

# 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

Let  $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  be a differentiable function.

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

## Change of Variables Theorem (CoV)

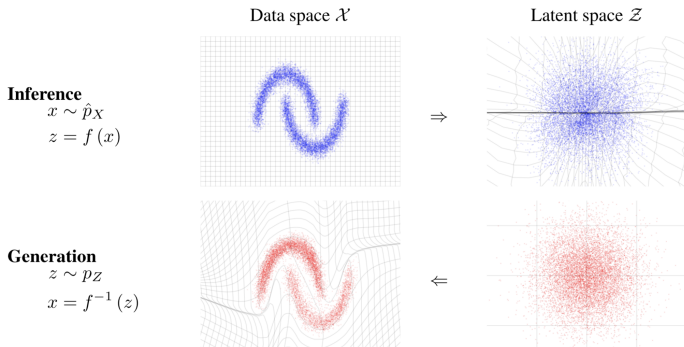
Let  $\mathbf{x} \in \mathbb{R}^m$  be a random vector with density  $p(\mathbf{x})$ , and let  $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  be a  $C^1$ -diffeomorphism ( $\mathbf{f}$  and  $\mathbf{f}^{-1}$  are continuously differentiable mappings). If  $\mathbf{z} = \mathbf{f}(\mathbf{x})$ , then

$$p(\mathbf{x}) = p(\mathbf{z}) |\det(\mathbf{J}_{\mathbf{f}})| = p(\mathbf{z}) \left| \det \left( \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) \right| = p(\mathbf{f}(\mathbf{x})) \left| \det \left( \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right) \right|$$
$$p(\mathbf{z}) = p(\mathbf{x}) |\det(\mathbf{J}_{\mathbf{f}^{-1}})| = p(\mathbf{x}) \left| \det \left( \frac{\partial \mathbf{x}}{\partial \mathbf{z}} \right) \right| = p(\mathbf{f}^{-1}(\mathbf{z})) \left| \det \left( \frac{\partial \mathbf{f}^{-1}(\mathbf{z})}{\partial \mathbf{z}} \right) \right|$$

# Recap of Previous Lecture

## Definition

A normalizing flow is a  $C^1$ -diffeomorphism that transforms data  $\mathbf{x}$  to noise  $\mathbf{z}$ .



## Log-Likelihood

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

# Recap of Previous Lecture

## Flow Log-Likelihood

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

One significant challenge is efficiently computing the Jacobian determinant.

## Linear Flows

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

- ▶ 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 (**P**, **L**, **U** or **Q**, **R**) are optimized.

## Recap of Previous Lecture

Consider an autoregressive model:

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

### Gaussian Autoregressive Normalizing Flow

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

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

- ▶ This transformation is both  $C^1$ -**diffeomorphism**, mapping  $p(\mathbf{z})$  to  $p_{\theta}(\mathbf{x})$ .
- ▶ The Jacobian matrix for this transformation is triangular.

The generative function  $\mathbf{f}_{\theta}^{-1}(\mathbf{z})$  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.

# Recap of Previous Lecture

## Posterior Distribution (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;
- ▶  $\theta$  – unobserved variables (latent parameters);
- ▶  $p_{\theta}(\mathbf{x}) = p(\mathbf{x}|\theta)$  – likelihood;
- ▶  $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$  – evidence;
- ▶  $p(\theta)$  – prior distribution;
- ▶  $p(\theta|\mathbf{x})$  – posterior distribution.

# Outline

1. Latent Variable Models (LVM) (continued)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference



# Outline

1. Latent Variable Models (LVM) (continued)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

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

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

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

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

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

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

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

## Extended Probabilistic Model

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

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

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

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

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

## Extended Probabilistic Model

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

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

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

# Latent Variable Models (LVM)

## Maximum Likelihood Estimation (MLE) Problem

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

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

## Extended Probabilistic Model

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

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

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

## Motivation

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

# Latent Variable Models (LVM)

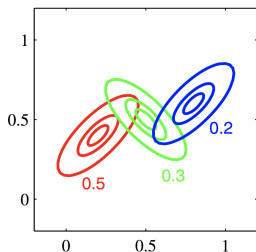
$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \rightarrow \max_{\theta}$$

