
On the Generative Utility of Cyclic Conditionals

Chang Liu^{1*}, Haoyue Tang^{2†}, Tao Qin¹, Jintao Wang², Tie-Yan Liu¹

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.

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 (**C**yclic-**o**nditional **G**enerative 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.

Denosing 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 $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$ ³, where \mathcal{X}, \mathcal{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 $\mathcal{X} \in \mathcal{X}$, $\mathcal{Z} \in \mathcal{Z}$ to denote measurable sets, and use “ $\stackrel{\xi}{=}$ ”, “ \subseteq^{ξ} ” 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 $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z}, \xi \otimes \zeta)$, where “ \times ” is the usual Cartesian product, $\mathcal{X} \otimes \mathcal{Z} := \sigma(\mathcal{X} \times \mathcal{Z})$ is the sigma-field generated by measurable rectangles from $\mathcal{X} \times \mathcal{Z}$, and $\xi \otimes \zeta$ is the product measure [9, Thm. 18.2]. Define the *slice* of $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$ at z as $\mathcal{W}_z := \{x \mid (x, z) \in \mathcal{W}\} \in \mathcal{X}$ [9, Thm. 18.1(i)], and its *projection* onto \mathbb{Z} as $\mathcal{W}^{\mathbb{Z}} := \{z \mid \exists x \in \mathbb{X} \text{ s.t. } (x, z) \in \mathcal{W}\} \in \mathcal{Z}$ (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 $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$ 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 $S \in \mathcal{X} \otimes \mathcal{Z}$, if there is only one joint distribution concentrated on 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

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

⁴There 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 $A := \int_{\mathbb{X}} a(x) \xi(dx)$, 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 $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, we say that a set $\mathcal{S} \in \mathcal{X} \otimes \mathcal{Z}$ is a $\xi \otimes \zeta$ -complete component of \mathcal{W} , if $\mathcal{S}^\# \cap \mathcal{W} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}$, where $\mathcal{S}^\# := \mathcal{S}^\mathbb{X} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^\mathbb{Z}$ is the *stretch* of \mathcal{S} .

Fig. 1 illustrates the concept. Roughly, the stretch $\mathcal{S}^\#$ 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.

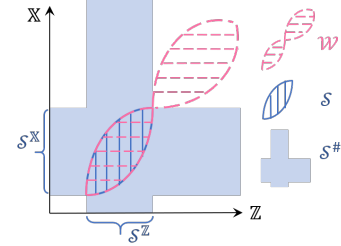


Figure 1: Illustration of a $\xi \otimes \zeta$ -complete component \mathcal{S} of \mathcal{W} .

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) $\mathcal{S}^\mathbb{X} \subseteq \xi \mathcal{W}_{q,p}^\mathbb{X}$, $\mathcal{S}^\mathbb{Z} \subseteq \zeta \mathcal{W}_{p,q}^\mathbb{Z}$, (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,

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

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

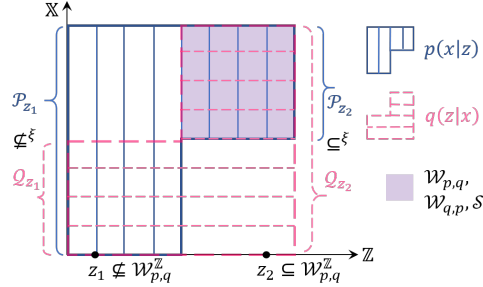


Figure 2: Illustration of our compatibility criterion in the absolutely continuous case (Thm. 2.3). The conditionals are uniform on the respective depicted slices. For condition (i), $\mathcal{P}_z \subseteq \xi \mathcal{Q}_z$ is *not* satisfied on the left half, e.g. z_1 , so $\mathcal{W}_{p,q}$ does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

⁵The 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.

⁶While 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.

⁷Formally, 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 $\mathcal{S}_z \stackrel{\xi}{=} \mathcal{S}^{\mathbb{X}}$, for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$, or $\mathcal{S}_x \stackrel{\zeta}{=} \mathcal{S}^{\mathbb{Z}}$, 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 $\mu(\mathcal{X}|z) = \delta_{f(z)}(\mathcal{X}) := \mathbb{I}[f(z) \in \mathcal{X}], \forall \mathcal{X} \in \mathcal{X}$ 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.

Theorem 2.6 (compatibility criterion, Dirac). *Suppose that \mathcal{X} contains all the single-point sets: $\{x\} \in \mathcal{X}, \forall x \in \mathbb{X}$. Conditional distribution $\nu(\mathbb{Z}|x)$ is compatible with $\mu(\mathcal{X}|z) := \delta_{f(z)}(\mathcal{X})$ where function $f : \mathbb{Z} \rightarrow \mathbb{X}$ is \mathcal{X}/\mathcal{Z} -measurable⁸, if and only if there exists $x_0 \in \mathbb{X}$ such that $\nu(f^{-1}(\{x_0\})|x_0) = 1$.*

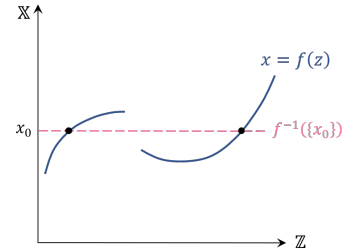


Figure 3: Illustration of our compatibility criterion in the Dirac case (Thm. 2.6).

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.

⁸For meaningful discussion, we require f to be \mathcal{X}/\mathcal{Z} -measurable, which includes any function between discrete sets and continuous functions when \mathcal{X} and \mathcal{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 $\mathbb{E}_{p(x)} \ell(x, f(g(x)))$, 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 **Cyclic** conditionals for **Generative** 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 $\mathbb{X} = \mathbb{R}^{d_x}$, $\mathbb{Z} = \mathbb{R}^{d_z}$, 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:

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

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, $p_\theta(x|z) := \mathcal{N}(x|f_\theta(z), \sigma_d^2 I_{d_x})$ and $q_\phi(z|x) := \mathcal{N}(z|g_\phi(x), \sigma_e^2 I_{d_z})$. 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]: $\text{tr}(A) = \mathbb{E}_{p(\eta)}[\eta^\top A \eta]$, where η is any random vector with zero mean and identity covariance (*e.g.*, a standard Gaussian). As the function within expectation is $\|\nabla_x \nabla_z^\top r\|_F^2 = \|\nabla_z \nabla_x^\top r\|_F^2 = \text{tr}((\nabla_z \nabla_x^\top r)^\top \nabla_z \nabla_x^\top r)$, applying the estimator yields a formulation that reduces gradient evaluation complexity to $O(d_{\mathbb{X}} + d_{\mathbb{Z}})$:

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

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:

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

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 $\mathbb{E}_{q_\phi(z)} \text{KL}(q_\phi(x|z) \| p_\theta(x|z))$ 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 $\text{argmin}_{z'} p_\theta(x|z')$, 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.)

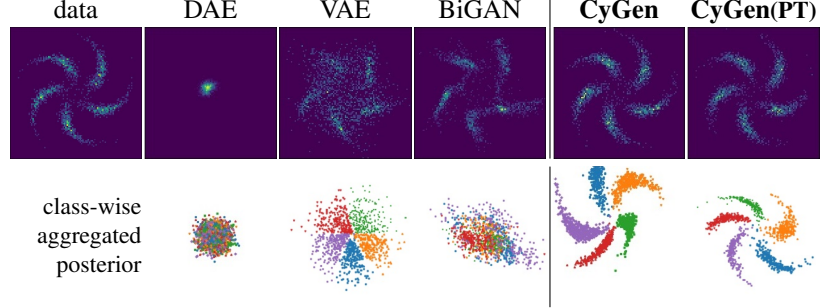
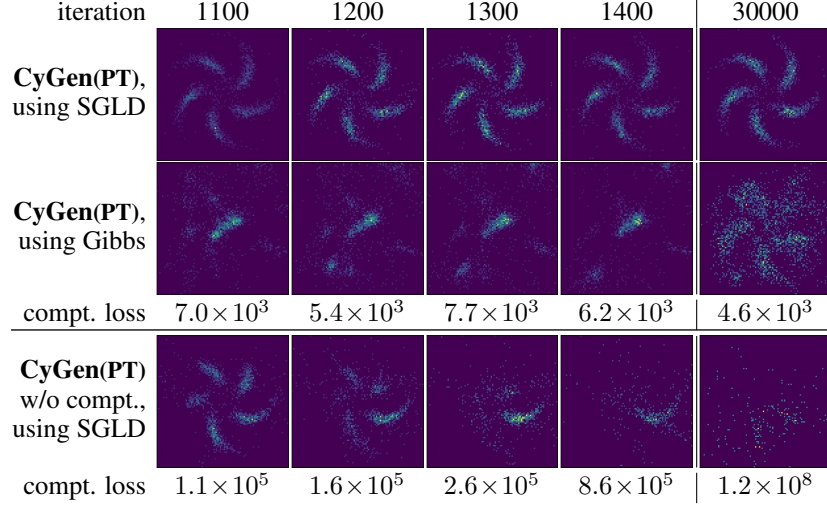


Figure 5: Generated data along the training process of CyGen after VAE pretraining (iteration 1000), using SGLD (rows 1,3) and Gibbs sampling (row 2) for generation, and with (rows 1,2) and without the compatibility loss (row 3) for training. See also Appx. Fig. 13.



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

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

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 *posterior*

prior 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).

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

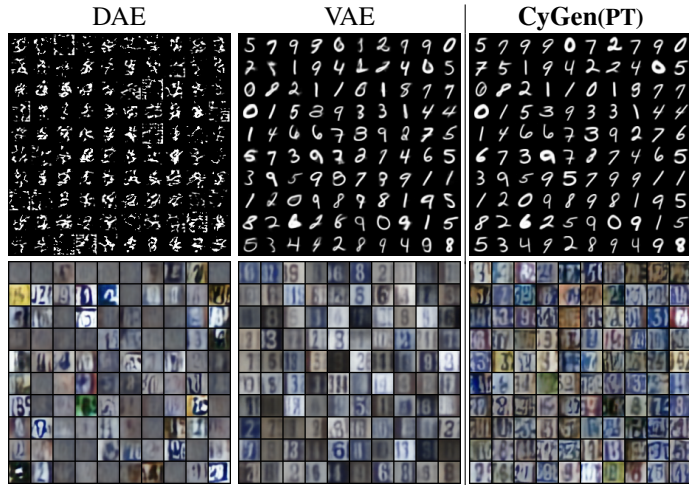


Figure 6: Generated data on the MNIST and SVHN datasets.

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

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. FFIORD: 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.

Supplementary Materials

A Background in Measure Theory

A.1 The Integral

The *integral* of a nonnegative measurable function f on a measure space $(\Omega, \mathcal{F}, \mu)$ is defined as:

$$\int f \, d\mu := \sup \sum_i \mu(\mathcal{W}^{(i)}) \inf_{\omega \in \mathcal{W}^{(i)}} f(\omega),$$

where the supremum is taken over all finite decompositions $\{\mathcal{W}^{(i)}\}$ of Ω into \mathcal{F} -sets [9, p.211]. For a general measurable function, its integral is defined as the subtraction from the integral of its positive part $f^+(\omega) := \max\{0, f(\omega)\}$ with the integral of its negative part $f^-(\omega) := \max\{0, -f(\omega)\}$. A measurable function is said to be μ -integrable [9, p.212] if both integrals of its positive and negative parts are finite.

(i) This is a general definition of integral. When Ω is an Euclidean space and μ is the Lebesgue measure on it, this integral reduces to the Lebesgue integral (which in turn coincides with the Riemann integral when the latter exists). When Ω is a discrete set (*i.e.*, a finite or countable set) and μ is the counting measure, this integral reduces to summation.

(ii) The integral satisfies common properties like linearity and monotonicity [9, Thm. 16.1], continuity under boundedness [9, Thm. 16.4, Thm. 16.5], *etc.* For a nonnegative function f , $\int f \, d\mu = 0$ if and only if $f = 0$, μ -a.e. [9, Thm. 15.2].

(iii) The integral over a set $\mathcal{W} \in \mathcal{F}$ is defined as $\int_{\mathcal{W}} f \, d\mu := \int \mathbb{I}_{\mathcal{W}} f \, d\mu$ [9, p.226], where $\mathbb{I}_{\mathcal{W}}$ is the indicator function.

(1) We thus sometimes also write $\int_{\Omega} f \, d\mu$ for $\int f \, d\mu$ to highlight the integral area. By this definition, $\int_{\mathcal{W}} f \, d\mu = 0$ if $\mu(\mathcal{W}) = 0$ [9, p.226].

(2) For two measurable functions f and g , if $f = g$, μ -a.e., then $\int_{\mathcal{W}} f \, d\mu = \int_{\mathcal{W}} g \, d\mu$ for any $\mathcal{W} \in \mathcal{F}$ [9, Thm. 15.2]. The inverse also holds if f and g are nonnegative and μ is sigma-finite, or f and g are integrable [9, Thm. 16.10(i,ii)]⁹.

(3) If f is a nonnegative measurable function, then $\nu(\mathcal{W}) := \int_{\mathcal{W}} f \, d\mu$, $\forall \mathcal{W} \in \mathcal{F}$, is a measure on (Ω, \mathcal{F}) [9, p.227]¹⁰. Such a measure ν is finite, if and only if f is μ -integrable.

A.2 Absolute Continuity and Radon-Nikodym Derivative

For two measures μ and ν on the same measurable space (Ω, \mathcal{F}) , ν is said to be *absolutely continuous* w.r.t μ , denoted as $\nu \ll \mu$, if $\mu(\mathcal{W}) = 0$ indicates $\nu(\mathcal{W}) = 0$ for $\mathcal{W} \in \mathcal{F}$ [9, p.448]. If μ and ν are sigma-finite and $\nu \ll \mu$, the *Radon-Nikodym theorem* [9, Thm. 32.2] asserts that there exists a μ -unique nonnegative function f on Ω , such that $\nu(\mathcal{W}) = \int_{\mathcal{W}} f(\omega) \mu(d\omega)$ for any $\mathcal{W} \in \mathcal{F}$. Such a function f is called the *Radon-Nikodym (R-N) derivative* of ν w.r.t μ , and is also denoted as $\frac{d\nu}{d\mu}$. It represents the density function of ν w.r.t base measure μ .

(i) Since the general definition of integral includes summation in the discrete case, this density function also includes the probability mass function in the discrete case.

(ii) The Dirac measure $\delta_{\omega_0}(\mathcal{W}) := \mathbb{I}_{\mathcal{W}}(\omega_0)$ (\mathbb{I} is the indicator function) at a single point $\omega_0 \in \Omega$ is not absolutely continuous on Euclidean spaces w.r.t the Lebesgue measure, which assigns measure 0 to the set $\{\omega_0\}$. To be strict, the Dirac delta function is not a proper density function, since its integrals covering ω_0 involve the indefinite $\infty \cdot 0$ on the component $\{\omega_0\}$ of the integral domain. Its characteristic that such integrals equal to one, is a standalone structure from being a function. So it is better treated as a measure of functional.

A.3 Product Measure Space

Two measure spaces $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$ induce a *product measure space* $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z}, \xi \otimes \zeta)$.

⁹9, Thm. 16.10(iii): $f = g$, μ -a.e., if $\int_{\mathcal{W}} f \, d\mu = \int_{\mathcal{W}} g \, d\mu$ for any \mathcal{W} from a pi-system Π that generates \mathcal{F} , and Ω is a finite or countable union of Π -sets.

¹⁰Its countable additivity is guaranteed by Billingsley [9, Thm. 16.9].

(i) The *product sigma-field* $\mathcal{X} \otimes \mathcal{Z} := \sigma(\mathcal{X} \times \mathcal{Z})$ is the smallest sigma-field on $\mathbb{X} \times \mathbb{Z}$ containing $\mathcal{X} \times \mathcal{Z}$ (24, Thm. 22; 65, Def. 7.1; equivalently, 42, Remark 14.10; 42, Def. 14.4). Note that the Cartesian product $\mathcal{X} \times \mathcal{Z}$, representing the set of *measurable rectangles*, is only a semiring (thus also a pi-system). So we need to extend for a sigma-field. For any $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, its *slice* (or section) at $z \in \mathbb{Z}$, defined by:

$$\mathcal{W}_z := \{x \mid (x, z) \in \mathcal{W}\},$$

lies in \mathcal{X} , and similarly $\mathcal{W}_x \in \mathcal{Z}$ [9, Thm. 18.1(i)]. We define the *projection* (or restriction) of \mathcal{W} onto \mathbb{Z} , as $\mathcal{W}^{\mathbb{Z}} := \{z \mid \exists x \in \mathbb{X} \text{ s.t. } (x, z) \in \mathcal{W}\}$. By definition, for any $z \in \mathbb{Z} \setminus \mathcal{W}^{\mathbb{Z}}$, $\mathcal{W}_z = \emptyset$.

(ii) The *product measure* $\xi \otimes \zeta$ is characterized by $(\xi \otimes \zeta)(\mathcal{X} \times \mathcal{Z}) = \xi(\mathcal{X})\zeta(\mathcal{Z})$ for measurable rectangles $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$. Some common conclusions require ξ and ζ to be sigma-finite on \mathcal{X} and \mathcal{Z} , respectively.

(1) In the characterization $(\xi \otimes \zeta)(\mathcal{X} \times \mathcal{Z}) = \xi(\mathcal{X})\zeta(\mathcal{Z})$, if the indefinite $0 \cdot \infty$ is met, it is zero. To see this, consider two sets \mathcal{X} and \mathcal{Z} that satisfy $\xi(\mathcal{X}) = 0$ and $\zeta(\mathcal{Z}) = 0$. Since ζ is sigma-finite, there are finite or countable disjoint \mathcal{Z} -sets $\mathcal{Z}^{(1)}, \mathcal{Z}^{(2)}, \dots$ such that $\zeta(\mathcal{Z}^{(i)}) < \infty$ for any $i \geq 1$ and $\bigcup_{i=1}^{\infty} \mathcal{Z}^{(i)} = \mathbb{Z}$. Redefining $\mathcal{Z}^{(i)}$ as $\mathcal{Z}^{(i)} \cap \mathcal{Z}$, we have $\bigcup_{i=1}^{\infty} \mathcal{Z}^{(i)} = \mathcal{Z}$ while still $\zeta(\mathcal{Z}^{(i)}) < \infty$. So $(\xi \otimes \zeta)(\mathcal{X} \times \mathbb{Z}) = (\xi \otimes \zeta)(\mathcal{X} \times \bigcup_{i=1}^{\infty} \mathcal{Z}^{(i)}) = (\xi \otimes \zeta)(\bigcup_{i=1}^{\infty} \mathcal{X} \times \mathcal{Z}^{(i)})$. Recalling that a measure is countably additive by definition, this is $= \sum_{i=1}^{\infty} (\xi \otimes \zeta)(\mathcal{X} \times \mathcal{Z}^{(i)}) = \sum_{i=1}^{\infty} \xi(\mathcal{X})\zeta(\mathcal{Z}^{(i)}) = 0$.

(2) In this case, such a $\xi \otimes \zeta$ is sigma-finite on $\mathcal{X} \times \mathcal{Z}$, and the characterization on the pi-system $\mathcal{X} \times \mathcal{Z}$ determines a unique sigma-finite measure on $\sigma(\mathcal{X} \times \mathcal{Z}) = \mathcal{X} \otimes \mathcal{Z}$ [9, Thm. 10.3]. See also Galambos [24, Thm. 22]; Klenke [42, Thm. 14.14]; Rinaldo [65, Thm. 7.9]. Moreover, we have [9, Thm. 18.2]:

$$(\xi \otimes \zeta)(\mathcal{W}) = \int_{\mathbb{Z}} \xi(\mathcal{W}_z) \zeta(dz) = \int_{\mathbb{X}} \zeta(\mathcal{W}_x) \xi(dx), \quad \forall \mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}. \quad (6)$$

Since for any $z \in \mathbb{Z} \setminus \mathcal{W}^{\mathbb{Z}}$, $\mathcal{W}_z = \emptyset$ (see (i)) thus $\xi(\mathcal{W}_z) = 0$, we also have (by leveraging the additivity of integrals over a countable partition [9, Thm. 16.9] and that an a.e. zero function gives a zero integral [9, Thm. 15.2(i)]):

$$(\xi \otimes \zeta)(\mathcal{W}) = \int_{\mathcal{W}^{\mathbb{Z}}} \xi(\mathcal{W}_z) \zeta(dz) = \int_{\mathcal{W}^{\mathbb{X}}} \zeta(\mathcal{W}_x) \xi(dx), \quad \forall \mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}. \quad (7)$$

(iii) For a function f on $\mathbb{X} \times \mathbb{Z}$, if it is $\mathcal{X} \otimes \mathcal{Z}$ -measurable, then $f(x, \cdot)$ is \mathcal{Z} -measurable for any $x \in \mathbb{X}$, and $f(\cdot, z)$ is \mathcal{X} -measurable for any $z \in \mathbb{Z}$ [9, Thm. 18.1(ii)]. When f is $\xi \otimes \zeta$ -integrable, Fubini's theorem [9, Thm. 18.3] asserts its integral on $\mathbb{X} \times \mathbb{Z}$ can be computed iteratedly in either order:

$$\int_{\mathbb{X} \times \mathbb{Z}} f(x, z) (\xi \otimes \zeta)(dxdz) = \int_{\mathbb{Z}} \left(\int_{\mathbb{X}} f(x, z) \xi(dx) \right) \zeta(dz) = \int_{\mathbb{X}} \left(\int_{\mathbb{Z}} f(x, z) \zeta(dz) \right) \xi(dx). \quad (8)$$

For any $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, the same equalities hold for function $\mathbb{I}_{\mathcal{W}} f$. For the first iterated integral, we have $\int_{\mathbb{X}} \mathbb{I}_{\mathcal{W}}(x, z) f(x, z) \xi(dx) = \int_{\mathbb{X}} \mathbb{I}_{\mathcal{W}_z}(x) f(x, z) \xi(dx) = \int_{\mathcal{W}_z} f(x, z) \xi(dx)$, and on the region $\mathbb{Z} \setminus \mathcal{W}^{\mathbb{Z}}$, the integral $\int_{\mathcal{W}_z} f(x, z) \xi(dx) = 0$ [9, p.226] since $\mathcal{W}_z = \emptyset$ on that region (see (i)). So we have a more general form of Fubini's theorem:

$$\int_{\mathcal{W}} f(x, z) (\xi \otimes \zeta)(dxdz) = \int_{\mathcal{W}^{\mathbb{Z}}} \left(\int_{\mathcal{W}_z} f(x, z) \xi(dx) \right) \zeta(dz) = \int_{\mathcal{W}^{\mathbb{X}}} \left(\int_{\mathcal{W}_x} f(x, z) \zeta(dz) \right) \xi(dx). \quad (9)$$

(iv) For a measure π on the product measurable space $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$, define its marginal distributions: $\pi^{\mathbb{X}}(\mathcal{X}) := \pi(\mathcal{X} \times \mathbb{Z}), \forall \mathcal{X} \in \mathcal{X}$, and $\pi^{\mathbb{Z}}(\mathcal{Z}) := \pi(\mathbb{X} \times \mathcal{Z}), \forall \mathcal{Z} \in \mathcal{Z}$.

A.4 Conditional Distributions

In the most general case, a distribution (probability measure) π on a measurable space (Ω, \mathcal{F}) gives a *conditional distribution* (conditional probability) $\pi(\mathcal{W}|\omega)$ for $\mathcal{W} \in \mathcal{F}$ w.r.t a sub-sigma-field $\mathcal{G} \subseteq \mathcal{F}$.

