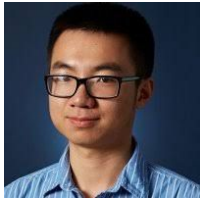
CS294-158 Deep Unsupervised Learning

Lecture 4a: Likelihood Models III: Latent variable models 2

Pieter Abbeel, Peter Chen, Jonathan Ho, Aravind Srinivas

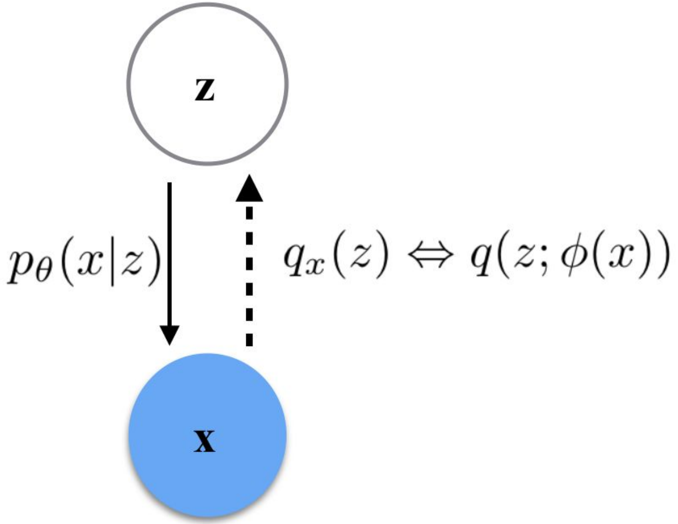
UC Berkeley

Outline

- Warm-up on Variational Inference
  - *Recap*
  - An importance sampling view
  - Variational Mutual Information Estimation/Maximization
  - Variational Dequantization
- Improving VAEs
  - Reducing variational gap
  - More flexible decoder & posterior collapse problem
  - More expressive architectures

Recap: Variational Inference (VI)

Key idea in variational inference is to approximate the true posterior with a variational distribution. This results in an optimization problem that minimizes the distance between these two distributions, ex: KL divergence.



Recap: Variational Inference (VI)

- Let $q_x(z)$ be our approximation of true posterior $p(z|x)$
  - Then we pick a divergence to minimize between these two distributions

$$
\begin{aligned}
D_{\mathrm{KL}}\left[q_x(z) \,\|\, p(z|x)\right] &= \mathbb{E}_{z\sim q_x(z)}\left[\log q_x(z) - \log p(z|x)\right] \\
&= \mathbb{E}_{z\sim q_x(z)}\left[\log q_x(z) - \log \frac{p(z,x)}{p(x)}\right] \\
&= \mathbb{E}_{z\sim q_x(z)}\left[\log q_x(z) - \log p(z) - \log p(x|z) + \log p(x)\right] \\
&= \underbrace{\mathbb{E}_{z\sim q_x(z)}\left[\log q_x(z) - \log p(z) - \log p(x|z)\right]}_{\text{Only this part depends on } z} + \log p(x)
\end{aligned}
$$

- note: the expectation can be approximated by stochastic samples, and every term in expectation can be computed in O(1) now

Outline

- Warm-up on Variational Inference
  - Recap
  - *An importance sampling view*
  - Variational Mutual Information Estimation/Maximization
  - Variational Dequantization
- Improving VAEs
  - Reducing variational gap
  - More flexible decoder & posterior collapse problem
  - More expressive architectures

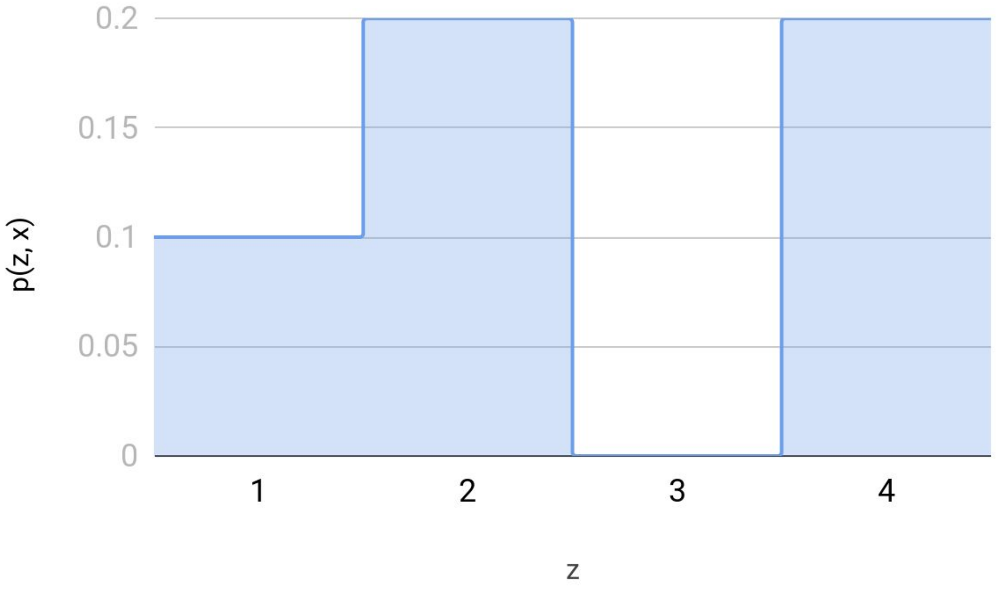
Marginal as enumeration

- To train a latent variable model, we want to evaluate marginal likelihood for any given x

