Failures of Variational Autoencoders and their Effects on Downstream Tasks

Yaniv Yacoby [1] **Weiwei Pan** [1] **Finale Doshi-Velez** [1]

Abstract

Variational Auto-encoders (VAEs) are deep generative latent variable models that are widely used for a number of downstream tasks. While it has been demonstrated that VAEs training can suffer from a number of pathologies, existing literature lacks characterizations of exactly when these pathologies occur and how they impact downstream task performance. In this paper we concretely characterize conditions under which VAE training exhibits pathologies and connect these failure modes to undesirable effects on specific downstream tasks – learning compressed and disentangled representations, adversarial robustness and semi-supervised learning.

1. Introduction

Variational Auto-encoders (VAEs) are deep generative latent variable models that transform simple distributions over a latent space to model complex data distributions (Kingma & Welling, 2013). Formally, a VAE has two components: a *generative* model that transforms a distribution over latent space $p(z)$ into a distribution over data $p(x)$, and an amortized *inference* model that provides an approximate posterior $q(z|x) \approx p(z|x)$. Due to the simplicity of their training procedure, they have been used for a wide range of downstream tasks, including: generating realistic looking synthetic data (e.g Pu et al. (2016)), learning compressed representations (e.g. Miao & Blunsom (2016); Gregor et al. (2016); Alemi et al. (2017)), adversarial defense using denoising (Luo & Pfister, 2018; Ghosh et al., 2018), and, when partial labels are available, generating counter-factual data using weak or semi-supervision (e.g. Kingma et al. (2014); Siddharth et al. (2017); Klys et al. (2018)).

The common choice of mean-field Gaussian (MFG) approximate posteriors for VAEs (MFG-VAE) results an inference procedure that is straight-forward to implement and stable in training. Unfortunately, a growing body of work has demonstrated that MFG-VAEs suffer from a variety of pathologies, including learning un-informative latent codes (posterior collapse) (e.g.van den Oord et al. (2017); Kim et al. (2018)) and unrealistic data distributions (mismatch between aggregate posterior and prior) (e.g. Makhzani et al. (2015); Tomczak & Welling (2017)). Recent work (Yacoby et al., 2020) attributes a number of these pathologies to properties of the training objective; in particular, the objective may compromise learning a good generative model in order to learn a good inference model. While this pathology has been noted in literature (Burda et al., 2016; Zhao et al., 2017; Cremer et al., 2018), no prior work has characterizes the conditions under which it occurs; more worrisomely, no prior work has related MFG-VAE pathologies with the performance of MFG-VAEs on downstream tasks. Rather, existing literature focuses on mitigating the over-regularizing effect of the VAE's inference model on it's generative model by using richer variational families (e.g. Kingma et al. (2016); Cremer et al. (2017); Nowozin (2018); Luo et al. (2020)). While promising, these methods introduce potentially significant additional computational costs to training, as well as new training issues (e.g. noisy gradients (Roeder et al., 2017; Tucker et al., 2018)). As such, it is important to understand precisely when MFG-VAEs exhibit pathologies and when alternative training methods are worth the computational trade-off. In this paper, we characterize the conditions under which MFG-VAEs perform poorly and link them directly to their performance on a variety of downstream tasks.

Our contributions are theoretical and empirical: (1) We characterize concrete conditions under which learning the inference model will compromise learning the generative model for MFG-VAEs. More problematically, we show that these bad solutions are globally optimal for the training objective, the ELBO. (2) We demonstrate that using the ELBO to select the output noise variance and the latent dimension results in biased estimates. (3) Furthermore, we demonstrate ways in which these pathologies affects key downstream tasks, including learning compressed and disentangled representations, adversarial robustness and semi-supervised learning. (4) Lastly, we show that while the use of richer variational families alleviate VAE pathologies on unsupervised learning tasks, they introduce new ones in the semi-supervised tasks.

[*]Equal contribution [1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. Correspondence to: Yaniv Yacoby <yanivyacoby@g.harvard.edu>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

Background on Unsupervised VAEs In this paper, we assume N observations of $x \in \mathbb{R}^D$. A VAE assumes the following generative process (Kingma & Welling, 2013): $p(z) = \mathcal{N}(0, I)$, $p_\theta(x|z) = \mathcal{N}(f_\theta(z), \sigma_\epsilon^2 I)$ where $z \in \mathbb{R}^K$ is a latent variable and f_θ is a neural network parametrized by θ. We learn θ while jointly approximating the posterior $p_\theta(z|x)$ with $q_\phi(z|x)$, by maximizing a lower bound of the marginal data likelihood:

$$\text{ELBO}(\theta, \phi) = \max_{\theta,\phi} \mathop{\mathbb{E}}_{p(x)} \left[\mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \right], \quad (1)$$

where $p(x)$ is the empirical data distribution, $p_\theta(x)$ is the learned data distribution, and $q_\phi(z|x)$ is a MFG with mean and variance $\mu_\phi(x), \sigma_\phi^2(x)$ that are outputs of a neural network with parameters ϕ.

The ELBO for can alternately be written as a sum of two objectives – the "MLE objective" (MLEO), which maximizes the $p_\theta(x)$, and the "posterior matching objective" (PMO), which encourages variational posteriors to match posteriors of the generative model. That is, we can write the minima of the negative ELBO over (θ, ϕ) as:

$$\operatorname*{argmin}_{\theta,\phi} (\underbrace{D_{\text{KL}}[p(x)||p_\theta(x)]}_{\text{MLEO}} + \underbrace{\mathop{\mathbb{E}}_{p(x)} [D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)]]}_{\text{PMO}}). \quad (2)$$

Background on Semi-Supervised VAEs We extend VAE model and inference to incorporate partial labels, allowing for some supervision of the latent space dimensions using the semi-supervised model introduced in (Kingma et al., 2014) as the "M2 model":

$$\begin{aligned} &z \sim \mathcal{N}(0, I),\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I), \\ &y \sim p(y),\ x|y, z = f_\theta(y, z) + \epsilon \end{aligned} \quad (3)$$

where y is observed only a portion of the time. The 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:

$$\begin{aligned} \mathcal{J}^\alpha(\theta, \phi) = \sum_{n=1}^{N} \mathcal{U}(x_n; \theta, \phi) + \gamma \cdot \sum_{m=1}^{M} \mathcal{L}(x_m, y_m; \theta, \phi) \\ + \alpha \cdot \sum_{m=1}^{M} \log q_\phi(y_m|x_m) \end{aligned} \quad (4)$$

where $\mathcal{U}$ and $\mathcal{L}$ lower bound the $p_\theta(x)$ and $p_\theta(x, y)$, respectively (Appendix D), and where the last term in $\mathcal{J}^\alpha$ explicitly increase discriminative power of the approximate posteriors $q_\phi(y_m|x_m)$. Following (Kingma et al., 2014), we assume MFG variational families for all approximate posteriors.

2. Pathologies of the VAE Training Objective

In Section 2.1 we identify two pathological properties of the VAE training objective. In Section 2.2 we provide empirical demonstrations where these pathologies manifest on datasets and in Section 3 we unpack how these pathologies affect a variety of downstream tasks.

2.1. Pathologies of the VAE Objective

We fix a set of feasible likelihood functions $\mathcal{F}$ and a variational family $\mathcal{Q}$, implied by our choice of the generative and inference model network architectures. We assume that $\mathcal{F}$ is expressive enough to contain any smooth function, including the ground truth generating function, and we assume $\mathcal{Q}$ contains all $q_{\phi^*}(z|x)$ corresponding to any $f_\theta \in \mathcal{F}$, where $\phi^* = \text{argmin}_\phi -\text{ELBO}(\theta, \phi)$.

Pathology I: The ELBO trades off generative model quality for simple posteriors

Theorem 1. *The global optima of the VAE objective correspond to incorrect generative models under the following two conditions: (1) the true posterior is difficult to approximate by a MFG for a large portion of x's, and (2) there does not exist a likelihood function f_θ in $\mathcal{F}$ with a simpler posterior that approximates $p(x)$ well.*

The formal statement and proof of Theorem 1 is in Appendix B. This theorem tells us that, under conditions (1) and (2), the ELBO can prefer learning likelihood functions f_θ that reconstruct $p(x)$ poorly, even when learning the ground truth likelihood $f_{\theta_{\text{GT}}}$ is possible!

Pathology II: The ELBO biases learning of the observation noise variance 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 ϵ either via hyper-parameter search or via direct optimization of the ELBO can be biased.

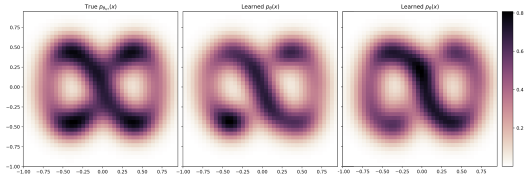
Theorem 2. *For fixed (θ, ϕ), the negative ELBO is minimized by setting ${\sigma^{(d)}}_\epsilon^2$ for each dimension d equal to:*

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\phi(z|x_n)} \left[(x_n^{(d)} - f_\theta(z)^{(d)})^2 \right]. \quad (5)$$

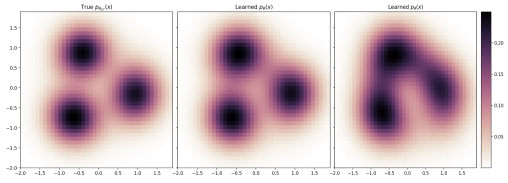
Proof in Appendix C. This theorem shows that the variance σ_ϵ^2 that minimizes the negative ELBO depends on the approximate posterior $q_\phi(z|x)$, and thus, even when θ is set to θ_{GT}, the learned σ_ϵ^2 may not equal ground truth σ_ϵ^2 if $q_\phi(z|x_n)$ is not the true posterior.

2.2. Empirical Demonstrations of VAE Pathologies

We empirically verify that learning σ_ϵ^2 using the ELBO yields biased estimate as shown in Theorem 2. We also give examples wherein global optima of the VAE training objective correspond to poor generative models. To show that this failure is due to pathologies identified in Theorem 1, we verify that: (A) the learned models have simple posteriors for high mass regions where the ground truth models do



(a) **Figure-8**. Left: true, Mid: IWAE, Right: VAE.



(b) **Clusters**. Left: true, Mid: IWAE, Right: VAE.

Figure 1. Comparison of true data distributions versus the corresponding learned distributions of VAE and IWAE. For these examples, since all conditions of Theorem 1 are satisfied, VAE training approximates p(x) poorly and IWAE performs better.

not, (B) training with IWAE (complex variational families) results in generally superior generative models and (C) VAE training cannot be improved meaningfully by methods designed to escape bad local optima, ex. Lagging Inference Networks (LIN) (He et al., 2019). While examples here are synthetic in order to provide intuition for general failures on down-stream tasks, in Section 4 we describe how each example typifies a class of real datasets on which VAEs can exhibit training pathologies. On all synthetic data, we compare models found as close as possible at the global optima of the ELBO as described in Appendix J.

Theorem 2 implies that the ELBO biases noise variance estimates. Consider the "Spiral Dots" Example in Appendix H.5. We perform two experiments. In the first, we initialize VAE training at ground-truth noise variance ($\sigma_\epsilon^2 = 0.01$), generative and inference models ($\theta_{\text{GT}}, \phi_{\text{GT}}$), as well at random values. We then select the learned model (θ, ϕ) with the highest ELBO and compute the noise variance that maximizes the ELBO fixing (θ, ϕ). The noise variance estimated by the best model (θ, ϕ) is 0.014 ± 0.001 across 5 trials. In the second experiment, we maximize the ELBO jointly over σ_ϵ^2, θ and ϕ (initialized at the ground truth as well as randomly). The σ_ϵ^2 corresponding to the restart with the highest ELBO is 0.020 ± 0.003. The ELBO therefore over-estimates the noise variance by 50% and 100%, respectively.

Approximation of $p(x)$ is poor when Conditions (1) and (2) of Theorem 1 both hold. Consider the "Figure-8" Example visualized in Figure 1a (details in Appendix H.1). Here, the posterior matching objective (PMO) is high for many x's, since in the neighborhood of $x \approx 0$ (where $p(x)$ is high), values of z in $[-\infty, -3.0] \cup [3.0, \infty]$ all map to similar values of x. As such, near $x = 0$, the posteriors $p_{\theta_{\text{GT}}}(z|x)$ are multi-modal (Appendix Figure 8d), satisfying condition (1) . We verify condition (2) is satisfied by considering all continuous parametrizations of the "Figure-8" curve: any such parametrization will result in a function f_θ for which distant values of z map to similar values near $x = 0$ and thus the PMO will be high in a neighborhood of $x = 0$. As predicted by Theorem 1, the learned generative model approximates $p(x)$ poorly (Appendix Figure 8a) in order to learn posteriors that are simpler than those of the ground truth model (Appendix Figures 8e vs. 8d)).

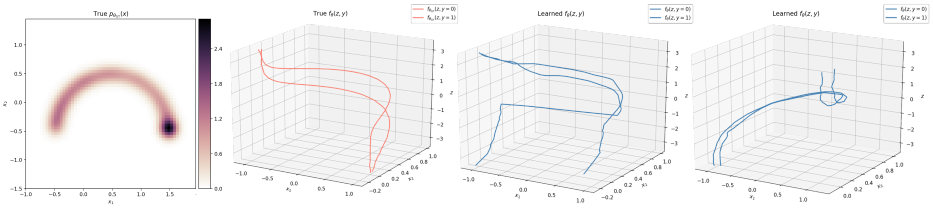
To show that these issues occur because the MFG variational family over-regularizes the generative model, we compare VAE with LIN and IWAE. As expected, IWAE learns $p(x)$ better than LIN, which outperforms the VAE (Figure 1a, Appendix Table 1). Like the VAE, LIN compromises learning the data distribution in order to learn simpler posteriors, since it also uses a MFG variational family (Appendix Figure 9). In contrast, IWAE is able to learn more complex posteriors and thus approximates $p(x)$ far better (Appendix Figure 10). See Appendix E for a more nuanced discussion of Theorem 1's conditions.

3. Impact on Downstream Tasks

Using real and synthetic datasets, we demonstrate concrete ways in which VAE training pathologies described in Theorems 1 & 2 negatively impact performance on downstream tasks. On unsupervised tasks, we show that IWAE does not suffer from the over-regularization of the generative model caused by the inference model, while LIN does. But surprisingly, IWAE does not always outperform the VAE on our semi-supervised tasks as its complex variational family allows the generative model to overfit.

Effects on Unsupervised Downstream Tasks. 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. In Appendix F, we show that, when the ground truth likelihood function has complex posteriors, the VAE objective *prefers* likelihoods that have simple posteriors with representations that entangle ground truth dimensions. This can render the learned representations uninterpretable.

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 σ_ϵ^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. Experiment results in Appendix F.



(a) True $p(x)$ (b) True vs. Learned $f_\theta(y, z)$. Left: true, Mid: IWAE, Right: VAE.

Figure 2. **Discrete Semi-Circle**. Comparison of VAE and IWAE on a semi-supervised example. The ground truth likelihood shows two distinct functions, one for each $y = 0, 1$. The VAE likelihood is over-regularized by an inflexible variational family and learns two nearly identical functions. The IWAE likelihood function is unregularized and learns two distinct but overfitted functions.

Lastly, we show how bias in VAE noise estimation impacts a task requiring data de-noising: manifold-based defense against adversarial attacks. In this task, classifiers makes predictions values projected onto the data manifold so as to be robust to adversarial perturbations (see Appendix F).

Effects on Semi-Supervised Downstream Tasks. In real datasets, we often have samples from multiple cohorts of the population. General characteristics of the population hold for all cohorts, but each cohort may have different distributions of these characteristics. We formalize this in our model by requiring the cohorts to lie on a shared fixed manifold, while each $p(x|y)$ has a different density on that manifold (Klys et al., 2018).

In our model, the ground truth posterior $p_{\theta_{\text{GT}}}(z|x) = \int_y p_{\theta_{\text{GT}}}(z, y|x)dy$ will be multi-modal, since for each value of y there are a number of different likely z's, each from a different cohort. As such, using a MFG variational family in the semi-supervised objective (Equation 28) will encourage inference to either compromise learning the data-distribution in order to better approximate the posterior, or to learn the data distribution well and poorly approximate the posterior, depending on our prioritization of the two objectives (indicated by our the choice of the hyperparameter γ). In the first case, data generation will be compromised but the model will be able to generate realistic counterfactuals. That is, fixing y will allow us to generate realistic data from different cohorts $p(x|y)$. In the latter case, the learned model will be able to generate realistic data but not realistic counterfactuals since the model will have collapse the conditional distributions $p_\theta(x|y) \approx p(x)$. That is, $p(x|y)$ will generates identical looking cohort regardless of our choice of y. In short, ***VAEs trades-off between generating realistic data and realistic counterfactuals.***

In Appendix G, we show that when y is discrete, VAEs struggle with balancing generating realistic data and realistic counterfactuals as expected. Surprisingly because the likelihood of IWAE is not regularized by an inflexible variational family and overfits - while IWAE generates realistic counterfactuals, it approximates $p(x)$ poorly. We also show (in Appendix G) that when y is continuous, IWAE surprisingly struggle with generating realistic counterfactuals when the discriminator $q_\phi(y|x)$ is a MFG.

