

# SUPPLEMENTARY MATERIAL: FAILURE MODES OF VARIATIONAL AUTOENCODERS AND THEIR EFFECTS ON DOWNSTREAM TASKS

**Anonymous authors**

Paper under double-blind review

## CONTENTS

<b>A</b>	<b>The Semi-Supervised VAE Training Objective</b>	<b>2</b>
<b>B</b>	<b>Proofs of Theorems</b>	<b>2</b>
B.1	Proof of Theorem 1 . . . . .	2
B.2	Proof of Theorem 2 . . . . .	3
<b>C</b>	<b>Experimental Details</b>	<b>4</b>
<b>D</b>	<b>Quantitative Results</b>	<b>5</b>
D.1	Approximation of $p(x)$ is poor when both conditions of Theorem 1 hold . . . . .	5
D.2	Failure to Learn Disentangled Representations due to Theorem 1 . . . . .	6
D.3	VAE training pathologies hinder learning compressed representations due to Theorem 1 . . . . .	6
D.4	VAEs trade-off between generating realistic data and realistic counterfactuals in semi-supervision due to Theorem 1 . . . . .	6
<b>E</b>	<b>Defense Against Adversarial Perturbations Requires the True Observation Noise and Latent Dimensionality</b>	<b>9</b>
<b>F</b>	<b>Unsupervised Pedagogical Examples</b>	<b>10</b>
F.1	Figure-8 Example . . . . .	10
F.2	Circle Example . . . . .	10
F.3	Absolute-Value Example . . . . .	11
F.4	Clusters Example . . . . .	11
F.5	Spiral Dots Example . . . . .	11
<b>G</b>	<b>Semi-Supervised Pedagogical Examples</b>	<b>12</b>
G.1	Discrete Semi-Circle Example . . . . .	12
G.2	Continuous Semi-Circle Example . . . . .	12
<b>H</b>	<b>Qualitative Results</b>	<b>14</b>
H.1	Qualitative results to support necessity of both conditions of Theorem 1 . . . . .	14
H.2	Qualitative Demonstration of Unsupervised VAE Pathologies . . . . .	16

H.3	Qualitative Demonstration of Semi-Supervised VAE Pathologies . . . . .	24
H.4	When learning compressed representations, posterior is simpler for mismatched models	30

## A THE SEMI-SUPERVISED VAE TRAINING OBJECTIVE

We extend VAE model and inference to incorporate partial labels, allowing for some supervision of the latent space dimensions. For this, we use the semi-supervised model first introduced by Kingma et al. (2014) as the “M2 model”. We assume the following generative process:

$$z \sim \mathcal{N}(0, I), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I), \quad y \sim p(y), \quad x|y, z = f_\theta(y, z) + \epsilon \quad (1)$$

where  $y$  is observed only a portion of the time. Inference objective for this model can be written as a sum of two objectives, a lower bound for the likelihood of  $M$  labeled observations and a lower bound for the likelihood for  $N$  unlabeled observations:

$$\mathcal{J}(\theta, \phi) = \sum_{n=1}^N \mathcal{U}(x_n; \theta, \phi) + \gamma \cdot \sum_{m=1}^M \mathcal{L}(x_m, y_m; \theta, \phi) \quad (2)$$

where  $\mathcal{U}$  and  $\mathcal{L}$  lower bound  $p_\theta(x)$  and  $p_\theta(x, y)$ , respectively:

$$\log p_\theta(x, y) \geq \underbrace{\mathbb{E}_{q_\phi(z|x, y)} [-\log p_\theta(x|y, z)] - \log p(y) + D_{\text{KL}}[q_\phi(z|x, y)||p(z)]}_{\mathcal{L}(x, y; \theta, \phi)} \quad (3)$$

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi(y|x)q_\phi(z|x)} [-\log p_\theta(x|y, z)] + D_{\text{KL}}[q_\phi(y|x)||p(y)] + D_{\text{KL}}[q_\phi(z|x)||p(z)]}_{\mathcal{U}(x; \theta, \phi)} \quad (4)$$

and  $\gamma$  controls their relative weight (as done by Siddharth et al. (2017)). When using IWAE, we substitute the IWAE lower bounds for  $\mathcal{U}$  and  $\mathcal{L}$  as follows:

$$\log p_\theta(x, y) \geq \underbrace{\mathbb{E}_{z_1, \dots, z_S \sim q_\phi(z|x, y)} \left[ \log \frac{1}{S} \frac{p_\theta(x, y, z_s)}{q_\phi(z_s|x, y)} \right]}_{\mathcal{L}(x, y; \theta, \phi)} \quad (5)$$

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{(y_1, z_1), \dots, (y_S, z_S) \sim q_\phi(y|x)q_\phi(z|x)} \left[ \log \frac{1}{S} \sum_{s=1}^S \frac{p_\theta(x, y_s, z_s)}{q_\phi(y_s|x)q_\phi(z_s|x)} \right]}_{\mathcal{U}(x; \theta, \phi)} \quad (6)$$

## B PROOFS OF THEOREMS

### B.1 PROOF OF THEOREM 1

Recall the decomposition the negative ELBO in Main Paper Equation 3. In the following discussion, we always set  $\phi$  to be optimal for our choice of  $\theta$ . Assuming that  $p(x)$  is continuous, then for any  $\eta \in \mathbb{R}$ , we can further decompose the PMO:

$$\begin{aligned} \mathbb{E}_{p(x)} [D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)]] &= \Pr[\mathcal{X}_{\text{Lo}}(\theta)] \mathbb{E}_{p(x)|\mathcal{X}_{\text{Lo}}(\theta)} [D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)]] \\ &\quad + \Pr[\mathcal{X}_{\text{Hi}}(\theta)] \mathbb{E}_{p(x)|\mathcal{X}_{\text{Hi}}(\theta)} [D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)]] \end{aligned} \quad (7)$$

where  $D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)] \leq \eta$  on  $\mathcal{X}_{\text{Lo}}(\theta)$ ,  $D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)] > \eta$  on  $\mathcal{X}_{\text{Hi}}(\theta)$ , with  $\mathcal{X}_i(\theta) \subseteq \mathcal{X}$ ; where  $\mathbb{E}_{p(x)|\mathcal{X}_i}$  is the expectation over  $p(x)$  restricted to  $\mathcal{X}_i(\theta)$  and renormalized, and  $\Pr[\mathcal{X}_i]$  is the probability of  $\mathcal{X}_i(\theta)$  under  $p(x)$ . Let us denote the expectation in first term on the right hand side of Equation 7 as  $D_{\text{Lo}}(\theta)$  and the expectation in the second term as  $D_{\text{Hi}}(\theta)$ .

Let  $f_{\theta_{\text{GT}}} \in \mathcal{F}$  be the ground truth likelihood function, for which we may assume that the MLE objective (MLEO) term is zero. We can now state our claim:

**Theorem.** Suppose that there exist an  $\eta \in \mathbb{R}$  such that  $\Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})] D_{\text{Hi}}(\theta_{\text{GT}})$  is greater than  $\Pr[\mathcal{X}_{\text{Lo}}(\theta_{\text{GT}})] D_{\text{Lo}}(\theta_{\text{GT}})$ . Suppose that (1) there exist an  $f_\theta \in \mathcal{F}$  such that  $D_{\text{Lo}}(\theta_{\text{GT}}) \geq D_{\text{Lo}}(\theta)$  and

$$\Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})] (D_{\text{Hi}}(\theta_{\text{GT}}) - D_{\text{Lo}}(\theta_{\text{GT}})) > \Pr[\mathcal{X}_{\text{Hi}}(\theta)] D_{\text{Hi}}(\theta) + D_{\text{KL}}[p(x)||p_\theta(x)];$$

suppose also that (2) that for no such  $f_\theta \in \mathcal{F}$  is the MLEO  $D_{\text{KL}}[p(x)||p_\theta(x)]$  equal to zero. Then at the global minima  $(\theta^*, \phi^*)$  of the negative ELBO, the MLEO will be non-zero.

*Proof.* The proof is straightforward. Condition (1) of the theorem implies that the negative ELBO of  $f_\theta$  will be lower than that of  $f_{\theta_{\text{GT}}}$ . That is, we can write:

$$-\text{ELBO}(\theta_{\text{GT}}, \phi_{\text{GT}}) = \Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})] D_{\text{Hi}}(\theta_{\text{GT}}) + \Pr[\mathcal{X}_{\text{Lo}}(\theta_{\text{GT}})] D_{\text{Lo}}(\theta_{\text{GT}}) \quad (8)$$

$$= \Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})] D_{\text{Hi}}(\theta_{\text{GT}}) + (1 - \Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})]) D_{\text{Lo}}(\theta_{\text{GT}}) \quad (9)$$

$$= \Pr[\mathcal{X}_{\text{Hi}}(\theta_{\text{GT}})] (D_{\text{Hi}}(\theta_{\text{GT}}) - D_{\text{Lo}}(\theta_{\text{GT}})) + D_{\text{Lo}}(\theta_{\text{GT}}) \quad (10)$$

$$> \underbrace{\Pr[\mathcal{X}_{\text{Hi}}(\theta)] D_{\text{Hi}}(\theta) + \Pr[\mathcal{X}_{\text{Lo}}(\theta)] D_{\text{Lo}}(\theta)}_{-\text{ELBO}(\theta, \phi)} + D_{\text{KL}}[p(x)||p_\theta(x)] \quad (11)$$

So we have that  $-\text{ELBO}(\theta_{\text{GT}}, \phi_{\text{GT}}) > -\text{ELBO}(\theta, \phi)$ . Note again that by construction  $\phi_{\text{GT}}$  and  $\phi$  are both optimal for  $\theta_{\text{GT}}$  and  $\theta$ , respectively.

Furthermore, if there is an  $f_{\theta'} \in \mathcal{F}$  such that  $-\text{ELBO}(\theta', \phi') < -\text{ELBO}(\theta, \phi)$ , then it must also satisfy the conditions in assumption (1) and, hence, the global minima of the negative ELBO satisfy the conditions in assumption (1). By assumption (2), at the global minima of the negative ELBO, the MLEO  $D_{\text{KL}}[p(x)||p_\theta(x)]$  cannot be equal to zero.  $\square$

## B.2 PROOF OF THEOREM 2

In practice, the noise variance of the dataset is unknown and it is common to estimate the variance as a hyper-parameter. Here, we show that learning the variance of  $\epsilon$  either via hyper-parameter search or via direct optimization of the ELBO can be biased.

**Theorem.** For an observation set of size  $N$ , we have that

$$\underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} -\text{ELBO}(\theta, \phi, \sigma^{(d)\epsilon^2}) = \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{q_\phi(z|x_n)} \left[ (x_n^{(d)} - f_\theta(z)^{(d)})^2 \right]. \quad (12)$$

*Proof.* We rewrite the negative ELBO:

$$\underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} -\text{ELBO}(\theta, \phi, \sigma_\epsilon^2) \quad (13)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(x|z)] + D_{\text{KL}}[q_\phi(z|x)||p(z)] \right] \quad (14)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(x|z)] \right] \quad (15)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} \left[ -\sum_{d=1}^D \log \left( \frac{1}{\sqrt{2\pi\sigma^{(d)\epsilon^2}}} \cdot \exp \left( \frac{-(x^{(d)} - f_\theta(z)^{(d)})^2}{2\sigma^{(d)\epsilon^2}} \right) \right) \right] \right] \quad (16)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \sum_{d=1}^D \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} \left[ \log \left( \sqrt{2\pi\sigma^{(d)\epsilon^2}} \right) + \frac{(x^{(d)} - f_\theta(z)^{(d)})^2}{2\sigma^{(d)\epsilon^2}} \right] \right] \quad (17)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \sum_{d=1}^D \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} \left[ \log \left( \sigma^{(d)\epsilon} \right) + \frac{(x^{(d)} - f_\theta(z)^{(d)})^2}{2\sigma^{(d)\epsilon^2}} \right] \right] \quad (18)$$

$$= \underset{\sigma^{(d)\epsilon^2}}{\operatorname{argmin}} \sum_{d=1}^D \log \left( \sigma^{(d)\epsilon} \right) + \frac{1}{2\sigma^{(d)\epsilon^2}} \cdot \underbrace{\mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} \left[ (x^{(d)} - f_\theta(z)^{(d)})^2 \right] \right]}_{C(\theta, \phi, d)} \quad (19)$$

Setting the gradient of the above with respect to  $\sigma_\epsilon^2$  equal to zero yields the following:

$$0 = -\frac{\partial}{\partial \sigma_\epsilon^{(d)}} \text{ELBO}(\theta, \phi, \sigma_\epsilon^{(d)}) \quad (20)$$

$$= \frac{\sigma_\epsilon^{(d)^2} - C(\theta, \phi, d)}{\sigma_\epsilon^{(d)^3}}. \quad (21)$$

Thus, we can write,

$$\sigma_\epsilon^{(d)^2} = C(\theta, \phi, d) = \mathbb{E}_{p(x)} \left[ \mathbb{E}_{q_\phi(z|x)} \left[ (x^{(d)} - f_\theta(z)^{(d)})^2 \right] \right] \quad (22)$$

$$\approx \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{q_\phi(z|x_n)} \left[ (x_n^{(d)} - f_\theta(z)^{(d)})^2 \right] \quad (23)$$

□

## C EXPERIMENTAL DETAILS

**Initialization at Global Optima of the VAE Objective** The decoder function  $f_\theta$  is initialized to the ground-truth using full supervision given the ground-truth  $z$ 's and  $f_{\theta_{\text{GT}}}$ . The encoder is initialized to  $\phi_{\text{GT}}$  by fixing the decoder at the ground-truth and maximizing the ELBO (with the 10 random restarts). We fix the observation error  $\sigma_\epsilon^2$  to that of the ground truth model, and we fix a sufficiently flexible architecture – one that is significantly more expressive than needed to capture  $f_{\theta_{\text{GT}}}$  – to ensure that, if there exists a  $f_\theta$  with simpler posteriors, it would be included in our feasible set  $\mathcal{F}$ . Lastly, we select the restart that yields the lowest value of the objective function.

**Synthetic Datasets** We use 4 synthetic data-sets for unsupervised VAEs (described in Appendix F), and 2 synthetic data-sets for semi-supervised VAEs (described in Appendix G), and generate 5 versions of each data-set (each with 5000/2000/2000 train/validation/test points). We use 3 real semi-supervised data-sets: Diabetic Retinopathy Debrecen (Antal & Hajdu, 2014), Contraceptive Method Choice (Alcala-Fdez et al., 2010; Dua & Graff, 2017) and the Titanic (Alcala-Fdez et al., 2010; Simonoff, 1997) datasets, each with 10% observed labels, split in 5 different ways equally into train/validation/test.