(i) For any $\mathcal{W} \in \mathcal{F}$, the function $\mathcal{G} \rightarrow \mathbb{R}^{\geq 0}, \mathcal{G} \mapsto \pi(\mathcal{G} \cap \mathcal{W})$ gives a measure on \mathcal{G} . It is absolutely continuous w.r.t $\pi^{\mathcal{G}} : \mathcal{G} \rightarrow \mathbb{R}^{\geq 0}, \mathcal{G} \mapsto \pi(\mathcal{G})$, the projection of π onto \mathcal{G} , due to the monotonicity (or (sub-)additivity) of measures. So the R-N derivative on \mathcal{G} exists, which defines the conditional

distribution [9, p.457]:

$$\pi(\mathcal{W}|\omega) := \frac{d\pi(\cdot \cap \mathcal{W})}{d\pi^{\mathcal{G}}(\cdot)}(\omega),$$

where $\omega \in \Omega$. Note that as defined as an R-N derivative, the conditional distribution is only $\pi^{\mathcal{G}}$ -unique.

(ii) As a function of ω , $\pi(\mathcal{W}|\omega)$ is \mathcal{G} -measurable and π -integrable, and satisfies [9, p.457, Thm. 33.1]:

$$\int_{\mathcal{G}} \pi(\mathcal{W}|\omega) \pi^{\mathcal{G}}(d\omega) = \pi(\mathcal{G} \cap \mathcal{W}), \forall \mathcal{G} \in \mathcal{G}. \quad (10)$$

This could serve as an alternative definition of conditional probability.

(iii) For $\pi^{\mathcal{G}}$ -a.e. ω , $\pi(\cdot|\omega)$ is a distribution (probability measure) on (Ω, \mathcal{F}) [9, Thm. 33.2].

(iv) Conditional distributions on a product measurable space $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$. Consider the sub-sigma-field $\mathcal{G} := \{\mathbb{X}\} \times \mathcal{Z}$. By construction, any $\mathcal{G} \in \mathcal{G}$ can be formed by $\mathcal{G} = \mathbb{X} \times \mathcal{Z}$ for some $\mathcal{Z} \in \mathcal{Z}$. So $\pi^{\mathcal{G}}(\mathcal{G}) := \pi(\mathcal{G}) = \pi(\mathbb{X} \times \mathcal{Z}) =: \pi^{\mathbb{Z}}(\mathcal{Z})$, and Eq. (10) becomes $\pi((\mathbb{X} \times \mathcal{Z}) \cap \mathcal{W}) = \int_{\mathbb{X} \times \mathcal{Z}} \pi(\mathcal{W}|x, z) \pi^{\mathcal{G}}(dx dz) = \int_{\mathbb{X} \times \mathcal{Z}} \pi(\mathcal{W}|x, z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{Z}} \pi(\mathcal{W}|x, z) \pi^{\mathbb{Z}}(dz)$. This indicates that the conditional probability $\pi(\mathcal{W}|x, z)$ in this case is constant w.r.t x . We hence denote it as $\pi(\mathcal{W}|z)$.

Consider $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$ in the form $\mathcal{W} = \mathcal{X} \times \mathbb{Z}$ for some $\mathcal{X} \in \mathcal{X}$. For any $\mathcal{G} = \mathbb{X} \times \mathcal{Z} \in \mathcal{G}$, we have from Eq. (10) that $\int_{\mathcal{G}} \pi(\mathcal{W}|z) \pi^{\mathcal{G}}(dx dz) = \pi(\mathcal{G} \cap \mathcal{W}) = \pi(\mathcal{X} \times \mathcal{Z})$. From the above deduction, the l.h.s is $\int_{\mathbb{X} \times \mathcal{Z}} \pi(\mathcal{W}|z) \pi^{\mathcal{G}}(dx dz) = \int_{\mathcal{Z}} \pi(\mathcal{X} \times \mathbb{Z}|z) \pi^{\mathbb{Z}}(dz)$. Defining $\pi(\mathcal{X}|z)$ as $\pi(\mathcal{X} \times \mathbb{Z}|z)$ for any $\mathcal{X} \in \mathcal{X}$, we have:

$$\pi(\mathcal{X} \times \mathbb{Z}) = \int_{\mathcal{Z}} \pi(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{X}} \pi(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx), \forall \mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}. \quad (11)$$

This is the conditional distribution in the usual sense. Note again that as defined as R-N derivatives, the conditional distributions $\pi(\mathcal{X}|z)$ and $\pi(\mathcal{Z}|x)$ are only $\pi^{\mathbb{Z}}$ -unique and $\pi^{\mathbb{X}}$ -unique, respectively.

For any $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, define $\tilde{\pi}(\mathcal{W}) := \int_{\mathbb{Z}} \pi(\mathcal{W}_z|z) \pi^{\mathbb{Z}}(dz)$. It is easy to verify that $\tilde{\pi}$ is a distribution (probability measure; thus finite and sigma-finite) on $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$ [9, p.227], since $\pi(\mathcal{X}|z)$ is a distribution (and thus nonnegative) on $(\mathbb{X}, \mathcal{X})$ for $\pi^{\mathbb{Z}}$ -a.e. z [9, Thm. 33.2]. For any $\mathcal{W} = \mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, $\tilde{\pi}(\mathcal{W}) = \int_{\mathbb{Z}} \pi(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{Z}} \pi(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = \pi(\mathcal{X} \times \mathcal{Z})$ due to Eq. (11). So $\tilde{\pi}$ and π agree on the pi-system $\mathcal{X} \times \mathcal{Z}$, which indicates that they agree on $\sigma(\mathcal{X} \times \mathcal{Z}) = \mathcal{X} \otimes \mathcal{Z}$ due to Billingsley [9, Thm. 10.3, Thm. 3.3]. This means that (see the argument in (ii) (2) in Supplement A.3 for the second line of the equation):

$$\begin{aligned} \pi(\mathcal{W}) &= \int_{\mathbb{Z}} \pi(\mathcal{W}_z|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathbb{X}} \pi(\mathcal{W}_x|x) \pi^{\mathbb{X}}(dx) \\ &= \int_{\mathcal{W}^{\mathbb{Z}}} \pi(\mathcal{W}_z|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{W}^{\mathbb{X}}} \pi(\mathcal{W}_x|x) \pi^{\mathbb{X}}(dx), \forall \mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}. \end{aligned} \quad (12)$$

Finally, we formalize some definitions in the main text below.

Definition A.1. Consider a general measure space $(\Omega, \mathcal{F}, \mu)$. (i) We say that two measurable sets $\mathcal{S}, \tilde{\mathcal{S}} \in \mathcal{F}$ are μ -a.s. the same, denoted as “ $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$ ”, if $\mu(\mathcal{S} \triangle \tilde{\mathcal{S}}) = 0$, where “ \triangle ” denotes the symmetric difference between two sets. (ii) We say that \mathcal{S} is a μ -a.s. subset of \mathcal{W} , denoted as “ $\mathcal{S} \subseteq^{\mu} \mathcal{W}$ ”, if $\mu(\mathcal{S} \setminus \mathcal{W}) = 0$.

B Lemmas

B.1 Lemmas for General Probability

Lemma B.1. Let \mathcal{O} be a measure-zero set, $\mu(\mathcal{O}) = 0$, on a measure space $(\Omega, \mathcal{F}, \mu)$. Then for any measurable set \mathcal{W} , we have $\mu(\mathcal{O} \setminus \mathcal{W}) = \mu(\mathcal{W} \cap \mathcal{O}) = 0$, and $\mu(\mathcal{W} \cup \mathcal{O}) = \mu(\mathcal{W} \setminus \mathcal{O}) = \mu(\mathcal{W})$.

Proof. Due to the monotonicity of a measure [9, Thm. 16.1], we have $\mu(\mathcal{O} \setminus \mathcal{W}) \leq \mu(\mathcal{O}) = 0$ and $\mu(\mathcal{W} \cap \mathcal{O}) \leq \mu(\mathcal{O}) = 0$, so we get $\mu(\mathcal{O} \setminus \mathcal{W}) = \mu(\mathcal{W} \cap \mathcal{O}) = 0$. Since $\mu(\mathcal{W} \cup \mathcal{O}) = \mu(\mathcal{W} \cup (\mathcal{O} \setminus \mathcal{W}))$ and the two sets are disjoint, it equals to $\mu(\mathcal{W}) + \mu(\mathcal{O} \setminus \mathcal{W})$, which is $\mu(\mathcal{W})$ by the above conclusion. So we get $\mu(\mathcal{W} \cup \mathcal{O}) = \mu(\mathcal{W})$. When applying this conclusion to $\mathcal{W} \setminus \mathcal{O}$, we have $\mu((\mathcal{W} \setminus \mathcal{O}) \cup \mathcal{O}) = \mu(\mathcal{W} \setminus \mathcal{O})$, while the l.h.s is $\mu(\mathcal{W} \cup \mathcal{O})$ which is $\mu(\mathcal{W})$ by the same conclusion. So we get $\mu(\mathcal{W} \setminus \mathcal{O}) = \mu(\mathcal{W})$. \square

Lemma B.2. Let π be an absolutely continuous distribution (probability measure) on a measure space $(\Omega, \mathcal{F}, \mu)$ with a density function f , and let $\mathcal{S} \in \mathcal{F}$ be a measurable set. Then $\pi(\mathcal{S}) = 1$ if and only if $\pi(\mathcal{W}) = \int_{\mathcal{W} \cap \mathcal{S}} f \, d\mu, \forall \mathcal{W} \in \mathcal{F}$.

Proof. “Only if”: Since $\mathcal{S} \subseteq \Omega$, we have $\pi(\Omega \setminus \mathcal{S}) = \pi(\Omega) - \pi(\mathcal{S}) = 0$. For any $\mathcal{W} \in \mathcal{F}$, we have $\pi(\mathcal{W}) = \pi(\mathcal{W} \cap \mathcal{S}) + \pi(\mathcal{W} \cap (\Omega \setminus \mathcal{S}))$, while $0 \leq \pi(\mathcal{W} \cap (\Omega \setminus \mathcal{S})) \leq \pi(\Omega \setminus \mathcal{S}) = 0$. So we have $\pi(\mathcal{W}) = \pi(\mathcal{W} \cap \mathcal{S}) = \int_{\mathcal{W} \cap \mathcal{S}} f \, d\mu$.

“If”: $1 = \pi(\Omega) = \int_{\Omega \cap \mathcal{S}} f \, d\mu = \int_{\mathcal{S}} f \, d\mu = \int_{\mathcal{S} \cap \mathcal{S}} f \, d\mu = \pi(\mathcal{S})$. \square

Lemma B.3. Let \mathcal{S} and $\tilde{\mathcal{S}}$ be two measurable sets on a measure space $(\Omega, \mathcal{F}, \mu)$ such that $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$. Then $\mu(\mathcal{S} \setminus \tilde{\mathcal{S}}) = \mu(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$, and $\mu(\mathcal{S}) = \mu(\tilde{\mathcal{S}}) = \mu(\mathcal{S} \cup \tilde{\mathcal{S}}) = \mu(\mathcal{S} \cap \tilde{\mathcal{S}})$.

Proof. Let $\mathcal{D}^+ := \tilde{\mathcal{S}} \setminus \mathcal{S}$ and $\mathcal{D}^- := \mathcal{S} \setminus \tilde{\mathcal{S}}$. By construction, we have $\mathcal{D}^+ \cap \mathcal{S} = \emptyset$ and $\mathcal{D}^- \subseteq \mathcal{S}$, so we also have $\mathcal{D}^+ \cap \mathcal{D}^- = \emptyset$, and $\tilde{\mathcal{S}} = (\mathcal{S} \setminus \mathcal{D}^-) \cup \mathcal{D}^+ = (\mathcal{S} \cup \mathcal{D}^+) \setminus \mathcal{D}^-$. By definition, $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$ indicates $0 = \mu(\mathcal{S} \Delta \tilde{\mathcal{S}}) = \mu(\mathcal{D}^+ \cup \mathcal{D}^-) = \mu(\mathcal{D}^+) + \mu(\mathcal{D}^-)$, so we have both $\mu(\mathcal{D}^+) = 0$ and $\mu(\mathcal{D}^-) = 0$. Subsequently, $\mu(\tilde{\mathcal{S}}) = \mu((\mathcal{S} \setminus \mathcal{D}^-) \cup \mathcal{D}^+) = \mu(\mathcal{S} \setminus \mathcal{D}^-) + \mu(\mathcal{D}^+) = \mu(\mathcal{S} \setminus \mathcal{D}^-) = \mu(\mathcal{S}) - \mu(\mathcal{D}^- \cap \mathcal{S}) = \mu(\mathcal{S}) - \mu(\mathcal{D}^-) = \mu(\mathcal{S})$, and $\mu(\mathcal{S} \cup \tilde{\mathcal{S}}) = \mu(\mathcal{S} \cup \mathcal{D}^+) = \mu(\mathcal{S}) + \mu(\mathcal{D}^+) = \mu(\mathcal{S})$. Noting also that $\mathcal{S} \cup \tilde{\mathcal{S}} = (\mathcal{S} \cap \tilde{\mathcal{S}}) \cup (\mathcal{S} \Delta \tilde{\mathcal{S}})$ and that this is a disjoint union, we have $\mu(\mathcal{S} \cup \tilde{\mathcal{S}}) = \mu(\mathcal{S} \cap \tilde{\mathcal{S}}) + \mu(\mathcal{S} \Delta \tilde{\mathcal{S}}) = \mu(\mathcal{S} \cap \tilde{\mathcal{S}})$. \square

Lemma B.4. On a measure space $(\Omega, \mathcal{F}, \mu)$, “ $\cdot \stackrel{\mu}{=} \cdot$ ” is an equivalence relation.

Proof. Symmetry and reflexivity are obvious. For transitivity, let \mathcal{A}, \mathcal{B} and \mathcal{C} be three measurable sets such that $\mathcal{A} \stackrel{\mu}{=} \mathcal{B}$ and $\mathcal{B} \stackrel{\mu}{=} \mathcal{C}$. Since $\mathcal{A} \setminus \mathcal{C} = ((\mathcal{A} \setminus \mathcal{C}) \cap \mathcal{B}) \cup ((\mathcal{A} \setminus \mathcal{C}) \setminus \mathcal{B}) = (\mathcal{A} \cap (\mathcal{B} \setminus \mathcal{C})) \cup ((\mathcal{A} \setminus \mathcal{B}) \setminus \mathcal{C}) \subseteq (\mathcal{B} \setminus \mathcal{C}) \cup (\mathcal{A} \setminus \mathcal{B})$, we have $\mu(\mathcal{A} \setminus \mathcal{C}) \leq \mu(\mathcal{B} \setminus \mathcal{C}) + \mu(\mathcal{A} \setminus \mathcal{B}) = 0$ due to Lemma B.3. Similarly, $\mu(\mathcal{C} \setminus \mathcal{A}) = 0$. So $\mu(\mathcal{A} \Delta \mathcal{C}) = \mu(\mathcal{A} \setminus \mathcal{C}) + \mu(\mathcal{C} \setminus \mathcal{A}) = 0$. \square

Lemma B.5. Let \mathcal{S} and $\tilde{\mathcal{S}}$ be two measurable sets on a measure space $(\Omega, \mathcal{F}, \mu)$ such that $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$. Then for any measurable set \mathcal{W} , we have $\mathcal{S} \cup \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \cup \mathcal{W}$, $\mathcal{S} \cap \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \cap \mathcal{W}$, $\mathcal{S} \setminus \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \setminus \mathcal{W}$ and $\mathcal{W} \setminus \mathcal{S} \stackrel{\mu}{=} \mathcal{W} \setminus \tilde{\mathcal{S}}$.

Proof. Let $\mathcal{D}^+ := \tilde{\mathcal{S}} \setminus \mathcal{S}$ and $\mathcal{D}^- := \mathcal{S} \setminus \tilde{\mathcal{S}}$. By Lemma B.3, we have $\mu(\mathcal{D}^+) = 0$ and $\mu(\mathcal{D}^-) = 0$.

For any measurable set \mathcal{W} , we have $(\tilde{\mathcal{S}} \cup \mathcal{W}) \setminus (\mathcal{S} \cup \mathcal{W}) = \tilde{\mathcal{S}} \setminus \mathcal{S} \setminus \mathcal{W} = \mathcal{D}^+ \setminus \mathcal{W}$, and similarly $(\mathcal{S} \cup \mathcal{W}) \setminus (\tilde{\mathcal{S}} \cup \mathcal{W}) = \mathcal{D}^- \setminus \mathcal{W}$. So $\mu((\mathcal{S} \cup \mathcal{W}) \Delta (\tilde{\mathcal{S}} \cup \mathcal{W})) = \mu(((\mathcal{S} \cup \mathcal{W}) \setminus (\tilde{\mathcal{S}} \cup \mathcal{W})) \cup ((\tilde{\mathcal{S}} \cup \mathcal{W}) \setminus (\mathcal{S} \cup \mathcal{W}))) = \mu((\mathcal{D}^- \setminus \mathcal{W}) \cup (\mathcal{D}^+ \setminus \mathcal{W})) = \mu(\mathcal{D}^- \setminus \mathcal{W}) + \mu(\mathcal{D}^+ \setminus \mathcal{W}) \leq \mu(\mathcal{D}^-) + \mu(\mathcal{D}^+) = 0$, that is $\mathcal{S} \cup \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \cup \mathcal{W}$.

Since $(\tilde{\mathcal{S}} \cap \mathcal{W}) \setminus (\mathcal{S} \cap \mathcal{W}) = (\tilde{\mathcal{S}} \setminus \mathcal{S}) \cap \mathcal{W} = \mathcal{D}^+ \cap \mathcal{W}$ and similarly $(\mathcal{S} \cap \mathcal{W}) \setminus (\tilde{\mathcal{S}} \cap \mathcal{W}) = \mathcal{D}^- \cap \mathcal{W}$, we have $\mu((\mathcal{S} \cap \mathcal{W}) \Delta (\tilde{\mathcal{S}} \cap \mathcal{W})) = \mu(((\mathcal{S} \cap \mathcal{W}) \setminus (\tilde{\mathcal{S}} \cap \mathcal{W})) \cup ((\tilde{\mathcal{S}} \cap \mathcal{W}) \setminus (\mathcal{S} \cap \mathcal{W}))) = \mu((\mathcal{D}^- \cap \mathcal{W}) \cup (\mathcal{D}^+ \cap \mathcal{W})) = \mu(\mathcal{D}^- \cap \mathcal{W}) + \mu(\mathcal{D}^+ \cap \mathcal{W}) \leq \mu(\mathcal{D}^-) + \mu(\mathcal{D}^+) = 0$, so $\mathcal{S} \cap \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \cap \mathcal{W}$.

Since $(\tilde{\mathcal{S}} \setminus \mathcal{W}) \setminus (\mathcal{S} \setminus \mathcal{W}) = \tilde{\mathcal{S}} \setminus \mathcal{W} \setminus \mathcal{S} = \tilde{\mathcal{S}} \setminus \mathcal{S} \setminus \mathcal{W} = \mathcal{D}^+ \setminus \mathcal{W}$ and similarly $(\mathcal{S} \setminus \mathcal{W}) \setminus (\tilde{\mathcal{S}} \setminus \mathcal{W}) = \mathcal{D}^- \setminus \mathcal{W}$, we have $\mu((\mathcal{S} \setminus \mathcal{W}) \Delta (\tilde{\mathcal{S}} \setminus \mathcal{W})) = \mu(((\mathcal{S} \setminus \mathcal{W}) \setminus (\tilde{\mathcal{S}} \setminus \mathcal{W})) \cup ((\tilde{\mathcal{S}} \setminus \mathcal{W}) \setminus (\mathcal{S} \setminus \mathcal{W}))) = \mu((\mathcal{D}^- \setminus \mathcal{W}) \cup (\mathcal{D}^+ \setminus \mathcal{W})) = \mu(\mathcal{D}^- \setminus \mathcal{W}) + \mu(\mathcal{D}^+ \setminus \mathcal{W}) \leq \mu(\mathcal{D}^-) + \mu(\mathcal{D}^+) = 0$, so $\mathcal{S} \setminus \mathcal{W} \stackrel{\mu}{=} \tilde{\mathcal{S}} \setminus \mathcal{W}$.

Since $(\mathcal{W} \setminus \tilde{\mathcal{S}}) \setminus (\mathcal{W} \setminus \mathcal{S}) = \mathcal{W} \setminus (\mathcal{W} \setminus \mathcal{S}) \setminus \tilde{\mathcal{S}} = (\mathcal{W} \cap \mathcal{S}) \setminus \tilde{\mathcal{S}} = (\mathcal{S} \setminus \tilde{\mathcal{S}}) \cap \mathcal{W} = \mathcal{D}^- \cap \mathcal{W}$ and similarly $(\mathcal{W} \setminus \mathcal{S}) \setminus (\mathcal{W} \setminus \tilde{\mathcal{S}}) = \mathcal{D}^+ \cap \mathcal{W}$, we have $\mu((\mathcal{W} \setminus \mathcal{S}) \Delta (\mathcal{W} \setminus \tilde{\mathcal{S}})) = \mu(((\mathcal{W} \setminus \mathcal{S}) \setminus (\mathcal{W} \setminus \tilde{\mathcal{S}})) \cup ((\mathcal{W} \setminus \tilde{\mathcal{S}}) \setminus (\mathcal{W} \setminus \mathcal{S}))) = \mu((\mathcal{D}^+ \cap \mathcal{W}) \cup (\mathcal{D}^- \cap \mathcal{W})) = \mu(\mathcal{D}^+ \cap \mathcal{W}) + \mu(\mathcal{D}^- \cap \mathcal{W}) \leq \mu(\mathcal{D}^+) + \mu(\mathcal{D}^-) = 0$, so $\mathcal{W} \setminus \mathcal{S} \stackrel{\mu}{=} \mathcal{W} \setminus \tilde{\mathcal{S}}$. \square

Definition B.6. We say that a set satisfying a certain condition is μ -unique, if for any two such sets \mathcal{S} and $\tilde{\mathcal{S}}$, it holds that $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$.

Lemma B.7. Let π be an absolutely continuous distribution (probability measure) on a measure space $(\Omega, \mathcal{F}, \mu)$ with a density function f . If a set $\mathcal{S} \in \mathcal{F}$ satisfies $\pi(\mathcal{S}) = 1$ and that $f > 0$, μ -a.e. on \mathcal{S} , then such an \mathcal{S} is μ -unique.

Proof. Suppose we have two such sets \mathcal{S} and $\tilde{\mathcal{S}}$. By Lemma B.2, we know that for any $\mathcal{W} \in \mathcal{F}$, $\pi(\mathcal{W}) = \int_{\mathcal{W} \cap \mathcal{S}} f \, d\mu = \int_{\mathcal{W}} \mathbb{I}_{\mathcal{S}} f \, d\mu = \int_{\mathcal{W}} \mathbb{I}_{\tilde{\mathcal{S}}} f \, d\mu$. So by Billingsley [9, Thm. 16.10(ii)], we know that $\mathbb{I}_{\mathcal{S}} f = \mathbb{I}_{\tilde{\mathcal{S}}} f$, μ -a.e.

