, 2014. IMLS.

[15] 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.

[16] B. Dai and D. Wipf. Diagnosing and enhancing VAE models. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*, 2019.

[17] 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.

[18] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems*, pages 3203–3211, Montréal, Canada, 2014. NIPS Foundation.

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

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

[21] Y. Du and I. Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1(Oct):49–75, 2000.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR 2017)*, 2017.

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

[34] R. Hofmann and V. Tresp. Nonlinear Markov networks for continuous variables. *Advances in Neural Information Processing Systems*, pages 521–527, 1998.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] C. Liu, J. Zhuo, P. Cheng, R. Zhang, J. Zhu, and L. Carin. Understanding and accelerating particle-based variational inference. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 4082–4092, Long Beach, California USA, 2019. IMLS, PMLR.

[50] C. Liu, J. Zhuo, and J. Zhu. Understanding MCMC dynamics as flows on the Wasserstein space. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 4093–4103, Long Beach, California USA, 2019. IMLS, PMLR.

[51] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, pages 2370–2378, Barcelona, Spain, 2016. NIPS Foundation.

[52] Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021.

[53] A. Nakagawa, K. Kato, and T. Suzuki. Quantitative understanding of VAE as a non-linearly scaled isometric embedding. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[54] 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.

[55] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2335–2344, 2017.

[56] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[57] 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.

[58] M. Ranzato, Y.-L. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. *Advances in Neural Information Processing Systems*, 20:1185–1192, 2007.

[59] 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.

[60] 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.

[61] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, pages 1530–1538, Lille, France, 2015. IMLS.

[62] 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.

[63] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the International Conference on Machine Learning*, 2011.

[64] S. Rifai, Y. Bengio, Y. N. Dauphin, and P. Vincent. A generative process for sampling contractive auto-encoders. In *Proceedings of the International Conference on Machine Learning*, pages 1811–1818, 2012.

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

[66] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[67] M. Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020. Version 0.2.1.

[68] 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.

[69] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[70] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[71] 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.

[72] A. Taghvaei and P. Mehta. Accelerated flow for probability distributions. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6076–6085. PMLR, 2019.

[73] T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *arXiv preprint arXiv:2006.11469*, 2020.

[74] A. Vahdat and J. Kautz. NVAE: A deep hierarchical variational autoencoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] Z. Xiao, K. Kreis, J. Kautz, and A. Vahdat. VAEBM: A symbiosis between variational autoencoders and energy-based models. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

[82] 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.

[83] 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.

[84] 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.