Deep Generative Models
Lecture 3

Roman Isachenko

Ozon Masters

Spring, 2021

Recap of previous lecture

MLE problem

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

Challenge

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

LVM

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

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

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

Motivation

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

Recap of previous lecture

Incomplete likelihood maximization

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

Variational lower bound

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

Evidence Lower Bound (ELBO)

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

Instead of maximizing incomplete likelihood, maximize ELBO (equivalently minimize KL)

$$\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) \quad \rightarrow \quad \max_{q, \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \equiv \min_{q, \boldsymbol{\theta}} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})).$$

Recap of previous lecture

EM algorithm (block-coordinate optimization)

- Initialize $\boldsymbol{\theta}^*$;
- E-step
$$q(\mathbf{z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}^*) = \arg\min_q KL(q||p) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^*);$$
  - $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^*)$ could be **intractable**;
  - $q(\mathbf{z})$ is different for each object $\mathbf{x}$.
- M-step
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta});$$
- Repeat E-step and M-step until convergence.

Amortized variational inference

Restrict a family of all possible distributions $q(\mathbf{z})$ to a particular parametric class $q(\mathbf{z}|\mathbf{x}, \phi)$ conditioned on samples $\mathbf{x}$ with parameters ϕ.

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 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 heed to derive unbiased stochastic gradients).

Monte-Carlo estimation

$$\sum_{i=1}^{n} \mathbb{E}_q f(\mathbf{z}_i) \approx n \cdot \mathbb{E}_q f(\mathbf{z}) = n \cdot \int q(\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \approx n \cdot f(\mathbf{z}^*), \text{where } \mathbf{z}^* \sim q(\mathbf{z}).$$

ELBO gradients

$$\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\phi}, \boldsymbol{\theta}) \approx n \cdot \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}); \quad \nabla_{\boldsymbol{\phi}} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\phi}, \boldsymbol{\theta}) \approx n \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta})$$

ELBO

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

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

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

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

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

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

Challenge

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

$$\begin{aligned} \nabla_{\phi}\mathcal{L}(\phi, \boldsymbol{\theta}) &= \nabla_{\phi} \int q(\mathbf{z}|\mathbf{x}, \phi) \left[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \phi)\right] d\mathbf{z} \\ &\neq \int q(\mathbf{z}|\mathbf{x}, \phi) \nabla_{\phi} \left[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \phi)\right] d\mathbf{z} \end{aligned}$$

Solution

Reparametrization trick for $q(\mathbf{z}|\mathbf{x}, \phi)$ to allow the expectation is independent of parameters ϕ.

Reparametrization trick

$$f(\xi) = \mathbb{E}_{q(\eta|\xi)} h(\eta) = \int q(\eta|\xi) h(\eta) d\eta$$

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

$$f(\xi) = \int q(\eta|\xi) h(\eta) d\eta = \int r(\epsilon) h(g(\xi, \epsilon)) d\epsilon \approx h(g(\xi, \epsilon^*)), \quad \epsilon^* \sim r(\epsilon).$$

Examples

- $r(\epsilon) = \mathcal{N}(\epsilon|0, 1)$, $\eta = \sigma \cdot \epsilon + \mu$, $q(\eta|\xi) = \mathcal{N}(\eta|\mu, \sigma^2)$, $\xi = [\mu, \sigma]$.
- $\boldsymbol{\epsilon}^* \sim r(\boldsymbol{\epsilon}), \quad \mathbf{z} = g(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\phi}), \quad \mathbf{z} \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$

$$\begin{aligned}
\nabla_{\boldsymbol{\phi}} \int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) f(\mathbf{z}) d\mathbf{z} &= \nabla_{\boldsymbol{\phi}} \int r(\boldsymbol{\epsilon}) f(\mathbf{z}) d\boldsymbol{\epsilon} \\
&= \int r(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\phi}} f(g(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\phi})) d\boldsymbol{\epsilon} \approx \nabla_{\boldsymbol{\phi}} f(g(\mathbf{x}, \boldsymbol{\epsilon}^*, \boldsymbol{\phi}))
\end{aligned}$$

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

$$
\begin{aligned}
\nabla_{\phi}\mathcal{L}(\phi, \boldsymbol{\theta}) &= \nabla_{\phi} \int q(\mathbf{z}|\mathbf{x}, \phi) \left[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) - \log q(\mathbf{z}|\mathbf{x}, \phi)\right] d\mathbf{z} \\
&= \int r(\epsilon) \nabla_{\phi} \big[\log p(\mathbf{x}, g(\mathbf{x}, \epsilon, \phi)|\boldsymbol{\theta}) - \log q\big(g(\mathbf{x}, \epsilon, \phi)|\mathbf{x}, \phi\big)\big] d\epsilon \\
&\approx \nabla_{\phi} \big[\log p(\mathbf{x}, g(\mathbf{x}, \epsilon^*, \phi)|\boldsymbol{\theta}) - \log q\big(g(\mathbf{x}, \epsilon^*, \phi)|\mathbf{x}, \phi\big)\big]
\end{aligned}
$$

Variational assumption

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

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

Here $\boldsymbol{\mu}_{\phi}(\cdot), \boldsymbol{\sigma}_{\phi}(\cdot)$ are parameterized functions (outputs of neural network).