$$p(x) = \sum_z p(z, x)$$

Marginal as enumeration

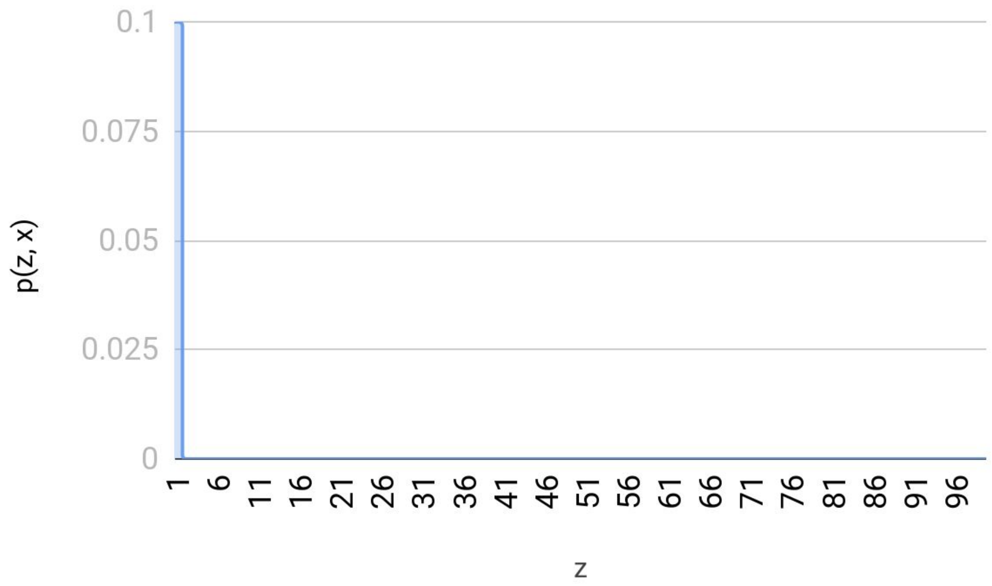
- Suppose we have a categorical z in {1, 2, 3, 4}
  - For a given x, we have the following PMF for joint



Marginal as enumeration

- In reality, z typically is higher dimensional {1, …., 100} graphed here
  - p(z, x) might have probability mass

    concentrated on one z



VI as Importance Sampling

- Intuition: the variational distribution q(z|x) samples the high density region of p(z, x)

$$
\begin{aligned}
\log p(x) &= \log \sum_z p(z, x) \\
&= \log \sum_z q(z|x) \frac{p(z, x)}{q(z|x)} \\
&= \log \mathbb{E}_{z \sim q(z|x)} \left[\frac{p(z, x)}{q(z|x)} \right] \\
&\geq \mathbb{E}_{z \sim q(z|x)} \left[\log \frac{p(z, x)}{q(z|x)} \right]
\end{aligned}
$$

VI as Importance Sampling

- We draw multiple z samples from q(z|x) and name them z_i
- Define w_i and L_k:

$$w_i = \frac{p(z_i, x)}{q(z_i|x)}$$

$$\mathcal{L}_k = \mathbb{E}\left[\log \frac{1}{k}\sum_{i=1}^{k} w_i\right] \leq \log \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} w_i\right] = \log p(\mathbf{x}),$$

[Burda et al., 2015]

VI as Importance Sampling

Theorem 1. *For all k, the lower bounds satisfy*

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

Moreover, if $p(\mathbf{h}, \mathbf{x})/q(\mathbf{h}|\mathbf{x})$ is bounded, then $\mathcal{L}_k$ approaches $\log p(\mathbf{x})$ as k goes to infinity.

[Burda et al., 2015]

VI as Importance Sampling

$$
\begin{aligned}
\mathcal{L}_k &= \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k}\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})}\right] \\
&= \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k}\left[\log \mathbb{E}_{I=\{i_1,\ldots,i_m\}}\left[\frac{1}{m}\sum_{j=1}^{m}\frac{p(\mathbf{x},\mathbf{h}_{i_j})}{q(\mathbf{h}_{i_j}|\mathbf{x})}\right]\right] \\
&\geq \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_k}\left[\mathbb{E}_{I=\{i_1,\ldots,i_m\}}\left[\log \frac{1}{m}\sum_{j=1}^{m}\frac{p(\mathbf{x},\mathbf{h}_{i_j})}{q(\mathbf{h}_{i_j}|\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{\mathbf{h}_1,\ldots,\mathbf{h}_m}\left[\log \frac{1}{m}\sum_{i=1}^{m}\frac{p(\mathbf{x},\mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{x})}\right] = \mathcal{L}_m
\end{aligned}
$$

[Burda et al., 2015]