**Real Datasets** We consider 3 UCI data-sets: Diabetic Retinopathy Debrecen (Antal & Hajdu, 2014), Contraceptive Method Choice (Alcala-Fdez et al., 2010; Dua & Graff, 2017) and the Titanic (Alcala-Fdez et al., 2010; Simonoff, 1997) datasets. In these, we treat the outcome as a partially observed label (observed 10% of the time). We split the data 5 different ways into equally sized train/validation/test. On each split of the data, we run 5 random restarts and select the run that yielded the best value on the training objective, computed on the validation set.

**Evaluation Metrics** To evaluate the quality of the generative model, we use the smooth  $k$ NN test statistic (Djolonga & Krause, 2017) on samples from the learned model vs. samples from the training set / ground truth model as an alternative to log-likelihood, since log-likelihood has been shown to be problematic for evaluation because of its numerical instability / high variance (Theis et al., 2016; Wu et al., 2017). In the semi-supervised case, we also use the smooth  $k$ NN test statistic to compare  $p(x|y)$  with the learned  $p_\theta(x|y)$ . Finally, in cases where we may have model mismatch, we also evaluate the mutual information between  $x$  and each dimension of the latent space  $z$ , using the estimator presented in (Kraskov et al., 2004).

**Architectures** On the synthetic data-sets, we use a leaky-ReLU encoder/decoder with 3 hidden layers, each 50 nodes. On the UCI data-sets, we use a leaky-ReLU encoder/decoder with 3 hidden layers, each 100 nodes.

**Optimization** For optimization, we use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and a mini-batch size of 100. We train for 100 epochs on synthetic data and for 20000 on real data (and verified convergence). We trained 5 random restarts on each of the split of the data.



For semi-supervised data-sets with discrete labels, we used continuous relaxations of the categorical distribution with temperature 2.2 (Jang et al., 2016) as the variational family in order to use the reparameterization trick (Kingma & Welling, 2013).

**Baselines** For our baselines, we compare the performance of a MFG-VAE with that of a VAE trained with the Lagging Inference Networks (LIN) algorithm (still with a MFG variational family), since the algorithm claims to be able to escape local optima in training. Since the pathologies we describe are global optima, we do not expect LIN to mitigate the issues. We use Importance Weighted Autoencoders (IWAE) as an example of a inference algorithm that uses a more complex variational family. Since the pathologies described are exacerbated by a limited variational family, we expect IWAE to out-perform the other two approaches. For each method, we select the hyper-parameters for which the best restart yields the best log-likelihood (using the smooth  $k$ NN test-statistic, described below).

**Hyper-parameters** When using IWAE, let  $S$  be the number of importance samples used. When using the Lagging Inference Networks, let  $T$  be the threshold for determining whether the inference network objective has converged, and let  $R$  be the number of training iterations for which the loss is averaged before comparing with the threshold. When using semi-supervision,  $\alpha$  determines the weight of the discriminator, and  $\gamma$  determines the weight of the labeled objective,  $\mathcal{L}$ . We grid-searched over all combination of the following sets of parameters:

**Unsupervised datasets:**

- IWAE:  $S \in \{3, 10, 20\}$
- Lagging Inference Networks:  $T \in \{0.05, 0.1\}$ ,  $R \in \{5, 10\}$

**Semi-supervised synthetic datasets:**

- IWAE:  $S \in \{3, 10, 20\}$
- Lagging Inference Networks:  $T \in \{0.05, 0.1\}$ ,  $R \in \{5, 10\}$
- All methods:  $\alpha \in \{0.0, 0.1, 1.0\}$ ,  $\gamma \in \{0.5, 1.0, 2.0, 5.0\}$

**Semi-supervised real datasets:**

- IWAE:  $S \in \{3, 10, 20\}$
- Lagging Inference Networks:  $T \in \{0.05, 0.1\}$ ,  $R \in \{5, 10\}$
- All methods:  $\alpha \in \{0.0, 0.1, 1.0\}$ ,  $\gamma \in \{0.5, 1.0, 2.0, 5.0\}$ ,  $\sigma_\epsilon^2 \in \{0.01, 0.5\}$ . On Titanic dimensionality of  $z$  is in  $\{1, 2\}$ , on Contraceptive and Diabetic Retinopathy  $\in \{2, 5\}$ .

**Hyper-parameters Selection** For each method, we selected the hyper-parameters that yielded the smallest value of the smooth  $k$ NN test statistic (indicating that they learned the  $p(x)$  best).

## D QUANTITATIVE RESULTS

In this section we present the quantitative results for the paper. For all data-sets, we fix a sufficiently flexible architecture (one that is significantly more expressive than needed to capture  $f_{\theta_{\text{GT}}}$ ) so that our feasible set  $\mathcal{F}$  is diverse enough to include likelihoods with simpler posteriors. For the synthetic data-sets, we then train each model to reach the global optima as follows: we train 10 restarts for each method and hyper-parameter settings – 5 random where we initialize randomly, and 5 random where the decoder and encoder are initialized to ground truth values. We select the restart with the best value of the objective function. See Appendix C for a complete detail on the experimental setup.

### D.1 APPROXIMATION OF $p(x)$ IS POOR WHEN BOTH CONDITIONS OF THEOREM 1 HOLD

Here we show that on data-sets for which Theorem 1 holds, VAEs approximate  $p(x)$  poorly. We do this on two data-sets, the “Figure-8” and the “Clusters” Examples (described in Appendices F.1 and

F.4, respectively). Table 1 shows that these data-sets, VAEs (even with a better training algorithm, LIN) approximate  $p(x)$  poorly, while methods with a more complex variational family (like IWAE) do not. Visualization of the posterior (in Appendix H.1) confirm that the VAE objective under-fits the generative model in order to learn a simpler posterior, whereas the IWAE objective does not: for the “Figure-8 Example” see Figures 4, 5 and 6, and for the “Clusters Example” see Figures 7, 8 and 9. In these two examples, we further see the ELBO’s regularizing effect on the learned  $f_\theta$ . On the “Figure-8 Example”, the learned  $f_\theta$  ensures that  $x$ ’s generated from  $z \in [-\infty, -3] \cup [3, \infty]$  are sufficiently different from  $x$ ’s generated from  $z \approx 0$ :  $f_\theta(z)$  curls away from the center  $z \approx 0$  and thus simplifies the posterior. On the “Clusters Examples”, the learned  $f_\theta$  has less pronounced changes in slope, and thus a simpler posterior.

## D.2 FAILURE TO LEARN DISENTANGLED REPRESENTATIONS DUE TO THEOREM 1

In disentangled representation learning, we suppose that each dimension of the latent space corresponds to a task-meaningful concept (Ridgeway, 2016; Chen et al., 2018). Our goal is to infer these meaningful ground truth latent dimensions. It’s been noted in literature that this inference problem is ill-posed - that is, there are an infinite number of likelihood functions (and hence latent codes) that can capture  $p(x)$  equally well (Locatello et al., 2018). Here, we show that, more problematically, the VAE objective can *prefer* learning the representations that *entangles* the ground-truth latent dimensions due to the pathology in Theorem 1.

Consider data generated by  $f_{\theta_{\text{GT}}}(z) = Az + b$ . If  $A$  is non-diagonal, then the posteriors of this model are correlated Gaussians (poorly approximated by MFGs). Let  $A' = AR$ , where we define  $R = (\Sigma V^\top)^{-1}(\Lambda - \sigma_\epsilon^2 I)^{1/2}$  with an arbitrary diagonal matrix  $\Lambda$  and matrices  $\Sigma, V$  taken from the SVD of  $A$ ,  $A = U\Sigma V^\top$ . In this case,  $f_\theta = A'z + b$  has the same marginal likelihood as  $f_{\theta_{\text{GT}}}$ , that is,  $p_\theta(x) = p_{\theta_{\text{GT}}}(x) = \mathcal{N}(b, \sigma_\epsilon^2 \cdot I + AA^\top)$ . However, since the posteriors of  $f_\theta$  are uncorrelated, the ELBO will prefer  $f_\theta$  over  $f_{\theta_{\text{GT}}}$ ! In the latent space corresponding to  $f_\theta$ , the original *interpretations* of the latent dimensions are now entangled.

Similarly, for more complicated likelihood functions, we expect the ELBO to prefer learning models with simpler posteriors which are not necessarily ones that are useful for constructing disentangled representations. This bias is reduced in the IWAE training objective.

## D.3 VAE TRAINING PATHOLOGIES HINDER LEARNING COMPRESSED REPRESENTATIONS DUE TO THEOREM 1

In practice, if the task does not require a specific latent space dimensionality,  $K$ , one chooses  $K$  that maximizes the  $\log p_\theta(x)$ . Note that using a higher  $K$  and a lower  $\sigma_\epsilon^2$  means we can capture the data distribution with a simpler function  $f_\theta(z)$  and hence get simpler posteriors. That is, increasing  $K$  alleviates the need to compromise the generative model in order to improve the inference model and leads to better approximation of  $p(x)$ . Thus, the ELBO will favor model mismatch ( $K$  larger than the ground truth) and prevent us from learning highly compressed representations when they are available.

We demonstrate this empirically by embedding the “Figure-8” and “Clusters” Examples into a 5D space using a linear transformation,  $A = \begin{pmatrix} 1.0 & 0.0 & 0.5 & 0.2 & -0.8 \\ 0.0 & 1.0 & -0.5 & 0.3 & -0.1 \end{pmatrix}$ , and then training a VAE with latent dimensionality  $K \in \{1, 2, 3\}$ , with  $K = 1$  corresponding to the ground-truth model. Training for  $K = 1$  is initialized at the ground truth model, and for  $K > 2$  we initialize randomly; in each case we optimize  $\sigma_\epsilon^2$  per-dimension to minimize the negative ELBO. The ELBO prefers models with larger  $K$  over the ground truth model ( $K = 1$ ), and that as  $K$  increases, the average informativeness of each latent code decreases (Table 2), since the latent space learns to generate the observation noise  $\epsilon$ . We confirm that the posteriors become simpler as  $K$  increases, lessening the incentive for the VAE to compromise on approximating  $p(x)$  (Figure 18). Lastly, we confirm that while LIN also shows preference for higher  $K$ ’s, IWAE does not (Table 2).

## D.4 VAES TRADE-OFF BETWEEN GENERATING REALISTIC DATA AND REALISTIC COUNTERFACTUALS IN SEMI-SUPERVISION DUE TO THEOREM 1

**Trade-offs when labels are discrete** The trade-off between realistic data and realistic counterfactuals generation is demonstrated in the “Discrete Semi-Circle” Example, visualized in Figure 11 (details

in Appendix G.1). The VAE is able to learn the data manifold and distribution well (Figure 11a). However, the learned model has a simple posterior in comparison to the true posterior (Figure 11f). In fact, the learned  $f_\theta(z, y)$  is collapsed to the same function for all values of  $y$  (Figure 11b). As a result,  $p_\theta(x|y) \approx p_\theta(x)$  under the learned model (Figure 11c). We call this phenomenon “functional collapse”. As expected, functional collapse occurs when training with LIN as well (Figure 12). In contrast, IWAE is able to learn two distinct data conditionals  $p_\theta(x|y)$ , but it does so at a cost. *Since IWAE does not regularize the generative model, it overfits* (Figure 13). Table 3 shows that IWAE learns  $p(x)$  worse than the VAE, while Table 4 shows that it learns  $p(x|y)$  significantly better. We see a similar pattern in the real data-sets (see Tables 5 and 6).

**Trade-offs when labels are continuous** When  $y$  is discrete, we can lower-bound the number of modes of  $p_\theta(z|x)$  by the number of distinct values of  $y$ , and choose a variational family that is sufficiently expressive. But when  $y$  is continuous, we cannot easily bound the complexity of  $p_\theta(z|x)$ . In this case, we show that the same trade-off between realistic data and realistic counterfactuals exists, and that there is an *additional* pathology introduced by the discriminator  $q_\phi(y|x)$  (Equation 2). Consider the “Continuous Semi-Circle” Example, visualized in Figure 14b (details in Appendix G.2). Here, since the posterior  $p_\theta(y|x)$  is bimodal, encouraging the MFG discriminator  $q_\phi(y|x)$  to be predictive will collapse  $f_\theta(y, z)$  to the same function for all  $y$  (Figure 14b). So as we increase  $\alpha$  (the priority placed on prediction), our predictive accuracy increases at the cost of collapsing  $p_\theta(x|y)$  towards  $p_\theta(x)$ . The latter will result in low quality counterfactuals (see Figure 14c). Like in the discrete case,  $\gamma$  still controls the tradeoff between realistic data and realistic counterfactuals; in the continuous case,  $\alpha$  *additionally* controls the tradeoff between realistic counterfactuals and predictive accuracy. Table 4 shows that IWAE is able to learn  $p(x)$  better than VAE and LIN, as expected, but *the naive addition of the discriminator to IWAE means that it learns  $p(x|y)$  no better than the other two models* (see below for an explanation); that is, with the naive discriminator, just like the VAE and LIN, IWAE suffers from functional collapse (see Figure 16).

**Naive application adaptation of IWAE for semi-supervision introduces new pathologies.** The variational family implied by the IWAE objective is not the one given by the IWAE decoder  $q_\phi$  (Cremer et al., 2017). As such, incorporating a discriminator term  $q_\phi(y|x)$  into an IWAE semi-supervised objective is non-trivial, since the real approximate variational family used is complex and requires intractable marginalization over  $z$ . Although some get around this intractability by working with lower bounds (Siddharth et al., 2017) on  $q_\phi(y, z|x)$  marginalized over  $z$ , the discriminator in these cases is nonetheless different from the variational posterior. This may be an additional factor of the poor performance of IWAE in the semi-supervised setting with continuous  $y$ .

Data	IWAE	LIN	VAE
Clusters	<b>0.057 ± 0.028</b>	0.347 ± 0.057	0.361 ± 0.083
Fig-8	<b>0.036 ± 0.013</b>	0.040 ± 0.081	0.066 ± 0.014

Table 1: Comparison unsupervised learned vs. true data distributions via the smooth  $k$ NN test (lower is better). Hyper-parameters selected via smaller value of the loss function on the validation set.