4. Discussion and Conclusion

VAE training pathologies negatively impact down stream tasks. In Section 3 we showed that due to the training pathology we identified in Theorems 1 & 2, VAEs may struggle with approximating $p(x)$, learning compressed and disentangled representations, and VAEs struggle with tasks requiring de-noising. In semi-supervised settings, VAEs trade-off generating realistic data with generating realistic counterfactual data. Moreover, these problems occur at global optima of the VAE training objective. While we show that on unsupervised tasks these issues are mitigated when we train with IWAE, the under-regularized generative models of IWAE can overfit and perform no better than VAEs on semi-supervised tasks.

VAE training pathologies can happen for many real datasets. We describe how the conditions of Theorem 1 manifests in real datasets. The "Figure-8" Example in Figure 1a generalizes to any data manifold where the Euclidean distance between two points in a high density region on manifold is (A) less than the length of the geodesic connecting and (B) within 2 standard deviation of observation noise. The "Clusters" Example in Figure 1b generalizes to data distributions that have distinct areas of high density connected by areas of low density. On these datasets, the VAE training objective prefers compromising the quality of the generative model for posteriors that are easy to approximate. As for the pathology noted in Theorem 2, we expect that the ELBO yields biased estimates of the observation noise whenever the learned model approximates $p(x)$ poorly.

Conclusion We concretely characterized conditions under which VAE training exhibits pathologies and connected these pathologies to undesirable effects on specific downstream tasks. We find that while inference with richer variational families (at a computational cost) can alleviate these issues on unsupervised tasks, they can introduce unexpected new pathologies in semi-supervised settings.

Acknowledgments

YY acknowledges support from NIH 5T32LM012411-04 and from IBM Research. WP acknowledges support from the Harvard Institute of Applied Computational Sciences.

References

Alcala-Fdez, J., Fernández, A., Luengo, J., Derrac, J., Garcia, S., Sanchez, L., and Herrera, F. 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.

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a Broken ELBO. *arXiv e-prints*, art. arXiv:1711.00464, November 2017.

Antal, B. and Hajdu, A. An ensemble-based system for automatic screening of diabetic retinopathy. *arXiv e-prints*, art. arXiv:1410.8576, October 2014.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance Weighted Autoencoders. *arXiv:1509.00519 [cs, stat]*, November 2016. URL http://arxiv.org/abs/1509.00519. arXiv: 1509.00519.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv e-prints*, art. arXiv:1802.04942, February 2018.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting Importance-Weighted Autoencoders. *arXiv:1704.02916 [stat]*, August 2017. URL http://arxiv.org/abs/1704.02916. arXiv: 1704.02916.

Cremer, C., Li, X., and Duvenaud, D. Inference Suboptimality in Variational Autoencoders. *arXiv:1801.03558 [cs, stat]*, May 2018. URL http://arxiv.org/abs/1801.03558. arXiv: 1801.03558.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models. pp. 42.

Dai, B., Wang, Z., and Wipf, D. The Usual Suspects? Reassessing Blame for VAE Posterior Collapse. *arXiv e-prints*, art. arXiv:1912.10702, December 2019.

Djolonga, J. and Krause, A. Learning Implicit Generative Models Using Differentiable Graph Tests. *arXiv e-prints*, art. arXiv:1709.01006, Sep 2017.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ghosh, P., Losalka, A., and Black, M. J. Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. *arXiv e-prints*, art. arXiv:1806.00081, May 2018.

Gregor, K., Besse, F., Jimenez Rezende, D., Danihelka, I., and Wierstra, D. Towards conceptual compression. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3549–3557. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6542-towards-conceptual-compression.pdf.

He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. *arXiv:1901.05534 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1901.05534. arXiv: 1901.05534.

Hwang, U., Park, J., Jang, H., Yoon, S., and Cho, N. I. PuVAE: A Variational Autoencoder to Purify Adversarial Examples. *arXiv e-prints*, art. arXiv:1903.00585, March 2019.

Jalal, A., Ilyas, A., Daskalakis, C., and Dimakis, A. G. The Robust Manifold Defense: Adversarial Training using Generative Models. *arXiv e-prints*, art. arXiv:1712.09196, December 2017.

Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. *arXiv e-prints*, art. arXiv:1611.01144, November 2016.

Jang, U., Jha, S., and Jha, S. ON THE NEED FOR TOPOLOGY-AWARE GENERATIVE MODELS FOR MANIFOLD-BASED DEFENSES. pp. 24, 2020.

Kim, Y., Wiseman, S., Miller, A. C., Sontag, D., and Rush, A. M. Semi-Amortized Variational Autoencoders. *arXiv e-prints*, art. arXiv:1802.02550, February 2018.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014.

Kingma, D. P. and Dhariwal, P. Glow: Generative Flow with Invertible 1x1 Convolutions. *arXiv:1807.03039 [cs, stat]*, July 2018. URL http://arxiv.org/abs/1807.03039. arXiv: 1807.03039.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv e-prints*, art. arXiv:1312.6114, December 2013.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-Supervised Learning with Deep Generative Models. *arXiv e-prints*, art. arXiv:1406.5298, June 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving Variational Inference with Inverse Autoregressive Flow. *arXiv e-prints*, art. arXiv:1606.04934, June 2016.

Klys, J., Snell, J., and Zemel, R. Learning Latent Subspaces in Variational Autoencoders. *arXiv e-prints*, art. arXiv:1812.06190, December 2018.

Klys, J., Snell, J., and Zemel, R. Learning latent subspaces in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 6444–6454, 2018.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. 69(6):066138, June 2004. doi: 10.1103/PhysRevE.69.066138.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv e-prints*, art. arXiv:1811.12359, November 2018.

Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse. *arXiv e-prints*, art. arXiv:1911.02469, November 2019.

Luo, Y. and Pfister, H. Adversarial Defense of Image Classification Using a Variational Auto-Encoder. *arXiv e-prints*, art. arXiv:1812.02891, December 2018.

Luo, Y., Beatson, A., Norouzi, M., Zhu, J., Duvenaud, D., Adams, R. P., and Chen, R. T. Q. SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models. *arXiv:2004.00353 [cs, stat]*, April 2020. URL http://arxiv.org/abs/2004.00353. arXiv: 2004.00353.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial Autoencoders. *arXiv e-prints*, art. arXiv:1511.05644, November 2015.

Meng, D. and Chen, H. MagNet: a Two-Pronged Defense against Adversarial Examples. *arXiv e-prints*, art. arXiv:1705.09064, May 2017.

Miao, Y. and Blunsom, P. Language as a Latent Variable: Discrete Generative Models for Sentence Compression. *arXiv e-prints*, art. arXiv:1609.07317, September 2016.

Nowozin, S. DEBIASING EVIDENCE APPROXIMATIONS: ON IMPORTANCE-WEIGHTED AUTOENCODERS AND JACKKNIFE VARIATIONAL INFERENCE. pp. 16, 2018.

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., and Carin, L. Variational autoencoder for deep learning of images, labels and captions. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2352–2360. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6528-variational-autoencoder-for-deep-learning-of.pdf.

Ridgeway, K. A Survey of Inductive Biases for Factorial Representation-Learning. *arXiv e-prints*, art. arXiv:1612.05299, December 2016.

Roeder, G., Wu, Y., and Duvenaud, D. Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference. *arXiv e-prints*, art. arXiv:1703.09194, March 2017.

Rolinek, M., Zietlow, D., and Martius, G. Variational Autoencoders Pursue PCA Directions (by Accident). *arXiv:1812.06775 [cs, stat]*, April 2019. URL http://arxiv.org/abs/1812.06775. arXiv: 1812.06775.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *arXiv e-prints*, art. arXiv:1805.06605, May 2018.

Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N. D., Kohli, P., Wood, F., and Torr, P. H. S. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. *arXiv e-prints*, art. arXiv:1706.00400, June 2017.

Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N. D., Kohli, P., Wood, F., and Torr, P. H. S. 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.

Simonoff, J. The "unusual episode" and a second statistics course. *Journal of Statistics Education*, 5, 03 1997. doi: 10.1080/10691898.1997.11910524.

Theis, L., Oord, A. v. d., and Bethge, M. 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.

Tomczak, J. M. and Welling, M. VAE with a VampPrior. *arXiv e-prints*, art. arXiv:1705.07120, May 2017.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives. *arXiv:1810.04152 [cs, stat]*, November 2018. URL http://arxiv.org/abs/1810.04152. arXiv: 1810.04152.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural Discrete Representation Learning. *arXiv e-prints*, art. arXiv:1711.00937, November 2017.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. 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.

Yacoby, Y., Pan, W., and Doshi-Velez, F. Characterizing and Avoiding Problematic Global Optima of Variational Autoencoders. *arXiv e-prints*, art. arXiv:2003.07756, March 2020.

Zhao, S., Song, J., and Ermon, S. Towards Deeper Understanding of Variational Autoencoding Models. *arXiv:1702.08658 [cs, stat]*, February 2017. URL http://arxiv.org/abs/1702.08658. arXiv: 1702.08658.

A. Related Work

Existing work that characterize MFG-VAEs pathologies primarily focus on relating local optima of the training objective to a single pathology: the un-informativeness of the learned latent codes (posterior collapse) (He et al., 2019; Lucas et al., 2019; Dai et al., 2019). In contrast, there has been little work to characterize pathologies at the global optima of the MFG-VAEs training objective. (Yacoby et al., 2020) shows that, when the decoder's capacity is restricted, posterior collapse and the mismatch between aggregated posterior and prior can occur as global optima of the training objective. In contrast to existing work, we focus on global optima of the MFG-VAE objective in fully general settings: with fully flexible generative and inference models, as well as with and without learned observation noise.

While works focused on improving VAEs have noted the over-regularizing effect of the variational family on the generative model (e.g. (Burda et al., 2016; Zhao et al., 2017; Cremer et al., 2018)), none have given a full characterization of the conditions under which the learned generative model is meaningfully compromised. Nor have they related the resulting bias in VAE training to potentially impactful effects on down-stream tasks. In particular, these works have shown that their proposed methods have higher test log-likelihood relative to a MFG-VAEs, but as we show in this paper, high test log-likelihood is not the only property needed for good performance on downstream tasks. Lastly, these works all propose fixes that require a potentially significant computational overhead. For instance, works that use complex variational families, such as normalizing flows (Kingma et al., 2016), require a significant number of parameters to scale (Kingma & Dhariwal, 2018). As in the case of the Importance Weighted Autoencoder (IWAE) objective (Burda et al., 2016), which can be interpreted as having a more complex variational family (Cremer et al., 2017), the complexity of the posterior scales with the number of importance samples used. Lastly, works that de-bias or reduce the variance of existing bounds (Nowozin, 2018; Luo et al., 2020; Tucker et al., 2018; Roeder et al., 2017) all require several evaluations of the objective.

Given that MFG-VAEs remain popular today due to the ease of their implementation, speed of training, and their theoretical connections to other dimensionality reduction approaches like probabilistic PCA (Rolinek et al., 2019; Dai et al.; Lucas et al., 2019), we believe such a characterization of MFG-VAE training pathologies is valuable and that concrete connections from these pathologies to down-stream effects are needed. More importantly, we believe this characterization will help clarify for which tasks / datasets a MFG-VAE suffices and for which the computational tradeoffs are worth it.

B. Theorem 1

We formalize these conditions in the following. Recall the decomposition the negative ELBO in Equation 2. In the following discussion, we alway set ϕ to be optimal for our choice of θ. Assuming that $p(x)$ is continuous, then for any $\alpha \in \mathbb{R}$, we can further decompose the PMO:

$$\begin{aligned}&\mathop{\mathbb{E}}_{p(x)}\left[D_{\mathrm{KL}}[q_\phi(z|x)||p_\theta(z|x)]\right]\\ &=\Pr[\mathcal{X}_{\mathrm{Lo}}(\theta)]\,\mathbb{E}_{p(x)|_{\mathcal{X}_{\mathrm{Lo}}}}\left[D_{\mathrm{KL}}[q_\phi(z|x)||p_\theta(z|x)]\right]\\ &\quad+\Pr[\mathcal{X}_{\mathrm{Hi}}(\theta)]\,\mathbb{E}_{p(x)|_{\mathcal{X}_{\mathrm{Hi}}}}\left[D_{\mathrm{KL}}[q_\phi(z|x)||p_\theta(z|x)]\right]\end{aligned} \tag{6}$$

where $D_{\mathrm{KL}}[q_\phi(z|x)||p_\theta(z|x)] \leq \alpha$ on $\mathcal{X}_{\mathrm{Lo}}(\theta)$, $D_{\mathrm{KL}}[q_\phi(z|x)||p_\theta(z|x)] > \alpha$ on $\mathcal{X}_{\mathrm{Hi}}(\theta)$, 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 6 as $D_{\mathrm{Lo}}(\theta)$ and the expectation in the second term as $D_{\mathrm{Hi}}(\theta)$.

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

$$\Pr[\mathcal{X}_{\mathrm{Hi}}(\theta_{\mathrm{GT}})]\,(D_{\mathrm{Hi}}(\theta_{\mathrm{GT}}) - D_{\mathrm{Lo}}(\theta_{\mathrm{GT}}))$$

is greater than

$$\Pr[\mathcal{X}_{\mathrm{Hi}}(\theta)]\,D_{\mathrm{Hi}}(\theta) + D_{KL}[p(x)||p_\theta(x)];$$

suppose also that (2) that for no such $f_\theta \in \mathcal{F}$ is the MLEO $D_{KL}[p(x)||p_\theta(x)]$ equal to zero. Then at the global minima (θ^*, ϕ^*) 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_θ will be lower than that of $f_{\theta_{\mathrm{GT}}}$. That is, we can write:

$$-\mathrm{ELBO}(\theta_{\mathrm{GT}}, \phi_{\mathrm{GT}}) \tag{7}$$

$$= \Pr[\mathcal{X}_{\mathrm{Hi}}(\theta_{\mathrm{GT}})]\,D_{\mathrm{Hi}}(\theta_{\mathrm{GT}}) + \Pr[\mathcal{X}_{\mathrm{Lo}}(\theta_{\mathrm{GT}})]\,D_{\mathrm{Lo}}(\theta_{\mathrm{GT}}) \tag{8}$$

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

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

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

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

C. 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 ϵ either via hyper-parameter search or via direct optimization of the ELBO can be biased.

Theorem 2. *For an observation set of size N, we have that*

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

Proof. We rewrite the argmin of the negative ELBO as follows: as follows:

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

$$= \underset{{\sigma^{(d)}}_\epsilon^2}{\text{argmin}} \sum_{d=1}^{D} \log\left({\sigma^{(d)}}_\epsilon\right) + \frac{1}{2{\sigma^{(d)}}_\epsilon^2} \cdot C(\theta, \phi, d) \tag{14}$$

where

$$C(\theta, \phi, d) = \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[(x^{(d)} - f_\theta(z)^{(d)})^2\right]\right] \tag{15}$$

(see Figure 3 for details). Setting the gradient of the above with respect to σ_ϵ^2 equal to zero yields the following:

$$0 = -\frac{\partial}{\partial \sigma_\epsilon^{(d)}} \text{ELBO}(\theta, \phi, \sigma_\epsilon^{(d)}) \tag{16}$$

$$= \frac{{\sigma^{(d)}}_\epsilon^2 - C(\theta, \phi, d)}{{\sigma^{(d)}}_\epsilon^3}. \tag{17}$$

Thus, we can write,

$${\sigma^{(d)}}_\epsilon^2 = C(\theta, \phi, d) = \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[(x^{(d)} - f_\theta(z)^{(d)})^2\right]\right] \tag{18}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_\phi(z|x_n)}\left[(x_n^{(d)} - f_\theta(z)^{(d)})^2\right] \tag{19}$$

□

D. 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 in (Kingma et al., 2014) as the "M2 model". We assume the following generative process:

$$\begin{gathered}z \sim \mathcal{N}(0, I), \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I), \quad y \sim p(y), \\ x|y, z = f_\theta(y, z) + \epsilon\end{gathered} \tag{27}$$

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:

$$\mathcal{J}(\theta, \phi) = \sum_{n=1}^{N} \mathcal{U}(x_n; \theta, \phi) + \gamma \cdot \sum_{m=1}^{M} \mathcal{L}(x_m, y_m; \theta, \phi) \tag{28}$$

where $\mathcal{U}$ and $\mathcal{L}$ lower bound $p_\theta(x)$ and $p_\theta(x, y)$, respectively, and γ controls their relative weight (as done in (Siddharth et al., 2017)). See Figures 4a and 4b for the definition of $\mathcal{U}, \mathcal{L}$ for a VAE and for IWAE, respectively.

