Expectation-Maximization Algorithm

Qinliang Su （苏勤亮）

Sun Yat-sen University

suqliang@mail.sysu.edu.cn

General Form of the Problem

- Given the joint distribution

$$p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}),$$

where $\boldsymbol{x}$ is the observed variable and $\boldsymbol{z}$ is the latent variable, we need to maximize the log likelihood *w.r.t.* $\boldsymbol{x}$, that is,

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x}; \boldsymbol{\theta}),$$

where

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$$

What we have is the joint pdf $p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$, but what we need to optimize is the marginal pdf $p(\boldsymbol{x}; \boldsymbol{\theta})$

Outline

- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models
- EM Variants

EM Algorithm

- Algorithm

  E-step: Evaluating the expectation

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

  M-step: Updating the parameter

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

- Key integrant in EM

  1) The posteriori distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

  2) The expectation of joint distribution $\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ *w.r.t.* the posteriori

  3) Maximization

Outline

- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models
- EM Variants

Re-representing the Log-likelihood

- The log-likelihood can be reformulated as

$\forall$ distribution $q(\boldsymbol{z})$

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log p(\boldsymbol{x})$$

$$= \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})q(\boldsymbol{z})}$$

$$= \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})}}_{\mathcal{L}(q,\boldsymbol{\theta})} + \underbrace{\sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})}}_{KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}))}$$

$$\boxed{= \mathcal{L}(q,\boldsymbol{\theta}) + KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta})), \qquad \text{for } \forall\ \boldsymbol{\theta}, q(\boldsymbol{z})}$$

Remark: The KL-divergence is used to *measure the distance* between two distributions q and p, which is defined as

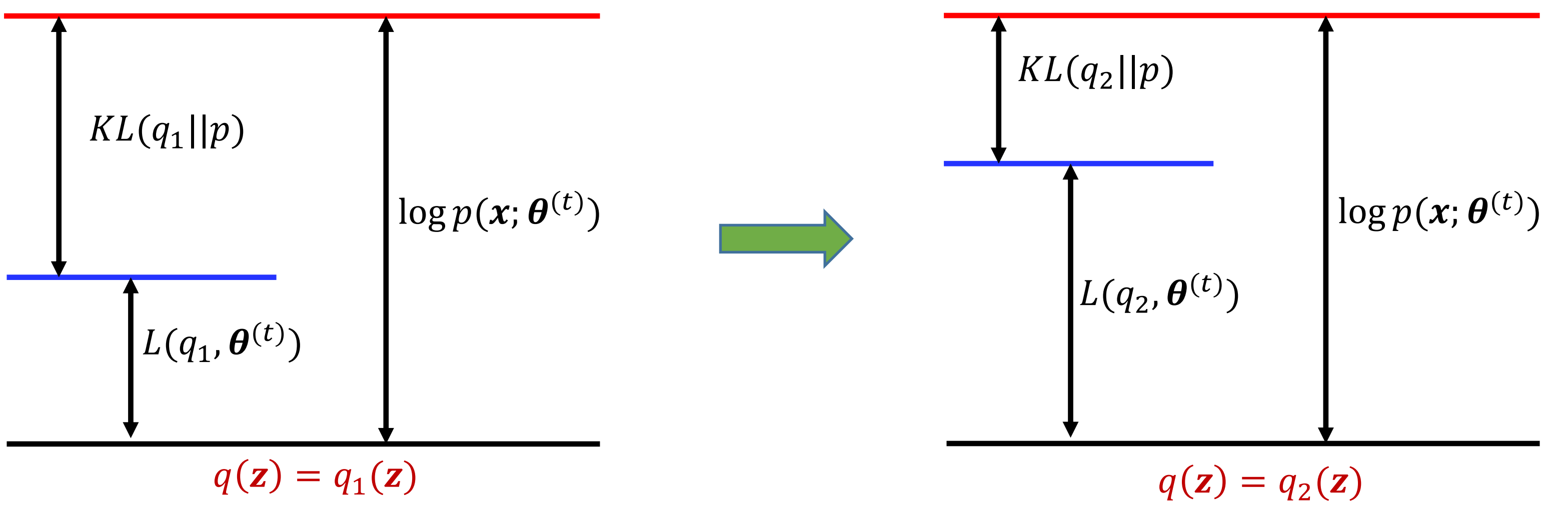
$$KL(q||p) \triangleq \int q(z) \log \frac{q(z)}{p(z)} dz \geq 0$$

- Thus, with the parameter at the t-th iteration denoted as $\boldsymbol{\theta}^{(t)}$, we have

$$\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) = \mathcal{L}\left(q,\boldsymbol{\theta}^{(t)}\right) + KL(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}))$$