VAE	Figure-8 Example			Clusters Example		
	$K = 1$ (ground-truth)	$K = 2$	$K = 3$	$K = 1$ (ground-truth)	$K = 2$	$K = 3$
Test – ELBO	−0.127 ± 0.057	−0.260 ± 0.040	<b>−0.234 ± 0.050</b>	4.433 ± 0.049	4.385 ± 0.034	<b>4.377 ± 0.024</b>
Test avg <sub>i</sub> $I(x; z_i)$	<b>2.419 ± 0.027</b>	1.816 ± 0.037	1.296 ± 0.064	<b>1.530 ± 0.011</b>	1.425 ± 0.019	1.077 ± 0.105

IWAE	Figure-8 Example			Clusters Example		
	$K = 1$ (ground-truth)	$K = 2$	$K = 3$	$K = 1$ (ground-truth)	$K = 2$	$K = 3$
Test – ELBO	<b>−0.388 ± 0.044</b>	−0.364 ± 0.051	−0.351 ± 0.045	<b>4.287 ± 0.047</b>	4.298 ± 0.054	4.295 ± 0.049
Test avg <sub>i</sub> $I(x; z_i)$	<b>2.159 ± 0.088</b>	1.910 ± 0.035	1.605 ± 0.087	1.269 ± 0.052	<b>1.321 ± 0.033</b>	1.135 ± 0.110

Table 2: The ELBO prefers learning models with more latent dimensions (and smaller  $\sigma_\epsilon^2$ ) over the ground truth model ( $k = 1$ ). Although the models preferred by the ELBO have a higher mutual information between the data and learned  $z$ ’s, the mutual information between dimension of  $z$  and the data decreases since with more latent dimensions, the latent space learns  $\epsilon$ . In contrast, IWAE does not suffer from this pathology. LIN was not included here because it was not able to minimize the negative ELBO as well as the VAE on these data-sets.

Data	IWAE	LIN	VAE
Discrete Semi-Circle	$0.694 \pm 0.096$	$0.703 \pm 0.315$	<b><math>0.196 \pm 0.078</math></b>
Continuous Semi-Circle	<b><math>0.015 \pm 0.011</math></b>	$0.128 \pm 0.094$	$0.024 \pm 0.014$

Table 3: Comparison of semi-supervised learned vs. true data distributions via the smooth  $k$ NN test (lower is better). Hyper-parameters selected via the smooth  $k$ NN test-statistic computed on the data marginals.

Data	IWAE		LIN		VAE	
	Cohort 1	Cohort 2	Cohort 1	Cohort 2	Cohort 1	Cohort 2
Discrete Semi-Circle	<b><math>1.426 \pm 1.261</math></b>	<b><math>1.698 \pm 0.636</math></b>	$18.420 \pm 1.220$	$10.118 \pm 0.996$	$15.206 \pm 1.200$	$11.501 \pm 1.300$
Continuous Semi-Circle	$15.951 \pm 3.566$	<b><math>14.416 \pm 1.402</math></b>	$15.321 \pm 1.507$	$17.530 \pm 1.509$	<b><math>13.128 \pm 0.825</math></b>	$16.046 \pm 1.019$

Table 4: Comparison of semi-supervised learned  $p_\theta(x|y)$  with ground truth  $p(x|y)$  via the smooth  $k$ NN test statistic (smaller is better). Hyper-parameters selected via smallest smooth  $k$ NN test statistic computed on the data marginals. For the discrete data, the cohorts are  $p(x|y = 0)$  and  $p(x|y = 1)$ , and for the continuous data, the cohorts are  $p(x|y = -3.5)$  and  $p(x|y = 3.5)$ .

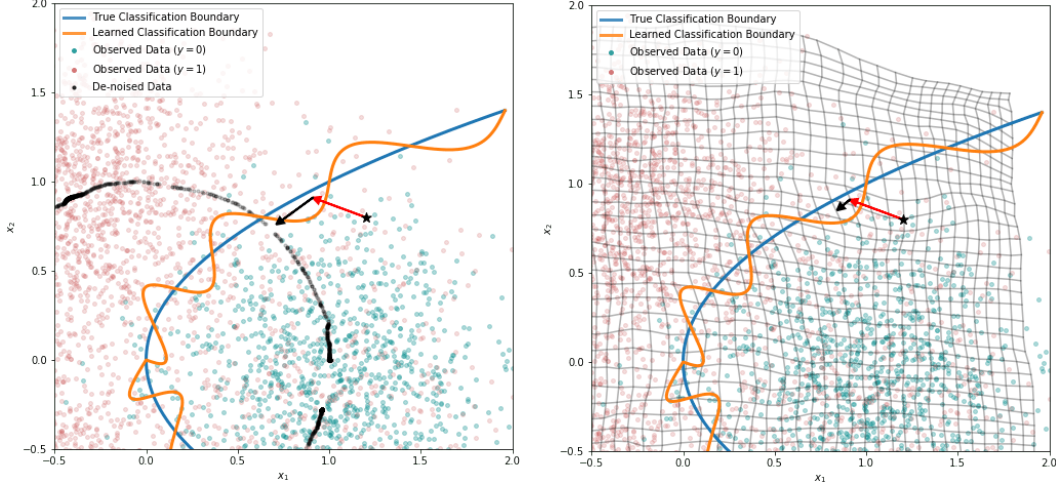
	IWAE	VAE
Diabetic Retinopathy	$3.571 \pm 2.543$	$6.206 \pm 1.035$
Contraceptive	$1.740 \pm 0.290$	$2.147 \pm 0.225$
Titanic	$2.794 \pm 1.280$	$1.758 \pm 0.193$

Table 5: Comparison of semi-supervised learned vs. true data distributions via the smooth  $k$ NN test (lower is better). Hyper-parameters selected via the smooth  $k$ NN test-statistic computed on the data marginals.

	IWAE			VAE		
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
Diabetic Retinopathy	$4.240 \pm 1.219$	$4.357 \pm 3.417$	N/A	$5.601 \pm 0.843$	$8.008 \pm 1.096$	N/A
Contraceptive	$7.838 \pm 1.138$	$5.521 \pm 3.519$	$6.626 \pm 2.571$	$5.388 \pm 0.788$	$4.994 \pm 0.932$	$3.722 \pm 0.488$
Titanic	$3.416 \pm 0.965$	$6.923 \pm 1.924$	N/A	$3.730 \pm 0.866$	$8.572 \pm 1.766$	N/A

Table 6: Comparison of semi-supervised learned vs. true conditional distributions  $p(x|y)$  via the smooth  $k$ NN test (lower is better). Hyper-parameters selected via the smooth  $k$ NN test-statistic computed on the data marginals.

## E DEFENSE AGAINST ADVERSARIAL PERTURBATIONS REQUIRES THE TRUE OBSERVATION NOISE AND LATENT DIMENSIONALITY



(a) Projection of adversarial example onto true manifold. (b) Projection of adversarial example onto manifold learned given model mismatch (higher dimensional latent space and smaller observation noise).

Figure 1: Comparison of projection of adversarial example onto ground truth vs. learned manifold. The star represents the original point, perturbed by the red arrow, and then projected onto the manifold by the black arrow.

As a defense against adversarial attacks, manifold-based approaches de-noise the data before feeding to a classifier with the hope that the de-noising will remove the adversarial perturbation from the data (Jalal et al., 2017; Meng & Chen, 2017; Samangouei et al., 2018; Hwang et al., 2019; Jang et al., 2020). In this section we argue that a correct decomposition of the data into  $f_\theta(z)$  and  $\epsilon$  (or “signal” and “noise”) is necessary to prevent against certain perturbation-based adversarial attacks.

Assume that our data was generated as follows:

$$\begin{aligned} z &\sim p(z) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &\sim f_{\theta_{\text{GT}}}(z) + \epsilon \\ y|z &\sim \text{Cat}(g_\psi \circ f_{\theta_{\text{GT}}}(z)) \end{aligned} \quad (24)$$

Let  $\mu_\phi(x)$  denote the mean of encoder and let  $M_{\theta,\phi}(x) = f_\theta \circ \mu_\phi(x)$  denote a projection onto the manifold. Our goal is to prevent adversarial attacks on a given discriminative classifier that predicts  $y|x$  – that is, we want to ensure that there does not exist any  $\eta$  such that  $x_n + \eta$  is classified with a different label than  $y_n$  by the learned classifier and not by the ground truth classifier. Since the labels  $y$  are computed as a function of the de-noised data,  $f_{\theta_{\text{GT}}}(z)$ , the true classifier is only defined on the manifold  $M$  (marked in blue in Figure 1). As such, any learned classifier (in orange) will intersect the true classifier on  $M$ , but may otherwise diverge from it away from the manifold. This presents a vulnerability against adversarial perturbations, since now any  $x$  can be perturbed to cross the learned classifier’s boundary (in orange) to flip its label, while its true label remains the same, as determined by the true classifier (in blue). To protect against this vulnerability, existing methods de-noise the data by projecting it onto the manifold before classifying. Since the true and learned classifiers intersect on the manifold, in order to flip an  $x$ ’s label, the  $x$  must be perturbed to cross the true classifier’s boundary (and not just the learned classifier’s boundary). This is illustrated in Figure 1a: the black star represents some data point, perturbed (by the red arrow) by an adversary to cross the learned classifier’s boundary but not the true classifier’s boundary. When projected onto the manifold (by the black arrow), the adversarial attack still falls on the same side of the true classifier and the learned classifier, rendering the attack unsuccessful and this method successful.

However, if the manifold is not estimated correctly from the data (i.e. if the ground truth dimensionality of the latent space and the observation noise  $\sigma_\epsilon^2$  are poorly estimated), this defense may fail. Consider, for example, the case in which  $f_\theta(z)$  is modeled with a VAE with a larger dimensional latent space and a smaller observation noise than the ground truth model. Figure 1b shows a uniform grid in  $x$ 's space projected onto the manifold learned by this mismatched model. The figure shows that the learned manifold barely differs from the original space, since the latent space of the VAE compensates for the observation noise  $\epsilon$  and thus does not de-noise the observation. When the adversarial attack is projected onto the manifold, it barely moves and is thus left as noisy. As the figure shows, the attack crosses the learned classifier's boundary but not the true boundary and is therefore successful.

## F UNSUPERVISED PEDAGOGICAL EXAMPLES

In this section we describe in detail the unsupervised pedagogical examples used in the paper and the properties that cause them to trigger the VAE pathologies. For each one of these example decoder functions, we fit a surrogate neural network  $f_\theta$  using full supervision (ensuring that the  $\text{MSE} < 1\text{e}-4$  and use that  $f_\theta$  to generate the actual data used in the experiments.

### F.1 FIGURE-8 EXAMPLE

**Generative Process:**

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= (0.6 + 1.8 \cdot \Phi(z)) \pi \\ x|z &= \underbrace{\begin{bmatrix} \frac{\sqrt{2}}{2} \cdot \frac{\cos(u(z))}{\sin(u(z))^2 + 1} \\ \sqrt{2} \cdot \frac{\cos(u(z)) \sin(u(z))}{\sin(u(z))^2 + 1} \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \quad (25)$$

where  $\Phi(z)$  is the Gaussian CDF and  $\sigma_\epsilon^2 = 0.02$  (see Figure 4).

**Properties:** In this example, values of  $z$  on  $[-\infty, -3.0]$ ,  $[3.0, \infty]$  and in small neighborhoods of  $z = 0$  all produce similar values of  $x$ , namely  $x \approx 0$ ; as such, the true posterior  $p_{\theta_{\text{GT}}}(z|x)$  is multi-modal in the neighborhood of  $x = 0$  (see Figure 4d), leading to high PMO. Additionally, in the neighborhood of  $x \approx 0$ ,  $p(x)$  is high. Thus, condition (1) of Theorem 1 is satisfied. One can verify condition (2) is satisfied by considering all continuous parametrizations of a figure-8 curve. Any such parametrization will result in a  $f_\theta$  for which far-away values of  $z$  lead to nearby values of  $x$  and thus in high PMO value for points near  $x = 0$ .

### F.2 CIRCLE EXAMPLE

**Generative Process:**

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &= \underbrace{\begin{bmatrix} \cos(2\pi \cdot \Phi(z)) \\ \sin(2\pi \cdot \Phi(z)) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \quad (26)$$

where  $\Phi(z)$  is the Gaussian CDF and  $\sigma_\epsilon^2 = 0.01$  (see Figure 2).

**Properties:** In this example, the regions of the data-space that have a non-Gaussian posterior are near  $x \approx [1.0, 0.0]$ , since in that neighborhood,  $z \in [-\infty, -3.0]$  and  $z \in [3.0, \infty]$  both generate nearby values of  $x$ . Thus, this model only satisfies condition 2 of Theorem 1. However, since overall the number of  $x$ 's for which the posterior is non-Gaussian are few, the VAE objective does not need to

trade-off capturing  $p(x)$  for easy posterior approximation. We see that traditional training is capable of recovering  $p(x)$ , regardless of whether training was initialized randomly or at the ground truth (see Figure 2).

### F.3 ABSOLUTE-VALUE EXAMPLE

#### Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &= \underbrace{\begin{bmatrix} \Phi(z) \\ \Phi(z) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \quad (27)$$

where  $\Phi(z)$  is the Gaussian CDF and  $\sigma_\epsilon^2 = 0.01$  (see Figure 3).

**Properties:** In this example, the posterior under  $f_{\theta_{\text{GT}}}$  cannot be well approximated using a MFG variational family (see Figure 3d). However, there does exist an alternative likelihood function  $f_\theta(z)$  (see 3b) that explains  $p(x)$  equally well and has simpler posterior 3e. As such, this model only satisfies condition 1 of Theorem 1.

### F.4 CLUSTERS EXAMPLE

#### Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= \frac{2\pi}{1 + e^{-\frac{1}{2}\pi z}} \\ t(u) &= 2 \cdot \tanh(10 \cdot u - 20 \cdot \lfloor u/2 \rfloor - 10) + 4 \cdot \lfloor u/2 \rfloor + 2 \\ x|z &= \underbrace{\begin{bmatrix} \cos(t(u(z))) \\ \sin(t(u(z))) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \quad (28)$$

where  $\sigma_\epsilon^2 = 0.2$ .

**Properties:** In this example,  $f_{\theta_{\text{GT}}}$  a step function embedded on a circle. Regions in which  $\frac{df_{\theta_{\text{GT}}}^{-1}}{dx}$  is high (i.e. the steps) correspond to regions in which  $p(x)$  is high. The interleaving of high density and low density regions on the manifold yield a multi-modal posterior (see Figure 7d). For this model, both conditions of Theorem 1 hold. In this example, we again see that the VAE objective learns a model with a simpler posterior (see Figure 7e) at the cost of approximating  $p(x)$  well (see Figure 7a).

