MeanFieldGaussian-VaritionalAutoEncoder

failurs in downstream tasks at global ELBO optimum

Failure Modes of Variational Autoencoders and Their Effects on Downstream Tasks

Yaniv Yacoby YANIVYACOBY@G.HARVARD.EDU
Weiwei Pan WEIWEIPAN@G.HARVARD.EDU
Finale Doshi-Velez FINALE@SEAS.HARVARD.EDU
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA

https://arxiv.org/pdf/2007.07124.pdf

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 VAE 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, such as learning compressed and disentangled representations, adversarial robustness, and semi-supervised learning.

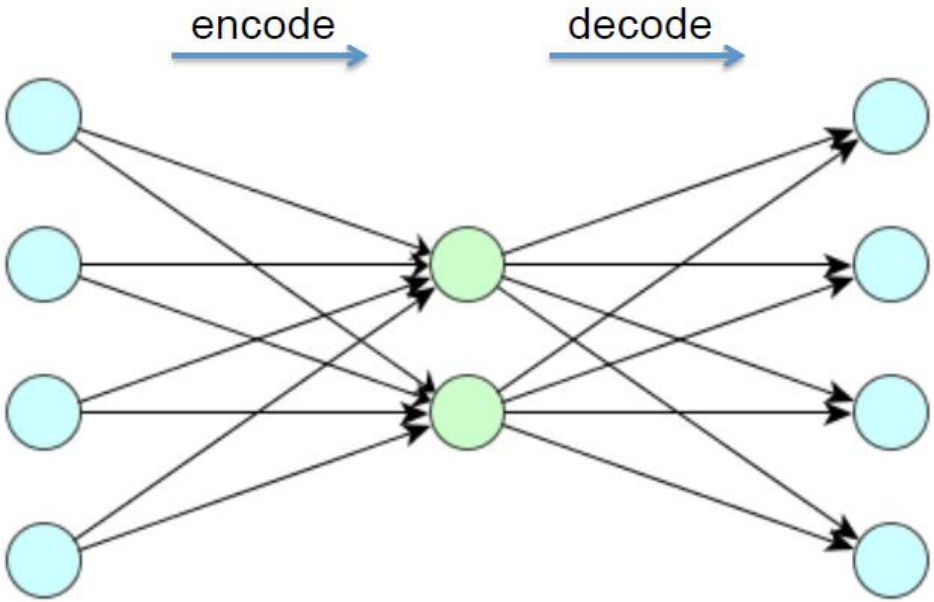
Keywords: Variational Autoencoder, Variational Inference, Approximate Inference, Latent Variable Models

Beate Sick

References

- Presented paper: https://arxiv.org/pdf/2007.07124.pdf
- ML talk of first author: https://www.youtube.com/watch?v=xUIuPA2QnDs
- 2014 VAE basics Kingma et al. https://arxiv.org/abs/1406.5298
- Blog explaining Kingma et al. https://bjlkeng.github.io/posts/semi-supervised-learning-with-variational-autoencoders/

Recall: Classical Auto Encoder (AE) without «variational»



Input	Hidden	Output
«observed»	«latent»	«observed»
x'	$z(x')$	$x(z(x'))$

x': observed data

z: latent variable

x: generated reconstruction of x'

https://tensorchiefs.github.io/bbs/files/vae.pdf
https://tensorchiefs.github.io/bbs/files/pca-ae-hashing-27012016.pdf

Paper focuses on MFG-VAE

VAE: Variational Autoencoder

- Conditioned on input x a distribution $p(z|x)$ is learned for latent z

MFG: Mean-Field Gaussian assumption:

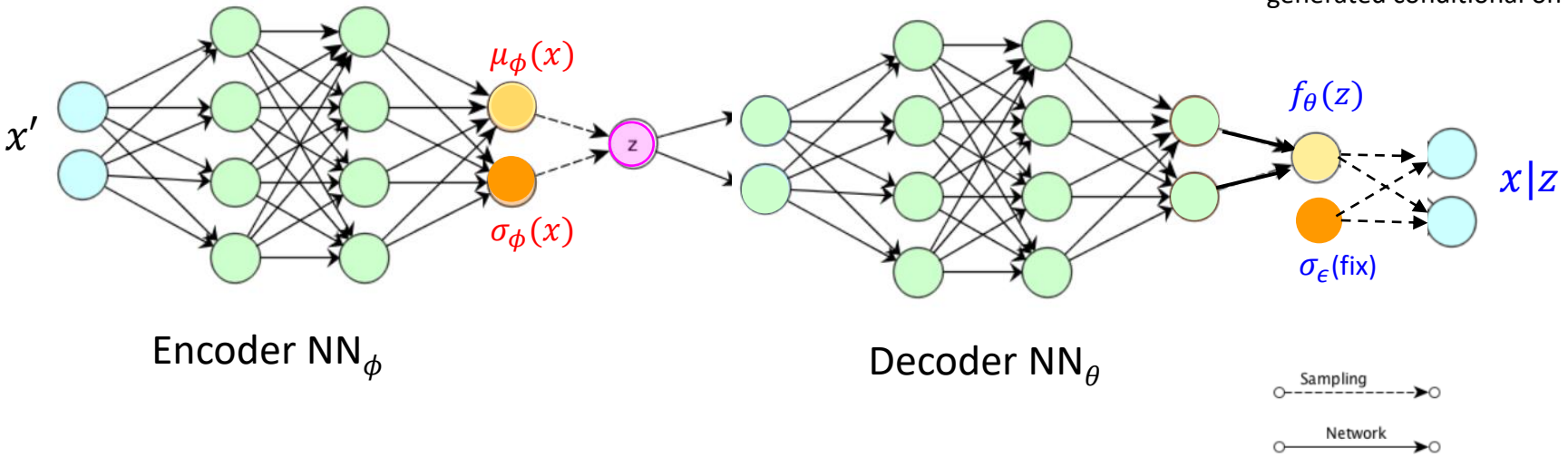
- Latent space $\mathrm{p}(z|x)$ is modeled with independent Gaussians
- Data distribution $\mathrm{p}(x|z)$ is modeled with independent Gaussians

MeanFieldGaussian-VariationalAutoEncoder: MFG-VAE

x': observed data (here 2D)

z: latent space (here 1D)

$x|z$: outcome data (here 2D) generated conditional on z



In a trained MFG-VAE

- the conditional latent Variable z with prior $z \sim N(0,1)$ follow a posterior:

  $z|x \sim p_\phi(z|x) \approx q_\phi(z|x) \overset{MFG}{=} N(\mu_\phi(x), \sigma_\phi^2(x) \cdot I)$

- the generated conditional outcome data $x_g\,|z$ follow also a Gaussian:

  $x|z \sim p_\theta(x|z) \overset{MFG}{\approx} N(f_\theta(x), \sigma_\epsilon^2 \cdot I)$

Note that for the modeled outcome $x|z$ only the mean-generator $f_\theta(x)$ for the Gaussian depends on z

