# Maximum likelihood learning and Expectation-Maximization (EM) algorithm

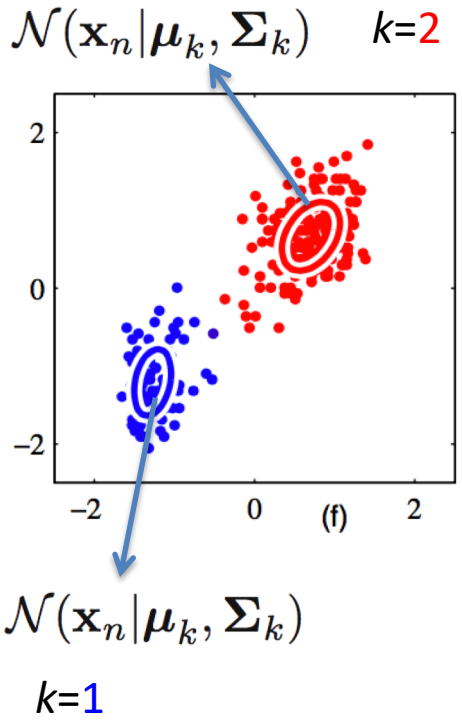
Shikui Tu

2018-03-15

# Outline

- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood
- General EM algorithm
- Bayesian learning

# Introduce a latent variable




We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster $k$

$$\mathbf{z} = (z_1, \ldots, z_K) \qquad z_k \in \{0, 1\} \qquad \sum_k z_k = 1$$

A mixing weight for each cluster:

$$p(z_k = 1) = \pi_k \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

*prior probability of point belonging to a cluster*

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

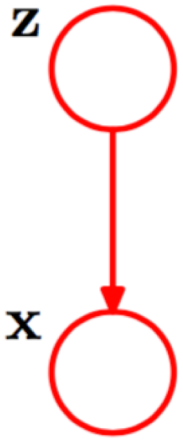
Assume the points in the same cluster follow a **Gaussian distribution**

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

# Gaussian Mixture Model (GMM)

## Generative process

- Randomly sample a $\mathbf{z}$ from a categorical distribution $[\pi_1, \ldots, \pi_K]$;
- Generate $\boldsymbol{x}$ according to Gaussian distribution $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$




Graphical representation of $\mathbf{p}(\mathbf{x}, \mathbf{z}) = \mathbf{p}(\boldsymbol{z})\mathbf{p}(\mathbf{x}|\mathbf{z})$

So, we get a distribution for the data point $\mathbf{x}$:

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

# Outline

- Gaussian Mixture Models (GMM)

- Expectation-Maximization (EM) for maximum likelihood

- General EM algorithm

- Bayesian learning

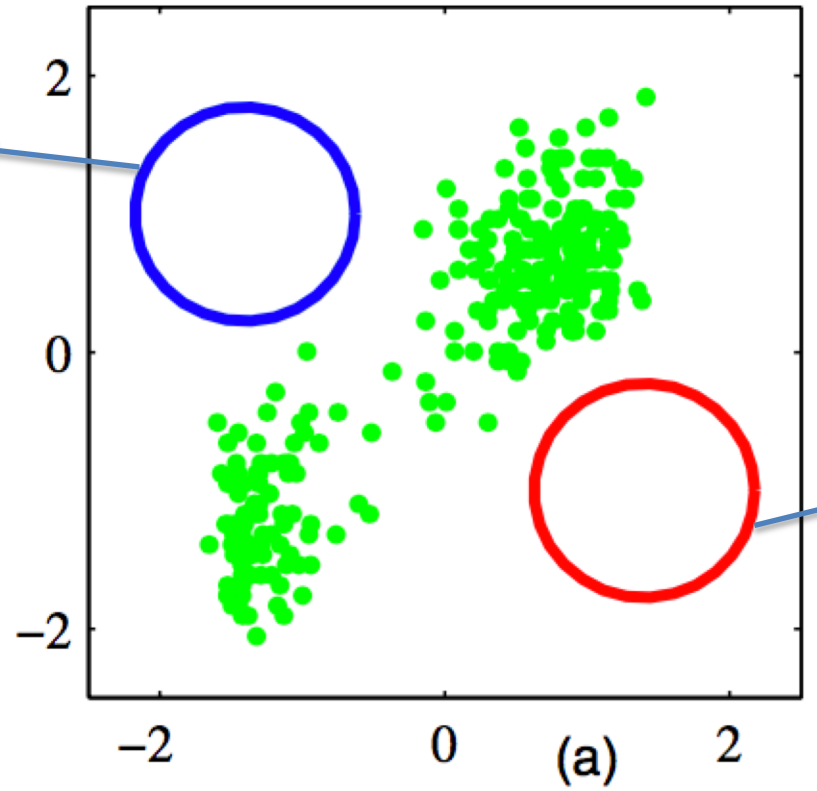
# Expectation-Maximization (EM) algorithm for maximum likelihood