E. Empirical Demonstrations of VAE Training Pathologies

Experiment setup We train each model to approximately 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 lowest value of the objective function. 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. Details in Appendix J.

Approximation of $p(x)$ may be fine when only condition (2) holds. What happens if the observations with highly

$$\underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} -\text{ELBO}(\theta, \phi, \sigma^2_\epsilon) \tag{20}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[-\log p_\theta(x|z)\right] + D_{\text{KL}}\left[q_\phi(z|x)||p(z)\right]\right] \tag{21}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[-\log p_\theta(x|z)\right]\right] \tag{22}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[-\sum_{d=1}^{D}\log\left(\frac{1}{\sqrt{2\pi\sigma^{(d)2}_\epsilon}}\cdot\exp\left(\frac{-(x^{(d)}-f_\theta(z)^{(d)})^2}{2\sigma^{(d)2}_\epsilon}\right)\right)\right]\right] \tag{23}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \sum_{d=1}^{D}\mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\sqrt{2\pi\sigma^{(d)2}_\epsilon}\right) + \frac{(x^{(d)}-f_\theta(z)^{(d)})^2}{2\sigma^{(d)2}_\epsilon}\right]\right] \tag{24}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \sum_{d=1}^{D}\mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\sigma^{(d)}{}_\epsilon\right) + \frac{(x^{(d)}-f_\theta(z)^{(d)})^2}{2\sigma^{(d)2}_\epsilon}\right]\right] \tag{25}$$

$$= \underset{\sigma^{(d)2}_\epsilon}{\text{argmin}} \quad \sum_{d=1}^{D}\log\left(\sigma^{(d)}{}_\epsilon\right) + \frac{1}{2\sigma^{(d)2}_\epsilon}\cdot\underbrace{\mathbb{E}_{p(x)}\left[\mathbb{E}_{q_\phi(z|x)}\left[(x^{(d)}-f_\theta(z)^{(d)})^2\right]\right]}_{C(\theta,\phi,d)} \tag{26}$$

Figure 3. Derivation in Theorem 2

$$\log p_\theta(x,y) \geq \underbrace{\mathbb{E}_{q_\phi(z|x,y)}\left[-\log p_\theta(x|y,z)\right] - \log p(y) + D_{\text{KL}}\left[q_\phi(z|x,y)||p(z)\right]}_{\mathcal{L}(x,y;\theta,\phi)} \tag{29}$$

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{q_\phi(y|x)q_\phi(z|x)}\left[-\log p_\theta(x|y,z)\right] + D_{\text{KL}}\left[q_\phi(y|x)||p(y)\right] + D_{\text{KL}}\left[q_\phi(z|x)||p(z)\right]}_{\mathcal{U}(x;\theta,\phi)} \tag{30}$$

(a) VAE Semi-Supervised Bounds

$$\log p_\theta(x,y) \geq \underbrace{\mathbb{E}_{z_1,\ldots,z_S\sim q_\phi(z|x,y)}\left[\log\frac{1}{S}\frac{p_\theta(x,y,z_s)}{q_\phi(z_s|x,y)}\right]}_{\mathcal{L}(x,y;\theta,\phi)} \tag{31}$$

$$\log p_\theta(x) \geq \underbrace{\mathbb{E}_{(y_1,z_1),\ldots,(y_S,z_S)\sim q_\phi(y|x)q_\phi(z|x)}\left[\log\frac{1}{S}\sum_{s=1}^{S}\frac{p_\theta(x,y_s,z_s)}{q_\phi(y_s|x)q_\phi(z_s|x)}\right]}_{\mathcal{U}(x;\theta,\phi)} \tag{32}$$

(b) IWAE Semi-Supervised Bounds

Figure 4.

non-Gaussian posterior were few in number? For instance, consider the "Circle" Example visualized in Figure 6 (details in Appendix H.2). In this example, the regions of the data-space that have a non-Gaussian posterior are near $x \approx (1.0, 0.0)$, since $z \in [-\infty, -3.0] \cup [3.0, \infty]$ map to points near $(1.0, 0.0)$. However, since overall number of such points is small, the VAE objective does not need to trade-off capturing $p(x)$ for easy posterior approximation. Indeed, we see that VAE training is capable of recovering $p(x)$, regardless of whether training was initialized randomly or at the ground truth (Appendix Figure 6).

Approximation of $p(x)$ may be fine when only condition (1) holds. We now study the case where the true posterior has a high PMO for a large portion of x's, but there exist a

f_θ in our feasible set $\mathcal{F}$ that approximates $p(x)$ well and has simple posteriors. Consider the "Absolute-Value" Example visualized in Figure 7 (details in Appendix H.3). That the true posteriors are complex (Appendix Figure 7d). However, there is an alternative likelihood $f_\theta(z)$ that explains $p(x)$ equally well and has simpler posteriors (Appendix Figure 7e) and this is the model selected by the VAE objective, regardless of whether training was initialized randomly or at the ground truth.

Approximation of $p(x)$ is poor when Conditions (1) and (2) of Theorem 1 both hold. Table 1 shows that on synthetic data-sets for which Theorem 1 hold, the VAE objective (even with a better training algorithm, LIN) approximates $p(x)$ poorly, while methods with a more complex variational family (IWAE) do not. Visualization of the posterior (in Appendix M) confirm that the VAE objective underfits the generative model in order to learn a simpler posterior, whereas the IWAE objective does not: for the "Figure-8 Example", see Figures 8, 9 and 10 and for the "Clusters Example", see Figures 11, 12 and 13). In these two examples, we further see the ELBO's regularizing effect on the learned f_θ. On the "Figure-8 Example", the learned f_θ 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_θ has less pronounced changes in slope, and thus a simpler posterior.

F. Effects of Pathologies on Unsupervised Learning Downstream Tasks

VAE training pathologies prevent learning disentangled representations 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 may *prefer* learning the representations that *entangles* the ground-truth latent dimensions.

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. Let $A' = AR$, where we define $R = (\Sigma V^\top)^{-1}(\Lambda - \sigma^2 I)^{1/2}$ with an arbitrary diagonal matrix Λ and matrices Σ, 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, $p_\theta(x) = p_{\theta_{\text{GT}}}(x) = \mathcal{N}(b, \sigma_\epsilon^2 \cdot I + AA^\intercal)$. However, since the posteriors of f_θ are uncorrelated, the ELBO will prefer f_θ over $f_{\theta_{\text{GT}}}$! In the latent space corresponding to f_θ , the original *interpretations* of the latent dimensions are now entangled. Similarly, for more complicated likelihood functions, we should 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.

VAE training pathologies prevent learning compressed representations In practice, if the task does not require a specific latent space dimensionality, K, one chooses K that maximizes the $\log p_\theta(x)$. As we demonstrate here, using a higher K and a lower σ_ϵ^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 embedding $A = \left(\begin{smallmatrix} 1.0 & 0.0 & 0.5 & 0.2 & -0.8 \\ 0.0 & 1.0 & -0.5 & 0.3 & -0.1 \end{smallmatrix}\right)$, 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 σ_ϵ^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 (Appendix Table 2), since the latent space learns to generate the observation noise ϵ. We confirm that the posteriors become simpler as K increases, lessening the incentive for the VAE to compromise on approximating $p(x)$ (Appendix Figures 22 and 21). We confirm that while LIN also shows preference for higher K's, IWAE does not (Appendix Table 2).

VAE training pathologies compromises defense against adversarial perturbations We've shown that VAEs prefer increasing the dimensionality of the latent space K and decreasing σ_ϵ^2. While these models have better approximate $p(x)$, they explain variance due to ϵ using variance due to z, and therefore do not correctly de-noise the data. Furthermore, even when K is fixed at the ground truth, Theorem 2 shows that the ELBO is unable to identify the correct σ_ϵ^2. Unfortunately, bias in the noise variance estimate will degrade the performance on tasks requiring correct estimation of the noise. An example of such task is manifold-based defense against adversarial attacks, in which classifiers makes predictions values projected onto the data manifold so as to be robust to adversarial perturbations (Jalal et al., 2017; Meng & Chen, 2017; Samangouei et al., 2018; Hwang et al., 2019; Jang et al., 2020). See Appendix K for full analysis.

G. Effect on Semi-Supervised Downstream Tasks

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 2a (details in Appendix I.1). the VAE is able to learn the data manifold and distribution well (Appendix Figure 15a). However, the learned model has a simple posterior in comparison to the true posterior (Appendix Figure 15f). In fact, the learned $f_\theta(z, y)$ is collapsed to the same function for all values of y (Figure 2b). As a result, $p_\theta(x|y) \approx p_\theta(x)$ under the learned model (Appendix Figure 15c). As expected, the same phenomenon occurs when training with LIN (Appendix Figure 16). In contrast, IWAE is able to learn two distinct data conditionals $p_\theta(x|y)$, but it does so at a cost. ***Since the IWAE does not regularize the generative model, this leads to overfitting*** (Figure 2b). Appendix Table 3 shows that IWAE learns $p(x)$ considerably worse than the VAE, while Appendix Table 4 shows that it learns the $p(x|y)$ significantly better. On real data we see similar patterns: Table 5 shows that while IWAE generally approximates $p(x)$ better than a VAE (and thus does not overfit in this case), both are unable to to learn $p(x|y)$ well (see Tables 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 4). Consider the "Continuous Semi-Circle" Example, visualized in Appendix Figure 18b (details in Appendix I.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 (Appendix Figure 18b). So as we increase α (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 18c). Like in the discrete case, γ still controls the tradeoff between realistic data and realistic counterfactuals; in the continuous case, α *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***.

H. 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_θ using full supervision (ensuring that the $\text{MSE} < 1\text{e}-4$ and use that f_θ to generate the actual data used in the experiments.

H.1. Figure-8 Example

Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= (0.6 + 1.8 \cdot \Phi(z))\,\pi \\ x|z &= \underbrace{\begin{bmatrix} \frac{\sqrt{2}}{2} \cdot \frac{\cos(u(z))}{\sin(u(z))^2+1} \\ \sqrt{2} \cdot \frac{\cos(u(z))\sin(u(z))}{\sin(u(z))^2+1} \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \tag{33}$$

where $\Phi(z)$ is the Gaussian CDF and $\sigma_\epsilon^2 = 0.02$ (see Figure 8).

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 8d), 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_θ for which far-away values of z lead to nearby values of x and thus in high PMO value for points near $x = 0$.

H.2. Circle Example

Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &= \underbrace{\begin{bmatrix} \cos(2\pi \cdot \Phi(z)) \\ \sin(2\pi \cdot \Phi(z)) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \tag{34}$$

where $\Phi(z)$ is the Gaussian CDF and $\sigma_\epsilon^2 = 0.01$ (see Figure 6).

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 6).

H.3. Absolute-Value Example

Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &= \underbrace{\begin{bmatrix} |\Phi(z)| \\ |\Phi(z)| \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \tag{35}$$

where $\Phi(z)$ is the Gaussian CDF and $\sigma_\epsilon^2 = 0.01$ (see Figure 7).

Properties: In this example, the posterior under $f_{\theta_{\text{GT}}}$ cannot be well approximated using a MFG variational family (see Figure 7d). However, there does exist an alternative likelihood function $f_\theta(z)$ (see 7b) that explains $p(x)$ equally well and has simpler posterior 7e. As such, this model only satisfies condition 1 of Theorem 1.

H.4. Clusters Example

Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= \frac{2\pi}{1 + e^{-\frac{1}{2}\pi z}} \\ t(u) &= 2 \cdot \tanh\left(10 \cdot u - 20 \cdot \lfloor u/2 \rfloor - 10\right) + 4 \cdot \lfloor u/2 \rfloor + 2 \\ x|z &= \underbrace{\begin{bmatrix} \cos(t(u(z))) \\ \sin(t(u(z))) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \tag{36}$$

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

Properties: In this example, $f_{\theta_{\text{GT}}}$ a step function embedded on a circle. Regions in which $\frac{df^{-1}_{\theta_{\text{GT}}}}{dx}$ 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 11d). 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 11e) at the cost of approximating $p(x)$ well (see Figure 11a).

H.5. Spiral Dots Example

Generative Model:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ u(z) &= \frac{4\pi}{1 + e^{-\frac{1}{2}\pi z}} \\ t(u) &= \tanh\left(10 \cdot u - 20 \cdot \lfloor u/2 \rfloor - 10\right) + 2 \cdot \lfloor u/2 \rfloor + 1 \\ x|z &= \underbrace{\begin{bmatrix} t(u(z)) \cdot \cos(t(u(z))) \\ t(u(z)) \cdot \sin(t(u(z))) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(z)} + \epsilon \end{aligned} \tag{37}$$

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

Properties: In this example, $f_{\theta_{\text{GT}}}$ a step function embedded on a spiral. Regions in which $\frac{df^{-1}_{\theta_{\text{GT}}}}{dx}$ 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 14d). In this example, we again see that the VAE objective learns a model with a simpler posterior (see Figure 14e) at the cost of approximating $p(x)$ well (see Figure 14a). Furthermore, for this model the VAE objective highly misestimates the observation noise.

I. 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_θ using full supervision (ensuring that the MSE $< 1\text{e}-4$ and use that f_θ to generate the actual data used in the experiments.

I.1. Discrete Semi-Circle Example

Generative Process:

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ y &\sim \text{Bern}\left(\frac{1}{2}\right) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|y,z &= \underbrace{\begin{bmatrix} \cos\left(\mathbb{I}(y=0) \cdot \pi \cdot \sqrt{\Phi(z)} + \mathbb{I}(y=1) \cdot \pi \cdot \Phi(z)^3\right) \\ \sin\left(\mathbb{I}(y=0) \cdot \pi \cdot \sqrt{\Phi(z)} + \mathbb{I}(y=1) \cdot \pi \cdot \Phi(z)^3\right) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(y,z)} + \epsilon \end{aligned} \tag{38}$$

where Φ 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 eeach 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 15b). When both functions collapse to the same function, $p_\theta(x|y) \approx p_\theta(x)$ (see Figure 15c). This will prevent the learned model from generating realistic counterfactuals.

I.2. Continuous Semi-Circle Example

Generative Process:

$$
\begin{aligned}
z &\sim \mathcal{N}(0,1) \\
y &\sim \mathcal{N}(0,1) \\
h(y) &= B^{-1}(\Phi(y); 0.2, 0.2) \\
\epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\
x|y,z &= \underbrace{\begin{bmatrix} \cos\left(h(y)\cdot\pi\cdot\sqrt{\Phi(z)} + (1-h(y))\cdot\pi\cdot\Phi(z)^3\right) \\ \sin\left(h(y)\cdot\pi\cdot\sqrt{\Phi(z)} + (1-h(y))\cdot\pi\cdot\Phi(z)^3\right) \end{bmatrix}}_{f_{\theta_{\text{GT}}}(y,z)} + \epsilon
\end{aligned}
\tag{39}
$$

where Φ is the CDF of a standard normal and $B^{-1}(.;\alpha,\beta)$ 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: when 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 18b) As such, as we increase α, the better our predictive accuracy will be but the more $p_\theta(x|y) \to p_\theta(x)$, causing the learned model to generate poor quality counterfactuals (see Figure 18c).

J. Experimental Details

Initialization at Global Optima of the VAE Objective On all synthetic data, we initialize the model at the ground truth $\theta_{\text{GT}}, \phi_{\text{GT}}$. The decoder function f_θ 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 ϕ_{GT} by fixing the decoder at the ground-truth and maximizing the ELBO (with the 10 random restarts). We fix the observation error σ_ϵ^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_θ 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 H), and 2 synthetic data-sets for semi-supervised VAEs (described in Appendix I), 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 Since existing work shows that on real data IWAE learns generative models with higher log data likelihood (Kingma et al., 2016; Cremer et al., 2017), we only considered synthetic data for the unsupervised tasks. For our semi-supervised tasks, we consider both synthetic data as well as 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). These datasets are selected because their classification is hard, and this is the regime in which we expect semi-supervised VAE training to struggle. For the UCI data-sets, 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 To evaluate the quality of the generative model the smooth kNN 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 (Theis et al., 2016; Wu et al., 2017). In the semi-supervised case, we also use the smooth kNN 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 reparametarization trick (Kingma & Welling, 2013).

Baselines For our baselines, we compare the performance of a vanilla 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 kNN 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, α determines the weight of the discriminator, and γ 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 kNN test statistic (indicating that they learned the $p(x)$ best).