Since $f > 0$, μ -a.e. on \mathcal{S} , we know that $\mathbb{I}_{\mathcal{S}} = \mathbb{I}_{\tilde{\mathcal{S}}}$, μ -a.e. on \mathcal{S} . This means that $\mu\{\omega \in \mathcal{S} \mid \mathbb{I}_{\mathcal{S}} \neq \mathbb{I}_{\tilde{\mathcal{S}}}\} = \mu\{\omega \in \mathcal{S} \mid \omega \notin \tilde{\mathcal{S}}\} = \mu(\mathcal{S} \setminus \tilde{\mathcal{S}}) = 0$. Symmetrically, since $f > 0$, μ -a.e. also on $\tilde{\mathcal{S}}$, we know that $\mu(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$. So we have $\mu(\mathcal{S} \triangle \tilde{\mathcal{S}}) = \mu((\mathcal{S} \setminus \tilde{\mathcal{S}}) \cup (\tilde{\mathcal{S}} \setminus \mathcal{S})) = \mu(\mathcal{S} \setminus \tilde{\mathcal{S}}) + \mu(\tilde{\mathcal{S}} \setminus \mathcal{S}) = 0$, which means that $\mathcal{S} \stackrel{\mu}{=} \tilde{\mathcal{S}}$. \square

The μ -unique set \mathcal{S} in the lemma serves as another form of the *support* of a distribution. The standard definition of the support requires a topological structure and \mathcal{F} is the corresponding Borel sigma-field. If given absolute continuity $\pi \ll \mu$, this lemma enables the generality that does not require a topological structure. The condition $\pi(\mathcal{S}) = 1$ prevents \mathcal{S} to be too small, while the condition that $f > 0$, μ -a.e. on \mathcal{S} prevents \mathcal{S} to be too large.

Definition B.8 (support of an absolutely continuous distribution (without topology)). Define the *support* of an absolutely continuous distribution (probability measure) π on a measure space $(\Omega, \mathcal{F}, \mu)$, as the μ -unique set $\mathcal{S} \in \mathcal{F}$ such that $\pi(\mathcal{S}) = 1$ and for any density function f of π , it holds that $f > 0$, μ -a.e. on \mathcal{S} .

B.2 Lemmas for Product Probability

In this subsection and the following, let $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z}, \xi \otimes \zeta)$ be the product measure space by the two individual ones $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$, where ξ and ζ are sigma-finite.

Lemma B.9. For a measure π on the product measure space $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z}, \xi \otimes \zeta)$, if $\pi \ll \xi \otimes \zeta$, then $\pi^{\mathbb{X}} \ll \xi$ and $\pi^{\mathbb{Z}} \ll \zeta$.

Proof. For any $\mathcal{X} \in \mathcal{X}$ such that $\xi(\mathcal{X}) = 0$, we have $(\xi \otimes \zeta)(\mathcal{X} \times \mathbb{Z}) = \xi(\mathcal{X})\zeta(\mathbb{Z}) = 0$, where the last equality is verified in (ii) (1) in Supplement A.3 when $\zeta(\mathbb{Z}) = \infty$. Since $\pi \ll \xi \otimes \zeta$, this means that $\pi(\mathcal{X} \times \mathbb{Z}) = \pi^{\mathbb{X}}(\mathcal{X}) = 0$. So $\pi^{\mathbb{X}} \ll \xi$. Similarly, $\pi^{\mathbb{Z}} \ll \zeta$. \square

Lemma B.10. For an assertion $t(x, z)$ on $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, $t(x, z)$ holds $\xi \otimes \zeta$ -a.e. on \mathcal{W} , if and only if $t(x, z)$ holds ξ -a.e. on \mathcal{W}_z , for ζ -a.e. z on $\mathcal{W}^{\mathbb{Z}}$.

Proof. By the definition of “ $t(x, z)$ holds $\xi \otimes \zeta$ -a.e. on \mathcal{W} ”, we have:

$$\begin{aligned} & (\xi \otimes \zeta)\{(x, z) \in \mathcal{W} \mid \neg t(x, z)\} = 0 \quad (\text{Since } \xi \text{ and } \zeta \text{ are sigma-finite, from Eq. (7),}) \\ \iff & \int_{\mathcal{W}^{\mathbb{Z}}} \xi\{x \in \mathcal{W}_z \mid \neg t(x, z)\} \zeta(dz) = 0 \\ & (\text{Since } \xi(\cdot) \text{ is nonnegative, from Billingsley [9, Thm. 15.2],}) \\ \iff & \xi\{x \in \mathcal{W}_z \mid \neg t(x, z)\} = 0, \text{ for } \zeta\text{-a.e. } z \text{ on } \mathcal{W}^{\mathbb{Z}}, \\ & \text{which is “} t(x, z) \text{ holds } \xi\text{-a.e. on } \mathcal{W}_z, \text{ for } \zeta\text{-a.e. } z \text{ on } \mathcal{W}^{\mathbb{Z}}\text{”}. \end{aligned} \quad \square$$

Lemma B.11. Let $\mathcal{X}, \tilde{\mathcal{X}} \in \mathcal{X}$ such that $\mathcal{X} \stackrel{\xi}{=} \tilde{\mathcal{X}}$. Then $\mathcal{X} \times \mathbb{Z} \stackrel{\xi \otimes \zeta}{=} \tilde{\mathcal{X}} \times \mathbb{Z}$.

Proof. Since $(\mathcal{X} \times \mathbb{Z}) \triangle (\tilde{\mathcal{X}} \times \mathbb{Z}) = ((\mathcal{X} \times \mathbb{Z}) \setminus (\tilde{\mathcal{X}} \times \mathbb{Z})) \cup ((\tilde{\mathcal{X}} \times \mathbb{Z}) \setminus (\mathcal{X} \times \mathbb{Z})) = ((\mathcal{X} \setminus \tilde{\mathcal{X}}) \cup (\tilde{\mathcal{X}} \setminus \mathcal{X})) \times \mathbb{Z}$, we can verify that $(\xi \otimes \zeta)((\mathcal{X} \times \mathbb{Z}) \triangle (\tilde{\mathcal{X}} \times \mathbb{Z})) = (\xi \otimes \zeta)((\mathcal{X} \setminus \tilde{\mathcal{X}}) \cup (\tilde{\mathcal{X}} \setminus \mathcal{X})) \times \mathbb{Z} = \xi((\mathcal{X} \setminus \tilde{\mathcal{X}}) \cup (\tilde{\mathcal{X}} \setminus \mathcal{X}))\zeta(\mathbb{Z}) = \xi(\mathcal{X} \triangle \tilde{\mathcal{X}})\zeta(\mathbb{Z}) = 0$, where the last equality is verified in (ii) (1) in Supplement A.3 when $\zeta(\mathbb{Z}) = \infty$. \square

B.3 Lemmas for $\xi \otimes \zeta$ -Complete Component

Echoing Def. 2.2, a set $\mathcal{S} \in \mathcal{X} \otimes \mathcal{Z}$ is called a $\xi \otimes \zeta$ -complete component of $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$, if

$$\mathcal{S}^{\#} \cap \mathcal{W} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}, \text{ where } \mathcal{S}^{\#} := \mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^{\mathbb{Z}}. \quad (13)$$

This means that \mathcal{S} is complete under *stretching* and intersecting with \mathcal{W} .

Lemma B.12. Let \mathcal{S} be a $\xi \otimes \zeta$ -complete component of \mathcal{W} . Then $\mathcal{S} \subseteq \stackrel{\xi \otimes \zeta}{=} \mathcal{W}$.

Proof. By construction, we have $\mathcal{S} \subseteq \mathcal{S}^{\#}$ so $\mathcal{S} \setminus \mathcal{W} = \mathcal{S} \setminus (\mathcal{S} \cap \mathcal{W}) = \mathcal{S} \setminus (\mathcal{S}^{\#} \cap \mathcal{W})$. Hence, $(\xi \otimes \zeta)(\mathcal{S} \setminus \mathcal{W}) = (\xi \otimes \zeta)(\mathcal{S} \setminus (\mathcal{S}^{\#} \cap \mathcal{W})) = 0$ by definition Eq. (13) and Lemma B.3. \square

Example B.13. Note that when \mathcal{S} is a $\xi \otimes \zeta$ -complete component of \mathcal{W} , it may not hold that $\mathcal{S}^{\mathbb{X}} \subseteq^{\xi} \mathcal{W}^{\mathbb{X}}$ and $\mathcal{S}^{\mathbb{Z}} \subseteq^{\zeta} \mathcal{W}^{\mathbb{Z}}$. Fig. 8 shows an example, where $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$ are the one dimensional Euclidean spaces with line Borel sigma-field and line Lebesgue measure, $(\mathbb{R}, \mathcal{R}, \lambda)$, and $\mathcal{W} := [0, 1]^2$ and $\mathcal{S} := [0, 1]^2 \cup ([1, 2] \times \{\frac{1}{2}\})$. We have $\mathcal{S}^{\mathbb{X}} = [0, 2]$ so $\mathcal{S}^{\#} = ([0, 2] \times \mathbb{R}) \cup (\mathbb{R} \times [0, 1])$ and $\mathcal{S}^{\#} \cap \mathcal{W} = \mathcal{W}$. Since $\mathcal{S} \Delta \mathcal{W} = [1, 2] \times \{\frac{1}{2}\}$ is a line segment that has measure zero under the plane Lebesgue measure $\xi \otimes \zeta = \lambda^2$, we have $\mathcal{S} \stackrel{\xi \otimes \zeta}{=} \mathcal{W}$ so \mathcal{S} is a $\xi \otimes \zeta$ -complete component of \mathcal{W} . But $\xi(\mathcal{S}^{\mathbb{X}} \setminus \mathcal{W}^{\mathbb{X}}) = \lambda([0, 2] \setminus [0, 1]) = \lambda(1, 2] = 1$ is not zero, so $\mathcal{S}^{\mathbb{X}} \subseteq^{\xi} \mathcal{W}^{\mathbb{X}}$ does not hold.

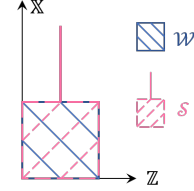


Figure 8: Example B.13 showing that a $\xi \otimes \zeta$ -complete component of \mathcal{W} may not have its projection be an a.s. subset of that of \mathcal{W} .

Lemma B.14. Let \mathcal{S} be a $\xi \otimes \zeta$ -complete component of \mathcal{W} , and $\tilde{\mathcal{S}}$ be a measurable set such that $\tilde{\mathcal{S}} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}$, $\tilde{\mathcal{S}}^{\mathbb{X}} \stackrel{\xi}{=} \mathcal{S}^{\mathbb{X}}$ and $\tilde{\mathcal{S}}^{\mathbb{Z}} \stackrel{\zeta}{=} \mathcal{S}^{\mathbb{Z}}$. Then this $\tilde{\mathcal{S}}$ is also a $\xi \otimes \zeta$ -complete component of \mathcal{W} .

Proof. By Lemma B.11, we know that $\tilde{\mathcal{S}}^{\mathbb{X}} \times \mathbb{Z} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}^{\mathbb{X}} \times \mathbb{Z}$, $\mathbb{X} \times \tilde{\mathcal{S}}^{\mathbb{Z}} \stackrel{\xi \otimes \zeta}{=} \mathbb{X} \times \mathcal{S}^{\mathbb{Z}}$. Repeatedly applying Lemma B.5, we have $\tilde{\mathcal{S}}^{\#} := \tilde{\mathcal{S}}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \tilde{\mathcal{S}}^{\mathbb{Z}} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^{\mathbb{Z}} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^{\mathbb{Z}} =: \mathcal{S}^{\#}$, and $\tilde{\mathcal{S}}^{\#} \cap \mathcal{W} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}^{\#} \cap \mathcal{W}$, which $\stackrel{\xi \otimes \zeta}{=} \mathcal{S} \stackrel{\xi \otimes \zeta}{=} \tilde{\mathcal{S}}$. From the transitivity (Lemma B.4), we have $\tilde{\mathcal{S}}^{\#} \cap \mathcal{W} \stackrel{\xi \otimes \zeta}{=} \tilde{\mathcal{S}}$. \square

Example B.15. Note that only the $\tilde{\mathcal{S}} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}$ condition is not sufficient. Fig. 9 shows such an example, where $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$ are the one dimensional Euclidean spaces with line Borel sigma-field and line Lebesgue measure, $(\mathbb{R}, \mathcal{R}, \lambda)$, and $\mathcal{W} := [0, 1]^2 \cup [1, 2]^2$, $\mathcal{S} := [0, 1]^2$, and $\tilde{\mathcal{S}} := [0, 1]^2 \cup ([1, 2] \times \{\frac{1}{2}\})$. We have $\mathcal{S}^{\#} = ([0, 1] \times \mathbb{R}) \cup (\mathbb{R} \times [0, 1])$ so $\mathcal{S}^{\#} \cap \mathcal{W} = \mathcal{S}$, justifying that \mathcal{S} is a $\xi \otimes \zeta$ -complete component of \mathcal{W} . On the other hand, since $\mathcal{S} \Delta \tilde{\mathcal{S}} = [1, 2] \times \{\frac{1}{2}\}$ is a line segment that has measure zero under the plane Lebesgue measure $\xi \otimes \zeta = \lambda^2$, we have $\tilde{\mathcal{S}} \stackrel{\xi \otimes \zeta}{=} \mathcal{S}$. But $\tilde{\mathcal{S}}^{\mathbb{X}} = [0, 2]$ so $\tilde{\mathcal{S}}^{\#} = ([0, 2] \times \mathbb{R}) \cup (\mathbb{R} \times [0, 1])$, which leads to $\tilde{\mathcal{S}}^{\#} \cap \mathcal{W} = \mathcal{W}$. Since $\tilde{\mathcal{S}} \Delta \mathcal{W} = ([1, 2] \times \{\frac{1}{2}\}) \cup ([1, 2] \times [1, 2])$ has a nonzero measure under λ^2 (it equals to 1), we know that $\tilde{\mathcal{S}}$ is not a $\xi \otimes \zeta$ -complete component of \mathcal{W} .

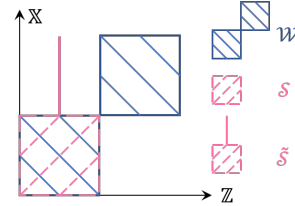


Figure 9: Example B.15 showing that in Lem. B.14, only being $\xi \otimes \zeta$ -a.s. the same as a $\xi \otimes \zeta$ -complete component $\tilde{\mathcal{S}}$ of \mathcal{W} is not sufficient for $\tilde{\mathcal{S}}$ to be also a $\xi \otimes \zeta$ -complete component of \mathcal{W} .

Lemma B.16. Let \mathcal{S} be a $\xi \otimes \zeta$ -complete component of \mathcal{W} , and f be an either nonnegative or $\xi \otimes \zeta$ -integrable function on $\mathbb{X} \times \mathbb{Z}$. Then for any measurable sets $\mathcal{Z} \subseteq \mathcal{S}^{\mathbb{Z}}$ and $\mathcal{X} \subseteq \mathcal{S}^{\mathbb{X}}$, we have:

$$\begin{aligned} \int_{\mathcal{Z}} \int_{\mathcal{W}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz) &= \int_{\mathcal{Z}} \int_{\mathcal{S}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz), \\ \int_{\mathcal{X}} \int_{\mathcal{W}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx) &= \int_{\mathcal{X}} \int_{\mathcal{S}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx). \end{aligned}$$

Particularly, $\int_{\mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{W}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz) = \int_{\mathcal{S}^{\mathbb{X}}} \int_{\mathcal{W}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx) = \int_{\mathcal{S}} f(x, z) (\xi \otimes \zeta)(dx dz)$.

Proof. Since \mathcal{S} is a $\xi \otimes \zeta$ -complete component of \mathcal{W} , Eq. (13) holds. By Lemma B.10, we know that for ζ -a.e. z on \mathbb{Z} , $\xi((\mathcal{S}^{\#} \cap \mathcal{W}) \Delta \mathcal{S})_z = \xi((\mathcal{S}_{\mathcal{Z}}^{\#} \cap \mathcal{W}_{\mathcal{Z}}) \Delta \mathcal{S}_{\mathcal{Z}}) = 0$. Noting that $\mathcal{S}_{\mathcal{Z}}^{\#} = \mathbb{X}$ for any $z \in \mathcal{S}^{\mathbb{Z}}$, this subsequently means that $\xi(\mathcal{W}_{\mathcal{Z}} \Delta \mathcal{S}_{\mathcal{Z}}) = 0$ for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$. By the additivity of integrals over a countable partition [9, Thm. 16.9] and that the integral over a measure-zero set is zero [9, p.226], we have $\int_{\mathcal{W}_{\mathcal{Z}}} f(x, z) \xi(dx) = \int_{\mathcal{S}_{\mathcal{Z}}} f(x, z) \xi(dx)$ for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$. Since a.e.-equal functions have the same integral [9, Thm. 15.2(v)], we have for any measurable $\mathcal{Z} \subseteq \mathcal{S}^{\mathbb{Z}}$, $\int_{\mathcal{Z}} \int_{\mathcal{W}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz) = \int_{\mathcal{Z}} \int_{\mathcal{S}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz)$. Similarly, for any measurable $\mathcal{X} \subseteq \mathcal{S}^{\mathbb{X}}$, $\int_{\mathcal{X}} \int_{\mathcal{W}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx) = \int_{\mathcal{X}} \int_{\mathcal{S}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx)$.

For $\mathcal{Z} = \mathcal{S}^{\mathbb{Z}}$, we have $\int_{\mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{W}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz) = \int_{\mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{S}_{\mathcal{Z}}} f(x, z) \xi(dx) \zeta(dz)$, which is $\int_{\mathcal{S}} f(x, z) (\xi \otimes \zeta)(dx dz)$ by the generalized form Eq. (9) of Fubini's theorem. Similarly, $\int_{\mathcal{S}^{\mathbb{X}}} \int_{\mathcal{W}_{\mathcal{X}}} f(x, z) \zeta(dz) \xi(dx) = \int_{\mathcal{S}} f(x, z) (\xi \otimes \zeta)(dx dz)$. \square

C Proofs

Recall that $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z}, \xi \otimes \zeta)$ is the product measure space by the two individual ones $(\mathbb{X}, \mathcal{X}, \xi)$ and $(\mathbb{Z}, \mathcal{Z}, \zeta)$, where ξ and ζ are sigma-finite.

C.1 The Joint-Conditional Absolute Continuity Lemma

Although this lemma is not formally presented in the main text, we highlight it here since it answers an important question and the answer is not straightforward.

The lemma reveals the relation between the absolute continuity of a joint π and that of its conditionals $\pi(\cdot|z)$, $\pi(\cdot|x)$. Roughly, the former guarantees the latter on the supports of the marginals, and the reverse also holds, allowing one to safely use density function formulae for deduction. But given two conditionals, one does not have the knowledge on the marginals *a priori*. For a more useful sufficient condition, one may consider the absolute continuity of the conditionals for ζ -a.e. z and ξ -a.e. x . Unfortunately this is not sufficient, and an example (C.2) is given after the proof. The lemma shows it is sufficient if the absolute continuity of one of the conditionals, say $\pi(\cdot|z)$, holds for *any* $z \in \mathbb{Z}$. The condition in the compatibility criterion Thm. 2.3 is also inspired from this lemma.

Lemma C.1 (joint-conditional absolute continuity). *(i) For a joint distribution π on $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$, it is absolutely continuous $\pi \ll \xi \otimes \zeta$ if and only if $\pi(\cdot|z) \ll \xi$ for $\pi^{\mathbb{Z}}$ -a.e. z and $\pi(\cdot|x) \ll \zeta$ for $\pi^{\mathbb{X}}$ -a.e. x . (ii) As a sufficient condition, $\pi \ll \xi \otimes \zeta$ if $\pi(\cdot|z) \ll \xi$ for ζ -a.e. z and $\pi(\cdot|x) \ll \zeta$ for any $x \in \mathbb{X}$ (or for any $z \in \mathbb{Z}$ and ξ -a.e. x).*

For conclusion (i):

Proof. “Only if”: Consider any $\mathcal{X} \in \mathcal{X}$ such that $\xi(\mathcal{X}) = 0$. From the definition of conditional distribution Eq. (11), we have $\pi^{\mathbb{X}}(\mathcal{X}) = \pi(\mathcal{X} \times \mathbb{Z}) = \int_{\mathbb{Z}} \pi(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = 0$, so $\pi(\mathcal{X}|z) = 0$ for $\pi^{\mathbb{Z}}$ -a.e. z since $\pi(\mathcal{X}|z)$ is nonnegative [9, Thm. 15.2(ii)]. This means that $\pi(\cdot|z) \ll \xi$ for $\pi^{\mathbb{Z}}$ -a.e. z . The same arguments apply symmetrically to $\pi(\cdot|x)$.

Note that since $\pi(\cdot|z)$ is defined as the R-N derivative, it is allowed to take any nonnegative value on a $\pi^{\mathbb{Z}}$ -measure-zero set. So we cannot guarantee its behavior for *any* $z \in \mathbb{Z}$.

“If”: Consider any $\mathcal{Z} \in \mathcal{Z}$ such that $\zeta(\mathcal{Z}) = 0$. Since $\pi(\cdot|x) \ll \zeta$ for $\pi^{\mathbb{X}}$ -a.e. x , we have $\pi(\mathcal{Z}|x) = 0$ for $\pi^{\mathbb{X}}$ -a.e. x . So from Eq. (11) we have $\pi^{\mathbb{Z}}(\mathcal{Z}) = \pi(\mathbb{X} \times \mathcal{Z}) = \int_{\mathbb{X}} \pi(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx) = 0$ [9, Thm. 15.2(i)]. This indicates that $\pi^{\mathbb{Z}} \ll \zeta$.

Now consider any $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$ such that $(\xi \otimes \zeta)(\mathcal{W}) = 0$. By the definition of product measure Eq. (6) [9, Thm. 18.2], we have $(\xi \otimes \zeta)(\mathcal{W}) = \int_{\mathbb{Z}} \xi(\mathcal{W}_z) \zeta(dz) = 0$, so $\xi(\mathcal{W}_z) = 0$ for ζ -a.e. z since $\xi(\mathcal{W}_z)$ is nonnegative [9, Thm. 15.2(ii)]. Due to that $\pi^{\mathbb{Z}} \ll \zeta$, this means that $\xi(\mathcal{W}_z) = 0$ for $\pi^{\mathbb{Z}}$ -a.e. z . Since $\pi(\cdot|z) \ll \xi$ for $\pi^{\mathbb{Z}}$ -a.e. z , this in turn means that $\pi(\mathcal{W}_z|z) = 0$ for $\pi^{\mathbb{Z}}$ -a.e. z . Subsequently, we have $\int_{\mathbb{Z}} \pi(\mathcal{W}_z|z) \pi^{\mathbb{Z}}(dz) = 0$ [9, Thm. 15.2(i)], which is $\pi(\mathcal{W}) = 0$ by Eq. (12). So we get $\pi \ll \xi \otimes \zeta$. \square

For conclusion (ii):

Proof. Consider any $\mathcal{Z} \in \mathcal{Z}$ such that $\zeta(\mathcal{Z}) = 0$. Since $\pi(\cdot|x) \ll \zeta$ for any $x \in \mathbb{X}$, we know that $\pi(\mathcal{Z}|x) = 0$ for any $x \in \mathbb{X}$. So from Eq. (11) we have $\pi^{\mathbb{Z}}(\mathcal{Z}) = \pi(\mathbb{X} \times \mathcal{Z}) = \int_{\mathbb{X}} \pi(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx) = 0$. This indicates that $\pi^{\mathbb{Z}} \ll \zeta$.

