Deep Generative Models

Lecture 5

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

EM-Algorithm

- E-Step:

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

- M-Step:

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

Amortized Variational Inference

Restrict the family of possible distributions $q(\mathbf{z})$ to a parameterized class $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$, conditioned on samples $\mathbf{x}$ and defined by $\boldsymbol{\phi}$.

Variational Bayes

- E-Step:

$$\boldsymbol{\phi}_k = \boldsymbol{\phi}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}_{k-1}}(\mathbf{x})\big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_{k-1}}$$

- M-Step:

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \eta \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\phi}_k,\boldsymbol{\theta}}(\mathbf{x})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{k-1}}$$

Recap of Previous Lecture

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

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

$$\nabla_\theta \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x},\phi) \nabla_\theta \log p(\mathbf{x}|\mathbf{z},\theta) d\mathbf{z} \approx$$

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

E-Step: $\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x})$, Reparameterization Trick

$$\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x}) = \int p(\boldsymbol{\epsilon}) \nabla_\phi \log p(\mathbf{x}|\mathbf{g}_\phi(\mathbf{x},\boldsymbol{\epsilon}),\theta) d\boldsymbol{\epsilon} - \nabla_\phi \mathrm{KL}$$

$$\approx \nabla_\phi \log p(\mathbf{x}|\mathbf{g}_\phi(\mathbf{x},\boldsymbol{\epsilon}^*),\theta) - \nabla_\phi \mathrm{KL}$$

Variational Assumption

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

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

Recap of Previous Lecture

Training (EM-Algorithm)

- Select a random sample $\mathbf{x}_i$, $i \sim \text{Uniform}\{1, n\}$ (or a minibatch).
- Compute the objective (using the reparameterization trick):

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

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

- Update parameters by taking a stochastic gradient step with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, leveraging autograd.