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

Initialization

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

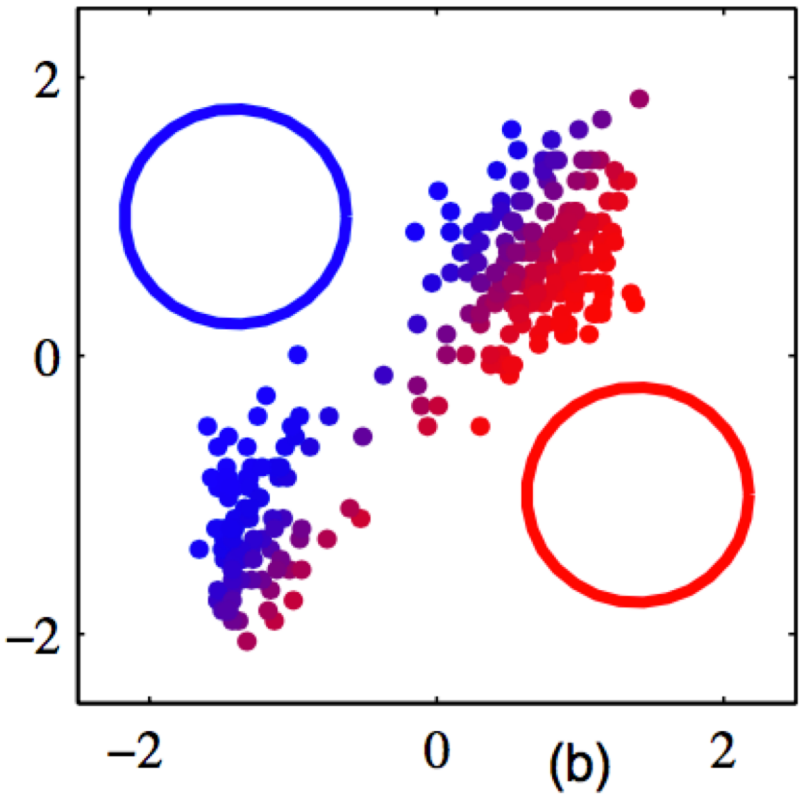
*k*=1




$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

*k*=2

# E-Step



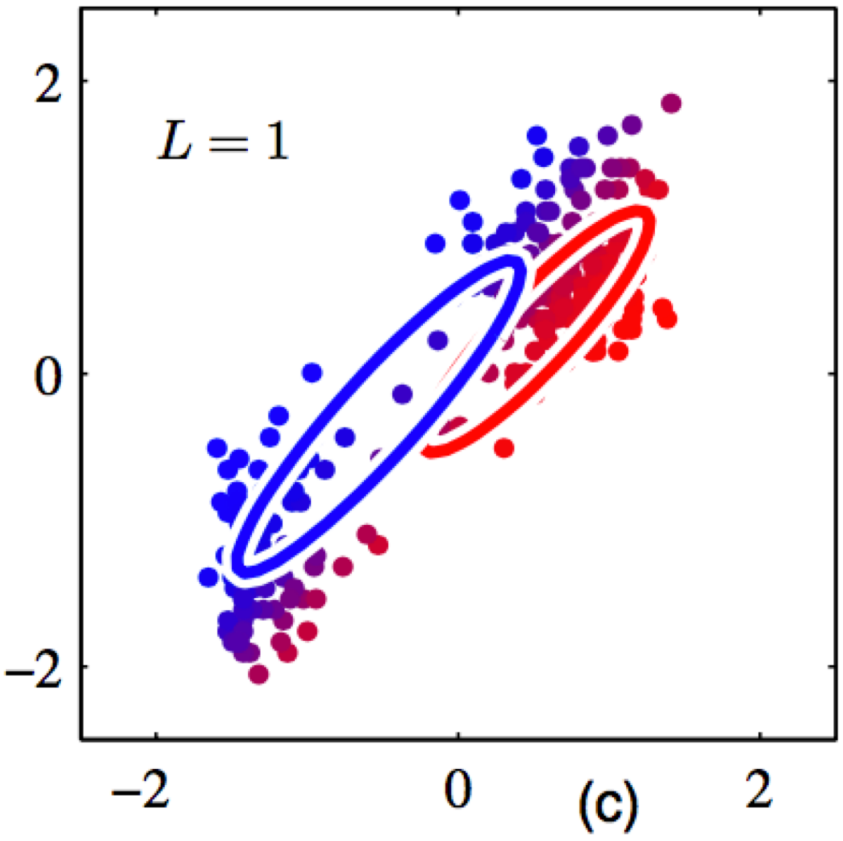

When the parameters are given, the assignments of the points can be calculated by the posterior probability, i.e., the probability of a data point belonging to a cluster once we have observed the data point.

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

Soft assignment:
A point fractionally belongs to two clusters.

For example,
0.2 belong to cluster 1
0.8 belong to cluster 2

# M-Step




When the assignments $\gamma(z_{nk})$ of the points to the clusters are known, parameters could be calculated for each cluster (Gaussian) separately.

Mixing weight $\pi_k$: the proportion of number of points in cluster k within all data points

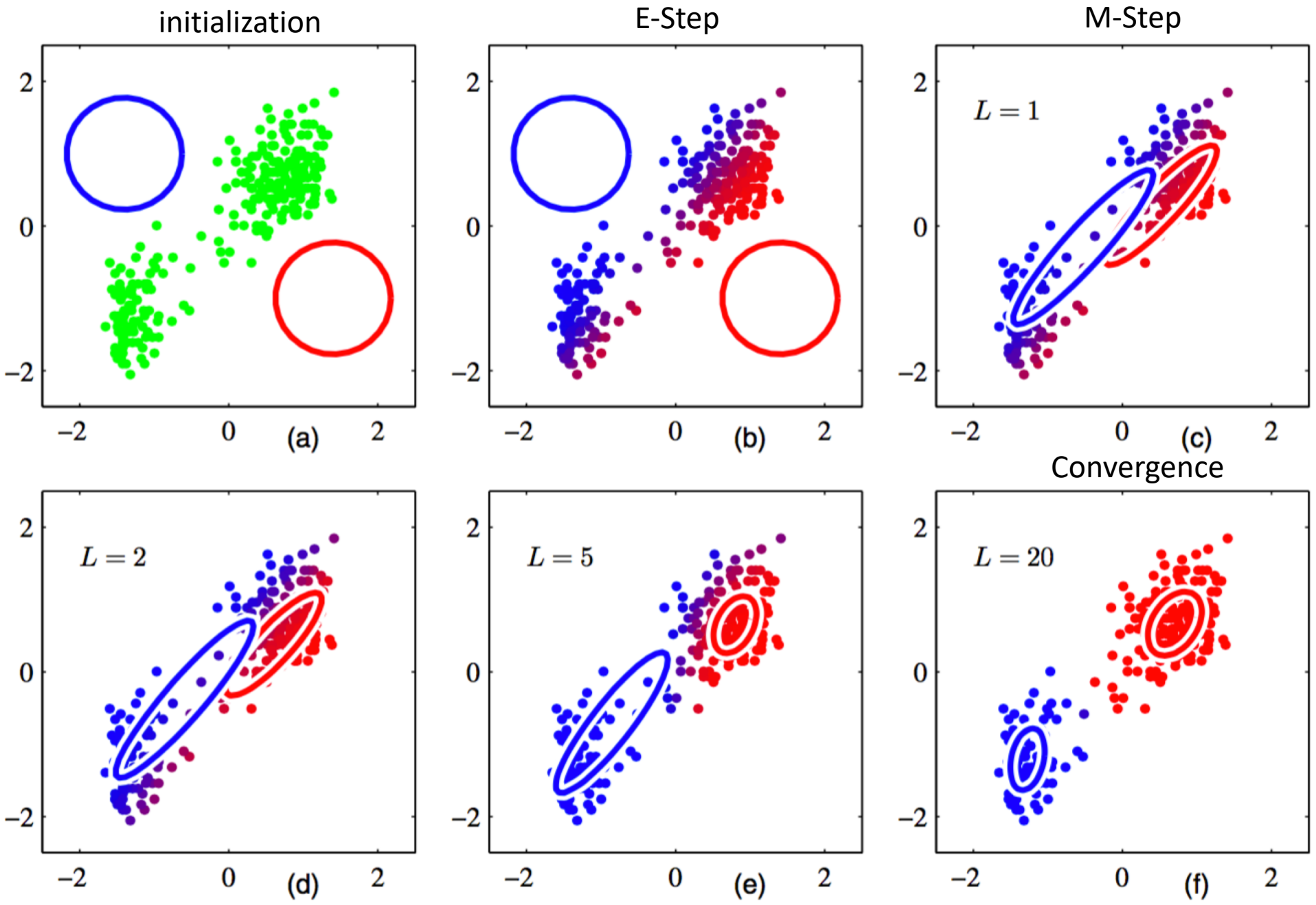
$$\pi_k = \frac{N_k}{N} \quad ; \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$: the mean and the covariance matrix are calculated for each cluster

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

*L* denotes the number of cycles of the EM algorithm.




*L* denotes the number of cycles of E-Step and M-Step.

