

# Deep Generative Models

## Lecture 6

Roman Isachenko



2024, Spring

## Recap of previous lecture

### EM-algorithm

- ▶ E-step

$$q^*(\mathbf{z}) = \arg \max_q \mathcal{L}(q, \theta^*) = \arg \min_q KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}, \theta^*));$$

- ▶ M-step

$$\theta^* = \arg \max_{\theta} \mathcal{L}(q^*, \theta);$$

### Amortized variational inference

Restrict a family of all possible distributions  $q(\mathbf{z})$  to a parametric class  $q(\mathbf{z}|\mathbf{x}, \phi)$  conditioned on samples  $\mathbf{x}$  with parameters  $\phi$ .

### Variational Bayes

- ▶ E-step

$$\phi_k = \phi_{k-1} + \eta \cdot \nabla_{\phi} \mathcal{L}(\phi, \theta_{k-1})|_{\phi=\phi_{k-1}}$$

- ▶ M-step

$$\theta_k = \theta_{k-1} + \eta \cdot \nabla_{\theta} \mathcal{L}(\phi_k, \theta)|_{\theta=\theta_{k-1}}$$

## Recap of previous lecture

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \left[ \log p(\mathbf{x}|\mathbf{z}, \theta) - \log \frac{q(\mathbf{z}|\mathbf{x}, \phi)}{p(\mathbf{z})} \right] \rightarrow \max_{\phi, \theta}.$$

M-step:  $\nabla_{\theta} \mathcal{L}(\phi, \theta)$ , Monte Carlo estimation

$$\begin{aligned} \nabla_{\theta} \mathcal{L}(\phi, \theta) &= \int q(\mathbf{z}|\mathbf{x}, \phi) \nabla_{\theta} \log p(\mathbf{x}|\mathbf{z}, \theta) d\mathbf{z} \approx \\ &\approx \nabla_{\theta} \log p(\mathbf{x}|\mathbf{z}^*, \theta), \quad \mathbf{z}^* \sim q(\mathbf{z}|\mathbf{x}, \phi). \end{aligned}$$

E-step:  $\nabla_{\phi} \mathcal{L}(\phi, \theta)$ , reparametrization trick

$$\begin{aligned} \nabla_{\phi} \mathcal{L}(\phi, \theta) &= \int r(\epsilon) \nabla_{\phi} \log p(\mathbf{x}|\mathbf{g}_{\phi}(\mathbf{x}, \epsilon), \theta) d\epsilon - \nabla_{\phi} \text{KL} \\ &\approx \nabla_{\phi} \log p(\mathbf{x}|\mathbf{g}_{\phi}(\mathbf{x}, \epsilon^*), \theta) - \nabla_{\phi} \text{KL} \end{aligned}$$

Variational assumption

$$r(\epsilon) = \mathcal{N}(0, \mathbf{I}); \quad q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\boldsymbol{\mu}_{\phi}(\mathbf{x}), \boldsymbol{\sigma}_{\phi}^2(\mathbf{x})).$$

$$\mathbf{z} = \mathbf{g}_{\phi}(\mathbf{x}, \epsilon) = \boldsymbol{\sigma}_{\phi}(\mathbf{x}) \odot \epsilon + \boldsymbol{\mu}_{\phi}(\mathbf{x}).$$

## Recap of previous lecture

### Final EM-algorithm

- ▶ pick random sample  $\mathbf{x}_i, i \sim U[1, n]$ .
- ▶ compute the objective:

$$\epsilon^* \sim r(\epsilon); \quad \mathbf{z}^* = \mathbf{g}_\phi(\mathbf{x}, \epsilon^*);$$

$$\mathcal{L}(\phi, \theta) \approx \log p(\mathbf{x}|\mathbf{z}^*, \theta) - KL(q(\mathbf{z}^*|\mathbf{x}, \phi)||p(\mathbf{z}^*)).$$