This equality holds for any distribution $q(\boldsymbol{z})$

- Different $q(\boldsymbol{z})$ will lead to different decomposition of $\log p\left(\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)$



Theoretical Justification for EM

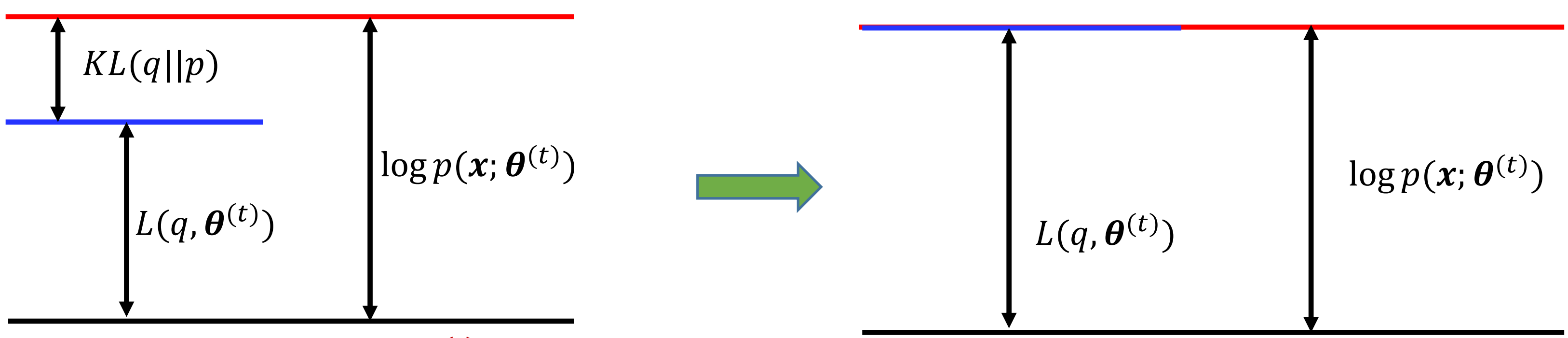
$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- If we set $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, then we have

$$KL\left(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right) = 0$$

Thus, we have

$$\begin{aligned}\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) &= \mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t)}\right) \\ &= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)}\end{aligned}$$



$q(\boldsymbol{z}) \neq p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$ $\qquad$ $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)}) = \mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t)}\right)$$

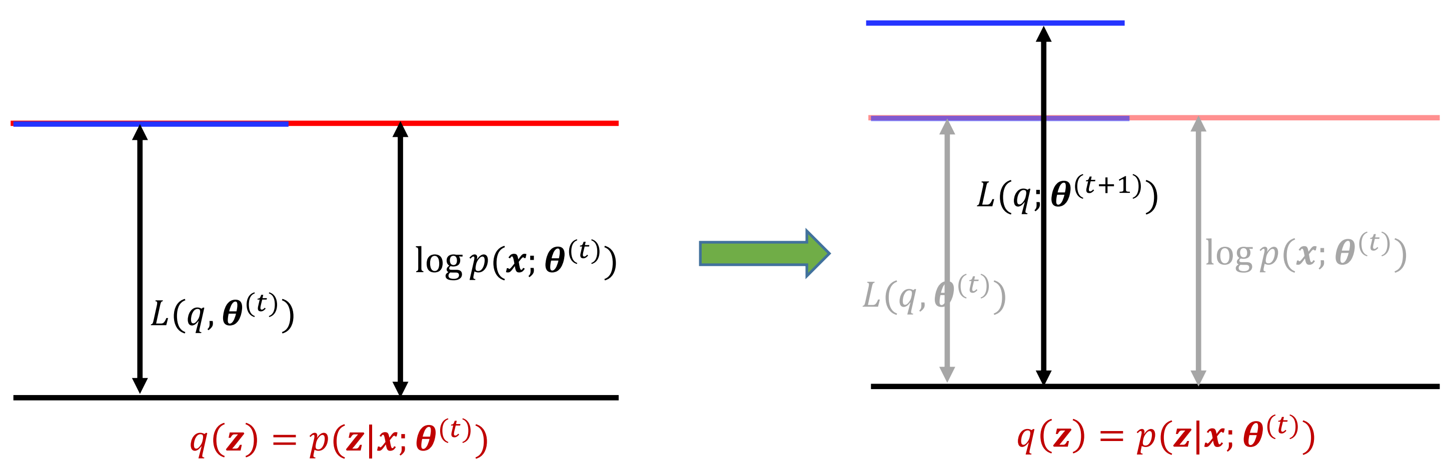
$$= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)\log\frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- If we update $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}}\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right),$$

then we must have the relation

$$\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t+1)}\right) \geq \underbrace{\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}^{(t)}\right)}_{=\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