# Details of the EM Algorithm

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients $\pi_k$, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

4. Evaluate the log likelihood

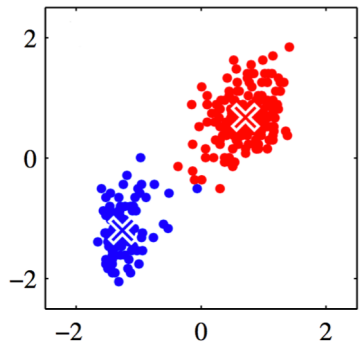
$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

# Relation to K-means

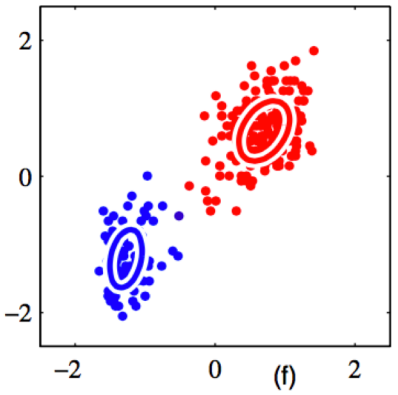
Fixed equal mixing weights

$$\|\mathbf{x} - \boldsymbol{\mu}_k\|^2$$




$$\{\boldsymbol{\mu}_k\}$$

$$\boldsymbol{\Sigma}_k = \varepsilon \mathbf{I}$$




GMM considers covariance and mixing weights.

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}$$

One-in-K assignment

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

Soft assignment

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

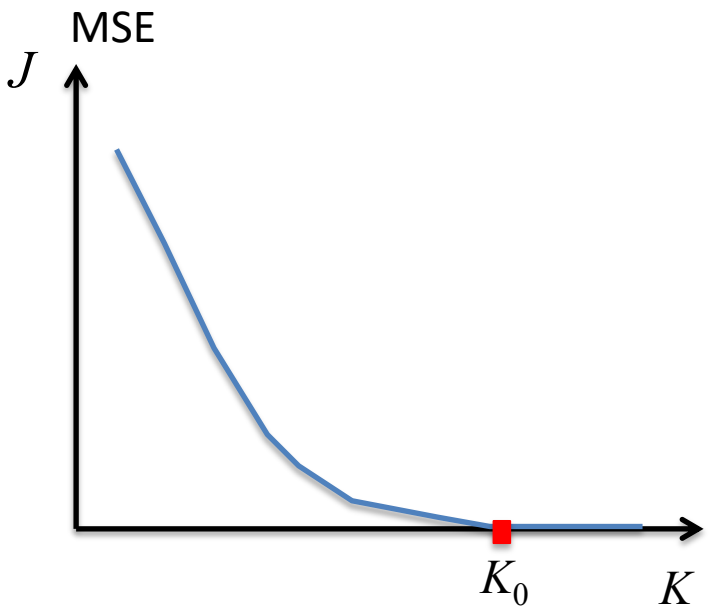
# Summary for the EM algorithm for GMM

- Does it find the global optimum?
  - No, like K-means, EM only finds the nearest local optimum and the optimum depends on the initialization
- GMM is more general than K-means by considering mixing weights, covariance matrices, and soft assignments.
- Like K-means, it does not tell you the best K.

# How to determine the cluster number K?

## K-mean

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

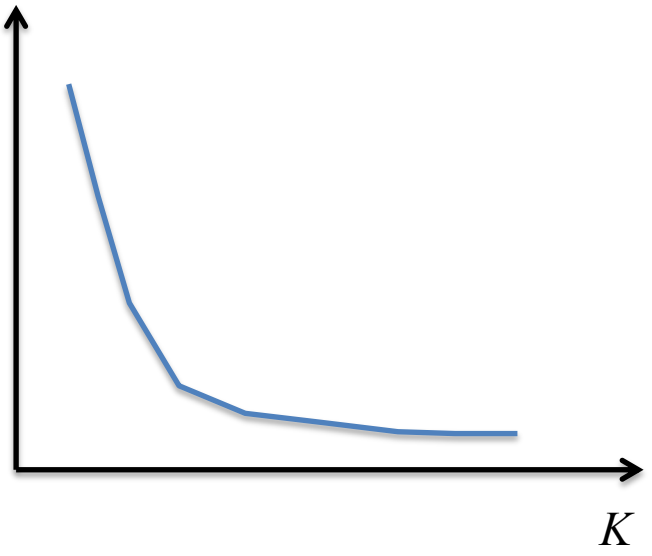



$J$ does not tell which $K$ is better.

## GMM

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

- Log-likelihood




Negative log-likelihood also decreases as $K$ increases.

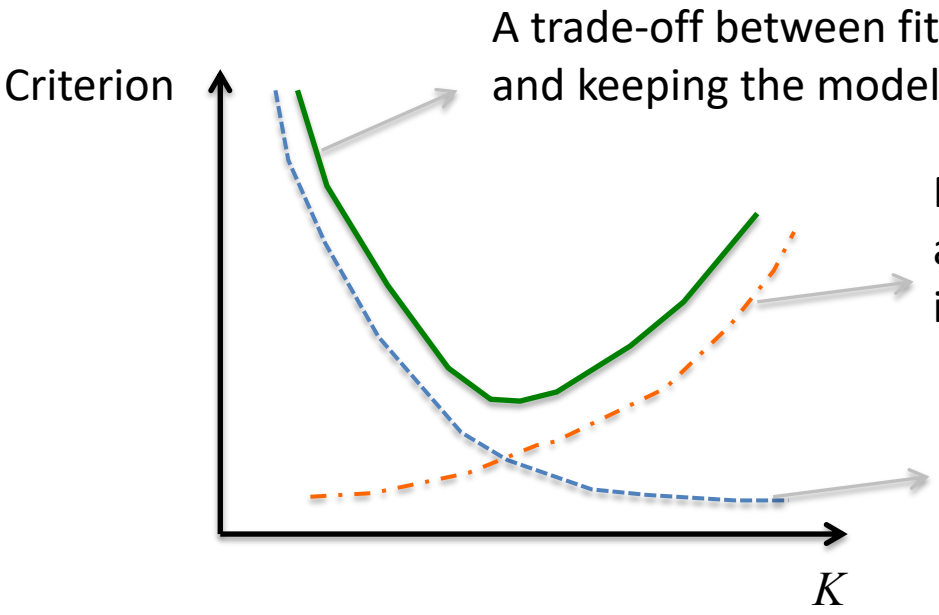
# Model selection in general

Probabilistic model

$$p(X_N \mid \Theta_K)$$

Candidate models:

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K \subseteq \cdots$$




A trade-off between fitting the data well and keeping the model simple

Models become more and more complex as K increases.

Negative log-likelihood (or fitting error) to make sure the model fit the data well.

Akaike's Information Criterion (AIC) $\ln p(X_N \mid \hat{\Theta}_K) - d_k$

Bayesian Information Criterion (BIC) $\ln p(X_N \mid \hat{\Theta}_K) - \frac{1}{2} d_k \ln N$

$d_k$: number of free parameters

$N$: sample size

# Outline

- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood
- General EM algorithm
- Bayesian learning

# The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

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

   and return to step 2.