- ▶ compute a stochastic gradients w.r.t.  $\phi$  and  $\theta$

$$\nabla_\phi \mathcal{L}(\phi, \theta) \approx \nabla_\phi \log p(\mathbf{x}|\mathbf{g}_\phi(\mathbf{x}, \epsilon^*), \theta) - \nabla_\phi KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}));$$

$$\nabla_\theta \mathcal{L}(\phi, \theta) \approx \nabla_\theta \log p(\mathbf{x}|\mathbf{z}^*, \theta).$$

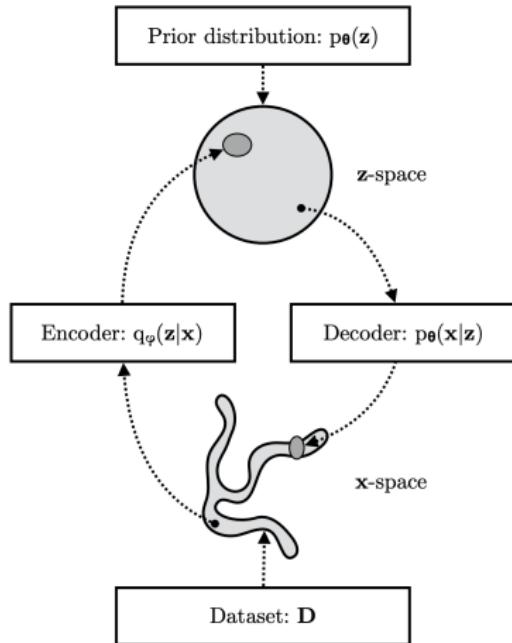
- ▶ update  $\theta, \phi$  according to the selected optimization method (SGD, Adam):

$$\begin{aligned}\phi &:= \phi + \eta \cdot \nabla_\phi \mathcal{L}(\phi, \theta), \\ \theta &:= \theta + \eta \cdot \nabla_\theta \mathcal{L}(\phi, \theta).\end{aligned}$$

# Recap of previous lecture

## Variational autoencoder (VAE)

- ▶ VAE learns stochastic mapping between  $\mathbf{x}$ -space, from  $\pi(\mathbf{x})$ , and a latent  $\mathbf{z}$ -space, with simple distribution.
- ▶ The generative model learns distribution  $p(\mathbf{x}, \mathbf{z}|\theta) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}, \theta)$ , with a prior distribution  $p(\mathbf{z})$ , and a stochastic decoder  $p(\mathbf{x}|\mathbf{z}, \theta)$ .
- ▶ The stochastic encoder  $q(\mathbf{z}|\mathbf{x}, \phi)$  (inference model), approximates the true but intractable posterior  $p(\mathbf{z}|\mathbf{x}, \theta)$ .



## Recap of previous lecture

	VAE	NF
<b>Objective</b>	ELBO $\mathcal{L}$	Forward KL/MLE
<b>Encoder</b>	stochastic $z \sim q(z x, \phi)$	deterministic $z = f_\theta(x)$ $q(z x, \theta) = \delta(z - f_\theta(x))$
<b>Decoder</b>	stochastic $x \sim p(x z, \theta)$	deterministic $x = g_\theta(z)$ $p(x z, \theta) = \delta(x - g_\theta(z))$
<b>Parameters</b>	$\phi, \theta$	$\theta \equiv \phi$

### Theorem

MLE for normalizing flow is equivalent to maximization of ELBO for VAE model with deterministic encoder and decoder:

$$p(x|z, \theta) = \delta(x - f_\theta^{-1}(z)) = \delta(x - g_\theta(z));$$

$$q(z|x, \theta) = p(z|x, \theta) = \delta(z - f_\theta(x)).$$

# Outline

## 1. Discrete VAE latent representations

Vector quantization

Gumbel-softmax for discrete VAE latents

## 2. ELBO surgery

## 3. Learnable VAE prior

# Discrete VAE latents

## Motivation

- ▶ Previous VAE models had **continuous** latent variables  $\mathbf{z}$ .
- ▶ **Discrete** representations  $\mathbf{z}$  are potentially a more natural fit for many of the modalities.
- ▶ Powerful autoregressive models (like PixelCNN) have been developed for modelling distributions over discrete variables.
- ▶ All cool transformer-like models work with discrete tokens.