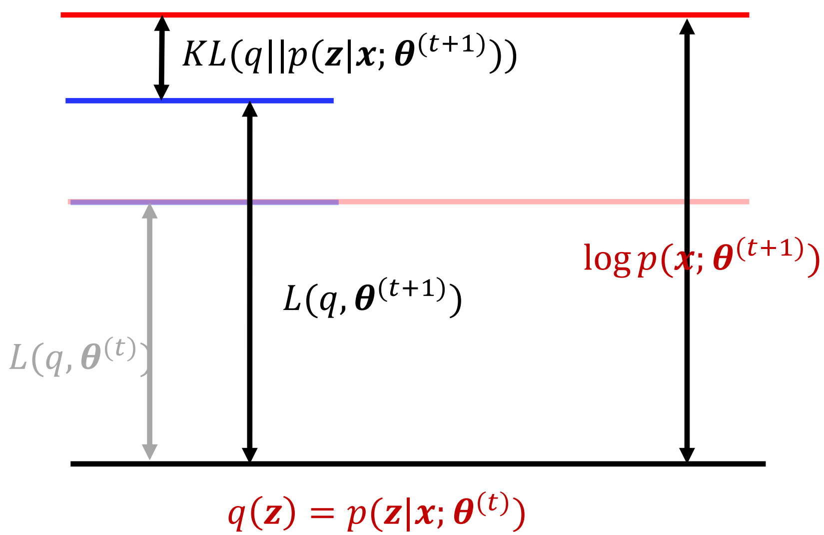


$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t+1)}\right)}{q(\boldsymbol{z})} + \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)})}$$

- By setting $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, we obtain

$$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}) = \underbrace{\mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta}^{(t+1)}\right)}_{\geq \log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})} + \underbrace{KL(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}))}_{\geq 0}$$

The KL-divergence is always non-negative



- Thus, we can see that

$$\boxed{\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})}$$

$\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z} \middle| \boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right)$ *can guarantee the increase of likelihood at each step*

- Equivalence between EM updating

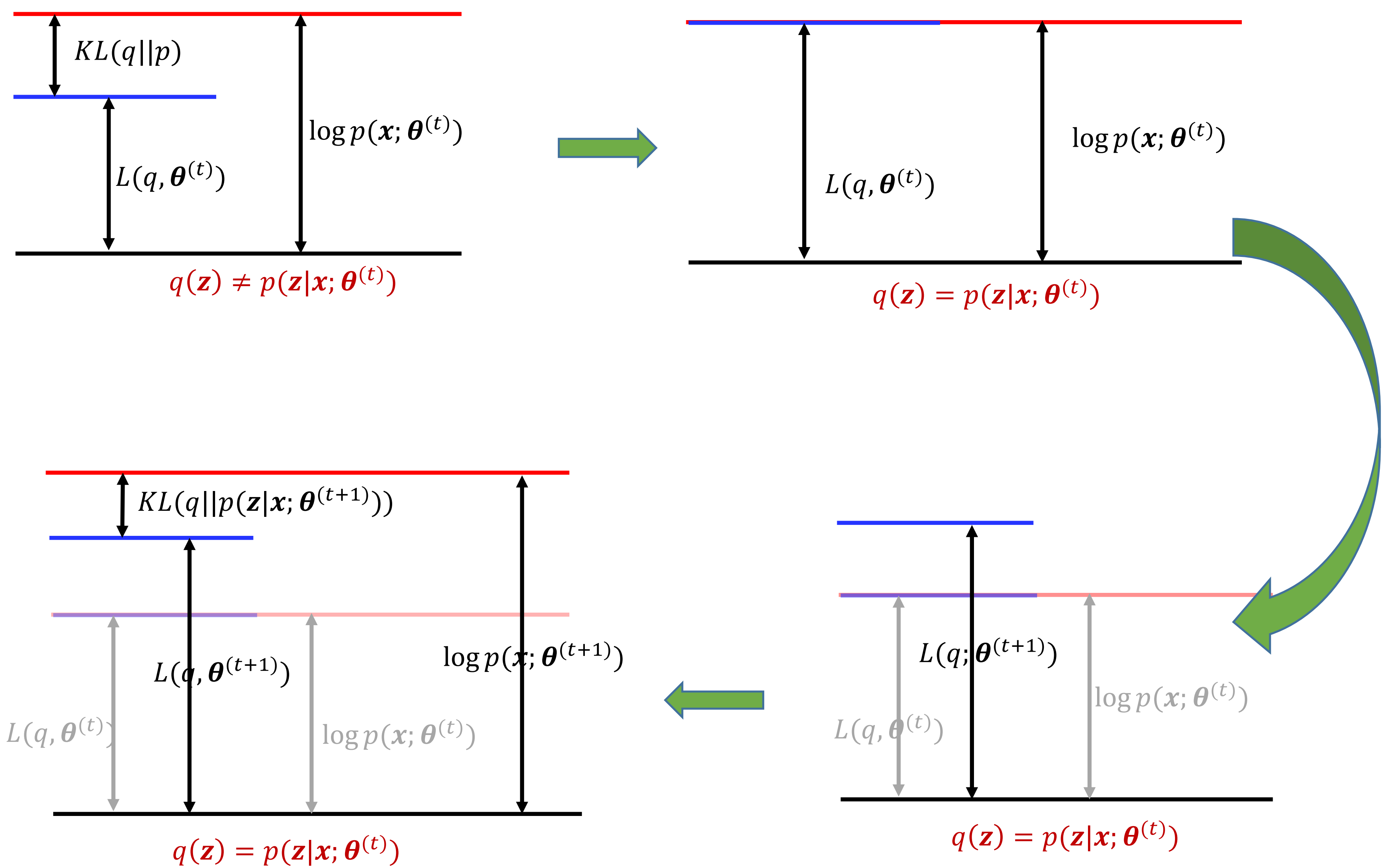
$$\arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) \quad \text{with} \quad \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) \triangleq \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