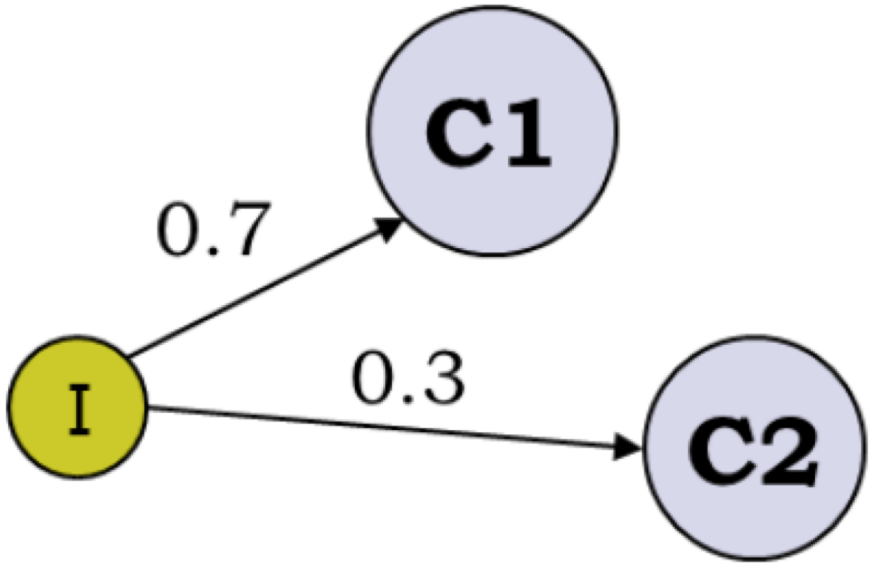
# EM的九层理解

1. EM 就是 E + M
2. EM 是一种局部下限构造
3. K-Means是一种Hard EM算法
4. 从EM 到 广义EM
5. 广义EM的一个特例是VBEM
6. 广义EM的另一个特例是WS算法
7. 广义EM的再一个特例是Gibbs抽样算法
8. WS算法是VAE和GAN组合的简化版
9. KL距离的统一

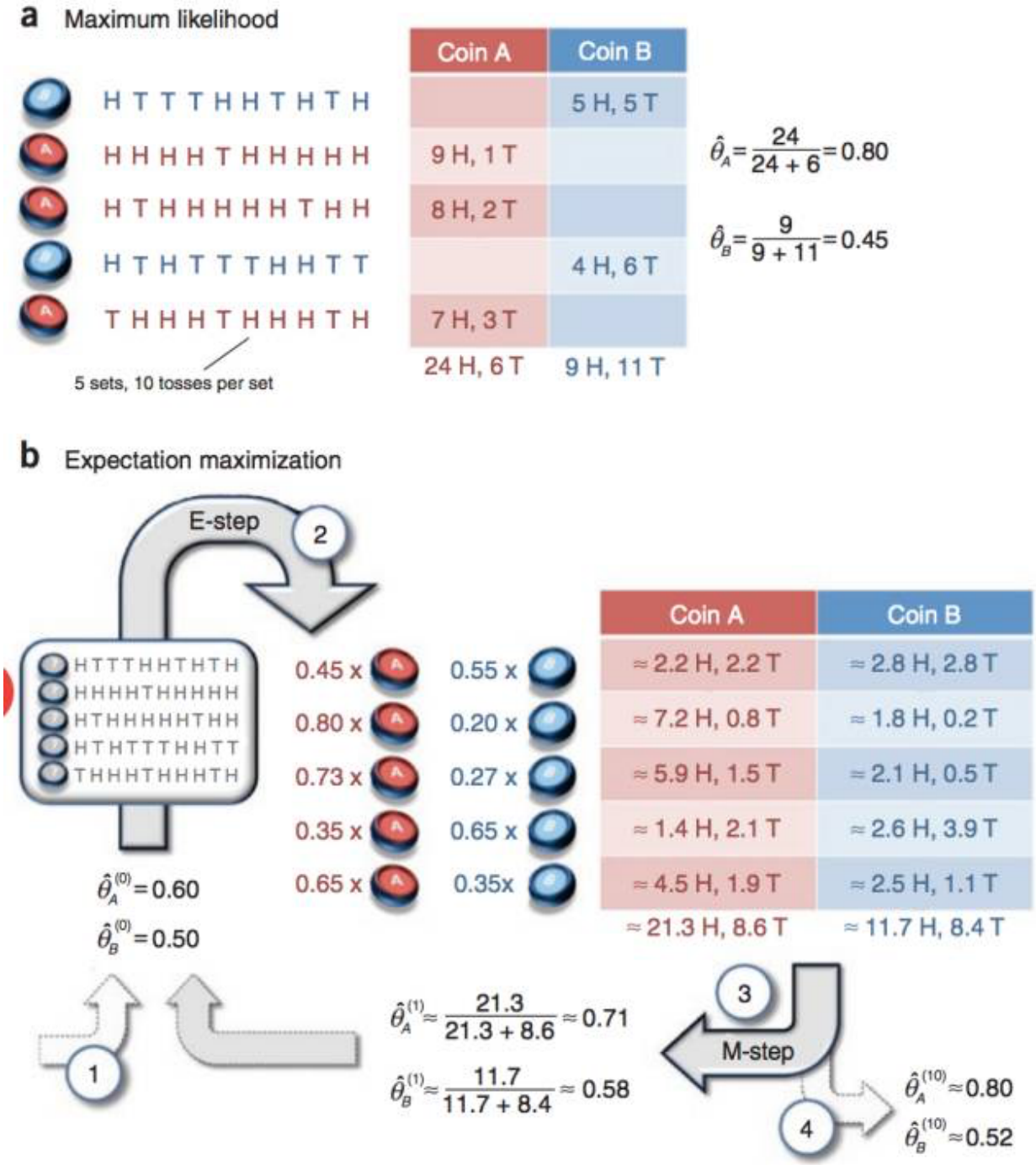
http://www.elecfans.com/d/604076.html

# （1）EM就是E 期望 + M 最大化

- 抛3个硬币，抛 I 硬币决定C1和C2，然后抛C1或者C2决定正反面， 然后估算3个硬币的正反面概率值。




**Bernoulli Mixture Model**


a Maximum likelihood
H T T T H H T H T H
H H H H T H H H H H
H T H H H H H T H H
H T H T T T H H T T
T H H H T H H H T H
5 sets, 10 tosses per set
Coin A
Coin B
5 H, 5 T
9 H, 1 T
8 H, 2 T
4 H, 6 T
7 H, 3 T
24 H, 6 T
9 H, 11 T
$\hat{\theta}_A = \frac{24}{24+6} = 0.80$
$\hat{\theta}_B = \frac{9}{9+11} = 0.45$
b Expectation maximization
E-step
2
HTTTHHTHTH
HHHHTHHHHH
HTHHHHHTHH
HTHTTTHHTT
THHHTHHHTH
0.45 x
0.55 x
0.80 x
0.20 x
0.73 x
0.27 x
0.35 x
0.65 x
0.65 x
0.35x
Coin A
Coin B
≈ 2.2 H, 2.2 T
≈ 2.8 H, 2.8 T
≈ 7.2 H, 0.8 T
≈ 1.8 H, 0.2 T
≈ 5.9 H, 1.5 T
≈ 2.1 H, 0.5 T
≈ 1.4 H, 2.1 T
≈ 2.6 H, 3.9 T
≈ 4.5 H, 1.9 T
≈ 2.5 H, 1.1 T
≈ 21.3 H, 8.6 T
≈ 11.7 H, 8.4 T
$\hat{\theta}_A^{(0)} = 0.60$
$\hat{\theta}_B^{(0)} = 0.50$
1
$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$
$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$
3
M-step
4
$\hat{\theta}_A^{(10)} \approx 0.80$
$\hat{\theta}_B^{(10)} \approx 0.52$