# Latent Variable Models (LVM)

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

## Examples

### *Mixture of Gaussians*



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

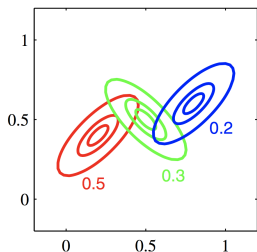


# Latent Variable Models (LVM)

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

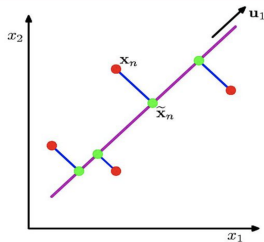
## Examples

### Mixture of Gaussians



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

### PCA Model



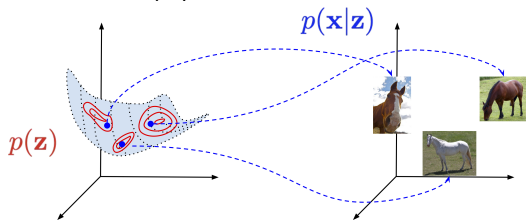
- ▶  $p_{\theta}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
- ▶  $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$

## MLE for LVM

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

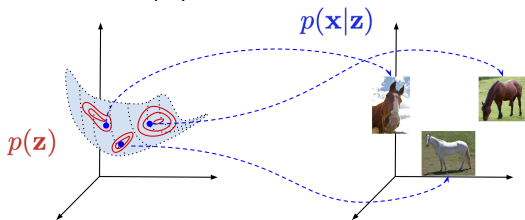
## MLE for LVM

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



## MLE for LVM

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



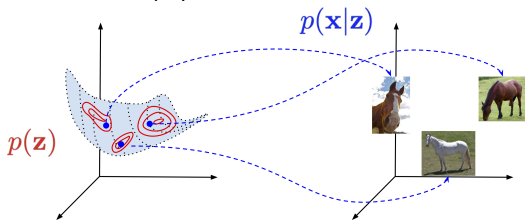
## Naive Approach

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

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

## MLE for LVM

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



## Naive Approach

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

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. Latent Variable Models (LVM) (continued)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

# ELBO Derivation I

## Inequality Derivation

$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

# ELBO Derivation I

## Inequality Derivation

$$\log p_{\theta}(\mathbf{x}) = \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$



# ELBO Derivation I

## Inequality Derivation

$$\begin{aligned}\log p_{\theta}(\mathbf{x}) &= \log \int p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \log \int \frac{q(\mathbf{z})}{q(\mathbf{z})} p_{\theta}(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \\ &= \log \mathbb{E}_q \left[ \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right]\end{aligned}$$

# ELBO Derivation I

## Inequality Derivation

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

# ELBO Derivation I

## Inequality Derivation

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

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

# ELBO Derivation I

## Inequality Derivation

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

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

## Variational Evidence Lower Bound (ELBO)

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

# ELBO Derivation I

## Inequality Derivation

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

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

## Variational Evidence Lower Bound (ELBO)

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

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

## ELBO Derivation II

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

### Equality Derivation

$$\mathcal{L}_{q,\theta}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

## ELBO Derivation II

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

### Equality Derivation

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

## ELBO Derivation II

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

### Equality Derivation

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



## ELBO Derivation II

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

### Equality Derivation

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

## ELBO Derivation II

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

### Equality Derivation

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

### Variational Decomposition

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

## ELBO Derivation II

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

### Equality Derivation

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

### Variational Decomposition

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

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

## Variational Evidence Lower Bound (ELBO)

$$\mathcal{L}_{q,\theta}(\mathbf{x}) = \int q(\mathbf{z}) \log \frac{p_{\theta}(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

## Variational Evidence Lower Bound (ELBO)

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

## Variational Evidence Lower Bound (ELBO)

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

## Variational Evidence Lower Bound (ELBO)

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

### Log-Likelihood Decomposition

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

## Variational Evidence Lower Bound (ELBO)

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

### Log-Likelihood Decomposition

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



# Variational Evidence Lower Bound (ELBO)

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

## Log-Likelihood Decomposition

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

- Instead of maximizing the likelihood, maximize the ELBO:

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

# Variational Evidence Lower Bound (ELBO)

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

## Log-Likelihood Decomposition

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

- Instead of maximizing the likelihood, maximize the ELBO:

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

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

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

# Outline

1. Latent Variable Models (LVM) (continued)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

# EM-Algorithm

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

# EM-Algorithm

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

## Block-Coordinate Optimization

- Initialize  $\theta^*$ ;

# EM-Algorithm

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

## Block-Coordinate Optimization

- ▶ Initialize  $\theta^*$ ;
- ▶ **E-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $q$ ):
$$\begin{aligned}q^*(\mathbf{z}) &= \arg \max_q \mathcal{L}_{q,\theta^*}(\mathbf{x}) = \\ &= \arg \min_q \text{KL}(q(\mathbf{z})\|p_{\theta^*}(\mathbf{z}|\mathbf{x})) = p_{\theta^*}(\mathbf{z}|\mathbf{x});\end{aligned}$$

# EM-Algorithm

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

## Block-Coordinate Optimization

- ▶ Initialize  $\theta^*$ ;
- ▶ **E-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $q$ ):
$$\begin{aligned}q^*(\mathbf{z}) &= \arg \max_q \mathcal{L}_{q,\theta^*}(\mathbf{x}) = \\ &= \arg \min_q \text{KL}(q(\mathbf{z})\|p_{\theta^*}(\mathbf{z}|\mathbf{x})) = p_{\theta^*}(\mathbf{z}|\mathbf{x});\end{aligned}$$
- ▶ **M-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $\theta$ ):
$$\theta^* = \arg \max_{\theta} \mathcal{L}_{q^*,\theta}(\mathbf{x});$$

# EM-Algorithm

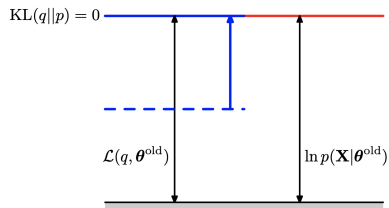
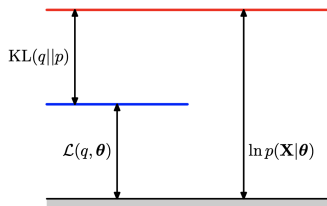
$$\begin{aligned}\mathcal{L}_{q,\theta}(\mathbf{x}) &= \mathbb{E}_q \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \text{KL}(q(\mathbf{z})\|p(\mathbf{z})) = \\ &= \mathbb{E}_q \left[ \log p_{\theta}(\mathbf{x}|\mathbf{z}) - \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right] d\mathbf{z} \rightarrow \max_{q,\theta}.\end{aligned}$$

## Block-Coordinate Optimization

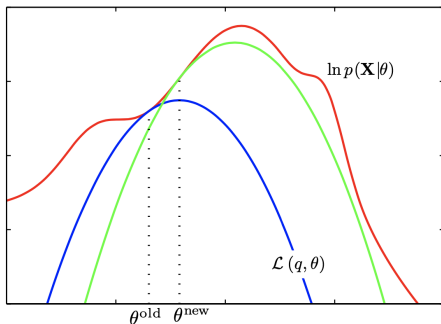
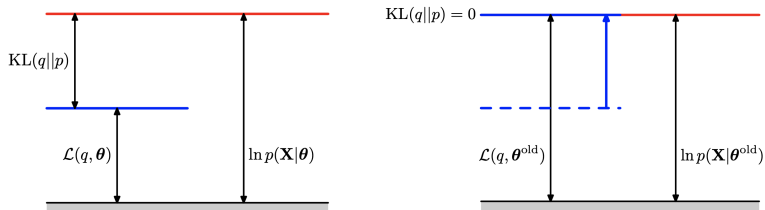
- ▶ Initialize  $\theta^*$ ;
- ▶ **E-step** (optimize  $\mathcal{L}_{q,\theta}(\mathbf{x})$  over  $q$ ):
$$\begin{aligned}q^*(\mathbf{z}) &= \arg \max_q \mathcal{L}_{q,\theta^*}(\mathbf{x}) = \\ &= \arg \min_q \text{KL}(q(\mathbf{z})\|p_{\theta^*}(\mathbf{z}|\mathbf{x})) = p_{\theta^*}(\mathbf{z}|\mathbf{x});\end{aligned}$$
- ▶ **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