### F.5 SPIRAL DOTS EXAMPLE

#### Generative Model:

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= \frac{4\pi}{1 + e^{-\frac{1}{2}\pi z}} \\ t(u) &= \tanh(10 \cdot u - 20 \cdot \lfloor u/2 \rfloor - 10) + 2 \cdot \lfloor u/2 \rfloor + 1 \\ x|z &= \underbrace{\begin{bmatrix} t(u(z)) \cdot \cos(t(u(z))) \\ t(u(z)) \cdot \sin(t(u(z))) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \quad (29)$$

where  $\sigma_\epsilon^2 = 0.01$ .

**Properties:** In this example,  $f_{\theta_{\text{GT}}}$  a step function embedded on a spiral. Regions in which  $\frac{df_{\theta_{\text{GT}}}^{-1}}{dx}$  is high (i.e. the steps) correspond to regions in which  $p(x)$  is high. The interleaving of high density and low density regions on the manifold yield a multi-modal posterior (see Figure 10d). In this example, we again see that the VAE objective learns a model with a simpler posterior (see Figure 10e) at the cost of approximating  $p(x)$  well (see Figure 10a). Furthermore, for this model the VAE objective highly misestimates the observation noise.

## G SEMI-SUPERVISED PEDAGOGICAL EXAMPLES

In this section we describe in detail the semi-supervised pedagogical examples used in the paper and the properties that cause them to trigger the VAE pathologies. For each one of these example decoder functions, we fit a surrogate neural network  $f_{\theta}$  using full supervision (ensuring that the  $\text{MSE} < 1e - 4$  and use that  $f_{\theta}$  to generate the actual data used in the experiments.

### G.1 DISCRETE SEMI-CIRCLE EXAMPLE

**Generative Process:**

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ y &\sim \text{Bern}\left(\frac{1}{2}\right) \\ \epsilon &\sim \mathcal{N}(0, \sigma_{\epsilon}^2 \cdot I) \\ x|y, z &= \underbrace{\begin{bmatrix} \cos\left(\mathbb{I}(y=0) \cdot \pi \cdot \sqrt{\Phi(z)} + \mathbb{I}(y=1) \cdot \pi \cdot \Phi(z)^3\right) \\ \sin\left(\mathbb{I}(y=0) \cdot \pi \cdot \sqrt{\Phi(z)} + \mathbb{I}(y=1) \cdot \pi \cdot \Phi(z)^3\right) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(y, z)} + \epsilon \end{aligned} \quad (30)$$

where  $\Phi$  is the CDF of a standard normal and  $\sigma_{\epsilon}^2 = 0.01$ .

**Properties:** We designed this data-set to specifically showcase issues with the semi-supervised VAE objective. As such, we made sure that the data marginal  $p(x)$  of this example will be learned well using unsupervised VAE (trained on the  $x$ 's only) This way we can focus on the new issues introduced by the semi-supervised objective.

For this ground-truth model, the posterior of the un-labeled data  $p_{\theta_{\text{GT}}}(z|x)$  is bimodal, since there are two functions that could have generated each  $x$ :  $f_{\theta_{\text{GT}}}(y=0, z)$  and  $f_{\theta_{\text{GT}}}(y=1, z)$ . As such, approximating this posterior with a MFG will encourage the semi-supervised objective to find a model for which  $f_{\theta_{\text{GT}}}(y=0, z) = f_{\theta_{\text{GT}}}(y=1, z)$  (see Figure 11b). When both functions collapse to the same function,  $p_{\theta}(x|y) \approx p_{\theta}(x)$  (see Figure 11c). This will prevent the learned model from generating realistic counterfactuals.

### G.2 CONTINUOUS SEMI-CIRCLE EXAMPLE

**Generative Process:**

$$\begin{aligned} z &\sim \mathcal{N}(0, 1) \\ y &\sim \mathcal{N}(0, 1) \\ h(y) &= B^{-1}(\Phi(y); 0.2, 0.2) \\ \epsilon &\sim \mathcal{N}(0, \sigma_{\epsilon}^2 \cdot I) \\ x|y, z &= \underbrace{\begin{bmatrix} \cos\left(h(y) \cdot \pi \cdot \sqrt{\Phi(z)} + (1 - h(y)) \cdot \pi \cdot \Phi(z)^3\right) \\ \sin\left(h(y) \cdot \pi \cdot \sqrt{\Phi(z)} + (1 - h(y)) \cdot \pi \cdot \Phi(z)^3\right) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(y, z)} + \epsilon \end{aligned} \quad (31)$$

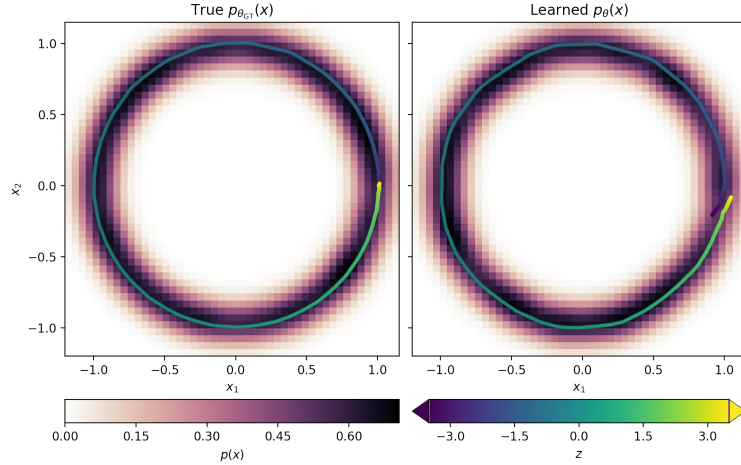
where  $\Phi$  is the CDF of a standard normal and  $B^{-1}(\cdot; \alpha, \beta)$  is the inverse CDF of the beta distribution.



**Properties:** As in the “Discrete Semi-Circle Example”, we designed this data-set to have a  $p(x)$  that the VAE objective would learn well so we can focus on the new issues introduced by the semi-supervised objective. The dataset demonstrates the same pathologies in the semi-supervised objective as shown by “Discrete Semi-Circle Example” with the addition of yet another pathology: since the posterior  $p_\theta(y|x)$  is bimodal in this example, encouraging a MFG  $q_\phi(y|x)$  discriminator to be predictive will collapse  $f_\theta(y, z)$  to the same function for all values of  $y$  (see Figure 14b) As such, as we increase  $\alpha$ , the better our predictive accuracy will be but the more  $p_\theta(x|y) \rightarrow p_\theta(x)$ , causing the learned model to generate poor quality counterfactuals (see Figure 14c).

## H QUALITATIVE RESULTS

### H.1 QUALITATIVE RESULTS TO SUPPORT NECESSITY OF BOTH CONDITIONS OF THEOREM 1



(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

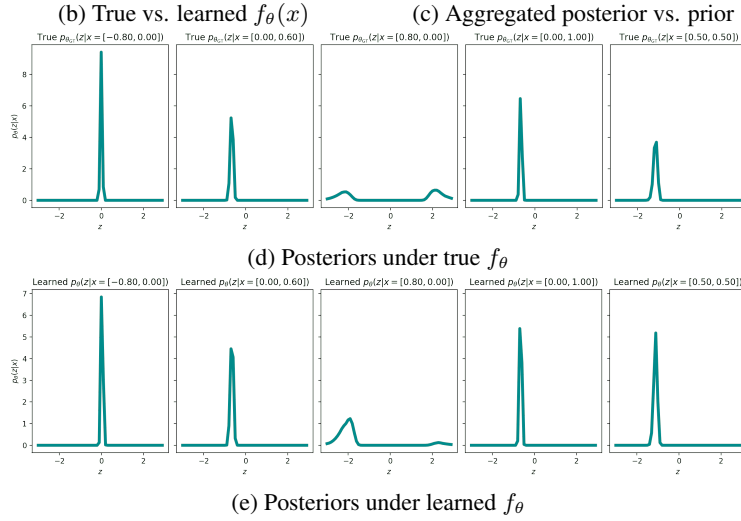
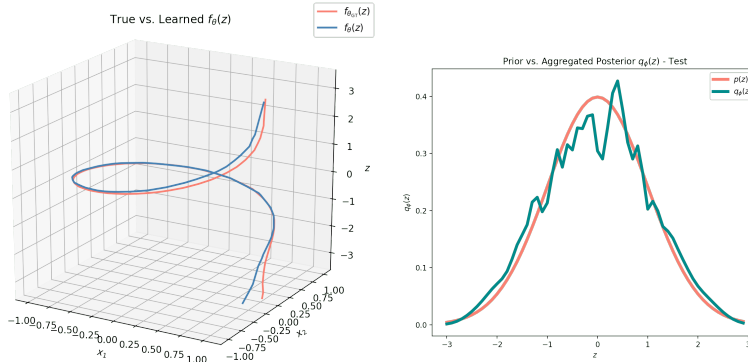


Figure 2: MFG-VAE trained on the Circle Example. In this toy data, condition (2) holds of Theorem 1 holds and condition (1) does not. To see this, notice that most examples of the posteriors are Gaussian-like, with the exception of the posteriors near  $x = [1.0, 0.0]$ , which are bimodal since in that neighborhood,  $x$  could have been generated using either  $z > 3.0$  or using  $z < -3.0$ . Since only a few training points have a high posterior matching objective, a VAE is able to learn the data distribution well.

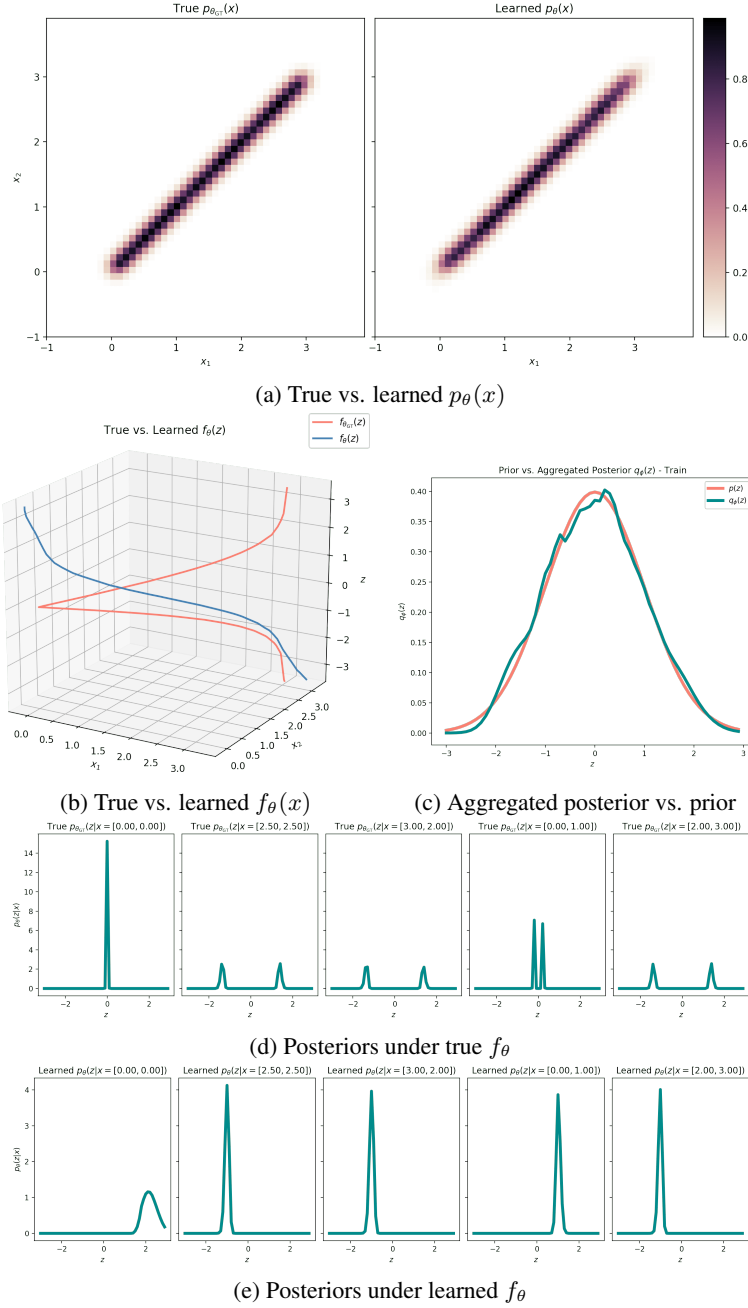
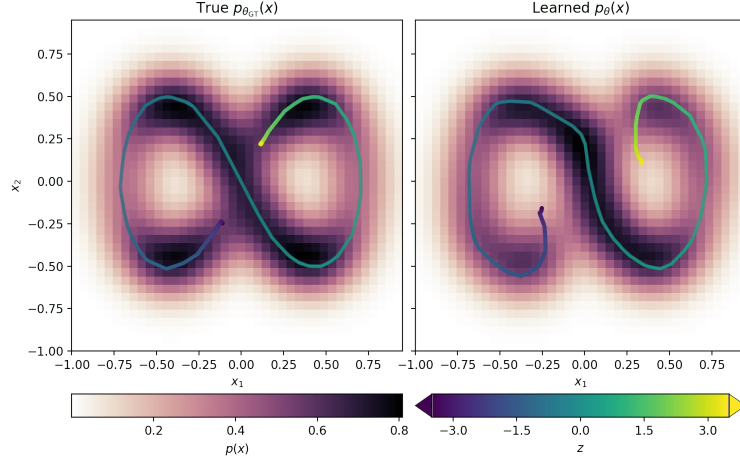
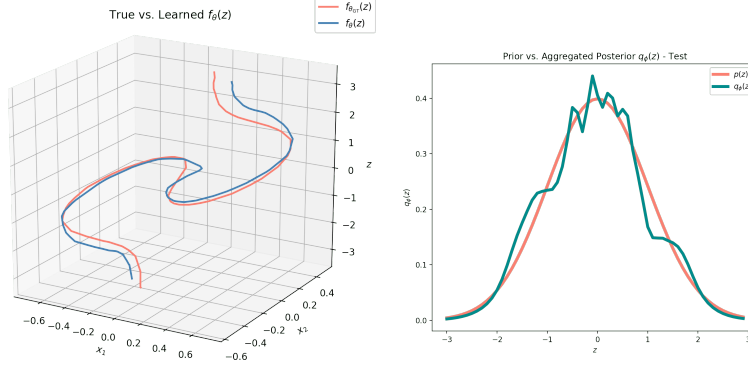


Figure 3: MFG-VAE trained on the Absolute-Value Example. In this toy data, condition (1) holds of Theorem 1 holds and condition (2) does not. To see this, notice that the function  $f_\theta$  learned with a VAE is completely different than the ground-truth  $f_\theta$ , and unlike the ground truth  $f_\theta$  which has bimodal posteriors, the learned  $f_\theta$  has unimodal posteriors (which are easier to approximate with a MFG). As such, a VAE is able to learn the data distribution well.

