

Deep Generative Models

Lecture 5

Roman Isachenko



2025, Spring

Recap of previous lecture

EM-algorithm

- ▶ E-step

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

- ▶ M-step

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

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 ϕ .

Variational Bayes

- ▶ E-step

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

- ▶ M-step

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

Recap of previous lecture

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

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

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

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

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

Variational assumption

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

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

Recap of previous lecture

Training (EM-algorithm)

- ▶ pick random sample $\mathbf{x}_i, i \sim \text{Uniform}\{1, n\}$ (or batch).
- ▶ compute the objective (using reparametrization trick):

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

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

- ▶ make gradient step using stochastic gradients w.r.t. ϕ and θ via autograd

Inference

- ▶ sample \mathbf{z}^* from the prior distribution $p(\mathbf{z}) (\mathcal{N}(0, \mathbf{I}))$;
- ▶ sample from the decoder $p(\mathbf{x}|\mathbf{z}^*, \theta)$.

Note: you do not need the encoder $q(\mathbf{z}|\mathbf{x}, \phi)$ during the generation.

Recap of previous lecture

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

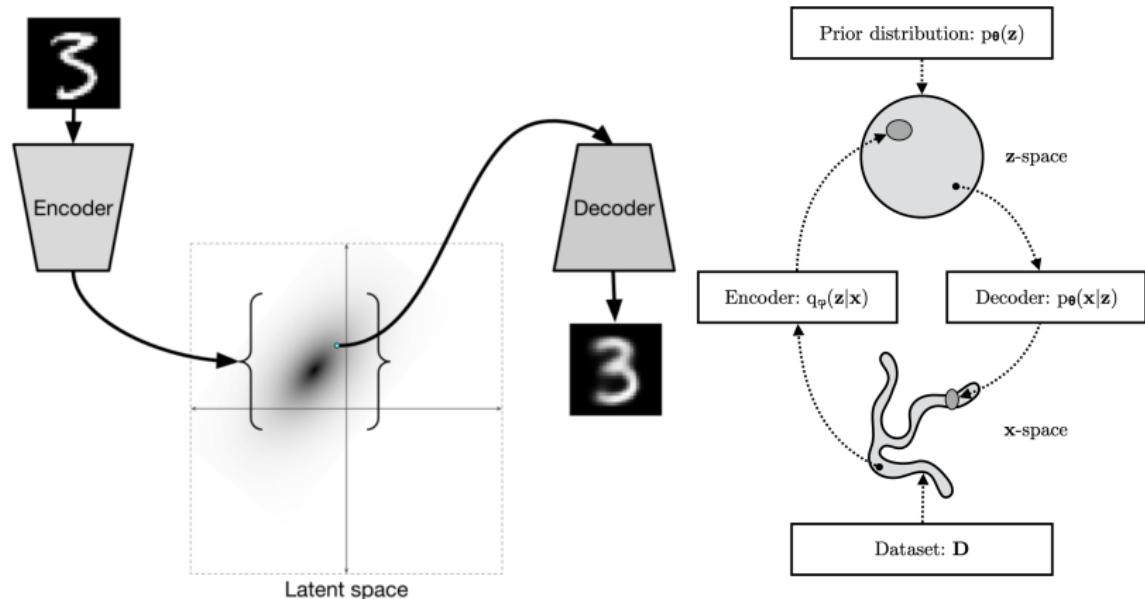


image credit: <http://ijdykeman.github.io/ml/2016/12/21/cvae.html>

Kingma D. P., Welling M. An introduction to variational autoencoders, 2019

Recap of previous lecture

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

Theorem

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

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

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

Recap of previous lecture

Assumptions

- ▶ Let $c \sim \text{Categorical}(\pi)$, where

$$\pi = (\pi_1, \dots, \pi_K), \quad \pi_k = P(c = k), \quad \sum_{k=1}^K \pi_k = 1.$$

- ▶ Let VAE model has discrete latent representation c with prior $p(c) = \text{Uniform}\{1, \dots, K\}$.

ELBO

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

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

- ▶ Our encoder should output discrete distribution $q(c|\mathbf{x}, \phi)$.
- ▶ We need the analogue of the reparametrization trick for the discrete distribution $q(c|\mathbf{x}, \phi)$.
- ▶ Our decoder $p(\mathbf{x}|c, \theta)$ should input discrete random variable c .