If we could calculate $\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ and $\log q(\mathbf{z}|\mathbf{x}, \phi)$, we are done. Could we?

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

$$\nabla_{\phi}\mathcal{L}(\phi, \boldsymbol{\theta}) \approx \nabla_{\phi}\Big[\log p(\mathbf{x}, g(\mathbf{x}, \boldsymbol{\epsilon}^*, \phi)|\boldsymbol{\theta}) - \log q\big(g(\mathbf{x}, \boldsymbol{\epsilon}^*, \phi)|\mathbf{x}, \phi\big)\Big]$$

First term

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

- $p(\mathbf{z})$ – prior distribution on latent variables $\mathbf{z}$. We could specify any distribution that we want. Let say $p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I})$.
- $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ - generative distibution. Since it parameterized function let it be neural network with parameters $\boldsymbol{\theta}$.

Second term

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

$$q(\mathbf{z}|\mathbf{x}, \phi) = r(\boldsymbol{\epsilon}) \cdot \left|\frac{\partial \boldsymbol{\epsilon}}{\partial \mathbf{z}}\right| \quad \Rightarrow \quad \log q(\mathbf{z}|\mathbf{x}, \phi) = \log r(\boldsymbol{\epsilon}) - \sum_{i=1}^{d} \log \left[\boldsymbol{\sigma}_{\phi}(\mathbf{x})\right]_i$$

Variational autoencoder (VAE)

Final algorithm

- pick $i \sim U[1, n]$;
- compute a stochastic gradient w.r.t. $\boldsymbol{\phi}$

$$
\begin{aligned}
\nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) \approx \nabla_{\boldsymbol{\phi}} \Big[\log p(\mathbf{x}, g(\mathbf{x}, \boldsymbol{\epsilon}^*, \boldsymbol{\phi}) | \boldsymbol{\theta}) - \\
- \log q\big(g(\mathbf{x}, \boldsymbol{\epsilon}^*, \boldsymbol{\phi}) | \mathbf{x}, \boldsymbol{\phi} \big) \Big], \quad \boldsymbol{\epsilon}^* \sim r(\boldsymbol{\epsilon});
\end{aligned}
$$

- compute a stochastic gradient w.r.t. $\boldsymbol{\theta}$

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

- update $\boldsymbol{\theta}, \boldsymbol{\phi}$ according to the selected optimization method (SGD, Adam, RMSProp):

$$
\begin{aligned}
\boldsymbol{\phi} &:= \boldsymbol{\phi} + \eta \nabla_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}), \\
\boldsymbol{\theta} &:= \boldsymbol{\theta} + \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}).
\end{aligned}
$$

Variational autoencoder (VAE)

- Encoder $q(\mathbf{z}|\mathbf{x}, \phi) = \text{NN}_e(\mathbf{x}, \phi)$ outputs $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}_\phi(\mathbf{x})$.
- Decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \text{NN}_d(\mathbf{z}, \boldsymbol{\theta})$ outputs parameters of the sample distribution.

