Expectation-Maximization Algorithm

DCS310

Sun Yat-sen University

Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models

General Form of the Concerned 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

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models

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

Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models

Re-representing the Log-likelihood

- The log-likelihood can be reformulated as

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

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

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

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

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

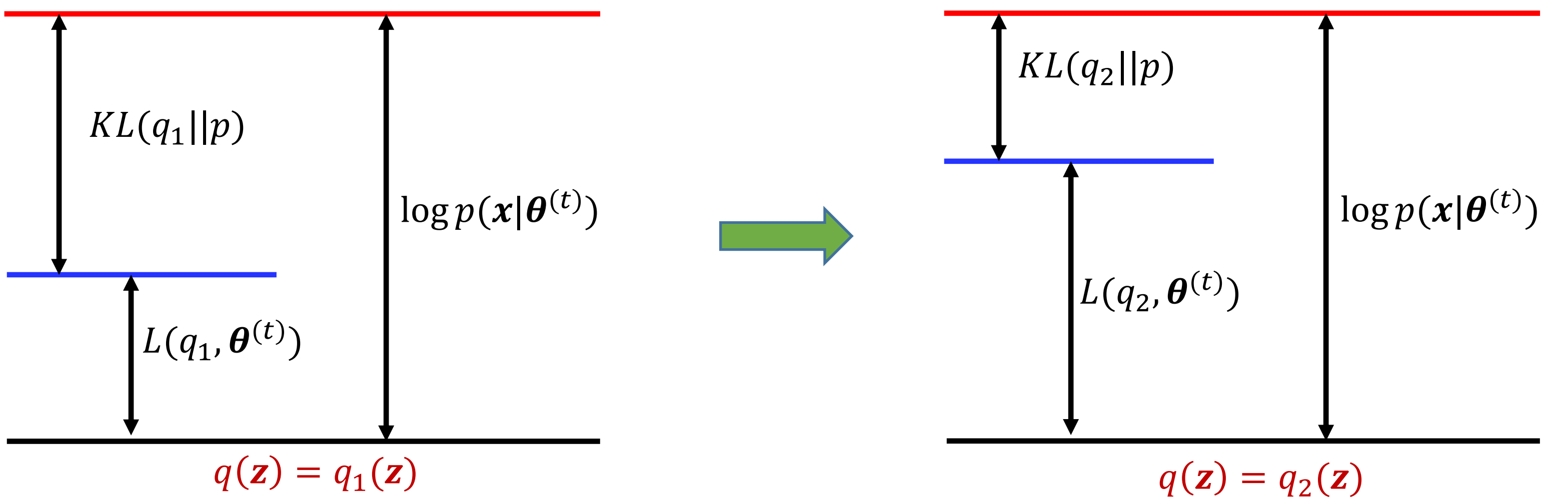
$$KL(q||p) \triangleq \int q(z) \log \frac{q(z)}{p(z)} dz \geq 0$$

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

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

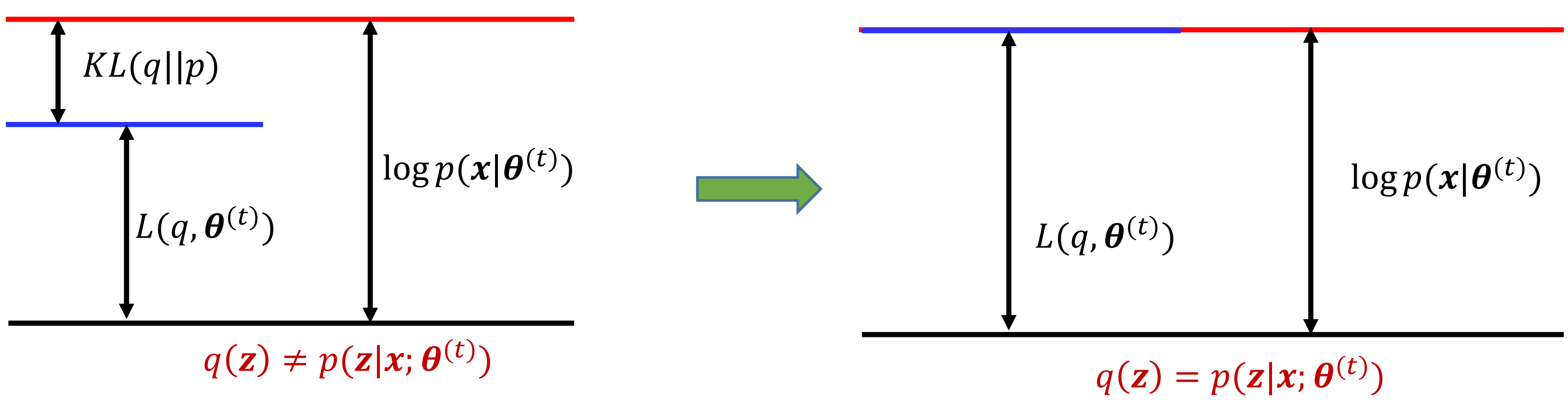
$$KL\big(q||p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})\big) = 0$$