## H.2 QUALITATIVE DEMONSTRATION OF UNSUPERVISED VAE PATHOLOGIES

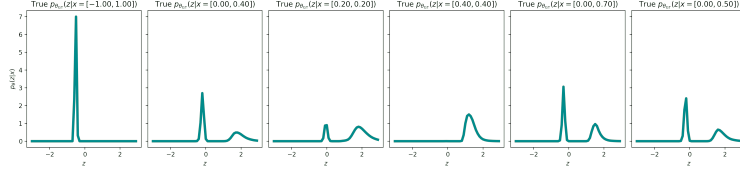


(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

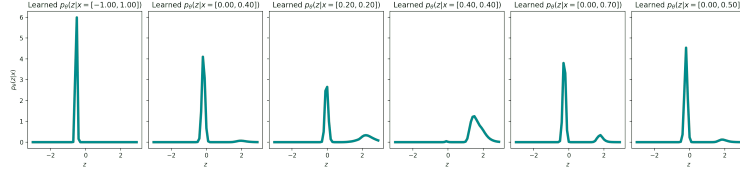


(b) True vs. learned  $f_{\theta}(x)$

(c) Aggregated posterior vs. prior

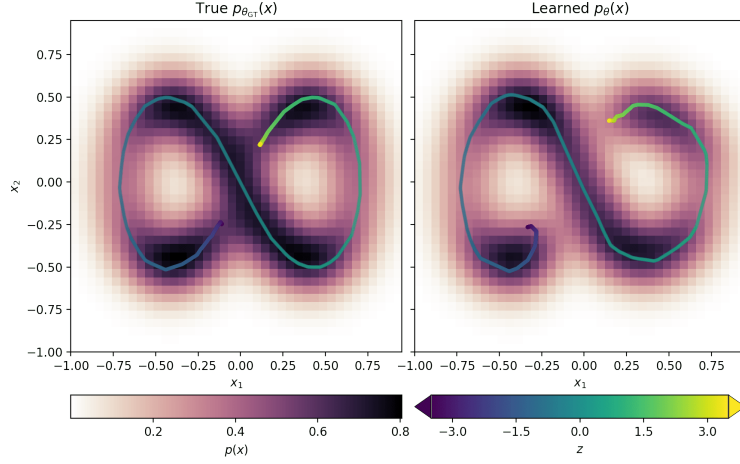


(d) Posteriors under true  $f_{\theta}$

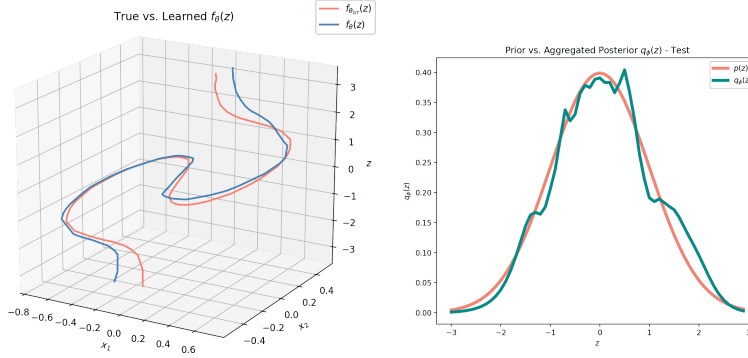


(e) Posteriors under learned  $f_{\theta}$

Figure 4: MFG-VAE trained on the Figure-8 Example. In this toy data, both conditions of Theorem 1 hold. The VAE learns a generative model with simpler posterior than that of the ground-truth, though it is unable to completely simplify the posterior as in the Absolute-Value Example. To learn a generative model with a simpler posterior, it curves the learned function  $f_{\theta}$  at  $z = -3.0$  and  $z = 3.0$  away from the region where  $z = 0$ . This is because under the true generative model, the true posterior  $p_{\theta}(z|x)$  in the neighborhood of  $x \approx 0$  has modes around either  $z = 0$  and  $z = 3.0$ , or around  $z = 0$  and  $z = -3.0$ , leading to a high posterior matching objective.

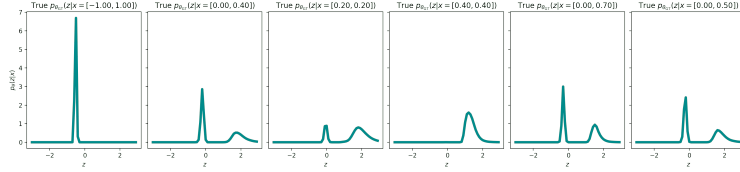


(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

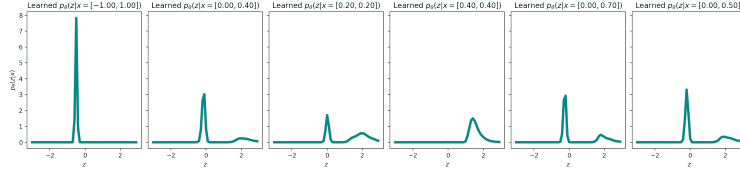


(b) True vs. learned  $f_{\theta}(x)$

(c) Aggregated posterior vs. prior

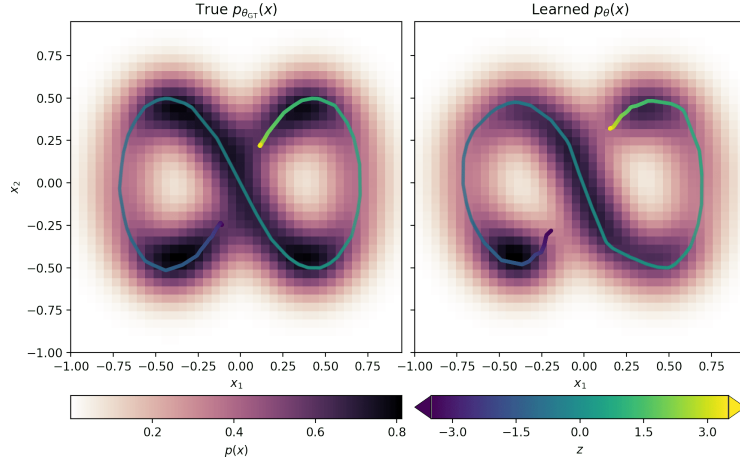


(d) Posteriors under true  $f_{\theta}$

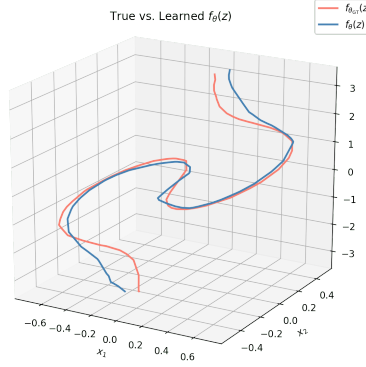


(e) Posteriors under learned  $f_{\theta}$

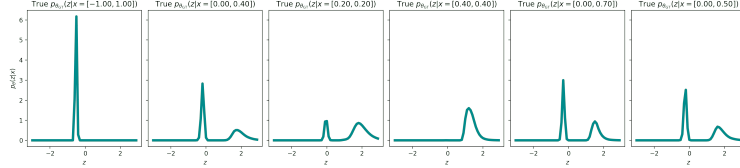
Figure 5: VAE with Lagging Inference Networks (LIN) trained on the Figure-8 Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a MFG-VAE does (see Figure 4).



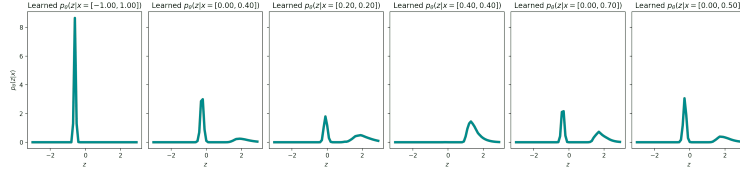
(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .



(b) True vs. learned  $f_{\theta}(x)$

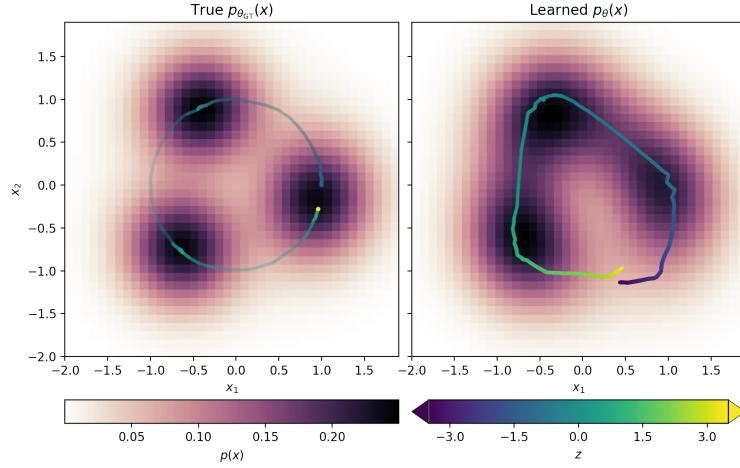


(c) Posteriors under true  $f_{\theta}$

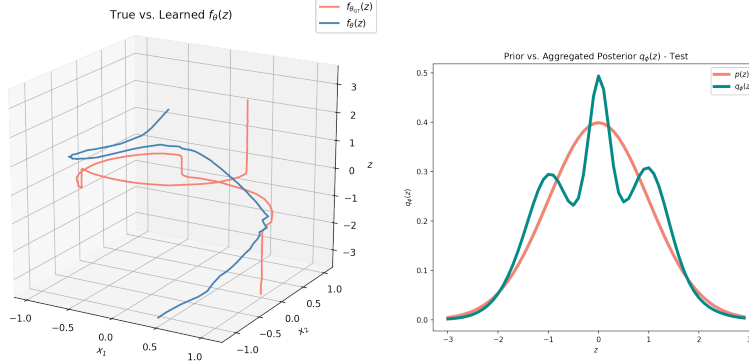


(d) Posteriors under learned  $f_{\theta}$

Figure 6: IWAE trained on the Figure-8 Example. In this toy data, both conditions of Theorem 1 hold. The IWAE learns a generative model with a slightly simpler posterior than that of the ground-truth. This is because even with the number of importance samples as large as  $S = 20$ , the variational family implied by the IWAE objective is not sufficiently expressive. The objective therefore prefers to learn a model with a lower data marginal likelihood. While increasing  $S \rightarrow \infty$  will resolve this issue, it is not clear how large a  $S$  is necessary and whether the additional computational overhead is worth it.

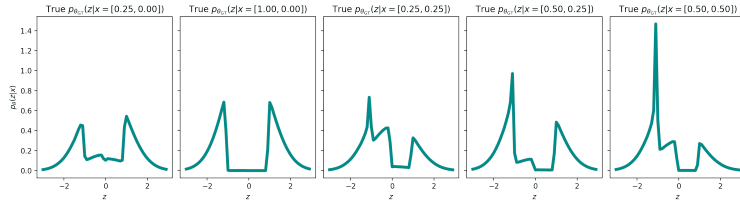


(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

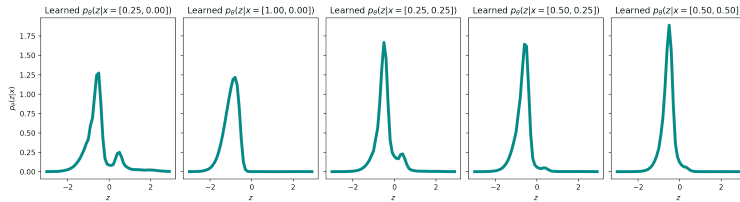


(b) True vs. learned  $f_{\theta}(x)$

(c) Aggregated posterior vs. prior



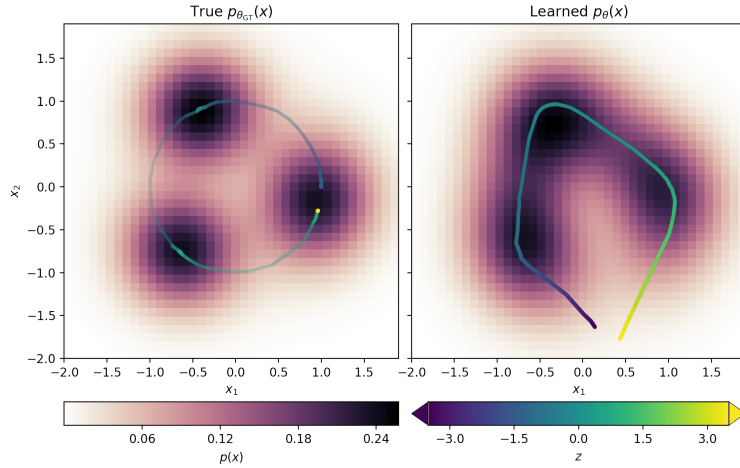
(d) Posteriors under true  $f_{\theta}$



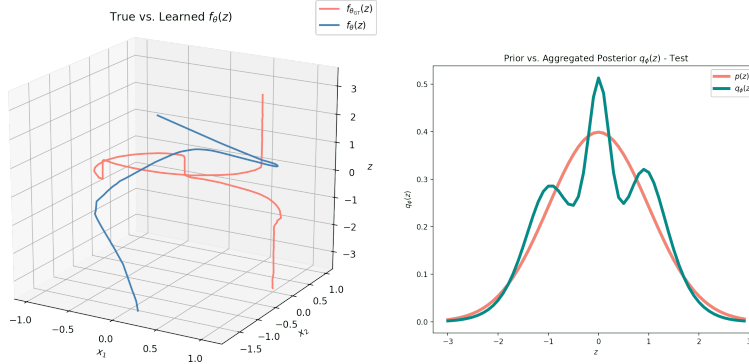
(e) Posteriors under learned  $f_{\theta}$

Figure 7: MFG-VAE trained on the Clusters Example. In this toy data, both conditions of Theorem 1 hold. The VAE learns a generative model with simpler posterior than that of the ground-truth, though it is unable to completely simplify the posterior as in the Absolute-Value Example. To learn a generative model with a simpler posterior, it learns a model with a function  $f_{\theta}(z)$  that, unlike the ground truth function, does not have steep areas interleaved between flat areas. As such, the learned model is generally more flat, causing the learned density to be “smeared” between the modes.



