

# Deep Generative Models

## Lecture 2

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

## Recap of previous lecture

We are given i.i.d. samples  $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$  (e.g.  $\mathcal{X} = \mathbb{R}^m$ ) from unknown distribution  $\pi(\mathbf{x})$ .

### Goal

We would like to learn a distribution  $\pi(\mathbf{x})$  for

- ▶ evaluating  $\pi(\mathbf{x})$  for new samples (how likely to get object  $\mathbf{x}$ ?);
- ▶ sampling from  $\pi(\mathbf{x})$  (to get new objects  $\mathbf{x} \sim \pi(\mathbf{x})$ ).

Instead of searching true  $\pi(\mathbf{x})$  over all probability distributions, learn function approximation  $p(\mathbf{x}|\theta) \approx \pi(\mathbf{x})$ .

### Divergence

- ▶  $D(\pi||p) \geq 0$  for all  $\pi, p \in \mathcal{S}$ ;
- ▶  $D(\pi||p) = 0$  if and only if  $\pi \equiv p$ .

### Divergence minimization task

$$\min_{\theta} D(\pi||p).$$

# Recap of previous lecture

## Forward KL

$$KL(\pi||p) = \int \pi(\mathbf{x}) \log \frac{\pi(\mathbf{x})}{p(\mathbf{x}|\theta)} d\mathbf{x} \rightarrow \min_{\theta}$$

## Reverse KL

$$KL(p||\pi) = \int p(\mathbf{x}|\theta) \log \frac{p(\mathbf{x}|\theta)}{\pi(\mathbf{x})} d\mathbf{x} \rightarrow \min_{\theta}$$

## Maximum likelihood estimation (MLE)

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

Maximum likelihood estimation is equivalent to minimization of the Monte-Carlo estimate of forward KL.

# Recap of previous lecture

## Likelihood as product of conditionals

Let  $\mathbf{x} = (x_1, \dots, x_m)$ ,  $\mathbf{x}_{1:j} = (x_1, \dots, x_j)$ . Then

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^m p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}); \quad \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^m \log p(x_j|\mathbf{x}_{1:j-1}, \boldsymbol{\theta}).$$

## MLE problem for autoregressive model

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \sum_{j=1}^m \log p(x_{ij}|\mathbf{x}_{i,1:j-1}\boldsymbol{\theta}).$$

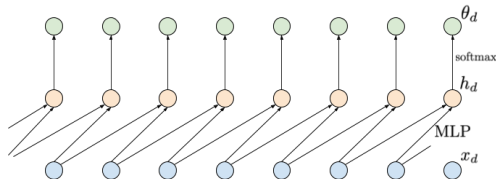
## Sampling

$$\hat{x}_1 \sim p(x_1|\boldsymbol{\theta}), \quad \hat{x}_2 \sim p(x_2|\hat{x}_1, \boldsymbol{\theta}), \quad \dots, \quad \hat{x}_m \sim p(x_m|\hat{\mathbf{x}}_{1:m-1}, \boldsymbol{\theta})$$

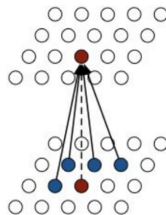
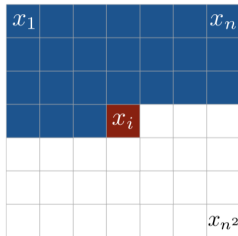
New generated object is  $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m)$ .

# Recap of previous lecture

## Autoregressive MLP



## Autoregressive CNN



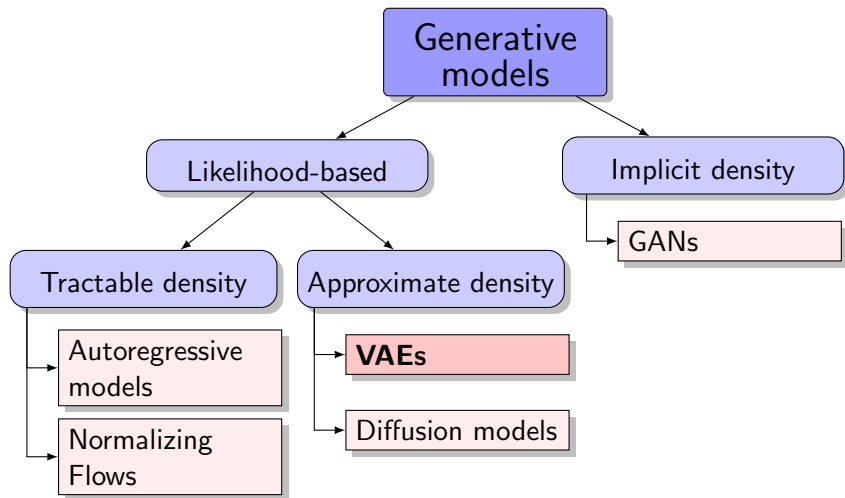
PixelCNN

image credit: [https://jmtomczak.github.io/blog/2/2\\_ARM.html](https://jmtomczak.github.io/blog/2/2_ARM.html)