Loss in unsupervised VAE: negative VAE ELBO

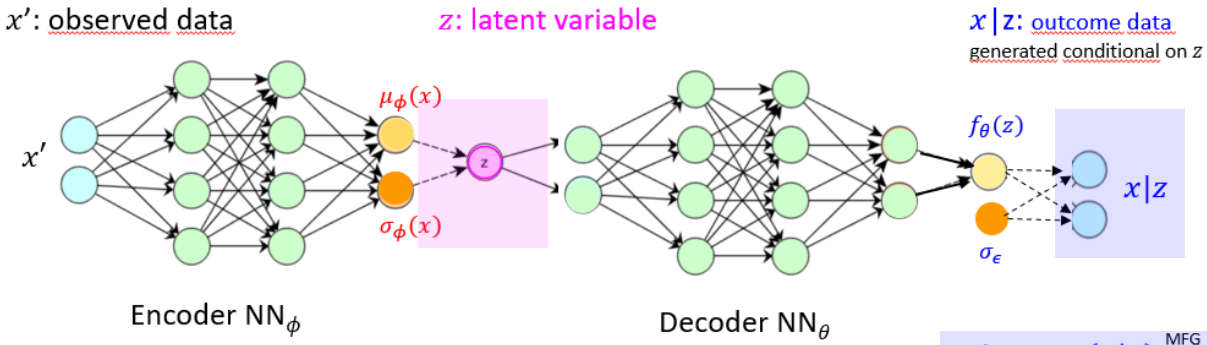
That is, we can write $\text{argmin}_{\theta,\phi} - \text{ELBO}(\theta, \phi)$ as follows (Zhao et al., 2017):

$$\text{argmin}_{\theta,\phi}(\underbrace{D_{\text{KL}}[p(x)||p_\theta(x)]}_{\text{MLEO}} + \underbrace{\mathbb{E}_{p(x)}\left[D_{\text{KL}}[q_\phi(z|x)||p_\theta(z|x)]\right]}_{\text{PMO}}) \quad (3)$$

Maximum
Likelihood
Estimation
Objective

Posterior
Matching
Objectiv

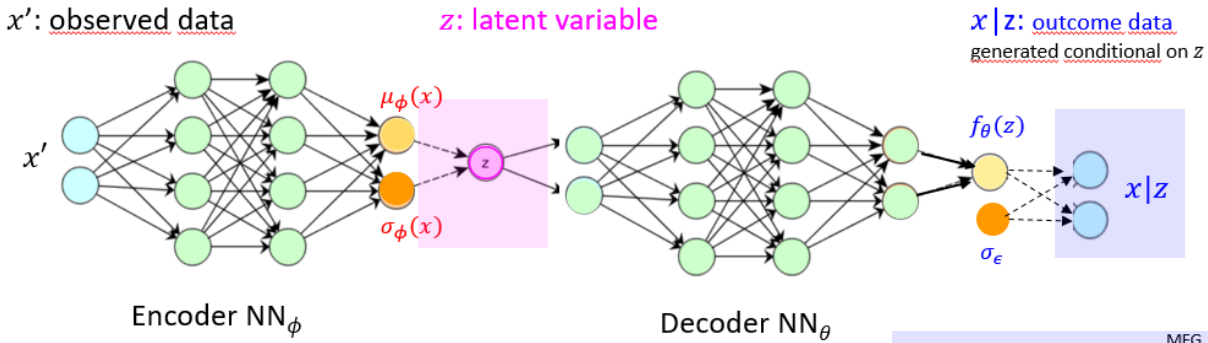
→ Often the global optimum of ELBO compromise between MLEO and PMO



$$z|x \sim p(z|x) \approx q_\phi(z|x) = N(\mu_\phi(x), \sigma_\phi^2(x) \cdot I)$$

$$x|z \sim p_\theta(x|z) \overset{\text{MFG}}{\approx} N(f_\theta(x), \sigma_\epsilon^2 \cdot I)$$

Learning objectives and downstream tasks



Learn a good representation of: latent space $p(z|x)$ data $p(x), p(x|z)$

Downstream tasks:

- Data compression
- Disentangled representation → GT data generating process

- $p(x)$ estimation
- data generation $p(x|z)$
- Outlier detection
- Signal/Noise separation → adversarial robustness

Main research questions of the paper

- Under which conditions fails the MFG-VAE solution at the global optimum of the ELBO-loss
  - to estimate the ground-truth data distribution $p(x), p(x|z)$?
  - to learn the ground-truth latent representation $p(z|x)$?
  - to decompose correctly between signal and noise?

  → which benchmark data sets trigger these failures?

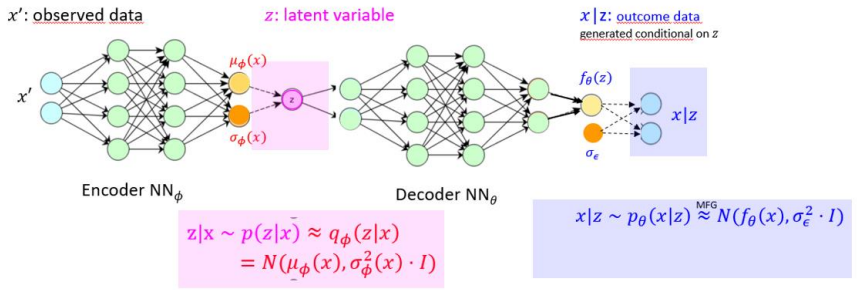
- Under which conditions would a more flexible variational posterior resolve the observed problems?

- How much are downstream tasks affected?

Side step: posterior collapse

Posterior collapse means that the latent variable is uniformat
ive and z is ignored → only the overall data distribution $p(x)$ is learned.

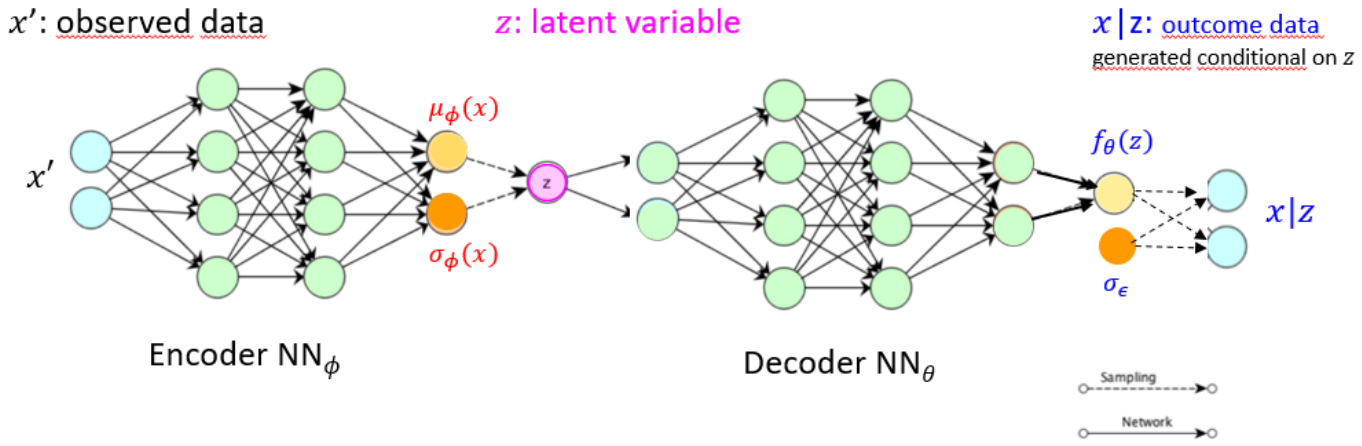
- Usual reasons for posterior collapse in VAE:
  - Signal of z is too weak, i.e. $\mu_\Phi(x_i) \approx \mu_\Phi(x_j), \sigma_\Phi(x_i) \approx \sigma_\Phi(x_j)\ \forall\ i, j$
  - Signal of z is too noisy, i.e. $\sigma_\Phi(x_j)$ is too large for most j
- Posterior collapse only occur if $p(z) = p_\theta\ (z|x) = q_\Phi(z|x)$ [He et al. 2019]



Posterior collapse in MFG-VAE globally optimal solutions only occur when $p(x)$ is Gaussian
→ not a relevant case and therefore ignored in the paper

Simulated data are used to investigat VAE failures

- Simulate Data → ground truth (GT) is known
  - GT dimension K of latent space z is known
  - GT variance σ_ϵ^2 of the observed outcome distribution for x is known



Experimental Setup

- Make sure that models, trained on simulated data, reaches global ELBO optimum
  1) Fix the hyperparameters K (=z-dim) and σ_ϵ^2 (=data-noise) at GT
  2) Train 10 MFG-VAE models with
     - Use flexible enough decoder NN for approximating p(x|z)
     - Use flexible enough encoder NN for approximating p(z|x)
     - Initialize 5 times the decoder at GT and best encoder NN
     - Initialize 5 times randomly
     - Pick trained model with the best ELBO

- Check if MFG restriction causes problems by comparing with IWAE that allows for more flexible variational posteriors.

