

# Deep Generative Models

## Lecture 3

Roman Isachenko

Moscow Institute of Physics and Technology  
Yandex School of Data Analysis

2025, Autumn

# Recap of Previous Lecture

## Jacobian Matrix

Given a differentiable function  $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ ,

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

## Change of Variables Theorem (CoV)

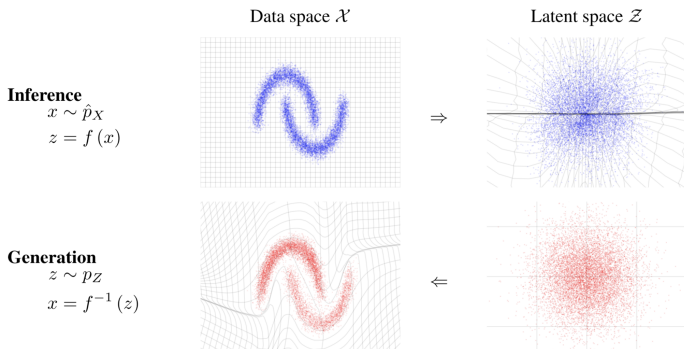
Let  $\mathbf{x}$  be a random variable with density  $p(\mathbf{x})$ , and  $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  a differentiable invertible mapping. If  $\mathbf{z} = \mathbf{f}(\mathbf{x})$  and  $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{z}) = \mathbf{g}(\mathbf{z})$ , then

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

# Recap of Previous Lecture

## Definition

A normalizing flow is a *differentiable, invertible* transformation that maps data  $\mathbf{x}$  to noise  $\mathbf{z}$ .



## Log-Likelihood

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log p(\mathbf{f}_K \circ \dots \circ \mathbf{f}_1(\mathbf{x})) + \sum_{k=1}^K \log |\det(\mathbf{J}_{\mathbf{f}_k})|$$

# Recap of Previous Lecture

## Flow Log-Likelihood

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

One significant challenge is efficiently computing the Jacobian determinant.

## Linear Flows

$$\mathbf{z} = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{W}\mathbf{x}, \quad \mathbf{W} \in \mathbb{R}^{m \times m}, \quad \boldsymbol{\theta} = \mathbf{W}, \quad \mathbf{J}_{\mathbf{f}} = \mathbf{W}^T$$

- ▶ LU Decomposition:

$$\mathbf{W} = \mathbf{P}\mathbf{L}\mathbf{U}.$$

- ▶ QR Decomposition:

$$\mathbf{W} = \mathbf{Q}\mathbf{R}.$$

Decomposition is performed only once during initialization. Then the decomposed matrices ( $\mathbf{P}$ ,  $\mathbf{L}$ ,  $\mathbf{U}$  or  $\mathbf{Q}$ ,  $\mathbf{R}$ ) are optimized.

## Recap of Previous Lecture

Consider an autoregressive model:

$$p(\mathbf{x}|\theta) = \prod_{j=1}^m p(x_j|\mathbf{x}_{1:j-1}, \theta), \quad p(x_j|\mathbf{x}_{1:j-1}, \theta) = \mathcal{N}(\mu_{j,\theta}(\mathbf{x}_{1:j-1}), \sigma_{j,\theta}^2(\mathbf{x}_{1:j-1})).$$

### Gaussian Autoregressive Normalizing Flow

$$\mathbf{x} = \mathbf{g}_{\theta}(\mathbf{z}) \quad \Rightarrow \quad x_j = \sigma_{j,\theta}(\mathbf{x}_{1:j-1}) \cdot z_j + \mu_{j,\theta}(\mathbf{x}_{1:j-1}).$$

$$\mathbf{z} = \mathbf{f}_{\theta}(\mathbf{x}) \quad \Rightarrow \quad z_j = (x_j - \mu_{j,\theta}(\mathbf{x}_{1:j-1})) \cdot \frac{1}{\sigma_{j,\theta}(\mathbf{x}_{1:j-1})}.$$

- ▶ This transformation is both **invertible** and **differentiable**, mapping  $p(\mathbf{z})$  to  $p(\mathbf{x}|\theta)$ .
- ▶ The Jacobian matrix for this transformation is triangular.

