

Deep Generative Models

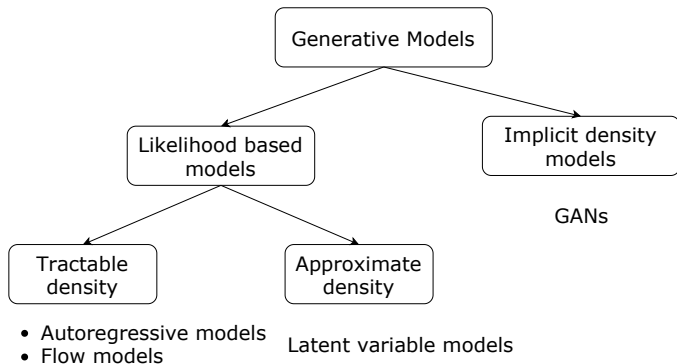
Lecture 3

Roman Isachenko

Ozon Masters

2021

Generative models zoo



Latent variable models

MLE problem

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

Challenge

$p(\mathbf{x}|\theta)$ could be intractable.

Extend probabilistic model

Introduce latent variable \mathbf{z} for each sample \mathbf{x}

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

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

Incomplete likelihood

MLE problem

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

Since \mathbf{Z} is unknown, maximize **incomplete likelihood**.

MILE problem

$$\begin{aligned}\theta^* &= \arg \max_{\theta} \log p(\mathbf{X} | \theta) = \arg \max_{\theta} \log \int p(\mathbf{X}, \mathbf{Z} | \theta) d\mathbf{Z} = \\ &= \arg \max_{\theta} \log \int p(\mathbf{X} | \mathbf{Z}, \theta) p(\mathbf{Z}) d\mathbf{Z}.\end{aligned}$$

Variational lower bound

ELBO

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

Instead of maximizing incomplete likelihood, maximize ELBO

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

EM-algorithm

- ▶ Initialize θ^* ;
- ▶ E-step

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

- ▶ M-step

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

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

Amortized variational inference

E-step

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

could be **intractable**.

Idea

Restrict the family of all possible distributions $q(\mathbf{z})$ to the particular parametric class conditioned of sample: $q(\mathbf{z}|\mathbf{x}, \phi)$.

Variational EM-algorithm

- ▶ 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

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

► 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 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 likelihood $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$.

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

Difficulty: number of samples n .

ELBO gradient (M-step, $\nabla_{\theta}\mathcal{L}(\phi, \theta)$)

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \theta) - KL(q(\mathbf{Z}|\mathbf{X}, \phi)||p(\mathbf{Z})) \rightarrow \max_{\phi, \theta}.$$

Optimization w.r.t. θ : **mini-batching** (1) + **Monte-Carlo** estimation (2)

$$\begin{aligned}\nabla_{\theta}\mathcal{L}(\phi, \theta) &= \sum_{i=1}^n \int q(\mathbf{z}_i|\mathbf{x}_i, \phi) \nabla_{\theta} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta) d\mathbf{z}_i \\ &\stackrel{(1)}{\approx} n \int q(\mathbf{z}_i|\mathbf{x}_i, \phi) \nabla_{\theta} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta) d\mathbf{z}_i, \quad i \sim U[1, n] \\ &\stackrel{(2)}{\approx} n \nabla_{\theta} \log p(\mathbf{x}_i|\mathbf{z}_i^*, \theta), \quad \mathbf{z}_i^* \sim q(\mathbf{z}_i|\mathbf{x}_i, \phi).\end{aligned}$$

Monte-Carlo estimation (2):

$$\int q(\mathbf{z})f(\mathbf{z})d\mathbf{z} \approx f(\mathbf{z}^*), \text{ where } \mathbf{z}^* \sim q(\mathbf{z}).$$

ELBO gradient (E-step, $\nabla_{\phi} \mathcal{L}(\phi, \theta)$)