Now consider any $\mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}$ such that $(\xi \otimes \zeta)(\mathcal{W}) = 0$. By the definition of product measure Eq. (6) [9, Thm. 18.2], we have $(\xi \otimes \zeta)(\mathcal{W}) = \int_{\mathbb{Z}} \xi(\mathcal{W}_z) \zeta(dz) = 0$, so $\xi(\mathcal{W}_z) = 0$ for ζ -a.e. z since $\xi(\mathcal{W}_z)$ is nonnegative [9, Thm. 15.2(ii)]. Due to that $\pi(\cdot|z) \ll \xi$ for ζ -a.e. z , this means that $\pi(\mathcal{W}_z|z) = 0$ for ζ -a.e. z . Since $\pi^{\mathbb{Z}} \ll \zeta$, this in turn means that $\pi(\mathcal{W}_z|z) = 0$ for $\pi^{\mathbb{Z}}$ -a.e. z . Subsequently, we have $\int_{\mathbb{Z}} \pi(\mathcal{W}_z|z) \pi^{\mathbb{Z}}(dz) = 0$ [9, Thm. 15.2(i)], which is $\pi(\mathcal{W}) = 0$ by Eq. (12). So we get $\pi \ll \xi \otimes \zeta$. The same arguments apply symmetrically when $\pi(\cdot|z) \ll \xi$ for *any* $z \in \mathbb{Z}$ and $\pi(\cdot|x) \ll \zeta$ for ξ -a.e. x . \square

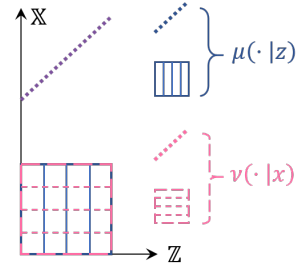


Figure 10: Illustration of the conditionals in Eq. (14) in Example C.2. Both conditionals are absolutely continuous for ζ -a.e. z or ξ -a.e. x , but they allow a compatible joint that is not absolutely continuous w.r.t $\xi \otimes \zeta$.

Example C.2. To see why it is not sufficient if the two conditionals are absolutely continuous only for ζ -a.e. z and ξ -a.e. x , we show an example below.

Consider the one-dimensional Euclidean space $\mathbb{X} = \mathbb{Z} = \mathbb{R}$ with line Borel sigma-field $\mathcal{X} = \mathcal{Z} = \mathcal{R}$ and line Lebesgue measure $\xi = \zeta = \lambda$. Let

$$\mu(\cdot|z) := \begin{cases} \delta_{z+2}, & z \in \mathbb{Q}[0, 1], \\ \text{Unif}[0, 1], & z \in \bar{\mathbb{Q}}[0, 1], \\ 0, & \text{otherwise}, \end{cases} \quad \nu(\cdot|x) := \begin{cases} \text{Unif}[0, 1], & x \in [0, 1], \\ \delta_{x-2}, & x \in \mathbb{Q}[2, 3], \\ 0, & \text{otherwise}, \end{cases} \quad (14)$$

where $\mathbb{Q}[0, 1] := [0, 1] \cap \mathbb{Q}$ and $\bar{\mathbb{Q}} := [0, 1] \setminus \mathbb{Q}$ are the rational and irrational numbers on $[0, 1]$. The conditionals are illustrated in Fig. 10. Since $\lambda(\mathbb{Q}) = 0$, the two conditionals are absolutely continuous for ζ -a.e. z and ξ -a.e. x . Consider the joint distribution on $\mathbb{X} \times \mathbb{Z} = \mathbb{R}^2$:

$$\pi := \frac{1}{2} \text{Unif}([0, 1] \times [0, 1]) + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \delta_{(z+2, z)},$$

where ϱ is a distribution on the rationals $\mathbb{Q}[0, 1]$ in $[0, 1]$ with the sigma-field of all the subsets of $\mathbb{Q}[0, 1]$. Such a distribution exists, for example, $\varrho(z) = 1/2^{n(z)}$ where $n : \mathbb{Q}[0, 1] \rightarrow \mathbb{N}^*$ bijective is a numbering function of the countable set $\mathbb{Q}[0, 1]$. In this way, each rational number $z \in \mathbb{Q}[0, 1]$ has a positive probability, meanwhile we have $\varrho(\mathbb{Q}[0, 1]) = \sum_{n=1}^{\infty} 1/2^n = 1$. Since $\pi(\{(z+2, z) \mid z \in \mathbb{Q}[0, 1]\}) = \frac{1}{2}$ but $\lambda^2(\{(z+2, z) \mid z \in \mathbb{Q}[0, 1]\}) = 0$ under the square Lebesgue measure λ^2 , π is not absolutely continuous.

To verify compatibility, note that μ and ν here satisfy the corresponding measurability and integrability. To verify Eq. (11) reduced from Eq. (10) for defining a conditional, we first derive the marginals:

$$\pi^{\mathbb{X}} = \frac{1}{2} \text{Unif}[0, 1] + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \delta_{z+2}, \quad \pi^{\mathbb{Z}} = \frac{1}{2} \text{Unif}[0, 1] + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \delta_z.$$

For any $\mathcal{X} \in \mathcal{X}$ and $\mathcal{Z} \in \mathcal{Z}$, we have:

$$\begin{aligned} \pi(\mathcal{X} \times \mathcal{Z}) &= \frac{1}{2} \text{Unif}[0, 1](\mathcal{X}) \text{Unif}[0, 1](\mathcal{Z}) + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[(z+2, z) \in \mathcal{X} \times \mathcal{Z}] \\ &= \frac{1}{2} \lambda(\mathcal{X}[0, 1]) \lambda(\mathcal{Z}[0, 1]) + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z \in (\mathcal{X} - 2) \cap \mathcal{Z}], \end{aligned}$$

where $\mathcal{X}[0, 1] := [0, 1] \cap \mathcal{X}$ and $\mathcal{Z}[0, 1] := [0, 1] \cap \mathcal{Z}$. To verify the conditional distribution $\mu(\mathcal{X}|z)$, we have:

$$\begin{aligned} \int_{\mathcal{Z}} \mu(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) &= \int_{\mathcal{Z} \cap \mathbb{Q}[0, 1]} \delta_{z+2}(\mathcal{X}) \pi^{\mathbb{Z}}(dz) + \int_{\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]} \text{Unif}[0, 1](\mathcal{X}) \pi^{\mathbb{Z}}(dz) \\ &= \int_{\mathcal{Z} \cap \mathbb{Q}[0, 1]} \mathbb{I}[z+2 \in \mathcal{X}] \pi^{\mathbb{Z}}(dz) + \int_{\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]} \lambda(\mathcal{X}[0, 1]) \pi^{\mathbb{Z}}(dz) \\ &= \pi^{\mathbb{Z}}((\mathcal{X} - 2) \cap \mathcal{Z} \cap \mathbb{Q}[0, 1]) + \lambda(\mathcal{X}[0, 1]) \pi^{\mathbb{Z}}(\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]) \end{aligned}$$

(Since a countable set has measure zero under $\text{Unif}[0, 1]$, i.e. $\lambda(\cdot \cap [0, 1])$.)

$$\begin{aligned} &= \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z \in (\mathcal{X} - 2) \cap \mathcal{Z} \cap \mathbb{Q}[0, 1]] \\ &\quad + \frac{1}{2} \lambda(\mathcal{X}[0, 1]) \left(\text{Unif}[0, 1](\mathcal{Z}[0, 1]) + \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z \in \mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]] \right) \\ &= \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z \in (\mathcal{X} - 2) \cap \mathcal{Z}] + \frac{1}{2} \lambda(\mathcal{X}[0, 1]) \lambda(\mathcal{Z}[0, 1]) = \pi(\mathcal{X} \times \mathcal{Z}). \end{aligned}$$

For the conditional distribution $\nu(\mathcal{Z}|x)$, we similarly have:

$$\begin{aligned} \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx) &= \int_{\mathcal{X}[0, 1]} \text{Unif}[0, 1](\mathcal{Z}) \pi^{\mathbb{X}}(dx) + \int_{\mathcal{X} \cap \mathbb{Q}[2, 3]} \delta_{x-2}(\mathcal{Z}) \pi^{\mathbb{X}}(dx) \\ &= \lambda(\mathcal{Z}[0, 1]) \pi^{\mathbb{X}}(\mathcal{X}[0, 1]) + \pi^{\mathbb{X}}(\mathcal{X} \cap \mathbb{Q}[2, 3] \cap (\mathcal{Z} + 2)) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \lambda(\mathcal{Z}[0, 1]) \lambda(\mathcal{X}[0, 1]) + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z + 2 \in \mathcal{X} \cap \mathbb{Q}[2, 3] \cap (\mathcal{Z} + 2)] \\
&= \frac{1}{2} \lambda(\mathcal{Z}[0, 1]) \lambda(\mathcal{X}[0, 1]) + \frac{1}{2} \sum_{z \in \mathbb{Q}[0, 1]} \varrho(z) \mathbb{I}[z \in (\mathcal{X} - 2) \cap \mathcal{Z}] = \pi(\mathcal{X} \times \mathcal{Z}).
\end{aligned}$$

So the two conditionals $\mu(\cdot|z)$ and $\nu(\cdot|x)$ are compatible and π is their joint distribution. This example illustrates that the absolute continuity of $\pi(\cdot|z)$ w.r.t ξ for ζ -a.e. z and that of $\pi(\cdot|x)$ w.r.t ζ for ξ -a.e. x , does not indicate the absolute continuity of π w.r.t $\xi \otimes \zeta$.

This example does not contradict result (i) of the Lemma. For any $z_0 \in \mathbb{Q}[0, 1]$, we have that $\mu(\cdot|z_0) = \delta_{z_0+2}$ is not absolutely continuous w.r.t $\xi = \lambda$. But $\pi^{\mathbb{Z}}(\{z_0\}) = \frac{1}{2} \varrho(z_0) > 0$. So it is not that $\mu(\cdot|z) \ll \xi$ for $\pi^{\mathbb{Z}}$ -a.e. z , which aligns with that π is not absolutely continuous w.r.t $\xi \otimes \zeta = \lambda^2$.

This example also shows that the absolute continuity of the compatible joint may depend on the joint itself, apart from the two conditionals. Consider another joint on $(\mathbb{R}^2, \mathcal{R}^2, \lambda^2)$:

$$\tilde{\pi} := \text{Unif}([0, 1] \times [0, 1]).$$

It is easy to see that $\tilde{\pi}^{\mathbb{X}}(\mathcal{X}) = \text{Unif}[0, 1](\mathcal{X}) = \lambda(\mathcal{X}[0, 1])$ and $\tilde{\pi}^{\mathbb{Z}}(\mathcal{Z}) = \lambda(\mathcal{Z}[0, 1])$. For any $\mathcal{X} \in \mathcal{X}$ and $\mathcal{Z} \in \mathcal{Z}$, we have $\tilde{\pi}(\mathcal{X} \times \mathcal{Z}) = \lambda(\mathcal{X}[0, 1]) \lambda(\mathcal{Z}[0, 1])$. To verify Eq. (11) for defining a conditional, we have:

$$\begin{aligned}
\int_{\mathcal{Z}} \mu(\mathcal{X}|z) \tilde{\pi}^{\mathbb{Z}}(dz) &= \int_{\mathcal{Z} \cap \mathbb{Q}[0, 1]} \delta_{z+2}(\mathcal{X}) \tilde{\pi}^{\mathbb{Z}}(dz) + \int_{\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]} \text{Unif}[0, 1](\mathcal{X}) \tilde{\pi}^{\mathbb{Z}}(dz) \\
&= \int_{\mathcal{Z} \cap \mathbb{Q}[0, 1]} \mathbb{I}[z + 2 \in \mathcal{X}] \tilde{\pi}^{\mathbb{Z}}(dz) + \int_{\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]} \lambda(\mathcal{X}[0, 1]) \tilde{\pi}^{\mathbb{Z}}(dz) \\
&= \lambda((\mathcal{X} - 2) \cap \mathcal{Z} \cap \mathbb{Q}[0, 1]) + \lambda(\mathcal{X}[0, 1]) \lambda(\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1])
\end{aligned}$$

(Since $\lambda((\mathcal{X} - 2) \cap \mathcal{Z} \cap \mathbb{Q}[0, 1]) \leq \lambda(\mathbb{Q}) = 0$ and $\lambda(\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]) = \lambda(\mathcal{Z} \cap \bar{\mathbb{Q}}[0, 1]) + \lambda(\mathbb{Q}[0, 1]) = \lambda(\mathcal{Z}[0, 1])$.)

$$= \lambda(\mathcal{X}[0, 1]) \lambda(\mathcal{Z}[0, 1]) = \tilde{\pi}(\mathcal{X} \times \mathcal{Z}).$$

For the conditional distribution $\nu(\mathcal{Z}|x)$, we similarly have:

$$\begin{aligned}
\int_{\mathcal{X}} \nu(\mathcal{Z}|x) \tilde{\pi}^{\mathbb{X}}(dx) &= \int_{\mathcal{X}[0, 1]} \text{Unif}[0, 1](\mathcal{Z}) \tilde{\pi}^{\mathbb{X}}(dx) + \int_{\mathcal{X} \cap \mathbb{Q}[2, 3]} \delta_{x-2}(\mathcal{Z}) \tilde{\pi}^{\mathbb{X}}(dx) \\
&= \lambda(\mathcal{Z}[0, 1]) \tilde{\pi}^{\mathbb{X}}(\mathcal{X}[0, 1]) + \tilde{\pi}^{\mathbb{X}}(\mathcal{X} \cap \mathbb{Q}[2, 3] \cap (\mathcal{Z} + 2))
\end{aligned}$$

(Since $\tilde{\pi}^{\mathbb{X}}(\mathcal{X} \cap \mathbb{Q}[2, 3] \cap (\mathcal{Z} + 2)) = \lambda(\mathcal{X} \cap \mathbb{Q}[2, 3] \cap (\mathcal{Z} + 2) \cap [0, 1]) \leq \lambda(\mathbb{Q}) = 0$.)

$$= \lambda(\mathcal{Z}[0, 1]) \lambda(\mathcal{X}[0, 1]) = \tilde{\pi}(\mathcal{X} \times \mathcal{Z}).$$

So $\tilde{\pi}$ is also a compatible joint of $\mu(\cdot|z)$ and $\nu(\cdot|x)$. In this case, the violation set for $\mu(\cdot|z) \ll \xi$, i.e. $\mathbb{Q}[0, 1]$, has measure zero under $\tilde{\pi}^{\mathbb{Z}}$, and the violation set for $\nu(\cdot|x) \ll \zeta$, i.e. $\mathbb{Q}[2, 3]$, has measure zero under $\tilde{\pi}^{\mathbb{X}}$. So result (i) of the Lemma asserts that $\tilde{\pi} \ll \xi \otimes \zeta$, which aligns with the example. This example shows that although both π and $\tilde{\pi}$ are the compatible joint of the same conditionals, they have different absolute continuity. So the condition in result (i) of the Lemma requires the knowledge of the marginals $\pi^{\mathbb{Z}}$ and $\pi^{\mathbb{X}}$.

C.2 Proof of Theorem 2.3

Example C.3. Before presenting the proof, we give an example showing that only being the intersection of $\mathcal{W}_{p,q}$ and $\mathcal{W}_{q,p}$ is not sufficient to make a valid support of a compatible joint. The example is illustrated in Fig. 11. The conditionals are uniform on the respective depicted slices, so conditions (iv) and (v) in the theorem (2.3) are satisfied. The sets $\mathcal{W}_{p,q}$ and $\mathcal{W}_{q,p}$ are depicted in the figure, and their intersection $\mathcal{W}_{p,q} \cap \mathcal{W}_{q,p}$ is the right half. Although on $\mathcal{W}_{p,q} \cap \mathcal{W}_{q,p}$, the conditionals do not render support conflict, the conditional $q(z|x)$ is unnormalized for a given x from the bottom half: it integrates to $1/2$ on $(\mathcal{W}_{p,q} \cap \mathcal{W}_{q,p})_x$. This means

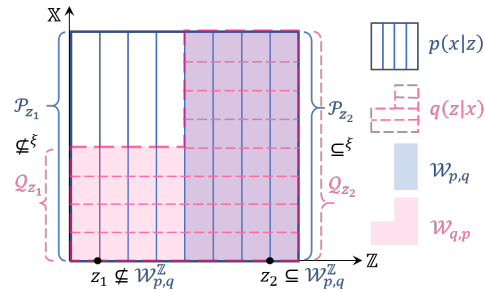


Figure 11: Illustration of Example C.3. The conditionals are uniform on the respective depicted slices. On $\mathcal{W}_{p,q} \cap \mathcal{W}_{q,p}$, the conditional $q(z|x)$ is not normalized.

that any compatible joint π would also have its conditional $\pi(\cdot|x)$ unnormalized for such x , which is impossible.

One may consider trying $\mathcal{W}_{q,p}$ as the joint support. This renders a support conflict similar to the example in the main text: a joint on $\mathcal{W}_{q,p}$ is required by $p(x|z)$ to also cover the top-left quadrant since z values in the left half are covered, but this contradicts with the absence of mass by $q(z|x)$. In fact, in this example, there is no $\xi \otimes \zeta$ -complete component of both $\mathcal{W}_{p,q}$ and $\mathcal{W}_{q,p}$, so the two conditionals are not compatible.

Proof. Let $\mu(\mathcal{X}|z)$ and $\nu(\mathcal{Z}|x)$ be the two everywhere absolutely continuous conditional distributions of whom $p(x|z)$ and $q(z|x)$ are the density functions. We begin with some useful conclusions.

(1) By construction, for any $(x, z) \in \mathcal{W}_{p,q}$, $p(x|z) > 0$. For any $z \in \mathcal{W}_{p,q}^{\mathbb{Z}}$, we have $\xi\{x \in (\mathcal{W}_{p,q})_z \mid q(z|x) = 0\} = \xi\{x \in \mathcal{P}_z \mid x \notin \mathcal{Q}_z\} = \xi(\mathcal{P}_z \setminus \mathcal{Q}_z) = 0$, which means that $q(z|x) > 0$, ξ -a.e. on $(\mathcal{W}_{p,q})_z$. By Lemma B.10, we have that $q(z|x) > 0$, $\xi \otimes \zeta$ -a.e. on $\mathcal{W}_{p,q}$. Symmetrically, $q(z|x) > 0$ on $\mathcal{W}_{q,p}$, and $p(x|z) > 0$, $\xi \otimes \zeta$ -a.e. on $\mathcal{W}_{q,p}$. Particularly, the ratio $\frac{p(x|z)}{q(z|x)}$ is well-defined and is positive and finite, both $\xi \otimes \zeta$ -a.e. on $\mathcal{W}_{p,q}$ and $\xi \otimes \zeta$ -a.e. on $\mathcal{W}_{q,p}$. The conclusions also hold ($\xi \otimes \zeta$ -a.e.) on any ($\xi \otimes \zeta$ -a.s.) subset of $\mathcal{W}_{p,q}$ or $\mathcal{W}_{q,p}$.

“Only if” (necessity):

Let π be a compatible joint distribution of such conditional distributions $\mu(\cdot|z)$ and $\nu(\cdot|x)$.

(2) Since “for any” indicates “a.e.” under any measure, the conditions in Lemma C.1 are satisfied, so we have $\pi \ll \xi \otimes \zeta$. By Lemma B.9, we also have $\pi^{\mathbb{X}} \ll \xi$ and $\pi^{\mathbb{Z}} \ll \zeta$. So there exist density functions (R-N derivatives; 9, Thm. 32.2) $u(x)$ and $v(z)$ such that for any $\mathcal{X} \in \mathcal{X}$ and $\mathcal{Z} \in \mathcal{Z}$, $\pi^{\mathbb{X}}(\mathcal{X}) = \int_{\mathcal{X}} u(x) \xi(dx)$ and $\pi^{\mathbb{Z}}(\mathcal{Z}) = \int_{\mathcal{Z}} v(z) \zeta(dz)$. This $u(x)$ is obviously ξ -integrable on any measurable subset of \mathbb{X} , since the integral is no larger (since u is nonnegative) than $\int_{\mathbb{X}} u(x) \xi(dx) = 1$ which is finite.

(3) By the definition of conditional distribution Eq. (11), for any $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, we have $\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mu(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{Z}} \mu(\mathcal{X}|z) v(z) \zeta(dz) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p(x|z) \xi(dx) v(z) \zeta(dz) = \int_{\mathcal{X} \times \mathcal{Z}} p(x|z) v(z) (\xi \otimes \zeta)(dxdz)$, where in the last equality, we have applied Fubini’s theorem Eq. (8) [9, Thm. 18.3]. Similarly, $\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{X} \times \mathcal{Z}} q(z|x) u(x) (\xi \otimes \zeta)(dxdz)$. Noting that $\mathcal{X} \times \mathcal{Z}$ is the pi-system that generates $\mathcal{X} \otimes \mathcal{Z}$, this indicates that $p(x|z)v(z) = q(z|x)u(x)$, $\xi \otimes \zeta$ -a.e. on $\mathbb{X} \times \mathbb{Z}$ [9, Thm. 16.10(iii)]¹¹. In other words, both $p(x|z)v(z)$ and $q(z|x)u(x)$ are density functions of π .

(3.1) Subsequently, by leveraging Lemma B.10, we have for ζ -a.e. z on \mathbb{Z} , $p(x|z)v(z) = q(z|x)u(x)$, ξ -a.e. on \mathbb{X} .

(4) Let $\mathcal{U} := \{x \mid u(x) > 0\}$ and $\mathcal{V} := \{z \mid v(z) > 0\}$, and define:

$$\mathcal{S} := (\mathcal{U} \times \mathcal{V}) \cap \mathcal{W}_{p,q}, \quad \tilde{\mathcal{S}} := (\mathcal{U} \times \mathcal{V}) \cap \mathcal{W}_{q,p}, \quad (15)$$

Since u and v are integrable thus measurable and $\mathbb{R}^{>0}$ is Lebesgue-measurable, we know that $\mathcal{U} \in \mathcal{X}$ and $\mathcal{V} \in \mathcal{Z}$ are also measurable. So \mathcal{S} and $\tilde{\mathcal{S}}$ are measurable.

(4.1) We can verify that \mathcal{S} is a $\xi \otimes \zeta$ -complete component of $\mathcal{W}_{p,q}$. Since $\mathcal{S} \subseteq \mathcal{W}_{p,q}$, we only need to verify that:

$$\begin{aligned} & (\xi \otimes \zeta)([(\mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{S}^{\mathbb{Z}}) \cap \mathcal{W}_{p,q}] \setminus \mathcal{S}) \\ & \text{(Since clearly } \mathcal{S}^{\mathbb{X}} \subseteq \mathcal{U}, \mathcal{S}^{\mathbb{Z}} \subseteq \mathcal{V}, \text{ and measures are monotone,)} \\ & \leq (\xi \otimes \zeta)([(\mathcal{U} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{V}) \cap \mathcal{W}_{p,q}] \setminus \mathcal{S}) \quad (\text{Since } (\mathcal{A} \cap \mathcal{C}) \setminus (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \setminus \mathcal{B}) \cap \mathcal{C},) \\ & = (\xi \otimes \zeta)([(\mathcal{U} \times \mathbb{Z} \cup \mathbb{X} \times \mathcal{V}) \setminus (\mathcal{U} \times \mathcal{V})] \cap \mathcal{W}_{p,q}) \quad (\text{Since } (\mathcal{A} \cup \mathcal{B}) \setminus \mathcal{C} = (\mathcal{A} \setminus \mathcal{C}) \cup (\mathcal{B} \setminus \mathcal{C}),) \\ & = (\xi \otimes \zeta)([(\mathcal{U} \times \mathbb{Z}) \setminus (\mathcal{U} \times \mathcal{V})] \cup [(\mathbb{X} \times \mathcal{V}) \setminus (\mathcal{U} \times \mathcal{V})] \cap \mathcal{W}_{p,q}) \\ & = (\xi \otimes \zeta)([\{(x, z) \mid u(x) > 0, v(z) = 0\} \cup \mathcal{W}_{p,q}] \cup [\{(x, z) \mid u(x) = 0, v(z) > 0\} \cup \mathcal{W}_{p,q}]) \end{aligned}$$

¹¹The requirement that $\mathbb{X} \times \mathbb{Z}$ is a finite or countable union of $\mathcal{X} \times \mathcal{Z}$ -sets is satisfied, since from the sigma-finiteness of ξ and ζ , \mathbb{X} and \mathbb{Z} are a finite or countable union of (ξ -finite) \mathcal{X} -sets and (ζ -finite) \mathcal{Z} -sets, respectively.