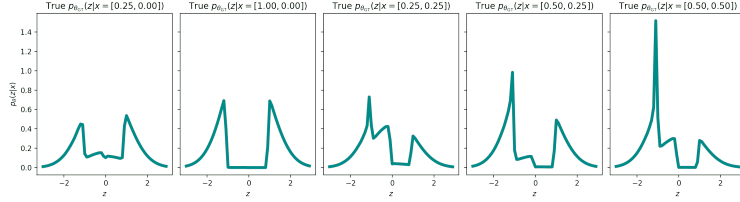


(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

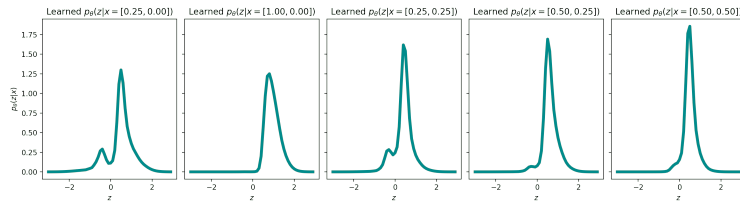


(b) True vs. learned  $f_{\theta}(x)$

(c) Aggregated posterior vs. prior

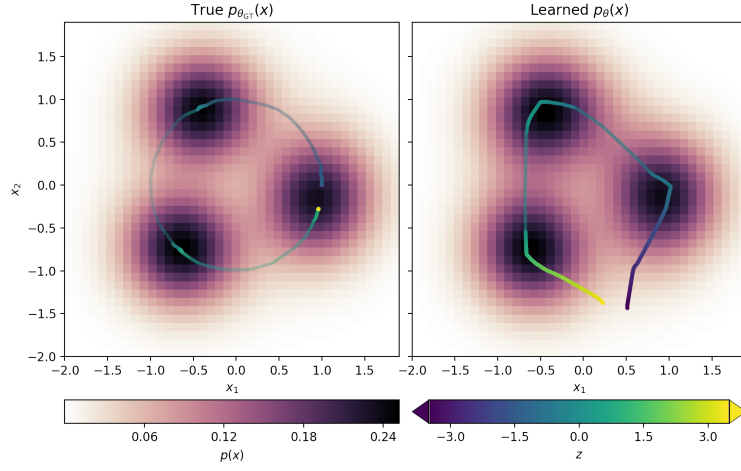


(d) Posteriors under true  $f_{\theta}$

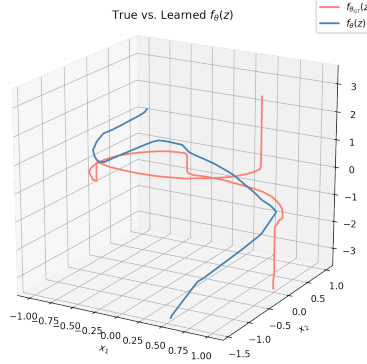


(e) Posteriors under learned  $f_{\theta}$

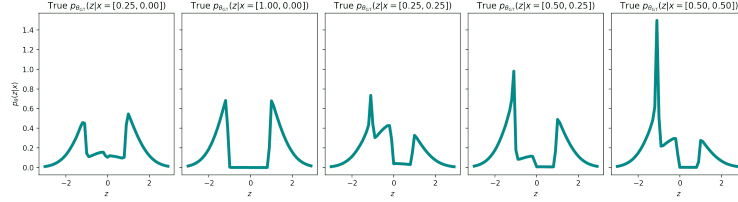
Figure 8: VAE with Lagging Inference Networks (LIN) trained on the Clusters Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a MFG-VAE does (see Figure 7).



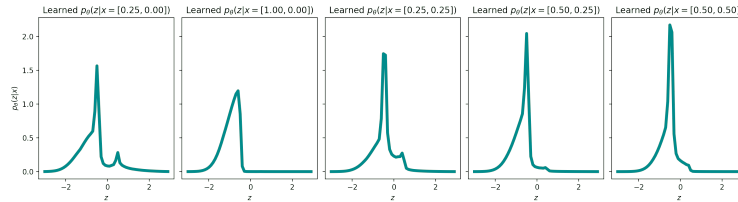
(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .



(b) True vs. learned  $f_{\theta}(x)$

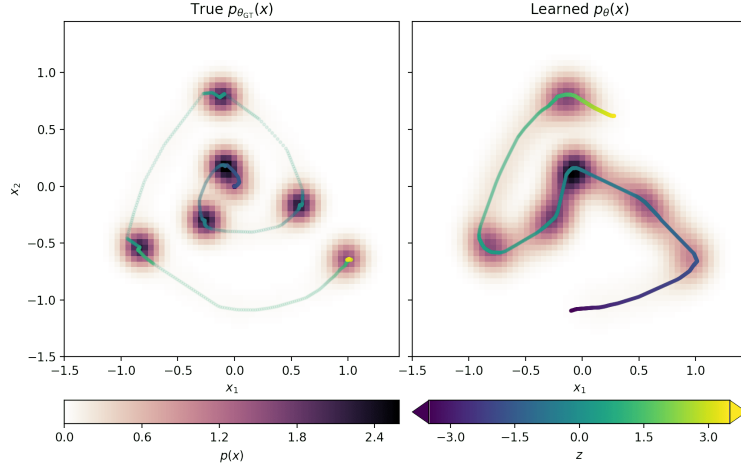


(c) Posteriors under true  $f_{\theta}$

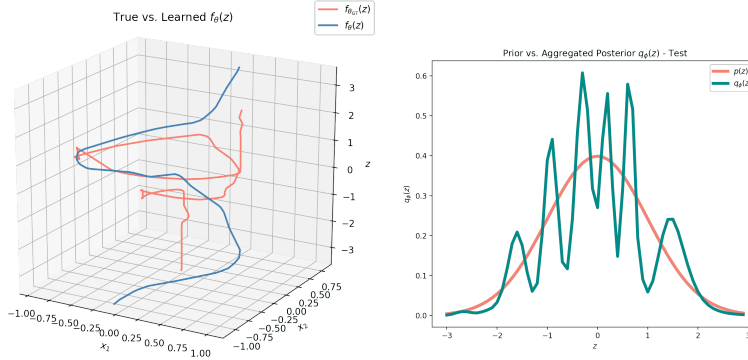


(d) Posteriors under learned  $f_{\theta}$

Figure 9: IWAE trained on the Clusters Example. In this toy data, both conditions of Theorem 1 hold. IWAE is able to learn the ground truth data distribution while finding a generative model with a simpler posterior than that of the ground-truth model.

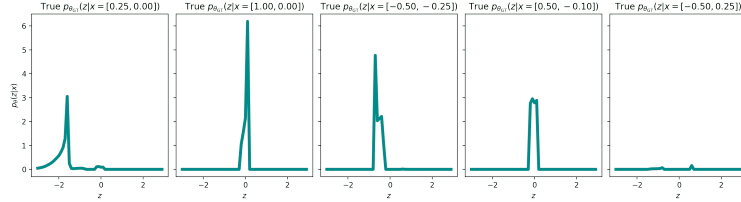


(a) True vs. learned  $p_{\theta}(x)$ , and learned vs. true  $f_{\theta}(z)$ , colored by the value of  $z$ .

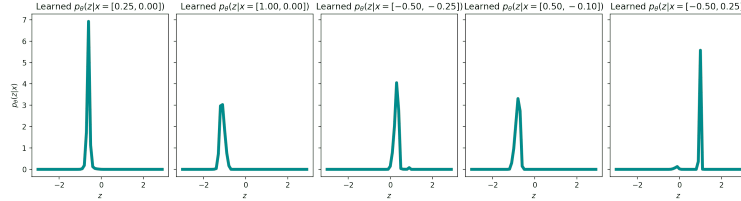


(b) True vs. learned  $f_{\theta}(x)$

(c) Aggregated posterior vs. prior



(d) Posteriors under true  $f_{\theta}$



(e) Posteriors under learned  $f_{\theta}$

Figure 10: MFG-VAE trained on the Spiral-Dots Example jointly over  $\theta, \phi, \epsilon_{\epsilon}^2$ . In this toy data, as Theorem 2 predicts, the ELBO drastically misestimates the observation noise. The VAE learns a generative model with simpler posterior than that of the ground-truth, though it is unable to completely simplify the posterior as in the Absolute-Value Example. To learn a generative model with a simpler posterior, it learns a model with a function  $f_{\theta}(z)$  that, unlike the ground truth function, does not have steep areas interleaved between flat areas. As such, the learned model is generally more flat, causing the learned density to be “smeared” between the modes. Moreover due to the error in approximating the true posterior with a MFG variational family, the ELBO misestimates  $\sigma_{\epsilon}^2$ .

### H.3 QUALITATIVE DEMONSTRATION OF SEMI-SUPERVISED VAE PATHOLOGIES

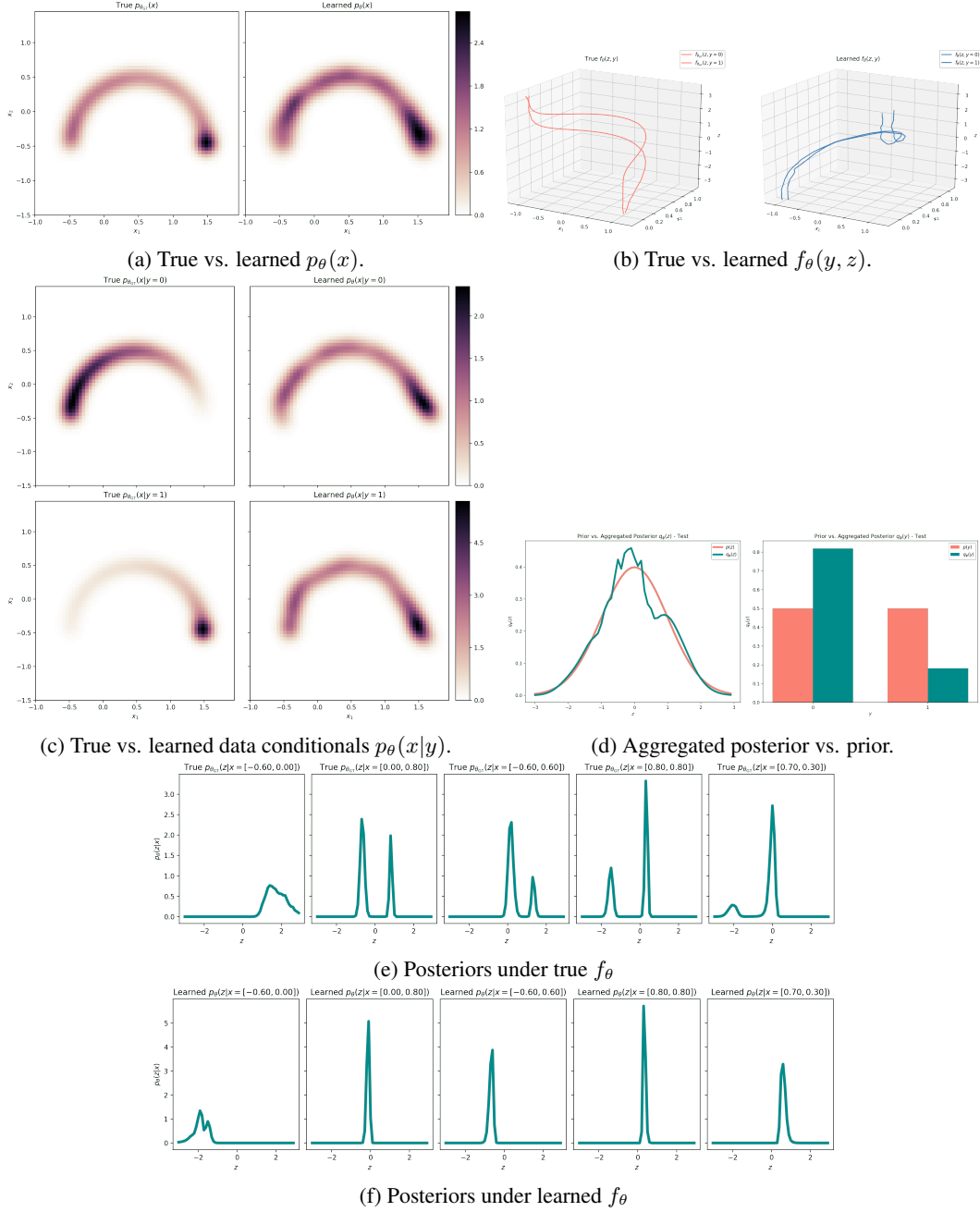


Figure 11: Semi-Supervised MFG-VAE trained on the Discrete Semi-Circle Example. While using semi-supervision, a VAE is still able to learn the  $p(x)$  relatively well. However, in this example, given  $x$  there is uncertainty as to whether it was generated from  $f_{\theta}(y=0, z)$  or from  $f_{\theta}(y=1, z)$ , the posterior  $p_{\theta}(z|x)$  is bimodal and will cause a high posterior matching objective. Since semi-supervised VAE objective prefers models with simpler posteriors, the VAE learns a unimodal posterior by collapsing  $f_{\theta}(y=0, z) = f_{\theta}(y=1, z)$ , causing  $p(x|y=0) \approx p(x|y=1) \approx p(x)$ . The learned model will therefore generate poor sample quality counterfactuals.

