

Lecture 9

Variational Inference

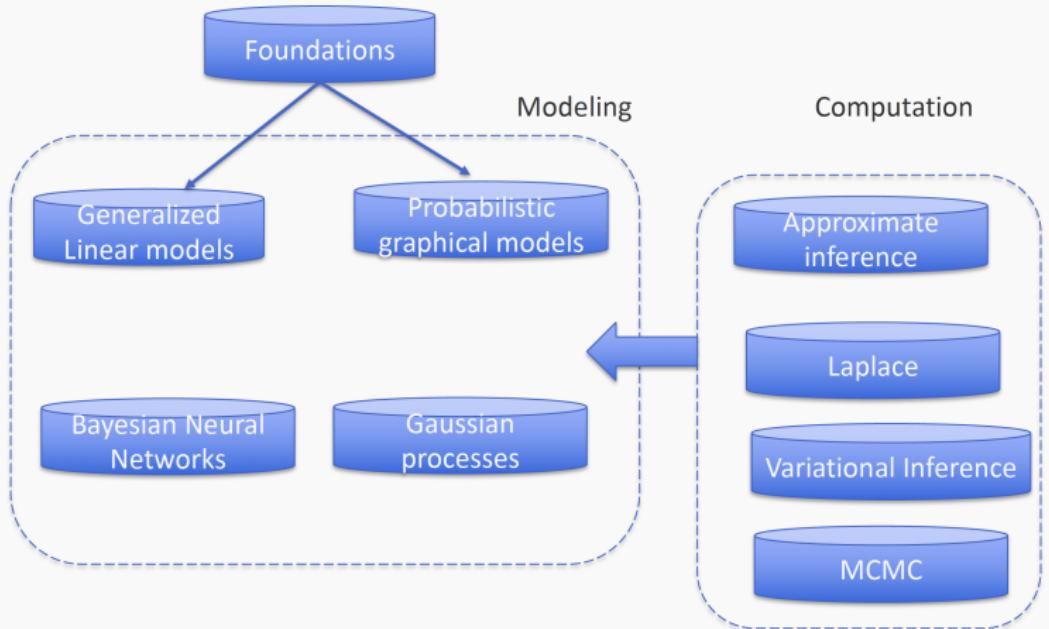
Instructor: Shibo Li

shiboli@cs.fsu.edu



Department of Computer Science
Florida State University

What we have learned so far



- We will discuss approximate inference schemes
- We will give examples about how to apply these schemes
- With these powerful computational tools in hand, we will discuss specific models as concrete examples

- Laplace approximation
- Variational Inference
- Markov Chain Monte-Carlo

- capture the key idea of these algorithms
- implement these approximate inference algorithms for a few well-known models!
- in the remaining section, see many practice tasks

- Gaussian Mixture Model and EM algorithm
- Variational Inference
 - Variational evidence lower bound
 - Mean-field variational inference
- Local variational inference
 - Convex conjugate
 - Logistic regression
 - Variational EM
- Variational message passing

- Gaussian Mixture Model and EM algorithm
- Global variational Inference
 - Mean-field variational inference
- Local variational inference
 - Logistic regression
- Variational message passing

- A probabilistic version of the k-means clustering algorithm
- Given a set of data points and a cluster number K , how do you group the data points into K clusters?
- Clustering is a fundamental data mining and pattern recognition task

K-means application

$K = 2$



$K = 3$



$K = 10$



Original image



- Given $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and cluster number K
- For each data point n
 - Sample the cluster membership $\mathbf{z}_n : K \times 1$ one-hot vector,
 $z_{nk}=1$ means \mathbf{x}_n belongs to cluster k