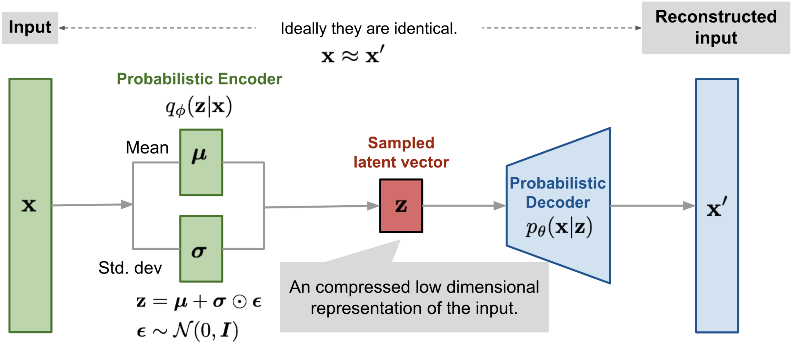

image credit: https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

Variational Autoencoder

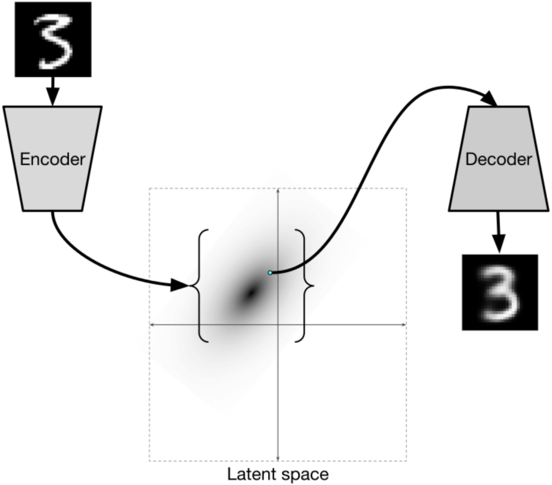

image credit: http://ijdykeman.github.io/ml/2016/12/21/cvae.html

Variational Autoencoder

Generated images for latent objects $\mathbf{z}$ sampled from prior $\mathcal{N}(0, \mathbf{I})$

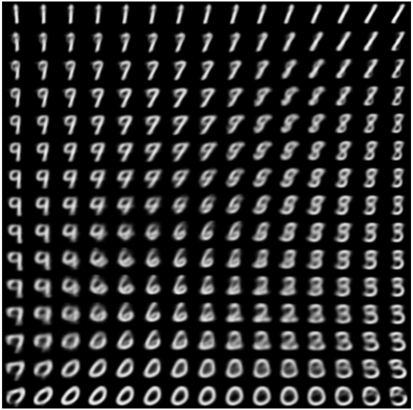

image credit: https://habr.com/ru/post/331664

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 $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^n$.

Evidence 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

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

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

$$\mathcal{L}(q) = \int q(\mathbf{T}) \log \frac{p(\mathbf{X}, \mathbf{T})}{q(\mathbf{T})} d\mathbf{T} = \int \prod_{i=1}^{k} q_i(\mathbf{T}_i) \log \frac{p(\mathbf{X}, \mathbf{T})}{\prod_{i=1}^{k} q_i(\mathbf{T}_i)} \prod_{i=1}^{k} d\mathbf{T}_i =$$

$$= \int \prod_{i=1}^{k} q_i \log p(\mathbf{X}, \mathbf{T}) \prod_{i=1}^{k} d\mathbf{T}_i - \sum_{i=1}^{k} \int \prod_{j=1}^{k} q_j \log q_i \prod_{i=1}^{k} d\mathbf{T}_i =$$

$$= \int q_j \left[\int \log p(\mathbf{X}, \mathbf{T}) \prod_{i \neq j} q_i d\mathbf{T}_i \right] d\mathbf{T}_j -$$

$$- \int q_j \log q_j d\mathbf{T}_j + \mathrm{const}(q_j) \to \max_{q_j}$$

Mean field approximation

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

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

$$= \int q_j \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j \log q_j d\mathbf{T}_j + \text{const}(q_j) \to \max_{q_j},$$

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

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

$$\mathcal{L}(q) = \int q_j(\mathbf{T}_j) \log \hat{p}(\mathbf{X}, \mathbf{T}_j) d\mathbf{T}_j - \int q_j(\mathbf{T}_j) \log q_j(\mathbf{T}_j) d\mathbf{T}_j + \text{const}(q_j) =$$

$$\int q_j(\mathbf{T}_j) \log \frac{\hat{p}(\mathbf{X}, \mathbf{T}_j)}{q_j(\mathbf{T}_j)} d\mathbf{T}_j + \text{const}(q_j) =$$

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

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

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

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

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

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

$$\log q(\mathbf{Z}) = \int q(\boldsymbol{\theta}) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} =$$

$$= \int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) d\boldsymbol{\theta} + \text{const} =$$

$$= \log p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta}_0) + \text{const.}$$

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

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

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)

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

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 $\boldsymbol{\theta}^*$;
- E-step

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

- M-step

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

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

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 introduce latent representation of observed samples to make model more interpretable.
- LVM maximizes variational evidence lower bound to find MLE of model parameters.
- ELBO maximization is performed by the general variational EM algorithm.
- Amortized inference allows to efficiently compute stochastic gradients for ELBO and to use deep neural networks for $q(\mathbf{z}|\mathbf{x}, \phi)$ and $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$.
- The VAE model is an LVM with an encoder network for $q(\mathbf{z}|\mathbf{x}, \phi)$ and a decoder network for $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$.

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 the variational lower bound.
- The EM-algorithm is an iterative algorithm which allows to optimize the variational lower bound.
- VAE model is an LVM, where the encoder gives the variational distribution, the decoder defines the likelihood model.
- The mean field approximation is a general form of variational inference (the EM-algorithm is just a special case).