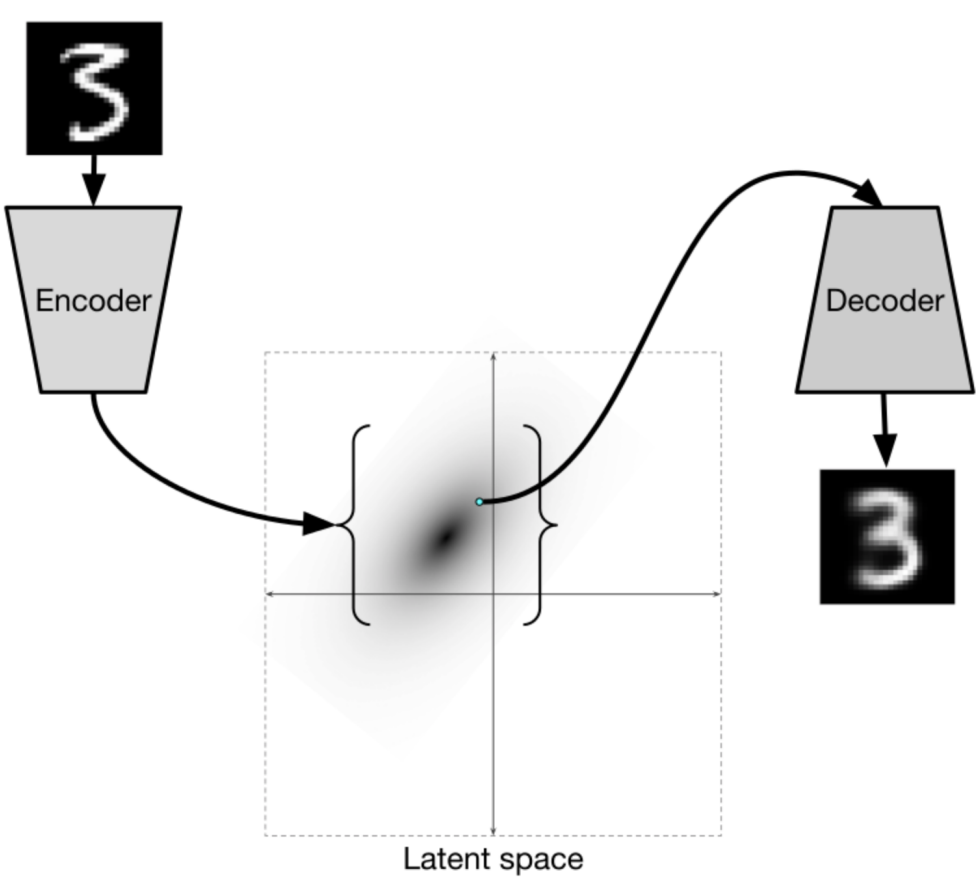
Inference

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

Note: The encoder $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$ isn't needed during generation.

Recap of Previous Lecture

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



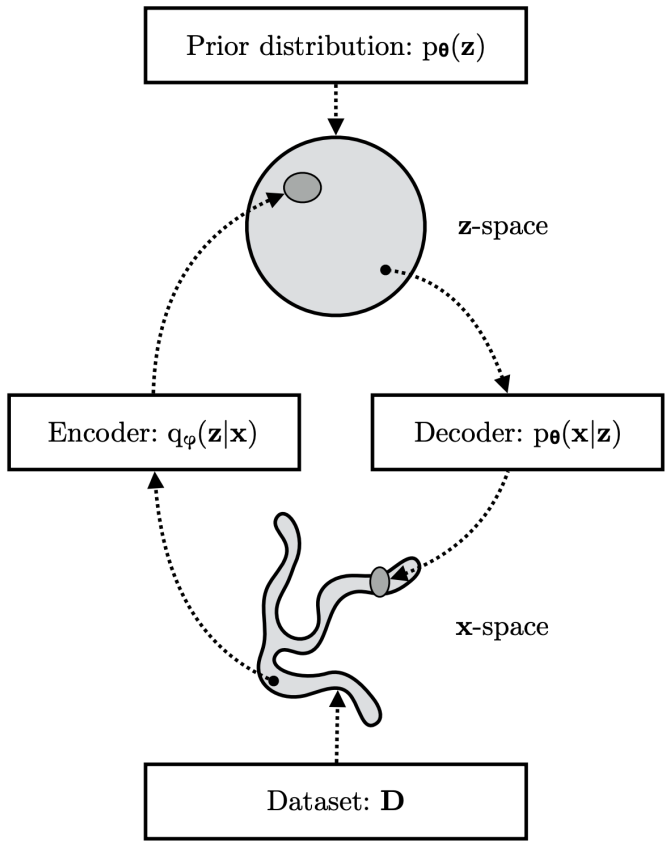


image credit: http://ijdykeman.github.io/ml/2016/12/21/cvae.html
Kingma D. P., Welling M. An Introduction to Variational Autoencoders, 2019

Recap of Previous Lecture

	VAE	**NF**
Objective	ELBO $\mathcal{L}$	Forward KL/MLE
Encoder	Stochastic $\mathbf{z} \sim q(\mathbf{z}\|\mathbf{x}, \boldsymbol{\phi})$	Deterministic $\mathbf{z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ $q(\mathbf{z}\|\mathbf{x}, \boldsymbol{\theta}) = \delta(\mathbf{z} - \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}))$
Decoder	Stochastic $\mathbf{x} \sim p(\mathbf{x}\|\mathbf{z}, \boldsymbol{\theta})$	Deterministic $\mathbf{x} = \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})$ $p(\mathbf{x}\|\mathbf{z}, \boldsymbol{\theta}) = \delta(\mathbf{x} - \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}))$
Parameters	$\boldsymbol{\phi}, \boldsymbol{\theta}$	$\boldsymbol{\theta} \equiv \boldsymbol{\phi}$

Theorem

MLE for normalizing flows is equivalent to maximizing the ELBO for a VAE model with deterministic encoder and decoder:

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \delta(\mathbf{x} - \mathbf{f}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})) = \delta(\mathbf{x} - \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}));$$

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = \delta(\mathbf{z} - \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})).$$

Nielsen D., et al. SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows, 2020

Recap of Previous Lecture

Assumptions

- Let $c \sim \text{Categorical}(\boldsymbol{\pi})$, where
$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K), \quad \pi_k = P(c = k), \quad \sum_{k=1}^{K} \pi_k = 1.$$
- Suppose the VAE includes a discrete latent variable c with prior $p(c) = \text{Uniform}\{1, \ldots, K\}$.