e.g., $[0, 0, 1, \dots, 0]^\top$

$$p(\mathbf{z}_n) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

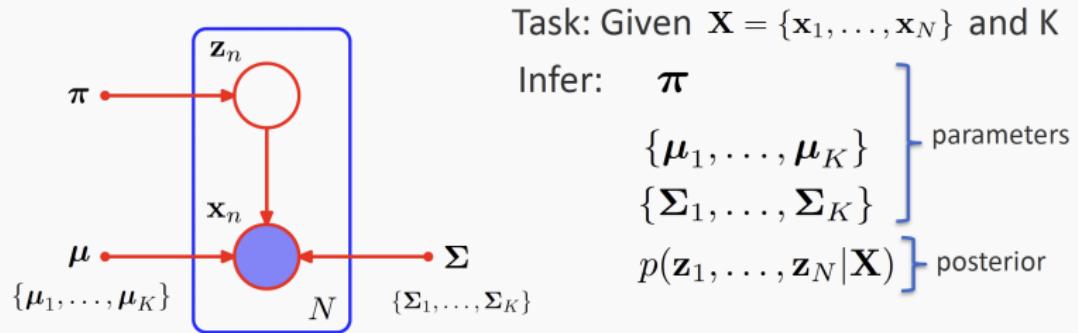
$$\begin{aligned}\boldsymbol{\pi} &= [\pi_1, \dots, \pi_K]^\top \\ \mathbf{z}_n &= [z_{n1}, \dots, z_{nK}]^\top\end{aligned}$$

- Given \mathbf{z}_n , sample \mathbf{x}_n from the cluster-specific Gaussian

$$p(\mathbf{x}_n | \mathbf{z}_n) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

cluster center

soft width



- Marginalize out \mathbf{z} and do MLE

$$p(\mathbf{x}_n) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



$$\ln p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$0 \leqslant \pi_k \leqslant 1$$

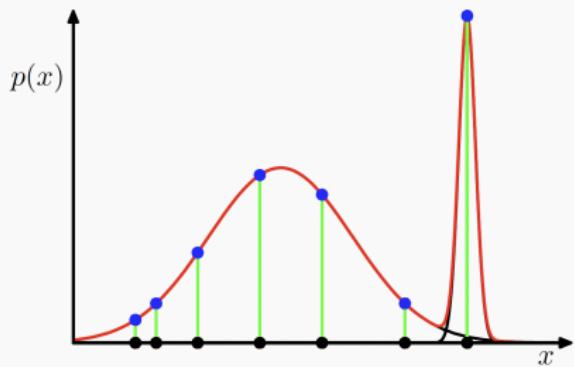
$$\text{s.t.} \quad \sum_{k=1}^K \pi_k = 1$$

- Given the parameters, we calculate the posterior of the cluster membership

$$p(\mathbf{z}_n | \mathbf{X}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Why ? Leave it as your exercise

- Singularity issues



$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

$$\sigma_j \rightarrow 0.$$

Suppose we use diagonal covariance, when the cluster center is close to a data point, it tends to collapse onto the point to increase the likelihood

- Can we get rid of the singularity issue?
- Can we jointly estimate the parameters and the posterior?



Suppose we have a model governed by parameters θ

$$p(\mathbf{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$$

Observations Latent random variables

The diagram illustrates a probabilistic graphical model. At the top is a red dot labeled θ . A red arrow points down from θ to a white circle labeled Z . Another red arrow points down from Z to a blue circle labeled X .

Question: what are θ and Z for GMMs?

$$\begin{aligned}\log(p(\mathbf{X}|\boldsymbol{\theta})) &= \log\left(\int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})d\mathbf{Z}\right) \\ &= \log\left(\int \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}q(\mathbf{Z})d\mathbf{Z}\right) \\ &\geq \int q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right)d\mathbf{Z}\end{aligned}$$

Jensen's inequality

$$\begin{aligned}\log(p(\mathbf{X}|\boldsymbol{\theta})) &= \underbrace{\int q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right)d\mathbf{Z}}_{L(\boldsymbol{\theta}, q(\mathbf{Z}))} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))} \\ &\geq 0\end{aligned}$$

$$\log(p(\mathbf{X}|\boldsymbol{\theta})) = \underbrace{\int q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right) d\mathbf{Z}}_{L(\boldsymbol{\theta}, q(\mathbf{Z}))} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))} \geq 0$$

$$\log(p(\mathbf{X}|\boldsymbol{\theta})) = L(\boldsymbol{\theta}, q^*(\mathbf{Z})) \text{ when } q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

Now fix $q^*(\mathbf{Z})$

$$\boldsymbol{\theta}^{\text{new}} = \underset{\widehat{\boldsymbol{\theta}}}{\operatorname{argmax}} L(\widehat{\boldsymbol{\theta}}, q^*(\mathbf{Z}))$$

$$\log(p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}})) \geq L(\boldsymbol{\theta}^{\text{new}}, q^*(\mathbf{Z})) \geq L(\boldsymbol{\theta}, q^*(\mathbf{Z})) = \log(p(\mathbf{X}|\boldsymbol{\theta}))$$

Like a bridge to improve the parameters!