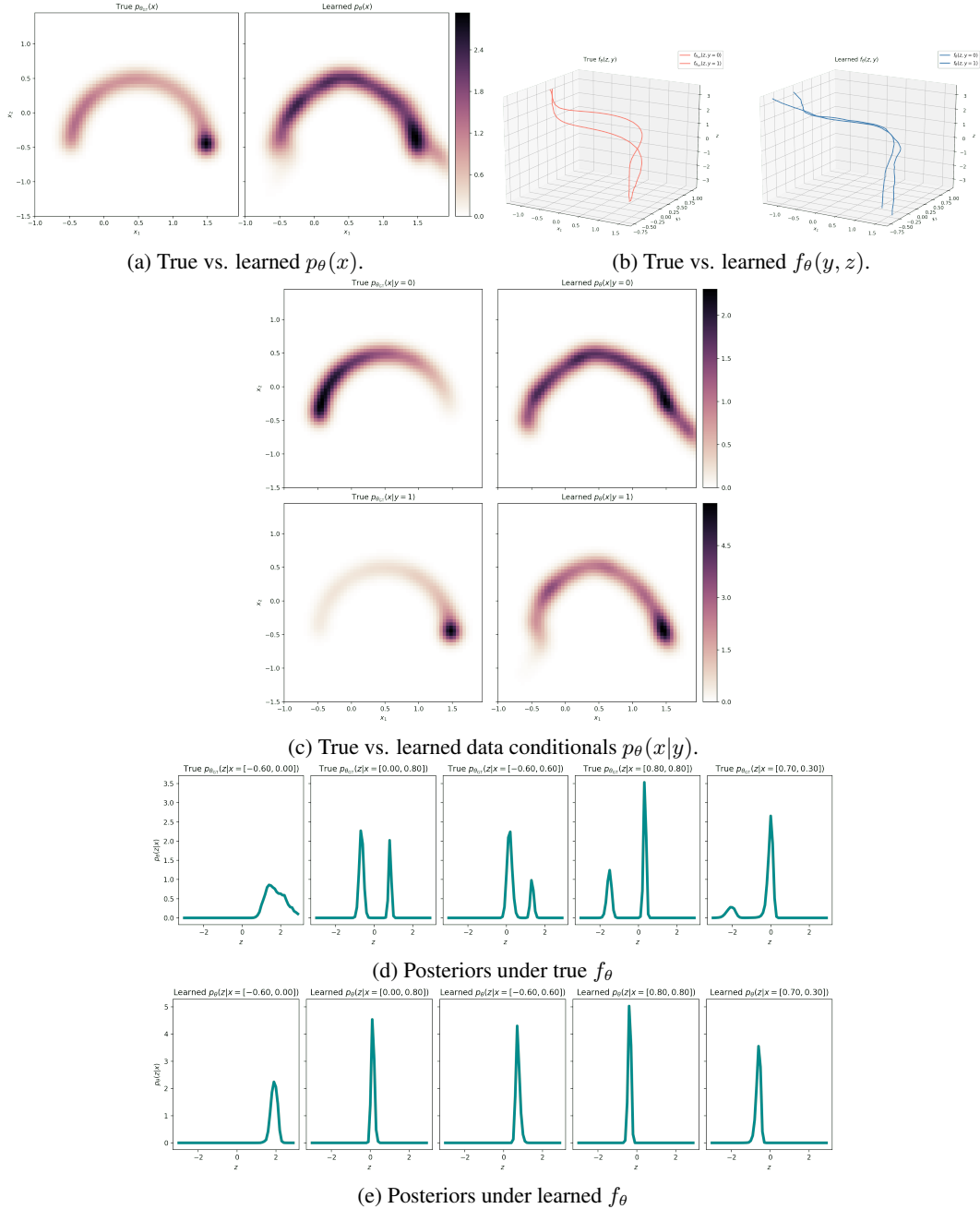


Figure 12: Semi-Supervised VAE trained with Lagging Inference Networks (LIN) trained on the Discrete Semi-Circle Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a MFG-VAE does (see Figure 11).

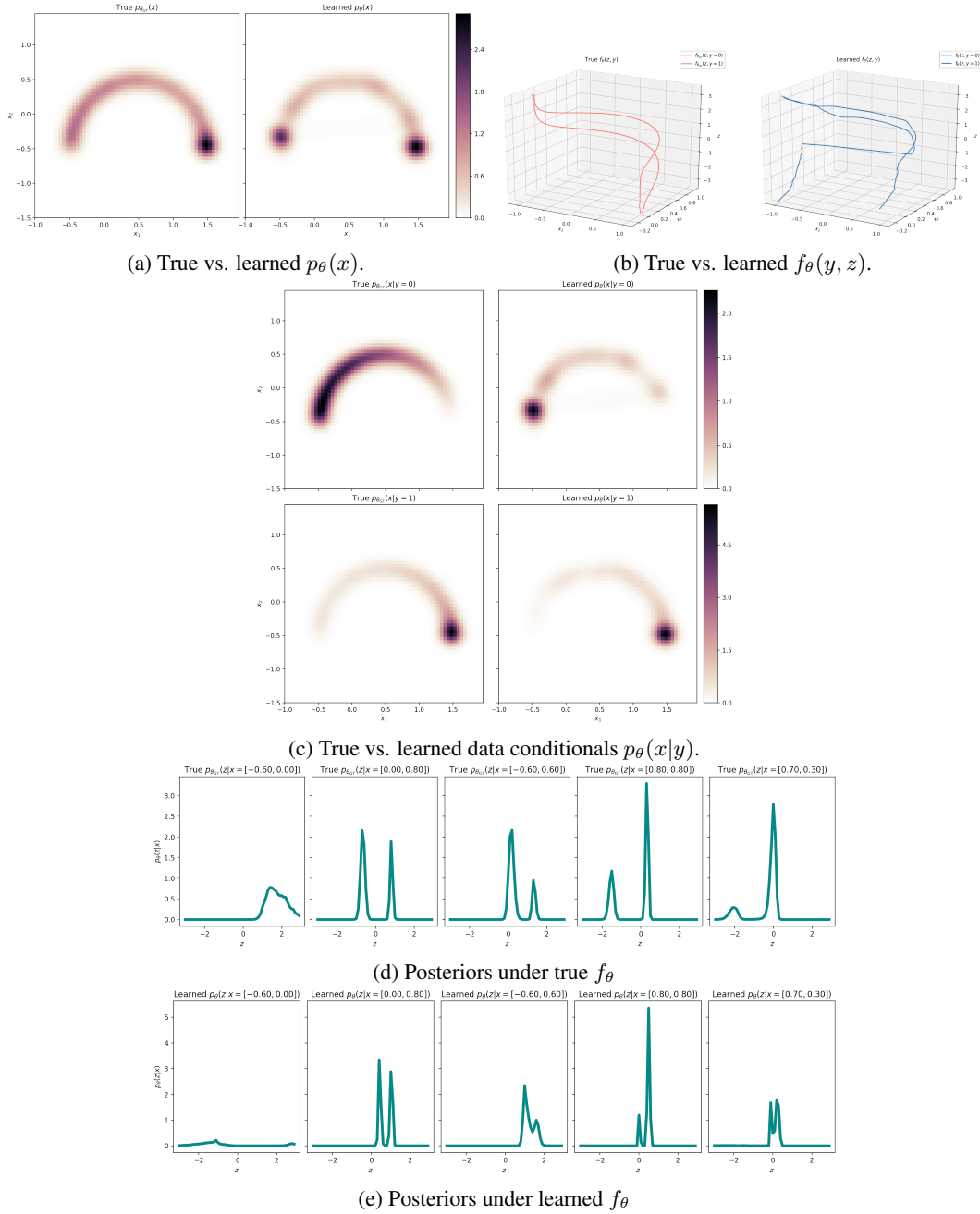


Figure 13: Semi-Supervised IWAE trained on the Discrete Semi-Circle Example. While using semi-supervision, a IWAE is still able to learn the  $p(x)$  and  $p(x|y)$  better than a VAE. This is because it allows for more complicated posteriors and therefore does not collapse  $f_{\theta}(y=0, z) = f_{\theta}(y=1, z)$ . However, since IWAE has a more complex variational family, the variational family no longer regularizes the function  $f_{\theta}$ . As such, in order to put enough mass on the left-side of the semi-circle,  $f_{\theta}$  jumps sharply from the right to the left, as opposed to preferring a simpler function such as the ground truth function.

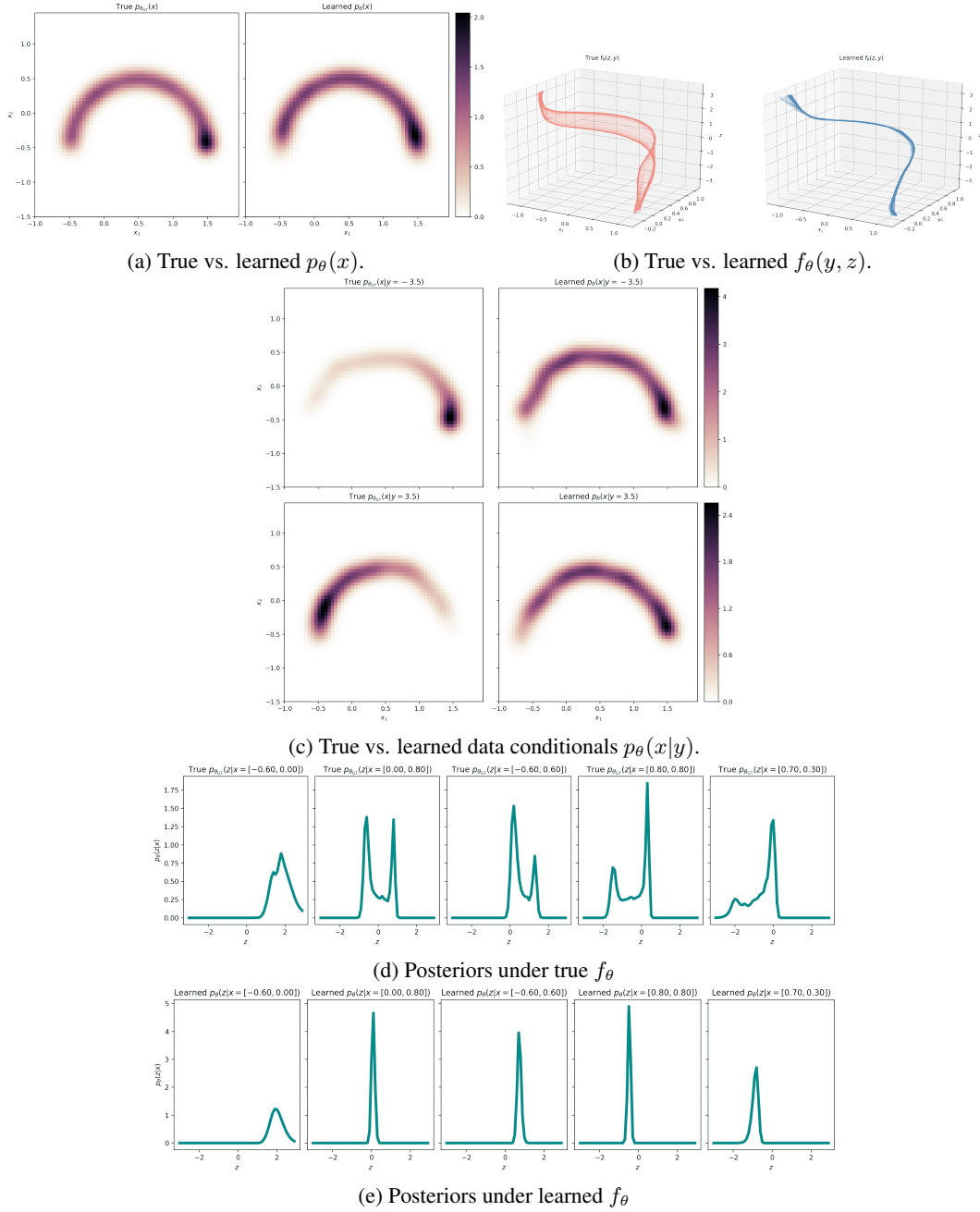


Figure 14: Semi-Supervised MFG-VAE trained on the Continuous Semi-Circle Example. In this example, the VAE exhibits the same problems as in the Discrete Semi-Circle Example (Figure 14). However, with since  $y$  is continuous, this poses an additional issue. Since  $q_\phi(y|x)$  (the discriminator) in the objective is a Gaussian, and the ground truth  $p_\theta(y|x)$  is multi-modal, the objective will select a function  $f_\theta$  under which  $p_\theta(y|x)$  is a MFG. This, again, leads to learning a model in which  $f_\theta(y = \cdot, z)$  are the same for all values of  $y$ , causing  $p(x|y = 0) \approx p(x|y = 1) \approx p(x)$ . The learned model will therefore generate poor sample quality counterfactuals.

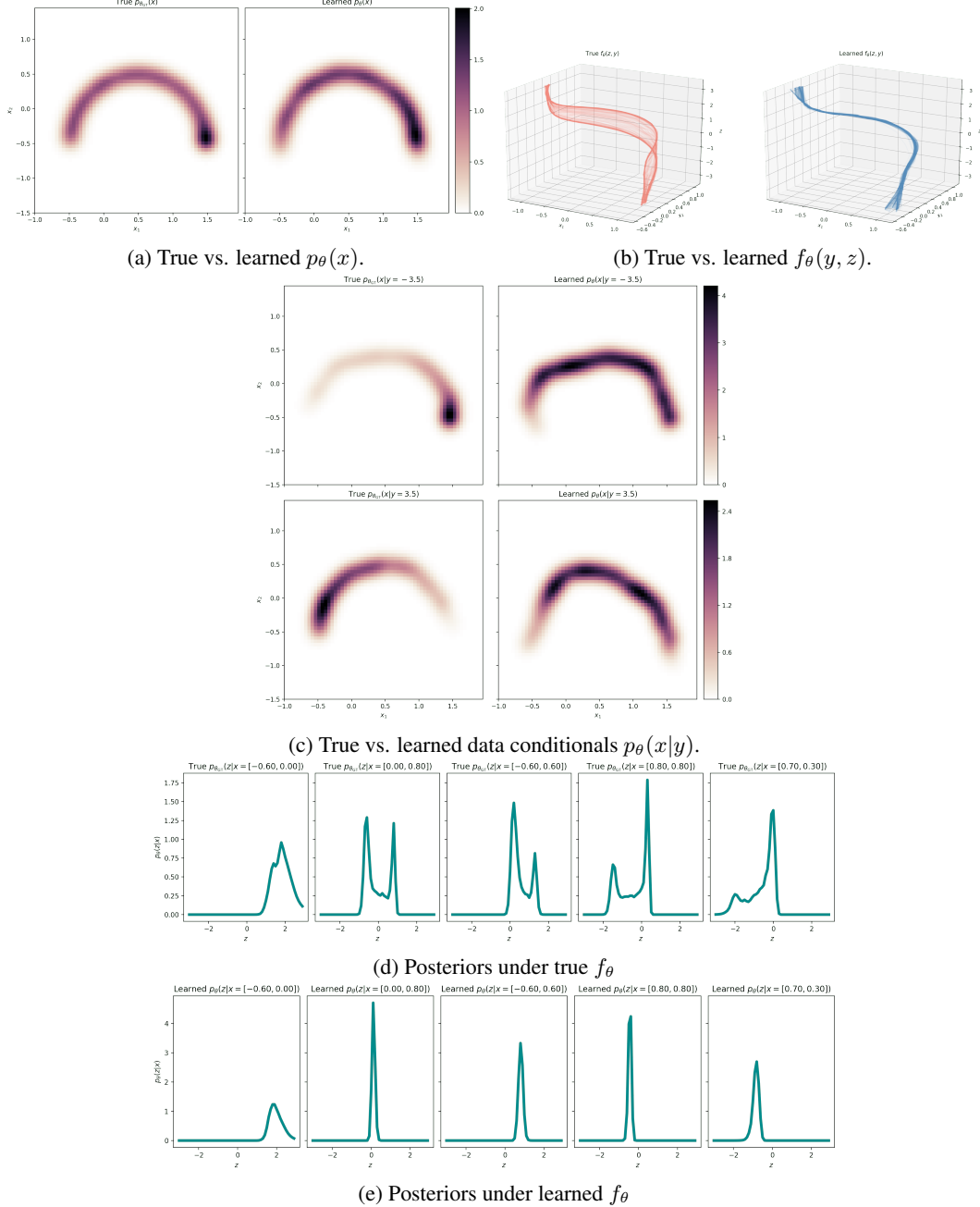


Figure 15: Semi-Supervised VAE trained with Lagging Inference Networks (LIN) trained on the Continuous Semi-Circle Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a MFG-VAE does (see Figure 14).