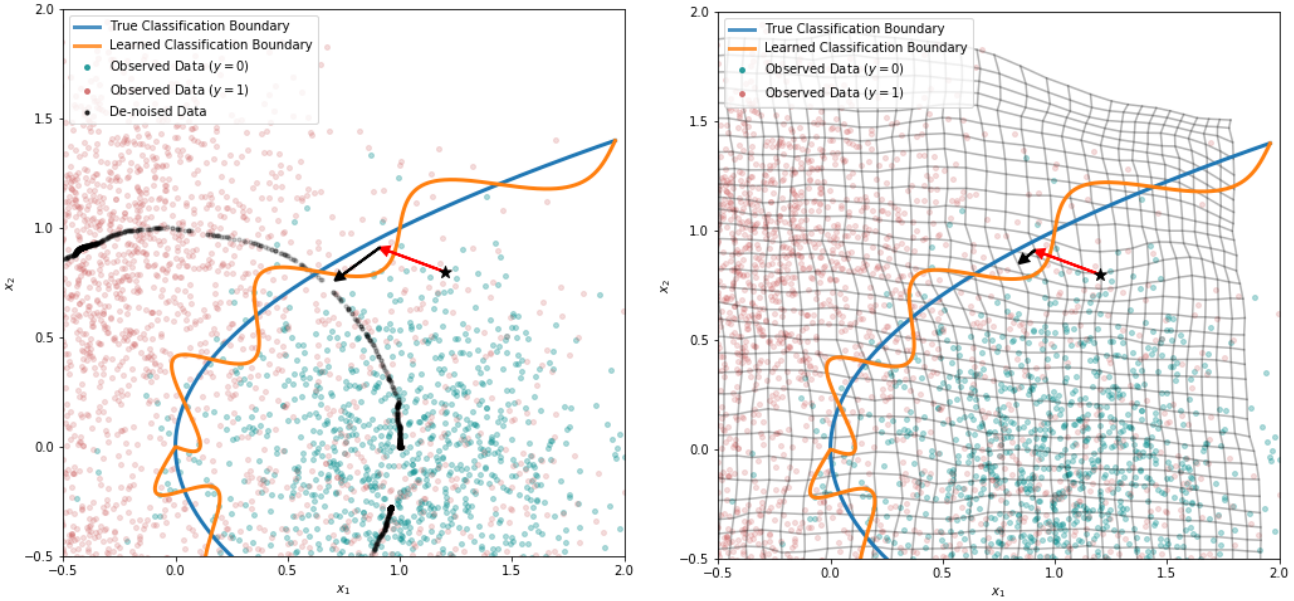
K. Defense Against Adversarial Perturbations Requires Learning the True Observation Noise

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 ϵ (or "signal" and "noise") is necessary to prevent against certain perturbation-based adversarial attacks.

Assume that our data was generated as follows:

$$\begin{aligned} z &\sim p(z) \\ \epsilon &\sim \mathcal{N}(0, \sigma_\epsilon^2 \cdot I) \\ x|z &\sim f_{\theta_{\text{GT}}}(z) + \epsilon \\ y|z &\sim \text{Cat}\left(g_\psi \circ f_{\theta_{\text{GT}}}(z)\right) \end{aligned} \tag{40}$$

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 η 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 5). 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



(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 5. 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.

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 5a: 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, 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 5b 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 ϵ 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.

L. Quantitative Results

Data	IWAE	LIN	VAE
Clusters	$\mathbf{0.057 \pm 0.028}$	0.347 ± 0.057	0.361 ± 0.083
Fig-8	$\mathbf{0.036 \pm 0.013}$	0.040 ± 0.081	0.066 ± 0.014

Table 1. Comparison unsupervised learned vs. true data distributions via the smooth kNN 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	$\mathbf{-0.234 \pm 0.050}$	4.433 ± 0.049	4.385 ± 0.034	$\mathbf{4.377 \pm 0.024}$
Test $\text{avg}_i I(x; z_i)$	$\mathbf{2.419 \pm 0.027}$	1.816 ± 0.037	1.296 ± 0.064	$\mathbf{1.530 \pm 0.011}$	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	$\mathbf{-0.388 \pm 0.044}$	-0.364 ± 0.051	-0.351 ± 0.045	$\mathbf{4.287 \pm 0.047}$	4.298 ± 0.054	4.295 ± 0.049
Test $\text{avg}_i I(x; z_i)$	$\mathbf{2.159 \pm 0.088}$	1.910 ± 0.035	1.605 ± 0.087	1.269 ± 0.052	$\mathbf{1.321 \pm 0.033}$	1.135 ± 0.110

Table 2. The VAE training objective prefers learning models with more latent dimensions (and smaller σ_ϵ^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 ϵ. 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 ± 0.096	0.703 ± 0.315	$\mathbf{0.196 \pm 0.078}$
Continuous Semi-Circle	$\mathbf{0.015 \pm 0.011}$	0.128 ± 0.094	0.024 ± 0.014

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

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

Table 4. Comparison of semi-supervised learned $p_\theta(x|y)$ with ground truth $p(x|y)$ via the smooth kNN test statistic (smaller is better). Hyper-parameters selected via smallest smooth kNN 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 $p(x|y = 3.5)$.

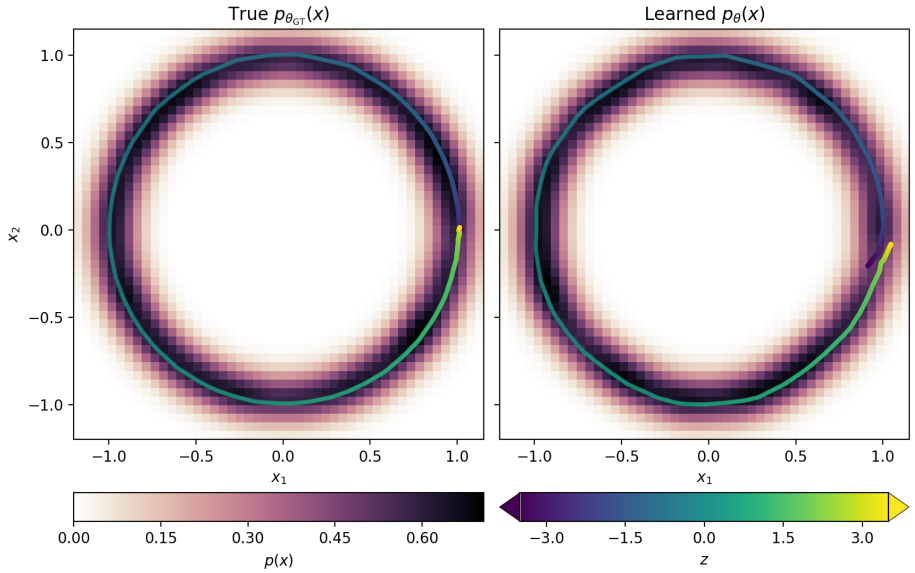
	IWAE	VAE
Diabetic Retinopathy	3.571 ± 2.543	6.206 ± 1.035
Contraceptive	1.740 ± 0.290	2.147 ± 0.225
Titanic	2.794 ± 1.280	1.758 ± 0.193

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

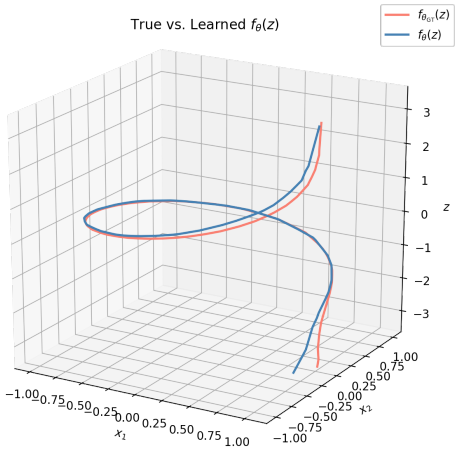
	IWAE			VAE		
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
Diabetic Retinopathy	4.240 ± 1.219	4.357 ± 3.417	N/A	5.601 ± 0.843	8.008 ± 1.096	N/A
Contraceptive	7.838 ± 1.138	5.521 ± 3.519	6.626 ± 2.571	5.388 ± 0.788	4.994 ± 0.932	3.722 ± 0.488
Titanic	3.416 ± 0.965	6.923 ± 1.924	N/A	3.730 ± 0.866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

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

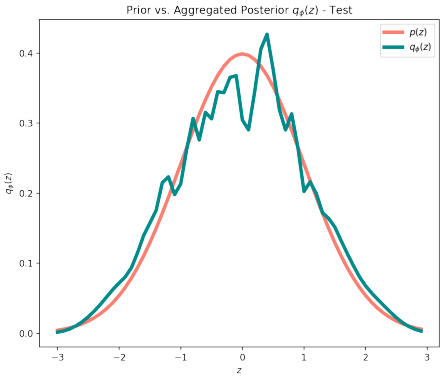
M. Qualitative examples to support necessity of both conditions of Theorem 1



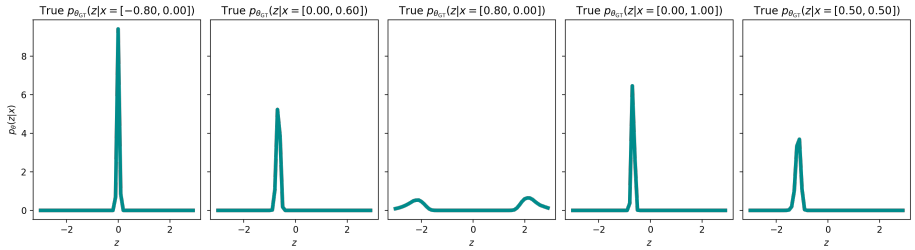
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



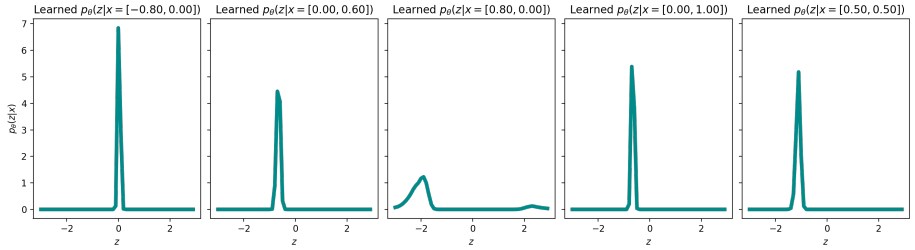
(b) True vs. learned $f_\theta(x)$



(c) Aggregated posterior vs. prior

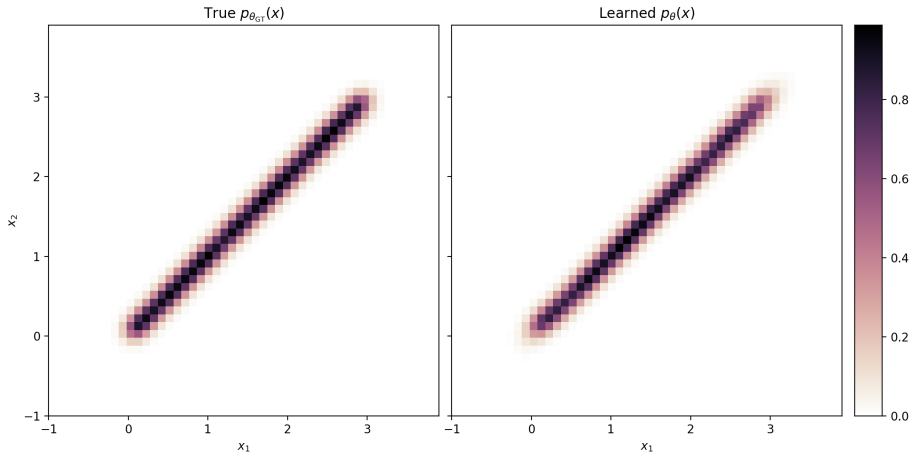


(d) Posteriors under true f_θ

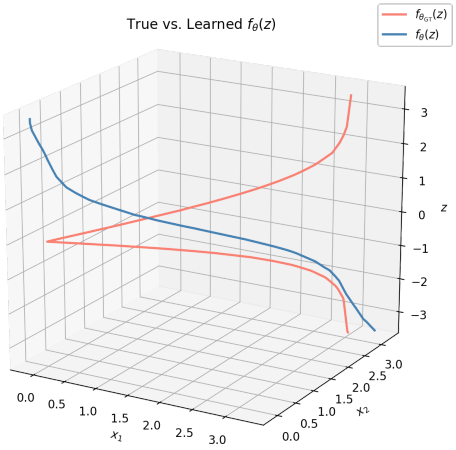


(e) Posteriors under learned f_θ

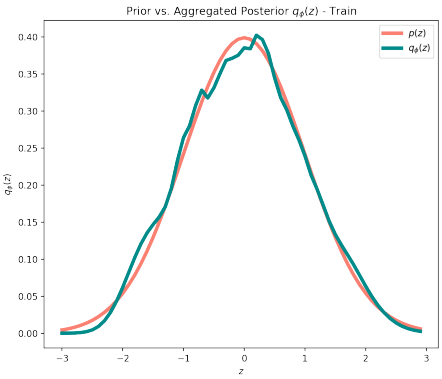
Figure 6. Vanilla VAE trained on the Circle Example. In this toy data, condition (2) holds of Theorem 1 holds and condition (1) does not. To see this, notice that most examples of the posteriors are Gaussian-like, with the exception of the posteriors near $x = [1.0, 0.0]$, which are bimodal since in 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.



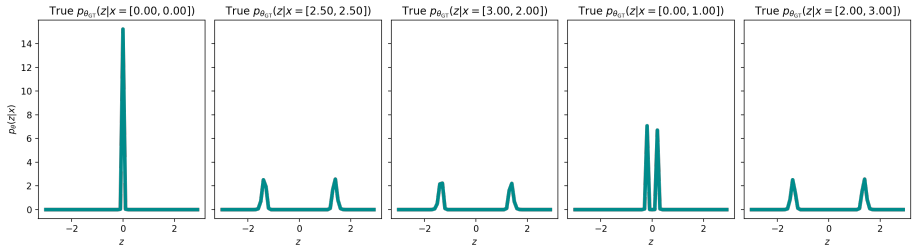
(a) True vs. learned $p_\theta(x)$



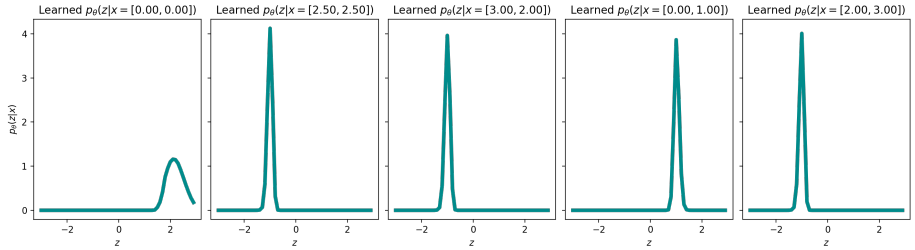
(b) True vs. learned $f_\theta(x)$



(c) Aggregated posterior vs. prior



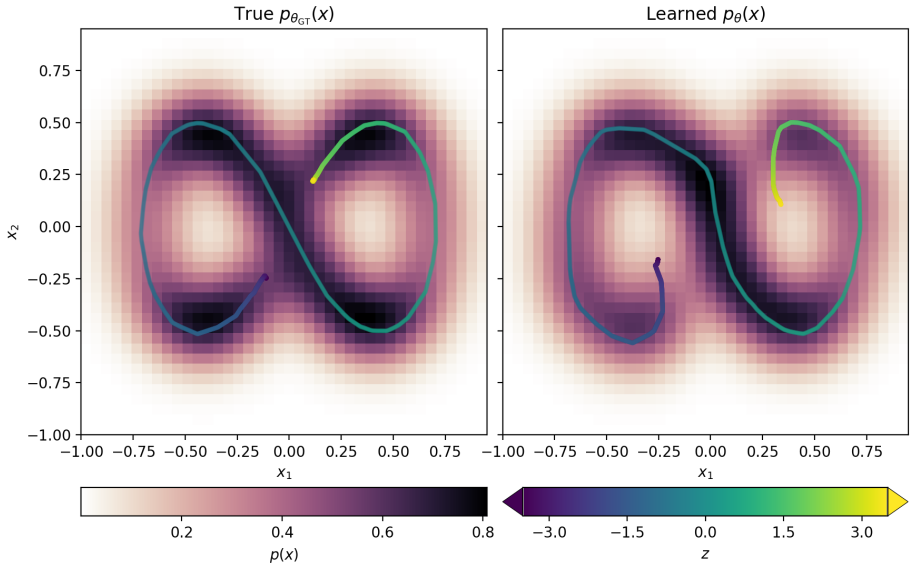
(d) Posteriors under true f_θ



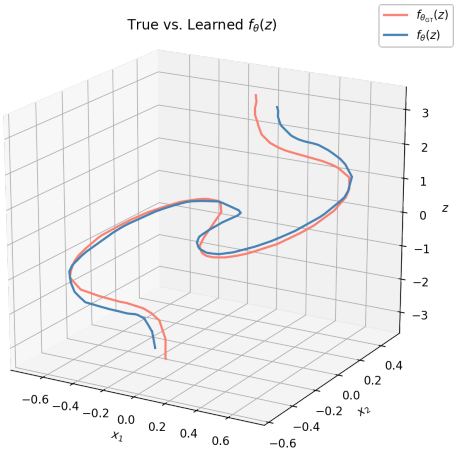
(e) Posteriors under learned f_θ

Figure 7. Vanilla 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_θ learned with a VAE is completely different than the ground-truth f_θ, and unlike the ground truth f_θ which has bimodal posteriors, the learned f_θ has unimodal posteriors (which are easier to approximate with a MFG). As such, a VAE is able to learn the data distribution well.