# Outline

1. Bayesian framework
2. Latent variable models (LVM)
3. Variational lower bound (ELBO)
4. EM-algorithm, amortized inference

# Generative models zoo



# Outline

1. Bayesian framework
2. Latent variable models (LVM)
3. Variational lower bound (ELBO)
4. EM-algorithm, amortized inference



# Bayesian framework

## Bayes theorem

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

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

## Meaning

We have unobserved variables  $\mathbf{t}$  and some prior knowledge about them  $p(\mathbf{t})$ . Then, the data  $\mathbf{x}$  has been observed. Posterior distribution  $p(\mathbf{t}|\mathbf{x})$  summarizes the knowledge after the observations.

## Bayesian framework

Let consider the case, where the unobserved variables  $\mathbf{t}$  is our model parameters  $\theta$ .

- ▶  $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$  – observed samples;
- ▶  $p(\theta)$  – prior parameters distribution (we treat model parameters  $\theta$  as random variables).

## Posterior distribution

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

## Bayesian inference

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

Note the difference from

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

# Bayesian framework

## Posterior distribution

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

## Bayesian inference

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

If evidence  $p(\mathbf{X})$  is intractable (due to multidimensional integration), we can't get posterior distribution and perform the precise inference.

## Maximum a posteriori (MAP) estimation

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

# Bayesian framework

## MAP estimation

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

Estimated  $\theta^*$  is a deterministic variable, but we could treat it as a random variable with density  $p(\theta|\mathbf{X}) = \delta(\theta - \theta^*)$ .

## Dirac delta function

$$\delta(x) = \begin{cases} +\infty, & x = 0; \\ 0, & x \neq 0; \end{cases} \quad \int \delta(x) dx = 1; \quad \int f(x) \delta(x-y) dx = f(y).$$

## MAP inference

$$p(\mathbf{x}|\mathbf{X}) = \int p(\mathbf{x}|\theta)p(\theta|\mathbf{X})d\theta \approx p(\mathbf{x}|\theta^*).$$

# Outline

1. Bayesian framework
2. Latent variable models (LVM)
3. Variational lower bound (ELBO)
4. EM-algorithm, amortized inference

# Latent variable models (LVM)

## 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  $p(\mathbf{x}|\theta)$  could be very complex and intractable (as well as real distribution  $\pi(\mathbf{x})$ ).

## Extended probabilistic model

Introduce latent variable  $\mathbf{z}$  for each 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

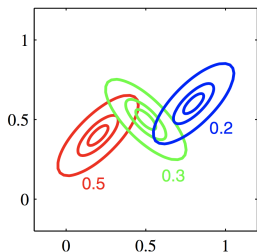
The distributions  $p(\mathbf{x}|\mathbf{z}, \theta)$  and  $p(\mathbf{z})$  could be quite simple.

# Latent variable models (LVM)

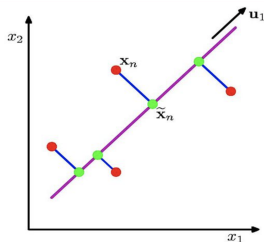
$$\log p(\mathbf{x}|\theta) = \log \int p(\mathbf{x}|\mathbf{z}, \theta) p(\mathbf{z}) d\mathbf{z} \rightarrow \max_{\theta}$$

## Examples

*Mixture of gaussians*



*PCA model*

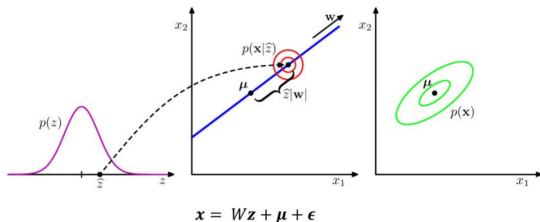


- ▶  $p(\mathbf{x}|\mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}|\mu_{\mathbf{z}}, \Sigma_{\mathbf{z}})$
- ▶  $p(\mathbf{x}|\mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$
- ▶  $p(\mathbf{z}) = \text{Categorical}(\boldsymbol{\pi})$
- ▶  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},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

# Latent variable models (LVM)

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

**PCA** projects original data  $\mathbf{X}$  onto a low dimensional latent space while maximizing the variance of the projected data.