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}, \phi) || p(\mathbf{z})) \rightarrow \max_{\phi, \theta} .$$

- ▶ Reparametrization trick to get unbiased gradients.
- ▶ Normal assumptions for  $q(\mathbf{z}|\mathbf{x}, \phi)$  and  $p(\mathbf{z})$  to compute KL analytically.

# Discrete VAE latents

## Assumptions

- ▶ Define dictionary (word book) space  $\{\mathbf{e}_k\}_{k=1}^K$ , where  $\mathbf{e}_k \in \mathbb{R}^C$ ,  $K$  is the size of the dictionary.
- ▶ Let  $c \sim \text{Categorical}(\boldsymbol{\pi})$ , where
$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K), \quad \pi_k = P(c = k), \quad \sum_{k=1}^K \pi_k = 1.$$
- ▶ Let VAE model has discrete latent representation  $c$  with prior  $p(c) = \text{Uniform}\{1, \dots, K\}$ .

## How it should work?

- ▶ Our variational posterior  $q(c|\mathbf{x}, \phi) = \text{Categorical}(\boldsymbol{\pi}_\phi(\mathbf{x}))$  (encoder) outputs discrete probabilities vector.
- ▶ We sample  $c^*$  from  $q(c|\mathbf{x}, \phi)$  (reparametrization trick analogue).
- ▶ Our generative distribution  $p(\mathbf{x}|\mathbf{e}_{c^*}, \theta)$  (decoder).

# Discrete VAE latents

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(c|\mathbf{x}, \phi)} \log p(\mathbf{x}|c, \theta) - KL(q(c|\mathbf{x}, \phi)||p(c)) \rightarrow \max_{\phi, \theta} .$$

### KL term

$$\begin{aligned} KL(q(c|\mathbf{x}, \phi)||p(c)) &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log \frac{q(k|\mathbf{x}, \phi)}{p(k)} = \\ &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log q(k|\mathbf{x}, \phi) - \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log p(k) = \\ &= -H(q(c|\mathbf{x}, \phi)) + \log K. \end{aligned}$$

- ▶ Is it possible to make reparametrization trick? (we sample from discrete distribution now!).
- ▶ Entropy term should be estimated.

# Outline

## 1. Discrete VAE latent representations

Vector quantization

Gumbel-softmax for discrete VAE latents

## 2. ELBO surgery

## 3. Learnable VAE prior

# Vector quantization

## Quantized representation

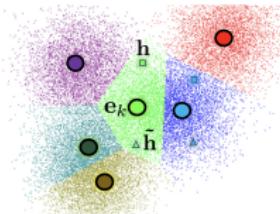
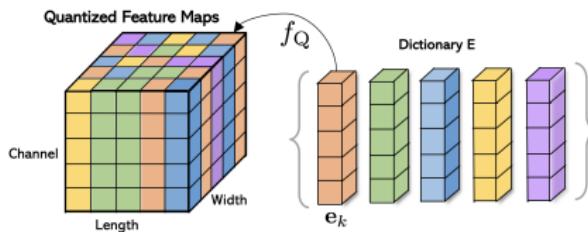
$\mathbf{z}_q \in \mathbb{R}^C$  for  $\mathbf{z} \in \mathbb{R}^C$  is defined by a nearest neighbor look-up using the shared dictionary space

$$\mathbf{z}_q = \mathbf{e}_{k^*}, \quad \text{where } k^* = \arg \min_k \|\mathbf{z} - \mathbf{e}_k\|.$$

- ▶ Let our encoder outputs continuous representation  $\mathbf{z}$ .
- ▶ Quantization will give us the discrete distribution  $q(c|x, \phi)$ .

## Quantization procedure

If we have tensor with the spatial dimensions we apply the quantization for each of  $W \times H$  locations.