i.e., *q* and *p* have no distance

Thus, we have

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

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



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

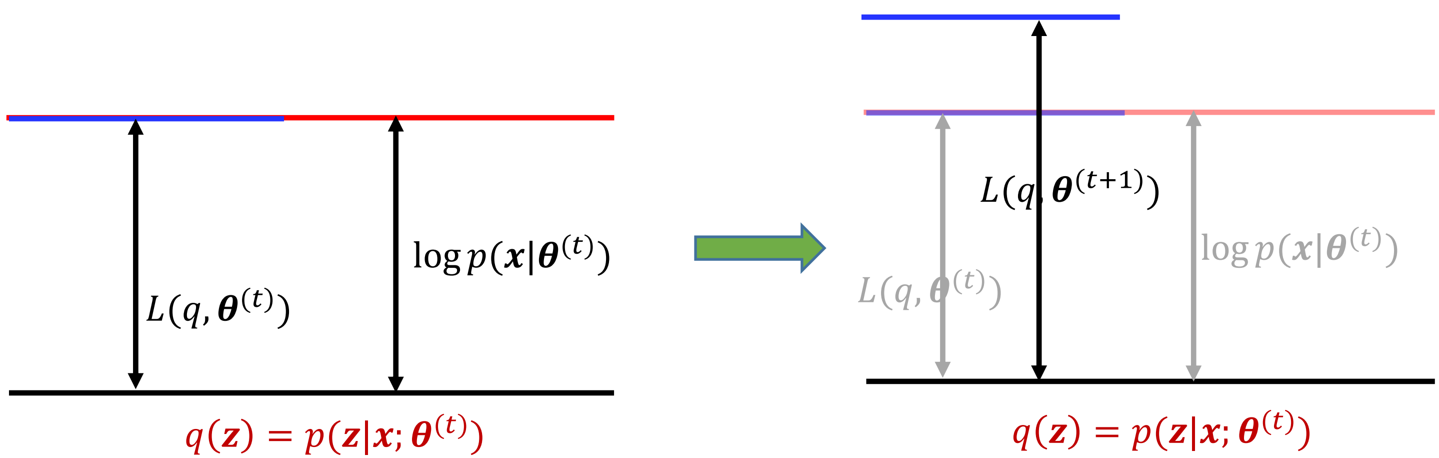
$$= \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}\middle|\boldsymbol{x};\boldsymbol{\theta}^{(t)}\right) \log \frac{p\left(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}^{(t)}\right)}{p(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})}$$

- 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

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



- From the nonnegative property of KL-divergence, we known that

$$KL\left(p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t)})||p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}^{(t+1)})\right) \geq 0$$