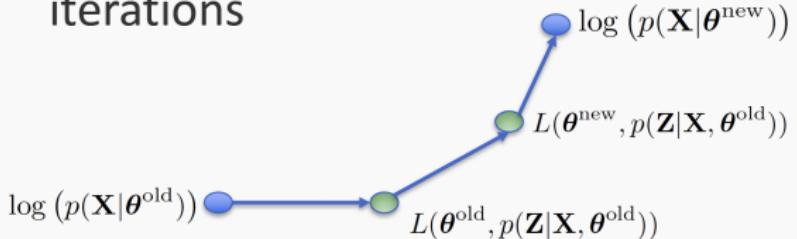
- Choose an initial setting θ^{new}
- Repeat

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

Evaluate $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ E step

Fix $q(\mathbf{Z})$, $\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}} L(\theta, q(\mathbf{Z}))$ M step

- Until $\|\theta^{\text{old}} - \theta^{\text{new}}\| \leq \epsilon$ or reach the maximum # of iterations



- Given $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and cluster number K
- For each data point n
 - Sample the cluster membership $\mathbf{z}_n : K \times 1$ one-hot vector,
 $z_{nk}=1$ means \mathbf{x}_n belongs to cluster k

e.g., $[0, 0, 1, \dots, 0]^\top$

$$p(\mathbf{z}_n) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

$$\begin{aligned}\boldsymbol{\pi} &= [\pi_1, \dots, \pi_K]^\top \\ \mathbf{z}_n &= [z_{n1}, \dots, z_{nK}]^\top\end{aligned}$$

- Given \mathbf{z}_n , sample \mathbf{x}_n from the cluster-specific Gaussian

$$p(\mathbf{x}_n | \mathbf{z}_n) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

cluster center

soft width

$$p(\mathbf{X}, \mathbf{Z} | \underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}_{\boldsymbol{\theta}}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

E step: $q(\mathbf{Z}) = \prod_{n=1}^N p(\mathbf{z}_n | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \prod_{n=1}^N \prod_{k=1}^K \gamma_{nk}^{z_{nk}}$

$$\gamma_{nk} \equiv p(\mathbf{z}_{nk} = 1 | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}})}$$

$$p(\mathbf{X}, \mathbf{Z} | \underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}_{\boldsymbol{\theta}}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

M step: $L(\boldsymbol{\theta}, q(\mathbf{Z})) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{q(\mathbf{Z})} \right)$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \right) + \text{const}$$
$$= \mathbb{E}_{q(\mathbf{Z})} \log \left(p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \right) + \text{const}$$
$$\mathbb{E}_{q(\mathbf{Z})} \left[\sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k + z_{nk} \log \left(\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right]$$

$$\sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \pi_k + \gamma_{nk} \log \left(\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

- M step:

$$\max \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \pi_k + \gamma_{nk} \log (\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^N \gamma_{nk}}{\sum_{k=1}^K \sum_{n=1}^N \gamma_{nk}}$$



$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \mathbf{x}_n \quad N_k = \sum_{n=1}^N \gamma_{nk}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top$$

- E step $\gamma_{nk} \equiv p(\mathbf{z}_{nk} = 1 | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}})}$