# （2）EM是一种局部下限构造

$$E[f(x)] \geq f(E[x])$$

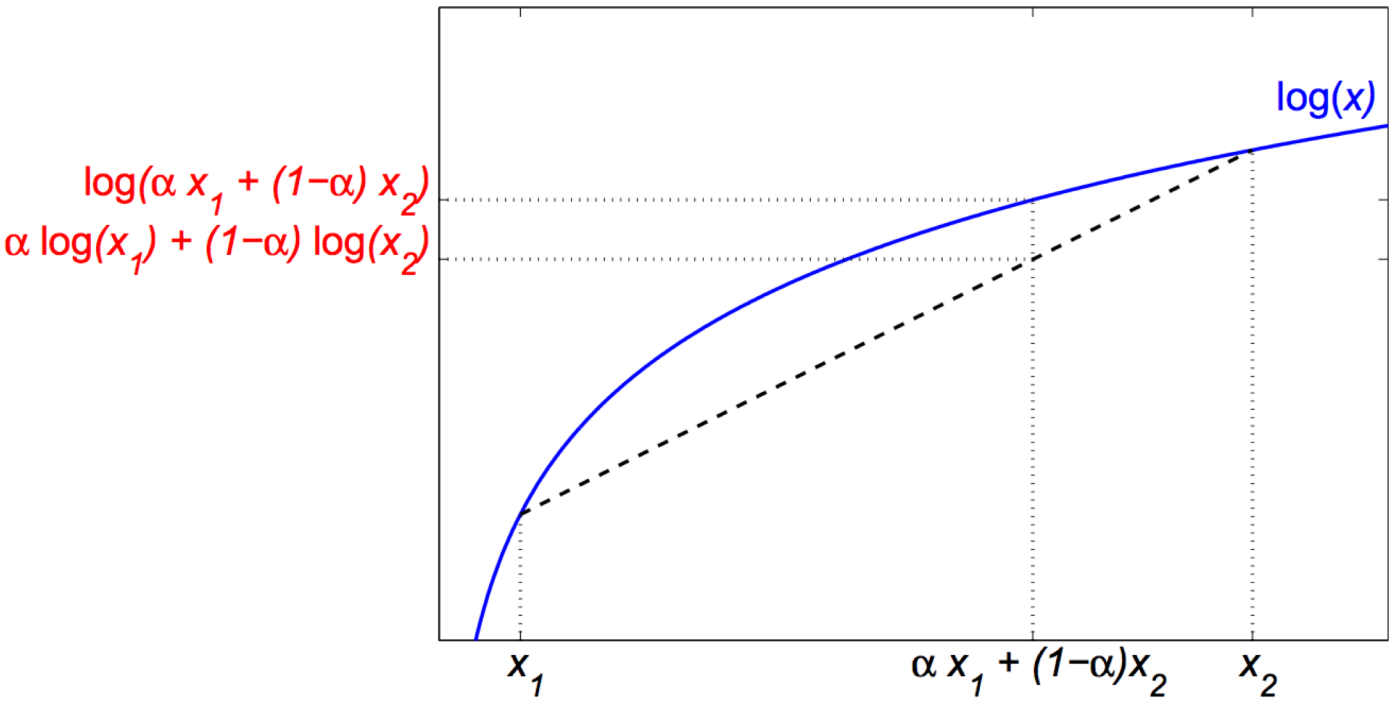
Jensen's Inequality due to convexity

$$\log(P(\mathbf{x}|\theta)) = \log(\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}|\theta))$$

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &= \log(\sum_{\mathbf{y}} q(\mathbf{y}) \frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}) \\
&\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(\frac{P(\mathbf{y}|\mathbf{x},\theta)P(\mathbf{x}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta)] - E_q[\log(\frac{q(\mathbf{y})}{P(\mathbf{y}|\mathbf{x},\theta)}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta))] - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta)) \\
&\geq \log(P(\mathbf{x}|\theta)) - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta))
\end{aligned}$$

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] - E_q[\log(q(\mathbf{y}))] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] + H(q(\mathbf{y}))
\end{aligned}$$

