Deep Generative Models

Lecture 4

Roman Isachenko

AI Masters

Autumn, 2022

Recap of previous lecture

Variational lower Bound (ELBO)

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

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

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 incomplete likelihood, maximize ELBO

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

▶ Maximization of ELBO by variational distribution q is equivalent to minimization of KL

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

Recap of previous lecture

EM-algorithm

▶ E-step

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

▶ M-step

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

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 ϕ.

Variational Bayes

▶ E-step

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

▶ M-step

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

Recap of previous lecture

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

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

$$\nabla_\theta \mathcal{L}(\phi, \theta) = \int q(z|x, \phi) \nabla_\theta \log p(x|z, \theta) dz \approx$$
$$\approx \nabla_\theta \log p(x|z^*, \theta), \quad z^* \sim q(z|x, \phi).$$

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

$$\nabla_\phi \mathcal{L}(\phi, \theta) = \int r(\epsilon) \nabla_\phi \log p(x|g(x, \epsilon, \phi), \theta) d\epsilon - \nabla_\phi \mathsf{KL}$$
$$\approx \nabla_\phi \log p(x|g(x, \epsilon^*, \phi), \theta) - \nabla_\phi \mathsf{KL}$$

Variational assumption

$$r(\epsilon) = \mathcal{N}(0, \mathbf{I}); \quad q(z|x, \phi) = \mathcal{N}(\mu_\phi(x), \sigma_\phi^2(x)).$$
$$z = g(x, \epsilon, \phi) = \sigma_\phi(x) \cdot \epsilon + \mu_\phi(x).$$

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 $p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}) = p(\mathbf{z}) p(\mathbf{x} | \mathbf{z}, \boldsymbol{\theta})$, with a prior distribution $p(\mathbf{z})$, and a stochastic decoder $p(\mathbf{x} | \mathbf{z}, \boldsymbol{\theta})$.

▶ The stochastic encoder $q(\mathbf{z} | \mathbf{x}, \phi)$ (inference model), approximates the true but intractable posterior $p(\mathbf{z} | \mathbf{x}, \boldsymbol{\theta})$.

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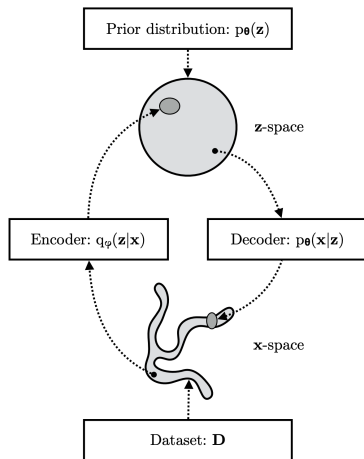
Outline

1. VAE limitations

2. Posterior collapse and decoder weakening techniques

3. Tighter variational bound

4. Normalizing flows prerequisities

Outline

VAE limitations

▶ Poor generative distribution (decoder)

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu_\theta}(\mathbf{z}), \boldsymbol{\sigma_\theta^2}(\mathbf{z})) \quad \text{or} \;\; = \text{Softmax}(\boldsymbol{\pi_\theta}(\mathbf{z})).$$

▶ Loose lower bound

$$\log p(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = (?).$$

▶ Poor prior distribution

$$p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}).$$

▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu_\phi}(\mathbf{x}), \boldsymbol{\sigma_\phi^2}(\mathbf{x})).$$

Outline

VAE limitations

▶ **Poor generative distribution (decoder)**

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}_{\boldsymbol{\theta}}^2(\mathbf{z})) \quad \text{or} \ = \text{Softmax}(\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{z})).$$

▶ Loose lower bound

$$\log p(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = (?).$$

▶ Poor prior distribution

$$p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}).$$

▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}_{\boldsymbol{\phi}}^2(\mathbf{x})).$$

Posterior collapse

LVM

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

ELBO objective

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

More powerful $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ leads to more powerful generative model $p(\mathbf{x}|\boldsymbol{\theta})$.

Extreme cast

$$p(\mathbf{x}|\boldsymbol{\theta}) \in \mathcal{P} = \{p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})| \, \forall \mathbf{z}, \boldsymbol{\theta}\}.$$

If the decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ is too powerful (it could model $p(\mathbf{x}|\boldsymbol{\theta})$), then ELBO avoids paying any cost $KL(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})||p(\mathbf{z}))$ ($q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \approx p(\mathbf{z})$), the variational posterior $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$ will not carry any information about $\mathbf{x}$, the latent variables $\mathbf{z}$ becomes irrelevant.