- M step

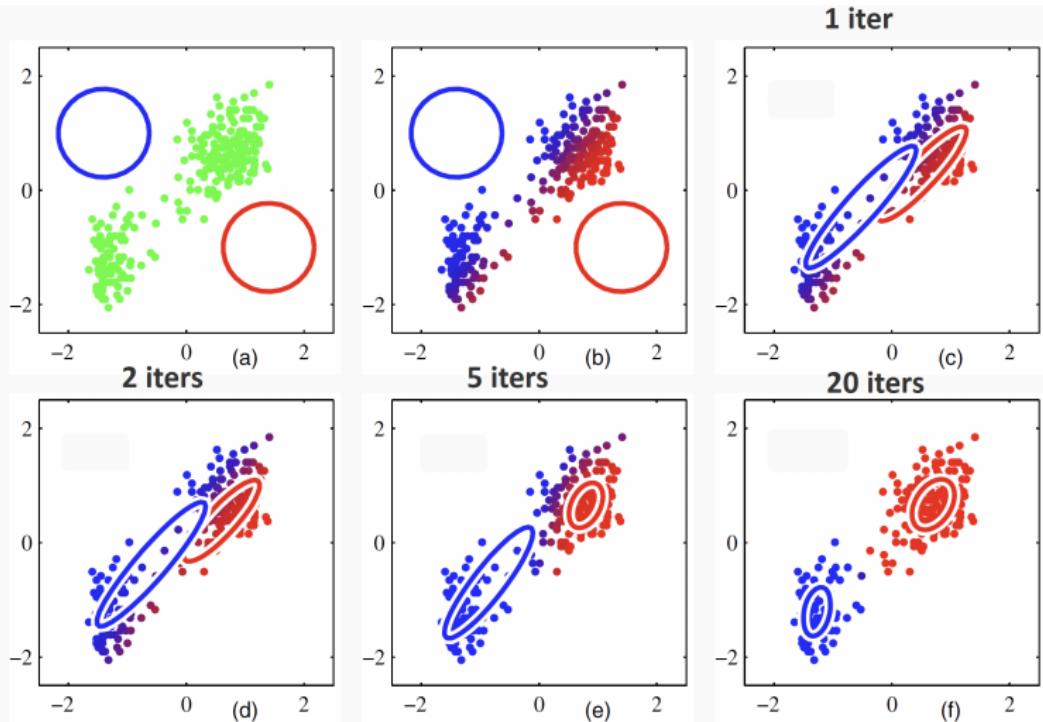
$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^N \gamma_{nk}}{\sum_{k=1}^K \sum_{n=1}^N \gamma_{nk}}$$

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\top}$$

We do not have any singularity issues!

EM algorithm for GMM



- Derive EM algorithm for mixture of Bernoulli distributions
- Derive EM algorithm for Bayesian linear regression

- Gaussian Mixture Model and EM algorithm
- **Global variational Inference**
 - Local variational inference
 - Variational message passing

- Consider a general model

$$p(\mathbf{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$$

Observations Latent random variables



Put aside the parameters first (either we use full Bayesian treatment to absorb θ into Z or consider θ as fixed constant)

Question: how to compute the posterior $p(\mathbf{Z}|\mathbf{X})$

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

Usually infeasible!!

GMMs, Bayesian linear regression are a few exceptions...

In most cases, you cannot get an analytical result

$$p(z) = \mathcal{N}(z|0, 1)$$

e.g.,

$$p(x|z) = \sigma(z)^x (1 - \sigma(x))^{1-z}$$

Question: how to compute the posterior $p(\mathbf{Z}|\mathbf{X})$

Idea: Now that the true posterior is complicated and tricky to compute, can we find *a simple form of distribution* (e.g., Gaussian) that *approximates* the true posterior? In other words, can we designate *a family of simple distributions*, from which *we find the best member that is closest to the true posterior?*

$$p(z) = \mathcal{N}(z|0, 1) \quad p(x|z) = \sigma(z)^x (1 - \sigma(x))^{1-z}$$

Let us use a Gaussian $q(z) = N(z|\mu, \sigma^2)$ to approximate the true posterior $p(z|x)$

The problem is how to determine the best μ, σ^2

- Suppose we assume the family (form) of approximate posterior $q(\mathbf{Z}|\boldsymbol{\alpha})$

$$\boldsymbol{\alpha}^* = \operatorname{argmin}_{\boldsymbol{\alpha}} \text{KL}(q(\mathbf{Z}|\boldsymbol{\alpha}) \| p(\mathbf{Z}|\mathbf{X}))$$

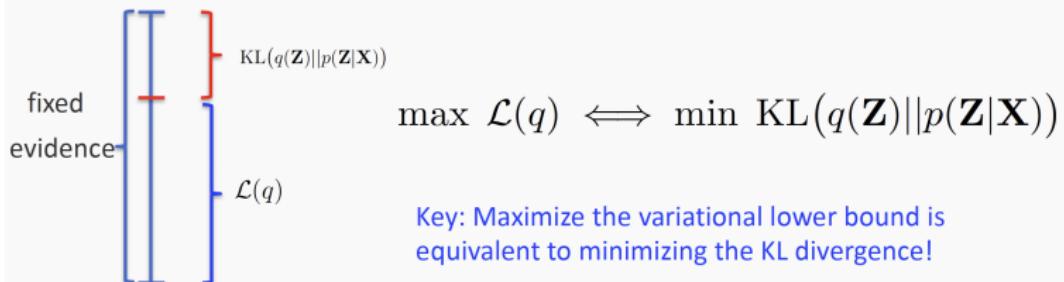


KL divergence is commonly used to measure the difference between distributions

What is the issue?