# Jensen's Inequality

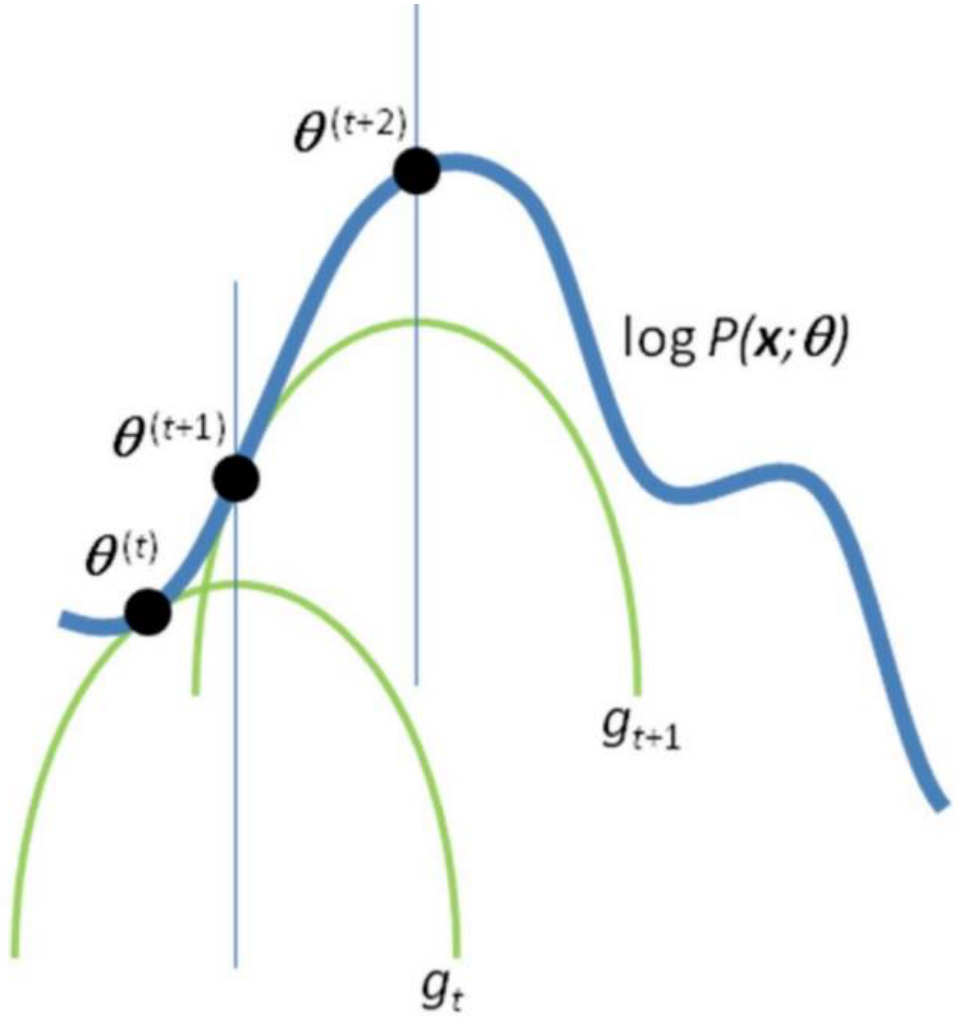



For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

$$\log\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i \log(x_i)$$

Equality if and only if $\alpha_i = 1$ for some $i$ (and therefore all others are 0).

# Maximize the lower bound




先固定当前参数， 计算得到当前隐变量分布的一个下届函数， 然后优化这个函数， 得到新的参数， 然后循环继续

# (3) K-均值方法是一种Hard-cut EM

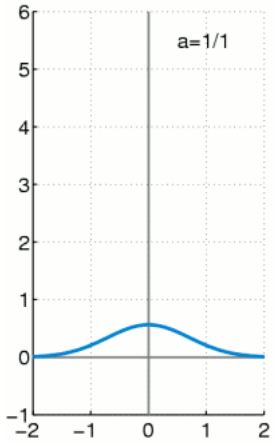
GMM joint distribution

$$P_\Theta(x_1,\ldots,x_n,z_1,\ldots z_n) = \prod_{t=1}^{N} P_\Theta(z_t)P_\Theta(x_t|z_t)$$

$$= \prod_{t=1}^{N} \frac{1}{K}\mathcal{N}(\mu^{z_t},I)(x_t)$$

$$(\mu^1,\ldots,\mu^K)^* = \underset{\mu^1,\ldots,\mu^k}{\operatorname{argmin}} \min_{z_1,\ldots,z_n} \sum_{t=1}^{N} ||\mu^{z_t} - x_t||^2$$

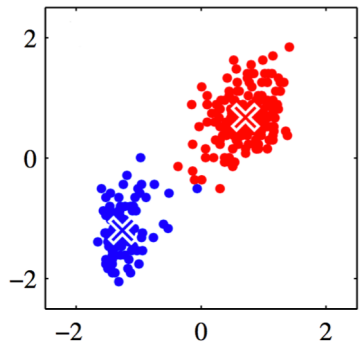
$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$



# K-means is a hard-cut EM

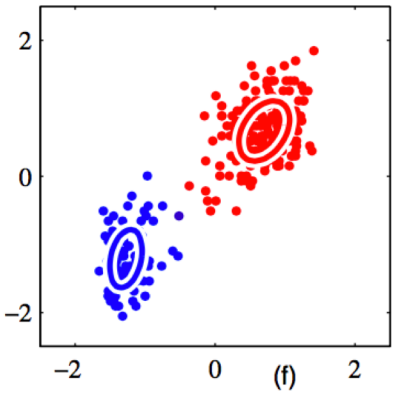
Fixed equal mixing weights

$\|\mathbf{x} - \boldsymbol{\mu}_k\|^2$




$\{\boldsymbol{\mu}_k\}$

$\boldsymbol{\Sigma}_k = \varepsilon \mathbf{I}$




GMM considers covariance and mixing weights.

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}$$

One-in-K assignment

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

Soft assignment

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

# (4) EM 是 广义EM的特例

$$\mathcal{L}(\theta) = \log P(\mathcal{Y}|\theta) = \log \int P(\mathcal{X}, \mathcal{Y}|\theta) d\mathcal{X}$$

$$\mathcal{L}(\theta) = \log \int q(\mathcal{X}) \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} \, d\mathcal{X} \geq \int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} \, d\mathcal{X} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta)$$

$$\begin{aligned} \int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} \, d\mathcal{X} &= \int q(\mathcal{X}) \log P(\mathcal{X}, \mathcal{Y}|\theta) \, d\mathcal{X} - \int q(\mathcal{X}) \log q(\mathcal{X}) \, d\mathcal{X} \\ &= \int q(\mathcal{X}) \log P(\mathcal{X}, \mathcal{Y}|\theta) \, d\mathcal{X} + \mathbf{H}[q], \end{aligned}$$

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{q(\mathcal{X})} + \mathbf{H}[q]$$

# Maximize free energy

自由能：

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q(\mathcal{X})} + \mathbf{H}[q]$$

**E - Step：**

$$q^{(k)}(\mathcal{X}) := \underset{q(\mathcal{X})}{\text{argmax}} \ \mathcal{F}\left(q(\mathcal{X}), \theta^{(k-1)}\right)$$

**M - Step：**

$$\theta^{(k)} := \underset{\theta}{\text{argmax}} \ \mathcal{F}\left(q^{(k)}(\mathcal{X}), \theta\right)$$

$$= \underset{\theta}{\text{argmax}} \ \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q^{(k)}(\mathcal{X})}$$

# Maximize free energy

自由能：

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q(\mathcal{X})} + \mathbf{H}[q]$$

**E - Step：**

$$q^{(k)}(\mathcal{X}) := \underset{q(\mathcal{X})}{\text{argmax}}\ \mathcal{F}\big(q(\mathcal{X}), \theta^{(k-1)}\big)$$

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{X}) \log \frac{P(\mathcal{X},\mathcal{Y}|\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \int q(\mathcal{X}) \log \frac{P(\mathcal{X}|\mathcal{Y},\theta)P(\mathcal{Y}|\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \int q(\mathcal{X}) \log P(\mathcal{Y}|\theta)\, d\mathcal{X} + \int q(\mathcal{X}) \log \frac{P(\mathcal{X}|\mathcal{Y},\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \mathcal{L}(\theta) - \mathbf{KL}[q(\mathcal{X}) \| P(\mathcal{X}|\mathcal{Y},\theta)]
\end{aligned}$$

$$q^{(k)}(\mathcal{X}) = P(\mathcal{X}|\mathcal{Y},\theta^{(k-1)})$$

## The $\mathbf{KL}[q(x)\|p(x)]$ is non-negative and zero iff $\forall x:\ p(x) = q(x)$

First let's consider discrete distributions; the Kullback-Liebler divergence is:

$$\mathbf{KL}[q\|p] = \sum_i q_i \log \frac{q_i}{p_i}.$$

To find the distribution $q$ which minimizes $\mathbf{KL}[q\|p]$ we add a Lagrange multiplier to enforce the normalization constraint:

$$E \stackrel{\text{def}}{=} \mathbf{KL}[q\|p] + \lambda\big(1 - \sum_i q_i\big) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda\big(1 - \sum_i q_i\big)$$

We then take partial derivatives and set to zero:

$$\left.\begin{aligned} \frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1) \\ \frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1 \end{aligned}\right\} \Rightarrow q_i = p_i.$$

Check that the curvature (Hessian) is positive (definite), corresponding to a minimum:

$$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \qquad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,$$

showing that $q_i = p_i$ is a genuine minimum.