## Vector Quantized VAE (VQ-VAE)

Let VAE latent variable  $\mathbf{c} \in \{1, \dots, K\}^{W \times H}$  is the discrete with spatial-independent variational posterior and prior distributions

$$q(\mathbf{c}|\mathbf{x}, \phi) = \prod_{i=1}^W \prod_{j=1}^H q(c_{ij}|\mathbf{x}, \phi); \quad p(\mathbf{c}) = \prod_{i=1}^W \prod_{j=1}^H \text{Uniform}\{1, \dots, K\}.$$

Let  $\mathbf{z}_e = \text{NN}_{e, \phi}(\mathbf{x}) \in \mathbb{R}^{W \times H \times C}$  is the encoder output.

### Deterministic variational posterior

$$q(c_{ij} = k^*|\mathbf{x}, \phi) = \begin{cases} 1, & \text{for } k^* = \arg \min_k \|[\mathbf{z}_e]_{ij} - \mathbf{e}_k\|; \\ 0, & \text{otherwise.} \end{cases}$$

$KL(q(c|\mathbf{x}, \phi)||p(c))$  term in ELBO is constant, entropy of the posterior is zero.

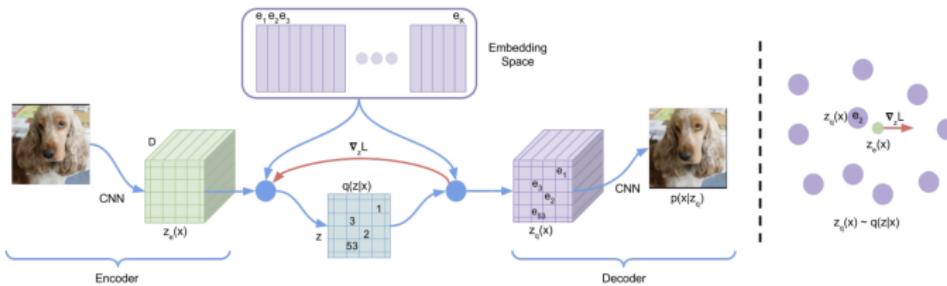
$$KL(q(c|\mathbf{x}, \phi)||p(c)) = -H(q(c|\mathbf{x}, \phi)) + \log K = \log K.$$

# Vector Quantized VAE (VQ-VAE)

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(c|x, \phi)} \log p(x|e_c, \theta) - \log K = \log p(x|z_q, \theta) - \log K,$$

where  $z_q = e_{k^*}$ ,  $k^* = \arg \min_k \|z_e - e_k\|$ .



**Problem:**  $\arg \min$  is not differentiable.

**Straight-through gradient estimation**

$$\frac{\partial \log p(x|z_q, \theta)}{\partial \phi} = \frac{\partial \log p(x|z_q, \theta)}{\partial z_q} \cdot \frac{\partial z_q}{\partial \phi} \approx \frac{\partial \log p(x|z_q, \theta)}{\partial z_q} \cdot \frac{\partial z_e}{\partial \phi}$$

# Vector Quantized VAE-2 (VQ-VAE-2)

Samples 1024x1024



Samples diversity



VQ-VAE (Proposed)

BigGAN deep

Razavi A., Oord A., Vinyals O. Generating Diverse High-Fidelity Images with VQ-VAE-2, 2019

# Outline

## 1. Discrete VAE latent representations

Vector quantization

Gumbel-softmax for discrete VAE latents

## 2. ELBO surgery

## 3. Learnable VAE prior

## Gumbel-softmax trick

- ▶ VQ-VAE has deterministic variational posterior (it allows to get rid of discrete sampling and reparametrization trick).
- ▶ There is no uncertainty in the encoder output.

## Gumbel-max trick

Let  $g_k \sim \text{Gumbel}(0, 1)$  for  $k = 1, \dots, K$ , i.e.  $g = -\log(-\log u)$ ,  $u \sim \text{Uniform}[0, 1]$ . Then a discrete random variable

$$c = \arg \max_k [\log \pi_k + g_k],$$

has a categorical distribution  $c \sim \text{Categorical}(\pi)$ .