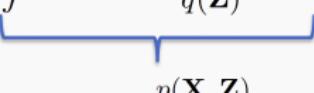
$$\log(p(\mathbf{X})) = \underbrace{\int q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\right\} d\mathbf{Z}}_{\text{Evidence}} + \underbrace{\int q(\mathbf{Z}) \log\left\{\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\right\} d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})) \geq 0}$$

Variational Lower Bound



- Given a family S of the approximate posterior $q(\mathbf{Z})$,

$$q^*(\mathbf{Z}) = \underset{q \in S}{\operatorname{argmax}} \mathcal{L}(q) = \int q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$


$$\mathbb{E}_q \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\}$$

Usually there is a trade-off: The larger the family S , the better the approximation quality, but the harder the optimization

- Assume the approximate posterior is factorized:

$$q(\mathbf{Z}) = \prod_i q(\mathbf{Z}_i)$$

$\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots\}$ nonoverlapping

Each $q(\mathbf{Z}_i)$ is a free form distribution

$$\max \mathcal{L}(q) = \int \prod_i q(\mathbf{Z}_i) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{\prod_i q(\mathbf{Z}_i)} \right\} d\mathbf{Z}$$

Solve this optimization by alternative updating

$$\max \mathcal{L}(q) = \int \prod_i q(\mathbf{Z}_i) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{\prod_i q(\mathbf{Z}_i)} \right\} d\mathbf{Z}$$

Update $q(\mathbf{Z}_j)$ giving $\{q(\mathbf{Z}_i)\}_{i \neq j}$ fixed

$$\mathcal{L}(q(\mathbf{Z}_j)) = \int q(\mathbf{Z}_j) \underbrace{\prod_{i \neq j} q(\mathbf{Z}_i) \log(p(\mathbf{X}, \mathbf{Z}))}_{\mathbb{E}_{q(\mathbf{Z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z}))} d\mathbf{Z} - \int q(\mathbf{Z}_j) \log(q(\mathbf{Z}_j)) d\mathbf{Z}_j + \text{const}$$

$$q(\mathbf{Z}_{-j}) = \prod_{i \neq j} q(\mathbf{Z}_i)$$

Solve this $q(\mathbf{Z}_j) \propto \exp \left\{ \mathbb{E}_{q(\mathbf{Z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\}$

$$q(\mathbf{Z}_j) = \frac{\exp \left\{ \mathbb{E}_{q(\mathbf{Z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\}}{\int \exp \left\{ \mathbb{E}_{q(\mathbf{Z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\} d\mathbf{Z}_j}$$

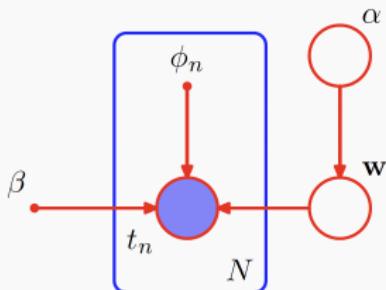
- Choose a factorized posterior form $q(\mathbf{Z}) = \prod_i q(\mathbf{z}_i)$
- Repeat
 - For each j
 - Fixed $\{q(\mathbf{z}_i)\}_{i \neq j}$ and update $q(\mathbf{z}_j) \propto \exp\{\mathbb{E}_{q(\mathbf{z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z}))\}$
 - End for
- Until each $q(\mathbf{z}_j)$ changes little or reach maximum # of iterations

$$p(\alpha) = \text{Gam}(\alpha | a_0, b_0)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^\top \phi_n, \beta^{-1})$$

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$$



Observed data $\mathbf{t} = [t_1, \dots, t_N]^\top$

Inference task: $p(\mathbf{w}, \alpha | \mathbf{t})$

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$

Alternative updating

$$q(\alpha) \propto \exp\{\mathbb{E}_{q(\mathbf{w})} \log p(\mathbf{t}, \mathbf{w}, \alpha)\}$$



$$q(\alpha) = \text{Gam}(\alpha | a_N, b_N)$$

$$a_N = a_0 + \frac{d}{2} \quad \text{d: dimension of } \mathbf{w}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\mathbf{w}^\top \mathbf{w}]$$