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_q \log p(\mathbf{X}|\mathbf{Z}, \theta) - KL(q(\mathbf{Z}|\mathbf{X}, \phi) || p(\mathbf{Z})) \rightarrow \max_{\phi, \theta}.$$

Difference from M-step: density function $q(\mathbf{z}|\mathbf{x}, \phi)$ depends on the parameters ϕ , it is impossible to use Monte-Carlo estimation:

$$\nabla_{\phi} \mathcal{L}(\phi, \theta) = \int \nabla_{\phi} q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z} - \nabla_{\phi} KL$$

Log-derivative trick

$$\nabla_{\xi} q(\eta|\xi) = q(\eta|\xi) \left(\frac{\nabla_{\xi} q(\eta|\xi)}{q(\eta|\xi)} \right) = q(\eta|\xi) \nabla_{\xi} \log q(\eta|\xi).$$

$$\nabla_{\phi} q(\mathbf{Z}|\mathbf{X}, \phi) = q(\mathbf{Z}|\mathbf{X}, \phi) \nabla_{\phi} \log q(\mathbf{Z}|\mathbf{X}, \phi).$$

ELBO gradient (E-step, $\nabla_{\phi} \mathcal{L}(\phi, \theta)$)

$$\begin{aligned}\nabla_{\phi} \mathcal{L}(\phi, \theta) &= \int \nabla_{\phi} q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z} - \nabla_{\phi} KL = \\ &= \int q(\mathbf{Z}|\mathbf{X}, \phi) [\nabla_{\phi} \log q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta)] d\mathbf{Z} - \nabla_{\phi} KL\end{aligned}$$

After applying log-reparametrization trick, we are able to use Monte-Carlo estimation:

$$\begin{aligned}\nabla_{\phi} \mathcal{L}(\phi, \theta) &\approx n \nabla_{\phi} \log q(\mathbf{z}_i^*|\mathbf{x}_i, \phi) \log p(\mathbf{x}_i|\mathbf{z}_i^*, \theta) - \nabla_{\phi} KL, \\ \mathbf{z}_i^* &\sim q(\mathbf{z}_i|\mathbf{x}_i, \phi).\end{aligned}$$

Problem

Unstable solution with huge variance.

Solution

Reparametrization trick

ELBO gradient (E-step, $\nabla_{\phi} \mathcal{L}(\phi, \theta)$)

Reparametrization trick

$$f(\xi) = \int q(\eta|\xi) h(\eta) d\eta$$

Let $\eta = g(\xi, \epsilon)$, where g is a deterministic function, ϵ is a random variable with a density function $r(\epsilon)$.

$$\begin{aligned} \nabla_{\xi} \int q(\eta|\xi) h(\eta) d\eta &= \nabla_{\xi} \int r(\epsilon) h(g(\xi, \epsilon)) d\epsilon \\ &\approx \nabla_{\xi} h(g(\xi, \epsilon^*)), \quad \epsilon^* \sim r(\epsilon). \end{aligned}$$

Example

$$q(\eta|\xi) = \mathcal{N}(\eta|\mu, \sigma^2), \quad r(\epsilon) = \mathcal{N}(\epsilon|0, 1), \quad \eta = \sigma \cdot \epsilon + \mu, \quad \xi = [\mu, \sigma].$$

ELBO gradient (E-step, $\nabla_{\phi} \mathcal{L}(\phi, \theta)$)

$$\begin{aligned}\nabla_{\phi} \mathcal{L}(\phi, \theta) &= \nabla_{\phi} \int q(\mathbf{Z}|\mathbf{X}, \phi) \log p(\mathbf{X}|\mathbf{Z}, \theta) d\mathbf{Z} - \nabla_{\phi} KL \\ &\approx n \nabla_{\phi} \int r(\epsilon) \log p(\mathbf{x}_i | g(\mathbf{x}_i, \epsilon, \phi), \theta) d\epsilon - \nabla_{\phi} KL \\ &\approx n \nabla_{\phi} \log p(\mathbf{x}_i | g(\mathbf{x}_i, \epsilon^*, \phi), \theta) - \nabla_{\phi} KL, \quad \epsilon^* \sim r(\epsilon).\end{aligned}$$