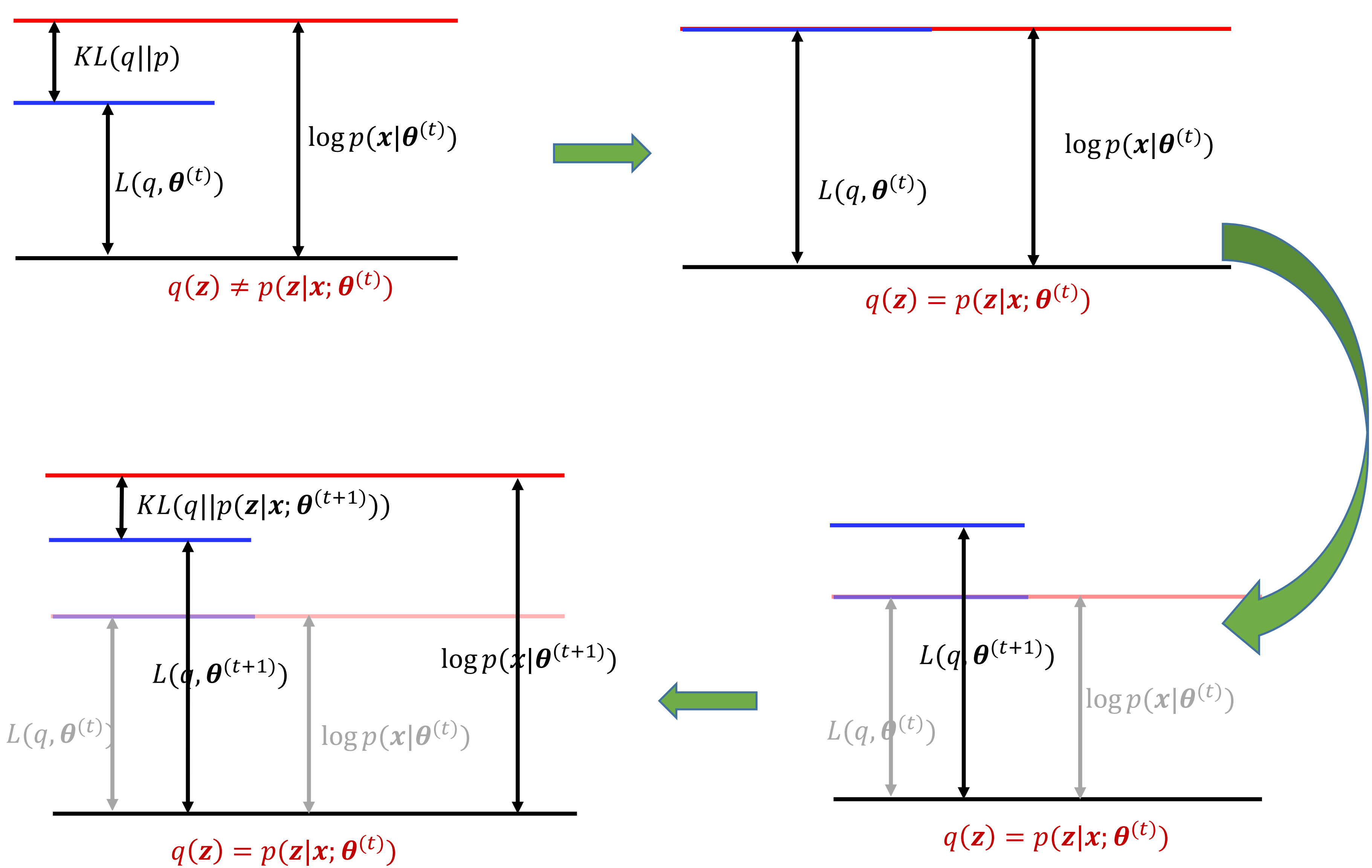
- Because $\log p(\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + KL(q||p(\mathbf{z}|\boldsymbol{x};\boldsymbol{\theta}))$ holds for any q, thus we have