(By Conclusion (3) and adjusting the set by a set of $\xi \otimes \zeta$ -measure-zero according to Lemma B.1.)

$$= (\xi \otimes \zeta)(\{ (x, z) \mid u(x) > 0, v(z) = 0, q(z|x) = 0 \} \cup \mathcal{W}_{p,q})$$

$$\cup \{ (x, z) \mid u(x) = 0, v(z) > 0, p(x|z) = 0 \} \cup \mathcal{W}_{p,q})$$

$$\leq (\xi \otimes \zeta)(\{ (x, z) \mid u(x) > 0, v(z) = 0, q(z|x) = 0 \} \cup \mathcal{W}_{p,q})$$

$$+ (\xi \otimes \zeta)(\{ (x, z) \mid u(x) = 0, v(z) > 0, p(x|z) = 0 \} \cup \mathcal{W}_{p,q}) \quad (\text{By Conclusion (1),})$$

$$= 0. \text{ Symmetrically, we can also verify that } \tilde{\mathcal{S}} \text{ is a } \xi \otimes \zeta\text{-complete component of } \mathcal{W}_{q,p}.$$

(4.2) We can also show that $\pi(\mathcal{S}) = \pi(\tilde{\mathcal{S}}) = 1$. From Conclusion (3), we have:

$$1 = \pi(\mathbb{X} \times \mathbb{Z}) = \int_{\mathbb{Z}} \int_{\mathbb{X}} p(x|z) \xi(dx) v(z) \zeta(dz)$$

(Since the integral on a region with an a.e. zero value is zero [9, Thm. 15.2(i)],)

$$= \int_{\mathcal{V}} \int_{\mathcal{P}_z} p(x|z) \xi(dx) v(z) \zeta(dz) \quad (\text{Since } \mathcal{S}^{\mathbb{Z}} \subseteq \mathcal{V} \text{ and due to Billingsley [9, Thm. 16.9],})$$

$$= \left(\int_{\mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{P}_z} + \int_{\mathcal{V} \setminus \mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{P}_z} \right) p(x|z) \xi(dx) v(z) \zeta(dz) \quad (\text{Since } \mathcal{V} \setminus \mathcal{S}^{\mathbb{Z}} = \mathcal{V} \setminus (\mathcal{V} \cap \mathcal{W}_{p,q}^{\mathbb{Z}}) = \mathcal{V} \setminus \mathcal{W}_{p,q}^{\mathbb{Z}},)$$

$$= \left(\int_{\mathcal{S}^{\mathbb{Z}}} \int_{\mathcal{P}_z} + \int_{\mathcal{V} \setminus \mathcal{W}_{p,q}^{\mathbb{Z}}} \int_{\mathcal{P}_z} \right) p(x|z) \xi(dx) v(z) \zeta(dz).$$

For the second iterated integral, note that for any z on \mathcal{V} , $v(z) > 0$, so from Conclusion (3.1), for ζ -a.e. z on \mathcal{V} , we have $p(x|z) > 0 \implies q(z|x) > 0$, ξ -a.e. on \mathbb{X} . This means that for ζ -a.e. z on \mathcal{V} , $\xi\{x \mid p(x|z) > 0, q(z|x) = 0\} = \xi(\mathcal{P}_z \setminus \mathcal{Q}_z) = 0$. So the set $\mathcal{V} \setminus \mathcal{W}_{p,q}^{\mathbb{Z}} = \{z \in \mathcal{V} \mid \xi(\mathcal{P}_z \setminus \mathcal{Q}_z) > 0 \text{ or } \mathcal{P}_z = \emptyset\}$ has the same measure under ζ as the set $\{z \in \mathcal{V} \mid \mathcal{P}_z = \emptyset\}$. This means that the second iterated integral $\int_{\mathcal{V} \setminus \mathcal{W}_{p,q}^{\mathbb{Z}}} \int_{\mathcal{P}_z} p(x|z) \xi(dx) v(z) \zeta(dz) = \int_{\{z \in \mathcal{V} \mid \mathcal{P}_z = \emptyset\}} \int_{\mathcal{P}_z} p(x|z) \xi(dx) v(z) \zeta(dz) = 0$ [9, p.226, Thm. 15.2(i)].

For the first iterated integral, note that by construction, $\mathcal{P}_z = (\mathcal{W}_{p,q})_z$ for any z on $\mathcal{S}^{\mathbb{Z}}$, since $\mathcal{S}^{\mathbb{Z}} \subseteq \mathcal{W}_{p,q}^{\mathbb{Z}}$. Moreover, from Conclusion (4.1) that \mathcal{S} is a $\xi \otimes \zeta$ -complete component of $\mathcal{W}_{p,q}$, we can subsequently apply Lemma B.16, so $\int_{\mathcal{S}^{\mathbb{Z}}} \int_{(\mathcal{W}_{p,q})_z} p(x|z) \xi(dx) v(z) \zeta(dz) = \int_{\mathcal{S}} p(x|z) v(z) (\xi \otimes \zeta)(dx dz)$, which is $\pi(\mathcal{S})$ by Conclusion (3). This means that $\pi(\mathcal{S}) = 1$. The same deduction applies symmetrically to $\tilde{\mathcal{S}}$, so we also have $\pi(\tilde{\mathcal{S}}) = 1$.

Main procedure (“only if”). We will verify that the set \mathcal{S} given in Eq. (15) satisfies all the necessary conditions.

Conclusion (4.2) shows that $\pi(\mathcal{S}) = 1 > 0$, and in Conclusion (2) we have verified that $\pi \ll \xi \otimes \zeta$. So we have $(\xi \otimes \zeta)(\mathcal{S}) > 0$, which verifies Condition (iii).

To verify Conditions (iv) and (v), by Conclusions (3) and (1), we have $p(x|z)v(z) = q(z|x)u(x)$ and $q(z|x) > 0$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} . By the construction of \mathcal{S} , we also have $v(z) > 0$ and $p(x|z) > 0$ everywhere on \mathcal{S} . So the ratio $\frac{p(x|z)}{q(z|x)}$ is finite and positive, $\xi \otimes \zeta$ -a.e. on \mathcal{S} , and it factorizes as $\frac{p(x|z)}{q(z|x)} = u(x) \frac{1}{v(z)}$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} . By Conclusion (2), $a(x) := u(x)$ is ξ -integrable on $\mathcal{S}^{\mathbb{X}}$.

To verify Condition (ii), note that by construction, $\mathcal{S}^{\mathbb{Z}} \subseteq \mathcal{V} \cap \mathcal{W}_{p,q}^{\mathbb{Z}} \subseteq \mathcal{W}_{p,q}^{\mathbb{Z}}$ so $\mathcal{S}^{\mathbb{Z}} \subseteq^{\zeta} \mathcal{W}_{p,q}^{\mathbb{Z}}$. Note that $\mathcal{S}^{\mathbb{X}} = \{x \in \mathcal{U} \mid \mathcal{V} \cap (\mathcal{W}_{p,q})_x \neq \emptyset\} = \{x \in \mathcal{U} \mid \exists z \in \mathcal{V} \text{ s.t. } x \in \mathcal{P}_z, \mathcal{P}_z \subseteq^{\xi} \mathcal{Q}_z\} \subseteq \{x \in \mathcal{U} \mid \exists z \in \mathcal{V} \text{ s.t. } x \in \mathcal{P}_z\} = \{x \mid u(x) > 0, \exists z \text{ s.t. } v(z) > 0, p(x|z) > 0\}$. By Conclusion (3.1), for ξ -a.e. x on $\mathcal{S}^{\mathbb{X}}$, we have $\exists z$ s.t. $q(z|x)u(x) = p(x|z)v(z) > 0$, so $q(z|x) > 0$ hence $\mathcal{Q}_x \neq \emptyset$. Moreover, for ξ -a.e. x on $\mathcal{S}^{\mathbb{X}}$ and ζ -a.e. z on \mathcal{Q}_x , we have $p(x|z)v(z) = q(z|x)u(x) > 0$, so $p(x|z) > 0$ hence $\mathcal{Q}_x \subseteq^{\zeta} \mathcal{P}_x$. These two conclusions means that for ξ -a.e. x on $\mathcal{S}^{\mathbb{X}}$, we have $x \in \{x \mid \mathcal{Q}_x \neq \emptyset, \mathcal{Q}_x \subseteq^{\zeta} \mathcal{P}_x\}$ which is exactly $\mathcal{W}_{q,p}^{\mathbb{X}}$. Hence $\mathcal{S}^{\mathbb{X}} \subseteq^{\xi} \mathcal{W}_{q,p}^{\mathbb{X}}$.

Now, all the left is to verify Condition (i). Conclusion (4.1) has verified that \mathcal{S} is a $\xi \otimes \zeta$ -complete component of $\mathcal{W}_{p,q}$. To verify that \mathcal{S} is a $\xi \otimes \zeta$ -complete component also of $\mathcal{W}_{q,p}$, note that Conclusion (4.1) has also verified that $\tilde{\mathcal{S}}$ is a $\xi \otimes \zeta$ -complete component of $\mathcal{W}_{q,p}$, so by Lemma B.14 it suffices to verify that $\mathcal{S} \stackrel{\xi \otimes \zeta}{\equiv} \tilde{\mathcal{S}}$, $\mathcal{S}^{\mathbb{X}} \stackrel{\xi}{\equiv} \tilde{\mathcal{S}}^{\mathbb{X}}$ and $\mathcal{S}^{\mathbb{Z}} \stackrel{\zeta}{\equiv} \tilde{\mathcal{S}}^{\mathbb{Z}}$.

By construction, for any $(x, z) \in \mathcal{S}$, we have $v(z) > 0$ and $p(x|z) > 0$, so from Conclusion (3) the density function $p(x|z)v(z)$ of π is positive. Similarly, the density function $q(z|x)u(x)$ of π is positive everywhere on $\tilde{\mathcal{S}}$. Moreover, Conclusion (4.2) has shown that both $\pi(\mathcal{S}) = 1$ and $\pi(\tilde{\mathcal{S}}) = 1$. So by Lemma B.7, we know that $\mathcal{S} \stackrel{\xi \otimes \zeta}{=} \tilde{\mathcal{S}}$. Also by construction, the density function $u(x)$ of $\pi^{\mathbb{X}}$ is positive everywhere on $\mathcal{S}^{\mathbb{X}}$ and on $\tilde{\mathcal{S}}^{\mathbb{X}}$. Moreover, we have $\pi^{\mathbb{X}}(\mathcal{S}^{\mathbb{X}}) = \pi(\mathcal{S}^{\mathbb{X}} \times \mathbb{Z}) \geq \pi(\mathcal{S}) = 1$ so $\pi^{\mathbb{X}}(\mathcal{S}^{\mathbb{X}}) = 1$ and similarly $\pi^{\mathbb{X}}(\tilde{\mathcal{S}}^{\mathbb{X}}) = 1$. Again by Lemma B.7, we have $\mathcal{S}^{\mathbb{X}} \stackrel{\xi}{=} \tilde{\mathcal{S}}^{\mathbb{X}}$. It follows similarly that $\mathcal{S}^{\mathbb{Z}} \stackrel{\xi}{=} \tilde{\mathcal{S}}^{\mathbb{Z}}$.

“If” (sufficiency):

(5) For Conditions (iv) and (v), denote $\tilde{a}(x) := |a(x)|$ and $\tilde{b}(z) := |b(z)|$, and let $\tilde{A} := \int_{\mathcal{S}^{\mathbb{X}}} \tilde{a}(x) \xi(dx)$.

(5.1) From the definition of integrability in Supplement A.1, Condition (v) is equivalent to that $\tilde{a}(x)$ is ξ -integrable on $\mathcal{S}^{\mathbb{X}}$. Particularly, $\tilde{A} < \infty$.

Due to Condition (i) and Lemma B.12, we have $\mathcal{S} \subseteq \xi \otimes \zeta \mathcal{W}_{p,q}$. So by Condition (iv) and Conclusion (1), we have $\frac{p(x|z)}{q(z|x)} = a(x)b(z) > 0$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} . This means both $(\xi \otimes \zeta)\{(x, z) \in \mathcal{S} \mid a(x)b(z) = 0\} = 0$ and $(\xi \otimes \zeta)\{(x, z) \in \mathcal{S} \mid a(x)b(z) < 0\} = 0$, since their summation is zero.

(5.2) Since $\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, a(x) = 0, z \in \mathcal{S}_x\} \subseteq \{(x, z) \in \mathcal{S} \mid a(x)b(z) = 0\}$, the second-last equation in Conclusion (5.1) above means that $(\xi \otimes \zeta)\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, a(x) = 0, z \in \mathcal{S}_x\} = 0$. So $(\xi \otimes \zeta)\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, a(x) \neq 0, z \in \mathcal{S}_x\} = (\xi \otimes \zeta)\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, z \in \mathcal{S}_x\} - (\xi \otimes \zeta)\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, a(x) = 0, z \in \mathcal{S}_x\} = (\xi \otimes \zeta)(\mathcal{S}) > 0$ by Condition (iii). Moreover, by Eq. (7), we have $(\xi \otimes \zeta)\{(x, z) \mid x \in \mathcal{S}^{\mathbb{X}}, a(x) \neq 0, z \in \mathcal{S}_x\} = \int_{\{x \in \mathcal{S}^{\mathbb{X}} \mid a(x) \neq 0\}} \zeta(\mathcal{S}_x) \xi(dx) > 0$, so $\xi\{x \in \mathcal{S}^{\mathbb{X}} \mid a(x) \neq 0\} > 0$ [9, p.226]. Particularly, $\tilde{A} > 0$ [9, Thm. 15.2(ii)].

(5.3) Since $\{(x, z) \in \mathcal{S} \mid a(x)b(z) \neq \tilde{a}(x)\tilde{b}(z)\} = \{(x, z) \in \mathcal{S} \mid a(x)b(z) < 0\}$, the last equation in Conclusion (5.1) above means that $a(x)b(z) = \tilde{a}(x)\tilde{b}(z)$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} . So we have $\frac{p(x|z)}{q(z|x)} = \tilde{a}(x)\tilde{b}(z)$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} , following Condition (iv).

(6) Based on Conclusions (5.1) and (5.2), we can define the following finite and nonnegative functions on \mathbb{X} and \mathbb{Z} :

$$u(x) := \begin{cases} \frac{1}{\tilde{A}} \tilde{a}(x), & \text{if } x \in \mathcal{S}^{\mathbb{X}} \text{ and } \tilde{a}(x) > 0, \\ 0, & \text{otherwise,} \end{cases} \quad v(z) := \begin{cases} \frac{1}{\tilde{A}\tilde{b}(z)}, & \text{if } z \in \mathcal{S}^{\mathbb{Z}} \text{ and } \tilde{b}(z) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

(6.1) By construction, $\tilde{a}(x)\tilde{b}(z) = u(x)/v(z)$ on $\mathcal{S} \cap \{(x, z) \mid \tilde{b}(z) > 0\}$. So from Conclusion (5.3), we have $p(x|z)v(z) = q(z|x)u(x)$, $\xi \otimes \zeta$ -a.e. on $\mathcal{S} \cap \{(x, z) \mid \tilde{b}(z) > 0\}$. Moreover, following a similar deduction as in Conclusion (5.2), we know that $(\xi \otimes \zeta)(\mathcal{S} \setminus \{(x, z) \mid \tilde{b}(z) > 0\}) = (\xi \otimes \zeta)\{(x, z) \mid z \in \mathcal{S}^{\mathbb{Z}}, b(z) = 0, x \in \mathcal{S}_z\} = 0$. So we have $p(x|z)v(z) = q(z|x)u(x)$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} .

(6.2) By construction, $\int_{\mathbb{X}} u(x) \xi(dx) = \int_{\mathcal{S}^{\mathbb{X}}} u(x) \xi(dx) = 1$.

Main procedure (“if”). We will show that the following function on $\mathcal{X} \otimes \mathcal{Z}$, which is the same as Eq. (1) in the theorem, is a distribution on $\mathbb{X} \times \mathbb{Z}$ such that $\mu(\mathcal{X}|z)$ and $\nu(\mathcal{Z}|x)$ are its conditional distributions:

$$\pi(\mathcal{W}) := \int_{\mathcal{W} \cap \mathcal{S}} q(z|x)u(x)(\xi \otimes \zeta)(dxdz), \quad \forall \mathcal{W} \in \mathcal{X} \otimes \mathcal{Z}. \quad (16)$$

Consider any measurable rectangle $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$. We have:

$$\begin{aligned} \pi(\mathcal{X} \times \mathcal{Z}) &= \int_{\mathcal{X} \times \mathcal{Z} \cap \mathcal{S}} q(z|x)u(x)(\xi \otimes \zeta)(dxdz) \\ &= \int_{\mathcal{X} \cap \mathcal{S}^{\mathbb{X}}} \int_{\mathcal{Z} \cap \mathcal{S}_x} q(z|x)\zeta(dz)u(x)\xi(dx) && \text{(By Condition (i) and applying Lemma B.16.)} \\ &= \int_{\mathcal{X} \cap \mathcal{S}^{\mathbb{X}}} \int_{\mathcal{Z} \cap (\mathcal{W}_{q,p})_x} q(z|x)\zeta(dz)u(x)\xi(dx) \end{aligned}$$

$$\begin{aligned}
& (\text{Since } \mathcal{S}^{\mathbb{X}} \subseteq^{\xi} \mathcal{W}_{q,p}^{\mathbb{X}} \text{ by Condition (ii) and } (\mathcal{W}_{q,p})_x = \mathcal{Q}_x \text{ on } \mathcal{W}_{q,p}^{\mathbb{X}}) \\
& = \int_{\mathcal{X} \cap \mathcal{S}^{\mathbb{X}}} \int_{\mathcal{Z} \cap \mathcal{Q}_x} q(z|x) \zeta(dz) u(x) \xi(dx) \\
& (\text{Since by construction, } u(x) = 0 \text{ outside } \mathcal{S}^{\mathbb{X}} \text{ and } q(z|x) = 0 \text{ outside } \mathcal{Q}_x) \\
& = \int_{\mathcal{X}} \int_{\mathcal{Z}} q(z|x) \zeta(dz) u(x) \xi(dx) \quad (\text{Recalling that } q(z|x) \text{ is the density function of } \nu(\cdot|x),) \\
& = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) u(x) \xi(dx). \tag{17}
\end{aligned}$$

Moreover, due to Conclusion (6.1), we have $\pi(\mathcal{W}) = \int_{\mathcal{W} \cap \mathcal{S}} p(x|z) v(z) (\xi \otimes \zeta)(dx dz)$ on $\mathcal{X} \otimes \mathcal{Z}$ [9, Thm. 15.2(v)]. Using this form of π and noting that the symmetrized conditions in the above deduction also hold, we have:

$$\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mu(\mathcal{X}|z) v(z) \zeta(dz). \tag{18}$$

Since both $q(z|x)$ and $u(x)$ are finite and nonnegative on $\mathbb{X} \times \mathbb{Z}$, π is a measure on $\mathcal{X} \otimes \mathcal{Z}$ by its definition Eq. (16) [9, p.227]. Moreover, from Eq. (17), we have $\pi(\mathbb{X} \times \mathbb{Z}) = \int_{\mathbb{X}} u(x) \xi(dx) = 1$ by Conclusion (6.2). So π is a distribution (probability measure) on $\mathbb{X} \times \mathbb{Z}$.

From Eq. (17), we have $\pi^{\mathbb{X}}(\mathcal{X}) := \pi(\mathcal{X} \times \mathbb{Z}) = \int_{\mathcal{X}} u(x) \xi(dx)$. So $u(x)$ is a density function of $\pi^{\mathbb{X}}$, and Eq. (17) in turn becomes $\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx)$. This indicates that $\nu(\mathcal{Z}|x)$ is a conditional distribution of π w.r.t sub-sigma-field $\mathcal{X} \times \{\mathcal{Z}\}$ for any $\mathcal{Z} \in \mathcal{Z}$, due to Eq. (11).

Similarly, from Eq. (18), we have $\pi^{\mathbb{Z}}(\mathcal{Z}) := \pi(\mathbb{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} v(z) \zeta(dz)$. So $v(z)$ is a density function of $\pi^{\mathbb{Z}}$, and Eq. (18) in turn becomes $\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mu(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz)$. This indicates that $\mu(\mathcal{X}|z)$ is a conditional distribution of π w.r.t sub-sigma-field $\{\mathcal{X}\} \times \mathcal{Z}$ for any $\mathcal{X} \in \mathcal{X}$, due to Eq. (11). The proof is completed. \square

C.3 Complete Support Proposition under the a.e.-Full Support Condition

Proposition C.4. *If $p(x|z)$ and $q(z|x)$ have a.e.-full supports, then $\mathcal{W}_{p,q} \stackrel{\xi \otimes \zeta}{\subseteq} \mathcal{W}_{q,p} \stackrel{\xi \otimes \zeta}{\subseteq} \mathbb{X} \times \mathbb{Z}$, and $\mathbb{X} \times \mathbb{Z}$ is the $\xi \otimes \zeta$ -unique complete support of them when compatible.*

Proof. By definition, $p(x|z) > 0$ and $q(z|x) > 0$, $\xi \otimes \zeta$ -a.e. By Lemma B.10, this means that for ξ -a.e. x , $p(x|z) > 0$ ζ -a.e. thus $\mathcal{P}_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$, and for ζ -a.e. z , $\mathcal{Q}_z \stackrel{\xi}{\subseteq} \mathbb{X}$. Similarly, we also have for ξ -a.e. x , $\mathcal{Q}_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$ thus $\mathcal{Q}_x \stackrel{\zeta}{\subseteq} \mathcal{P}_x$ by the transitivity (Lemma B.4) and subsequently $\mathcal{Q}_x \subseteq^{\zeta} \mathcal{P}_x$. Similarly, for ζ -a.e. z , $\mathcal{P}_z \subseteq^{\xi} \mathcal{Q}_z$. This means that for ξ -a.e. x , $(\mathcal{W}_{q,p})_x = \mathcal{Q}_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$ thus $z \in (\mathcal{W}_{q,p})_x$, ζ -a.e. By Lemma B.10, this means that for $\xi \otimes \zeta$ -a.e. (x, z) , we have $(x, z) \in \mathcal{W}_{q,p}$, so $\mathcal{W}_{q,p} \stackrel{\xi \otimes \zeta}{\subseteq} \mathbb{X} \times \mathbb{Z}$. Similarly, we have $\mathcal{W}_{p,q} \stackrel{\xi \otimes \zeta}{\subseteq} \mathbb{X} \times \mathbb{Z}$.