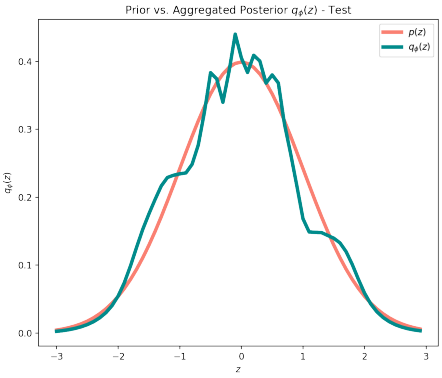
N. Qualitative Demonstration of Unsupervised VAE Pathologies



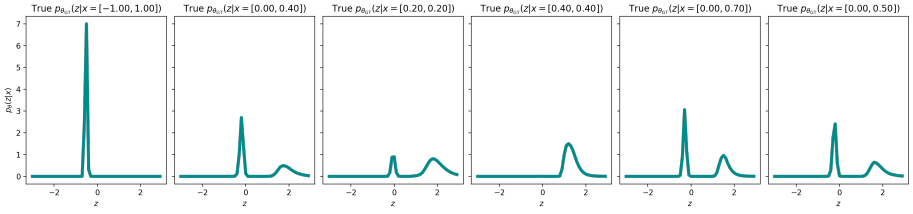
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



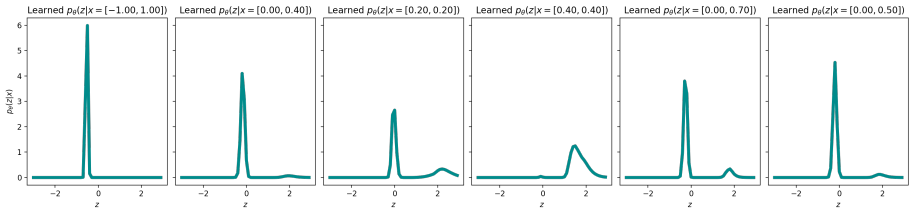
(b) True vs. learned $f_\theta(x)$



(c) Aggregated posterior vs. prior

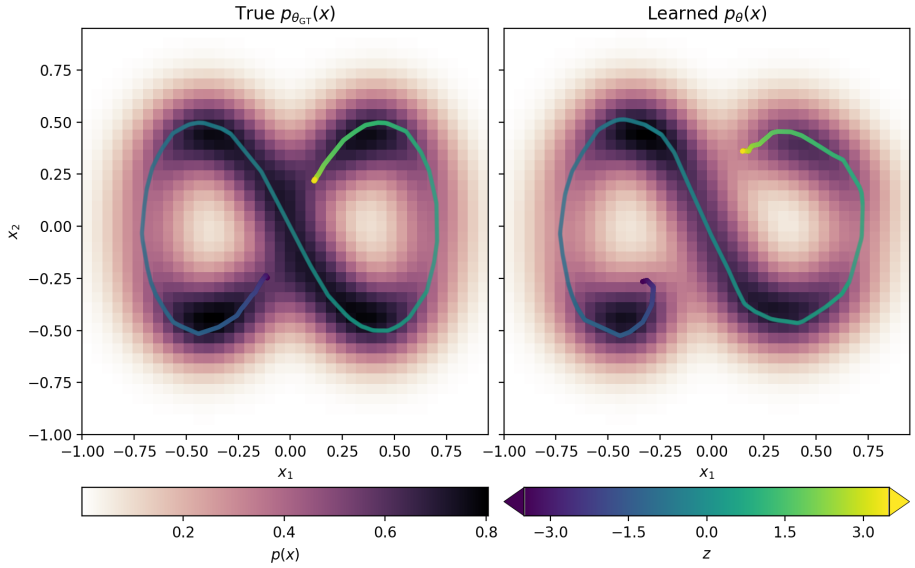


(d) Posteriors under true f_θ

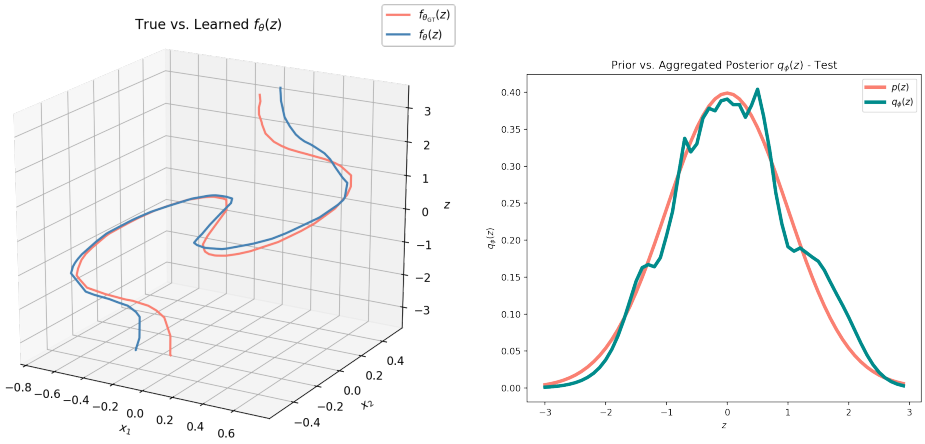


(e) Posteriors under learned f_θ

Figure 8. Vanilla 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_θ 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.

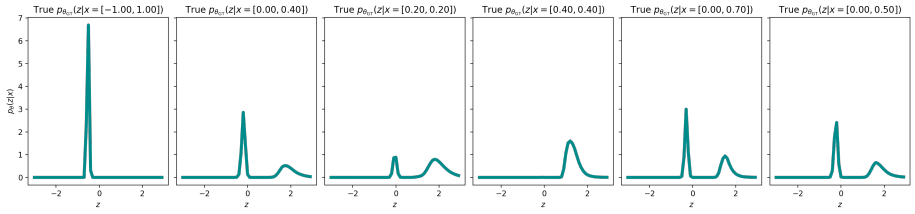


(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.

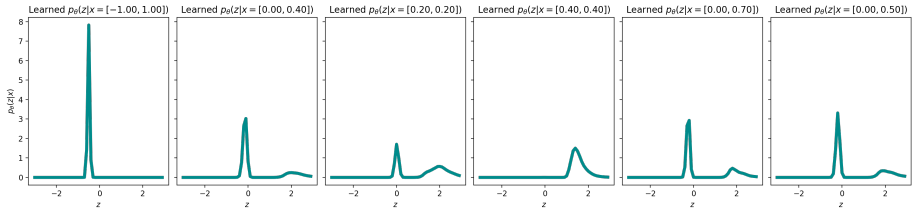


(b) True vs. learned $f_\theta(x)$

(c) Aggregated posterior vs. prior

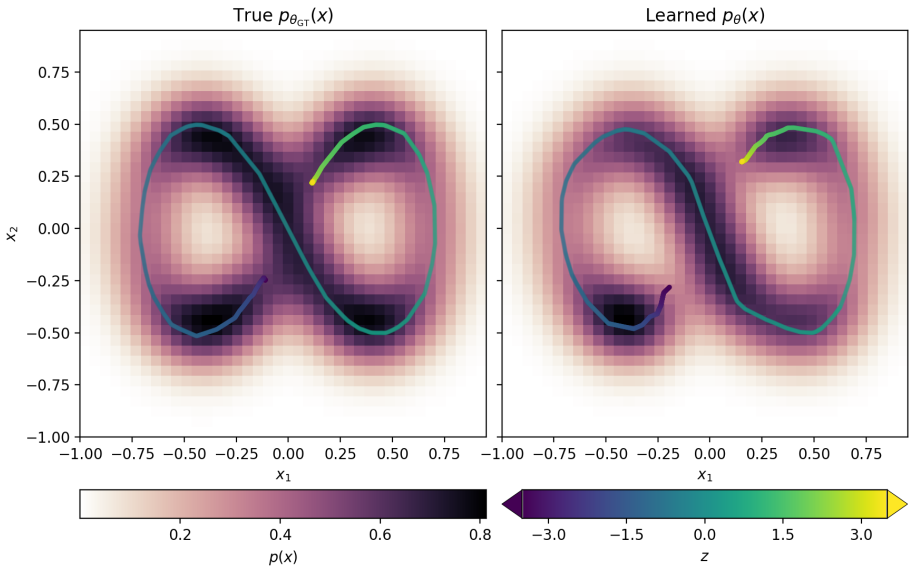


(d) Posteriors under true f_θ

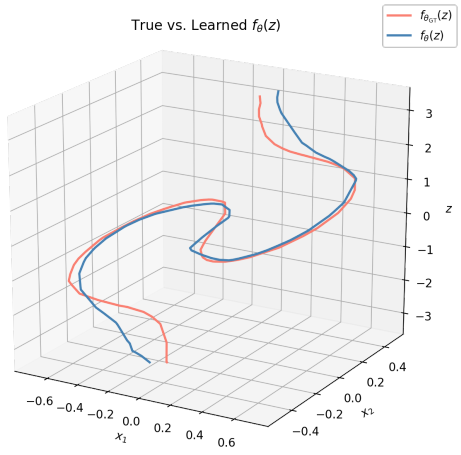


(e) Posteriors under learned f_θ

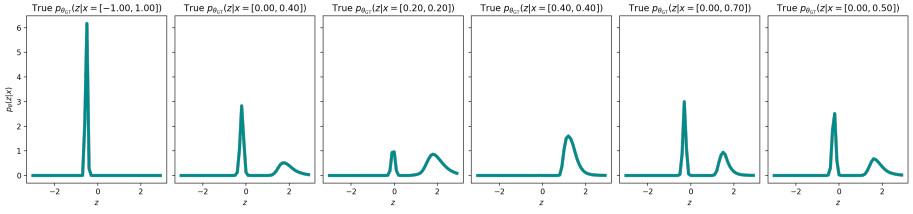
Figure 9. VAE with Lagging Inference Networks (LIN) trained on the Figure-8 Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a Vanilla VAE does (see Figure 8).



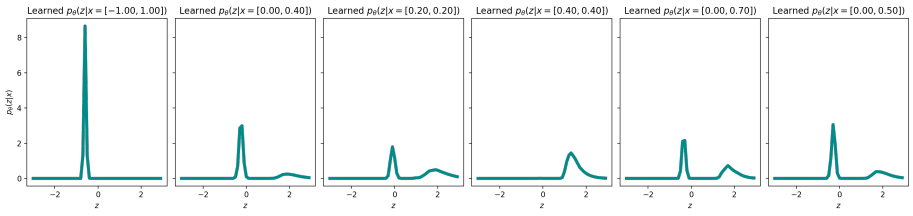
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



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

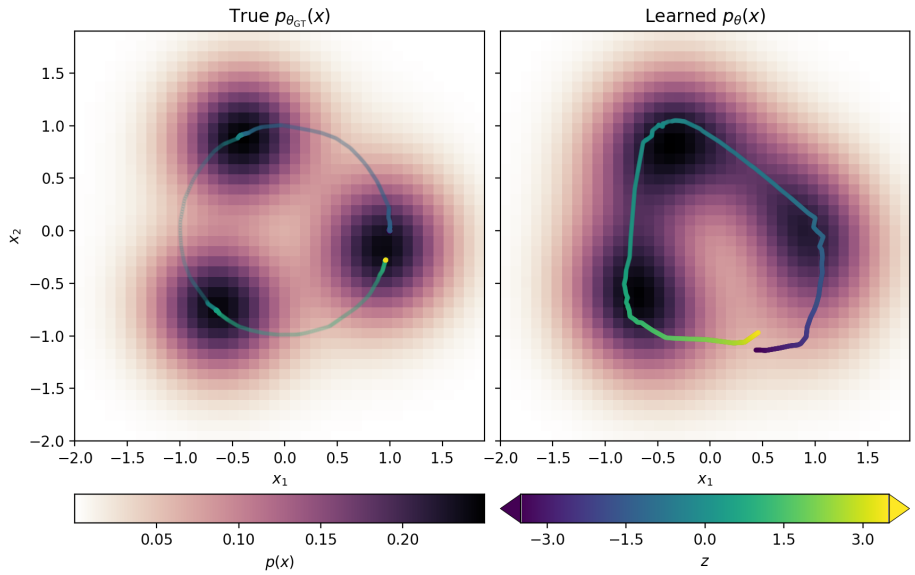


(c) Posteriors under true f_θ

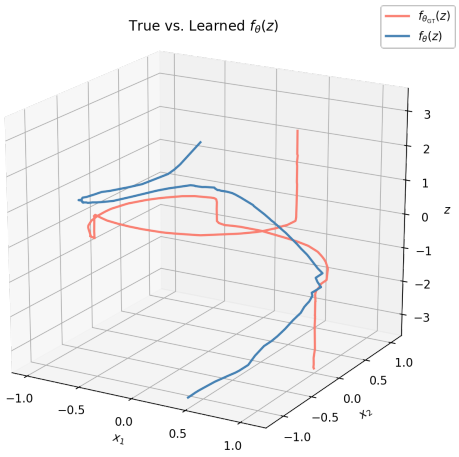


(d) Posteriors under learned f_θ

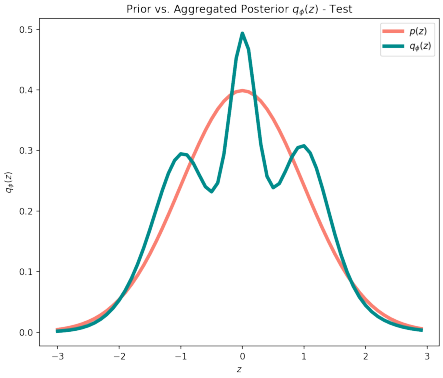
Figure 10. IWAE trained on the Figure-8 Example. In this toy data, both conditions of Theorem 1 hold. The IWAE learns a generative model with a slightly simpler posterior than that of the ground-truth. This is because even with the number of importance samples as large as $S = 20$, the variational family 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 $S \to \infty$ will resolve this issue, it is not clear how large a S is necessary and whether the additional computational overhead is worth it.