Variational assumption

$$\begin{aligned}q(\mathbf{z}|\mathbf{x}, \phi) &= \mathcal{N}(\mu(\mathbf{x}), \Sigma(\mathbf{x})). \\ \mathbf{z} = g(\mathbf{x}, \epsilon, \phi) &= \sqrt{\Sigma(\mathbf{x})} \cdot \epsilon + \mu(\mathbf{x}).\end{aligned}$$

$\nabla_{\phi} KL(q(\mathbf{Z}|\mathbf{X}, \phi) || p(\mathbf{Z}))$ has an analytical solution.

Variational autoencoder (VAE)

Final algorithm

- ▶ pick $i \sim U[1, n]$;
- ▶ compute stochastic gradient w.r.t. ϕ

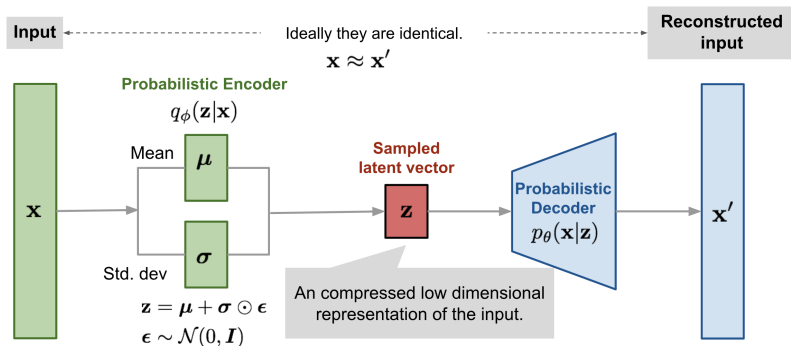
$$\nabla_{\phi} \mathcal{L}(\phi, \theta) = n \nabla_{\phi} \log p(\mathbf{x}_i | g(\mathbf{x}_i, \epsilon^*, \phi), \theta) - \nabla_{\phi} KL(q(\mathbf{z}_i | \mathbf{x}_i, \phi) || p(\mathbf{z}_i)), \quad \epsilon^* \sim r(\epsilon);$$

- ▶ compute stochastic gradient w.r.t. θ

$$\nabla_{\theta} \mathcal{L}(\phi, \theta) = n \nabla_{\theta} \log p(\mathbf{x}_i | \mathbf{z}_i^*, \theta), \quad \mathbf{z}_i^* \sim q(\mathbf{z}_i | \mathbf{x}_i, \phi);$$

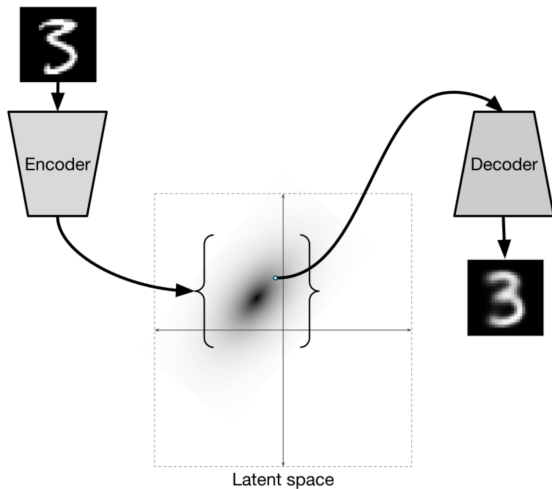
- ▶ update θ, ϕ according to the selected optimization method (SGD, Adam, RMSProp).

