Deep Generative Models

Lecture 4

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of previous lecture

Forward KL for NF

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_\mathbf{f})|$$

Reverse KL for NF

$$KL(p||\pi) = \mathbb{E}_{p(\mathbf{z})} \left[\log p(\mathbf{z}) - \log |\det(\mathbf{J}_\mathbf{g})| - \log \pi(\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}))\right]$$

Flow KL Duality

$$\arg\min_{\boldsymbol{\theta}} KL(\pi(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\theta})) = \arg\min_{\boldsymbol{\theta}} KL(p(\mathbf{z}|\boldsymbol{\theta})||p(\mathbf{z}))$$

- $p(\mathbf{z})$ is the base distribution; $\pi(\mathbf{x})$ is the data distribution;
- $\mathbf{z} \sim p(\mathbf{z})$, $\mathbf{x} = \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})$, so $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta})$;
- $\mathbf{x} \sim \pi(\mathbf{x})$, $\mathbf{z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$, so $\mathbf{z} \sim p(\mathbf{z}|\boldsymbol{\theta})$.

Papamakarios G. et al. Normalizing flows for probabilistic modeling and inference, 2019

Recap of previous lecture

Posterior Distribution (Bayes Theorem)

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

- $\mathbf{x}$ – observed variables;
- $\boldsymbol{\theta}$ – unobserved variables (latent variables/parameters);
- $p(\mathbf{x}|\boldsymbol{\theta})$ – likelihood;
- $p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ – evidence;
- $p(\boldsymbol{\theta})$ – prior distribution;
- $p(\boldsymbol{\theta}|\mathbf{x})$ – posterior distribution.

Recap of previous lecture

Latent Variable Models (LVM)

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}.$$

MLE Problem for LVM

$$\begin{aligned}\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{X}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}) = \\ = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i) d\mathbf{z}_i.\end{aligned}$$

Naive Monte-Carlo Estimation

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z} = \mathbb{E}_{p(\mathbf{z})} p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) \approx \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}_k, \boldsymbol{\theta}),$$

where $\mathbf{z}_k \sim p(\mathbf{z})$.

Recap of previous lecture

ELBO derivation 1 (inequality)

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \; = \; \log \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} \; \geq \; \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \; = \; \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

ELBO derivation 2 (equality)

$$\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z} = \int q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z} =$$
$$= \log p(\mathbf{x}|\boldsymbol{\theta}) - KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))$$

Variational decomposition

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

Recap of previous lecture

Variational lower Bound (ELBO)

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

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

Log-likelihood decomposition

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \mathbb{E}_q \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z})||p(\mathbf{z})) + KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})).$$

- Instead of maximizing the likelihood, maximize the ELBO:

$$\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) \quad \rightarrow \quad \max_{q, \boldsymbol{\theta}} \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})$$

- Maximizing the ELBO with respect to the variational distribution q is equivalent to minimizing the KL divergence:

$$\arg\max_q \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) \equiv \arg\min_q KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})).$$

Recap of previous lecture

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

EM-algorithm (block-coordinate optimization)

- Initialize $\boldsymbol{\theta}^*$;
- **E-step:** $(\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) \rightarrow \max_q)$

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

- **M-step:** $(\mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x}) \rightarrow \max_{\boldsymbol{\theta}})$

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

- Repeat E-step and M-step until convergence.

Outline

1. EM-algorithm
   Amortized Inference
   ELBO Gradients, Reparametrization Trick
2. Variational Autoencoder (VAE)
3. Discrete VAE Latent Representations

Outline

1. EM-algorithm
   Amortized Inference
   ELBO Gradients, Reparametrization Trick

2. Variational Autoencoder (VAE)

3. Discrete VAE Latent Representations

Outline

1. EM-algorithm
   - Amortized Inference
   - ELBO Gradients, Reparametrization Trick
2. Variational Autoencoder (VAE)
3. Discrete VAE Latent Representations

Amortized Variational Inference

E-step

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

$q(\mathbf{z})$ approximates the true posterior distribution $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^*)$, which is why it is called the **variational posterior**.

- $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^*)$ may be **intractable**;
- $q(\mathbf{z})$ is unique for each data point $\mathbf{x}$.