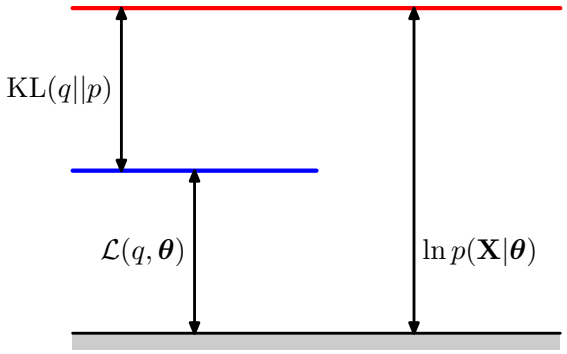
At the minimum is it easily verified that $\mathbf{KL}[p\|p] = 0$.

# EM never decreases the likelihood

- The E-Step brings the free energy to the likelihood.
- The M-Step maximizes the free energy with respect to $\Theta$.

$$\mathcal{L}\left(\theta^{(k-1)}\right) \underset{\text{E step}}{=} \mathcal{F}\left(q^{(k)}, \theta^{(k-1)}\right) \underset{\text{M step}}{\leq} \mathcal{F}\left(q^{(k)}, \theta^{(k)}\right) \underset{\text{Jensen}}{\leq} \mathcal{L}\left(\theta^{(k)}\right)$$

# EM never decreases the likelihood




New log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t+1)})]$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x,\theta^{(t+1)})]$

New lower bound

$\mathcal{F}(q_y^{(t+1)},\theta^{(t+1)})$

Log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$

$\mathrm{KL}[q_y^{(t)}||p(y|x,\theta^{(t)})]$

$\mathcal{F}(q_y^{(t)},\theta^{(t)})$

Lower bound

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$

$= \mathcal{F}(q_y^{(t+1)},\theta^{(t)})$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x,\theta^{(t)})] = 0$

(t)

**E-Step**

**M-Step**

(t+1)

# (5) 广义EM的一个特例是VBEM

$$
\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{X}) \log \frac{P(\mathcal{X},\mathcal{Y}|\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \int q(\mathcal{X}) \log \frac{P(\mathcal{X}|\mathcal{Y},\theta)P(\mathcal{Y}|\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \int q(\mathcal{X}) \log P(\mathcal{Y}|\theta)\, d\mathcal{X} + \int q(\mathcal{X}) \log \frac{P(\mathcal{X}|\mathcal{Y},\theta)}{q(\mathcal{X})}\, d\mathcal{X} \\
&= \mathcal{L}(\theta) - \mathbf{KL}[q(\mathcal{X}) \| P(\mathcal{X}|\mathcal{Y},\theta)]
\end{aligned}
$$

$$
q^{(k)}(\mathcal{X}) = P(\mathcal{X}|\mathcal{Y},\theta^{(k-1)})
$$

$$
\mathbf{KL}[q(\mathcal{X}) \| P(\mathcal{X}|\mathcal{Y},\theta)] = 0
$$

However, if there are **constraints** on the variational posterior *q*, the **KL = 0** may not be achieved.

# Variational Bayes (VB)

Marginal likelihood

$$\begin{aligned}\ln p(\mathbf{y}\,|\,m) &= \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; p(\mathbf{x},\mathbf{y},\boldsymbol{\theta}\,|\,m) \\ &= \ln \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x},\boldsymbol{\theta}) \frac{p(\mathbf{x},\mathbf{y},\boldsymbol{\theta}\,|\,m)}{q(\mathbf{x},\boldsymbol{\theta})} \\ &\geq \int d\boldsymbol{\theta}\, d\mathbf{x}\; q(\mathbf{x},\boldsymbol{\theta}) \ln \frac{p(\mathbf{x},\mathbf{y},\boldsymbol{\theta}\,|\,m)}{q(\mathbf{x},\boldsymbol{\theta})}\end{aligned}$$

Usually, it is very difficult to calculate the Bayesian posterior for the joint distribution of hidden variables and parameters. We may consider the factorized form instead:

$$q(\mathbf{x},\boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

$$\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x},\mathbf{y},\boldsymbol{\theta}\,|\,m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

# VBEM algorithm

自由能：

$$\mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) = \int d\boldsymbol{\theta}\, d\mathbf{x}\; q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

$$q(\mathbf{x}, \boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

**E - Step：**

**VBE step:** $$q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i) = \frac{1}{\mathcal{Z}_{\mathbf{x}_i}} \exp\left[\int d\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \ln p(\mathbf{x}_i, \mathbf{y}_i \,|\, \boldsymbol{\theta}, m)\right]$$

**M - Step：** $$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = \prod_{i=1}^{n} q_{\mathbf{x}_i}^{(t+1)}(\mathbf{x}_i)$$

**VBM step:** $$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}} p(\boldsymbol{\theta} \,|\, m) \exp\left[\int d\mathbf{x}\; q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}, m)\right]$$

# (6) 广义EM的另一个特例是WS算法

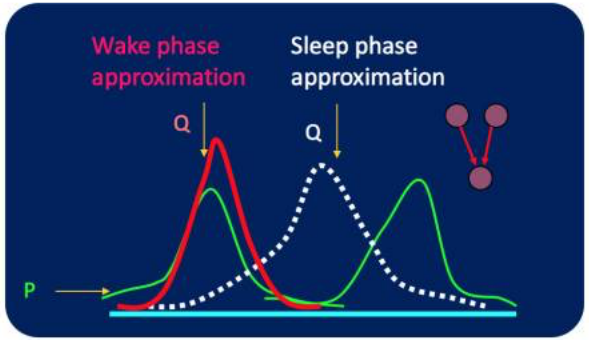
Wake-Sleep Algorithm

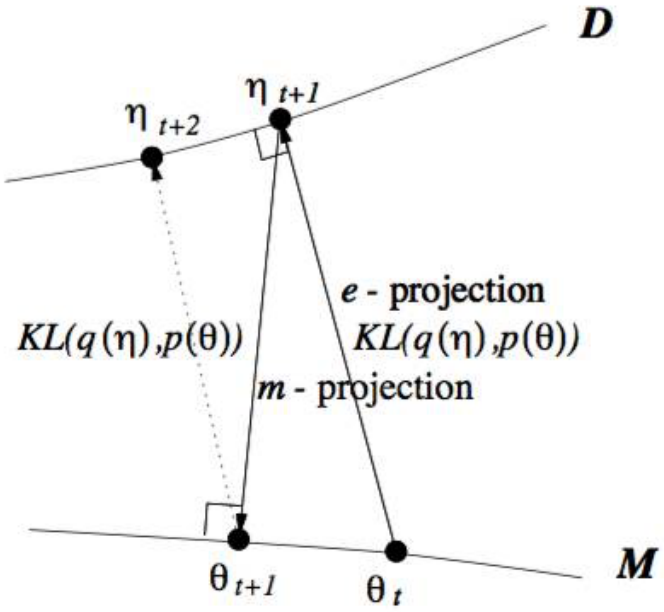
$$\log P[d;\mathcal{G}] \geq \log P[d;\mathcal{G}] - \mathrm{KL}(\mathrm{Q}[\mathrm{h}|\mathrm{d};\mathcal{R}],\mathrm{P}[\mathrm{h}|\mathrm{d};\mathcal{G}])$$