and the updating rule $\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right)$

$$\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right) = \underbrace{\sum_{\boldsymbol{z}} p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}_{\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]} - \underbrace{\sum_{\boldsymbol{z}} p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)\log p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right)}_{constant}$$

Therefore, $\arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right) \Leftrightarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$

EM algorithm can guarantee the increase of likelihood at each step


$KL(q||p)$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) \neq p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$KL(q||p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}))$
$L(q,\boldsymbol{\theta}^{(t+1)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t+1)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q;\boldsymbol{\theta}^{(t+1)})$
$\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})$
$L(q,\boldsymbol{\theta}^{(t)})$
$q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

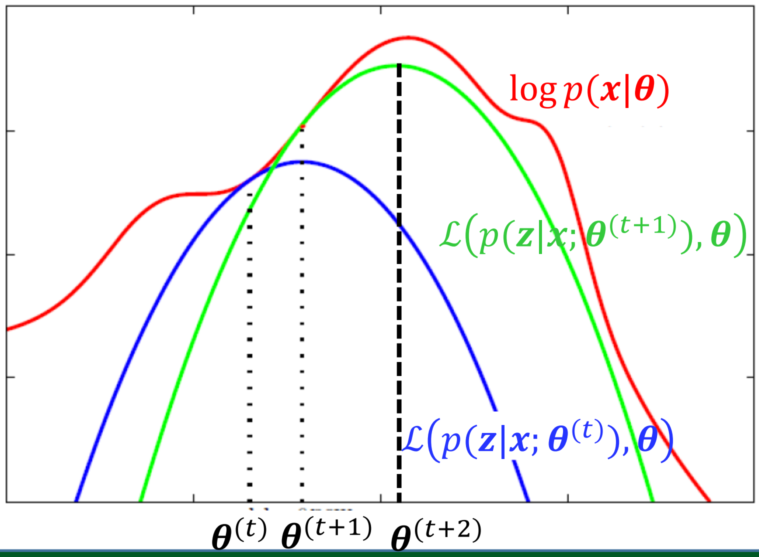
A View in the Parameter Space

1) E-step (*t*): deriving the expression $\mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$ given the model parameter $\boldsymbol{\theta}^{(t)}$

2) M-step (*t*): computing the optimal value $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}),\boldsymbol{\theta})$

3) E-step (*t+1*): deriving the expression for $\mathcal{L}(p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)}),\boldsymbol{\theta})$ given the model parameter $\boldsymbol{\theta}^{(t+1)}$

4) Repeating the above process until convergence



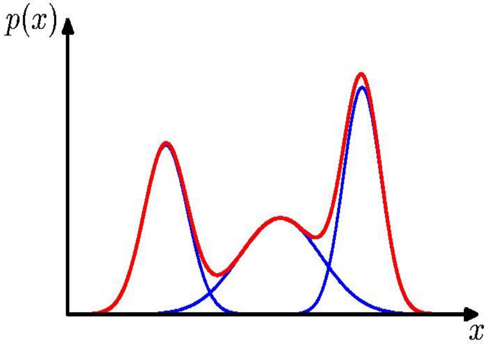
Outline

- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models
- EM Variants

Gaussian Mixture Model Review

- For a Gaussian mixture distribution, *i.e.*,

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



it can be represented as the marginal distribution of the joint distribution

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$$

$$= \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$

  - $\boldsymbol{z} = [z_1, z_2, \cdots, z_K]$ follows the categorical distribution with parameter $\boldsymbol{\pi}$