Variational Bayes

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

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

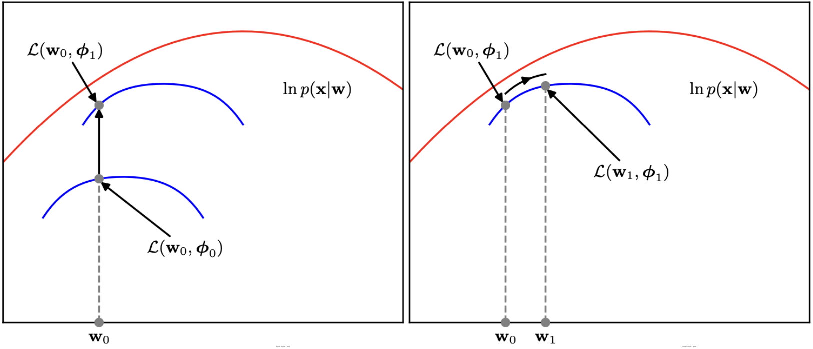
Variational EM Illustration

- 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:

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



Bishop C. Deep Learning: Foundations and Concepts, 2024

Variational EM-algorithm

ELBO

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

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

- **E-step:**

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

  where ϕ are the parameters of the variational posterior distribution $q(\mathbf{z}|\mathbf{x},\phi)$.
- **M-step:**

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

  where $\boldsymbol{\theta}$ are the parameters of the generative distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$.

All that remains is to obtain **unbiased** Monte Carlo estimates of the gradients: $\nabla_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x})$ and $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{x})$.

Outline

1. EM-algorithm
   - Amortized Inference
   - ELBO Gradients, Reparametrization Trick
2. Variational Autoencoder (VAE)
3. Discrete VAE Latent Representations

ELBO Gradients, (M-step, $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\phi,\boldsymbol{\theta}}(\mathbf{x})$)

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

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

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

Naive Monte-Carlo estimation

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z} \approx \frac{1}{K}\sum_{k=1}^{K} p(\mathbf{x}|\mathbf{z}_k,\boldsymbol{\theta}), \quad \mathbf{z}_k \sim p(\mathbf{z}).$$

The variational posterior $q(\mathbf{z}|\mathbf{x},\phi)$ usually concentrates more probability mass in a smaller region than the prior $p(\mathbf{z})$.

---

image credit: https://jmtomczak.github.io/blog/4/4_VAE.html

ELBO Gradients, (E-step, $\nabla_\phi \mathcal{L}_{\phi,\boldsymbol{\theta}}(\mathbf{x})$)

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

Unlike in the M-step, the density $q(\mathbf{z}|\mathbf{x}, \phi)$ now depends on ϕ, so standard Monte Carlo estimation cannot be applied:

$$
\begin{aligned}
\nabla_\phi \mathcal{L}_{\phi,\boldsymbol{\theta}}(\mathbf{x}) &= \nabla_\phi \int q(\mathbf{z}|\mathbf{x}, \phi) \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} - \nabla_\phi KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z})) \\
&\neq \int q(\mathbf{z}|\mathbf{x}, \phi) \nabla_\phi \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) d\mathbf{z} - \nabla_\phi KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}))
\end{aligned}
$$

Reparametrization trick (LOTUS trick)

Suppose $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}, \phi)$ is generated by a random variable $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ through the deterministic mapping $\mathbf{z} = \mathbf{g}_\phi(\mathbf{x}, \boldsymbol{\epsilon})$. Then

$$
\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}, \phi)} \mathbf{f}(\mathbf{z}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})} \mathbf{f}(\mathbf{g}_\phi(\mathbf{x}, \boldsymbol{\epsilon}))
$$

Note: The LHS expectation is with respect to the parametric distribution $q(\mathbf{z}|\mathbf{x}, \phi)$, while the RHS uses the non-parametric distribution $p(\boldsymbol{\epsilon})$.

ELBO Gradients, (E-step, $\nabla_{\phi}\mathcal{L}_{\phi,\theta}(\mathbf{x})$)

Reparametrization trick (LOTUS trick)

$$\nabla_{\phi}\int q(\mathbf{z}|\mathbf{x},\phi)\mathbf{f}(\mathbf{z})d\mathbf{z} = \nabla_{\phi}\int p(\boldsymbol{\epsilon})\mathbf{f}(\mathbf{g}_{\phi}(\mathbf{x},\boldsymbol{\epsilon}))d\boldsymbol{\epsilon}$$
$$= \int p(\boldsymbol{\epsilon})\nabla_{\phi}\mathbf{f}(\mathbf{g}_{\phi}(\mathbf{x},\boldsymbol{\epsilon}))d\boldsymbol{\epsilon} \approx \nabla_{\phi}\mathbf{f}(\mathbf{g}_{\phi}(\mathbf{x},\boldsymbol{\epsilon}^*)),$$

where $\boldsymbol{\epsilon}^* \sim p(\boldsymbol{\epsilon})$.