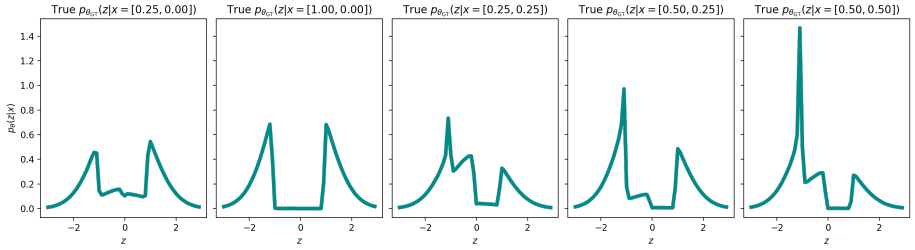
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



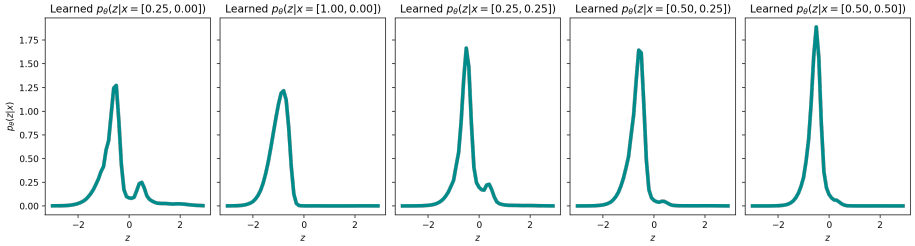
(b) True vs. learned $f_\theta(x)$



(c) Aggregated posterior vs. prior

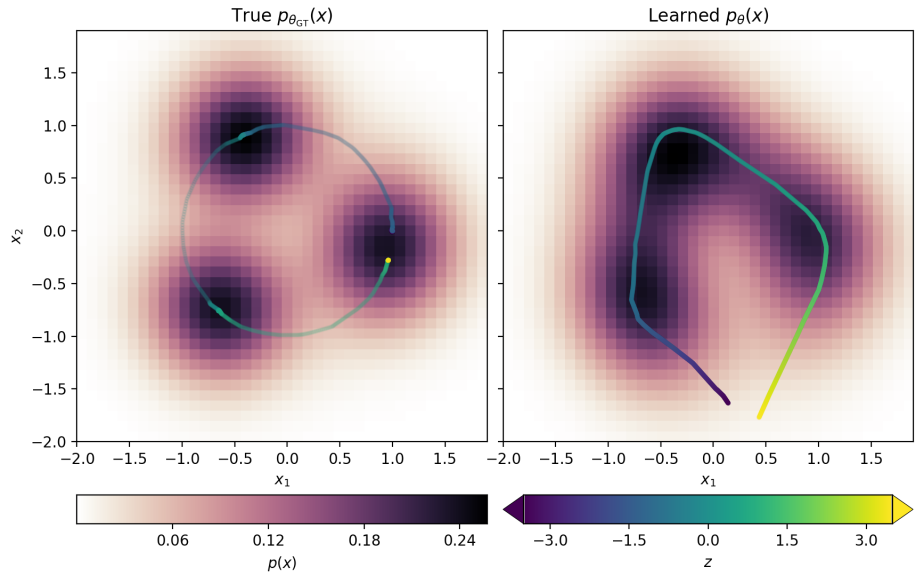


(d) Posteriors under true f_θ

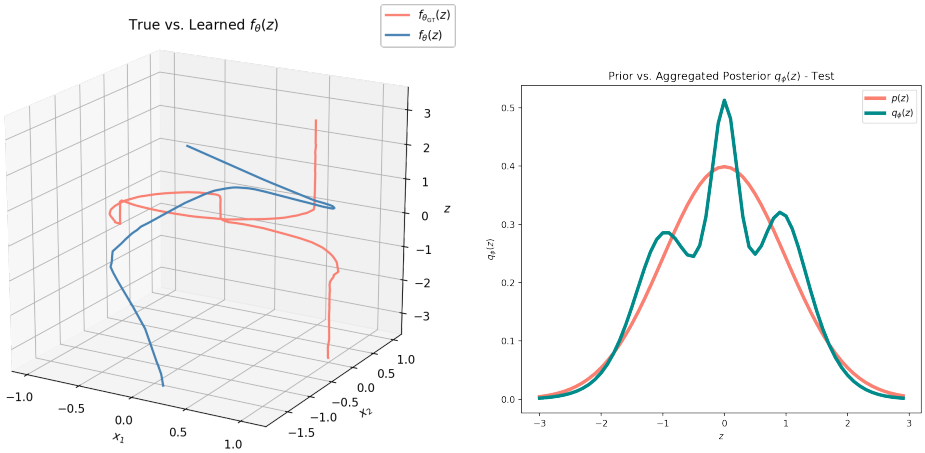


(e) Posteriors under learned f_θ

Figure 11. Vanilla 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.

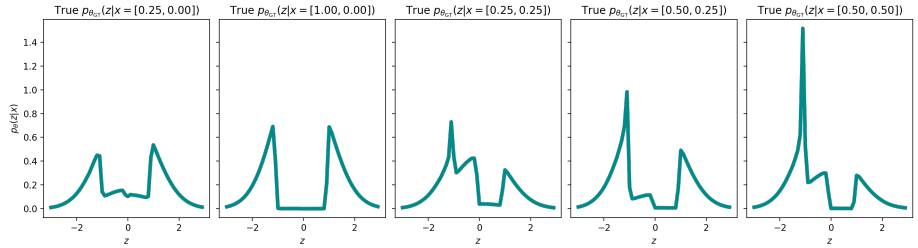


(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.

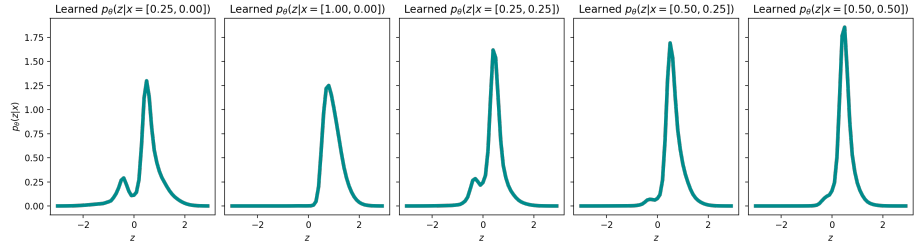


(b) True vs. learned $f_\theta(x)$

(c) Aggregated posterior vs. prior

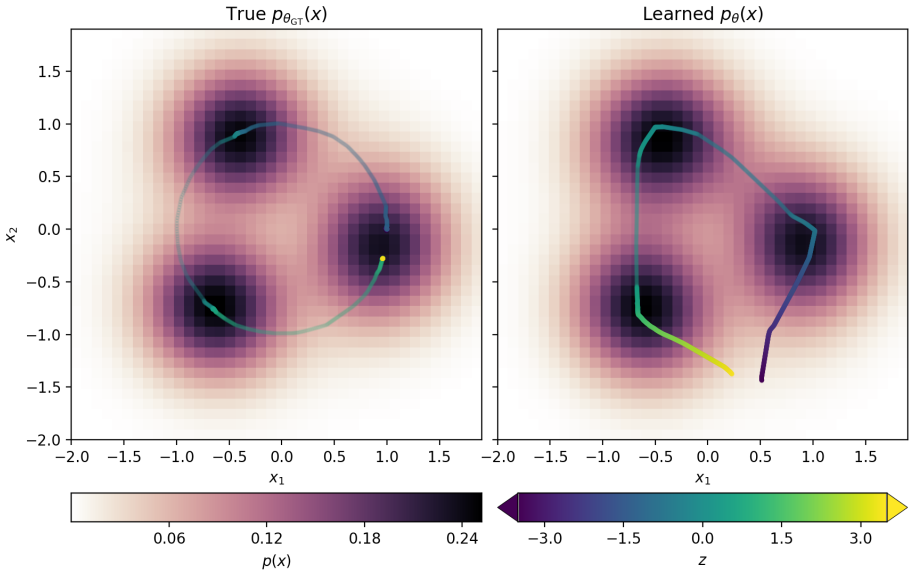


(d) Posteriors under true f_θ

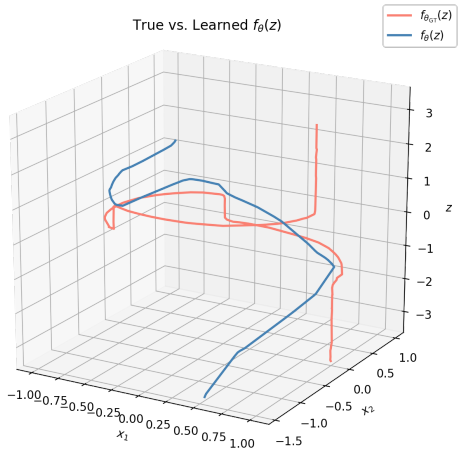


(e) Posteriors under learned f_θ

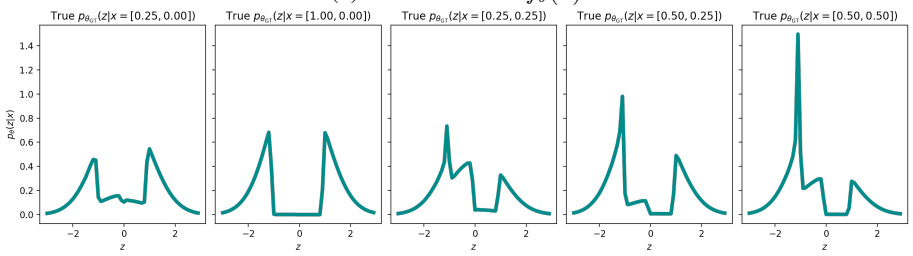
Figure 12. VAE with Lagging Inference Networks (LIN) trained on the Clusters Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribsution. As such, LIN fails in the same way a Vanilla VAE does (see Figure 11).



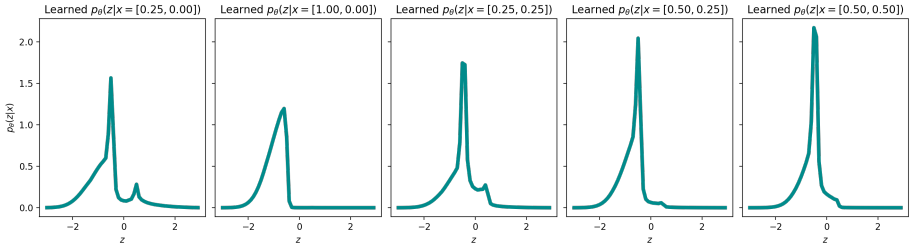
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



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

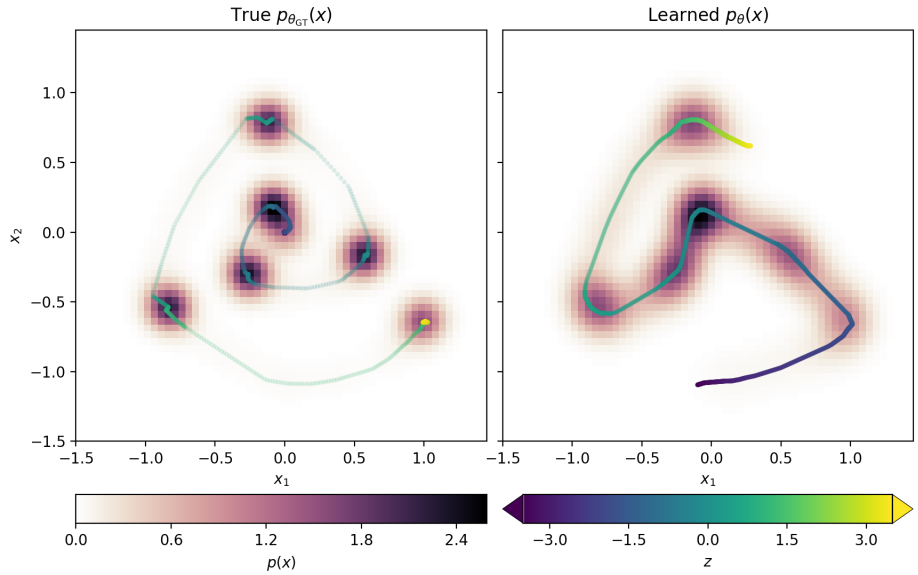


(c) Posteriors under true f_θ

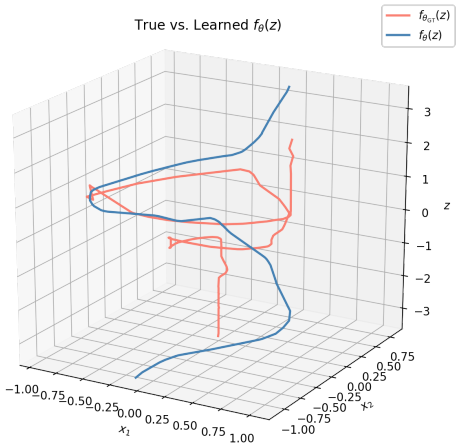


(d) Posteriors under learned f_θ

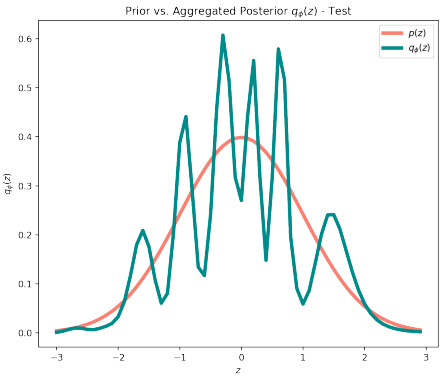
Figure 13. IWAE trained on the Clusters Example. In this toy data, both conditions of Theorem 1 hold. IWAE is able to learn the ground truth data distribution while finding a generative model with a simpler posterior than that of the ground-truth model.



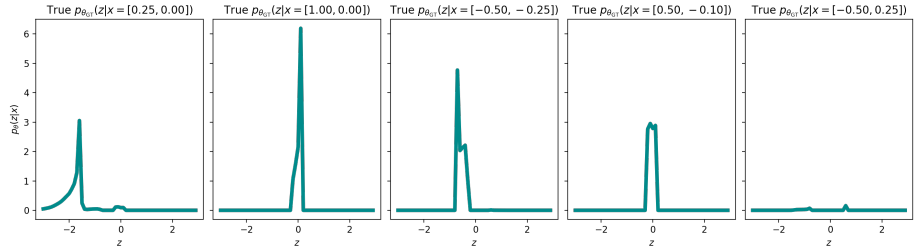
(a) True vs. learned $p_\theta(x)$, and learned vs. true $f_\theta(z)$, colored by the value of z.



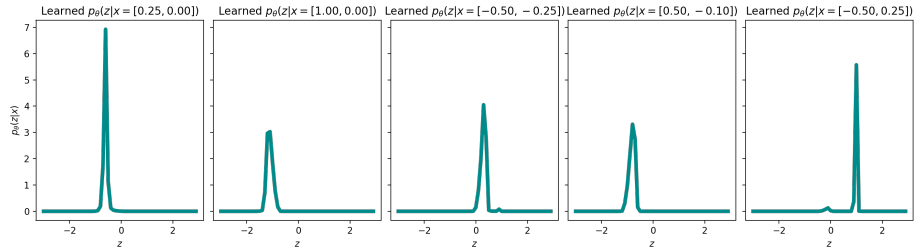
(b) True vs. learned $f_\theta(x)$



(c) Aggregated posterior vs. prior



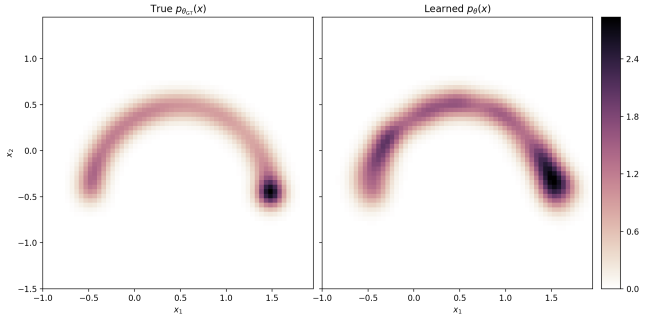
(d) Posteriors under true f_θ



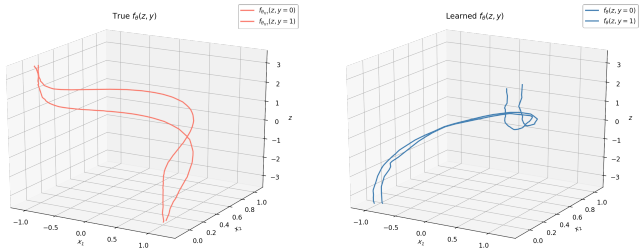
(e) Posteriors under learned f_θ

Figure 14. Vanilla VAE trained on the Spiral-Dots Example jointly over $\theta, \phi, \epsilon^2_\epsilon$. 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 σ^2_ϵ.

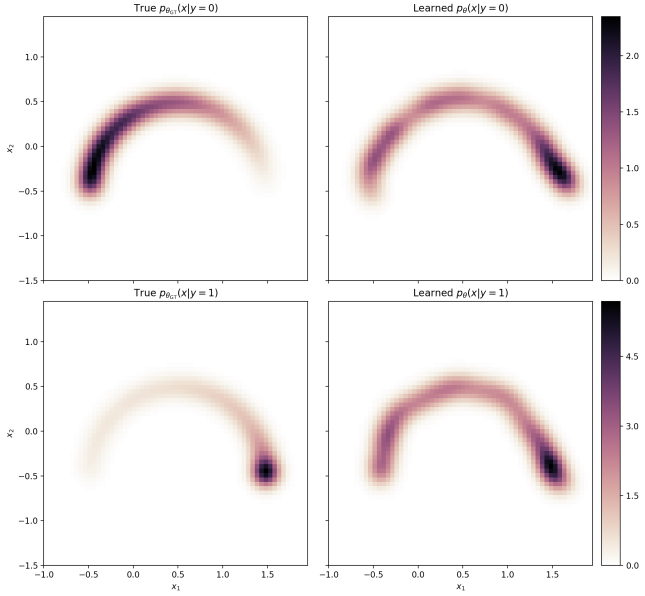
O. Qualitative Demonstration of Semi-Supervised VAE Pathologies



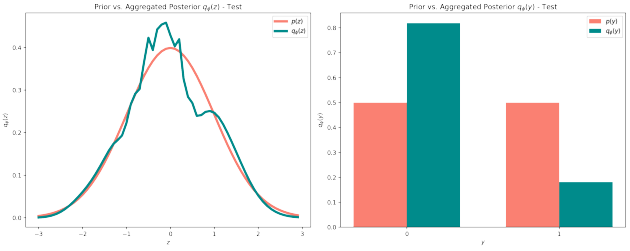
(a) True vs. learned $p_\theta(x)$.



