Deep Generative Models

Lecture 4

Roman Isachenko

AI Masters

Autumn, 2022

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

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

- M-step

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

Recap of previous lecture

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

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

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

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

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

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

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

Variational assumption

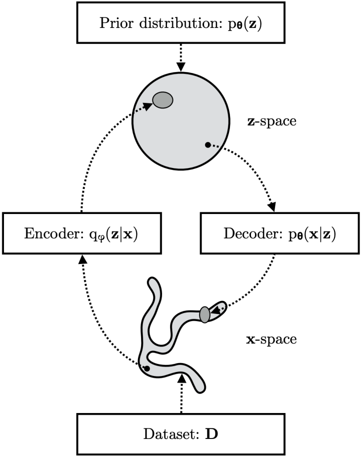
$$r(\boldsymbol{\epsilon}) = \mathcal{N}(0, \mathbf{I}); \quad q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})).$$

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

Recap of previous lecture

Variational autoencoder (VAE)

- VAE learns stochastic mapping between **x**-space, from $\pi(\mathbf{x})$, and a latent **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 $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$.


Prior distribution: $p_{\boldsymbol{\theta}}(\mathbf{z})$
z-space
Encoder: $q_{\varphi}(\mathbf{z}|\mathbf{x})$
Decoder: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$
x-space
Dataset: **D**

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

Outline

1. Posterior collapse and decoder weakening techniques

2. Tighter variational bound

3. Normalizing flows

Outline

1. Posterior collapse and decoder weakening techniques
2. Tighter variational bound
3. Normalizing flows

VAE limitations

- **Poor generative distribution (decoder)**

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

- 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)

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

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

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

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

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

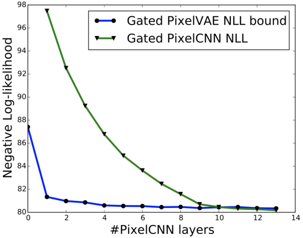
PixelVAE

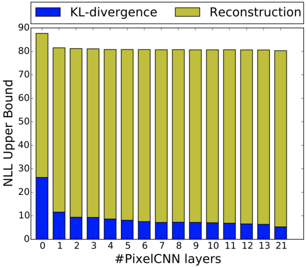
Autoregressive decoder

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

- Global structure is captured by latent variables.
- Local statistics are captured by limited receptive field autoregressive model.

MNIST results





Gulrajani I. et al. PixelVAE: A Latent Variable Model for Natural Images, 2016

Decoder weakening techniques

- Powerful decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ makes the model expressive, but posterior collapse is possible.
- PixelVAE model uses the autoregressive PixelCNN model with small number of layers to limit receptive field.

How to force the model encode information about $\mathbf{x}$ into $\mathbf{z}$?

KL annealing

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

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

Free bits

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

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

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

1. Posterior collapse and decoder weakening techniques
2. Tighter variational bound
3. Normalizing flows

VAE limitations

- Poor generative distribution (decoder)

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

- **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)

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

Importance sampling

LVM

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

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

ELBO: derivation 1

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

Could we choose better $f(\mathbf{x}, \mathbf{z})$?

Importance Weighted Autoencoders (IWAE)

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

Let define

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

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

ELBO

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

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

Importance Weighted Autoencoders (IWAE)

VAE objective

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

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

IWAE objective

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

If $K = 1$, these objectives coincide.

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

Importance Weighted Autoencoders (IWAE)

Theorem

1. $\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}_M(q, \boldsymbol{\theta}), \quad \text{for } K \geq M;$
2. $\log p(\mathbf{x}|\boldsymbol{\theta}) = \lim_{K \to \infty} \mathcal{L}_K(q, \boldsymbol{\theta})$ if $\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \phi)}$ is bounded.

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

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

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

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

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

Gradient

$$\Delta_K = \nabla_{\boldsymbol{\theta}, \phi} \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p(\mathbf{x}, \mathbf{z}_k | \boldsymbol{\theta})}{q(\mathbf{z}_k | \mathbf{x}, \phi)} \right), \quad \mathbf{z}_k \sim q(\mathbf{z}|\mathbf{x}, \phi).$$

Theorem

$$\text{SNR}_K = \frac{\mathbb{E}[\Delta_K]}{\sigma(\Delta_K)}; \quad \text{SNR}_K(\boldsymbol{\theta}) = O(\sqrt{K}); \quad \text{SNR}_K(\phi) = O\left(\sqrt{\frac{1}{K}}\right).$$

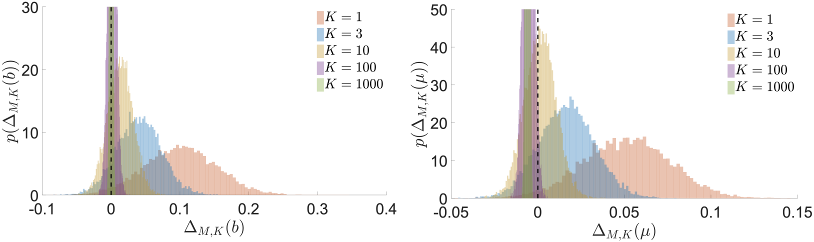
Hence, increasing K vanishes gradient signal of inference network $q(\mathbf{z}|\mathbf{x}, \phi)$.