Variational assumption

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

Here $\boldsymbol{\mu}_{\phi}(\cdot)$ and $\boldsymbol{\sigma}_{\phi}(\cdot)$ are parameterized functions (outputs of a neural network).
Thus, we can write $q(\mathbf{z}|\mathbf{x},\phi) = \mathrm{NN}_e(\mathbf{x},\phi)$, known as the **encoder**.

ELBO Gradient (E-step, $\nabla_\phi \mathcal{L}_{\phi,\theta}(\mathbf{x})$)

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

Reconstruction term

$$\nabla_\phi \int q(\mathbf{z}|\mathbf{x},\phi) \log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) d\mathbf{z} = \int p(\boldsymbol{\epsilon}) \nabla_\phi \log p(\mathbf{x}|\mathbf{g}_\phi(\mathbf{x},\boldsymbol{\epsilon}),\boldsymbol{\theta}) d\boldsymbol{\epsilon} \approx$$
$$\approx \nabla_\phi \log p\left(\mathbf{x}|\boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}^* + \boldsymbol{\mu}_\phi(\mathbf{x}),\boldsymbol{\theta}\right), \quad \text{where } \boldsymbol{\epsilon}^* \sim \mathcal{N}(0,\mathbf{I})$$

The generative distribution $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta})$ can be a neural network.
We write $p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) = \mathrm{NN}_d(\mathbf{z},\boldsymbol{\theta})$, called the **decoder**.

KL term

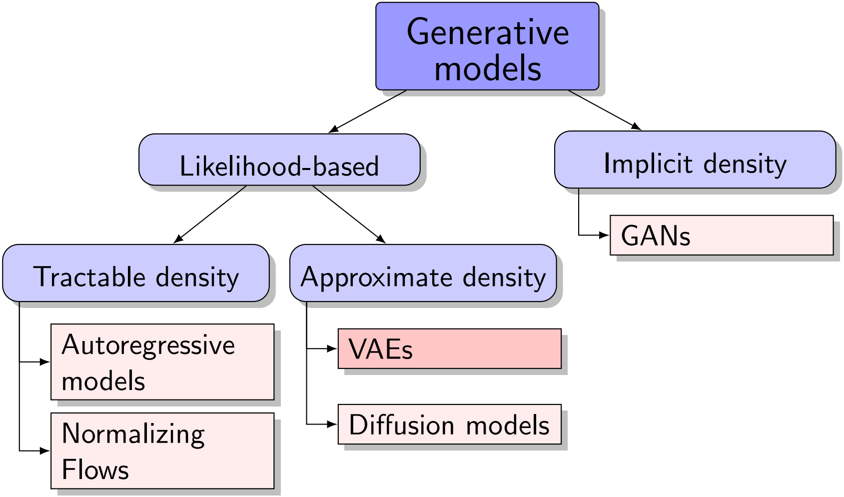
$p(\mathbf{z})$ is the prior distribution on the latent variables $\mathbf{z}$, often taken as $p(\mathbf{z}) = \mathcal{N}(0,\mathbf{I})$.

$$\nabla_\phi KL(q(\mathbf{z}|\mathbf{x},\phi)||p(\mathbf{z})) = \nabla_\phi KL\left(\mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}),\boldsymbol{\sigma}^2_\phi(\mathbf{x}))||\mathcal{N}(0,\mathbf{I})\right)$$

This expression has a closed-form analytical solution.

Outline

Generative Models Zoo



Variational Autoencoder (VAE)

Training (EM-algorithm)

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

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

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

- Update parameters via gradient steps with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ using stochastic gradients (as in autograd).

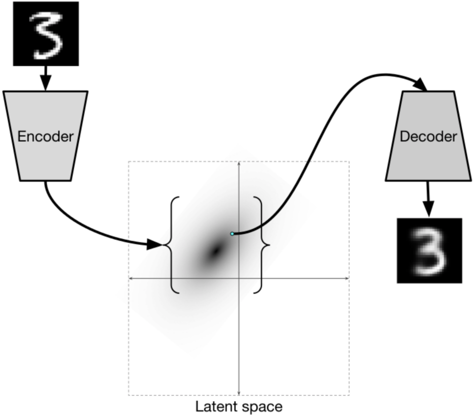
Inference

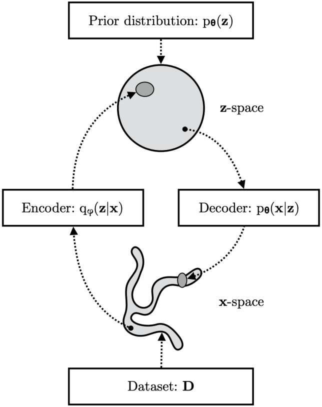
- Sample $\mathbf{z}^*$ from the prior $p(\mathbf{z})$ ($\mathcal{N}(0, \mathbf{I})$);
- Generate data from the decoder $p(\mathbf{x}|\mathbf{z}^*, \boldsymbol{\theta})$.