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$

Alternative updating

$$q(\mathbf{w}) \propto \exp\{\mathbb{E}_{q(\alpha)} \log p(\mathbf{t}, \mathbf{w}, \alpha)\}$$



$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}$$

$$\mathbf{S}_N = (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi)^{-1}$$

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$

The required moments

$$\begin{aligned}\mathbb{E}[\alpha] &= a_N/b_N \\ \mathbb{E}[\mathbf{w}\mathbf{w}^T] &= \mathbf{m}_N\mathbf{m}_N^T + \mathbf{S}_N\end{aligned}$$

Predictive distribution

$$\begin{aligned}p(t|\mathbf{x}, \mathbf{t}) &= \int p(t|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathbf{t}) d\mathbf{w} \simeq \int p(t|\mathbf{x}, \mathbf{w})q(\mathbf{w}) d\mathbf{w} \\ &= \mathcal{N}(t|\mathbf{m}_N^T \phi(\mathbf{x}), \sigma^2(\mathbf{x}))\end{aligned}$$

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0)g(\boldsymbol{\eta})^{\nu_0} \exp \left\{ \nu_0 \boldsymbol{\eta}^T \boldsymbol{\chi}_0 \right\}$$

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n, \mathbf{z}_n)g(\boldsymbol{\eta}) \exp \left\{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\}$$

Task: $p(\boldsymbol{\eta}, \mathbf{Z}|\mathbf{X})$

Assume: $q(\mathbf{Z}, \boldsymbol{\eta}) = q(\mathbf{Z})q(\boldsymbol{\eta})$

The updates are analytical

$$q^*(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g(\mathbb{E}[\boldsymbol{\eta}]) \exp \left\{ \mathbb{E}[\boldsymbol{\eta}^T] \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\}$$

$$q^*(\boldsymbol{\eta}) = f(\nu_N, \boldsymbol{\chi}_N) g(\boldsymbol{\eta})^{\nu_N} \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{\chi}_N \right\}$$

$$\begin{aligned}\nu_N &= \nu_0 + N \\ \boldsymbol{\chi}_N &= \boldsymbol{\chi}_0 + \sum_{n=1}^N \mathbb{E}_{\mathbf{z}_n} [\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)]\end{aligned}$$

- Gaussian Mixture Model and EM algorithm
- Global variational Inference
- Local variational inference
- Variational message passing

- Seeks an bound for a factor function of individual variables or a subset of variables
- Convex conjugate

$$f(x) = \max_{\lambda} \quad \lambda x - g(\lambda)$$

$$g(\lambda) = \max_{\mathbf{x}} \quad \lambda x - f(x)$$

- Key idea: if a factor is convex, use the convex conjugate obtain an bound (easier form)

In general

$$\log(p(\mathbf{X}, \boldsymbol{\theta})) = \log(p(\boldsymbol{\theta})) + \sum_n \log(p(\mathbf{x}_n | \boldsymbol{\theta}))$$

If it is convex to $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}) \geq \boldsymbol{\lambda}^\top \boldsymbol{\theta} - g(\boldsymbol{\lambda})$

- Let us consider the sigmoid function in the likelihood

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$\log(\sigma(x)) = -\log(1 + e^{-x})$ is concave

Let's verify it

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\begin{aligned}\log(\sigma(x)) &= -\log(1 + e^{-x}) \\ &= -\log \{e^{-x/2}(e^{x/2} + e^{-x/2})\} \\ &= x/2 - \log \{e^{x/2} + e^{-x/2}\} \\ &\quad \underbrace{_{f(x)}}\end{aligned}$$

$$f(x) = -\ln(e^{x/2} + e^{-x/2}) \quad \text{Symmetric}$$

$$f(x) = f(|x|) = f(\sqrt{x^2})$$

We can show that f is convex to x^2 , so we can use convex conjugate

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f(\sqrt{x^2}) \right\}$$

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f \left(\sqrt{x^2} \right) \right\}$$



$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$0 = \lambda - \frac{dx}{dx^2} \frac{d}{dx} f(x) = \lambda + \frac{1}{4x} \tanh \left(\frac{x}{2} \right)$$



ξ is the optimal x corresponding to λ

$$\lambda(\xi) = -\frac{1}{4\xi} \tanh \left(\frac{\xi}{2} \right) = -\frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2} \right]$$