IWAE: Importance Weighted Autoencoder arXiv:1509.00519

Failure on data distribution

Conditions when MFG-VAE compromise learning data distribution

The data distribution $p(x)$ is «seriously» misestimated by the global ELBO optimal MGF-VAE solution if both of the following 2 conditions hold:

Cx1:

The true posterior $p(z|x)$ for the GT $f_{\theta\,GT}$ [with $x|z \sim N(f_{\theta\,GT}(z), \sigma_\epsilon^2 I)$] is difficult to approximate by an MFG for a large proportion of x.

Cx2:

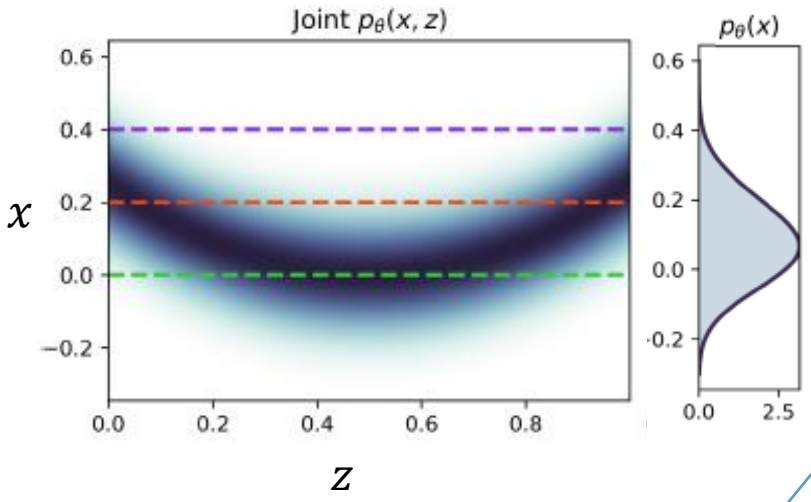
There does not exist a (likelihood function) f_θ with [a corresponding] simpler [more MFG like] posterior that approximates $p(x)$ well.

Proof: see appendix in paper.

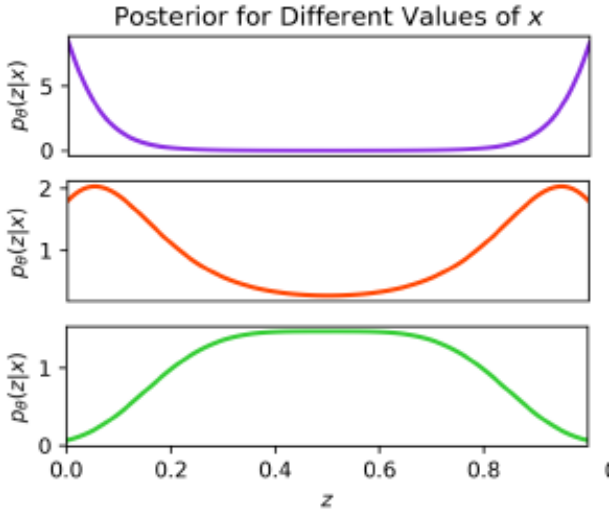
Example 1: Only condition 1 is not sufficient for a failure on $p(x)$

GT: data generating process for simulating data:

$$z \sim U(0,1), \quad (x|z) \sim N(f_{\theta_{GT}} = (z - 0.5)^2, \sigma_\epsilon^2)$$



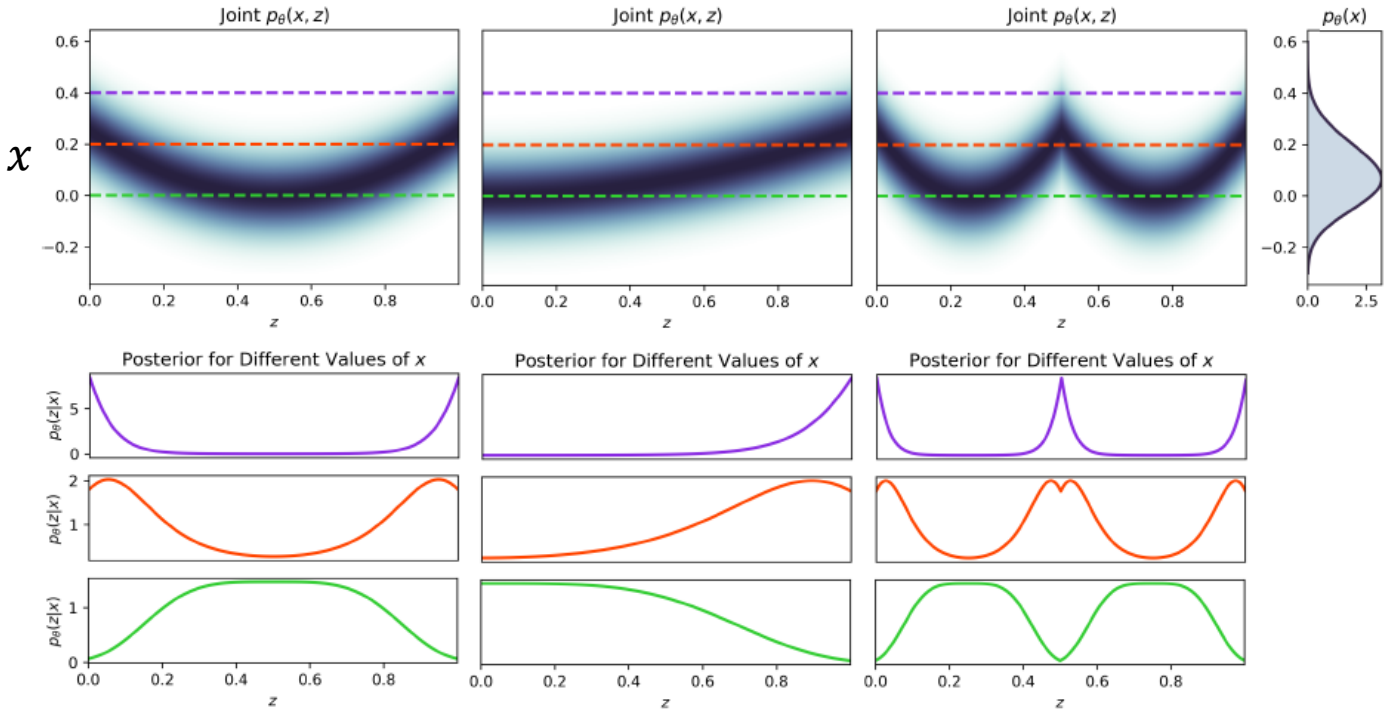
GT posterior $p(z|x)$ examples



For many x-values the posterior is not similar to a Gaussian

Cx1 is fulfilled, since the GT posterior $p(z|x)$ for the GT $f_{\theta_{GT}}$ $[x|z \sim N(f_{\theta_{GT}}(z), \sigma_\epsilon^2 I)]$ is difficult to approximate by an MFG for a large proportion of x.