# EM-Algorithm Illustration



# Outline

1. Latent Variable Models (LVM) (continued)
2. Variational Evidence Lower Bound (ELBO)
3. EM-Algorithm
4. Amortized Inference

# Amortized Variational Inference

## E-step

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

# Amortized Variational Inference

## E-step

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

$q(\mathbf{z})$  approximates the true posterior  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$ , hence it is called **variational posterior**.

# Amortized Variational Inference

## E-step

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

$q(\mathbf{z})$  approximates the true posterior  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$ , hence it is called **variational posterior**.

- ▶  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$  may be **intractable**;
- ▶  $q(\mathbf{z})$  is individual for each data point  $\mathbf{x}$ .

# Amortized Variational Inference

## E-step

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

$q(\mathbf{z})$  approximates the true posterior  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$ , hence it is called **variational posterior**.

- ▶  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$  may be **intractable**;
- ▶  $q(\mathbf{z})$  is individual for each data point  $\mathbf{x}$ .

## Variational Bayes

We restrict the family of possible distributions  $q(\mathbf{z})$  to a parametric class  $q_{\phi}(\mathbf{z} | \mathbf{x})$ , **conditioned on data  $\mathbf{x}$**  and **parameterized by  $\phi$** .

# Amortized Variational Inference

## E-step

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

$q(\mathbf{z})$  approximates the true posterior  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$ , hence it is called **variational posterior**.

- ▶  $p_{\theta^*}(\mathbf{z} | \mathbf{x})$  may be **intractable**;
- ▶  $q(\mathbf{z})$  is individual for each data point  $\mathbf{x}$ .

## Variational Bayes

We restrict the family of possible distributions  $q(\mathbf{z})$  to a parametric class  $q_{\phi}(\mathbf{z} | \mathbf{x})$ , **conditioned on data  $\mathbf{x}$**  and **parameterized by  $\phi$** .

- ▶ 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

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



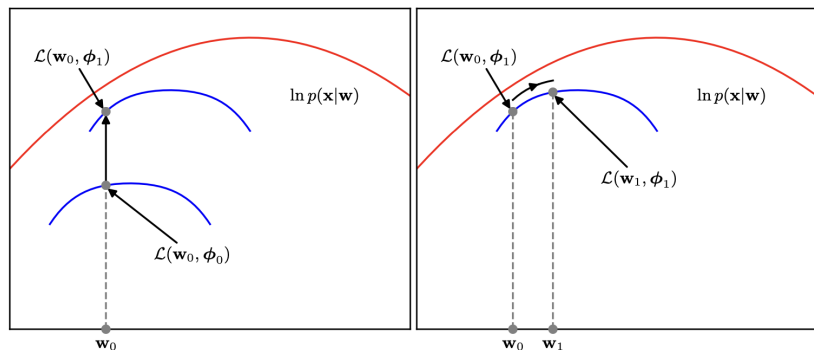
# Variational EM Illustration

- 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:

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



# Variational EM Algorithm

## ELBO

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

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

# Variational EM Algorithm

## ELBO

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

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

### ► E-step:

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

where  $\phi$  denotes the parameters of the variational posterior  $q_{\phi}(\mathbf{z}|\mathbf{x})$ .

### ► M-step:

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

where  $\theta$  represents the parameters of the generative model  $p_{\theta}(\mathbf{x}|\mathbf{z})$ .

# Variational EM Algorithm

## ELBO

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

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

### ► E-step:

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

where  $\phi$  denotes the parameters of the variational posterior  $q_{\phi}(\mathbf{z}|\mathbf{x})$ .

### ► M-step:

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

where  $\theta$  represents the parameters of the generative model  $p_{\theta}(\mathbf{x}|\mathbf{z})$ .

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

# Summary

- ▶ 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  $\theta$ .
- ▶ Amortized variational inference enables efficient estimation of the ELBO via Monte Carlo estimation.