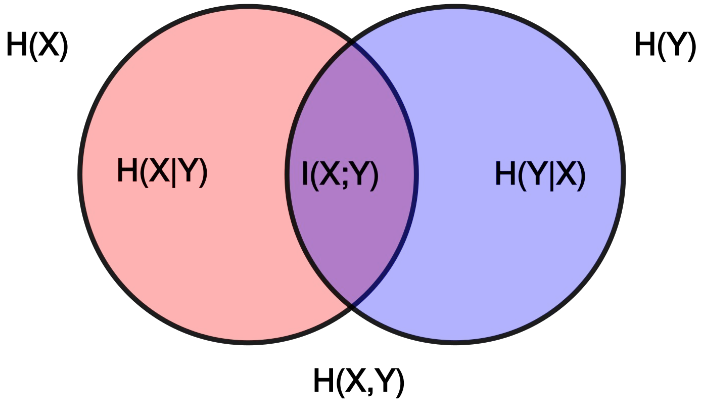
Outline

- Warm-up on Variational Inference
  - Recap
  - An importance sampling view
  - *Variational Mutual Information Estimation/Maximization*
  - Variational Dequantization
- Improving VAEs
  - Reducing variational gap
  - More flexible decoder & posterior collapse problem
  - More expressive architectures

Mutual Information

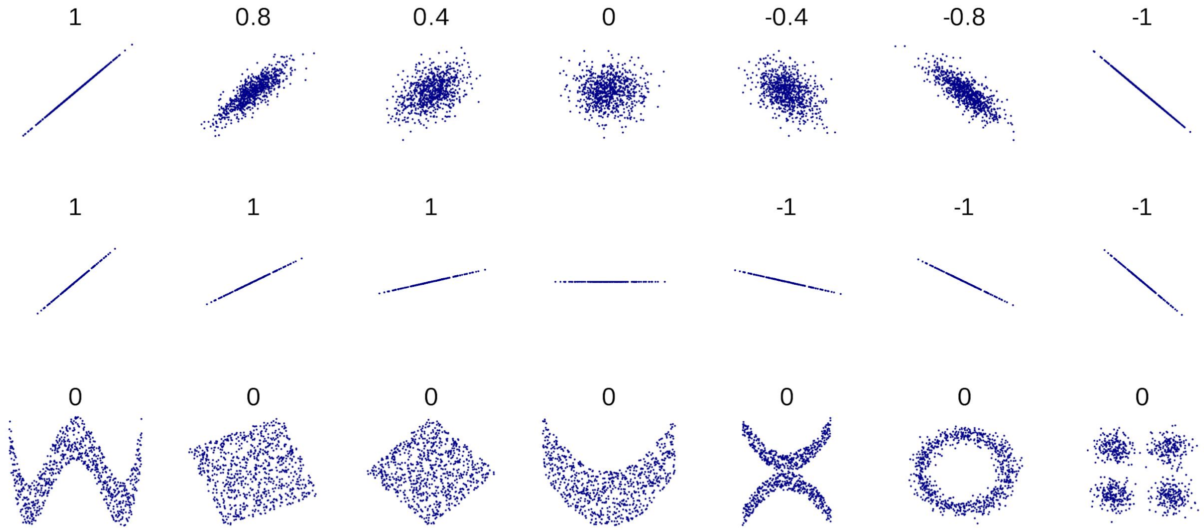
- Mutual information between two random variables X, Y: I(X; Y) is defined as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$



Mutual Information

- Mutual Information is a general way to measure dependency between two random variables
  - Unlike the more commonly used covariance



Mutual Information

- Useful in a lot of settings where one wants to maximize dependency between two variables or estimate their dependencies:
  - Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning
  - InfoGAN
  - CPC
  - ...

Estimating Mutual Information

- We can try to estimate the mutual information between z and x in a latent variable model

$$
\begin{aligned}
I(z;x) =& H(z) - H(z|x) \\
=& H(z) - \mathbb{E}_{(z,x)\sim p(z,x)}\left[-\log p(z|x)\right] \\
=& H(z) + \mathbb{E}_{(z,x)\sim p(z,x)}\left[\log p(z|x) - \log q(z|x) + \log q(z|)\right] \\
\geq& H(z) + \mathbb{E}_{(z,x)\sim p(z,x)}\left[\log q(z|x)\right]
\end{aligned}
$$

- Has intractable posterior p(z|x) but we can estimate by introducing a variational distribution q(z|x)

Outline

- Warm-up on Variational Inference
  - Recap
  - An importance sampling view
  - Variational Mutual Information Estimation/Maximization
  - *Variational Dequantization*
- Improving VAEs
  - Reducing variational gap
  - More flexible decoder & posterior collapse problem
  - More expressive architectures

Recap: Uniform Dequantization

- **Uniform Dequantization**. Add noise to data.
  - $\mathbf{x} \in \{0, 1, 2, \ldots, 255\}$
  - We draw noise u uniformly from $[0, 1)^D$

$$
\begin{aligned}
\mathbb{E}_{\mathbf{y} \sim p_{\text{data}}} \left[\log p_{\text{model}}(\mathbf{y})\right] &= \sum_{\mathbf{x}} P_{\text{data}}(\mathbf{x}) \int_{[0,1)^D} \log p_{\text{model}}(\mathbf{x} + \mathbf{u})\, d\mathbf{u} \\
&\leq \sum_{\mathbf{x}} P_{\text{data}}(\mathbf{x}) \log \int_{[0,1)^D} p_{\text{model}}(\mathbf{x} + \mathbf{u})\, d\mathbf{u} \\
&= \mathbb{E}_{\mathbf{x} \sim P_{\text{data}}} \left[\log P_{\text{model}}(\mathbf{x})\right]
\end{aligned}
$$