Autoregressive VAE decoder

How to make the generative model $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ more powerful?

PixelVAE/VLAE

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{j=1}^{m} p(x_j|\mathbf{x}_{1:j-1}, \mathbf{z}, \boldsymbol{\theta})$$

▶ Global structure is captured by latent variables $\mathbf{z}$.

▶ Local statistics are captured by limited receptive field of autoregressive context $\mathbf{x}_{1:j-1}$.

PixelVAE/VLAE models use the autoregressive PixelCNN decoder model with small number of layers to limit receptive field.

https://arxiv.org/abs/1611.05013Gulrajani I. et al. PixelVAE: A Latent Variable Model for Natural Images, 2016,
https://arxiv.org/abs/1611.02731Chen X. et al. Variational Lossy Autoencoder, 2016

Decoder weakening techniques

How to force the model encode information about **x** into **z**?

KL annealing

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

Start training with $\beta = 0$, increase it until $\beta = 1$ during training.

Free bits

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

It ensures the use of less than λ bits of information and results in $KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z})) \geq \lambda$.

Bowman S. R. et al. *Generating Sentences from a Continuous Space*, 2015
Kingma D. P. et al. *Improving Variational Inference with Inverse Autoregressive Flow*, 2016

Outline

VAE limitations

▶ Poor generative distribution (decoder)

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu_\theta}(\mathbf{z}), \boldsymbol{\sigma_\theta^2}(\mathbf{z})) \quad \text{or } = \text{Softmax}(\boldsymbol{\pi_\theta}(\mathbf{z})).$$

▶ **Loose lower bound**

$$\log p(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = (?).$$

▶ Poor prior distribution

$$p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}).$$

▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu_\phi}(\mathbf{x}), \boldsymbol{\sigma_\phi^2}(\mathbf{x})).$$

Importance sampling

LVM

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z} = \int \left[\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})}\right] q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})d\mathbf{z}$$

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

Here $f(\mathbf{x}, \mathbf{z}) = \frac{p(\mathbf{x},\mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})}$.

ELBO: derivation 1

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

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

$f(\mathbf{x}, \mathbf{z})$ could be any function that satisfies $p(\mathbf{x}|\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim q}f(\mathbf{x}, \mathbf{z})$.
Could we choose better $f(\mathbf{x}, \mathbf{z})$?

Importance Weighted Autoencoders (IWAE)

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int \left[\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \right] q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} f(\mathbf{x}, \mathbf{z})$$

Let define

$$f(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_K) = \frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi})}$$

$$\mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} f(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_K) = p(\mathbf{x}|\boldsymbol{\theta})$$

ELBO

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} f(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_K) \geq$$

$$\geq \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log f(\mathbf{x}, \mathbf{z}_1, \ldots, \mathbf{z}_K) =$$

$$= \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log \left[\frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi})} \right] = \mathcal{L}_K(q, \boldsymbol{\theta}).$$

Burda Y., Grosse R., Salakhutdinov R. Importance Weighted Autoencoders, 2015

Importance Weighted Autoencoders (IWAE)

VAE objective

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

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \left(\frac{1}{K} \sum_{k=1}^{K} \log \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi})} \right) \to \max_{q, \boldsymbol{\theta}}.$$

IWAE objective

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi})} \right) \to \max_{q, \boldsymbol{\theta}}.$$

If $K = 1$, these objectives coincide.

Burda Y., Grosse R., Salakhutdinov R. Importance Weighted Autoencoders, 2015

Importance Weighted Autoencoders (IWAE)

Theorem

1. $\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}_M(q, \boldsymbol{\theta}), \quad$ for $K \geq M$;

2. $\log p(\mathbf{x}|\boldsymbol{\theta}) = \lim_{K \to \infty} \mathcal{L}_K(q, \boldsymbol{\theta})$ if $\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})}$ is bounded.

If $K > 1$ the bound could be tighter.

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

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi})} \right).$$

- $\mathcal{L}_1(q, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta})$;
- $\mathcal{L}_\infty(q, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$.
- Which $q^*(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$ gives $\mathcal{L}(q^*, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$?