EM Two Steps

- It is a latent-variable model, thus we can use EM to optimize it

Remark: maximizing $\max_{\boldsymbol{\theta}} \mathcal{L}\left(p\left(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right),\boldsymbol{\theta}\right)$ is equivalent to $\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$

- *Reminder:* Key integrant in EM
  - E-step: Expectation *w.r.t.* the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{p\left(\boldsymbol{z}^{(n)}|\boldsymbol{x}^{(n)};\boldsymbol{\theta}^{(t)}\right)}\left[\log p\left(\boldsymbol{x}^{(n)},\boldsymbol{z}^{(n)};\boldsymbol{\theta}\right)\right]$$

  - M-step: Maximization

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

EM: E-step

- The posteriori distribution

$$p\left(\boldsymbol{z} = \mathbf{1}_k \middle| \boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) = \frac{p\left(\boldsymbol{x}, \boldsymbol{z} = \mathbf{1}_k; \boldsymbol{\theta}^{(t)}\right)}{\sum_{i=1}^{K} p\left(\boldsymbol{x}, \boldsymbol{z} = \mathbf{1}_i; \boldsymbol{\theta}^{(t)}\right)}$$

$$\boxed{= \frac{\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right) \pi_k^{(t)}}{\sum_{i=1}^{K} \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_i^{(t)}, \boldsymbol{\Sigma}_i^{(t)}\right) \pi_i^{(t)}}}$$

  - $\mathbf{1}_k$ denotes the one-hot vector with the k-th element being 1

- The log of the joint distribution $\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$

$$\boxed{\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) = \sum_{k=1}^{K} z_k \cdot [\log \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \log \pi_k]}$$

Note that $\boldsymbol{z}$ can only be a one-hot vector

- The expectation

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

$$= \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[z_k][\log \mathcal{N}(\boldsymbol{x}_n;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \log \pi_k]$$

  - Due to $p\left(\boldsymbol{z}=\mathbf{1}_k \middle| \boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) = \frac{\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)\pi_k^{(t)}}{\sum_{i=1}^{K}\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_i^{(t)},\boldsymbol{\Sigma}_i^{(t)}\right)\pi_i^{(t)}}$, we have

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[z_k] = \frac{\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}\right)\pi_k^{(t)}}{\sum_{i=1}^{K}\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu}_i^{(t)},\boldsymbol{\Sigma}_i^{(t)}\right)\boldsymbol{\pi}_i^{(t)}} \triangleq \gamma_k^{(t)}$$

- Therefore, we have

$$\boxed{\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) = \sum_{k=1}^{K} \gamma_k^{(t)}[\log \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \log \pi_k]}$$

- Substituting $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\frac{1}{(2\mathrm{Pi})^{D/2}|\boldsymbol{\Sigma}_k|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$ into $\mathcal{Q}(\cdot)$ gives

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\sum_{k=1}^{K}\gamma_k^{(t)}\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C$$

  - C is the constant

- So far, only one data example $\boldsymbol{x}$ is considered

- If data $\boldsymbol{x}^{(n)}$ for $n=1,2,\cdots N$ are considered, the $\mathcal{Q}(\cdot)$ becomes

$$\boxed{\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}^{(t)}\left[-\frac{1}{2}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C}$$

EM: M-step

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}^{(t)}\left[-\frac{1}{2}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k\right)-\frac{1}{2}\log|\boldsymbol{\Sigma}_k|+\log\pi_k\right]+C$$

- By taking derivatives *w.r.t.* $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and setting them to zero, we obtain the optimal $\boldsymbol{\theta}$ as

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t)}\boldsymbol{x}^{(n)}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t)}\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k^{(t+1)}\right)\left(\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_k^{(t+1)}\right)^T$$

  - For π_k, we need to consider the optimization under constraint $\sum_{k=1}^{K}\pi_k = 1$, leading to the solution

$$\pi_k^{(t+1)} = \frac{N_k}{N}$$

  where $N_k = \sum_{n=1}^{N}\gamma_{nk}^{(t)}$ is the effective number of examples assigned to the *k*-th class

Summary of EM Algorithm

- Given the current estimate $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{K}$, update γ_{nk} as

$$\gamma_{nk} \leftarrow \frac{\mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{i=1}^{K} \mathcal{N}(\boldsymbol{x}^{(n)}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i}$$