Example 1: Only condition 1 is not sufficient for a failure on $p(x)$



(a) **Variant 1**

$f_\theta(z) = (z - 0.5)^2$

GT

(b) **Variant 2**

$f_\theta(z) = 0.25 \cdot z^2$

counterexample

(c) **Variant 3**

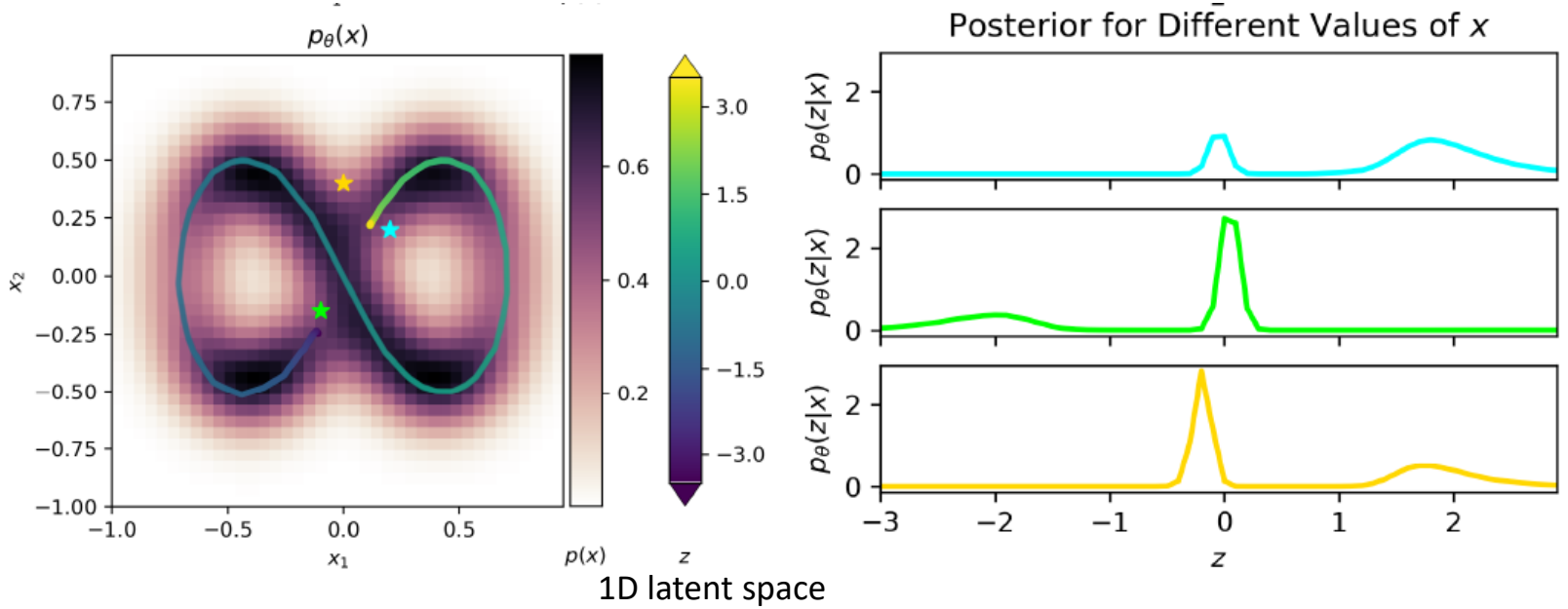
$f_\theta(z) = \mathbb{I}(z < 0.5) \cdot (2z - 0.5)^2 + \mathbb{I}(z \geq 0.5) \cdot (2z - 1.5)^2$

Cx2 is **not fulfilled**, since it is not true that there does not exist a (likelihood function) f_θ with [a corrosponding] simpler [more MFG like] posterior that approximates $p(x)$ well.

→ Thanks to ”non-identiability” a non-MFG GT posterior does not necessarily prevent MFG-VAE to model the data distribution $p(x)$ well.

8-shape example where MFG-VAE fails to model 2D $p(x)$

GT:



The largest density of the GT $\mathrm{p(x|z)}$ is on an 8-shaped manifold

→ In the GT 8-shape manifold we have a cross point, where many x exist that could have been generated from different z

→ **Cx1 fulfilled:** many x with multi-modal GT posterior $p(z|x)$

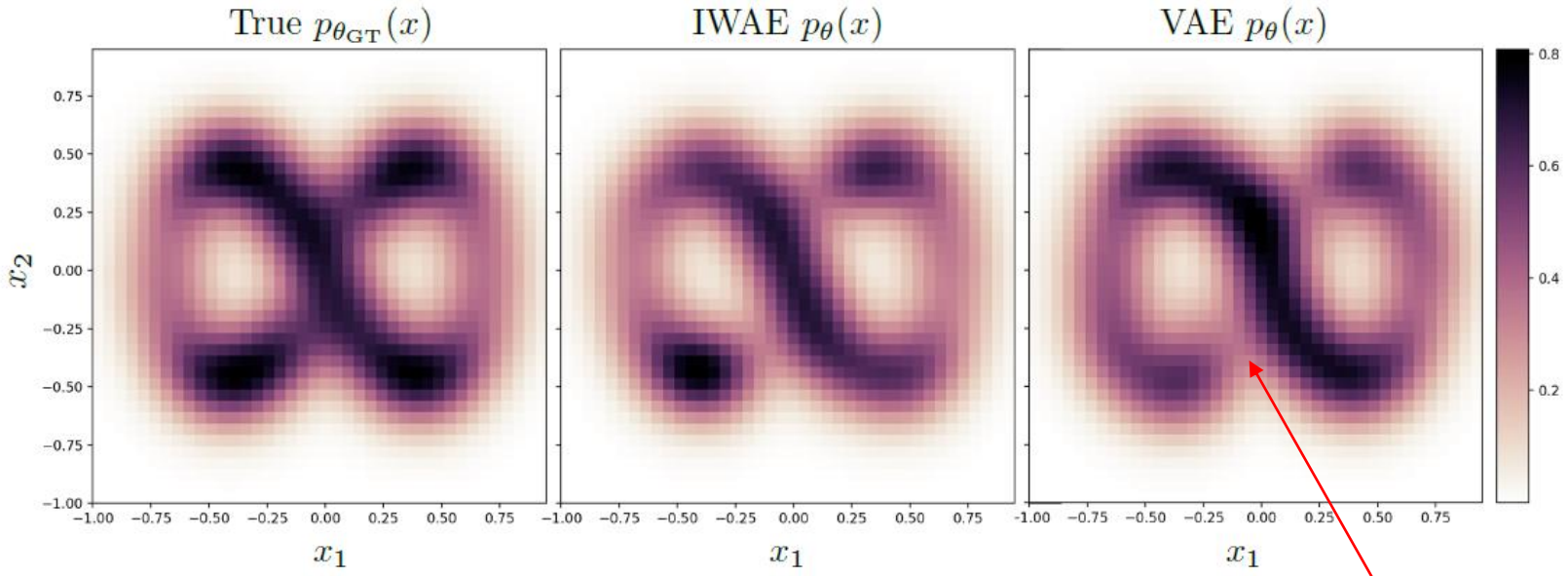
In MFG-VAE $\mathrm{p(x|z)} \approx N(f_\theta(z), \sigma_\epsilon^2 I)$ the highest density at $x = f_\theta(\mathrm{z})$