Variational autoencoder (VAE)



<https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html>

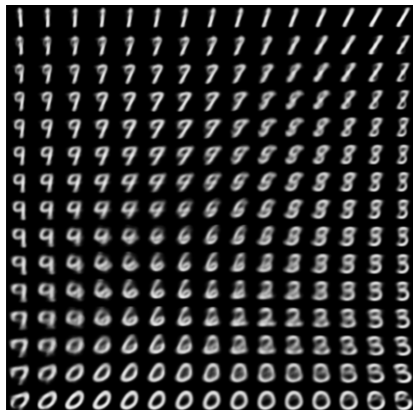
Variational Autoencoder



Isaac Dykeman, <http://ijdykeman.github.io/ml/2016/12/21/cvae.html>

Variational Autoencoder

Generation objects by sampling the latent space $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$



<http://bit.ly/2w73aXB>

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;
- ▶ $p(\mathbf{x})$ – evidence;
- ▶ $p(\mathbf{t})$ – prior;
- ▶ $p(\mathbf{t}|\mathbf{x})$ – posterior.

Variational Lower Bound

We are given the set of objects $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$. The goal is to perform bayesian inference on the latent variables $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^n$.

Empirical Lower BOund (ELBO)

$$\begin{aligned}\log p(\mathbf{X}) &= \log \frac{p(\mathbf{X}, \mathbf{T})}{p(\mathbf{T}|\mathbf{X})} = \\ &= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{p(\mathbf{T}|\mathbf{X})} d\mathbf{T} = \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})q(\mathbf{T})}{p(\mathbf{T}|\mathbf{X})q(\mathbf{T})} d\mathbf{T} = \\ &= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{q(\mathbf{T})} d\mathbf{T} + \int q(\mathbf{T}) \log \frac{q(\mathbf{T})}{p(\mathbf{T}|\mathbf{X})} d\mathbf{T} = \\ &= \mathcal{L}(q) + KL(q(\mathbf{T})||p(\mathbf{T}|\mathbf{X})) \geq \mathcal{L}(q).\end{aligned}$$

We would like to maximize lower bound $\mathcal{L}(q)$.

Mean field approximation

Independence assumption

$$q(\mathbf{T}) = \prod_{i=1}^k q_i(\mathbf{T}_i), \quad \mathbf{T} = [\mathbf{T}_1, \dots, \mathbf{T}_k], \quad \mathbf{T}_j = \{\mathbf{t}_{ij}\}_{i=1}^n, \quad \mathbf{t}_i = \{\mathbf{T}_{ij}\}_{j=1}^k.$$

Block coordinate optimization of ELBO for $q_j(\mathbf{T}_j)$

$$\begin{aligned} \mathcal{L}(q) &= \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{q(\mathbf{T})} d\mathbf{T} = \int \prod_{i=1}^k q_i(\mathbf{T}_i) \log \frac{p(\mathbf{X}, \mathbf{T})}{\prod_{i=1}^k q_i(\mathbf{T}_i)} \prod_{i=1}^k d\mathbf{T}_i = \\ &= \int \prod_{i=1}^k q_i \log p(\mathbf{X}, \mathbf{T}) \prod_{i=1}^k d\mathbf{T}_i - \sum_{i=1}^k \int \prod_{j=1}^k q_j \log q_i \prod_{i=1}^k d\mathbf{T}_i = \\ &= \int q_j \left[\int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i \right] d\mathbf{T}_j - \\ &\quad - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) \rightarrow \max_{q_j} \end{aligned}$$

Mean field approximation

Block coordinate optimization of ELBO for $q_j(\mathbf{T}_j)$

$$\begin{aligned}\mathcal{L}(q) &= \int q_j \left[\int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i \right] d\mathbf{T}_j - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) = \\ &= \int q_j \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) \rightarrow \max_{q_j},\end{aligned}$$

$$\text{where } \log \hat{p}(\mathbf{X}, \mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}(q_j)$$

$$\mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) = \int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i.$$

$$\begin{aligned}\mathcal{L}(q) &= \int q_j(\mathbf{T}_j) \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j(\mathbf{T}_j) \log q_j(\mathbf{T}_j) d\mathbf{T}_j + \text{const}(q_j) = \\ &= \int q_j(\mathbf{T}_j) \log \frac{\hat{p}(\mathbf{X}, \mathbf{T}_j)}{q_j(\mathbf{T}_j)} d\mathbf{T}_j + \text{const}(q_j) = \\ &= -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.\end{aligned}$$