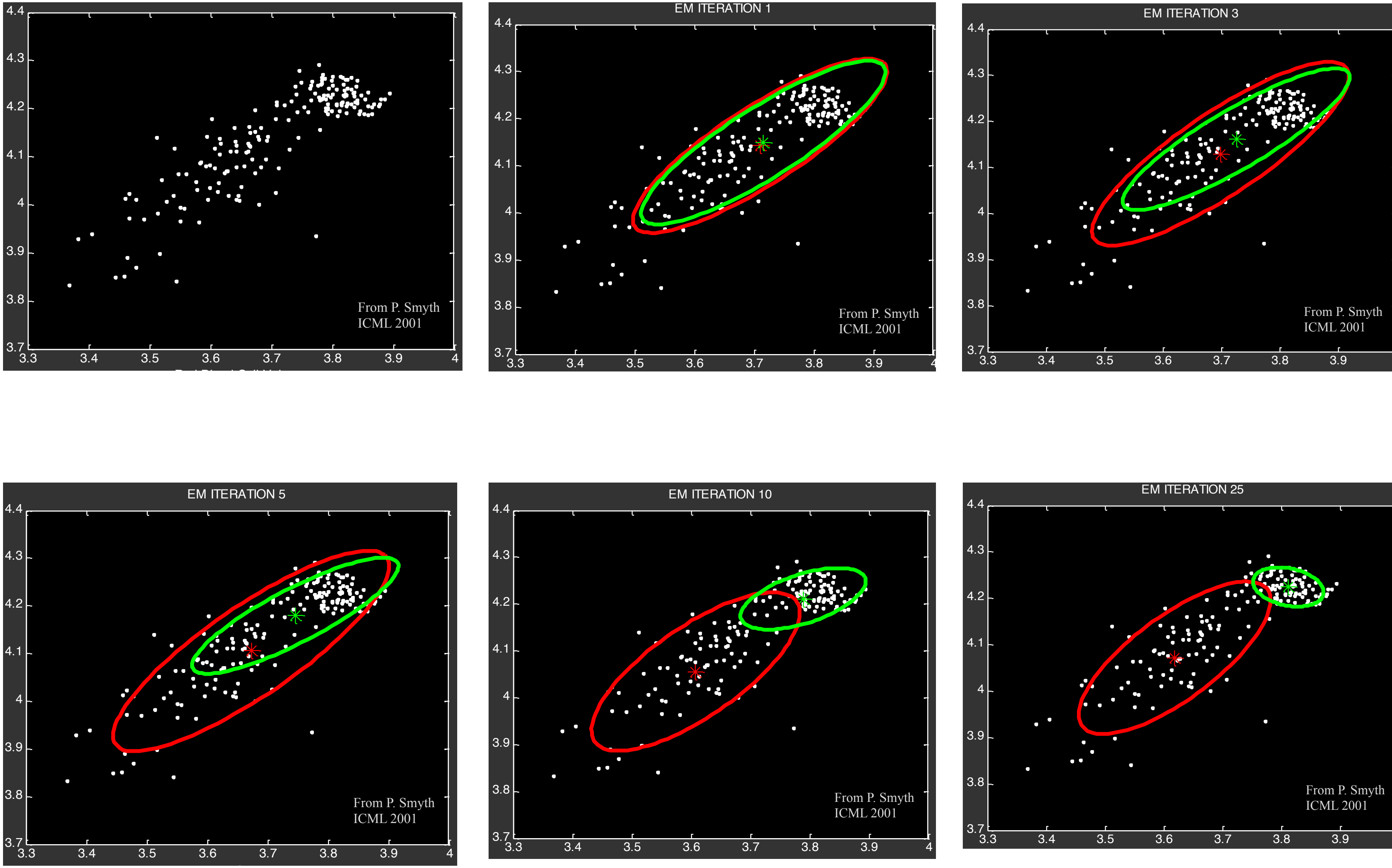
- Given the γ_{nk}, update $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ and π_k as

$$N_k \leftarrow \sum_{n=1}^{N} \gamma_{nk}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}^{(n)}$$

$$\boldsymbol{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k)(\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k)^T$$

$$\pi_k \leftarrow \frac{N_k}{N}$$


From P. Smyth
ICML 2001
EM ITERATION 1
From P. Smyth
ICML 2001
EM ITERATION 3
From P. Smyth
ICML 2001
EM ITERATION 5
From P. Smyth
ICML 2001
EM ITERATION 10
From P. Smyth
ICML 2001
EM ITERATION 25
From P. Smyth
ICML 2001