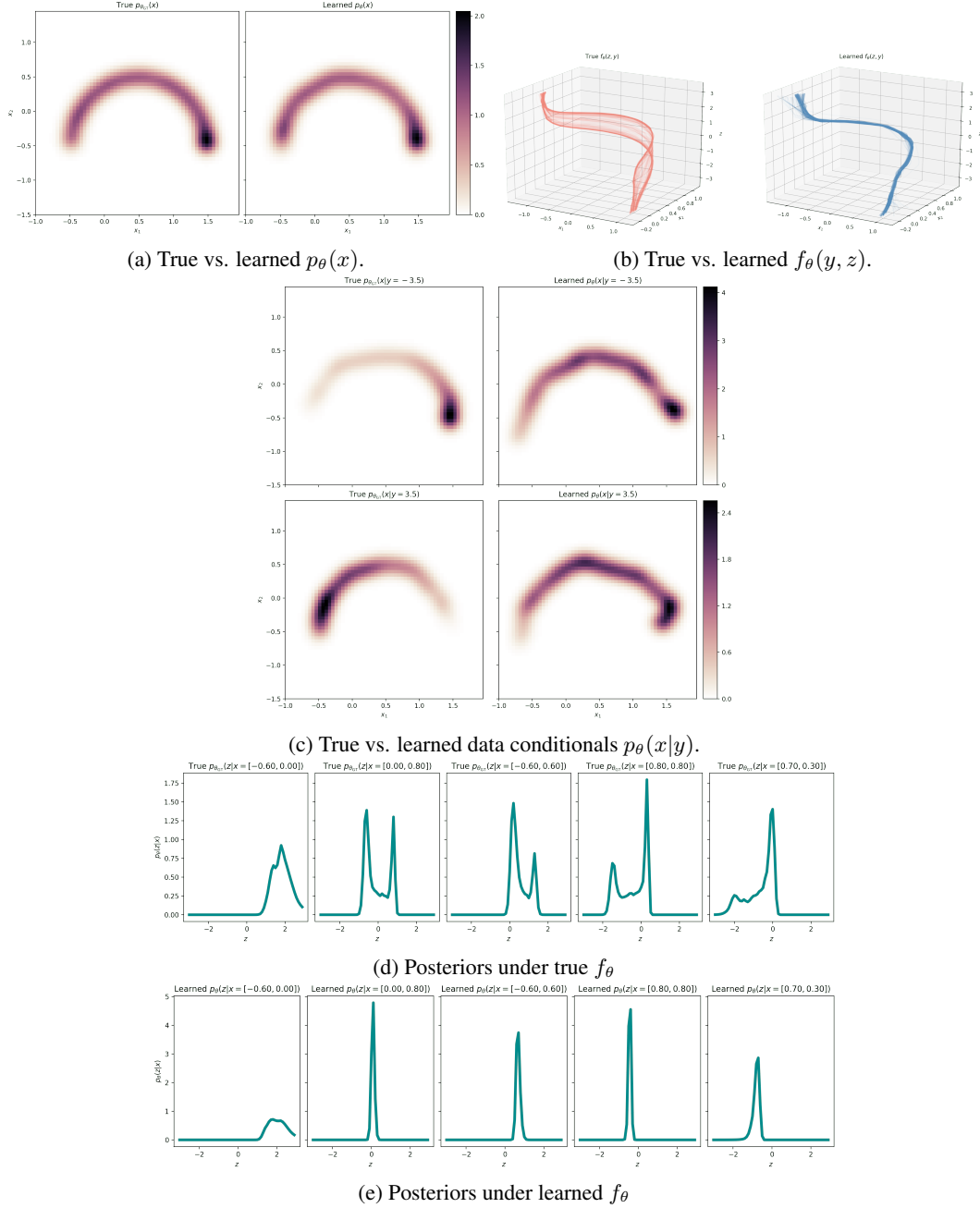


Figure 16: Semi-Supervised IWAE trained on the Continuous Semi-Circle Example. While using semi-supervision, a IWAE is still able to learn the  $p(x)$  and  $p(x|y)$  better than a VAE. However, since  $q_\phi(y|x)$  (the discriminator) in the objective is a Gaussian, and the ground truth  $p_\theta(y|x)$  is multi-modal, the objective will select a function  $f_\theta$  under which  $p_\theta(y|x)$  is a MFG. This, again, leads to learning a model in which  $f_\theta(y = \cdot, z)$  are the same for all values of  $y$ , causing  $p(x|y = 0) \approx p(x|y = 1) \approx p(x)$ . The learned model will therefore generate poor sample quality counterfactuals.

#### H.4 WHEN LEARNING COMPRESSED REPRESENTATIONS, POSTERIOR IS SIMPER FOR MISMATCHED MODELS

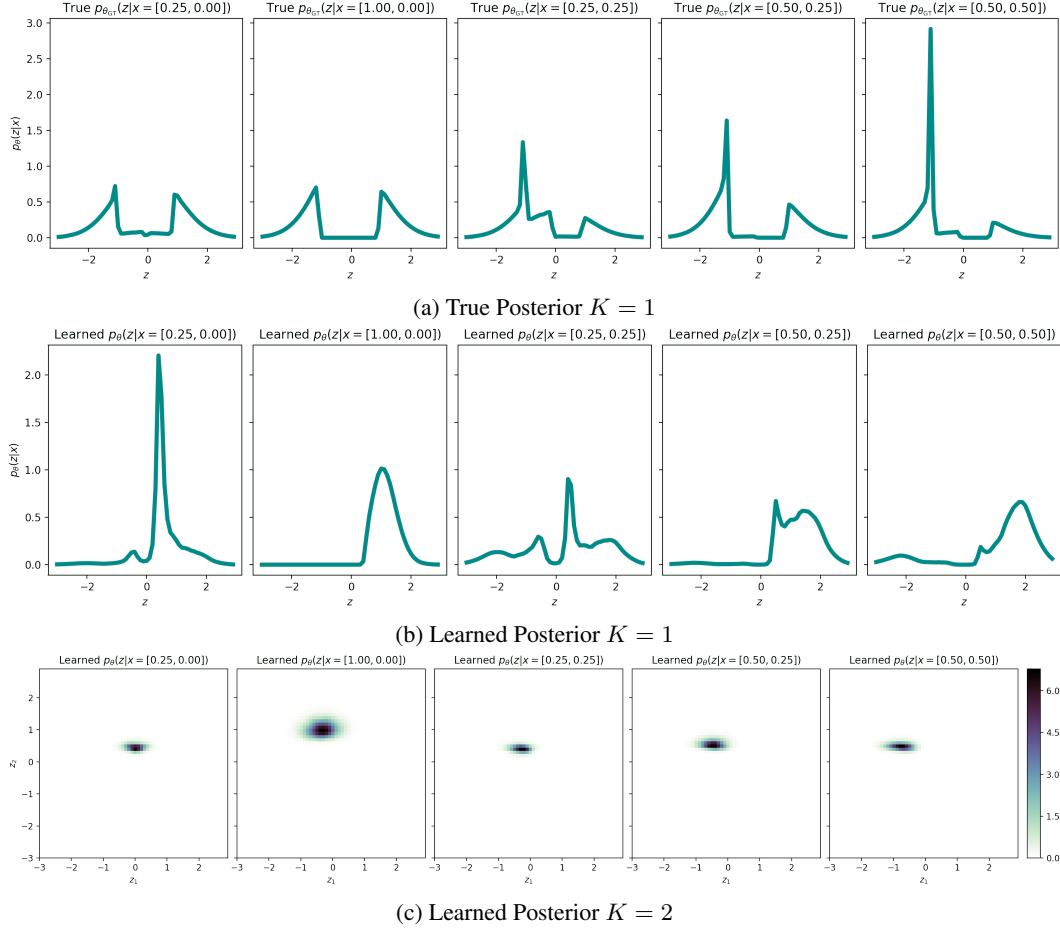


Figure 17: VAEs learn simpler posteriors as latent dimensionality  $K$  increases and as the observation noise  $\sigma_\epsilon^2$  decreases on “Clusters Example” (projected into 5D space).

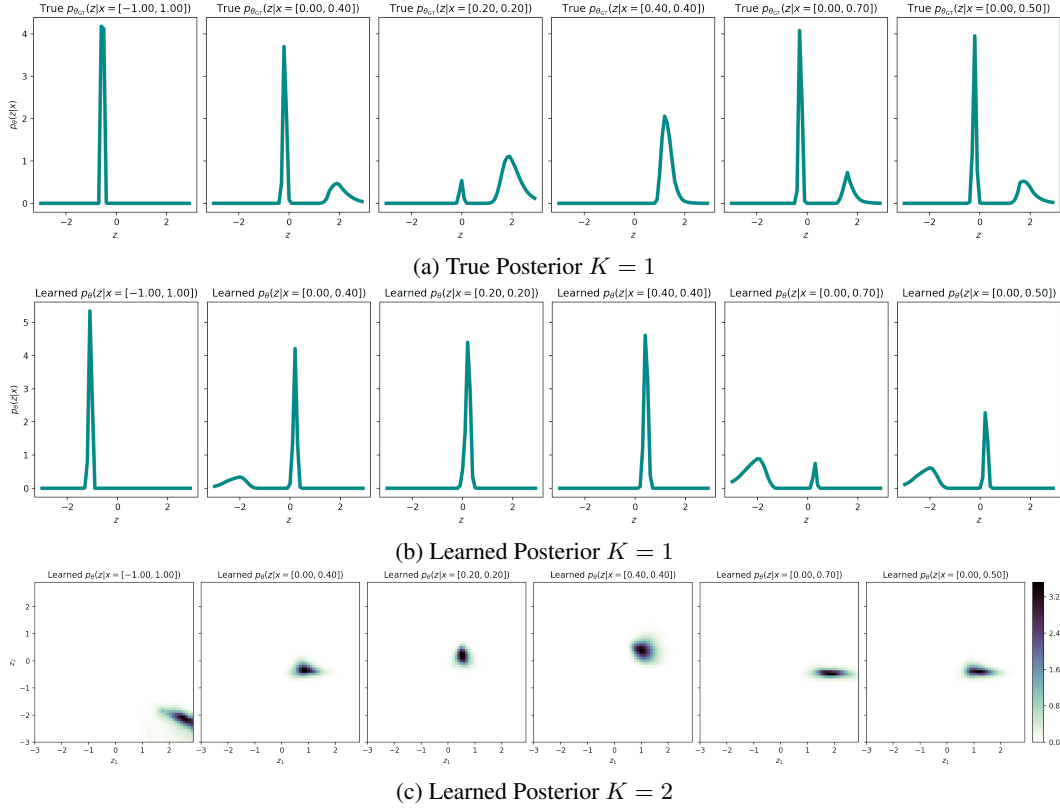


Figure 18: VAEs learn simpler posteriors as latent dimensionality  $K$  increases and as the observation noise  $\sigma_{\epsilon}^2$  decreases on “Figure-8 Example” (projected into 5D space).

## REFERENCES

- Jesus Alcala-Fdez, Alberto Fernández, Julián Luengo, J. Derrac, S Garcia, Luciano Sanchez, and Francisco Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17: 255–287, 01 2010.
- Balint Antal and Andras Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *arXiv e-prints*, art. arXiv:1410.8576, October 2014.
- Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv e-prints*, art. arXiv:1802.04942, February 2018.
- Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting Importance-Weighted Autoencoders. *arXiv:1704.02916 [stat]*, August 2017. URL <http://arxiv.org/abs/1704.02916>. arXiv: 1704.02916.
- Josip Djolonga and Andreas Krause. Learning Implicit Generative Models Using Differentiable Graph Tests. *arXiv e-prints*, art. arXiv:1709.01006, Sep 2017.
- Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Uiwon Hwang, Jaewoo Park, Hyemi Jang, Sungroh Yoon, and Nam Ik Cho. PuVAE: A Variational Autoencoder to Purify Adversarial Examples. *arXiv e-prints*, art. arXiv:1903.00585, March 2019.
- Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G. Dimakis. The Robust Manifold Defense: Adversarial Training using Generative Models. *arXiv e-prints*, art. arXiv:1712.09196, December 2017.

- Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. *arXiv e-prints*, art. arXiv:1611.01144, November 2016.
- Uyeong Jang, Somesh Jha, and Susmit Jha. ON THE NEED FOR TOPOLOGY-AWARE GENERATIVE MODELS FOR MANIFOLD-BASED DEFENSES. pp. 24, 2020.
- Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014.
- Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv:1312.6114, December 2013.
- Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models. *arXiv e-prints*, art. arXiv:1406.5298, June 2014.
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. 69(6): 066138, June 2004. doi: 10.1103/PhysRevE.69.066138.
- Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv e-prints*, art. arXiv:1811.12359, November 2018.
- Dongyu Meng and Hao Chen. MagNet: a Two-Pronged Defense against Adversarial Examples. *arXiv e-prints*, art. arXiv:1705.09064, May 2017.
- Karl Ridgeway. A Survey of Inductive Biases for Factorial Representation-Learning. *arXiv e-prints*, art. arXiv:1612.05299, December 2016.
- Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *arXiv e-prints*, art. arXiv:1805.06605, May 2018.
- N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H. S. Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. *arXiv e-prints*, art. arXiv:1706.00400, June 2017.
- N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H. S. Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. *arXiv:1706.00400 [cs, stat]*, November 2017. URL <http://arxiv.org/abs/1706.00400>. arXiv: 1706.00400.
- Jeffrey Simonoff. The "unusual episode" and a second statistics course. *Journal of Statistics Education*, 5, 03 1997. doi: 10.1080/10691898.1997.11910524.
- Lucas Theis, Aaron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv:1511.01844 [cs, stat]*, April 2016. URL <http://arxiv.org/abs/1511.01844>. arXiv: 1511.01844.
- Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the Quantitative Analysis of Decoder-Based Generative Models. *arXiv:1611.04273 [cs]*, June 2017. URL <http://arxiv.org/abs/1611.04273>. arXiv: 1611.04273.