Outline

1. Vector quantization: (discrete VAE latent representations)
2. ELBO surgery
3. Learnable VAE prior
4. Likelihood-free learning

Outline

1. Vector quantization: (discrete VAE latent representations)
2. ELBO surgery
3. Learnable VAE prior
4. Likelihood-free learning

Vector quantization

Define the dictionary space $\{\mathbf{e}_k\}_{k=1}^K$, where $\mathbf{e}_k \in \mathbb{R}^L$, K is the size of the dictionary.

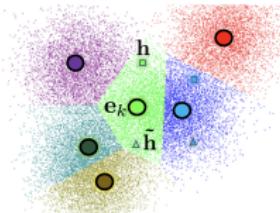
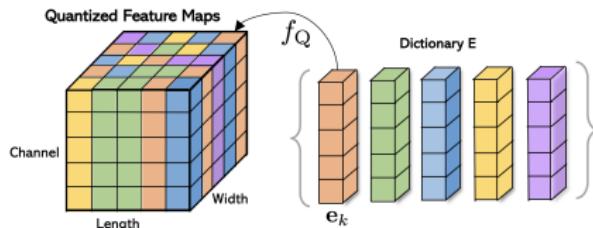
Quantized representation

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

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

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 our encoder outputs continuous representation $\mathbf{z}_e = \text{NN}_{e,\phi}(\mathbf{x}) \in \mathbb{R}^L$.
- ▶ Quantization will give us the deterministic mapping from the encoder output \mathbf{z}_e to its quantized representation \mathbf{z}_q .
- ▶ Let use the dictionary elements \mathbf{e}_c in the decoder distribution $p(\mathbf{x}|\mathbf{e}_c, \theta)$ (instead of $p(\mathbf{x}|c, \theta)$).

Deterministic variational posterior

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

$$KL(q(c|\mathbf{x}, \phi) || p(c)) = - \underbrace{H(q(c|\mathbf{x}, \phi))}_{=0} + \log K = \log K.$$

Note: KL term (regularizer) does not affect the ELBO objective.

Vector Quantized VAE (VQ-VAE): forward

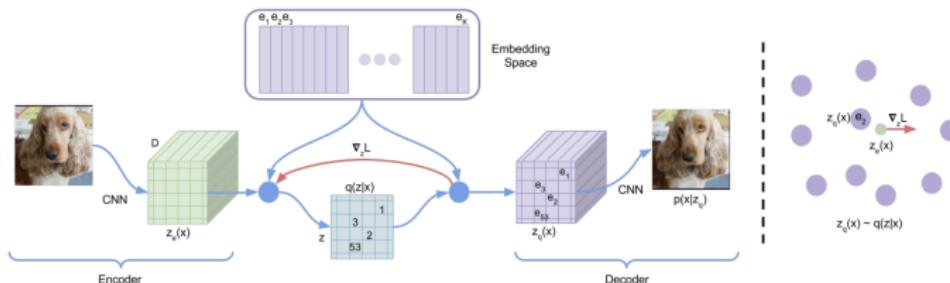
Deterministic variational posterior

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

ELBO

$$\mathcal{L}_{\phi, \theta}(\mathbf{x}) = \mathbb{E}_{q(c|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{e}_c, \theta) - \log K = \log p(\mathbf{x}|\mathbf{z}_q, \theta) - \log K,$$

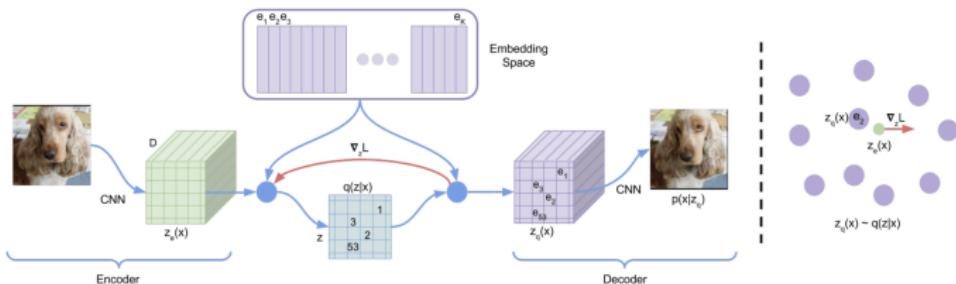
where $\mathbf{z}_q = \mathbf{e}_{k^*}$, $k^* = \arg \min_k \|\mathbf{z}_e - \mathbf{e}_k\|$.