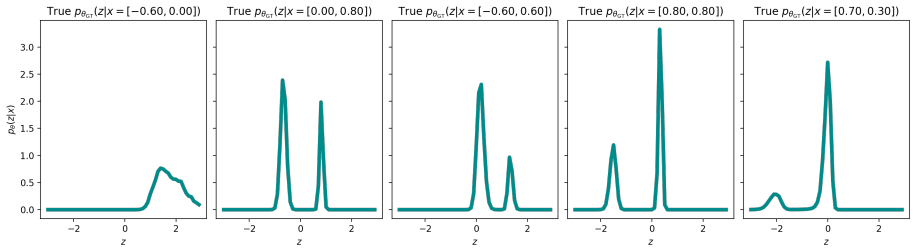
(b) True vs. learned $f_\theta(y, z)$.



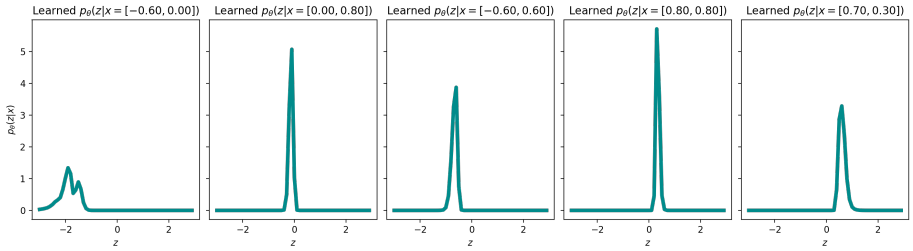
(c) True vs. learned data conditionals $p_\theta(x|y)$.



(d) Aggregated posterior vs. prior.

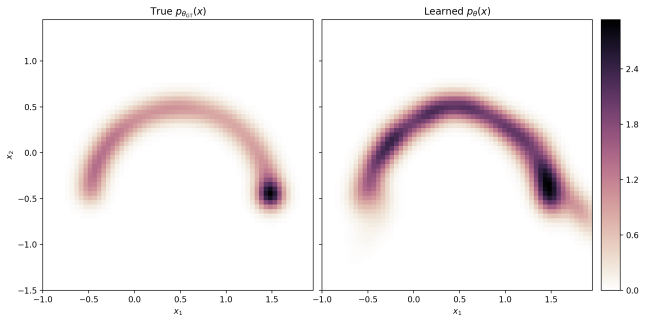


(e) Posteriors under true f_θ

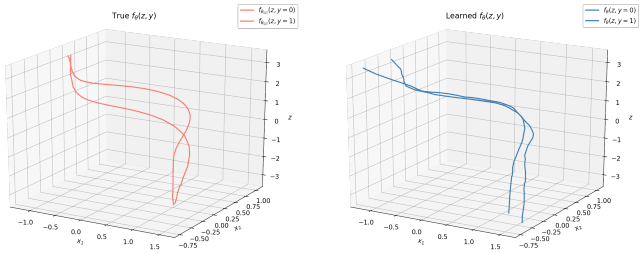


(f) Posteriors under learned f_θ

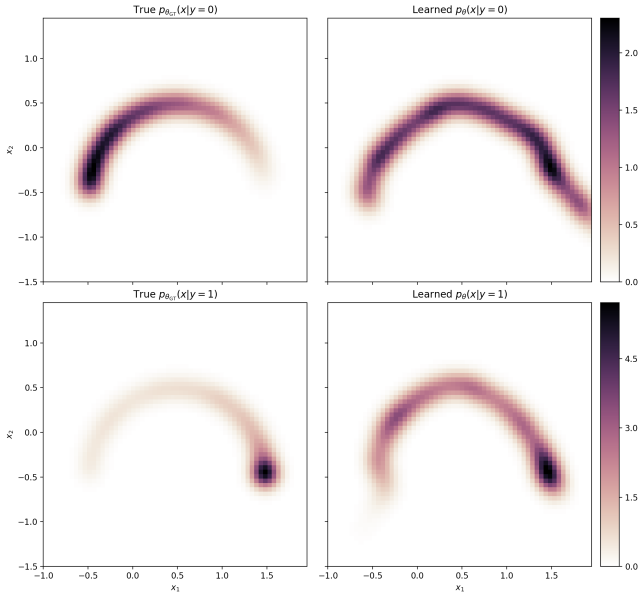
Figure 15. Vanilla Semi-Supervised VAE trained on the Discrete Semi-Circle Example. While using semi-supervision, a VAE is still able to learn the $p(x)$ relatively well. However, in this example, given x there is uncertainty as to whether it was generated from $f_\theta(y = 0, z)$ or from $f_\theta(y = 1, z)$, 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.



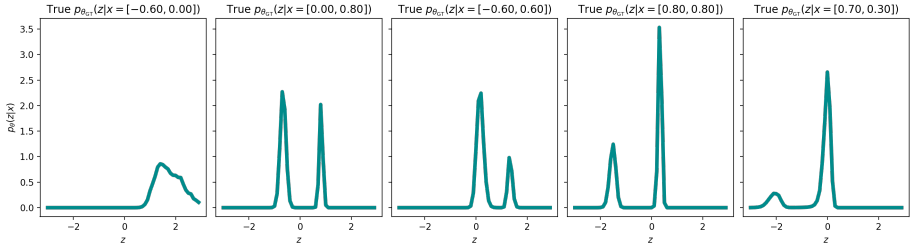
(a) True vs. learned $p_\theta(x)$.



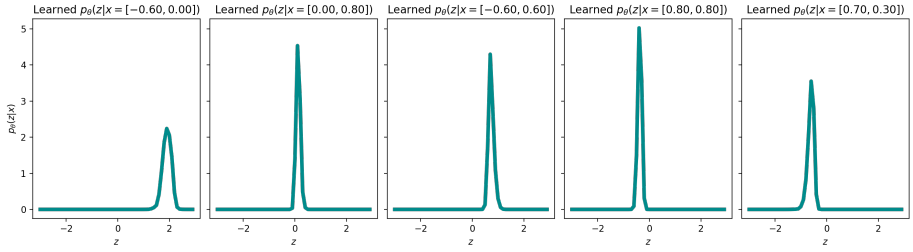
(b) True vs. learned $f_\theta(y, z)$.



(c) True vs. learned data conditionals $p_\theta(x|y)$.

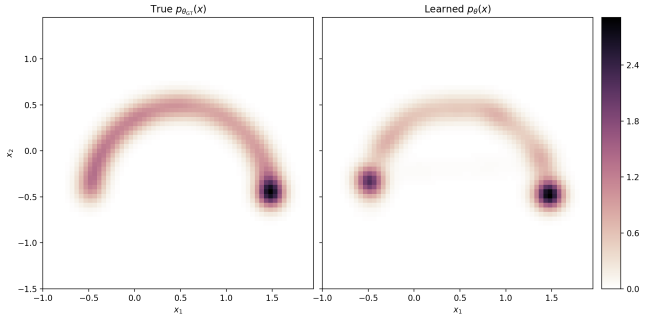


(d) Posteriors under true f_θ

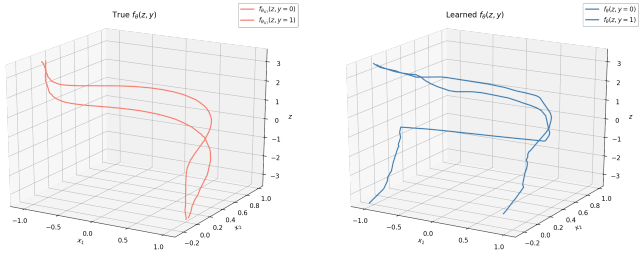


(e) Posteriors under learned f_θ

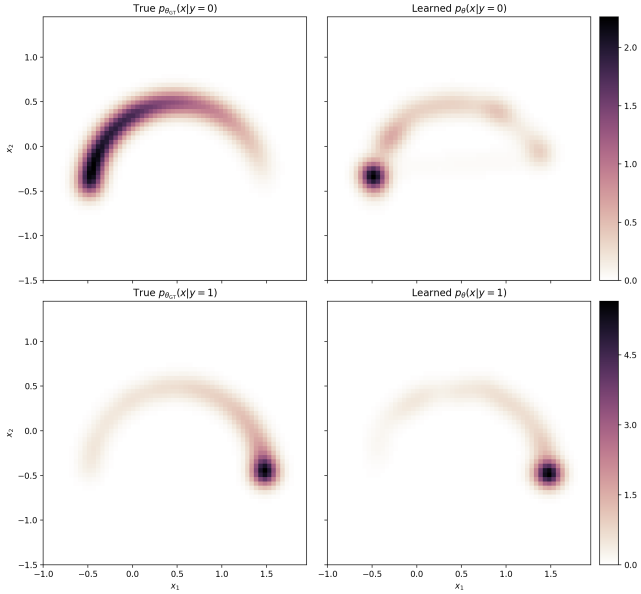
Figure 16. Semi-Supervised VAE trained with Lagging Inference Networks (LIN) trained on the Discrete Semi-Circle Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a Vanilla VAE does (see Figure 15).



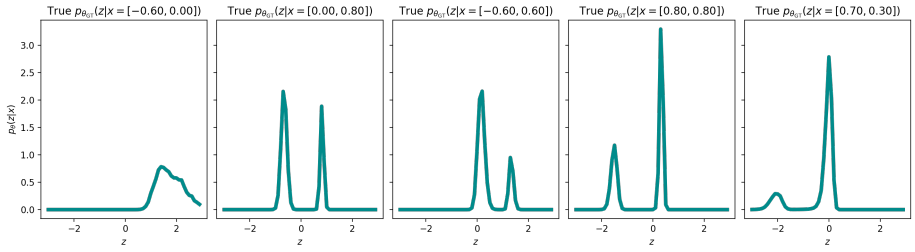
(a) True vs. learned $p_\theta(x)$.



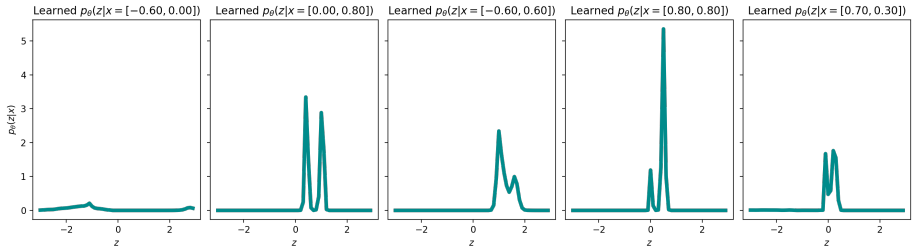
(b) True vs. learned $f_\theta(y, z)$.



(c) True vs. learned data conditionals $p_\theta(x|y)$.

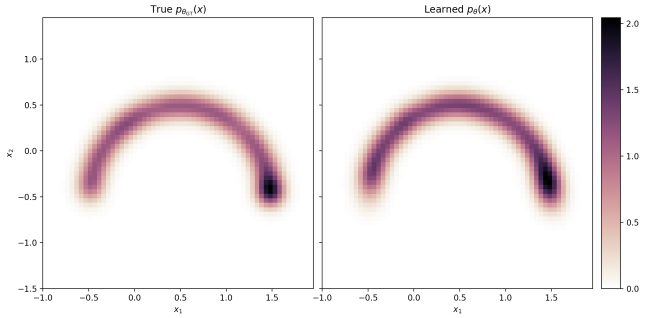


(d) Posteriors under true f_θ

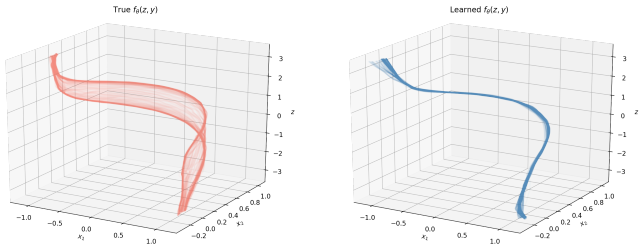


(e) Posteriors under learned f_θ

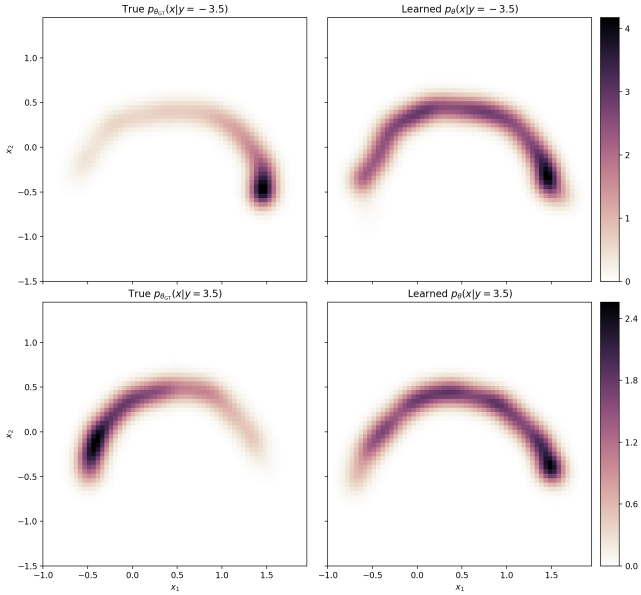
Figure 17. Semi-Supervised IWAE trained on the Discrete Semi-Circle Example. While using semi-supervision, a IWAE is still able to learn the $p(x)$ and $p(x|y)$ better than a VAE. This is because it allows for more complicated posteriors and therefore does not collapse $f_\theta(y = 0, z) = f_\theta(y = 1, z)$. However, since IWAE has a more complex variational family, the variational family no longer regularizes the function f_θ. As such, in order to put enough mass on the left-side of the semi-circle, f_θ jumps sharply from the right to the left, as opposed to preferring a simpler function such as the ground truth function.



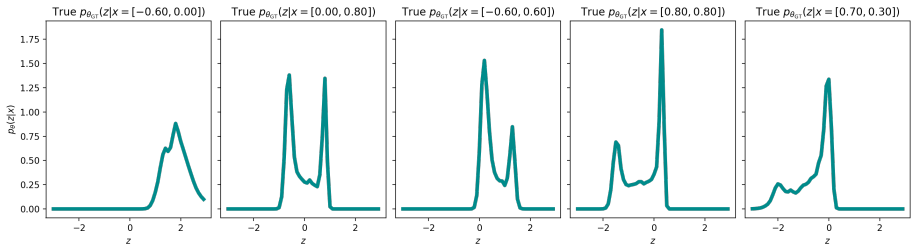
(a) True vs. learned $p_\theta(x)$.



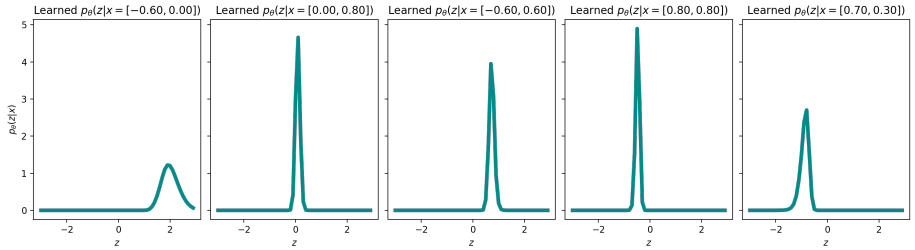
(b) True vs. learned $f_\theta(y, z)$.



(c) True vs. learned data conditionals $p_\theta(x|y)$.