ELBO

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

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

- Our encoder must output the discrete distribution $q(c|\mathbf{x},\phi)$.
- We'll require an analogue of the reparameterization trick for discrete $q(c|\mathbf{x},\phi)$.
- Our decoder $p(\mathbf{x}|c,\boldsymbol{\theta})$ must accept the discrete variable c as input.

Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

Outline

1. Vector Quantization: (Discrete VAE Latent Representations)
2. ELBO Surgery
3. Learnable VAE Prior

Outline

1. Vector Quantization: (Discrete VAE Latent Representations)
2. ELBO Surgery
3. Learnable VAE Prior

Vector Quantization

Define the codebook (dictionary) space $\{\mathbf{e}_k\}_{k=1}^{K}$ with $\mathbf{e}_k \in \mathbb{R}^L$ and K the number of codebook entries.

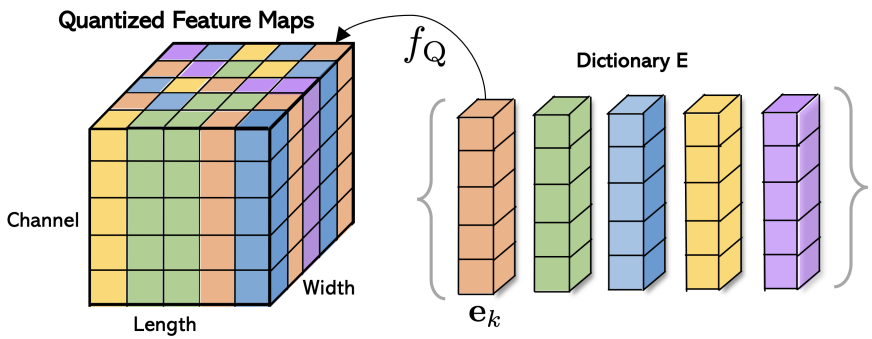
Quantized Representation

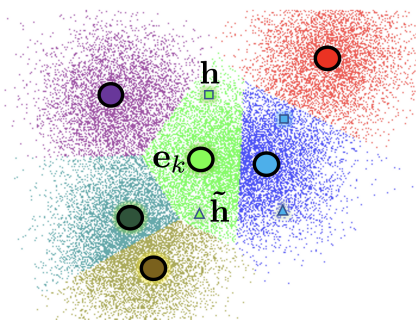
A quantized vector $\mathbf{z}_q \in \mathbb{R}^L$, for any $\mathbf{z} \in \mathbb{R}^L$, is defined via nearest-neighbor lookup in the codebook:

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

Quantization Procedure

If the encoded tensor has spatial dimensions, quantization is independently applied to each of the $W \times H$ locations.





Zhao Y. et al. Feature Quantization Improves GAN Training, 2020

Vector Quantized VAE (VQ-VAE)

- The encoder outputs a continuous vector $\mathbf{z}_e = \mathrm{NN}_{e,\phi}(\mathbf{x}) \in \mathbb{R}^L$.
- Quantization deterministically maps $\mathbf{z}_e$ to its quantized codebook vector $\mathbf{z}_q$.
- The decoder is conditioned on codebook entries $\mathbf{e}_c$, i.e., via $p(\mathbf{x}|\mathbf{e}_c, \boldsymbol{\theta})$ (instead of $p(\mathbf{x}|c, \boldsymbol{\theta})$).

Deterministic Variational Posterior

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

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

Note: The KL regularizer becomes constant and has no direct effect on the ELBO objective in this case.

Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

Vector Quantized VAE (VQ-VAE): Forward

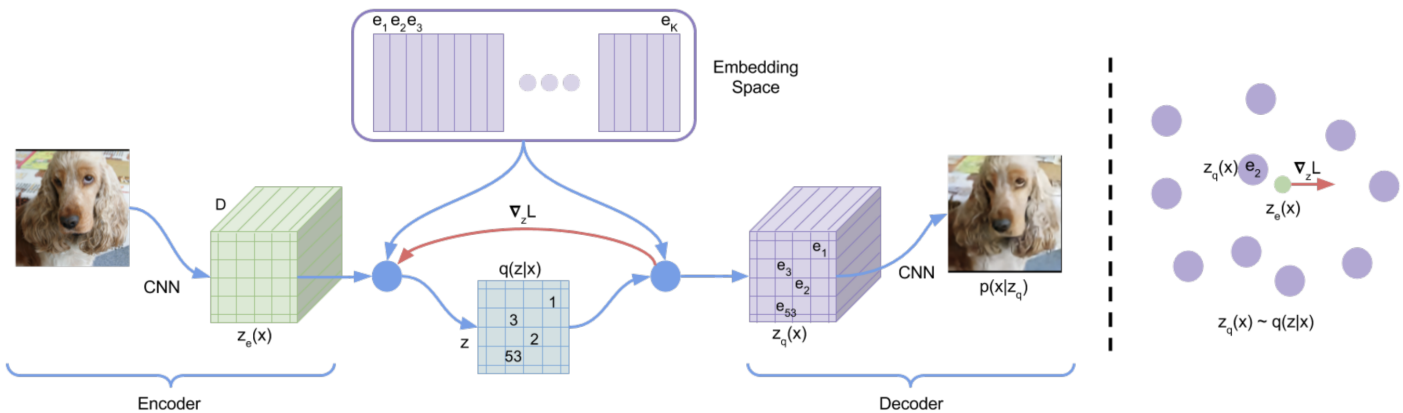
Deterministic Variational Posterior

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

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



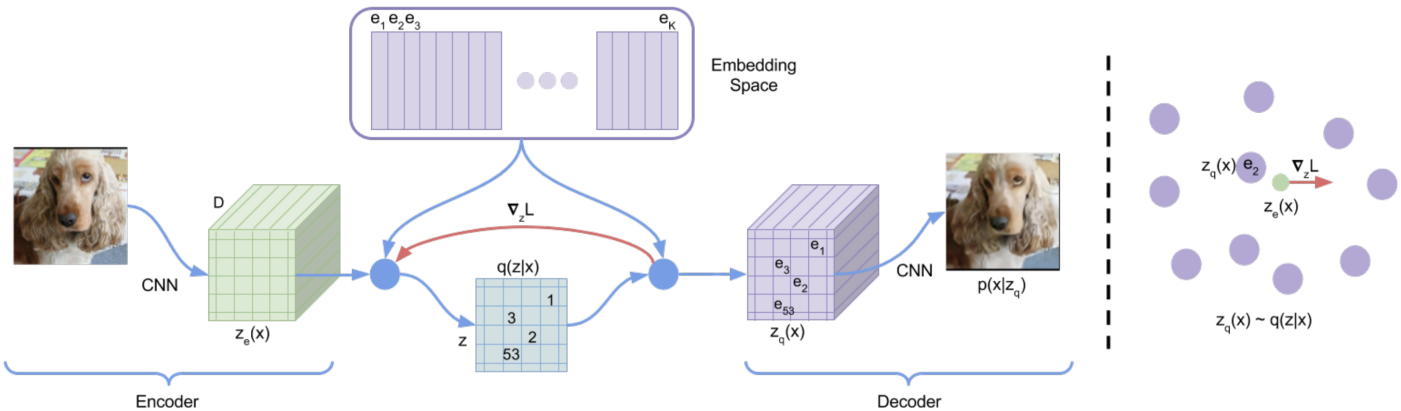
Challenge: The arg min operation is non-differentiable.

Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

Vector Quantized VAE (VQ-VAE): Backward

ELBO

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}) = \log p(\mathbf{x}|\mathbf{z}_q, \theta) - \log K, \quad \mathbf{z}_q = \mathbf{e}_{k^*}, \; k^* = \arg\min_k \|\mathbf{z}_e - \mathbf{e}_k\|.$$



Straight-Through Gradient Estimator

$$\frac{\partial \log p(\mathbf{x}|\mathbf{z}_q, \theta)}{\partial \phi} = \frac{\partial \log p(\mathbf{x}|\mathbf{z}_q, \theta)}{\partial \mathbf{z}_q} \cdot \frac{\partial \mathbf{z}_q}{\partial \phi} =$$

$$= \frac{\partial \log p(\mathbf{x}|\mathbf{z}_q, \theta)}{\partial \mathbf{z}_q} \cdot \frac{\partial \mathbf{z}_q}{\partial \mathbf{z}_e} \cdot \frac{\partial \mathbf{z}_e}{\partial \phi} \approx \frac{\partial \log p(\mathbf{x}|\mathbf{z}_q, \theta)}{\partial \mathbf{z}_q} \cdot \frac{\partial \mathbf{z}_e}{\partial \phi}$$

Oord A., Vinyals O., Kavukcuoglu K. Neural Discrete Representation Learning, 2017

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

Extension to the spatial domain: $\mathbf{c} \in \{1, \ldots, K\}^{W \times H}$

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

Sample Diversity

VQ-VAE (Proposed) **BigGAN deep**

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

Outline

1. Vector Quantization: (Discrete VAE Latent Representations)
2. ELBO Surgery
3. Learnable VAE Prior

ELBO Surgery

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

Theorem

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

- $q_{\text{agg}}(\mathbf{z}|\phi) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i,\phi)$ denotes the **aggregated** variational posterior.
- $\mathbb{I}_q[\mathbf{x},\mathbf{z}]$ is the mutual information between $\mathbf{x}$ and $\mathbf{z}$ under the data distribution $\pi(\mathbf{x})$ and $q(\mathbf{z}|\mathbf{x},\phi)$.
- The first term encourages $q_{\text{agg}}(\mathbf{z}|\phi)$ to match the prior $p(\mathbf{z})$.
- The second term reduces the information about $\mathbf{x}$ encoded in $\mathbf{z}$.

Hoffman M. D., Johnson M. J. ELBO Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound, 2016

ELBO Surgery

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

Proof

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}(q(\mathbf{z}|\mathbf{x}_i,\phi)\|p(\mathbf{z})) = \frac{1}{n}\sum_{i=1}^{n}\int q(\mathbf{z}|\mathbf{x}_i,\phi)\log\frac{q(\mathbf{z}|\mathbf{x}_i,\phi)}{p(\mathbf{z})}d\mathbf{z} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int q(\mathbf{z}|\mathbf{x}_i,\phi)\log\frac{q_{\text{agg}}(\mathbf{z}|\phi)q(\mathbf{z}|\mathbf{x}_i,\phi)}{p(\mathbf{z})q_{\text{agg}}(\mathbf{z}|\phi)}d\mathbf{z} =$$

$$= \int\frac{1}{n}\sum_{i=1}^{n}q(\mathbf{z}|\mathbf{x}_i,\phi)\log\frac{q_{\text{agg}}(\mathbf{z}|\phi)}{p(\mathbf{z})}d\mathbf{z} + \frac{1}{n}\sum_{i=1}^{n}\int q(\mathbf{z}|\mathbf{x}_i,\phi)\log\frac{q(\mathbf{z}|\mathbf{x}_i,\phi)}{q_{\text{agg}}(\mathbf{z}|\phi)}d\mathbf{z} =$$

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

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

Hoffman M. D., Johnson M. J. ELBO Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound, 2016

ELBO Surgery

Revisiting the ELBO

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

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i,\phi)}\log p(\mathbf{x}_i|\mathbf{z},\theta)}_{\text{Reconstruction Loss}} - \underbrace{\mathbb{I}_q[\mathbf{x},\mathbf{z}]}_{\text{Mutual Information}} \underbrace{- \mathrm{KL}(q_{\text{agg}}(\mathbf{z}|\phi)\|p(\mathbf{z}))}_{\text{Marginal KL}}$$

The prior distribution $p(\mathbf{z})$ only appears in the last term.