→ If 8-shaped GT $\mathrm{p(x)}$ is well estimated, then $f_\theta(z)$ would be 8-shaped

→ Such an 8-shaped $f_\theta(z)$ would lead to non-MFG posteriors in area of the cross point

→ **Cx2 fulfilled:** $\nexists\ f_\theta$ with MFG like posteriors that estimates data distribution well

8-shape example where MFG-VAE fails to model 2D $p(x)$

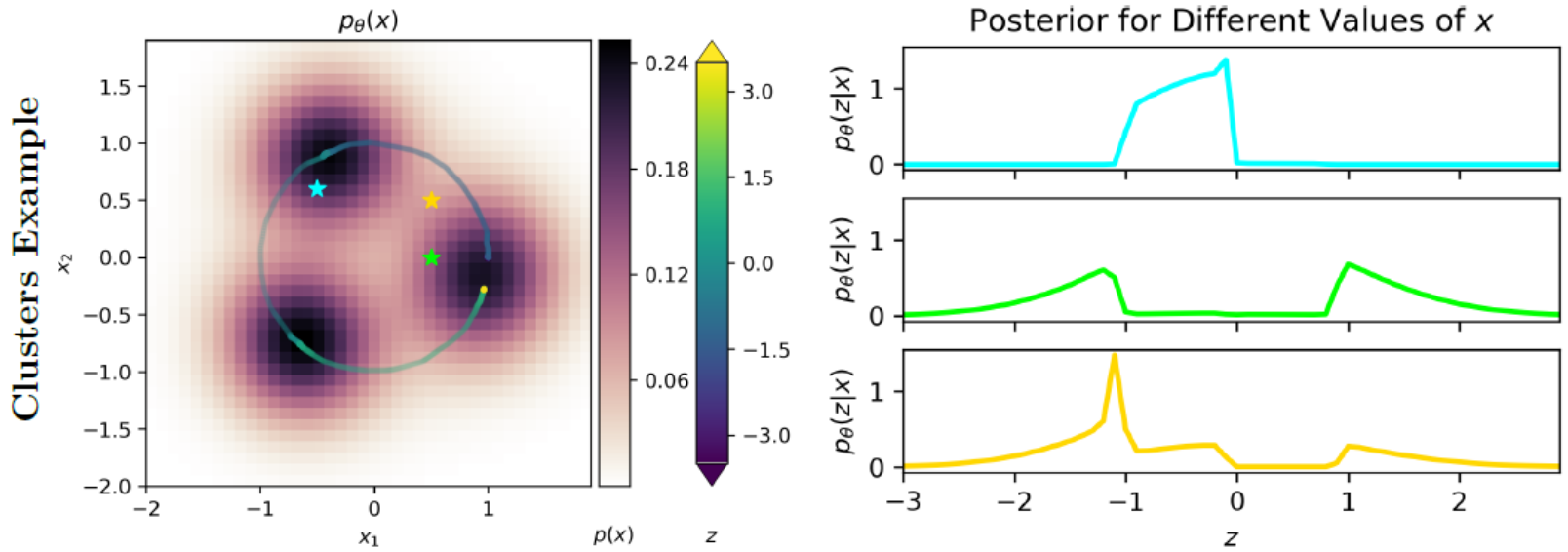


$f_\theta(z)$ of MFG-VAE does not resemble "crossing"

→ As expected MFG-VAE fails to model the data distribution

Cluster example where MFG-VAE fails to model 2D $p(x)$

GT:



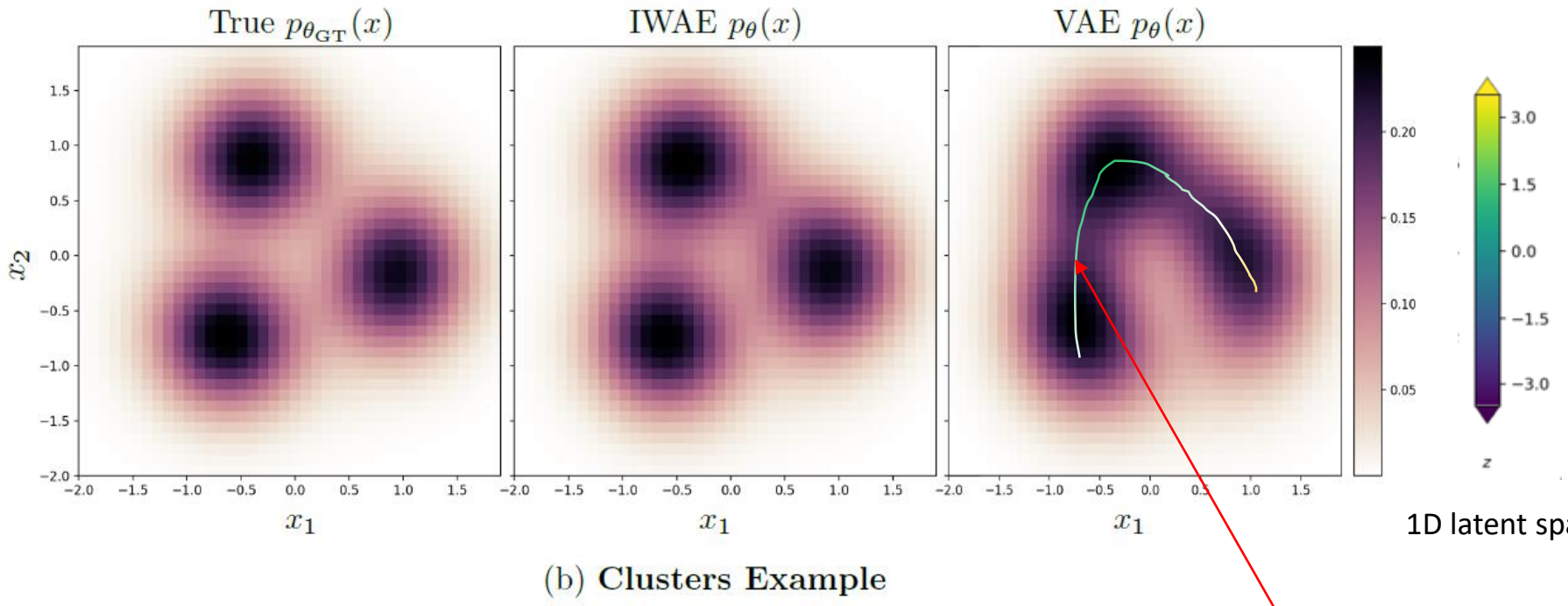
1D latent space

The largest density of the GT $p(x|z)$ is clustered

→ Between cluster many x could have been generated from different z

→ **Cx1 & Cx2 fullfilled**

Cluster example where MFG-VAE fails to model 2D $p(x)$



(b) **Clusters Example**

The $f_\theta(z)$ of MFG-VAE does not resemble "clusters" but is a continuous function (along a triangle) → for every point it is quite clear which z has generated the point

→ As expected MFG-VAE fails to model the clustered data distribution

Failure on latent space

8-shaped and cluster example revisited

VAE	Figure-8 Example			Clusters Example		
	$K = 1$ (GT)	$K = 2$	$K = 3$	$K = 1$ (GT)	$K = 2$	$K = 3$
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		
	$K = 1$ (GT)	$K = 2$	$K = 3$	$K = 1$ (GT)	$K = 2$	$K = 3$
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 1: **The ELBO prefers learning models with more latent dimensions (and smaller σ_ϵ^2) over the ground-truth (GT) model ($k = 1$).** Although the models preferred by the ELBO have higher mutual information between the data and learned z's, the average mutual information between dimensions of z and the data decreases since with more latent dimensions, the latent space learns ϵ. In contrast, IWAE does not suffer from

→ MFG-VAE prefers larger K (z-dimension) and smaller σ_ϵ^2 than used in GT.