The generative function  $\mathbf{g}_{\theta}(\mathbf{z})$  is **sequential**, while the inference function  $\mathbf{f}_{\theta}(\mathbf{x})$  is **not sequential**.

# Recap of Previous Lecture

Let us partition  $\mathbf{x}$  and  $\mathbf{z}$  into two groups:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] = [\mathbf{x}_{1:d}, \mathbf{x}_{d+1:m}]; \quad \mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2] = [\mathbf{z}_{1:d}, \mathbf{z}_{d+1:m}].$$

## Coupling Layer

$$\begin{cases} \mathbf{x}_1 = \mathbf{z}_1; \\ \mathbf{x}_2 = \mathbf{z}_2 \odot \sigma_{\theta}(\mathbf{z}_1) + \mu_{\theta}(\mathbf{z}_1). \end{cases} \quad \begin{cases} \mathbf{z}_1 = \mathbf{x}_1; \\ \mathbf{z}_2 = (\mathbf{x}_2 - \mu_{\theta}(\mathbf{x}_1)) \odot \frac{1}{\sigma_{\theta}(\mathbf{x}_1)}. \end{cases}$$

Both density estimation and sampling require just a single pass!

## Jacobian

$$\det \left( \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) = \det \begin{pmatrix} \mathbf{I}_d & \mathbf{0}_{d \times (m-d)} \\ \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_2} \end{pmatrix} = \prod_{j=1}^{m-d} \frac{1}{\sigma_{j,\theta}(\mathbf{x}_1)}.$$

A coupling layer is a special instance of an gaussian autoregressive normalizing flow.

# Outline

1. Forward and Reverse KL for NF
2. Latent Variable Models (LVM)
3. Variational Evidence Lower Bound (ELBO)
4. EM-Algorithm

# Outline

1. Forward and Reverse KL for NF
2. Latent Variable Models (LVM)
3. Variational Evidence Lower Bound (ELBO)
4. EM-Algorithm



# Forward KL vs. Reverse KL

Forward KL ( $\equiv$  Maximum Likelihood Estimation)

$$\begin{aligned}\text{KL}(\pi\|p) &= \int \pi(\mathbf{x}) \log \frac{\pi(\mathbf{x})}{p(\mathbf{x}|\boldsymbol{\theta})} d\mathbf{x} \\ &= -\mathbb{E}_{\pi(\mathbf{x})} \log p(\mathbf{x}|\boldsymbol{\theta}) + \text{const} \rightarrow \min_{\boldsymbol{\theta}}\end{aligned}$$

Forward KL for Normalizing Flows

$$\begin{aligned}\log p(\mathbf{x}|\boldsymbol{\theta}) &= \log p(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_{\mathbf{f}})| \\ \text{KL}(\pi\|p) &= -\mathbb{E}_{\pi(\mathbf{x})} \left[ \log p(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det(\mathbf{J}_{\mathbf{f}})| \right] + \text{const}\end{aligned}$$

- ▶ We need to compute  $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$  and its Jacobian.
- ▶ Access to the density  $p(\mathbf{z})$  is required.
- ▶ The inverse function  $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{f}_{\boldsymbol{\theta}}^{-1}(\mathbf{z})$  is required only for sampling from the normalizing flow.