Optimal VAE Prior

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

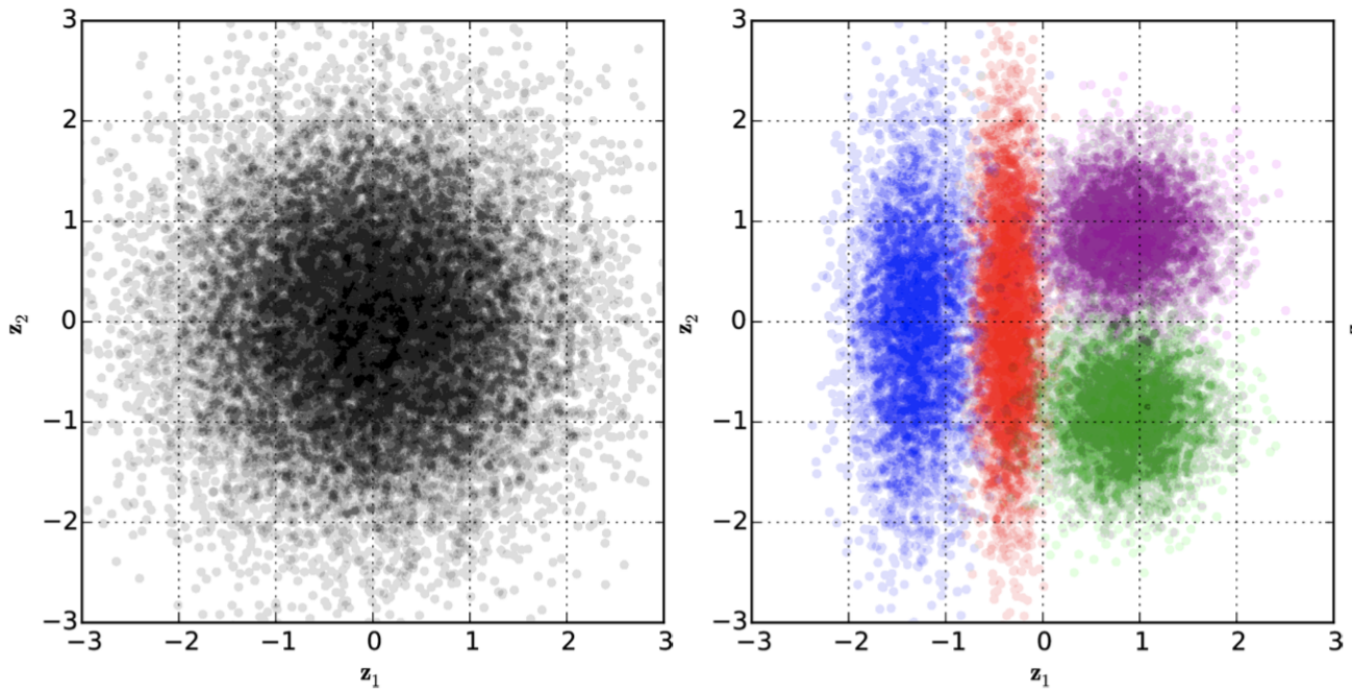
Hence, the optimal prior $p(\mathbf{z})$ is the aggregated variational posterior $q_{\text{agg}}(\mathbf{z}|\phi)$.

Hoffman M. D., Johnson M. J. ELBO Surgery: Yet Another Way to Carve Up the Variational Evidence Lower Bound, 2016

Marginal KL

$$\mathrm{KL}(q_{\text{agg}}(\mathbf{z}|\phi)\|p(\mathbf{z}))$$

- $q(\mathbf{z}|\mathbf{x},\phi) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x}))$ is unimodal.
- It is generally believed that the **mismatch between** $p(\mathbf{z})$ **and** $q_{\text{agg}}(\mathbf{z}|\phi)$ is the primary explanation for blurry VAE-generated images.