[Theis, Oord, Bethge, 2016]

Variational Dequantization

- **Variational Dequantization**. Add a learnable noise q to data.

$$\mathbb{E}_{\mathbf{x}\sim P_{\text{data}}}\left[\log P_{\text{model}}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim P_{\text{data}}}\left[\log \int_{[0,1)^D} q(\mathbf{u}|\mathbf{x}) \frac{p_{\text{model}}(\mathbf{x}+\mathbf{u})}{q(\mathbf{u}|\mathbf{x})}\, d\mathbf{u}\right]$$

$$\geq \mathbb{E}_{\mathbf{x}\sim P_{\text{data}}}\left[\int_{[0,1)^D} q(\mathbf{u}|\mathbf{x}) \log \frac{p_{\text{model}}(\mathbf{x}+\mathbf{u})}{q(\mathbf{u}|\mathbf{x})}\, d\mathbf{u}\right]$$

$$= \mathbb{E}_{\mathbf{x}\sim P_{\text{data}}}\mathbb{E}_{\mathbf{u}\sim q(\cdot|\mathbf{x})}\left[\log \frac{p_{\text{model}}(\mathbf{x}+\mathbf{u})}{q(\mathbf{u}|\mathbf{x})}\right]$$

[Ho et al., 2019]

Outline

- Warm-up on Variational Inference
  - Recap
  - An importance sampling view
  - Variational Mutual Information Estimation/Maximization
  - Variational Dequantization
- Improving VAEs
  - *Reducing variational gap*
  - More flexible decoder & posterior collapse problem
  - More expressive architectures

Reducing variational gap

- Gap between marginal log-likelihood and VLB: mismatch between approximate posterior and true posterior
- To reduce the gap
  - Importance Sampling: IWAE
  - More expressive approximate posterior
  - More expressive prior

Importance Sampling: IWAE

- Trained with Importance Weighted objective L_k

$$w_i = \frac{p(z_i, x)}{q(z_i|x)}$$

$$\mathcal{L}_k = \mathbb{E}\left[\log \frac{1}{k}\sum_{i=1}^{k} w_i\right] \leq \log \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} w_i\right] = \log p(\mathbf{x}),$$

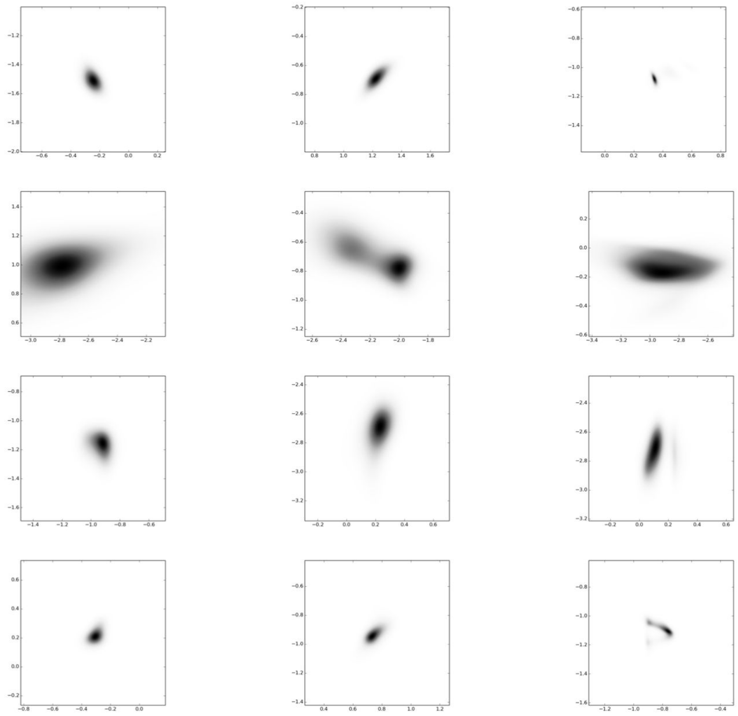
[Burda et al., 2015]

Importance Sampling: IWAE

	MNIST			
	VAE		IWAE	
k	NLL	active units	NLL	active units
1	86.76	19	86.76	19
5	86.47	20	85.54	22
50	86.35	20	84.78	25

Importance Sampling: IWAE

Left: VAE. **Middle:** IWAE, with $k = 5$. **Right:** IWAE, with $k = 50$. The IWAE prefers less regular posteriors and more spread out posterior predictions.