- ▶ Let our encoder  $q(c|\mathbf{x}, \phi) = \text{Categorical}(\pi_\phi(\mathbf{x}))$  outputs logits of  $\pi_\phi(\mathbf{x})$ .
- ▶ We could sample from the discrete distribution using Gumbel-max reparametrization.

---

*Maddison C. J., Mnih A., Teh Y. W. The Concrete distribution: A continuous relaxation of discrete random variables, 2016*

*Jang E., Gu S., Poole B. Categorical reparameterization with Gumbel-Softmax, 2016*

## Gumbel-softmax trick

### Reparametrization trick (LOTUS)

$$\nabla_{\phi} \mathbb{E}_{q(c|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{e}_c, \theta) = \mathbb{E}_{\text{Gumbel}(0,1)} \nabla_{\phi} \log p(\mathbf{x}|\mathbf{e}_{k^*}, \theta),$$

where  $k^* = \arg \max_k [\log q(k|\mathbf{x}, \phi) + g_k]$ .

**Problem:** We still have non-differentiable  $\arg \max$  operation.

### Gumbel-softmax relaxation

Concrete distribution = continuous + discrete

$$\hat{c}_k = \frac{\exp\left(\frac{\log q(k|\mathbf{x}, \phi) + g_k}{\tau}\right)}{\sum_{j=1}^K \exp\left(\frac{\log q(j|\mathbf{x}, \phi) + g_j}{\tau}\right)}, \quad k = 1, \dots, K.$$

Here  $\tau$  is a temperature parameter. Now we have differentiable operation, but the gradient estimate is biased now.

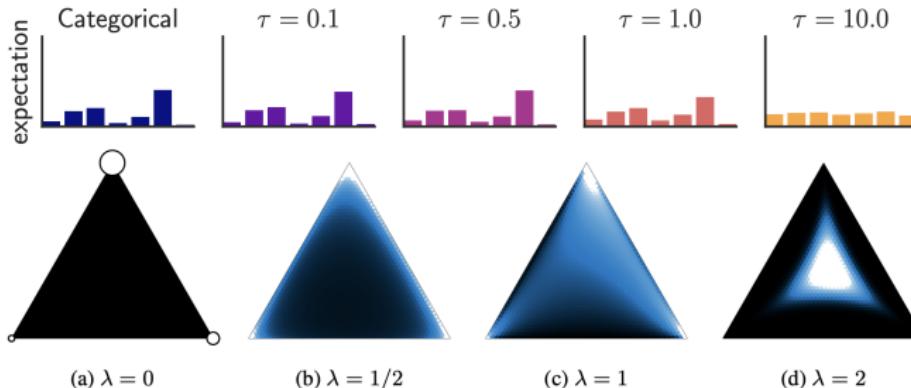
---

Maddison C. J., Mnih A., Teh Y. W. *The Concrete distribution: A continuous relaxation of discrete random variables*, 2016

Jang E., Gu S., Poole B. *Categorical reparameterization with Gumbel-Softmax*, 2016

# Gumbel-softmax trick

## Concrete distribution



## Reparametrization trick

$$\nabla_{\phi} \mathbb{E}_{q(c|x, \phi)} \log p(\mathbf{x}|\mathbf{e}_c, \theta) = \mathbb{E}_{\text{Gumbel}(0,1)} \nabla_{\phi} \log p(\mathbf{x}|\mathbf{z}, \theta),$$

where  $\mathbf{z} = \sum_{k=1}^K \hat{c}_k \mathbf{e}_k$  (all operations are differentiable now).

---

Maddison C. J., Mnih A., Teh Y. W. *The Concrete distribution: A continuous relaxation of discrete random variables*, 2016

Jang E., Gu S., Poole B. *Categorical reparameterization with Gumbel-Softmax*, 2016

# DALL-E/dVAE

## Deterministic VQ-VAE posterior

$$q(\hat{z}_{ij} = k^* | \mathbf{x}) = \begin{cases} 1, & \text{for } k^* = \arg \min_k \|[\mathbf{z}_e]_{ij} - \mathbf{e}_k\| \\ 0, & \text{otherwise.} \end{cases}$$

- ▶ Gumbel-Softmax trick allows to make true categorical distribution and sample from it.
- ▶ Since latent space is discrete we could train autoregressive transformers in it.
- ▶ It is a natural way to incorporate text and image token spaces.