Downstream effects: Bad data compression & bad signal/noise decomposition

Intuition: larger K (z-dimension) than GT- K and smaller σ_ϵ^2 Csn2 there exist alternative (non-GT) generative models that explain p(x) well but have more MFG-like posteriors.

E.g. in 8-shape the problem was that noisy points close to 1D z-curve crossing had bimodal posteriors. Higher K alleviates need for crossing, small σ_ϵ^2 leads to more uni-modal posterior

Conditions when MFG-VAE fails to learn the GT latent space

The GT latent representation can not be recovered by the global ELBO optimal MGF-VAE solution if both of the following 2 conditions hold:

Cz1 (=Cx1):

The true posterior $p(z|x)$ for the GT $f_{\theta_{GT}}$ [with $x|z \sim N(f_{\theta_{GT}}(z), \sigma_\epsilon^2 I)$] is difficult to approximate by an MFG for a large proportion of x.

Cz2 (opposite of Cx1):

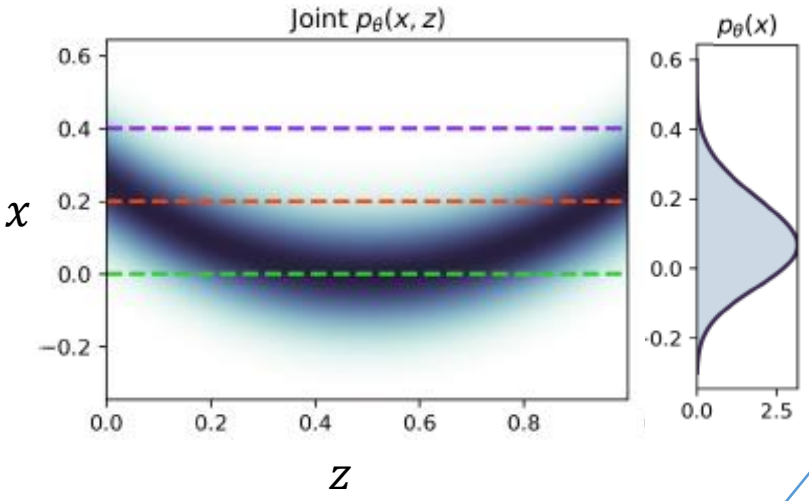
There exists a (likelihood function) f_θ with [a corresponding] simpler [more MFG like] posterior that approximates $p(x)$ well.

Proof: see appendix in paper.

Example 1 revisited

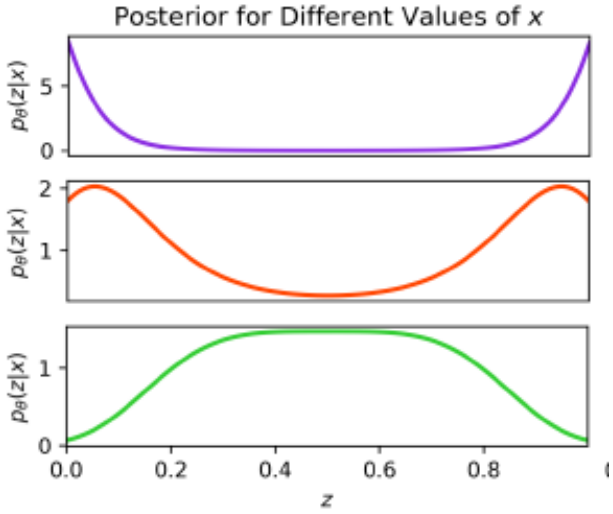
GT: data generating process for simulating data:

$$z \sim U(0,1), \quad (x|z) \sim N(f_{\theta_{GT}} = (z - 0.5)^2, \sigma_\epsilon^2)$$



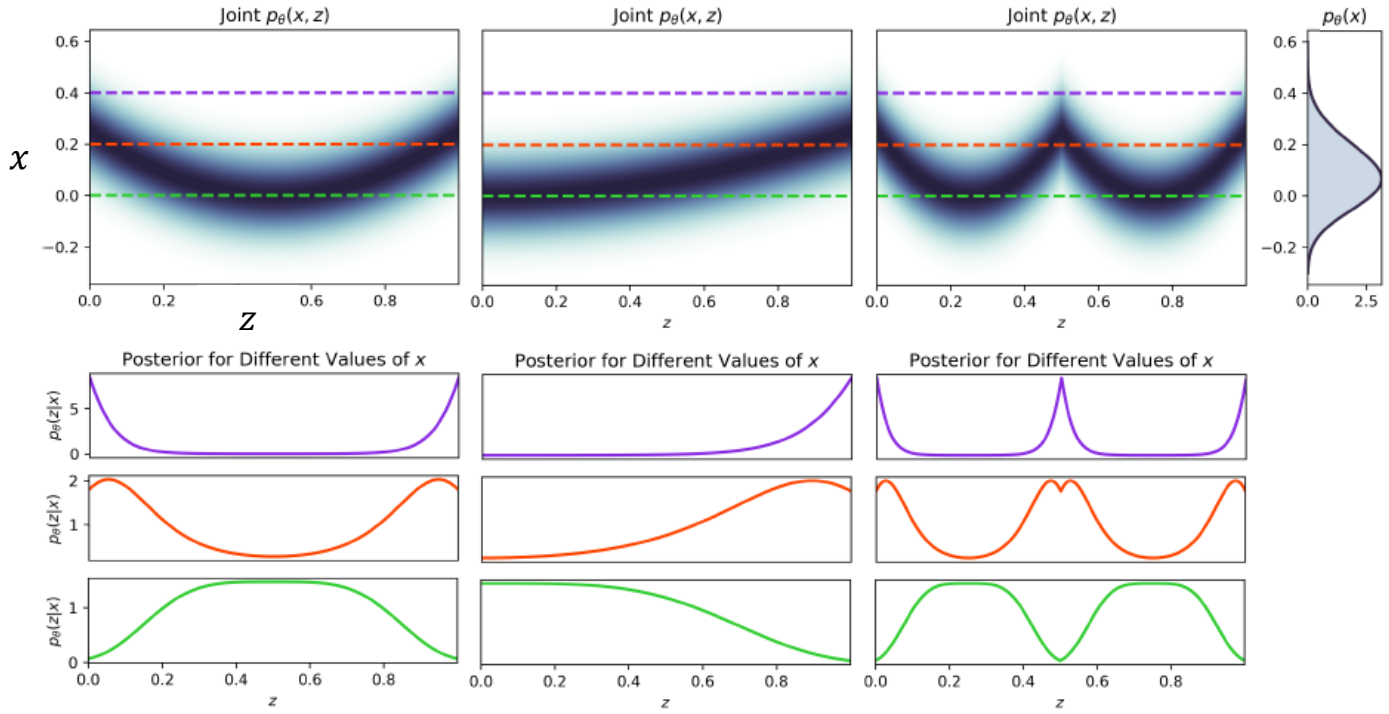
For many x-values the posterior is not similar to a Gaussian

GT posterior $p(z|x)$ examples



Cz1 (=Cx1) is fulfilled,
since the GT posterior $p(z|x)$ for the GT $f_{\theta_{GT}}$ $[x|z \sim N(f_{\theta_{GT}}(z), \sigma_\epsilon^2 I)]$
is difficult to approximate by an MFG for a large proportion of x.