(a) Prior distribution (b) Posteriors in standard VAE

Rezende D. J., Mohamed S. Variational Inference with Normalizing Flows, 2015

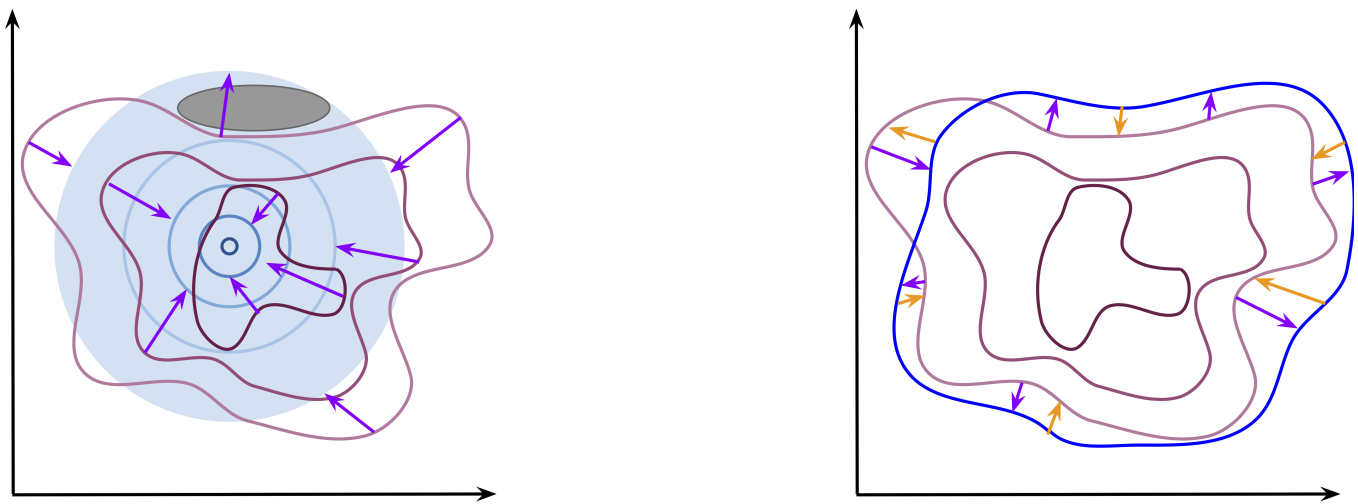
Outline

1. Vector Quantization: (Discrete VAE Latent Representations)
2. ELBO Surgery
3. Learnable VAE Prior

Optimal VAE Prior

- Standard Gaussian $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$ often leads to over-regularization.
- $p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}|\phi) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i, \phi)$ risks overfitting and incurs high computational cost.

Non-Learnable Prior $p(\mathbf{z})$ Learnable Prior $p(\mathbf{z}|\boldsymbol{\lambda})$



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

This is the forward KL divergence with respect to $p(\mathbf{z}|\boldsymbol{\lambda})$.

image credit: https://jmtomczak.github.io/blog/7/7_priors.html

NF-Based VAE Prior

NF Model in Latent Space

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

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

- For example, RealNVP with coupling layers,
- Autoregressive normalizing flows (efficient $\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{z})$, but $\mathbf{g}_{\boldsymbol{\lambda}}(\mathbf{z}^*)$ can be slow).

ELBO with NF-Based VAE Prior

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

$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\phi)} \left[\log p(\mathbf{x}|\mathbf{z},\theta) + \log p(\mathbf{z}|\boldsymbol{\lambda}) - \log q(\mathbf{z}|\mathbf{x},\phi)\right]$$

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

Chen X. et al. Variational Lossy Autoencoder, 2016

Summary

- Vector quantization provides a way to construct VAEs with discrete latent variables and deterministic variational posteriors.
- The straight-through gradient estimator allows gradients to pass as if quantization were an identity operation during backpropagation.
- ELBO surgery gives insights into the prior's influence in VAEs; the optimal prior is the aggregated variational posterior.
- The mismatch between $p(\mathbf{z})$ and $q_{\text{agg}}(\mathbf{z}|\phi)$ is widely regarded as the principal reason for VAE-generated image blurriness.
- Normalizing flow-based priors, including autoregressive flows, can be incorporated directly into VAEs.