TEXT PROMPT

an armchair in the shape of an avocado [...]

AI-GENERATED IMAGES



# Outline

## 1. Discrete VAE latent representations

Vector quantization

Gumbel-softmax for discrete VAE latents

## 2. ELBO surgery

## 3. Learnable VAE prior

# Outline

## 1. Discrete VAE latent representations

Vector quantization

Gumbel-softmax for discrete VAE latents

## 2. ELBO surgery

## 3. Learnable VAE prior

# ELBO surgery

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) = \frac{1}{n} \sum_{i=1}^n \left[ \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})) \right].$$

## Theorem

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})) = KL(q_{\text{agg}}(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_q[\mathbf{x}, \mathbf{z}];$$

- ▶  $q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i)$  – **aggregated** posterior distribution.
- ▶  $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$  – mutual information between  $\mathbf{x}$  and  $\mathbf{z}$  under empirical data distribution and distribution  $q(\mathbf{z}|\mathbf{x})$ .
- ▶ **First term** pushes  $q_{\text{agg}}(\mathbf{z})$  towards the prior  $p(\mathbf{z})$ .
- ▶ **Second term** reduces the amount of information about  $\mathbf{x}$  stored in  $\mathbf{z}$ .

# ELBO surgery

## Theorem

$$\frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || p(\mathbf{z})) = KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z})) + \mathbb{I}_q[\mathbf{x}, \mathbf{z}].$$

## Proof

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || p(\mathbf{z})) &= \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z})} d\mathbf{z} = \\ &= \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q_{\text{agg}}(\mathbf{z}) q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z}) q_{\text{agg}}(\mathbf{z})} d\mathbf{z} = \int \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i) \log \frac{q_{\text{agg}}(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} + \\ &+ \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q(\mathbf{z}|\mathbf{x}_i)}{q_{\text{agg}}(\mathbf{z})} d\mathbf{z} = KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z})) + \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || q_{\text{agg}}(\mathbf{z})) \end{aligned}$$

Without proof:

$$\mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || q_{\text{agg}}(\mathbf{z})) \in [0, \log n].$$

# ELBO surgery

## ELBO revisiting

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) &= \frac{1}{n} \sum_{i=1}^n [\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z}))] = \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}, \theta)}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x}, \mathbf{z}]}_{\text{MI}} - \underbrace{KL(q_{\text{agg}}(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}} \end{aligned}$$

Prior distribution  $p(\mathbf{z})$  is only in the last term.

## Optimal VAE prior

$$KL(q_{\text{agg}}(\mathbf{z})||p(\mathbf{z})) = 0 \Leftrightarrow p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i).$$

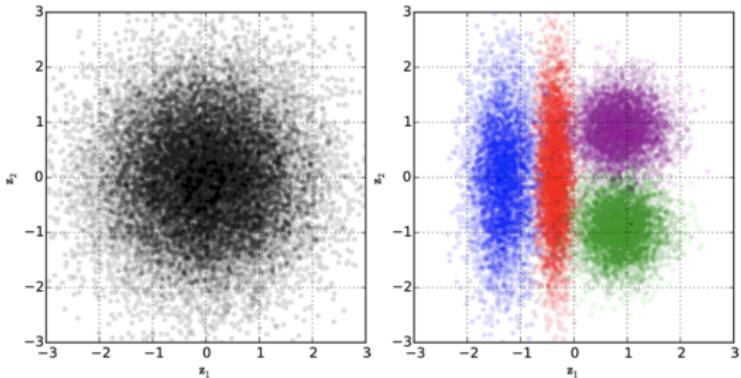
The optimal prior  $p(\mathbf{z})$  is the aggregated posterior  $q_{\text{agg}}(\mathbf{z})$ !