Note: $\lambda(\xi) = \lambda(-\xi)$

Variational logistic regression

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f\left(\sqrt{x^2}\right) \right\}$$

$$g(\lambda) = \lambda(\xi)\xi^2 - f(\xi) = \lambda(\xi)\xi^2 + \ln(e^{\xi/2} + e^{-\xi/2})$$



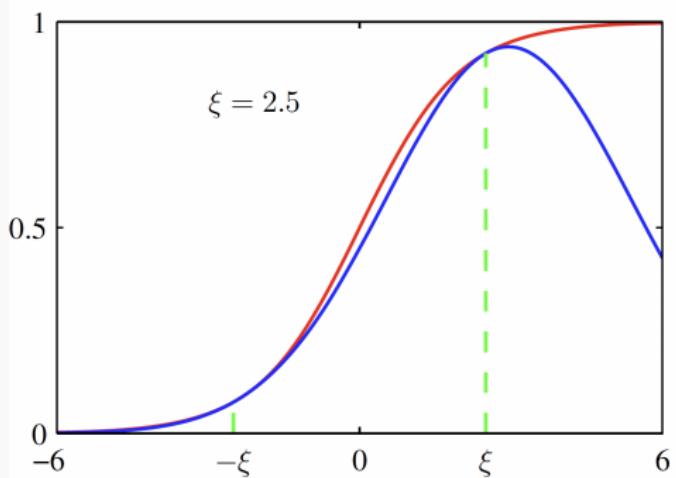
$$f(x) \geq g(\lambda) = \lambda x^2 - \lambda \xi^2 - \ln(e^{\xi/2} + e^{-\xi/2})$$



$$\log(\sigma(x)) = x/2 + f(x)$$

$$\sigma(x) \geq \sigma(\xi) \exp\{(x - \xi)/2 + \lambda(\xi)(x^2 - \xi^2)\}$$

Lower-bound of $\sigma(x)$



- Given an arbitrary feature vector ϕ , the binary response t is sampled from

$$p(t|\mathbf{w}) = \sigma(a)^t \{1 - \sigma(a)\}^{1-t} \quad a = \mathbf{w}^T \phi$$

$$\begin{aligned} &= \left(\frac{1}{1 + e^{-a}} \right)^t \left(1 - \frac{1}{1 + e^{-a}} \right)^{1-t} \\ &= e^{at} \frac{e^{-a}}{1 + e^{-a}} = e^{at} \sigma(-a) \end{aligned}$$

From the previous result

$$\sigma(z) \geq \sigma(\xi) \exp \left\{ (z - \xi)/2 - \lambda(\xi)(z^2 - \xi^2) \right\}$$

where $\lambda(\xi) = \frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2} \right]$

Note: We omit – in
the previous symbol



$$p(t|\mathbf{w}) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}$$

Variational logistic regression

$$p(t|\mathbf{w}) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}$$

Given the design matrix (features after appropriate (nonlinear) transformations)

Binary output $\mathbf{t} = [t_1, \dots, t_N]^\top$
Each $t_n \in \{0, 1\}$

$$\Phi = [\phi_1, \dots, \phi_N]^\top$$

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) d\mathbf{w} = \int \left[\prod_{n=1}^N p(t_n|\mathbf{w}) \right] p(\mathbf{w}) d\mathbf{w}$$



$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geq h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w})$$

$$\begin{aligned} h(\mathbf{w}, \boldsymbol{\xi}) &= \prod_{n=1}^N \sigma(\xi_n) \exp \left\{ \mathbf{w}^\top \phi_n t_n - (\mathbf{w}^\top \phi_n + \xi_n)/2 \right. \\ &\quad \left. - \lambda(\xi_n)([\mathbf{w}^\top \phi_n]^2 - \xi_n^2) \right\}. \end{aligned}$$

- Consider approximate posterior $q(\mathbf{w})$

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})} \right\}$$





$$\min \text{KL}(q(\mathbf{w}) \| p(\mathbf{w}|\mathbf{t}))$$

Infeasible! Also, if you directly optimize w.r.t a free
from, you obtain the true posterior

Solution: We maximize its variational lower bound!

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geq h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w})$$

Why called ``variational LB''?