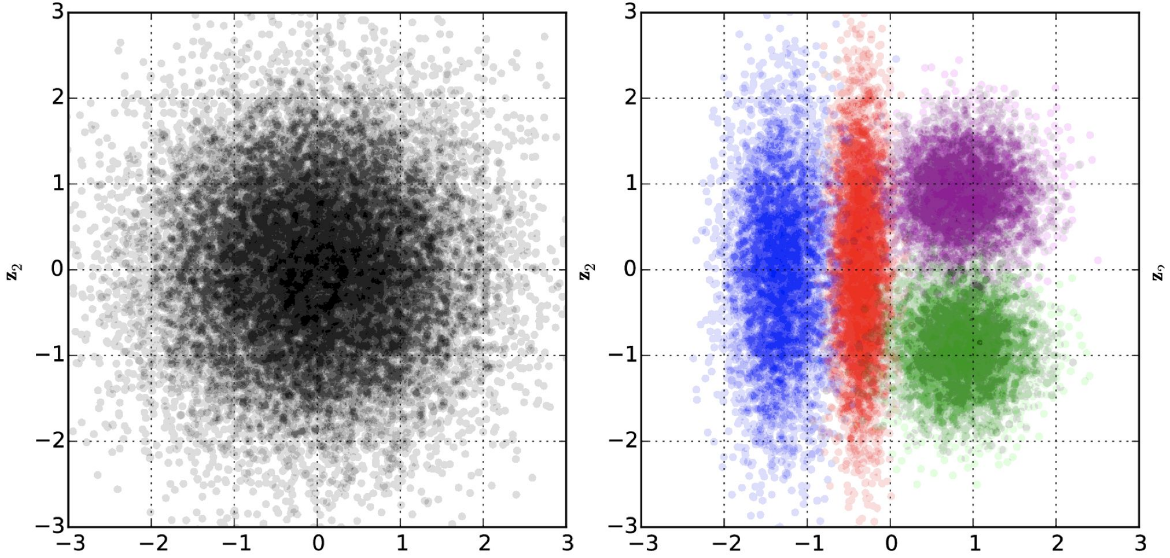


More expressive posterior

- Assuming we have a fixed prior $p(z)$, the approximate posterior $q(z|x)$ has a "bin-packing" problem
  - For each data point x, $q(z|x)$ finds a distinct region in $p(z)$, so $p(x|z)$ can reconstruct that datapoint with as little information loss as possible
  - With all data points, $q(z|x)$ should "tile" $p(z)$ efficiently
    - because [E_z [$p(z|x)$] = $p(z)$]
  - * this is only an intuition; in practice we want our models to generalize instead of just memorizing known data points

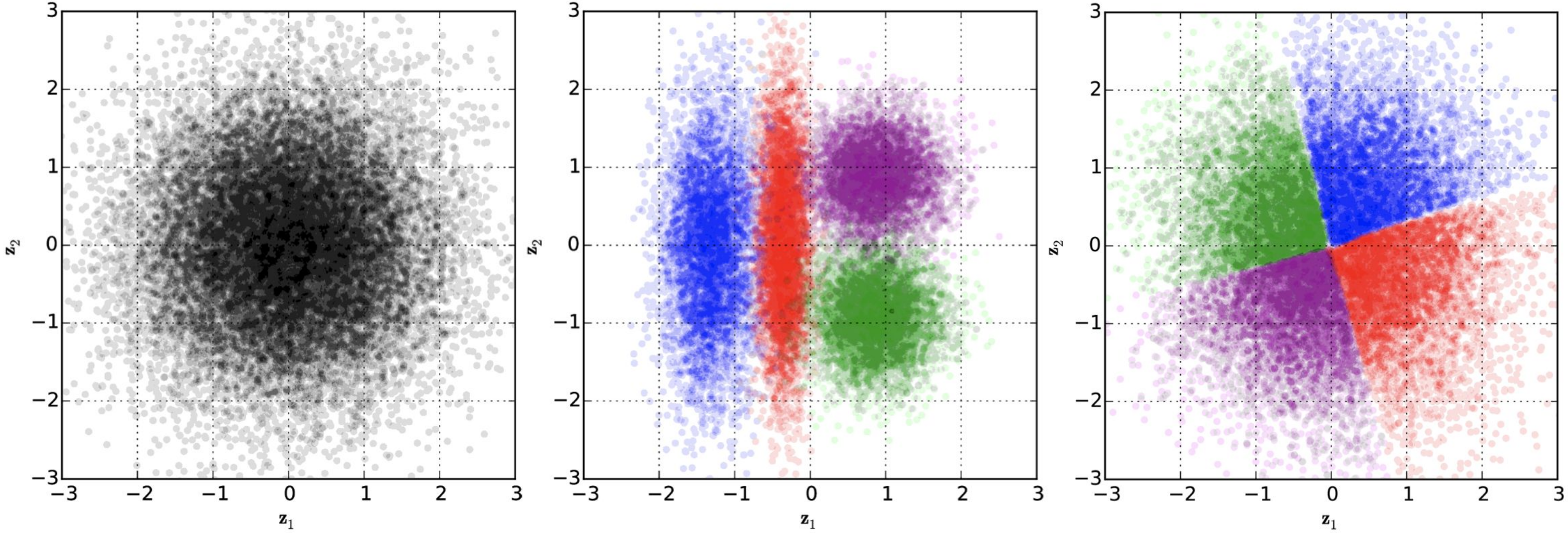
More expressive posterior

- For example, if we have a dataset with 4 datapoints {A, B, C, D}
  - 2D isotropic gaussian as p(z); diagonal covariance gaussian as q(z|x)



More expressive posterior

- Wouldn't it be nice to have q(z|x) that's much more flexible than gaussian with diagonal covariance and pack the space much better?