# Variational posterior

## ELBO decomposition

$$\log p(\mathbf{x}|\theta) = \mathcal{L}(q, \theta) + KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\mathbf{x}, \theta)).$$

- ▶  $q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x}))$  is a unimodal distribution.
- ▶ The optimal prior  $p(\mathbf{z})$  is the aggregated posterior  $q_{\text{agg}}(\mathbf{z})$ .



(a) Prior distribution

(b) Posteriors in standard VAE

It is widely believed that **mismatch between  $p(\mathbf{z})$  and  $q_{\text{agg}}(\mathbf{z})$  is the main reason of blurry images of VAE.**

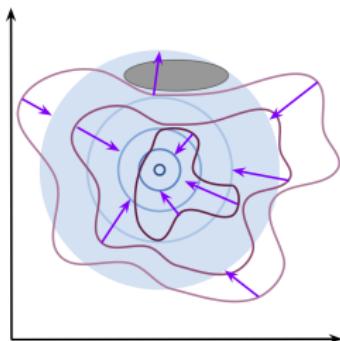
# Outline

1. Discrete VAE latent representations
  - Vector quantization
  - Gumbel-softmax for discrete VAE latents
2. ELBO surgery
3. Learnable VAE prior

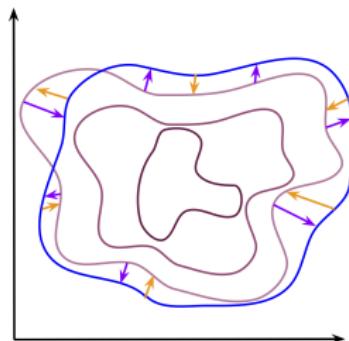
## Optimal VAE prior

- ▶ Standard Gaussian  $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}) \Rightarrow$  over-regularization;
- ▶  $p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i) \Rightarrow$  overfitting and highly expensive.

Non learnable prior  $p(\mathbf{z})$



Learnable prior  $p(\mathbf{z}|\boldsymbol{\lambda})$



ELBO revisiting

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) = \text{RL} - \text{MI} - KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z}|\boldsymbol{\lambda}))$$

It is Forward KL with respect to  $p(\mathbf{z}|\boldsymbol{\lambda})$ .

## NF-based VAE prior

### NF model in latent space

$$\log p(\mathbf{z}|\boldsymbol{\lambda}) = \log p(\mathbf{z}^*) + \log \left| \det \left( \frac{d\mathbf{z}^*}{d\mathbf{z}} \right) \right| = \log p(f(\mathbf{z}, \boldsymbol{\lambda})) + \log |\det(\mathbf{J}_f)|$$
$$\mathbf{z} = \mathbf{g}_{\boldsymbol{\lambda}}(\mathbf{z}^*) = \mathbf{f}_{\boldsymbol{\lambda}}^{-1}(\mathbf{z}^*)$$

- ▶ RealNVP with coupling layers.
- ▶ Autoregressive NF (fast  $\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{z})$ , slow  $\mathbf{g}_{\boldsymbol{\lambda}}(\mathbf{z}^*)$ ).

### ELBO with NF-based VAE prior

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} [\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \log p(\mathbf{z}|\boldsymbol{\lambda}) - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \left[ \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \underbrace{\left( \log p(f_{\boldsymbol{\lambda}}(\mathbf{z})) + \log |\det(\mathbf{J}_f)| \right)}_{\text{NF-based prior}} - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \right]\end{aligned}$$

## Summary

- ▶ The ELBO surgery reveals insights about a prior distribution in VAE. The optimal prior is the aggregated posterior. It is widely believed that mismatch between  $p(\mathbf{z})$  and  $q_{\text{agg}}(\mathbf{z})$  is the main reason of blurry images of VAE.
- ▶ We could use NF-based prior in VAE (even autoregressive).
- ▶ Vector Quantization is the way to create VAE with discrete latent space and deterministic variational posterior.
- ▶ Straight-through gradient ignores quantize operation in backprop.
- ▶ Gumbel-softmax trick relaxes discrete problem to continuous one using Gumbel-max reparametrization trick.