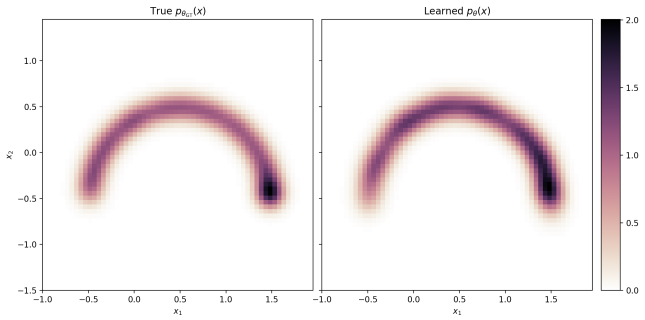


(d) Posteriors under true f_θ

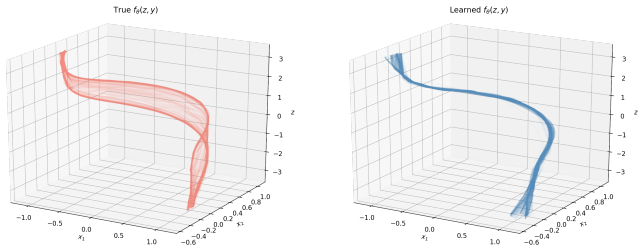


(e) Posteriors under learned f_θ

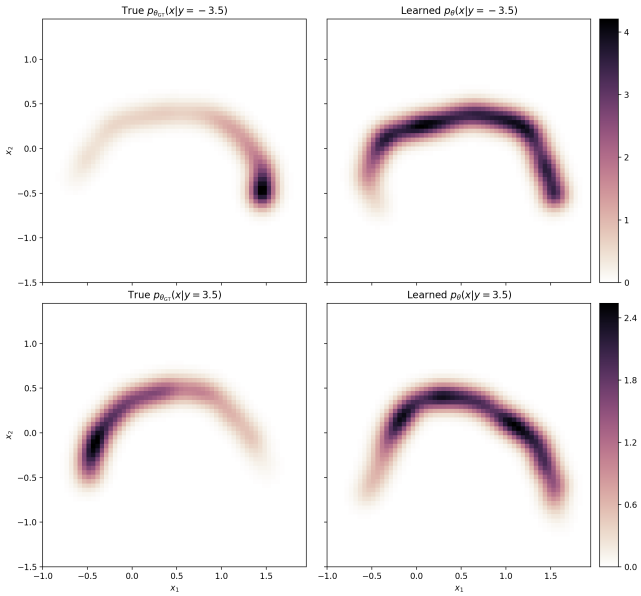
Figure 18. Vanilla Semi-Supervised VAE trained on the Continuous Semi-Circle Example. In this example, the VAE exhibits the same problems as in the Discrete Semi-Circle Example (Figure 18). However, with since y is continuous, this poses an additional issue. Since $q_\phi(y|x)$ (the discriminator) in the objective is a Gaussian, and the ground truth $p_\theta(y|x)$ is multi-modal, the objective will select a function f_θ 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.



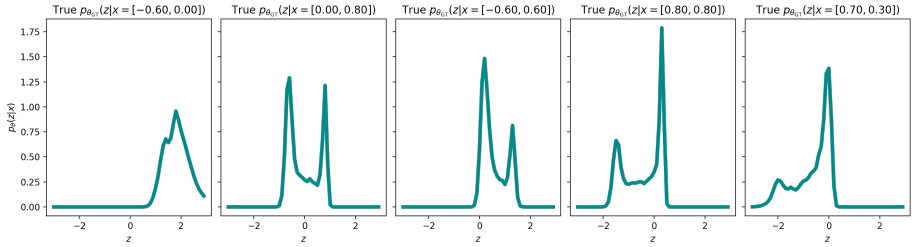
(a) True vs. learned $p_\theta(x)$.



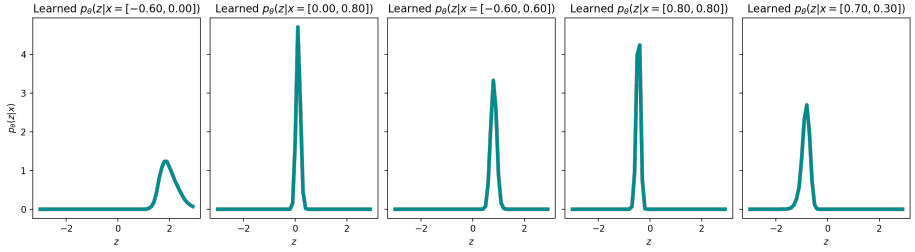
(b) True vs. learned $f_\theta(y, z)$.



(c) True vs. learned data conditionals $p_\theta(x|y)$.

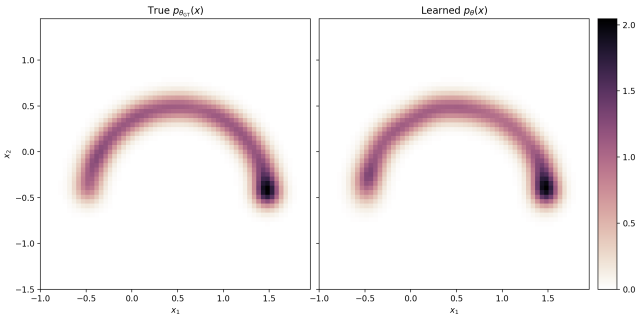


(d) Posteriors under true f_θ

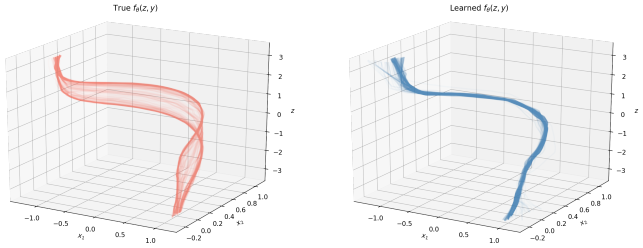


(e) Posteriors under learned f_θ

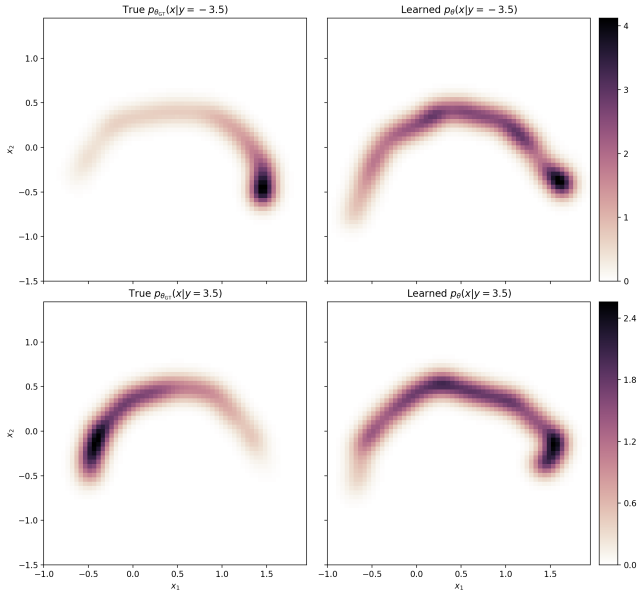
Figure 19. Semi-Supervised VAE trained with Lagging Inference Networks (LIN) trained on the Continuous Semi-Circle Example. While LIN may help escape local optima, on this data, the training objective is still biased away from learning the true data distribution. As such, LIN fails in the same way a Vanilla VAE does (see Figure 18).



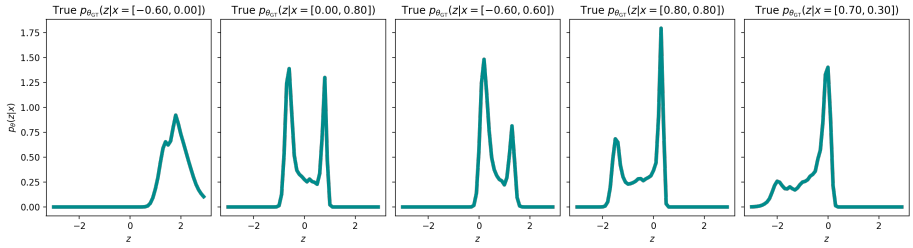
(a) True vs. learned $p_\theta(x)$.



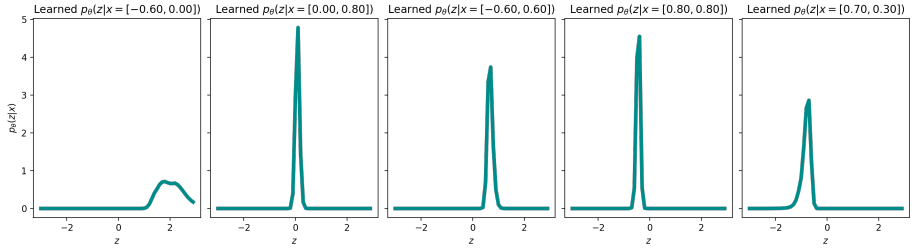
(b) True vs. learned $f_\theta(y, z)$.



(c) True vs. learned data conditionals $p_\theta(x|y)$.



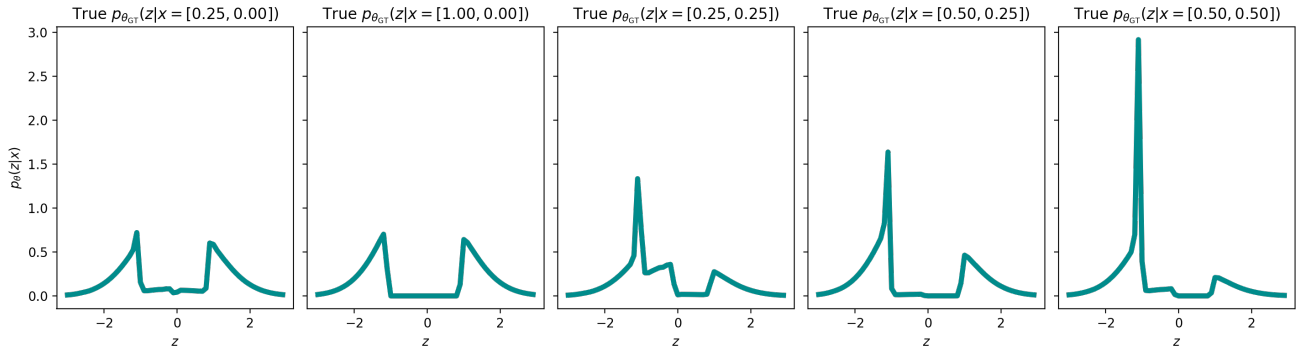
(d) Posteriors under true f_θ



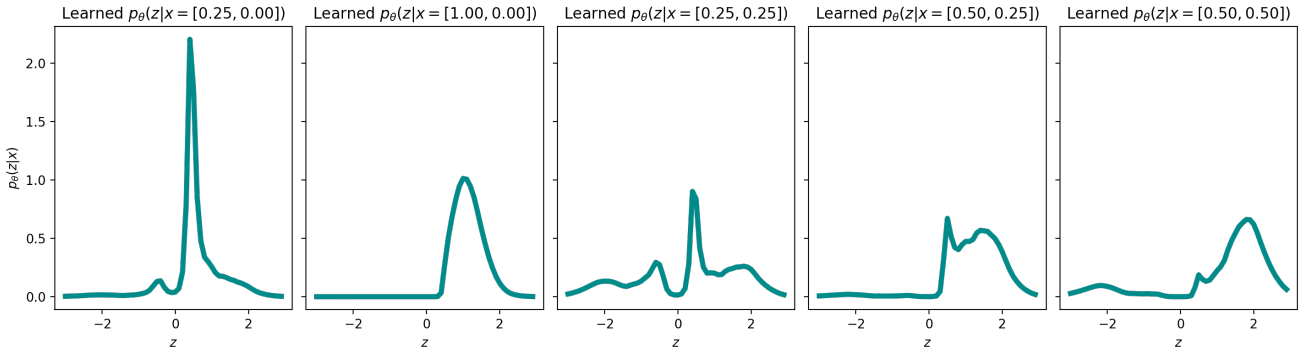
(e) Posteriors under learned f_θ

Figure 20. 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_θ 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.

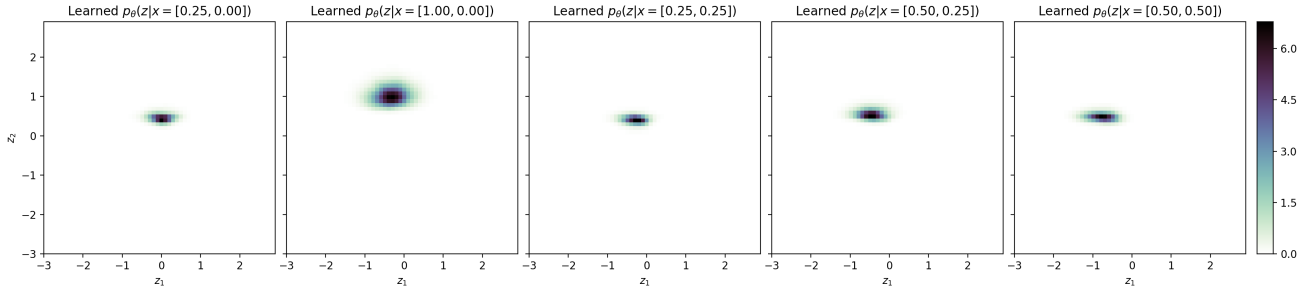
P. Posterior Is Simper Under Model Mismatch



(a) True Posterior $K = 1$

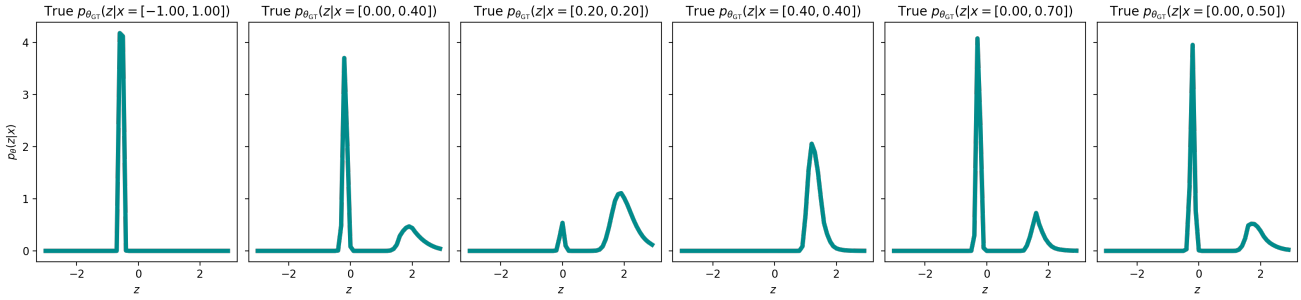


(b) Learned Posterior $K = 1$

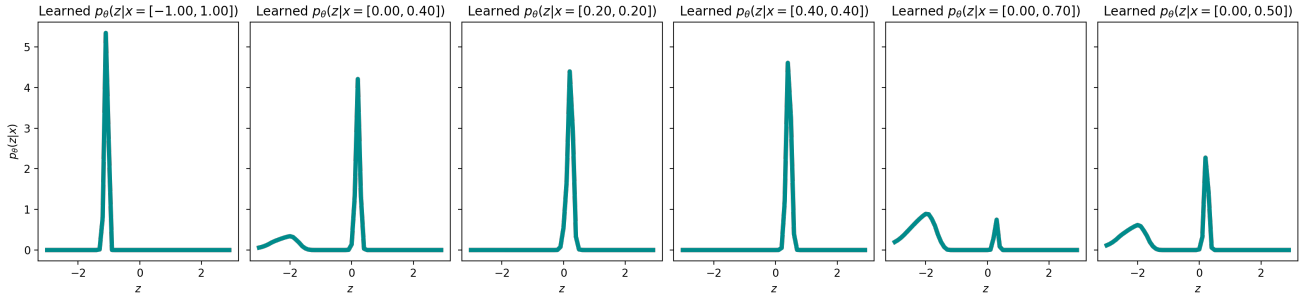


(c) Learned Posterior $K = 2$

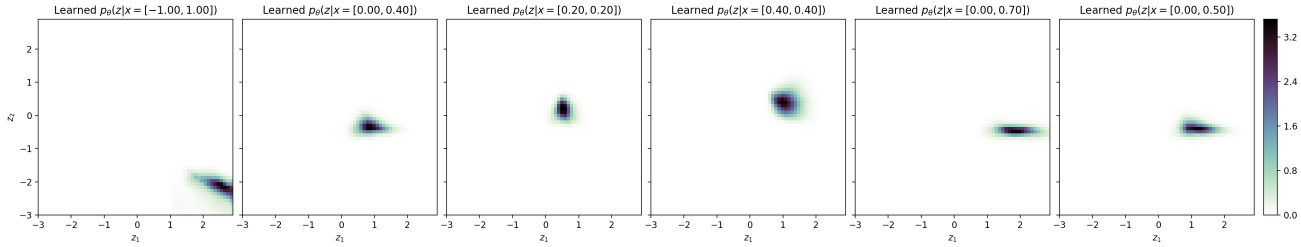
Figure 21. VAEs learn simpler posteriors as latent dimensionality K increases and as the observation noise σ_ϵ^2 decreases on "Clusters Example" (projected into 5D space).



(a) True Posterior $K = 1$



(b) Learned Posterior $K = 1$



(c) Learned Posterior $K = 2$

Figure 22. VAEs learn simpler posteriors as latent dimensionality K increases and as the observation noise σ_ϵ^2 decreases on "Figure-8 Example" (projected into 5D space).