More expressive posterior

- Core requirements:
  - Computationally efficient to generate: $z \sim q(z|x)$
  - Re-parameterizable: $z = f(\varepsilon; \varphi)$
  - Expressive: $p(z|x)$ can have very difficult form, multi-modal, etc..
- Many works dedicated to finding more expressive $q(z|x)$
  - Normalizing Flow (2015), Hamiltonian Variational Inference (2015)
  - Inverse Autoregressive Flow (2016), Variational Boosting (2016)
  - Householder Flow (2017)
  - Sylvester Normalizing Flows (2018)
  - ……

Recap: Inverse autoregressive flows

- The inverse of an autoregressive flow is also a flow, called the **inverse autoregressive flow (IAF)**
  - **x** $\longrightarrow$ **z** has the same structure as the **sampling** in an autoregressive model
  - **z** $\longrightarrow$ **x** has the same structure as **log likelihood** computation of an autoregressive model. So, **IAF sampling is fast**

$$
\begin{aligned}
z_1 &= f_\theta^{-1}(x_1) \\
z_2 &= f_\theta^{-1}(x_2; z_1) \\
z_3 &= f_\theta^{-1}(x_3; z_1, z_2)
\end{aligned}
$$

$$
\begin{aligned}
x_1 &= f_\theta(z_1) \\
x_2 &= f_\theta(z_2; z_1) \\
x_3 &= f_\theta(z_3; z_1, z_2)
\end{aligned}
$$

IAF-VAE

$$[\mathbf{m}_t, \mathbf{s}_t] \leftarrow \texttt{AutoregressiveNN}[t](\mathbf{z}_t, \mathbf{h}; \boldsymbol{\theta}) \tag{12}$$

and compute $\mathbf{z}_t$ as:

$$\boldsymbol{\sigma}_t = \text{sigmoid}(\mathbf{s}_t) \tag{13}$$

$$\mathbf{z}_t = \boldsymbol{\sigma}_t \odot \mathbf{z}_{t-1} + (1 - \boldsymbol{\sigma}_t) \odot \mathbf{m}_t \tag{14}$$

- s_t initialized to be +2, so sigmoid(s) is close to identity

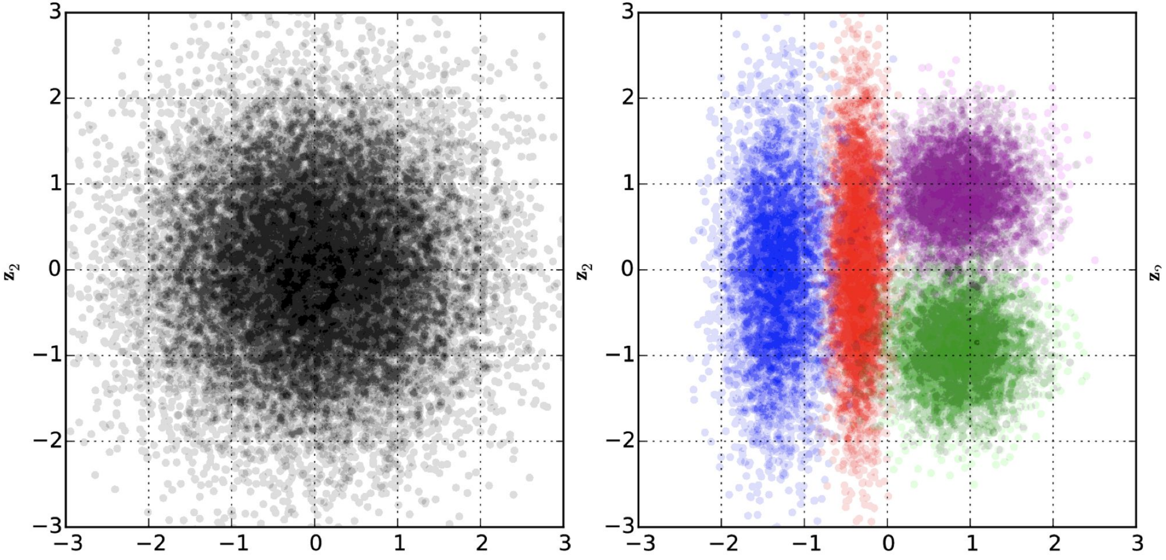
IAF-VAE

- AutoregressiveNN parameterized as 2-layer MADE

Model	VLB	$\log p(\mathbf{x}) \approx$
Convolutional VAE + HVI [1]	-83.49	-81.94
DLGM 2hl + IWAE [2]		-82.90
LVAE [3]		-81.74
DRAW + VGP [4]	-79.88	
Diagonal covariance	-84.08 ($\pm$ 0.10)	-81.08 ($\pm$ 0.08)
IAF (Depth $=2$, Width $=320$)	-82.02 ($\pm$ 0.08)	-79.77 ($\pm$ 0.06)
IAF (Depth $=2$, Width $=1920$)	-81.17 ($\pm$ 0.08)	-79.30 ($\pm$ 0.08)
IAF (Depth $=4$, Width $=1920$)	-80.93 ($\pm$ 0.09)	-79.17 ($\pm$ 0.08)
IAF (Depth $=8$, Width $=1920$)	-80.80 ($\pm$ 0.07)	**-79.10** ($\pm$ 0.07)

More expressive prior

- Recall the bin-packing analogy, we can make tighter packing possible by changing the “bin” shape