Let \mathcal{S} be a complete support of $p(x|z)$ and $q(z|x)$ when they are compatible. Since $(\xi \otimes \zeta)(\mathcal{S}) = 1 > 0$ from Condition (iii) in Theorem 2.3, we know that $\xi(\mathcal{S}^{\mathbb{X}}) > 0$ by Eq. (7) and Billingsley [9, p.226]. So there is an $x \in \mathcal{S}^{\mathbb{X}}$ such that $(\mathcal{W}_{q,p})_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$, otherwise there would be a non-measure-zero set of x violating $(\mathcal{W}_{q,p})_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$. As a $\xi \otimes \zeta$ -complete component of $\mathcal{W}_{q,p}$ by Condition (i) in Theorem 2.3, we have $\mathcal{S} \stackrel{\xi \otimes \zeta}{\subseteq} (\mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cap \mathcal{W}_{q,p}) \cup (\mathbb{Z} \times \mathcal{S}^{\mathbb{Z}} \cap \mathcal{W}_{q,p}) \supseteq \mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \cap \mathcal{W}_{q,p} \stackrel{\xi \otimes \zeta}{\subseteq} \mathcal{S}^{\mathbb{X}} \times \mathbb{Z}$. This means that $(\xi \otimes \zeta)(\mathcal{S}^{\mathbb{X}} \times \mathbb{Z} \setminus \mathcal{S}) = \int_{\mathcal{S}^{\mathbb{X}}} \zeta(\mathbb{Z} \setminus \mathcal{S}_x) \xi(dx) = 0$ by Eq. (7), so for ξ -a.e. x on $\mathcal{S}^{\mathbb{X}}$, $\zeta(\mathbb{Z} \setminus \mathcal{S}_x) = 0$ [9, Thm. 15.2(ii)] thus $\mathcal{S}_x \stackrel{\zeta}{\subseteq} \mathbb{Z}$. So $\mathcal{S}^{\mathbb{Z}} \stackrel{\zeta}{\subseteq} \mathbb{Z}$. Moreover, we also have $\mathcal{S} \supseteq^{\xi \otimes \zeta} \mathbb{X} \times \mathcal{S}^{\mathbb{Z}}$ which $\stackrel{\xi \otimes \zeta}{\subseteq} \mathbb{X} \times \mathbb{Z}$, so we have $\mathcal{S} \stackrel{\xi \otimes \zeta}{\subseteq} \mathbb{X} \times \mathbb{Z}$. Similarly, from that \mathcal{S} is a $\xi \otimes \zeta$ -complete component also of $\mathcal{W}_{p,q}$, we have the same conclusion. \square

C.4 Proof of Theorem 2.4

Proof. Let π and $\tilde{\pi}$ be two compatible joints of $p(x|z)$ and $q(z|x)$, and they are supported on the same complete support \mathcal{S} . By Conclusions (2) and (3) in the proof (Supplement C.2) of Theorem 2.3, there exist functions $u(x)$, $v(z)$ and $\tilde{u}(x)$, $\tilde{v}(z)$ such that $p(x|z)v(z)$ and $q(z|x)u(x)$ are the densities of π , and $p(x|z)\tilde{v}(z)$ and $q(z|x)\tilde{u}(x)$ of $\tilde{\pi}$, and $p(x|z)v(z) = q(z|x)u(x)$ and $p(x|z)\tilde{v}(z) = q(z|x)\tilde{u}(x)$,

$\xi \otimes \zeta$ -a.e. By the definition of a support in Lemma B.7, we know that the densities of π and $\tilde{\pi}$ are positive $\xi \otimes \zeta$ -a.e. on \mathcal{S} , and $\int_{\mathcal{S}^x} u \, d\xi = \int_{\mathcal{S}^x} \tilde{u} \, d\xi = \int_{\mathcal{S}^z} v \, d\zeta = \int_{\mathcal{S}^z} \tilde{v} \, d\zeta = 1$.

Consequently, we have $\frac{p(x|z)}{q(z|x)} = \frac{u(x)}{v(z)} = \frac{\tilde{u}(x)}{\tilde{v}(z)}$, $\xi \otimes \zeta$ -a.e. on \mathcal{S} . By Lemma B.10, for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$, we have $\frac{u(x)}{v(z)} = \frac{\tilde{u}(x)}{\tilde{v}(z)}$ for ξ -a.e. x on \mathcal{S}_z , which means that $\int_{\mathcal{S}_z} \frac{u(x)}{v(z)} \xi(dx) = \int_{\mathcal{S}_z} \frac{\tilde{u}(x)}{\tilde{v}(z)} \xi(dx)$. Since for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$, $\mathcal{S}_z \stackrel{\xi}{=} \mathcal{S}^{\mathbb{X}}$, we have by Lemma B.3 that $\int_{\mathcal{S}^x} \frac{u(x)}{v(z)} \xi(dx) = \int_{\mathcal{S}^x} \frac{\tilde{u}(x)}{\tilde{v}(z)} \xi(dx)$, which in turn gives $\frac{1}{v(z)} \int_{\mathcal{S}^x} u \, d\xi = \frac{1}{v(z)} = \frac{1}{\tilde{v}(z)} \int_{\mathcal{S}^x} \tilde{u} \, d\xi = \frac{1}{\tilde{v}(z)}$. So $v(z) = \tilde{v}(z)$ for ζ -a.e. z on $\mathcal{S}^{\mathbb{Z}}$, and similarly $u(x) = \tilde{u}(x)$ for ξ -a.e. x on $\mathcal{S}^{\mathbb{X}}$. Subsequently, the density $p(x|z)v(z)$ or $q(z|x)u(x)$ of π is $\xi \otimes \zeta$ -a.e. the same as the density $p(x|z)\tilde{v}(z)$ or $q(z|x)\tilde{u}(x)$ of $\tilde{\pi}$. Hence, π and $\tilde{\pi}$ are the same distribution. \square

C.5 The Dirac Compatibility Lemma

Before proving the main instructive compatibility theorem (2.6) for the Dirac case, we first present an existential equivalent criterion for compatibility, which provides insights to the problem.

Lemma C.5 (Dirac compatibility, existential). *Conditional distribution $\nu(\mathcal{Z}|x)$ is compatible with $\mu(\mathcal{X}|z) := \delta_{f(z)}(\mathcal{X})$ where function $f : \mathbb{Z} \rightarrow \mathbb{X}$ is \mathcal{X}/\mathcal{Z} -measurable, if and only if there is a distribution β on $(\mathbb{Z}, \mathcal{Z})$ such that $\nu(\mathcal{Z}|x) = \frac{d\beta(\mathcal{Z} \cap f^{-1}(\cdot))}{d\beta(f^{-1}(\cdot))}(x)$, and this β is the marginal $\pi^{\mathbb{Z}}$ of a compatible joint π of them.*

Proof. We first show the validity of the R-N derivative. Since β is a distribution thus a finite measure, $\beta(\mathcal{Z} \cap f^{-1}(\cdot))$ and $\beta(f^{-1}(\cdot))$ are also finite thus sigma-finite. For any $\mathcal{X} \in \mathcal{X}$ such that $\beta(f^{-1}(\mathcal{X})) = 0$, we have $\beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) \leq \beta(f^{-1}(\mathcal{X})) = 0$ since $\mathcal{Z} \cap f^{-1}(\mathcal{X}) \subseteq f^{-1}(\mathcal{X})$ and measures are monotone. So $\beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = 0$ and $\beta(\mathcal{Z} \cap f^{-1}(\cdot)) \ll \beta(f^{-1}(\cdot))$. By the R-N theorem [9, Thm. 32.2], the R-N derivative exists.

“Only if” (necessity): Let π be a compatible joint. Since μ and ν are its conditional distributions, by Eq. (11), we have:

$$\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mu(\mathcal{X}|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx), \quad \forall \mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}.$$

The first integral is $\int_{\mathcal{Z}} \mathbb{I}[f(z) \in \mathcal{X}] \pi^{\mathbb{Z}}(dz) = \int_{\mathcal{Z}} \mathbb{I}[z \in f^{-1}(\mathcal{X})] \pi^{\mathbb{Z}}(dz) = \pi^{\mathbb{Z}}(\mathcal{Z} \cap f^{-1}(\mathcal{X}))$. Particularly, $\pi^{\mathbb{X}}(\mathcal{X}) = \pi(\mathcal{X} \times \mathbb{Z}) = \pi^{\mathbb{Z}}(f^{-1}(\mathcal{X}))$, i.e. $\pi^{\mathbb{X}}$ is the transformed (pushed-forward) distribution from $\pi^{\mathbb{Z}}$ by measurable function f [9, p.196]. On the other hand, the equality to the second integral means that $\pi^{\mathbb{Z}}(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{X}}(dx) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{Z}}(f^{-1}(dx))$. This means that $\nu(\mathcal{Z}|x)$ is the R-N derivative of $\mathcal{X} \mapsto \pi^{\mathbb{Z}}(\mathcal{Z} \cap f^{-1}(\mathcal{X}))$ w.r.t $\mathcal{X} \mapsto \pi^{\mathbb{Z}}(f^{-1}(\mathcal{X}))$. Taking β as $\pi^{\mathbb{Z}}$, which is a distribution on $(\mathbb{Z}, \mathcal{Z})$, yields the necessary condition.

“If” (sufficiency): For any measurable rectangle $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, define $\pi(\mathcal{X} \times \mathcal{Z}) := \int_{\mathcal{Z}} \mu(\mathcal{X}|z) \beta(dz)$ and $\tilde{\pi}(\mathcal{X} \times \mathcal{Z}) := \beta(\mathcal{Z} \cap f^{-1}(\mathcal{X}))$. Since for any $z \in \mathbb{Z}$, $f(z) \in \mathbb{X}$, so $\pi(\mathbb{X} \times \mathbb{Z}) = \int_{\mathbb{Z}} \mu(\mathbb{X}|z) \beta(dz) = \int_{\mathbb{Z}} \beta(dz) = 1$. Since $f^{-1}(\mathbb{X}) = \mathbb{Z}$, we have $\tilde{\pi}(\mathbb{X} \times \mathbb{Z}) = \beta(\mathbb{Z}) = 1$. So both π and $\tilde{\pi}$ are finite thus sigma-finite. Moreover, for any $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, we have $\pi(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mathbb{I}[f(z) \in \mathcal{X}] \beta(dz) = \int_{\mathcal{Z}} \mathbb{I}[z \in f^{-1}(\mathcal{X})] \beta(dz) = \int_{\mathcal{Z} \cap f^{-1}(\mathcal{X})} \beta(dz) = \beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \tilde{\pi}(\mathcal{X} \times \mathcal{Z})$, i.e. π and $\tilde{\pi}$ agree on the pi-system $\mathcal{X} \times \mathcal{Z}$. So by Billingsley [9, Thm. 10.3], π and $\tilde{\pi}$ extend to the same distribution (probability measure) on $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$.

On the other hand, we have $\pi^{\mathbb{Z}}(\mathcal{Z}) = \pi(\mathbb{X} \times \mathcal{Z}) = \int_{\mathcal{Z}} \mu(\mathbb{X}|z) \beta(dz) = \int_{\mathcal{Z}} \beta(dz) = \beta(\mathcal{Z})$, and furthermore from this, μ is a conditional distribution of π due to its construction and Eq. (11). Moreover, $\tilde{\pi}^{\mathbb{X}}(\mathcal{X}) = \tilde{\pi}(\mathcal{X} \times \mathbb{Z}) = \beta(\mathbb{Z} \cap f^{-1}(\mathcal{X})) = \beta(f^{-1}(\mathcal{X}))$, and by the definition of $\nu(\mathcal{Z}|x)$ as an R-N derivative, we have $\beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \beta(f^{-1}(dx))$, which is $\tilde{\pi}(\mathcal{X} \times \mathcal{Z}) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \tilde{\pi}^{\mathbb{X}}(dx)$. So again due to Eq. (11), ν is a conditional distribution of $\tilde{\pi}$. Since π and $\tilde{\pi}$ are the same distribution on $(\mathbb{X} \times \mathbb{Z}, \mathcal{X} \otimes \mathcal{Z})$, we know that μ and ν are compatible. \square

Key insights. Let π be a compatible joint of $\mu(\mathcal{X}|z) := \delta_{f(z)}(\mathcal{X})$ and $\nu(\mathcal{Z}|x)$. For any $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, we have:

$$\pi(\mathcal{X} \times \mathcal{Z}) = \pi^{\mathbb{Z}}(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \pi^{\mathbb{Z}}(f^{-1}(dx)) = \int_{f^{-1}(\mathcal{X})} \nu(\mathcal{Z}|f(z)) \pi^{\mathbb{Z}}(dz),$$

where the last equality holds due to the rule of change of variables [9, Thm. 16.13]. Let $f^{-1}(\mathcal{X}) := \sigma(\{f^{-1}(\mathcal{X}) \mid \mathcal{X} \in \mathcal{Z}\})$ be the pulled-back sigma-field from \mathcal{Z} by f . It is a sub-sigma-field of \mathcal{Z} as every $f^{-1}(\mathcal{X}) \in \mathcal{Z}$ since f is measurable. So the last equality means that:

$$\nu(\mathcal{Z}|f(z)) = \frac{d\pi^{\mathbb{Z}}(\mathcal{Z} \cap \cdot)}{d\pi^{\mathbb{Z}}(\cdot)} \Big|_{f^{-1}(\mathcal{X})} (z).$$

The expression on the left makes sense since for all values of z that yield the same value of $f(z)$, the R-N derivative is the same. The second equality also gives:

$$\nu(\mathcal{Z}|x) = \frac{d\pi^{\mathbb{Z}}(\mathcal{Z} \cap f^{-1}(\cdot))}{d\pi^{\mathbb{Z}}(f^{-1}(\cdot))} \Big|_{\mathcal{X}} (x).$$

C.6 Proof of Theorem 2.6

Proof. “Only if” (necessity): Suppose that $\nu(\mathcal{Z}|x)$ and $\mu(\mathcal{X}|z) := \delta_{f(z)}(\mathcal{X})$ are compatible but for any $x \in \mathbb{X}$, $\nu(f^{-1}(\{x\})|x) < 1$. Consider the set $\mathcal{S} := \{(f(z), z) \mid z \in \mathbb{Z}\}$. Since f is \mathcal{X}/\mathcal{Z} -measurable, this set \mathcal{S} is $\mathcal{X} \otimes \mathcal{Z}$ -measurable. It is also easy to verify that $\mathcal{S}_z = \{f(z)\}$ and $\mathcal{S}_x = f^{-1}(\{x\})$. Now let π be any of their compatible joint distribution. From Eq. (12), we know that $\pi(\mathcal{S}) = \int_{\mathbb{Z}} \mu(\mathcal{S}_z|z) \pi^{\mathbb{Z}}(dz) = \int_{\mathbb{Z}} \delta_{f(z)}(\{f(z)\}) \pi^{\mathbb{Z}}(dz) = \int_{\mathbb{Z}} \pi^{\mathbb{Z}}(dz) = \pi^{\mathbb{Z}}(\mathbb{Z}) = 1$. On the other hand, also from Eq. (12) and due to the compatibility, we have $\pi(\mathcal{S}) = \int_{\mathbb{X}} \nu(\mathcal{S}_x|x) \pi^{\mathbb{X}}(dx) = \int_{\mathbb{X}} \nu(f^{-1}(\{x\})|x) \pi^{\mathbb{X}}(dx) < \int_{\mathbb{X}} \pi^{\mathbb{X}}(dx) = \pi^{\mathbb{X}}(\mathbb{X}) = 1$, which leads to a contradiction. So if $\nu(\mathcal{Z}|x)$ and $\mu(\mathcal{X}|z)$ are compatible, then there is $x_0 \in \mathbb{X}$ such that $\nu(f^{-1}(\{x_0\})|x_0) = 1$.

“If” (sufficiency): Let $\beta(\mathcal{Z}) := \nu(f^{-1}(\{x_0\}) \cap \mathcal{Z}|x_0)$ be a set function on \mathcal{Z} . We can verify that this β is a distribution (probability measure) on $(\mathbb{Z}, \mathcal{Z})$ since $\nu(\cdot|x_0)$ is. Particularly, since f is \mathcal{X}/\mathcal{Z} -measurable and $\{x_0\} \in \mathcal{X}$ due to the assumption, we know that $f^{-1}(\{x_0\})$ thus $f^{-1}(\{x_0\}) \cap \mathcal{Z}$ for any $\mathcal{Z} \in \mathcal{Z}$ are in \mathcal{Z} ; $\beta(\emptyset) = \nu(\emptyset|x_0) = 0$; $\beta(\mathbb{Z}) = \nu(f^{-1}(\{x_0\})|x_0) = 1$ according to the assumption; for any countable disjoint \mathcal{Z} -sets $\mathcal{Z}^{(1)}, \mathcal{Z}^{(2)}, \dots$, it holds that $\mathcal{Z}^{(1)} \cap f^{-1}(\{x_0\}), \mathcal{Z}^{(2)} \cap f^{-1}(\{x_0\}), \dots$ are also disjoint \mathcal{Z} -sets, so $\beta(\bigcup_{i=1}^{\infty} \mathcal{Z}^{(i)}) = \nu(f^{-1}(\{x_0\}) \cap \bigcup_{i=1}^{\infty} \mathcal{Z}^{(i)}|x_0) = \nu(\bigcup_{i=1}^{\infty} f^{-1}(\{x_0\}) \cap \mathcal{Z}^{(i)}|x_0) = \sum_{i=1}^{\infty} \nu(f^{-1}(\{x_0\}) \cap \mathcal{Z}^{(i)}|x_0) = \sum_{i=1}^{\infty} \beta(\mathcal{Z}^{(i)})$.

Now we prove that $\beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \int_{\mathcal{X}} \nu(\mathcal{Z}|x) \beta(f^{-1}(dx))$, $\forall \mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$ which is sufficient due to Lemma C.5. For any $\mathcal{X} \times \mathcal{Z} \in \mathcal{X} \times \mathcal{Z}$, the l.h.s is $\beta(\mathcal{Z} \cap f^{-1}(\mathcal{X})) = \nu(f^{-1}(\{x_0\}) \cap f^{-1}(\mathcal{X}) \cap \mathcal{Z}|x_0) = \nu(f^{-1}(\{x_0\}) \cap \mathcal{Z}|x_0) \mathbb{I}[x_0 \in \mathcal{X}]$, where the last equality holds since $z \in f^{-1}(\{x_0\}) \cap f^{-1}(\mathcal{X})$ if and only if $f(z) = x_0 \in \mathcal{X}$. The integral on the r.h.s is $\int_{\mathcal{X}} \nu(\mathcal{Z}|x) \nu(f^{-1}(\{x_0\}) \cap f^{-1}(dx)|x_0)$. Since the measure $\mathcal{X} \mapsto \nu(f^{-1}(\{x_0\}) \cap f^{-1}(\mathcal{X})|x_0)$ is zero on the set $\mathcal{X} \setminus \{x_0\}$ (if there is any $z \in f^{-1}(\{x_0\}) \cap f^{-1}(\mathcal{X} \setminus \{x_0\})$, then we have $f(z) = x_0$ and $f(z) \in \mathcal{X} \setminus \{x_0\}$, which is a contradiction), the integral can be reduced on $\{x_0\} \cap \mathcal{X}$ [9, Thm. 16.9]: $\int_{\{x_0\} \cap \mathcal{X}} \nu(\mathcal{Z}|x) \nu(f^{-1}(\{x_0\}) \cap f^{-1}(dx)|x_0) = \mathbb{I}[x_0 \in \mathcal{X}] \nu(\mathcal{Z}|x_0) \nu(f^{-1}(\{x_0\})|x_0) = \nu(\mathcal{Z}|x_0) \mathbb{I}[x_0 \in \mathcal{X}]$. Moreover, $\nu(\mathcal{Z}|x_0) = \nu(\mathcal{Z} \cap f^{-1}(\{x_0\})|x_0) + \nu(\mathcal{Z} \setminus f^{-1}(\{x_0\})|x_0)$ where $\nu(\mathcal{Z} \setminus f^{-1}(\{x_0\})|x_0) \leq \nu(\mathbb{Z} \setminus f^{-1}(\{x_0\})|x_0) = 1 - \nu(f^{-1}(\{x_0\})|x_0) = 0$, we have $\nu(\mathcal{Z} \setminus f^{-1}(\{x_0\})|x_0) = 0$ and $\nu(\mathcal{Z}|x_0) = \nu(\mathcal{Z} \cap f^{-1}(\{x_0\})|x_0)$. So the integral on the r.h.s is $\nu(\mathcal{Z} \cap f^{-1}(\{x_0\})|x_0) \mathbb{I}[x_0 \in \mathcal{X}]$, which is the same as the l.h.s. So the equality is verified. \square

D Topics on the Methods of CyGen

D.1 Relation to other auto-encoder regularizations

There are methods that consider regularizing the standard auto-encoder (AE) [66, 4] with deterministic encoder $g(x)$ and decoder $f(z)$ for certain robustness. These regularizations are introduced in addition to the standard AE loss, *i.e.* the reconstruction loss: $\mathbb{E}_{p^*(x)} \ell(x, f(g(x)))$, where $\ell(x, x')$ is a measure of similarity between x and x' . If $\ell(x, f(z))$ can be treated as a (scaled) negative log-likelihood $-\log p(x|z)$ on \mathbb{X} (*e.g.*, squared 2-norm ℓ for a Gaussian $p(x|z)$, cross entropy ℓ for a Bernoulli/categorical $p(x|z)$), then we can adopt a distributional view of the decoder as $p(x|z)$ and the encoder as $\delta_{g(x)}(z)$ ¹², and reformulate the reconstruction loss also under the distributional view: $\mathbb{E}_{p^*(x)} [-\log p(x|g(x))] = \mathbb{E}_{p^*(x) \delta_{g(x)}(z)} [-\log p(x|z)]$.

¹²This is the notation of a Dirac’s delta function, which is not a function in the usual sense. We adopt this form for the similarity to the DAE loss.

Comparison with Jacobian norm regularizations. Contractive AE (CAE) [63, 64] regularizes the Jacobian norm of the encoder, $\lambda \mathbb{E}_{p^*(x)} \|\nabla_x g^\top(x)\|_F^2$ (λ controls the scale), in hope to encourage the robustness of the encoded representation against local changes around training data. When it is combined with the reconstruction loss which preserves data variation in the representation for reconstruction, the robustness is confined to the orthogonal direction to the data manifold, which often does not reflect semantic meanings of interest. In other words, the variation in this orthogonal direction is contracted in the representation, hence the name. When applied to a linear encoder, this becomes the well-known weight-decay regularizer. Note that CAE does not have a generative modeling utility, as it uses a deterministic encoder which leads to insufficient determinacy (Sec. 2.2.2).

Denosing AE (DAE) [77, 6, 7] considers the robustness to random corruption/perturbation on data, so its encoding process is $z = g(x + \epsilon_e)$ where $\epsilon_e \sim \mathcal{N}(0, \sigma_e^2 I_{d_x})$ (or any other distribution with $\mathbb{E}[\epsilon_e] = 0$ and $\text{Var}[\epsilon_e] = \sigma_e^2 I_{d_x}$), which defines a probabilistic encoder $q(z|x)$ (note that this is different from an additive Gaussian encoder). The goal for training a DAE is thus to try to reconstruct the input under the random corruption, by minimizing the DAE loss:

$$\mathbb{E}_{p^*(x)q(z|x)} [-\log p(x|z)], \quad (19)$$

which resembles the distributional form of the standard reconstruction loss. For infinitesimal corruption variance σ_e^2 and squared 2-norm ℓ , the DAE loss Eq. (19) is roughly equivalent to regularizing the standard reconstruction loss with $\sigma_e^2 \mathbb{E}_{p^*(x)} \|\nabla_x (f \circ g)^\top(x)\|_F^2$, *i.e.* the Jacobian norm of the reconstruction function [63, 1]. So DAE can be viewed to promote the robustness of reconstruction while CAE of the representation [63].