Burda Y., Grosse R., Salakhutdinov R. Importance Weighted Autoencoders, 2015

Importance Weighted Autoencoders (IWAE)

Objective

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x}, \phi)} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k | \boldsymbol{\theta})}{q(\mathbf{z}_k | \mathbf{x}, \phi)} \right) \to \max_{\phi, \boldsymbol{\theta}}.$$

Theorem

Gradient signal of $q(\mathbf{z}|\mathbf{x}, \phi)$ vanishes as K increases:

$$\Delta_K = \nabla_{\boldsymbol{\theta}, \phi} \mathcal{L}_K(q, \boldsymbol{\theta}); \quad \mathsf{SNR}_K = \frac{\mathbb{E}[\Delta_K]}{\sigma(\Delta_K)};$$

$$\mathsf{SNR}_K(\boldsymbol{\theta}) = O(\sqrt{K}); \quad \mathsf{SNR}_K(\phi) = O\left(\sqrt{K^{-1}}\right).$$

- ▶ IWAE makes the variational bound tighter and extends the class of variational distributions.
- ▶ Gradient signal becomes really small, training is complicated.
- ▶ IWAE is a standard quality measure for VAE models.

Rainforth T. et al. Tighter variational bounds are not necessarily better, 2018

Outline

Likelihood-based models so far...

Autoregressive models

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{m} p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta})$$
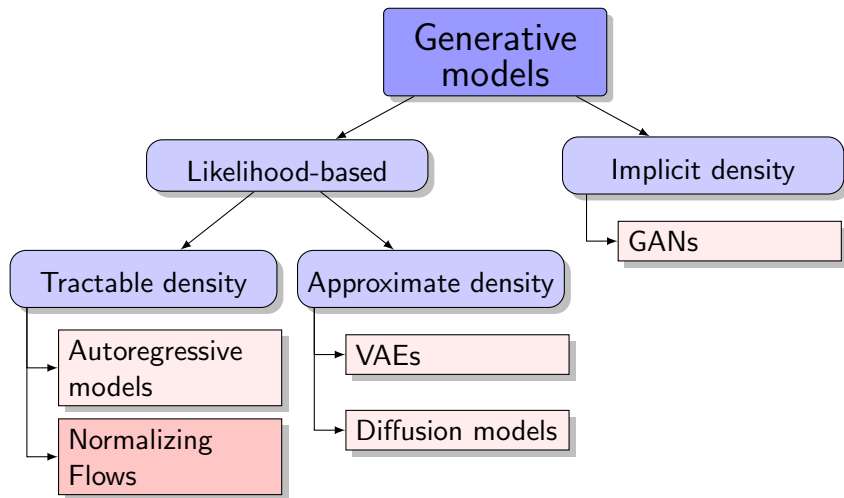
▶ tractable likelihood,

▶ no inferred latent factors.

Latent variable models

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

▶ latent feature representation,

▶ intractable likelihood.

How to build model with latent variables and tractable likelihood?

Generative models zoo

Normalizing flows prerequisites

Jacobian matrix
Let $f : \mathbb{R}^m \to \mathbb{R}^m$ is a differentiable function.

$$\mathbf{z} = f(\mathbf{x}), \quad \mathbf{J} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \cdots & \frac{\partial z_1}{\partial x_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial z_m}{\partial x_1} & \cdots & \frac{\partial z_m}{\partial x_m} \end{pmatrix} \in \mathbb{R}^{m \times m}$$

Change of variable theorem (CoV)
Let $\mathbf{x}$ be a random variable with density function $p(\mathbf{x})$ and $f : \mathbb{R}^m \to \mathbb{R}^m$ is a differentiable, **invertible** function (diffeomorphism). If $\mathbf{z} = f(\mathbf{x})$, $\mathbf{x} = f^{-1}(\mathbf{z}) = g(\mathbf{z})$, then

$$p(\mathbf{x}) = p(\mathbf{z})|\det(\mathbf{J}_f)| = p(\mathbf{z})\left|\det\left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right)\right| = p(f(\mathbf{x}))\left|\det\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\right)\right|$$

$$p(\mathbf{z}) = p(\mathbf{x})|\det(\mathbf{J}_g)| = p(\mathbf{x})\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}}\right)\right| = p(g(\mathbf{z}))\left|\det\left(\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}}\right)\right|.$$

Summary

▶ Standart VAE has several limitations that we will address later in the course.

▶ More powerful decoder in VAE leads to more expressive generative model. However, too expressive decoder could lead to the posterior collapse.

▶ The decoder weakening is a set of techniques to avoid the posterior collapse.

▶ The IWAE could get the tighter lower bound to the likelihood, but the training of such model becomes more difficult.

▶ Change of variable theorem allows to get the density function of the random variable under the invertible transformation.