# Forward KL vs. Reverse KL

## Reverse KL

$$\begin{aligned}\text{KL}(p\|\pi) &= \int p(\mathbf{x}|\boldsymbol{\theta}) \log \frac{p(\mathbf{x}|\boldsymbol{\theta})}{\pi(\mathbf{x})} d\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} [\log p(\mathbf{x}|\boldsymbol{\theta}) - \log \pi(\mathbf{x})] \rightarrow \min_{\boldsymbol{\theta}}\end{aligned}$$

## Reverse KL for Normalizing Flows (LOTUS Trick)

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

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

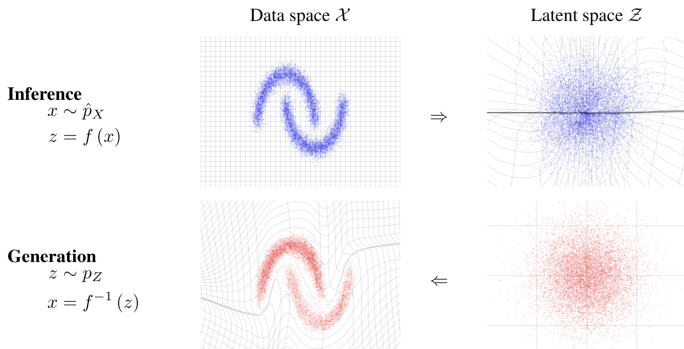
- ▶ We need to compute  $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})$  and its Jacobian.
- ▶ Sampling from  $p(\mathbf{z})$  is required (though direct evaluation is not), along with evaluating  $\pi(\mathbf{x})$ .
- ▶ Evaluating  $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$  is not required.

# Normalizing Flows KL Duality

## Theorem

Fitting the NF model  $p(\mathbf{x}|\boldsymbol{\theta})$  to a target distribution  $\pi(\mathbf{x})$  via the forward KL (MLE) is equivalent to fitting the induced distribution  $p(\mathbf{z}|\boldsymbol{\theta})$  to the base distribution  $p(\mathbf{z})$  via the reverse KL:

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



# Normalizing Flows KL Duality

## Theorem

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

## Proof

- ▶  $\mathbf{z} \sim p(\mathbf{z})$ ,  $\mathbf{x} = \mathbf{g}_{\theta}(\mathbf{z})$ , so  $\mathbf{x} \sim p(\mathbf{x}|\theta)$ ;
- ▶  $\mathbf{x} \sim \pi(\mathbf{x})$ ,  $\mathbf{z} = \mathbf{f}_{\theta}(\mathbf{x})$ , so  $\mathbf{z} \sim p(\mathbf{z}|\theta)$ ;

$$\log p(\mathbf{z}|\theta) = \log \pi(\mathbf{g}_{\theta}(\mathbf{z})) + \log |\det(\mathbf{J}_{\mathbf{g}})|;$$

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

$$\begin{aligned} \text{KL}(p(\mathbf{z}|\theta) \| p(\mathbf{z})) &= \mathbb{E}_{p(\mathbf{z}|\theta)} [\log p(\mathbf{z}|\theta) - \log p(\mathbf{z})] = \\ &= \mathbb{E}_{p(\mathbf{z}|\theta)} [\log \pi(\mathbf{g}_{\theta}(\mathbf{z})) + \log |\det(\mathbf{J}_{\mathbf{g}})| - \log p(\mathbf{z})] = \\ &= \mathbb{E}_{\pi(\mathbf{x})} [\log \pi(\mathbf{x}) - \log |\det(\mathbf{J}_{\mathbf{f}})| - \log p(\mathbf{f}_{\theta}(\mathbf{x}))] = \\ &= \mathbb{E}_{\pi(\mathbf{x})} [\log \pi(\mathbf{x}) - \log p(\mathbf{x}|\theta)] = \text{KL}(\pi(\mathbf{x}) \| p(\mathbf{x}|\theta)). \end{aligned}$$

# Outline

1. Forward and Reverse KL for NF
2. Latent Variable Models (LVM)
3. Variational Evidence Lower Bound (ELBO)
4. EM-Algorithm

# Bayesian Framework

## Bayes' Theorem

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

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

## Interpretation

- ▶ We begin with unknown variables  $\theta$  and a prior belief  $p(\theta)$ .
- ▶ Once data  $\mathbf{x}$  is observed, the posterior  $p(\theta|\mathbf{x})$  incorporates both prior beliefs and evidence from the data.