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

Importance Weighted Autoencoders (IWAE)

Theorem

$$\mathrm{SNR}_K = \frac{\mathbb{E}[\Delta_K]}{\sigma(\Delta_K)}; \quad \mathrm{SNR}_K(\boldsymbol{\theta}) = O(\sqrt{K}); \quad \mathrm{SNR}_K(\phi) = O\left(\sqrt{\frac{1}{K}}\right).$$



- 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

1. Posterior collapse and decoder weakening techniques

2. Tighter variational bound

3. Normalizing flows

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.

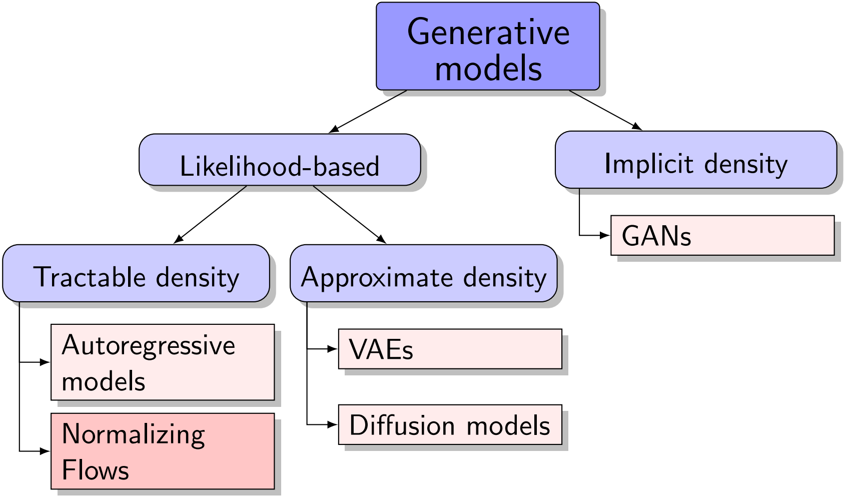
Latent variable models

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

- latent feature representation,
- intractable likelihood.

How to build model with latent variables and tractable likelihood?

Generative models zoo



Normalizing flows prerequisites

Jacobian matrix

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

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

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

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

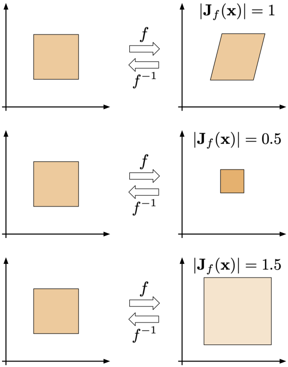
Jacobian determinant

Inverse function theorem

If function f is invertible and Jacobian matrix is continuous and non-singular, then

$$\mathbf{J}_f = \mathbf{J}_{g^{-1}} = \mathbf{J}_g^{-1}, \quad |\det(\mathbf{J}_f)| = \frac{1}{|\det(\mathbf{J}_g)|}$$

- $\mathbf{x}$ and $\mathbf{z}$ have the same dimensionality ($\mathbb{R}^m$).
- $f(\mathbf{x}, \boldsymbol{\theta})$ could be parametric function.
- Determinant of Jacobian matrix $\mathbf{J} = \frac{\partial f(\mathbf{x},\boldsymbol{\theta})}{\partial \mathbf{x}}$ shows how the volume changes under the transorfmation.


$|\mathbf{J}_f(\mathbf{x})| = 1$
f
f^{-1}
$|\mathbf{J}_f(\mathbf{x})| = 0.5$
f
f^{-1}
$|\mathbf{J}_f(\mathbf{x})| = 1.5$
f
f^{-1}

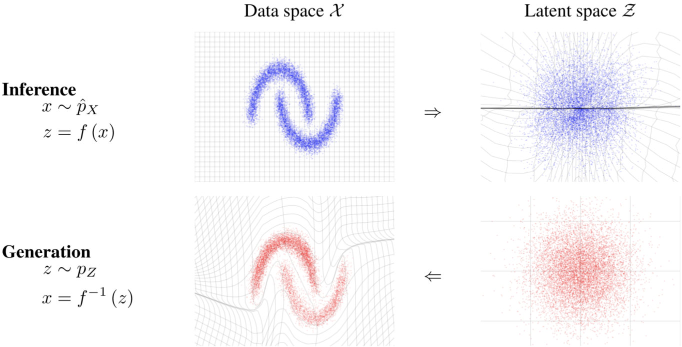
https://jmtomczak.github.io/blog/3/3_flows.html

Fitting flows

MLE problem

$$p(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{z}) \left| \det \left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) \right| = p(f(\mathbf{x}, \boldsymbol{\theta})) \left| \det \left(\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \mathbf{x}} \right) \right|$$

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(f(\mathbf{x}, \boldsymbol{\theta})) + \log |\det(\mathbf{J}_f)| \rightarrow \max_{\boldsymbol{\theta}}$$



Dinh L., Sohl-Dickstein J., Bengio S. Density estimation using Real NVP, 2016

Summary

- 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.
- Flow models transform a simple base distribution to a complex one via a sequence of invertible transformations with tractable Jacobian.
- Flow models have a tractable likelihood that is given by the change of variable theorem.