Relation to Soft K-Means

- When restricting $\boldsymbol{\Sigma}_k$ to the form $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$, the EM updating rules for GMM are

$$\pi_k \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$$

$$\gamma_{nk} \leftarrow \frac{\pi_k e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\right\|^2}}{\sum_{i=1}^{K} \pi_i e^{-\frac{1}{2\sigma^2}\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_i\right\|^2}}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}^{(n)}}{\sum_{n=1}^{N} \gamma_{nk}}$$

- Updates in soft K-means

Setting π_k and β as
$\pi_k = \frac{1}{K}$, $\beta = \frac{1}{2\sigma^2}$

$$r_{nk} = \frac{e^{-\beta\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\right\|^2}}{\sum_{i=1}^{K} e^{-\beta\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_i\right\|^2}}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} r_{nk}\, \boldsymbol{x}^{(n)}}{\sum_{n=1}^{N} r_{nk}}$$

Outline

- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models
- EM Variants

Review of the EM Algorithms

- To use EM algorithms, the key steps below are required

1) Computing the posteriori distribution

$$p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$$

2) Evaluating the expectation of $\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ *w.r.t.* the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$, *i.e.*,

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)=\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$$

3) Maximizing

$$\boldsymbol{\theta}^{(t+1)}=\arg\max_{\boldsymbol{\theta}}\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$$

However, not all of them are always achievable

Two Issues in EM Algorithm

- Issue one

The maximization is not achievable

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$$

- Issue two

1) The posteriori $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ cannot be derived analytically

2) Even if $p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$ can be obtained, we still cannot derive the close-form expression for the expectation

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$$

Generalized EM

- It is quite often in training LVMs that the optimization $\max_{\boldsymbol{\theta}} \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ cannot be solved

How to address this issue?

- Maximizing $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ is not necessary. Increasing $\mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$ is sufficient to guarantee the EM algorithm to work

- That is, if we adopt SGD to update the parameter as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \gamma \cdot \left.\frac{\partial \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

we can also guarantee the monotonic increase of log-likelihood

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

- Sketch of proof

  - First, after the SGD update, it can be easily seen that

$$\mathcal{Q}\left(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\right) \geq \mathcal{Q}\left(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\right)$$

  - From $\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right), \boldsymbol{\theta}\right) = \int p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right)} d\boldsymbol{z} = \mathcal{Q}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) - \int p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) \log p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right) d\boldsymbol{z}$, we further have

$$\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right); \boldsymbol{\theta}^{(t+1)}\right) \geq \underbrace{\mathcal{L}\left(p\left(\boldsymbol{z}\middle|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}\right); \boldsymbol{\theta}^{(t)}\right)}_{=\log p(\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

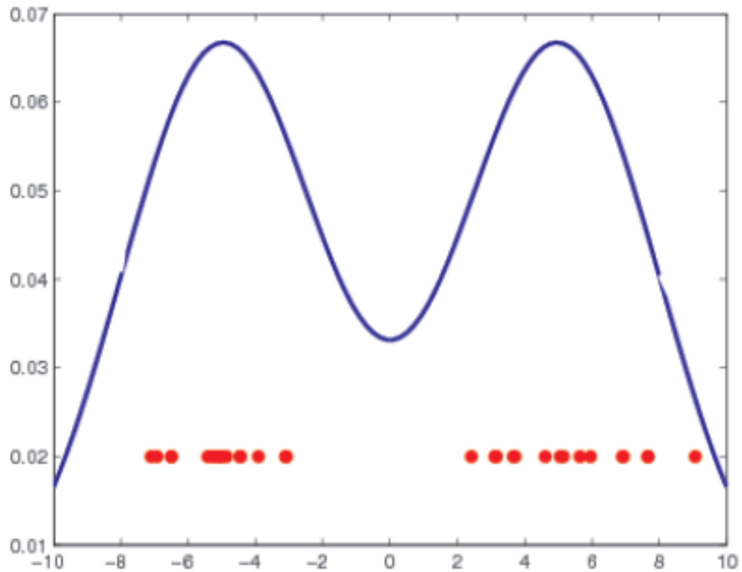
  - Due to

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) = \underbrace{\mathcal{L}\left(p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)}), \boldsymbol{\theta}^{(t+1)}\right)}_{\geq \log p(\boldsymbol{x}|\boldsymbol{\theta}^{(t)})} + \underbrace{KL(p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})||p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}))}_{\geq 0}$$

$$\Longrightarrow \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t+1)}) \geq \log p(\boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

MCMC EM

- For any probability distributions, we can always draw samples from it, *e.g.*, using Markov chain Monte Carlo (MCMC) methods



- Although the exact expression of the posteriori $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$ is not known, we can use samples drawn from it to approximate it