# Bayesian Framework

Consider the case where the unobserved variables  $\theta$  are model parameters (i.e.,  $\theta$  are random variables).

- ▶  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ : observed samples;
- ▶  $p(\theta)$ : prior distribution.

## Posterior Distribution

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

If the evidence  $p(\mathbf{X})$  is intractable (due to high-dimensional integration), the posterior cannot be computed exactly.

## Maximum a Posteriori (MAP) Estimation

$$\theta^* = \arg \max_{\theta} p(\theta|\mathbf{X}) = \arg \max_{\theta} (\log p(\mathbf{X}|\theta) + \log p(\theta))$$

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

$$\theta^* = \arg \max_{\theta} p(\mathbf{X}|\theta) = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i|\theta) = \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i|\theta).$$

The distribution  $p(\mathbf{x}|\theta)$  can be highly complex and often intractable (just like the true data distribution  $\pi(\mathbf{x})$ ).

## Extended Probabilistic Model

Introduce a latent variable  $\mathbf{z}$  for each observed sample  $\mathbf{x}$ :

$$p(\mathbf{x}, \mathbf{z}|\theta) = p(\mathbf{x}|\mathbf{z}, \theta)p(\mathbf{z}); \quad \log p(\mathbf{x}, \mathbf{z}|\theta) = \log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(\mathbf{z}).$$

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

## Motivation

Both  $p(\mathbf{x}|\mathbf{z}, \theta)$  and  $p(\mathbf{z})$  are usually much simpler than  $p(\mathbf{x}|\theta)$ .

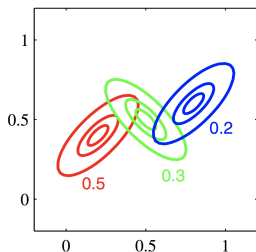


# Latent Variable Models (LVM)

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

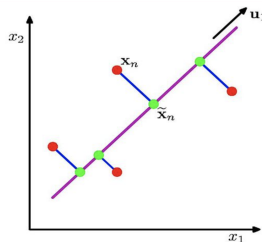
## Examples

### Mixture of Gaussians



- ▶  $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
- ▶  $p(\mathbf{z}) = \text{Categorical}(\boldsymbol{\pi})$

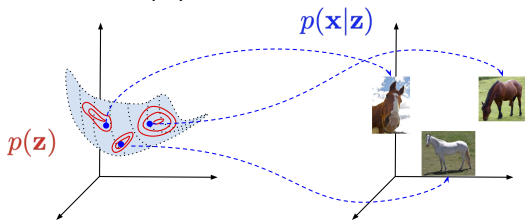
### PCA Model



- ▶  $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
- ▶  $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$

## MLE for LVM

$$\sum_{i=1}^n \log p(\mathbf{x}_i | \theta) = \sum_{i=1}^n \log \int p(\mathbf{x}_i | \mathbf{z}_i, \theta) p(\mathbf{z}_i) d\mathbf{z}_i \rightarrow \max_{\theta}.$$



## Naive Approach

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

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

**Challenge:** As the dimensionality of  $\mathbf{z}$  increases, the number of samples needed to adequately cover the latent space grows exponentially.

image credit: [https://jmtomczak.github.io/blog/4/4\\_VAE.html](https://jmtomczak.github.io/blog/4/4_VAE.html)

# Outline

1. Forward and Reverse KL for NF
2. Latent Variable Models (LVM)
3. Variational Evidence Lower Bound (ELBO)
4. EM-Algorithm

# ELBO Derivation I

## Inequality Derivation

$$\begin{aligned}\log p(\mathbf{x}|\boldsymbol{\theta}) &= \log \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \\ &= \log \mathbb{E}_q \left[ \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} \right] \geq \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z})} = \mathcal{L}_{q,\boldsymbol{\theta}}(\mathbf{x})\end{aligned}$$

Here,  $q(\mathbf{z})$  is any distribution such that  $\int q(\mathbf{z}) d\mathbf{z} = 1$ .