In contrast, for additive Gaussian decoder (*i.e.*, squared 2-norm ℓ) and encoder, our compatibility regularization Eq. (2) is $\mathbb{E}_{p(x,z)} \left\| \frac{1}{\sigma_d^2} (\nabla_z f^\top(z))^\top - \frac{1}{\sigma_e^2} \nabla_x g^\top(x) \right\|_F^2$, which is different from CAE and DAE regularizations. Ideologically, the compatibility loss is an intrinsic constraint to make use of the distributional nature of the encoder and decoder, and is not motivated from the additional requirement of robustness in some sense.

Comparison with a more accurate DAE reformulation. In fact, the analysis in [63, 1] for DAE as a regularization of the reconstruction loss is inaccurate. Key ingredients for the analysis are the Taylor expansions: $\|x + \varepsilon\|_2^2 = \|x\|_2^2 + 2x^\top \varepsilon + \varepsilon^\top \varepsilon$, $\exp\{x + \varepsilon\} = \exp\{x\}(1 + \varepsilon + \frac{1}{2}\varepsilon^2) + o(\varepsilon^2)$, and $\log(1 + \varepsilon) = \varepsilon - \frac{1}{2}\varepsilon^2 + o(\varepsilon^2)$. In the following, we consider $\ell(x, x') = \|x - x'\|_2^2$, corresponding to an additive Gaussian decoder $p(x|z)$.

First consider the additive Gaussian encoder, $q(z|x) = \mathcal{N}(z|g(x), \sigma_e^2 I_{d_z})$, or $z = g(x) + \epsilon_e$, $\epsilon_e \sim \mathcal{N}(0, \sigma_e^2 I_{d_z})$. Consider the case for infinitesimal σ_e . For the DAE loss Eq. (19), we have (omitting the expectation over $p^*(x)$):

$$\begin{aligned} \mathbb{E}_{q(z|x)} [-\log p(x|z)] &= \mathbb{E}_{q(z|x)} [\ell(x, f(z))] = \mathbb{E}_{q(z|x)} \|x - f(z)\|_2^2 = \mathbb{E}_{p(\epsilon_e)} \|x - f(g(x) + \epsilon_e)\|_2^2 \\ &= \mathbb{E}_{p(\epsilon_e)} \left[\left\| x - f(g(x)) - (\nabla f^\top)^\top \epsilon_e - \frac{1}{2} \epsilon_e^\top (\nabla^2 f) \epsilon_e + o(\epsilon_e^2) \right\|_2^2 \right] \\ &= \mathbb{E}_{p(\epsilon_e)} \left[\|x - f(g(x))\|_2^2 - 2(x - f(g(x)))^\top \left((\nabla f^\top)^\top \epsilon_e + \frac{1}{2} \epsilon_e^\top (\nabla^2 f) \epsilon_e \right) \right. \\ &\quad \left. + \epsilon_e^\top (\nabla f^\top) (\nabla f^\top)^\top \epsilon_e + o(\epsilon_e^2) \right] \\ &= \ell(x, f(g(x))) - \sigma_e^2 (x - f(g(x)))^\top \Delta f + \sigma_e^2 \|\nabla f^\top\|_F^2 + o(\sigma_e^2), \end{aligned} \quad (20)$$

where $(\nabla f^\top)^\top$ is the Jacobian of f , $(\epsilon_e^\top (\nabla^2 f) \epsilon_e)_i := \sum_{j,k=1..d_z} (\epsilon_e)_i (\epsilon_e)_j \partial_{z_i} \partial_{z_j} f_i$, and $(\Delta f)_i := \sum_{j=1..d_z} \partial_{z_j} \partial_{z_j} f_i$, and they are evaluated at $z = g(x)$. Note that in addition to the Jacobian norm regularization term discovered in [63, 1], there is a second regularization term $-\sigma_e^2 (x - f(g(x)))^\top \Delta f$ that DAE imposes.

For the data-fitting loss of CyGen Eq. (4), a similar approximation can be derived (again, omitting the expectation over $p^*(x)$):

$$\begin{aligned} \log \mathbb{E}_{q(z|x)} [1/p(x|z)] &= \log \mathbb{E}_{q(z|x)} \exp\{-\log p(x|z)\} = \log \mathbb{E}_{q(z|x)} \exp\{\ell(x, f(z))\} \\ &= \log \mathbb{E}_{q(z|x)} \exp\{\|x - f(z)\|_2^2\} = \log \mathbb{E}_{p(\epsilon_e)} \exp\{\|x - f(g(x) + \epsilon_e)\|_2^2\} \end{aligned}$$

$$\begin{aligned}
&= \log \mathbb{E}_{p(\epsilon_e)} \exp \left\{ \left\| x - f(g(x)) - (\nabla f^\top)^\top \epsilon_e - \frac{1}{2} \epsilon_e^\top (\nabla^2 f) \epsilon_e + o(\epsilon_e^2) \right\|_2^2 \right\} \\
&= \log \mathbb{E}_{p(\epsilon_e)} \exp \left\{ \|x - f(g(x))\|_2^2 - 2(x - f(g(x)))^\top \left((\nabla f^\top)^\top \epsilon_e + \frac{1}{2} \epsilon_e^\top (\nabla^2 f) \epsilon_e \right) \right. \\
&\quad \left. + \epsilon_e^\top (\nabla f^\top) (\nabla f^\top)^\top \epsilon_e + o(\epsilon_e^2) \right\} \\
&= \log \mathbb{E}_{p(\epsilon_e)} \left[\exp \{ \|x - f(g(x))\|_2^2 \} \left(1 - 2(x - f(g(x)))^\top \left((\nabla f^\top)^\top \epsilon_e + \frac{1}{2} \epsilon_e^\top (\nabla^2 f) \epsilon_e \right) \right. \right. \\
&\quad \left. \left. + \epsilon_e^\top (\nabla f^\top) (\nabla f^\top)^\top \epsilon_e + 2 \left((x - f(g(x)))^\top (\nabla f^\top)^\top \epsilon_e \right)^2 + o(\epsilon_e^2) \right) \right] \\
&= \log \left[\exp \{ \|x - f(g(x))\|_2^2 \} \left(1 - \sigma_e^2 (x - f(g(x)))^\top \Delta f \right. \right. \\
&\quad \left. \left. + \sigma_e^2 \|\nabla f^\top\|_F^2 + 2\sigma_e^2 \|(\nabla f^\top)(x - f(g(x)))\|_2^2 + o(\sigma_e^2) \right) \right] \\
&= \ell(x - f(g(x)) - \sigma_e^2 (x - f(g(x)))^\top \Delta f + \sigma_e^2 \|\nabla f^\top\|_F^2 + 2\sigma_e^2 \|(\nabla f^\top)(x - f(g(x)))\|_2^2 + o(\sigma_e^2)).
\end{aligned}$$

This is different from the regularization interpretation of DAE Eq. (20) as a third regularization term $2\sigma_e^2 \|(\nabla f^\top)(x - f(g(x)))\|_2^2$ is presented.

The compatibility loss Eq. (2) in CyGen becomes: $\mathbb{E}_{\rho(x,z)} \left\| \frac{1}{\sigma_d^2} (\nabla_z f^\top(z))^\top - \frac{1}{\sigma_e^2} \nabla_x g^\top(x) \right\|_F^2$, where σ_d^2 is the Gaussian variance of the decoder $p(x|z)$ (inverse scale for $\ell(\cdot, \cdot)$). When $\rho(x, z) = p^*(x)q(z|x)$, this can be further reduced to (omitting the expectation over $p^*(x)$):

$$\begin{aligned}
&\mathbb{E}_{q(z|x)} \left\| \frac{1}{\sigma_d^2} (\nabla_z f^\top(z))^\top - \frac{1}{\sigma_e^2} \nabla_x g^\top(x) \right\|_F^2 = \mathbb{E}_{p(\epsilon_e)} \left\| \frac{1}{\sigma_d^2} (\nabla f^\top(g(x) + \epsilon_e))^\top - \frac{1}{\sigma_e^2} \nabla g^\top \right\|_F^2 \\
&= \mathbb{E}_{p(\epsilon_e)} \left\| \frac{1}{\sigma_d^2} \left((\nabla f^\top)^\top + (\nabla^2 f^\top)^\top \epsilon_e + \frac{1}{2} \epsilon_e^\top (\nabla^3 f^\top)^\top \epsilon_e + o(\epsilon_e^2) \right) - \frac{1}{\sigma_e^2} \nabla g^\top \right\|_F^2 \\
&= \left\| \frac{1}{\sigma_d^2} (\nabla f^\top)^\top - \frac{1}{\sigma_e^2} \nabla g^\top \right\|_F^2 + \frac{\sigma_e^2}{\sigma_d^4} (\nabla f^\top) : (\nabla \Delta f^\top) + \frac{\sigma_e^2}{\sigma_d^4} \|\nabla^2 f^\top\|_F^2 + o(\sigma_e^2),
\end{aligned}$$

where $((\nabla^2 f^\top)^\top \epsilon_e)_{ij} := \sum_{k=1..d_z} (\epsilon_e)_k \partial_{z_k} \partial_{z_j} f_i$, $(\epsilon_e^\top (\nabla^3 f^\top)^\top \epsilon_e)_{ij} := \sum_{k,k'=1..d_z} (\epsilon_e)_k (\epsilon_e)_{k'} \partial_{z_k} \partial_{z_{k'}} \partial_{z_j} f_i$, and $(\nabla f^\top) : (\nabla \Delta f^\top) := \sum_{i=1..d_x, j,k=1..d_z} (\partial_{z_j} f_i) (\partial_{z_j} \partial_{z_k} f_i)$, $\|\nabla^2 f^\top\|_F^2 := \sum_{i=1..d_x, j,k=1..d_z} (\partial_{z_k} \partial_{z_j} f_i)^2$, and all terms are evaluated at $z = g(x)$. This is different from the regularization of CAE and the regularization explanation of DAE.

For the corruption encoder, $z = g(x + \epsilon_e)$, $\epsilon_e \sim \mathcal{N}(0, \sigma_e^2 I_{d_x})$, approximations of the DAE loss Eq. (19) and the data-fitting loss of CyGen Eq. (4) (i.e., negative data likelihood loss) are similar to the above expansions except that derivatives of f are replaced with those of $f \circ g$. Particularly, from Eq. (20), we find that the conclusion in [63, 1] missed the term $-\sigma_e^2 (x - f(g(x)))^\top \Delta(f \circ g)$ that is also of order σ_e^2 . For the compatibility loss, as there is no explicit expression of $\log q(z|x)$ (unless $g(x)$ is invertible), the above expression does not hold. But anyway, it is different from CAE and DAE regularizations.

Relation to the tied weights trick. The compatibility loss also explains the ‘‘tied weights’’ trick in AE, which is widely adopted and is vital for the success of AE [58, 77, 76, 63, 1]. The trick is considered when components of x and z are binary, and a one-layer, product-of-Bernoulli encoder $q(z|x) = \prod_{j=1}^{d_z} \text{Bern}(z_j | s((W_e)_{j,:}x + (b_e)_j))$ and decoder $p(x|z) = \prod_{i=1}^{d_x} \text{Bern}(x_i | s((W_d)_{i,:}z + (b_d)_i))$ are used, where $s(l) := 1/(1 + \exp\{-l\})$ denotes the sigmoid activation function. For the encoder, we have $q(z|x) = \frac{\exp\{z^\top (W_e x + b_e)\}}{\prod_{j=1}^{d_z} (1 + \exp\{(W_e)_{j,:}x + (b_e)_j\})}$ thus $\log q(z|x) = z^\top (W_e x + b_e) - \sum_{j=1}^{d_z} \log(1 + \exp\{(W_e)_{j,:}x + (b_e)_j\})$, so $\nabla_x \nabla_z^\top \log q(z|x) = W_e^\top$. Similarly for the decoder, we have $\nabla_x \nabla_z^\top \log p(x|z) = W_d$. So the compatibility loss Eq. (2) in this case is $\|W_d - W_e^\top\|_F^2$, which leads to $W_d = W_e^\top$ when it is zero. This recovers the tied weight trick.

In this Bernoulli case, the CAE regularizer is $\mathbb{E}_{p^*(x)} \sum_{j=1}^{d_z} \frac{\exp\{-2((W_e)_{j,:}x + (b_e)_j)\} \sum_{i=1}^{d_x} (W_e)_{ji}^2}{(1 + \exp\{-((W_e)_{j,:}x + (b_e)_j)\})^4} = \mathbb{E}_{p^*(x)} \sum_{j=1}^{d_z} s((W_e)_{j,:}x + (b_e)_j)^2 (1 - s((W_e)_{j,:}x + (b_e)_j))^2 \sum_{i=1}^{d_x} (W_e)_{ji}^2$ and DAE does not have the Jacobian-norm regularization explanation, so they are different from the compatibility loss.

D.2 Gradient estimation for flow-based models without tractable inverse

Flow-based density models. As the insight we draw from the analysis on Gaussian VAE in Sec. 3.1, it is inappropriate to implement both conditionals $p_\theta(x|z)$, $q_\phi(z|x)$ using additive Gaussian models. So we need more flexible and expressive probabilistic models that also allow explicit density evaluation (so implicit models like GANs are not suitable). Flow-based models [19, 55, 39, 5, 13] are a good choice. They also allow direct sampling with reparameterization for efficiently estimating and optimizing the data-fitting loss Eq. (4) (for which energy-based models are costly), and have been used as the inference model $q_\phi(z|x)$ of VAEs [61, 41, 75, 26]. For a connection to these prior works, we use a flow-based model also for the inference model $q_\phi(z|x)$. An additive-Gaussian likelihood model $p_\theta(x|z)$ is then allowed for learning nonlinear representations.

To define the distribution $q_\phi(z|x)$, a flow-based model uses a parameterized *invertible* differentiable transformation $z = T_\phi(e|x)$ to map a random seed e (of the same dimension d_z) following a simple base distribution $p(e)$ ¹³ (e.g., a standard Gaussian) to $\mathbb{Z} = \mathbb{R}^{d_z}$. By deliberately designed architectures, the transformation $T_\phi(\cdot|x)$ is guaranteed to be invertible, yet still being expressive, with some examples that are even universal approximators [73]. Benefited from the invertibility, the defined density can be explicitly given by the rule of change of variables [9, Thm. 17.2]:

$$q_\phi(z|x) = p(e = T_\phi^{-1}(z|x)) \left| \nabla_z T_\phi^{-\top}(z|x) \right|,$$

where $\left| \nabla_z T_\phi^{-\top}(z|x) \right|$ is the absolute value of the determinant of the Jacobian of $T_\phi^{-1}(z|x)$ (w.r.t z).

Problem for evaluating the compatibility loss. Although $T_\phi(z|x)$ is guaranteed to be invertible, in common instances computing its inverse is intractable [61, 41, 75] or costly [26, 5, 13] (however, they all guarantee an easy calculation of the Jacobian determinant $\left| \nabla_z T_\phi^{-\top}(z|x) \right|$ for efficient density evaluation). This means that density estimation of $q_\phi(z|x)$ is intractable for an arbitrary z value, but is only possible for a generated z value, whose inverse e is known in advance (the generated z is computed from this e). This however, introduces problems when computing the gradients $\nabla_x \log q_\phi(z|x)$, $\nabla_z \log q_\phi(z|x)$ for the compatibility loss (Eq. (2) or Eq. (3)).

To see this, it is important to distinguish the “formal arguments” and “actual arguments” of a function. It makes a difference when taking derivatives if the actual arguments are fed to formal arguments in an involved way. What we need is the derivatives w.r.t the formal arguments, but automatic differentiation tools (e.g., the autograd utility in PyTorch [56]) could only compute the derivatives w.r.t the actual arguments. We use capital subscripts for formal arguments and lowercase letters for actual arguments. Following this rule, we denote $\log q_{Z|X}^\phi(z|x)$ for $\log q_\phi(z|x)$ above, so $\nabla_Z \log q_{Z|X}^\phi$ denotes the gradient function that differentiates the first formal argument Z of $\log q_{Z|X}^\phi$, and similarly for $\nabla_X \log q_{Z|X}^\phi$. Then at a generated value of $z = T_\phi(e|x)$ from a random seed e , the required gradients in the compatibility loss are w.r.t to the formal arguments $\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)$ and $\nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x)$, while automatic differentiation tools could only give the gradients w.r.t the actual arguments $\nabla_e \log q_{Z|X}^\phi(T_\phi(e|x)|x)$ and $\nabla_x \log q_{Z|X}^\phi(T_\phi(e|x)|x)$, which are not the desired gradients. Note that we do not know the exact calculation rule of $\log q_{Z|X}^\phi(z|x)$ for arbitrary z and x , but can only evaluate $h^\phi(e, x) := \log q_{Z|X}^\phi(T_\phi(e|x)|x)$ from a given e and x . Automatic differentiation could only evaluate the gradients of this $h^\phi(e, x)$ but not of $\log q_{Z|X}^\phi(z|x)$.

¹³Although some flow-based models (e.g., the Sylvester flow [75]) also incorporate the dependency on x in the base distribution $\tilde{p}_\phi(\tilde{e}|x)$ (e.g., $\mathcal{N}(\tilde{e}|\mu_\phi(x), \Sigma_\phi(x))$), we can reparameterize this distribution as transformed from a “more basic” parameter-free base distribution $p(e)$ (e.g., $\mathcal{N}(0, I_{d_z})$) by an x -dependent transformation (e.g., $\tilde{e} = \mu_\phi(x) + \Sigma_\phi(x)^{1/2}e$) and concatenate this transformation to the original one as $z = T_\phi(e|x)$.

Solution. An explicit deduction is thus required for an expression of the correct gradients in terms of what automatic differentiation could evaluate. From the chain rule, we have:

$$\begin{aligned}\nabla_e h^\phi(e, x) &= \nabla_e \log q_{Z|X}^\phi(T_\phi(e|x)|x) = (\nabla_e T_\phi^\top(e|x)) (\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)), \\ \nabla_x h^\phi(e, x) &= \nabla_x \log q_{Z|X}^\phi(T_\phi(e|x)|x) = (\nabla_x T_\phi^\top(e|x)) (\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)) \\ &\quad + \nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x).\end{aligned}$$

The first equation gives one of the desired gradients: $\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x) = (\nabla_e T_\phi^\top(e|x))^{-1} (\nabla_e h^\phi(e, x))$. The term $\nabla_e h^\phi(e, x)$ can be evaluated using automatic differentiation, as mentioned. The other term, *i.e.* the Jacobian $\nabla_e T_\phi^\top(e|x)$, can also use automatic differentiation by tracking the forward flow computation $z = T_\phi(e|x)$, but it is often available in closed-form for flow-based models, as flow-based models need to evaluate its determinant anyway so the architecture is designed to give its closed-form expression.

The second equation gives an expression of the other desired gradient: $\nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x) = \nabla_x h^\phi(e, x) - (\nabla_x T_\phi^\top(e|x)) (\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x))$. Again, the first term $\nabla_x h^\phi(e, x)$ can be evaluated using automatic differentiation. The term $(\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x))$ can be evaluated using the expression we just derived above. For the rest term, *i.e.* the Jacobian $\nabla_x T_\phi^\top(e|x)$, it can also be evaluated using automatic differentiation by tracking the forward flow computation $z = T_\phi(e|x)$. For computation efficiency, this can be implemented by taking the gradient of $z = T_\phi(e|x)$ w.r.t x with the `grad_outputs` argument of `torch.autograd.grad` fed by $\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)$ (gradients w.r.t x will not be back-propagated through this $\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)$). This reduces computation complexity from $O(d_{\mathbb{X}} d_{\mathbb{Z}})$ down to $O(d_{\mathbb{X}} + d_{\mathbb{Z}})$. In summary, the desired gradients can be computed via the following expressions:

$$\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x) = (\nabla_e T_\phi^\top(e|x))^{-1} (\nabla_e h^\phi(e, x)), \quad (21)$$

$$\nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x) = \nabla_x h^\phi(e, x) - (\nabla_x T_\phi^\top(e|x)) (\nabla_Z \log q_{Z|X}^\phi(T_\phi(e|x)|x)). \quad (22)$$

Second-order differentiations for the compatibility loss can be done in a similar way.

A simplified compatibility loss. For the compatibility loss Eq. (3) in the form of Hutchinson's trace estimator, a further simplification is possible. The loss is given by:

$$C(\theta, \phi) = \mathbb{E}_{\rho(x,z)} \mathbb{E}_{p(\eta_x)} \left\| \nabla_Z (\eta_x^\top \nabla_X \log p_{X|Z}^\theta(x|z) - \eta_x^\top \nabla_X \log q_{Z|X}^\phi(z|x)) \right\|_2^2,$$

with any random vector η_x satisfying $\mathbb{E}[\eta_x] = 0$, $\text{Var}[\eta_x] = I_{d_{\mathbb{X}}}$. The reference distribution $\rho(x, z)$ can be taken as $p^*(x)q_\phi(z|x)$ for practical sampling for estimating the expectation. For a flow-based $q_\phi(z|x)$, sampling from $(x, z) \sim p^*(x)q_\phi(z|x)$ is equivalent to $(x, T_\phi(e|x)), e \sim p(e)$. So the loss can be reformulated as:

$$C(\theta, \phi) = \mathbb{E}_{p^*(x)p(e)} \mathbb{E}_{p(\eta_x)} \left\| \nabla_Z (\eta_x^\top \nabla_X \log p_{X|Z}^\theta(x|T_\phi(e|x)) - \eta_x^\top \nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x)) \right\|_2^2.$$

Note from Eq. (21), the gradient w.r.t Z is an invertible linear transformation of the gradient w.r.t e , so its norm equals zero if and only if the gradient w.r.t e has a zero norm. So to avoid this matrix inversion, we consider a simpler loss that achieves the same optimal solution:

$$\tilde{C}(\theta, \phi) := \mathbb{E}_{p^*(x)p(e)} \mathbb{E}_{p(\eta_x)} \left\| \nabla_e (\eta_x^\top \nabla_X \log p_{X|Z}^\theta(x|T_\phi(e|x)) - \eta_x^\top \nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x)) \right\|_2^2. \quad (23)$$

For additive Gaussian $p_{X|Z}^\theta$, its gradient $\nabla_X \log p_{X|Z}^\theta$ is available in closed-form. For $\nabla_X \log q_{Z|X}^\phi(T_\phi(e|x)|x)$ in the second term, it can be estimated using Eq. (22) we just developed. The subsequent gradient w.r.t e can be evaluated by automatic differentiation. So this loss is tractable to estimate and optimize.

E Experiment Details

E.1 Baseline Methods

We compare our proposed CyGen with bi-directional models (composed of both a likelihood model and an inference model) including Denoising Auto-Encoder (DAE), Variational Auto-Encoder (VAE) and BiGAN. Sketches of these models are introduced below.

DAE [77] first corrupts a real input data x with a local noise and then pass it through an encoder to define $q_\phi(z|x)$. The latent code z is then decoded to the data space by a decoder $p_\theta(x|z)$. The objective is to minimize the reconstruction error (Eq. (19)): $\mathbb{E}_{p^*(x)} \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(x|z)]$. Compared with VAE, it can be seen as the version with $\beta = 0$, *i.e.* it does not involve a prescribed prior $p(z)$. Nevertheless, optimizing the objective w.r.t ϕ may lead to undesired results. Particularly, for any given x , it may drive $q_\phi(z|x)$ to only concentrate on z values that maximizes $\log p_\theta(x|z)$. This renders incompatibility and an insufficient determinacy (see Sec. 3.2).