More expressive prior

- Core requirements:
  - Computationally efficient to evaluate p(z) for arbitrary z
  - Expressive: p(z) can have very difficult form, multi-modal, etc..
- Many different works
  - AF prior in VLAE (2016)
  - PixelCNN prior in VQ-VAE (2017)

AF Prior

- AF prior = Unconditional IAF posterior
  - Only difference is that AF prior has deeper generative path: log p(x|f(ε)) versus log p(x|ε)

$$
\begin{aligned}
\mathcal{L}(\mathbf{x};\theta) &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}),\epsilon=f^{-1}(\mathbf{z})}\left[\log p(\mathbf{x}|f(\epsilon)) + \log u(\epsilon) + \log\det\frac{d\epsilon}{d\mathbf{z}} - \log q(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}),\epsilon=f^{-1}(\mathbf{z})}\left[\log p(\mathbf{x}|f(\epsilon)) + \log u(\epsilon) - \underbrace{(\log q(\mathbf{z}|\mathbf{x}) - \log\det\frac{d\epsilon}{d\mathbf{z}})}_{\text{IAF Posterior}}\right]
\end{aligned}
$$

AF Prior

Table 1: Statically Binarized MNIST

Model	**NLL Test**
Normalizing flows (Rezende & Mohamed, 2015)	85.10
DRAW (Gregor et al., 2015)	< 80.97
Discrete VAE (Rolfe, 2016)	81.01
PixelRNN (van den Oord et al., 2016a)	79.20
IAF VAE (Kingma et al., 2016)	79.88
AF VAE	79.30
VLAE	**79.03**

Outline

- Warm-up on Variational Inference
  - Recap
  - An importance sampling view
  - Variational Mutual Information Estimation/Maximization
  - Variational Dequantization
- Improving VAEs
  - Reducing variational gap
  - *More flexible decoder & posterior collapse problem*
  - More expressive architectures

Decoder distribution

- So far all models use simple distribution for $p(x|z)$
- Due to lack of expressivity itself, all entropy is pushed to z and z needs to convey a lot of information

Powerful decoder

- What's the maximum VLB?

$$\mathbb{E}_{x\sim p_{\text{data}}(x)}[VLB] \leq \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log p_\theta(x)]$$
$$\leq \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log p_{\text{data}}(x)]$$

- What if $p(x|z) = p_{\text{data}}(x)$?

$$\mathbb{E}_{x\sim p_{\text{data}}(x)}[VLB] = \mathbb{E}_{x\sim p_{\text{data}}(x), z\sim q(z|x)}[\log p(x|z) + \log p(z) - \log q(z|x)]$$
$$= \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log p_{\text{data}}(x) + \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log p(z) - \log q(z|x)]\right]$$

  - q(z|x) would be set to p(z) -> z has no information

Powerful decoder

- Having information in z incurs vlb penalty of KL(q ll p) which is usually non-zero
- “Ignoring latent code” problems well documented in literature
  - (Fabius & van Amersfoort,2014; Chung et al., 2015; Bowman et al., 2015; Serban et al., 2016; Fraccaro et al., 2016; Xu &Sun, 2016)
  - Many proposed solutions

Weakening models

- Adding dropout in autoregressive conditioning (Bowman et al., 2015)
- PixelCNN with limited receptive field (Chen et al., 2016)
- Constant bit rate $D_{KL}(q_\varphi(z|x) \parallel p_\theta(z)) = c$ (Guu et al., 2017), (Xu & Durrett, 2018), (Davidson et al., 2018)
- Minimum bit rate $D_{KL}(q_\varphi(z|x) \parallel p_\theta(z)) \geq \delta$ (Razavi et al., 2019)

Changing training dynamics

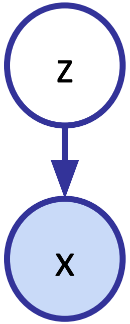
- $D_{KL}(q_\varphi(z|x) \parallel p_\theta(z))$ warmup (Bowman et al., 2015); (Yang et al., 2017); (Kim et al., 2018); (Gulrajani et al., 2016)
- "Free-bits" (Kingma et al., 2016); (Chen et al., 2016)
- More training updates to $q(z|x)$ (He et al., 2019)

Outline

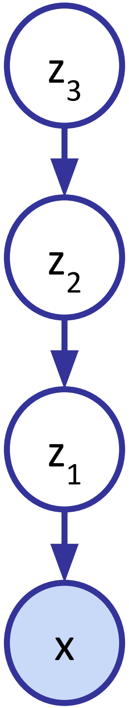
- Warm-up on Variational Inference
  - Recap
  - An importance sampling view
  - Variational Mutual Information Estimation/Maximization
  - Variational Dequantization
- Improving VAEs
  - Reducing variational gap
  - More flexible decoder & posterior collapse problem
  - *More expressive architectures*

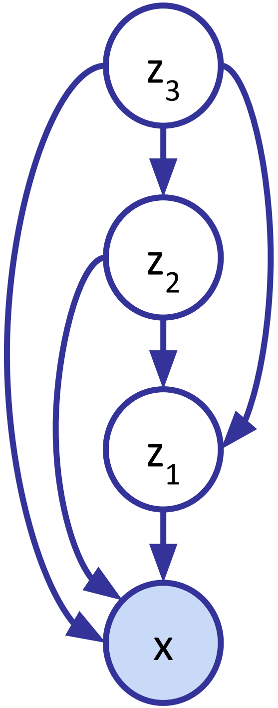
Hierarchical latent variable models

$$p(x, z) = p(x|z)p(z)$$



$$p(x, z_{1:L}) = p(x|z_{1:L}) \left(\prod_{i=1}^{L-1} p(z_i \mid z_{i+1:L}) \right) p(z_L)$$