Note: The encoder $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$ is not required during generation.

Variational Autoencoder

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



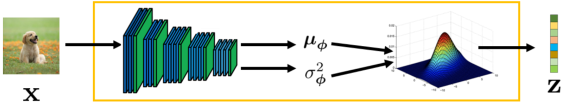


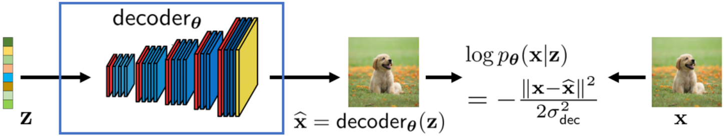
---

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

Variational Autoencoder

- The encoder $q(\mathbf{z}|\mathbf{x}, \phi) = \mathrm{NN}_e(\mathbf{x}, \phi)$ outputs $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}_\phi(\mathbf{x})$.
- The decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathrm{NN}_d(\mathbf{z}, \boldsymbol{\theta})$ outputs parameters of the data distribution.





Chan S. Tutorial on Diffusion Models for Imaging and Vision, 2024

VAE vs Normalizing Flows

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

Theorem

MLE for a normalizing flow is equivalent to maximizing the ELBO for a VAE model with a:

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

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

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

Outline

Discrete VAE Latents

Motivation

- Previous VAE models used **continuous** latent variables $\mathbf{z}$.
- **Discrete** representations $\mathbf{z}$ can be a more natural fit for certain modalities.
- Powerful autoregressive models (such as PixelCNN) are effective for modeling distributions over discrete variables.
- Modern transformer-like models process discrete tokens.

ELBO

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

- Use the reparametrization trick to obtain unbiased gradients.
- Adopt normal distributions for $q(\mathbf{z}|\mathbf{x},\phi)$ and $p(\mathbf{z})$ to enable analytical computation of the KL term.

Discrete VAE Latents

Assumptions

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

ELBO

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

$$KL(q(c|\mathbf{x}, \phi)||p(c)) = \sum_{k=1}^{K} q(k|\mathbf{x}, \phi) \log \frac{q(k|\mathbf{x}, \phi)}{p(k)} =$$

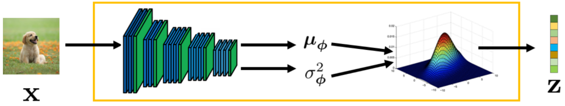
$$= \sum_{k=1}^{K} q(k|\mathbf{x}, \phi) \log q(k|\mathbf{x}, \phi) - \sum_{k=1}^{K} q(k|\mathbf{x}, \phi) \log p(k) =$$

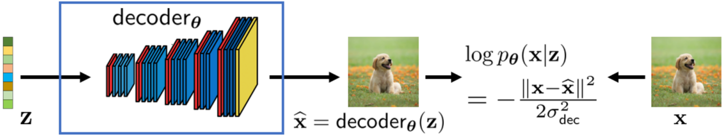
$$= -H(q(c|\mathbf{x}, \phi)) + \log K.$$

Discrete VAE Latents

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

- The encoder should output a discrete distribution $q(c|\mathbf{x},\phi)$.
- We need an analogue of the reparametrization trick for discrete $q(c|\mathbf{x},\phi)$.
- The decoder $p(\mathbf{x}|c,\boldsymbol{\theta})$ must take a discrete random variable c as input.





Chan S. Tutorial on Diffusion Models for Imaging and Vision, 2024

Summary

- Amortized variational inference enables efficient for the ELBO via Monte Carlo estimation.
- The reparametrization trick yields unbiased gradients with respect to the variational posterior $q(\mathbf{z}|\mathbf{x}, \phi)$.
- The VAE model is a latent variable model with two neural networks: a stochastic encoder $q(\mathbf{z}|\mathbf{x}, \phi)$ and a stochastic decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$.
- NF models can be viewed as VAE models with deterministic encoder and decoder.
- Discrete VAE representations constitute a natural form of latent variables.