VAE [40] defines a joint distribution $p_\theta(x, z) = p(z)p_\theta(x|z)$ using a specified prior $p(z)$. It learns $p_\theta(x|z)$ to match data distribution $p^*(x)$ with the help of an inference model $q_\phi(z|x)$, using the Evidence Lower Bound (ELBO) objective:

$$\min_{\theta, \phi} \mathbb{E}_{p^*(x)} \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(x|z)] + \beta \mathbb{E}_{p^*(x)} [\text{KL}(q_\phi(z|x) \| p(z))]. \quad (24)$$

When $\beta = 1$, the negative objective is a lower bound of the data likelihood (evidence) $\mathbb{E}_{p^*(x)} [\log p_\theta(x)]$ where $p_\theta(x) := \int_{\mathbb{Z}} p(z) p_\theta(x|z) dz$, hence the name. Optimizing it w.r.t ϕ also drives $q_\phi(z|x)$ to the true posterior $p_\theta(z|x)$ and also makes the bound tighter. A β other than 1 is considered when there is some desideratum on the latent variable, *e.g.*, disentanglement [32].

BiGAN [20, 22]. In addition to learning the data distribution $p^*(x)$ using GAN [25], BiGAN also aims to learn a representation extractor, so it introduces an inference model $q_\phi(z|x)$ which is often deterministic (*i.e.*, a Dirac distribution). The likelihood model (generator) $p_\theta(x|z)$ defines a joint $p(z)p_\theta(x|z)$ with the help of a prescribed prior $p(z)$, while the inference model also defines a joint $p^*(x)q_\phi(z|x)$. Samples from both distributions can be easily drawn, so BiGAN seeks to match them using the GAN loss (Jensen-Shannon divergence) with the help of a discriminator $D(x, z)$. In each training step, the discriminator $D(x, z)$ is first updated on a mini-batch of $p^*(x)q_\phi(z|x)$ data $x^+ \sim p^*(x), z^+ \sim q_\phi(\cdot|x^+)$ with positive labels $y^+ = 1$ and a mini-batch of $p(z)p_\theta(x|z)$ data $x^- \sim p(z), x^- \sim p_\theta(\cdot|z^-)$ with negative labels $y^- = 0$. The goal of the training the discriminator is to minimize the binary cross entropy loss $\text{BCE}(D(x^+, z^+), y^+) + \text{BCE}(D(x^-, z^-), y^-)$. The conditional models $q_\phi(z|x)$ and $p_\theta(x|z)$ are then updated to maximize the same loss $\text{BCE}(D(x^+, z^+), y^+) + \text{BCE}(D(x^-, z^-), y^-)$.

GibbsNet [45] is also considered, which is similar to BiGAN, except that BiGAN’s prior-driven joint sample generation $z^- \sim p(z), x^- \sim p(x|z^-)$ is modified to run through multiple cycles: $z_0 \sim p(z), x_0 \sim p(x|z_0), z_1 \sim q(z|x_0), x_1 \sim p(x|z_1), \dots, z^- \sim q(z|x_{K-1}), x^- \sim p(x|z^-)$. This resembles a Gibbs chain, and is considered in GibbsNet for reducing the influence of a specified prior, as the stationary distribution of the Markov chain does not rely on the initial distribution but is determined by the two conditional models (see Sec. 1.1 (paragraph 2) for its limitation). But this iterated application of the likelihood and inference models makes gradient back-propagation involved. The gradient accumulates with the cycling iteration, resulting in a scale much larger than usual. This makes gradient-based optimization unstable and even leads to numerical instability. We did not find a reasonable result in any experiment using the same architecture so we omit the comparison.

E.2 Model Architecture

Our code is developed based on the repositories of the Sylvester flow¹⁴ [75] and FFJORD¹⁵ [26] for the task environment and flow architectures. VAE, DAE and CyGen share the same architecture of $p_\theta(x|z)$ and of $q_\phi(z|x)$, which are detailed in Table 16. The inference model $q_\phi(z|x)$ adopts the architecture of Sylvester flow [75], illustrated in Fig. 12. It consists of a neural network (denoted as C-QNN) that outputs q_{nn} , a reparameterization module, and a set of consecutive N flows. The outputs q_μ and q_σ are used to parameterize the diagonal Gaussian distribution for initializing z_0 , the input to the flows. For implementation simplicity, we choose the Householder version of the Sylvester flow.

¹⁴<https://github.com/rianevdborg/sylvester-flows>

¹⁵<https://github.com/rtqichen/ffjord>

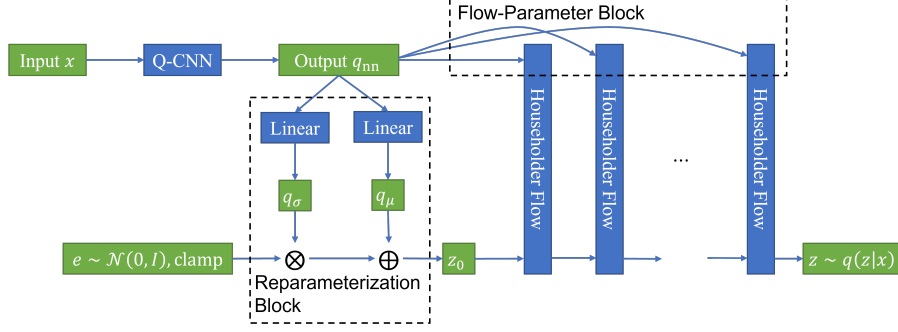


Figure 12: Flow architecture of the inference model $q_\phi(z|x)$. See Table 16 for detailed specification.

For each flow layer, the output \mathbf{z}_t of the flow given input \mathbf{z}_{t-1} is:

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \mathbf{A}_t h(\mathbf{B}_t \mathbf{z}_{t-1} + \mathbf{b}_t),$$

where $\mathbf{A}_t = \mathbf{Q}_t \mathbf{R}_t$, $\mathbf{B}_t = \tilde{\mathbf{R}}_t \mathbf{Q}_t$ and \mathbf{b}_t are parameters of the t -th flow and h is the hyperbolic-tangent activation function. Let $\mathbf{A} = \mathbf{Q}\mathbf{R}$, $\mathbf{B} = \tilde{\mathbf{R}}\mathbf{Q}^\top$, where \mathbf{R} and $\tilde{\mathbf{R}}$ are upper triangular matrices, and $\mathbf{Q} = \prod_{i=1}^H (\mathbf{I} - \frac{2\mathbf{v}_i \mathbf{v}_i^\top}{\mathbf{v}_i^\top \mathbf{v}_i})$ is a sequence of $H = 8$ Householder transformations. All the flow parameters, *i.e.* $\mathbf{v}_{1:H}$, \mathbf{b} , \mathbf{R} and $\tilde{\mathbf{R}}$, depend on q_{nn} via a flow-parameter block.

E.3 The Synthetic Dataset

The synthetic datasets (“pinwheel” in the main text and “8gaussians” in this appendix) are adopted from the above mentioned repositories of the Sylvester flow and FFJORD. The dimension of the data space is $d_x = 2$, and we take the latent space to be of the same dimension $d_z = 2$.

For the inference model $q_\phi(z|x)$, we use a three-layer MLP for the C-QNN component, with 8 hidden nodes in each layer. After the reparameterization block, consecutive $N = 32$ Householder flow layers are concatenated. Each flow layer has $H = 2$ Householder transformations¹⁶. For the likelihood model $p_\theta(x|z)$, it is implemented as an additive Gaussian model. Its mean function is a three-layer MLP with 16 hidden nodes in each layer, and its variance is taken isotropic with fixed scale 0.01.

For training, we use the Adam optimizer [38] with batch-size 1000 and weight decay parameter 1×10^{-5} for all methods. All methods use a learning rate of 1×10^{-3} except for DAE which uses 1×10^{-4} . For BiGAN, the generator is updated once per 128 updates of the discriminator using step size 1×10^{-4} . For CyGen, conditional models $p_\theta(x|z)$ and $q_\phi(z|x)$ are trained by minimizing: $1 \times 10^{-5} \times \text{compatibility loss Eq. (23)} + \text{data-fitting loss Eq. (4)}$, where the expectation in the data-fitting loss is estimated using 16 samples from $q_\phi(z|x)$ with reparameterization [40]. For the version CyGen(PT) with PreTraining, the conditional models are first pretrained as in a VAE by minimizing the ELBO objective Eq. (24) (with $\beta = 1$) using the standard Gaussian prior for 1000 epochs, and are then trained as in CyGen by minimizing the above objective with a 10-times smaller learning rate for the likelihood model $p_\theta(x|z)$ (same learning rate for the inference model $q_\phi(z|x)$). DAE is also pretrained in this way.

For data generation, VAE and BiGAN use ancestral sampling: first draw a sample of z from the standard Gaussian prior $p(z)$, and then draw a data sample x from the likelihood model $p_\theta(x|z)$. For DAE and CyGen, they do not have a prior model for ancestral sampling. They use MCMC methods, including Gibbs sampling and SGLD (see Sec. 3.3). One difference for the synthetic experiment is that the SGLD version is done by passing through the likelihood model $p_\theta(x|z)$ with prior samples drawn via SGLD in the latent space \mathbb{Z} similar to Eq. (5):

$$z^{(t+1)} = z^{(t)} + \varepsilon \nabla_{z^{(t)}} \log \frac{q_\phi(z^{(t)}|x^{(t)})}{p_\theta(x^{(t)}|z^{(t)})} + \sqrt{2\varepsilon} \eta_z^{(t)}, \text{ where } x^{(t)} \sim p_\theta(x|z^{(t)}), \eta_z^{(t)} \sim \mathcal{N}(0, I_{d_z}), \quad (25)$$

and ε is a step size parameter, taken as 3×10^{-4} . Both Gibbs sampling and SGLD are run for 100 iterations (transition steps). Their comparison for CyGen is shown in Fig. 5, where SGLD is much better. For DAE, using Gibbs sampling and using SGLD produce similar data generation results.

¹⁶To make an invertible transformation, the number of Householder transformations H needs to be no larger than the dimension of z , which is 2 in this synthetic experiment

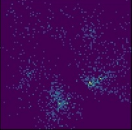
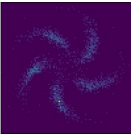
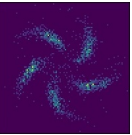
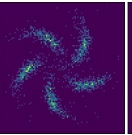
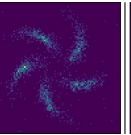
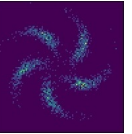
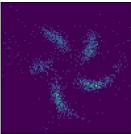
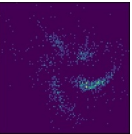
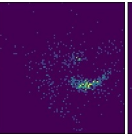
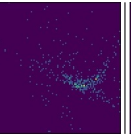
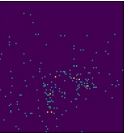
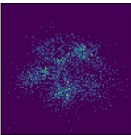
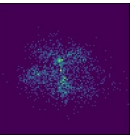
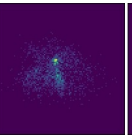
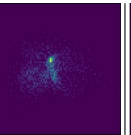
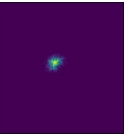
iteration 1000	iteration	1100	1200	1300	1400	30000
VAE pretraining 	CyGen(PT)					
	compt. loss	7.0×10^3	5.4×10^3	7.7×10^3	6.2×10^3	4.6×10^3
	CyGen(PT) w/o compt. loss					
compt. loss 1.6×10^4	compt. loss	1.1×10^5	1.6×10^5	2.6×10^5	8.6×10^5	1.2×10^8
	DAE					
	compt. loss	6.3×10^3	2.2×10^3	1.9×10^3	9.7×10^2	2.2×10^2

Figure 13: Generated data using \mathbb{Z} -space SGLD (Eq. (25)) along the training process after VAE pretraining (iter. 1000) using CyGen(PT), CyGen(PT) without compatibility loss, and DAE. The last DAE result is at iteration 9200, after which numerical overflow occurs.

We show more results next.

Impact of the compatibility loss. Fig. 13 shows the training process of CyGen(PT), CyGen(PT) without compatibility loss, and DAE, all after pretrained as a VAE (iter. 1000; leftmost panel), in terms of generated data distribution. We see that the normal CyGen(PT) behaves stably to the end and well approximates the data distribution along the training process. Its compatibility loss is indeed decreasing. On the other hand, CyGen(PT) without the compatibility loss diverges eventually, with an exploding compatibility loss. Although it well optimizes the data-fitting loss Eq. (4), if compatibility is not enforced, the loss is not the data likelihood that we want to optimize.

Note that CyGen(PT) without compatibility loss improves generation quality upon the VAE-pretrained model in the first few training iterations (*e.g.*, at iter. 1100). This is because the ELBO objective (Eq. (24)) of VAE also drives $q_\phi(z|x)$ towards the true posterior $p_\theta(z|x) \propto p(z)p_\theta(x|z)$ (defined with a specified prior $p(z)$) so compatibility approximately holds in the first few iterations, which makes the data-fitting loss (Eq. (4)) effective.

The collapse process of DAE. Fig. 13 (rows 1,3) also shows the comparison with DAE. We see that after pretraining, DAE quickly shrinks its data distribution, ending up with a collapsed data distribution, and even finally comes to a numerical problem. This is due to the mode-collapse behavior of the inference model $q_\phi(z|x)$ from minimizing the DAE loss Eq. (19) and the subsequent insufficient determinacy, as explained in Sec. 3.2. Although its compatibility loss is also decreasing, this comes at the cost of the insufficient determinacy which hinders capturing the data distribution and also makes the training process unstable.

Incorporating knowledge into the conditionals. We plot the prior distributions in Fig. 14 of VAE, CyGen, and CyGen(PT) with VAE pretraining, in the form of the histogram of the drawn z samples. For CyGen/CyGen(PT), samples are drawn by \mathbb{Z} -space SGLD (Eq. (25)) using the same step size $\varepsilon = 3 \times 10^{-4}$ and number of iterations 100. Compared with VAE, the priors learned by CyGen and CyGen(PT) are more expressive. For CyGen which is not subjected to any further constraints, there may be multiple $p_\theta(x|z)$ and $q_\phi(z|x)$ that are compatible and well match the given data distribution. Using a standard Gaussian prior for pretraining successfully incorporates the knowledge of a centered and centrosymmetric prior into the conditional model $p_\theta(x|z)$. The arbitrariness of possible $p_\theta(x|z)$ and $q_\phi(z|x)$ is largely mitigated in this way. This observation meets the discussion in Sec. 4.1 (paragraph 4) on the aggregated posteriors in Fig. 4.

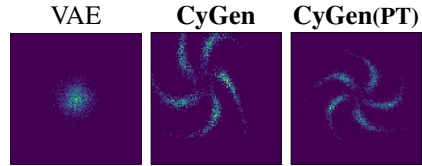
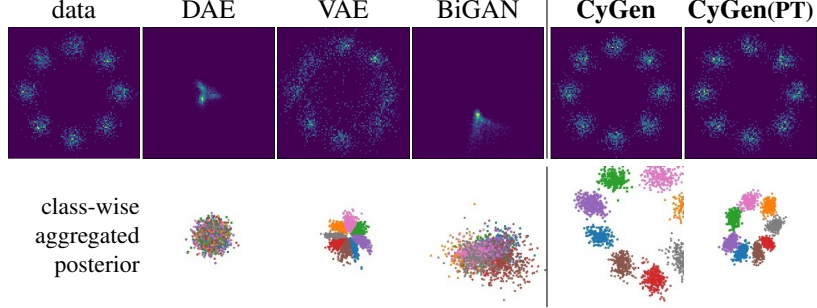


Figure 14: Prior distributions $p(z)$ of VAE, CyGen (without pretraining), and CyGen(PT) (with VAE pretraining). Prior samples of CyGen/CyGen(PT) are drawn using \mathbb{Z} -space SGLD (Eq. (25)).

Figure 15: Generated data (DAE and CyGen use \mathbb{Z} -space SGLD Eq. (25)) and class-wise aggregated posteriors of DAE, VAE, BiGAN and CyGen, on the “8gaussians” dataset. Also shows results of CyGen(PT) that is PreTrained as a VAE. (Best view in color.)



Results on another similar dataset (8gaussians). We repeat all the settings above (for the “pinwheel” dataset) on another similar synthetic dataset “8gaussians” (Fig. 15 top-left). Its data distribution also has a (nearly) non-connected support, with 8 connected components. The only difference in settings is that BiGAN uses a learning rate 1×10^{-4} for updating its generator and 3×10^{-5} for its discriminator. We see similar observations from the results in Fig. 15.

For data generation, DAE again produces a collapsed data distribution. VAE’s data distribution is blurred and the 8 clusters are touching, making the support connected due to using the standard Gaussian prior. BiGAN is unstable on this dataset and does not produce a reasonable data distribution, despite some parameter tuning. In contrast, CyGen and CyGen(PT) still recover the data distribution faithfully, with clear 8 clusters. Again, the advantage to overcome manifold mismatch is demonstrated.

For representation learning, DAE again collapses the class-wise aggregated posteriors of all classes and puts them in the same place. VAE identifies the latent clusters but which are compressed to the origin and squeezed together to border each other. BiGAN’s aggregated posteriors of different classes largely overlap each other. In contrast, CyGen and CyGen(PT) separate the latent clusters clearly. CyGen(PT) additionally embodies the knowledge from the VAE-pretrained likelihood model that the prior hence the (all-class) aggregated posterior is centered and centrosymmetric, without suffering fitting the data distribution.

In all, these observations again verify that CyGen achieves both superior generation and representation learning performances.

E.4 Real-World Datasets

Model architecture. All methods use the same architecture of the inference model (encoder) and the likelihood model (decoder), illustrated in Fig. 12 and detailed in Table 16 (except for the reported results from [45] of BiGAN and GibbsNet listed in Table 7, which are around random guess using the same flow architecture). The Gaussian variance of the likelihood model $p_\theta(x|z)$ is selected to be 0.01 for all dimensions. All methods use the Adam optimizer [38] with learning rate 1×10^{-4} and batch-size 100 for 100 epochs.

Data generation. VAE uses ancestral sampling, and DAE uses its standard Gibbs sampling procedure initialized from $z_0 \sim \mathcal{N}(0, I_{d_z})$. CyGen(PT) generates data by running \mathbb{X} -space SGLD (Eq. (5)) with step size $\varepsilon = 1 \times 10^{-3}$ initialized from $x_0 \sim p_\theta(\cdot|z_0)$ where $z_0 \sim \mathcal{N}(0, I_{d_z})$.

Downstream classification. For the downstream classification task on the latent space \mathbb{Z} , we sample one latent representation z for each data point x directly from the learned inference model $q_\phi(z|x)$, and then train a 2-layer MLP classifier with 10 hidden nodes on top of the latent representation. Results of DAE, VAE and CyGen(PT) in Table 7 are averaged over 10 random trials. All downstream classifiers are trained for 100 epochs.

Training strategy of our method. CyGen(PT) first pretrains its likelihood and inference models as in a VAE using the ELBO objective Eq. (24) (with $\beta = 1$ on MNIST and $\beta = 0.01$ on SVHN) for 100 epochs, and then trains them using CyGen ($1 \times 10^{-3} \times$ compatibility loss Eq. (23) + data-fitting loss Eq. (4)) with a 10-times smaller learning rate for the likelihood model (same learning rate for the inference model). On SVHN, we clamp the standard Gaussian random seed e in the reparameterization step that initializes z_0 to be within the interval $[-0.1, 0.1]$ (element-wise).

Remark on the VAE pretraining. For real-world images, the optimization process of CyGen from a cold start is unstable, possibly because the data distribution roughly concentrates on a low-dimensional space thus is nearly not absolutely continuous, while the CyGen-defined distribution is absolutely continuous (see Lem. C.1 in Appx. C.1). Improved techniques to handle this issue are important future work. Moreover, the VAE pretraining downweighs the effect of the prior (*i.e.*, using a small β on SVHN), so it is approximately a DAE, which does not show a reasonable result (Fig. 6). So CyGen is the key to the high-quality results.

Table 16: Inference and likelihood model architectures for MNIST and SVHN

Layers	In-Out Size	Stride
Inference Model $q_\phi(z x)$ Architecture for MNIST-C-QNN		
Input x	$1 \times 28 \times 28$	
5×5 GatedConv2d (32), Sigmoid	$32 \times 28 \times 28$	1
5×5 GatedConv2d (32), Sigmoid	$32 \times 14 \times 14$	2
5×5 GatedConv2d (64), Sigmoid	$64 \times 14 \times 14$	1
5×5 GatedConv2d (64), Sigmoid	$64 \times 7 \times 7$	2
5×5 GatedConv2d (64), Sigmoid	$64 \times 7 \times 7$	1
7×7 GatedConv2d (256), Sigmoid	$256 \times 1 \times 1$	1
Output q_{nn} , squeeze	256	
Inference Model $q_\phi(z x)$ Architecture for SVHN-C-QNN		
Input x	$3 \times 32 \times 32$	
5×5 Conv2d (32), LReLU	$32 \times 28 \times 28$	1
4×4 Conv2d (64), LReLU	$64 \times 13 \times 13$	2
4×4 Conv2d (128), LReLU	$128 \times 10 \times 10$	1
4×4 Conv2d (256), LReLU	$256 \times 4 \times 4$	2
4×4 Conv2d (512), LReLU	$512 \times 1 \times 1$	1
4×4 Conv2d (256), Sigmoid	$256 \times 1 \times 1$	1
Output q_{nn} , squeeze	256	
Reparameterization Block for $q_\phi(z x)$ for MNIST and SVHN		
Input q_{nn}	256	
Output-1 q_μ : Linear 256×64	64	
Draw $e \sim \mathcal{N}(0, I_{d_z})$ and output $z_0 = q_\mu + e \odot q_\sigma$	64	
Flow-Parameter Block for $q_\phi(z x)$ for MNIST and SVHN		
Input q_{nn}	256	
Output-1 $\mathbf{v}_{1:8}$: Linear 256×512	512	
Output-2 \mathbf{b} : Linear 256×8	512	
Output-3 \mathbf{R} : Linear $256 \times (64 \times 64)$	(64×64)	
Output-4 $\tilde{\mathbf{R}}$: Linear $256 \times (64 \times 64)$	(64×64)	
Likelihood Model $p_\theta(z x)$ Architecture for MNIST		
Input z	$64 \times 1 \times 1$	
7×7 GatedConvT2d (64), Sigmoid	$64 \times 7 \times 7$	1
5×5 GatedConvT2d (64), Sigmoid	$64 \times 7 \times 7$	1
5×5 GatedConvT2d (64), Sigmoid	$64 \times 14 \times 14$	2
5×5 GatedConvT2d (32), Sigmoid	$32 \times 14 \times 14$	1
5×5 GatedConvT2d (32), Sigmoid	$32 \times 28 \times 28$	2
5×5 GatedConvT2d (32), Sigmoid	$32 \times 28 \times 28$	1
1×1 GatedConv2d (1), Sigmoid	$1 \times 28 \times 28$	2
Output x	$1 \times 28 \times 28$	
Likelihood Model $p_\theta(z x)$ Architecture for SVHN		
Input z	$64 \times 1 \times 1$	
4×4 ConvT2d (256), LReLU	$256 \times 4 \times 4$	1
4×4 ConvT2d (128), LReLU	$128 \times 10 \times 10$	1
4×4 ConvT2d (64), LReLU	$64 \times 13 \times 13$	1
4×4 ConvT2d (32), LReLU	$32 \times 28 \times 28$	2
5×5 ConvT2d (32), LReLU	$32 \times 32 \times 32$	1
1×1 ConvT2d (32), LReLU	$32 \times 32 \times 32$	1
1×1 Conv2d (32), Sigmoid	$32 \times 32 \times 32$	1
Output x	$3 \times 32 \times 32$	