- ▶  $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
- ▶  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$
- ▶  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$
- ▶  $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^2\mathbf{M}), \text{ where } \mathbf{M} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$



# Maximum likelihood estimation for LVM

## MLE for extended problem

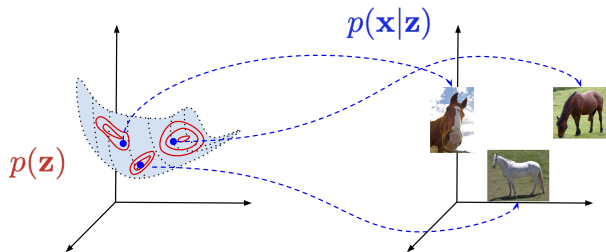
$$\begin{aligned}\theta^* &= \arg \max_{\theta} p(\mathbf{X}, \mathbf{Z} | \theta) = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i, \mathbf{z}_i | \theta) = \\ &= \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i, \mathbf{z}_i | \theta).\end{aligned}$$

However,  $\mathbf{Z}$  is unknown.

## MLE for original problem

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

# Naive approach



## Monte-Carlo estimation

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

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

**Challenge:** to cover the space properly, the number of samples grows exponentially with respect to dimensionality of  $\mathbf{z}$ .

# Outline

1. Bayesian framework
2. Latent variable models (LVM)
3. Variational lower bound (ELBO)
4. EM-algorithm, amortized inference

# Variational lower bound (ELBO)

## Derivation 1 (inequality)

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

## Derivation 2 (equality)

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

## Variational decomposition

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

## Variational lower bound (ELBO)

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

### Log-likelihood decomposition

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

- Instead of maximizing incomplete likelihood, maximize ELBO

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

- Maximization of ELBO by **variational** distribution  $q$  is equivalent to minimization of KL

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

# Outline

1. Bayesian framework
2. Latent variable models (LVM)
3. Variational lower bound (ELBO)
4. EM-algorithm, amortized inference

# EM-algorithm

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

## Block-coordinate optimization

- ▶ Initialize  $\theta^*$ ;
- ▶ **E-step** ( $\mathcal{L}(q, \theta) \rightarrow \max_q$ )

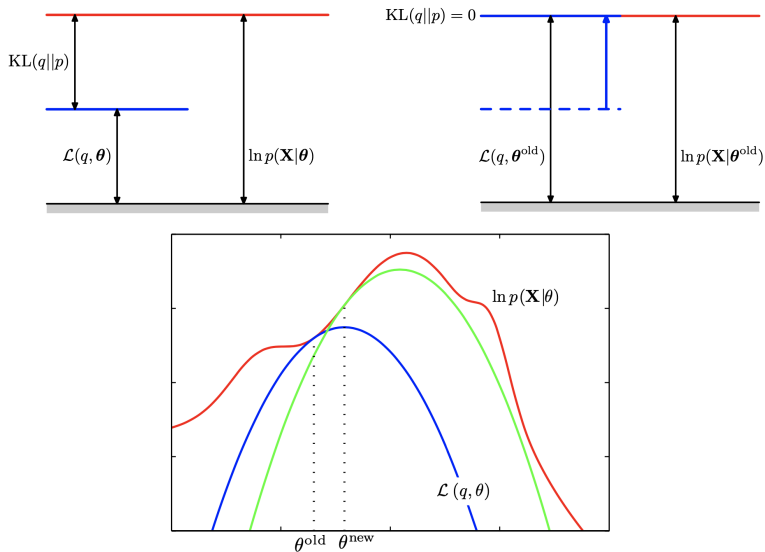
$$\begin{aligned}q^*(\mathbf{z}) &= \arg \max_q \mathcal{L}(q, \theta^*) = \\ &= \arg \min_q KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \theta^*)) = p(\mathbf{z}|\mathbf{x}, \theta^*);\end{aligned}$$

- ▶ **M-step** ( $\mathcal{L}(q, \theta) \rightarrow \max_\theta$ )

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

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

# EM-algorithm illustration





# Amortized variational inference

## E-step

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

- ▶  $q(\mathbf{z})$  approximates true posterior distribution  $p(\mathbf{z}|\mathbf{x}, \theta^*)$ , that is why it is called **variational posterior**;
- ▶  $p(\mathbf{z}|\mathbf{x}, \theta^*)$  could be **intractable**;
- ▶  $q(\mathbf{z})$  is different for each object  $\mathbf{x}$ .

## Idea

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  $\phi$ .

## Variational Bayes

- ▶ E-step

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

- ▶ M-step

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

# Variational EM-algorithm

## ELBO

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

### ► E-step

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

where  $\boldsymbol{\phi}$  – parameters of variational posterior distribution  $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$ .

### ► M-step

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

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

Now all we have to do is to obtain two gradients  $\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta})$ ,  $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta})$ .

**Challenge:** Number of samples  $n$  could be huge (we need to derive unbiased stochastic gradients).

# Summary

- ▶ Bayesian inference is a generalization of most common machine learning tasks. It allows to construct MLE, MAP and bayesian inference, to compare models complexity and many-many more cool stuff.
- ▶ LVM introduces latent representation of observed samples to make model more interpretable.
- ▶ LVM maximizes variational evidence lower bound (ELBO) to find MLE for the parameters.
- ▶ The general variational EM algorithm maximizes ELBO objective for LVM model to find MLE for parameters  $\theta$ .