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

- Thus, we can see that

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

EM algorithm can guarantee the increase of likelihood at each step



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

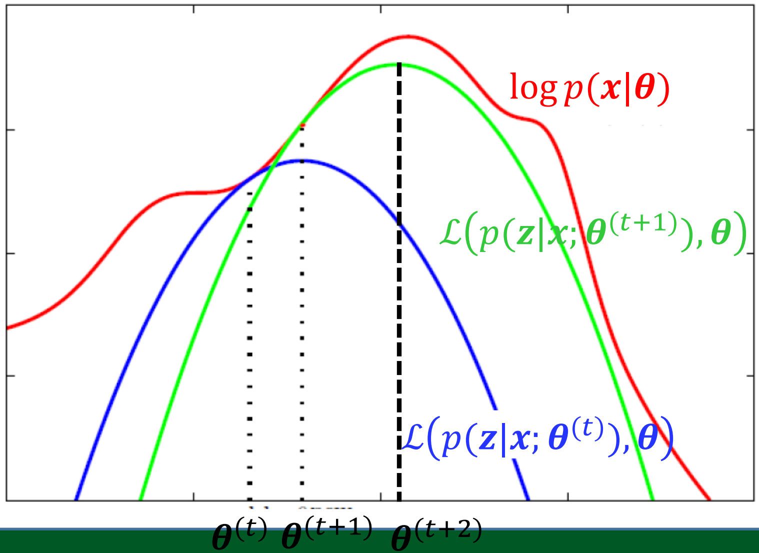
A View in the Parameter Space

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

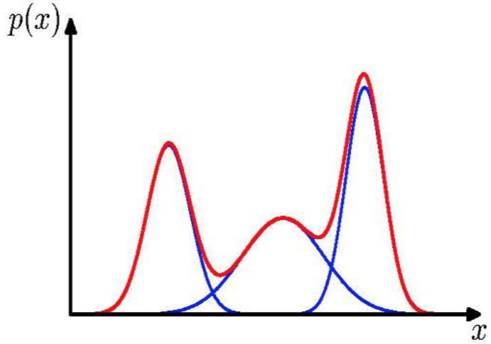


Outline

- The Concerned Problem
- EM Algorithm
- Theoretical Guarantees
- Example: Training Gaussian Mixture Models

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: E-step

- The posteriori distribution

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

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

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

- The log of the joint distribution $p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) = \prod_{k=1}^{K}[\pi_k \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)]^{z_k}$

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

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

- The expectation

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

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

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

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

- Therefore, we have

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

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

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

  - 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, $\mathcal{Q}(\cdot)$ becomes

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

EM: M-step

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

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

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

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

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

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

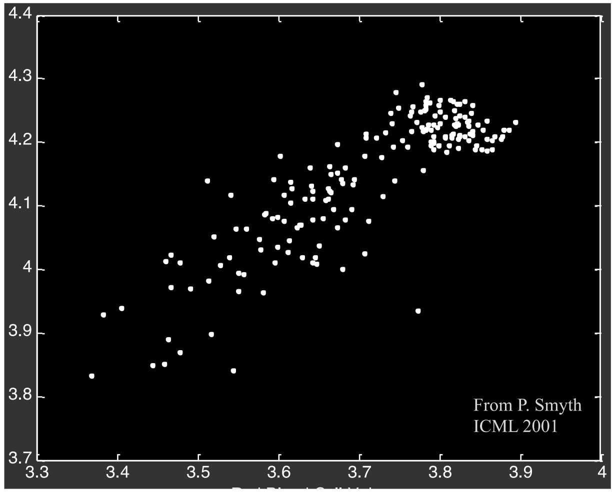
- Given the γ_{nk}, update $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ and π_k as

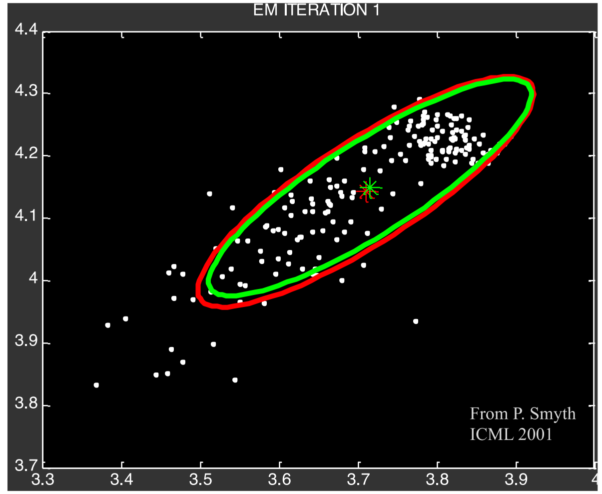
$$N_k \leftarrow \sum_{n=1}^{N} \gamma_{nk}$$

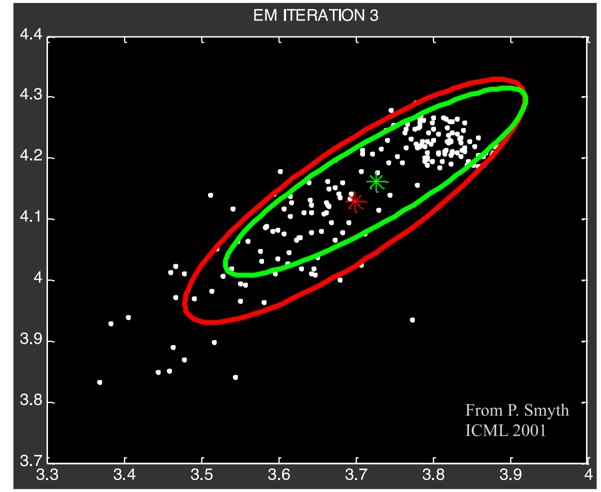
$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n$$