Example 1 revisited



(a) **Variant 1**

$f_\theta(z) = (z - 0.5)^2$

GT

(b) **Variant 2**

$f_\theta(z) = 0.25 \cdot z^2$

example

(c) **Variant 3**

$f_\theta(z) = \mathbb{I}(z < 0.5) \cdot (2z - 0.5)^2 + \mathbb{I}(z \geq 0.5) \cdot (2z - 1.5)^2$

Cz2 (opposite of Cx2 is fullfileed,
since there exists a (likelihood function) f_θ with [a corrosponding] simpler [more MFG like] posterior that approximates $p(x)$ well.

Compromised downstream tasks if latent space is misestimated

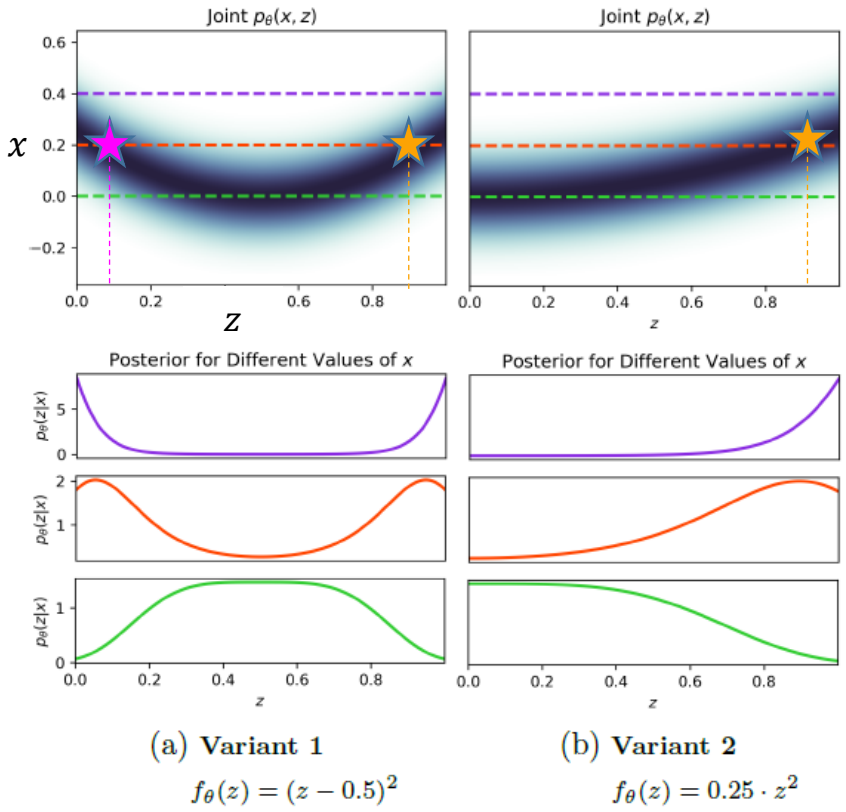
MFG-VAE fails to provide correct inference of the underlying latent factors.

Assume:
Z: underlying diseases
X: observable symptoms

A patient with symptoms x=0.2

- Could have 2 different diseases, **z=0.1** and **z=0.9**
  → correct GT interpretation
- Has disease **z=0.9**
  → wrong MFG-VAE interpretation

GT MFG-VAE solution



(a) **Variant 1** $f_\theta(z) = (z - 0.5)^2$

(b) **Variant 2** $f_\theta(z) = 0.25 \cdot z^2$

Failure on counterfactuals

Recall semi-supvervised MFG-VAE

$$\mathcal{J}(\theta,\phi) = \sum_{n=1}^{N} \mathcal{U}(x_n;\theta,\phi) + \gamma \cdot \sum_{m=1}^{M} \mathcal{L}(x_m,y_m;\theta,\phi) + \alpha \cdot \sum_{m=1}^{M} \log q_\phi(y_m|x_m) \qquad (5)$$

where the $\mathcal{U}$ and $\mathcal{L}$ lower bound $p_\theta(x)$ and $p_\theta(x,y)$, respectively (see Appendix B); the last

N: number of unlabeled data (only x observed)
M: number of labeled data (x and y observed)
γ, α: tuning parameters

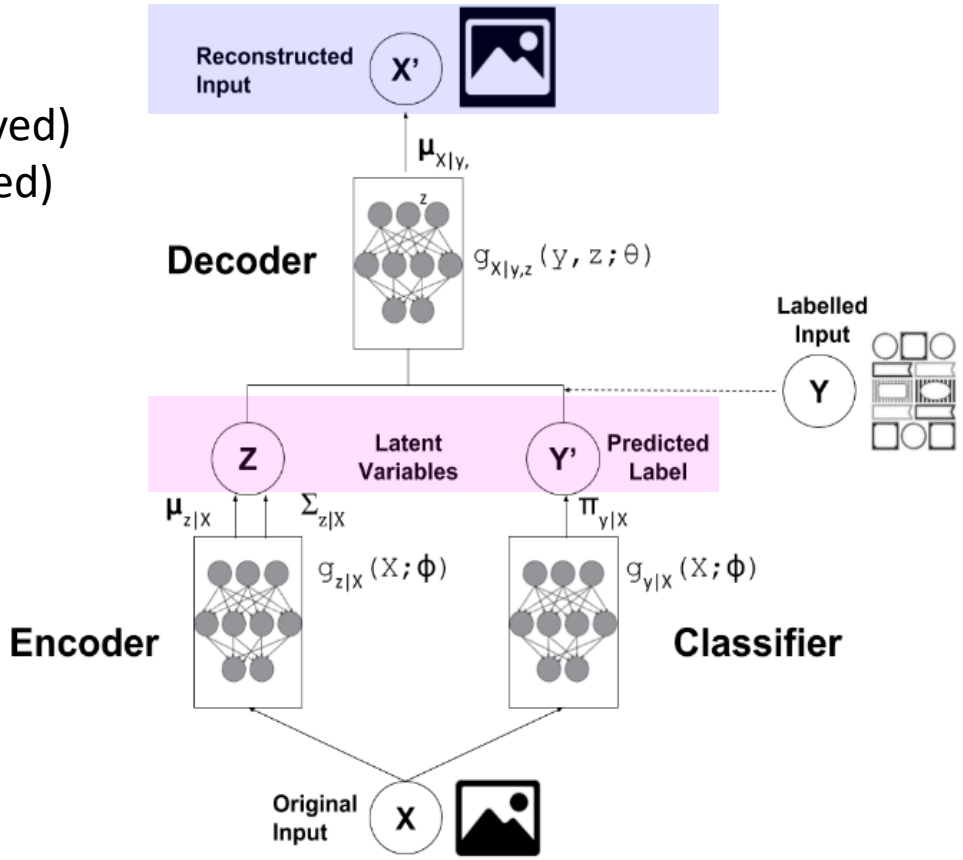


Figure 2: M2 Variational Autoencoder for Semi-Supervised Learning

https://bjlkeng.github.io/posts/semi-supervised-learning-with-variational-autoencoders/

Example: 2 patient types (y=0,1) with overlapping symptoms (x)

GT:

Each patient type ($y = 0$ or $y = 1$) has another dependency of the latent representation on x (z|y,x)

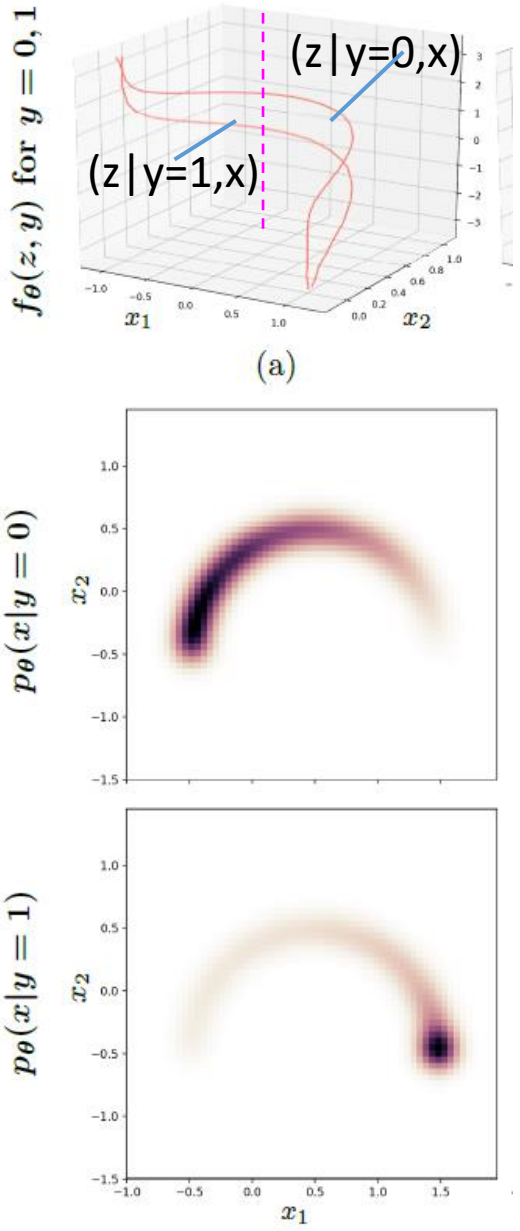
→ Posterior $p(z|x)$ is bimodal for many x

$p(x|y = 0)$ and $p(x|y = 1)$ lie on same manifold

→ Symptoms overlap,

however the distributions is different

$p(x|y = 0) \neq p(x|y = 1)$



Example: 2 patient types (y=0,1) with overlapping symptoms (x)

MFG collapses the z-versus-x-curves for $y = 0$ and $y = 1$ in the z-x-space, «rewarded» by a MFG like posterior $p(z|x)$.

MFG-VAE fails to

- recover the GT latent space
- the different conditional distributions (symptom frequencies in patient type $y = 0$ and $y = 1$)
- The symptoms distribution $p(x)$

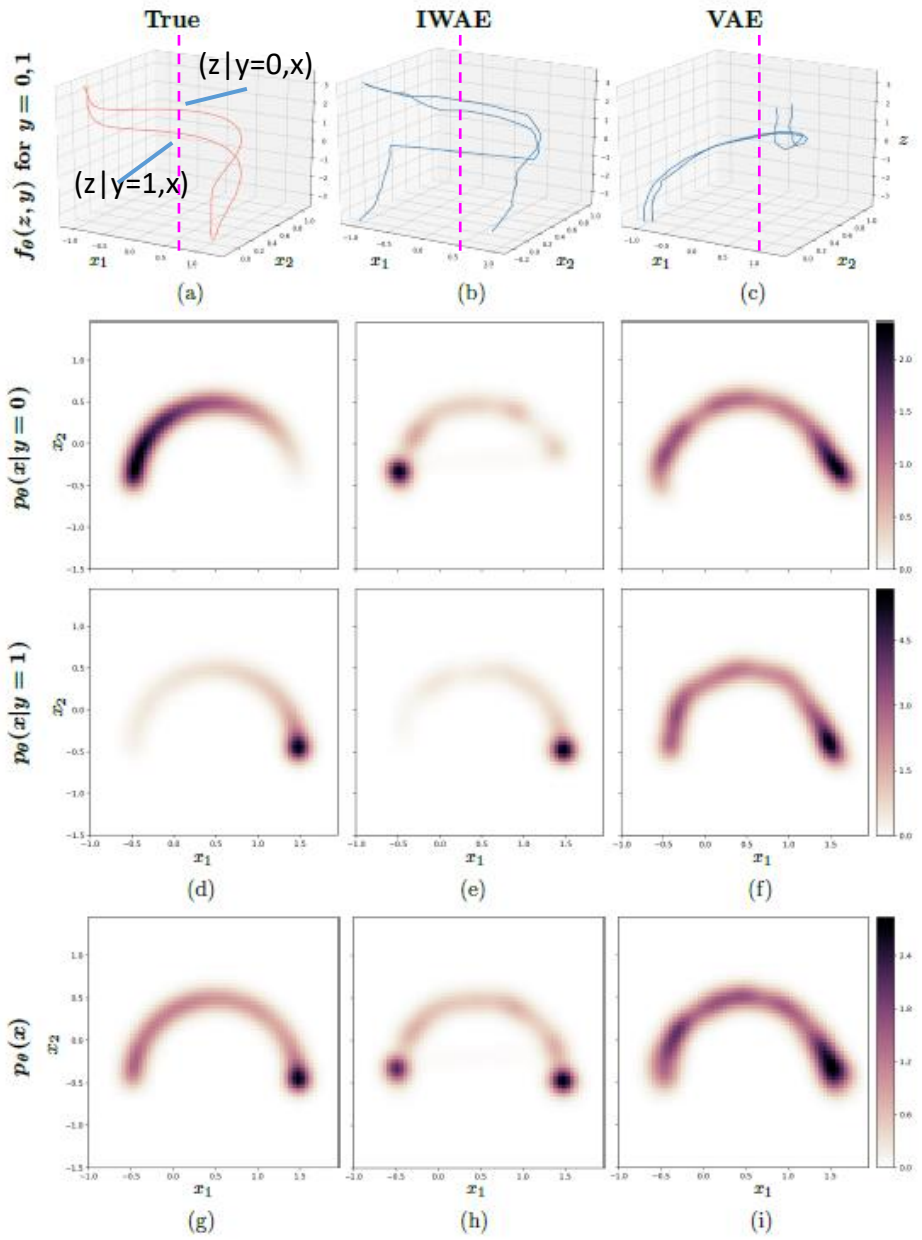


Figure 6: **Semi-supervised VAE exhibits "functional collapse" on the Discrete Semi-**

Guidelines for partitioners

- Set σ_ϵ using domain expertise of via unbiased hyper-parameter selection
- If the data is clustered or lives on a manifold in distorted euclidean space (e.g. 8-shaped example), then use a more flexible variational approximation
    - When using a more flexible variational approximation, regularize the decoder network weights to prevent overfitting
- If the data topology can not be investigated, check if a more flexible variational approximation leads to better results
- If the data topology can not be investigated, check if a more flexible variational approximation leads to better results

Summary

- Necessary conditions for MFG-VAE failures are defined and discussed for
  - failures on data distribution estimates (unsupervised)
  - failures on GT latent representation estimates (unsupervised & semi-supervised)

→Benchmark data sets were developed with which these failures were triggered and demonstrated on different downstream tasks

→ It was discussed and demonstrated in which situations a more flexible variational approximation helps to avoid these failures

→ Guidelines for partitioners are given