Mean field approximation

Independence assumption

$$q(\mathbf{T}) = \prod_{i=1}^k q_i(\mathbf{T}_i), \quad \mathbf{T} = [\mathbf{T}_1, \dots, \mathbf{T}_k], \quad \mathbf{T}_j = \{\mathbf{t}_{ij}\}_{i=1}^n.$$

ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) \parallel \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

Solution

$$q_j(\mathbf{T}_j) = \hat{p}(\mathbf{X}, \mathbf{T}_j)$$

$$\log \hat{p}(\mathbf{X}, \mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

Mean field approximation

ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

Solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

Let assume the following factorization: $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2] = [\mathbf{Z}, \boldsymbol{\theta}]$, and restrict the class of probability distribution for $\boldsymbol{\theta}$ to Dirac delta functions:

$$q_2 = q(\mathbf{T}_2) = q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Under the restrictions the exact solution for q_2 is not reached.

Mean field approximation

General solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

Solution for $q_1 = q(\mathbf{Z})$

$$\begin{aligned} \log q(\mathbf{Z}) &= \int q(\boldsymbol{\theta}) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} = \\ &= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} = \\ &= \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) + \text{const}. \end{aligned}$$

EM-algorithm (E-step)

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

Mean field approximation

ELBO

$$\mathcal{L}(q) = -KL(q_j(\mathbf{T}_j) || \hat{p}(\mathbf{X}, \mathbf{T}_j)) + \text{const}(q_j) \rightarrow \max_{q_j}.$$

ELBO maximization w.r.t. $q_2 \equiv \theta_0$

$$\begin{aligned}\mathcal{L}(q_2) &= -KL(q(\boldsymbol{\theta}) || \hat{p}(\mathbf{X}, \boldsymbol{\theta})) + \text{const}(\boldsymbol{\theta}_0) \\&= \int q(\boldsymbol{\theta}) \log \frac{\hat{p}(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}_0) \\&= \int q(\boldsymbol{\theta}) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}_0) \\&= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} - \int \delta \log \delta d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}_0) \\&= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}_0)\end{aligned}$$

Mean field approximation

ELBO maximization w.r.t. $q_2 \equiv \theta_0$

$$\mathcal{L}(q_2) = \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const}(\boldsymbol{\theta}_0) = \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}_0).$$

$$\log \hat{p}(\mathbf{X}, \mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

$$\begin{aligned} \log \hat{p}(\mathbf{X}, \boldsymbol{\theta}) &= \mathbb{E}_{q_1} \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) + \text{const} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta}) d\mathbf{Z} + \log p(\boldsymbol{\theta}) + \text{const} \end{aligned}$$

EM-algorithm (M-step)

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \rightarrow \max_{\boldsymbol{\theta}}$$

Mean field approximation

Solution

$$\log q_j(\mathbf{T}_j) = \mathbb{E}_{i \neq j} \log p(\mathbf{X}, \mathbf{T}) + \text{const}$$

EM algorithm (special case)

- ▶ Initialize θ^* ;
- ▶ E-step

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

- ▶ M-step

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

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

Summary

- ▶ Latent variable models introduce latent variables to the initial probabilistic model to make distribution $p(\mathbf{x}|\boldsymbol{\theta})$ tractable.
- ▶ To solve the MLE problem LVM optimizes variational lower bound.
- ▶ EM-algorithm is an iterative algorithm which allows to optimize the variational lower bound.
- ▶ VAE model is a LVM, encoder is a variational distribution, decoder is a likelihood model.
- ▶ Mean field approximation is a general form of variational inference (EM-algorithm is just a special case).