Training with multiple latent variables

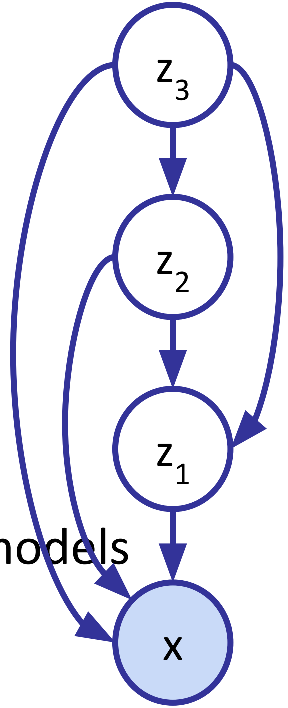
- Generation path

$$p(x, z_{1:L}) = p(x|z_{1:L}) \left(\prod_{i=1}^{L-1} p(z_i \mid z_{i+1:L}) \right) p(z_L)$$

- Variational lower bound

$$\log p(x) \geq \mathbb{E}_{z_{1:L} \sim q(z_{1:L}|x)} \left[\log \frac{p(x, z_{1:L})}{q(z_{1:L}|x)} \right]$$

  - Note: evaluating/differentiating p(x, z) is fast, just like AR models



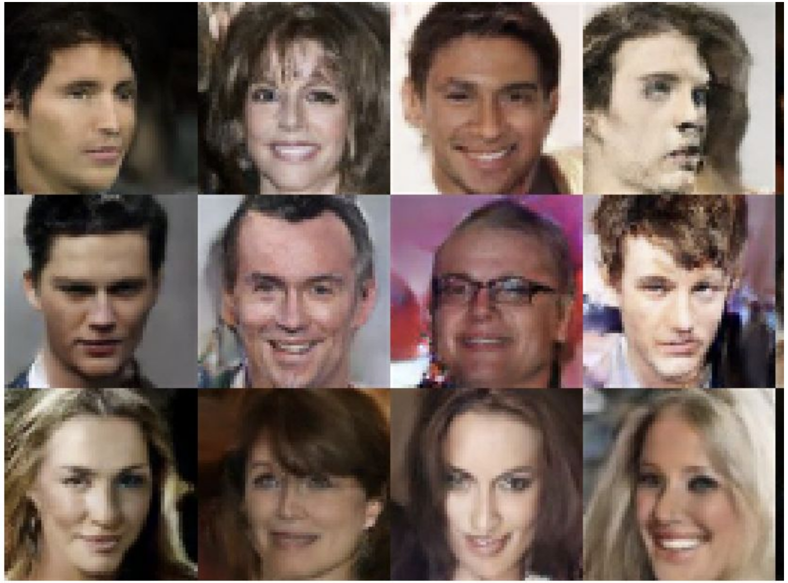
Inference networks for hierarchical models

- $q(z_{1:L}|x)$ should be as flexible as possible, yet fast to sample for fast training
- Example designs
  - IAF-VAE (Kingma et al. 2016)
    - Inverse autoregressive flow for each z, stitched together in an autoregressive fashion over layers 1:L
  - Bidirectional-Inference Variational Autoencoder (BIVA) (Maaløe et al. 2019)
    - Uses autoregressive flows over 1:L
    - Very effective: SOTA on many benchmarks for latent variable models
  - Note: above, autoregressive structure is over layers (not dimension of data), so sampling speed is acceptable.

SOTA

	BITS/DIM
Results with autoregressive components	
CONVDRAW (GREGOR ET AL., 2016)	< 3.58
IAFVAE $\mathcal{L}_1$ (KINGMA ET AL., 2016)	≤ 3.15
IAFVAE $\mathcal{L}_{1e3}$ (KINGMA ET AL., 2016)	≤ 3.11
GATEDPIXELCNN (VAN DEN OORD ET AL., 2016B)	$= 3.03$
PIXELRNN (VAN DEN OORD ET AL., 2016C)	$= 3.00$
VLAE (CHEN ET AL., 2017)	≤ 2.95
PIXELCNN++ (SALIMANS ET AL., 2017)	$= 2.92$
Results without autoregressive components	
NICE (DINH ET AL., 2014)	$= 4.48$
DEEPGMMS (VAN DEN OORD & SCHRAUWEN, 2014)	$= 4.00$
REALNVP (DINH ET AL., 2016)	$= 3.49$
DISCRETEVAE++ (VAHDAT ET AL., 2018)	≤ 3.38
GLOW (KINGMA & DHARIWAL, 2018)	$= 3.35$
BIVA L=10, $\mathcal{L}_1$	≤ 3.17
BIVA L=15, $\mathcal{L}_1$	≤ 3.12
BIVA L=15, $\mathcal{L}_{1e3}$	≤ 3.08

Table 4. Test log-likelihood on CIFAR-10 for different number of importance weighted samples. We evaluated two different BIVA with various number of layers (L).



[Maaløe et al. 2019]