## Variational Evidence Lower Bound (ELBO)

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

This inequality holds for any choice of  $q(\mathbf{z})$ .

## ELBO Derivation II

$$p(\mathbf{z}|\mathbf{x}, \theta) = \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{p(\mathbf{x}|\theta)}$$

### Equality Derivation

$$\begin{aligned}\mathcal{L}_{q,\theta}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z})} d\mathbf{z} = \\ &= \int q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)}{q(\mathbf{z})} d\mathbf{z} = \\ &= \int q(\mathbf{z}) \log p(\mathbf{x}|\theta) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \theta)}{q(\mathbf{z})} d\mathbf{z} = \\ &= \log p(\mathbf{x}|\theta) - \text{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}, \theta))\end{aligned}$$

### Variational Decomposition

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

Here,  $\text{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}, \theta)) \geq 0$ .

# Variational Evidence Lower Bound (ELBO)

$$\begin{aligned}\mathcal{L}_{q,\theta}(\mathbf{x}) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z})} d\mathbf{z} = \\ &= \int q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z}, \theta) d\mathbf{z} + \int q(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \mathbb{E}_q \log p(\mathbf{x}|\mathbf{z}, \theta) - \text{KL}(q(\mathbf{z})\|p(\mathbf{z}))\end{aligned}$$

## Log-Likelihood Decomposition

$$\begin{aligned}\log p(\mathbf{x}|\theta) &= \mathcal{L}_{q,\theta}(\mathbf{x}) + \text{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}, \theta)) \\ &= \mathbb{E}_q \log p(\mathbf{x}|\mathbf{z}, \theta) - \text{KL}(q(\mathbf{z})\|p(\mathbf{z})) + \text{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}, \theta)).\end{aligned}$$

- Instead of maximizing the likelihood, maximize the ELBO:

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

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

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

# Outline

1. Forward and Reverse KL for NF
2. Latent Variable Models (LVM)
3. Variational Evidence Lower Bound (ELBO)
4. EM-Algorithm

# EM-Algorithm

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

## Block-Coordinate Optimization

- ▶ Initialize  $\theta^*$ ;
- ▶ **E-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $q$ ):

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

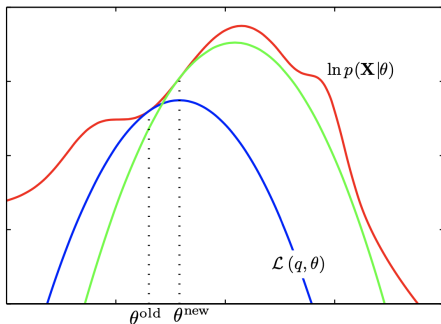
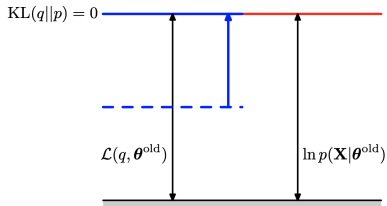
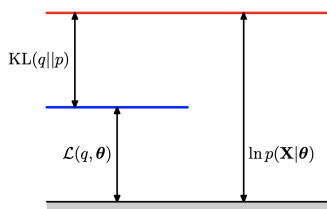
- ▶ **M-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $\theta$ ):

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

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



# EM-Algorithm Illustration



# Summary

- ▶ Flow duality establishes the relationship between the data and latent spaces using forward and reverse KL formulations.
- ▶ The Bayesian framework generalizes nearly all standard machine learning methods.
- ▶ LVMs introduce latent representations for observed data, enabling more interpretable models.
- ▶ LVMs maximize the variational evidence lower bound (ELBO) to obtain maximum likelihood estimates for the parameters.
- ▶ The general variational EM algorithm optimizes the ELBO within LVMs to recover the MLE for the parameters  $\theta$ .