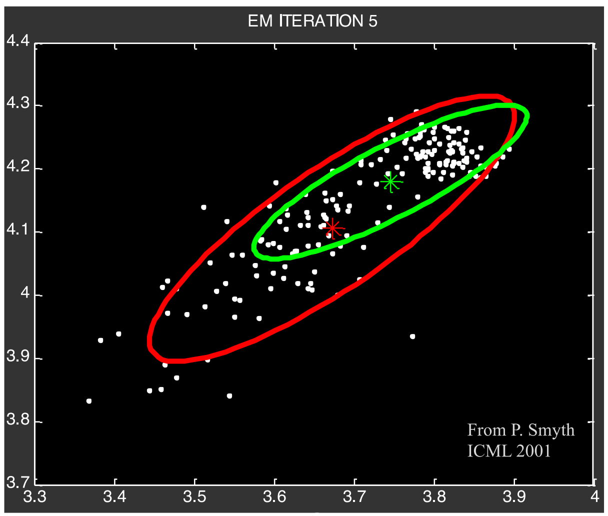
$$\boldsymbol{\Sigma}_k \leftarrow \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T$$

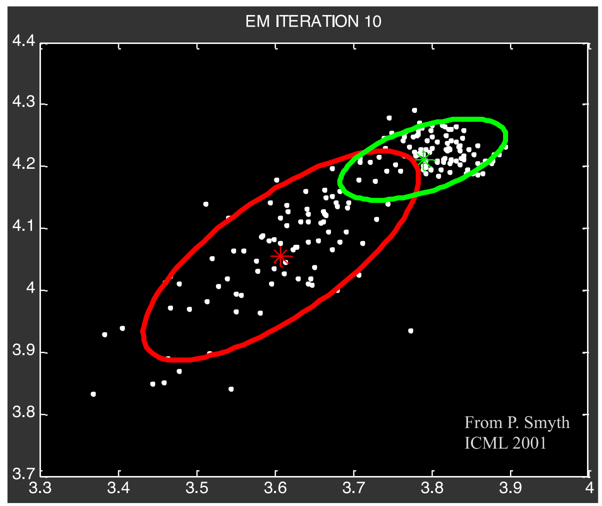
$$\pi_k \leftarrow \frac{N_k}{N}$$

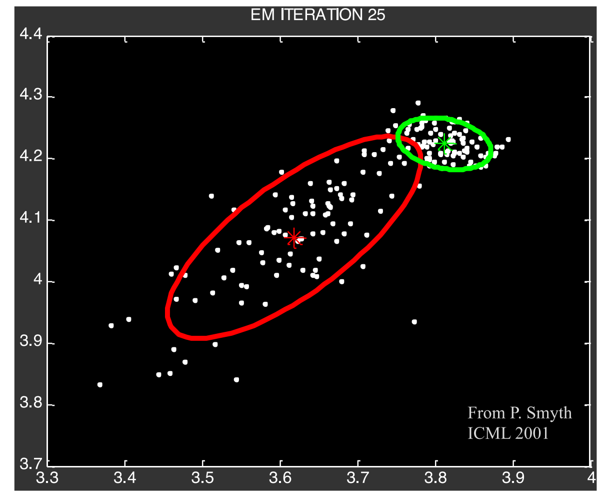

4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001


EM ITERATION 1
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001


EM ITERATION 3
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001


EM ITERATION 5
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001


EM ITERATION 10
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001


EM ITERATION 25
4.4
4.3
4.2
4.1
4
3.9
3.8
3.7
3.3 3.4 3.5 3.6 3.7 3.8 3.9 4
From P. Smyth
ICML 2001

Relation to Soft K-Means

- When restricting $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$, the updating of GMM becomes

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

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

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

where $\beta_i = \frac{\ln \pi_i}{2\sigma^2}$

- Updates in soft K-means

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

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