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

Vector Quantized VAE (VQ-VAE): backward ELBO

$$\mathcal{L}_{\phi, \theta}(x) = \log p(x|z_q, \theta) - \log K, \quad z_q = e_{k^*}, k^* = \arg \min_k \|z_e - e_k\|.$$



Straight-through gradient estimation

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

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

Generalization to the spatial dimension: $\mathbf{c} \in \{1, \dots, K\}^{W \times H}$

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

Samples diversity



VQ-VAE (Proposed)

BigGAN deep

Outline

1. Vector quantization: (discrete VAE latent representations)
2. ELBO surgery
3. Learnable VAE prior
4. Likelihood-free learning

ELBO surgery

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

Theorem

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

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

ELBO surgery

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

Proof

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

ELBO surgery

ELBO revisiting

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

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

Optimal VAE prior

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

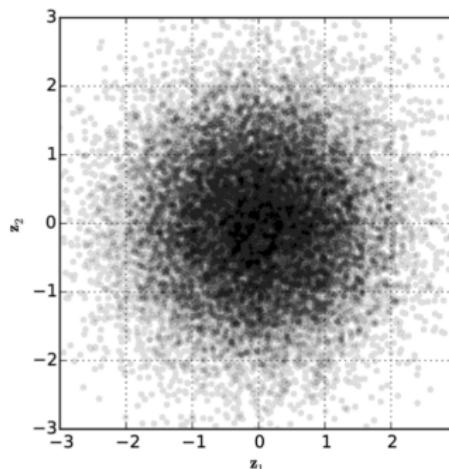
The optimal prior $p(\mathbf{z})$ is the aggregated variational posterior distribution $q_{\text{agg}}(\mathbf{z}|\phi)$!

Variational posterior

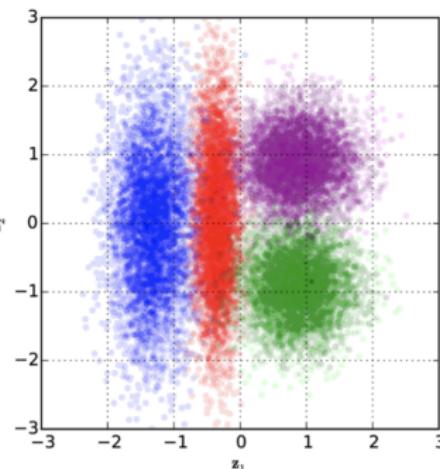
ELBO decomposition

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

- ▶ $q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x}))$ is a unimodal distribution.
- ▶ It is widely believed that **mismatch between $p(\mathbf{z})$ and $q_{\text{agg}}(\mathbf{z}|\phi)$ is the main reason of blurry images of VAE.**



(a) Prior distribution



(b) Posteriors in standard VAE

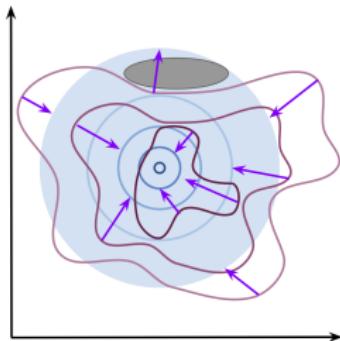
Outline

1. Vector quantization: (discrete VAE latent representations)
2. ELBO surgery
3. Learnable VAE prior
4. Likelihood-free learning

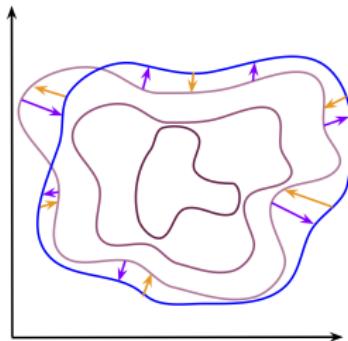
Optimal VAE prior

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

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



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



ELBO revisiting

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

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

NF-based VAE prior

NF model in latent space

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

$$\mathbf{z} = \mathbf{g}_{\boldsymbol{\lambda}}(\mathbf{z}^*) = \mathbf{f}_{\boldsymbol{\lambda}}^{-1}(\mathbf{z}^*)$$

- ▶ 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

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

Summary

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