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w})h(\mathbf{w}, \boldsymbol{\xi})}{q(\mathbf{w})} \right\}$$

It's possible to
take equality

Variational logistic regression

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \boldsymbol{\xi})}{q(\mathbf{w})} \right\}$$

The same as Mean-Field

$$q(\mathbf{w}) \propto \exp(\log\{p(\mathbf{w}) h(\mathbf{w}, \boldsymbol{\xi})\}) \quad p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

$$-\frac{1}{2} (\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)$$

$$+ \sum_{n=1}^N \left\{ \mathbf{w}^T \boldsymbol{\phi}_n (t_n - 1/2) - \lambda(\xi_n) \mathbf{w}^T (\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T) \mathbf{w} \right\} + \text{const}$$

Complete squares

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$
$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \sum_{n=1}^N (t_n - 1/2) \boldsymbol{\phi}_n \right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + 2 \sum_{n=1}^N \lambda(\xi_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T. \end{aligned}$$

- No, we haven't identified the variational parameters ξ

$$\max_{\xi} \quad \mathcal{L}(q, \xi) = \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \xi)}{q(\mathbf{w})} \right\}$$

$$\sum_{n=1}^N \left\{ \ln \sigma(\xi_n) - \xi_n/2 - \lambda(\xi_n) (\phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n - \xi_n^2) \right\} + \text{const}$$

$$0 = \lambda'(\xi_n) (\phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n - \xi_n^2)$$

$$(\xi_n^{\text{new}})^2 = \phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n = \phi_n^T (\mathbf{S}_N + \mathbf{m}_N \mathbf{m}_N^T) \phi_n$$

- We conduct an EM procedure

$$\max \quad \mathcal{L}(q, \xi) = \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w})h(\mathbf{w}, \xi)}{q(\mathbf{w})} \right\}$$

E step: update $q(\mathbf{w})$

M step: update ξ

Alternatively maximize the variational lower bound

- The variational bound is variational transformation, it means, if you do NOT restrict the range of the variational parameters, they always have settings that reach equality

$$\log(p(\mathbf{t})) \geq \mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})}\right\}$$

$$\mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})}\right\} \geq \mathcal{L}(q, \epsilon) = \mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{w})h(\mathbf{w}, \boldsymbol{\xi})}{q(\mathbf{w})}\right\}$$

In practice, we often restrict the family/range of the variational parameters to gain the computational easiness

- In general, if we also need to estimate hyperparameters.

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$$

$$\max_{q(\mathbf{Z}), \boldsymbol{\theta}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{Z})} \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

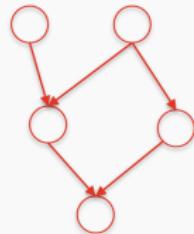
E step: $q(\mathbf{Z}) \leftarrow \operatorname{argmax}_{q \in \mathcal{S}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta})$ fix $\boldsymbol{\theta}$



M step: $\boldsymbol{\theta} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta})$ fix $q(\mathbf{Z})$

- Consider a Bayesian network

$$p(\mathbf{x}) = \prod_i p(\mathbf{x}_i | \text{pa}(\mathbf{x}_i))$$



- Assume a factorized posterior over the nodes

$$q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$$

- Consider the update on each node

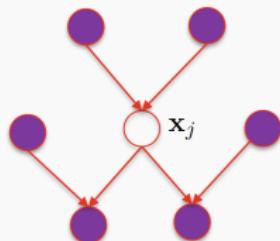
$$q(\mathbf{x}_j) \propto \exp\{\mathbb{E}_{q(\mathbf{x}_{\neg j})}[\log p(\mathbf{x})]\}$$

Questions: which factors involve \mathbf{x}_j ?

The conditional probabilities where \mathbf{x}_j is a parent/child

$$\begin{aligned} q(\mathbf{x}_j) &\propto \exp\{\mathbb{E}[\log p(\mathbf{x}_j | \text{pa}(\mathbf{x}_j))]\} \\ &+ \sum_{\mathbf{x}_j \in \text{pa}(\mathbf{x}_t)} \mathbb{E}[\log p(\mathbf{x}_t | \mathbf{x}_j, \text{pa}(\mathbf{x}_t) \setminus \{\mathbf{x}_j\})] \end{aligned}$$

Markov blanket



- Given a graphical model, the mean-field variational update only requires a local computation on the graph

- What is EM algorithm
- $\text{Log}(\text{Evidence}) = \text{Variational Lower Bound} + \text{KL}$
- What is EM algorithm
- Global variational inference, mean-field
- General update in exponential family
- Local variational inference, convex conjugate
- Variational message passing
- Being able to derive and implement variational inference