- Thus, we can draw lots of samples $\boldsymbol{z}_s$ for $s = 1, \cdots, S$ from the posteriori distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$ such that

$$\boldsymbol{z}_s \sim p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$$

- Then, the expectation $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})]$ can be approximated as

$$\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right) \approx \frac{1}{S} \cdot \sum_{s=1}^{S} \log p(\boldsymbol{x},\boldsymbol{z}_s;\boldsymbol{\theta})$$

- We can optimize the approximate $\mathcal{Q}\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}\right)$ with SGD algorithm

The two sub-problems in the Issue Two are both solved. Thus, latent-variable models can always be trained with MCMC EM

VB-EM

- Drawing samples from a distribution is *computationally expensive*

- An alternative approach is to use a simple distribution $q(\boldsymbol{z};\boldsymbol{\phi})$ to approximate the exact posterior distribution $p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})$

How to get the approximate simple distribution $q(\boldsymbol{z};\boldsymbol{\phi})$?

- Idea

1) Assuming a simple form for $q(\boldsymbol{z};\boldsymbol{\phi})$, *e.g.*,

$$q(\boldsymbol{z};\boldsymbol{\phi}) = \prod \mathcal{N}\left(z_i;\mu_i,\sigma_i^2\right)$$

2) Finding the best $\boldsymbol{\phi}$ that minimizes the KL-divergence

$$KL\left(q(\boldsymbol{z};\boldsymbol{\phi}) \| p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\right)$$

- Steps to update the model parameter $\boldsymbol{\theta}$

1) Finding the best approximate $q(\boldsymbol{z}; \boldsymbol{\phi})$ such that

$$\boldsymbol{\phi}^{(t)} = \arg\min_{\boldsymbol{\phi}} KL\left(q(\boldsymbol{z}; \boldsymbol{\phi}) \| p(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\theta}^{(t)})\right)$$

2) Using $q\left(\boldsymbol{z}; \boldsymbol{\phi}^{(t)}\right)$ to compute expectation $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$ approximately as

$$\tilde{Q}\left(\boldsymbol{\theta};\ \boldsymbol{\phi}^{(t)}\right) = \mathbb{E}_{q\left(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\right)}[\log p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})]$$

3) Obtaining the new value $\boldsymbol{\theta}^{(t+1)}$ as

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \tilde{Q}\left(\boldsymbol{\theta};\ \boldsymbol{\phi}^{(t)}\right)$$

- The two optimization problems can be equivalently written as

$$\min_{\boldsymbol{\phi}} \; KL\big(q(\boldsymbol{z};\boldsymbol{\phi}) \| p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\big) \Leftrightarrow \max_{\boldsymbol{\phi}} \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

$$\Leftrightarrow \max_{\boldsymbol{\phi}} \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{z};\boldsymbol{\phi}^{(t)})}[\log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})] \Leftrightarrow \max_{\boldsymbol{\theta}} \int q\big(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\big) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q\big(\boldsymbol{z};\boldsymbol{\phi}^{(t)}\big)} d\boldsymbol{z}$$

- The algorithm to optimize $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can be understood as solving the following optimization problem in an alternative way

$$\max_{\boldsymbol{\phi},\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$

with

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

- Instead of updating $\boldsymbol{\theta}, \boldsymbol{\phi}$ alternatively, we can also update them *simultaneously* with the SGD algorithm, that is,

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \gamma \cdot \left. \frac{\partial \mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

$$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \gamma \cdot \left. \frac{\partial \mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^{(t)}}$$

The method is dubbed *variational Bayesian EM* (VB-EM)

- In general, we optimize $\mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ *w.r.t.* the two parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ simultaneously

- Actually, it can be proved that $\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$ is a *lower bound* of the log-likelihood $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, that is,

$$\ln p(\boldsymbol{x};\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$

(Proof can be found in the next slide)

When the log-likelihood $\ln p(\boldsymbol{x};\boldsymbol{\theta})$ cannot be directly maximized, we can seek to optimize its lower bound

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) \triangleq \int q(\boldsymbol{z};\boldsymbol{\phi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\phi})} d\boldsymbol{z}$$

where $q(\boldsymbol{z};\boldsymbol{\phi})$ can be set to be any distribution form, *e.g.*,

$$q(\boldsymbol{z};\boldsymbol{\phi}) = \prod \mathcal{N}\left(z_i;\mu_i,\sigma_i^2\right)$$

Proof of $\ln p(\boldsymbol{x};\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$

$$\ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{z} = \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\ \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z} + \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

$$= \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z} + KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$$

$$\geq \int_{\boldsymbol{z}} q_{\boldsymbol{\phi}}(\boldsymbol{z}) \ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} d\boldsymbol{z}$$

$$\triangleq \mathcal{L}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})$$