Tutorial 8 Variational Inference
Olga Mikheeva

Excellent concise description of theory and Example of using mean field coordinate ascent to solve a GMM.

- We want to find: $p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$
- But: $$p(\boldsymbol{x}) = \int p(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{z}.$$
- is intractable.

Lets find an I-Projection

$$q^*(z) = argmin_{q(z) \in \mathcal{Q}} KL(q(z)||p(z|\boldsymbol{x}))$$

$$\begin{aligned} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}|\boldsymbol{x})] \\ &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned}$$

- Hard term is there again but now without any q.
- Evidence Lower Bound:

$$\begin{aligned} ELBO(q) &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] + E_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \end{aligned}$$

- Match prior + match data + reduce spread

Lets find an I-Projection

$$q^*(\boldsymbol{z}) = argmin_{q(\boldsymbol{z})\in\mathcal{Q}} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

$$\begin{aligned} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}|\boldsymbol{x})] \\ &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned}$$

$$\begin{aligned} ELBO(q) &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] + E_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \end{aligned}$$

$$\log p(\boldsymbol{x}) \;=\; ELBO(q) \;+\; KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

If x is set to a data value this is a log liklihood of that data

Mean Field Approximation

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_i(z_i)$$

$$ELBO(q) = E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})]$$

- Coordinate ascent says take one term, j, to maximize by plugging in and setting the variational derivative to 0:

(note to self: show on board)

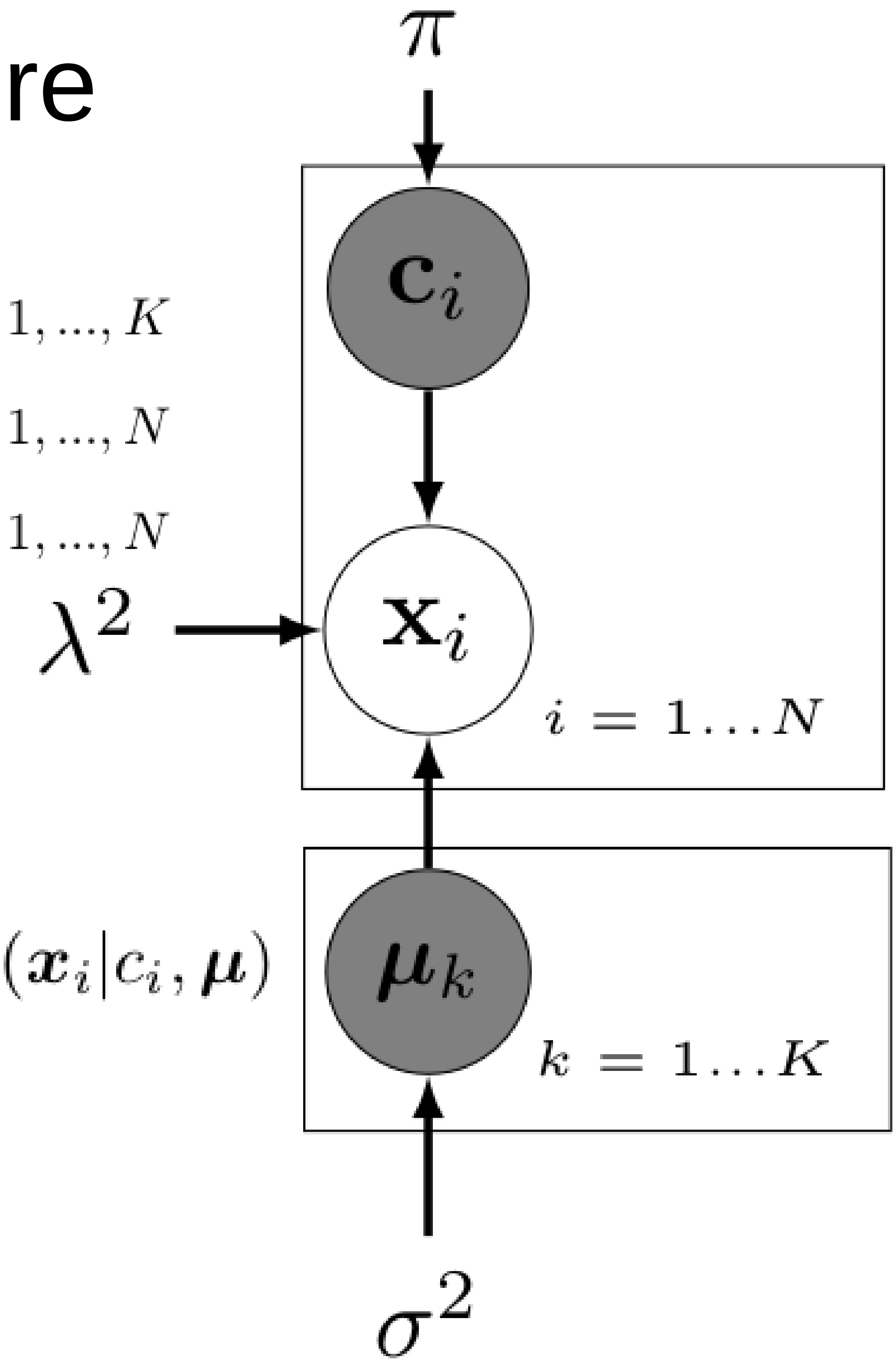
$$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j \mid \boldsymbol{z}_{-j}, \boldsymbol{x})]\}$$

Gaussian Mixture Model

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}) \qquad k = 1, ..., K$$

$$c_i \sim Categorical\Big(\frac{1}{K}, ..., \frac{1}{K}\Big) \qquad i = 1, ..., N$$

$$\boldsymbol{x}_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(c_i^T \boldsymbol{\mu}, \lambda^2 \boldsymbol{I}) \qquad i = 1, ..., N$$



$$p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{i=1}^{N} p(c_i) p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})$$

Categorical dist. = multinomial dist.

Mean Field

$$p(\boldsymbol{\mu}, \boldsymbol{c}) \approx q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(c_i)$$

$$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_k, s_k^2 \boldsymbol{I})$$

$$q(c_i) \sim Categorical(\boldsymbol{\phi}_i)$$

ELBO

Alert: remember this slide exists if you are dong the tutorial as you are asked to do this.

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \phi) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

$$\sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] = -\tfrac{1}{2} \sum_{k=1}^{K} [p \log(2\pi\sigma^2) + \int d\boldsymbol{\mu}\, \mathrm{N}(\mu_k, m_k, s_k^2 I)(\mu_k - \alpha)^2 / \sigma^2]$$

$$= -\tfrac{1}{2}\Big[K\rho \log(2\pi\sigma^2) + \sum_{k=1}^{K} m_k^2 + \rho S_k^2 + \alpha^2 - 2\boldsymbol{m}_k \bullet \boldsymbol{\alpha} \Big) / \sigma^2\Big]$$

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

$$\sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] = \quad \text{-N log K} \quad \text{: all terms are 1/K}$$

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

- Third term is sort of like first, but now we have $c_i{}^T\boldsymbol{\mu}$ together and λ replaces σ and more...
- That leads to a sum over k and a ϕ_{ik} factor.
- Also the sum of i is to N.

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

- Forth term is easy as it ends up as moments of a normal distribution. End up with an expression with p, K and s.

- Last term is easy too and will only involve the ϕ_{ik}