$$\equiv -\mathcal{F}(d;\mathcal{R},\mathcal{G})$$

Free energy

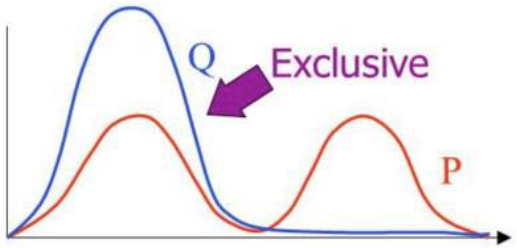
$$\mathcal{F}(d;\mathcal{R},\mathcal{G}) = -\log P[d;\mathcal{G}] + \mathrm{KL}(\mathrm{Q}[\mathrm{h}|\mathrm{d};\mathcal{R}],\mathrm{P}[\mathrm{h}|\mathrm{d};\mathcal{G}])$$



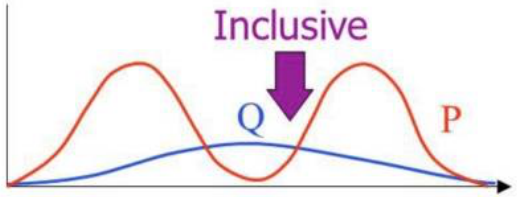



# (7) 广义EM的再一个特例是Gibbs Sampling

Minimising
$KL(Q||P)$
$= \sum_H Q(H) \ln \frac{Q(H)}{P(H|V)}$




Minimising
$KL(P||Q)$
$= \sum_H P(H|V) \ln \frac{P(H|V)}{Q(H)}$




KL 散度：

$$D_{KL}(f||g) = \int_{-\infty}^{\infty} f(x) \log \left( \underbrace{\frac{f(x)}{g(x)}}_{=:r} \right) dx,$$

抽样估算：

$$\widehat{D_{KL}}(f||g) = \frac{1}{N} \sum_i^N \log \left( \frac{f_u(x_i)}{g_u(x_i)} \right) + \log(\hat{r})$$

Importance Sampling：

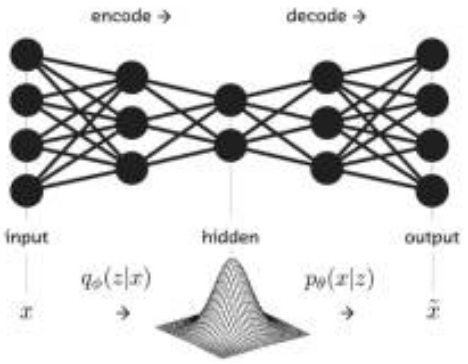
$$\hat{r} = \frac{1/n}{1/n} \frac{\sum_j f_u(x_j)/\pi_f(x_j)}{\sum_j g_u(x_j)/\pi_g(x_j)}.$$

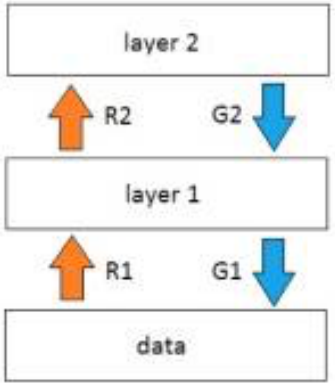
# (8) WS算法是VAE和GAN组合的简化版

**Wake 阶段**

自下而上从数据学习
认知模型

最大KL(Q, P)，衍生到VAE

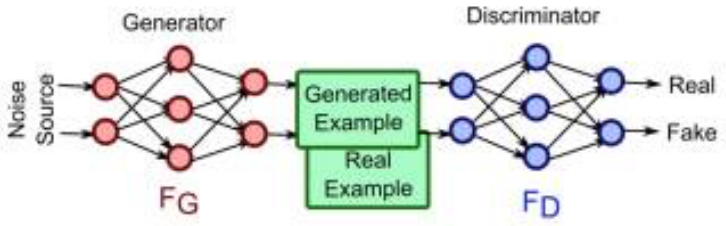



**Sleep 阶段**

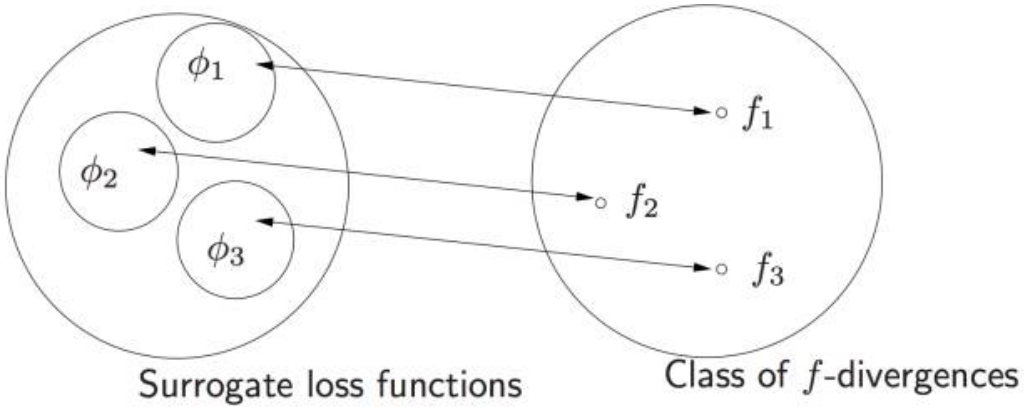
自上而下从模型
再生数据

最大KL(P, Q)，衍生到GAN



# (9) KL距离的统一

- 正反KL距离全部统一到 f 散度的框架




$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]$$

# Outline

- Gaussian Mixture Models (GMM)
- Expectation-Maximization (EM) for maximum likelihood
- General EM algorithm
- Bayesian learning

# Bayesian learning

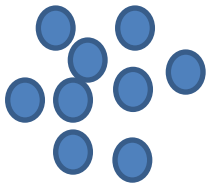
- Maximum A Posteriori (MAP)

$$\max_{\Theta} p(\Theta|X)$$

Equivalent to:

$$\log p(X, \Theta) = \log p(X|\,\Theta) + \log p(\,\Theta)$$

Consider a simple example:



$$p(x|\Theta) = G(x|\mu, \Sigma)$$

$$p(\mu) = G(\mu|\mu_0, \sigma_0^2)$$

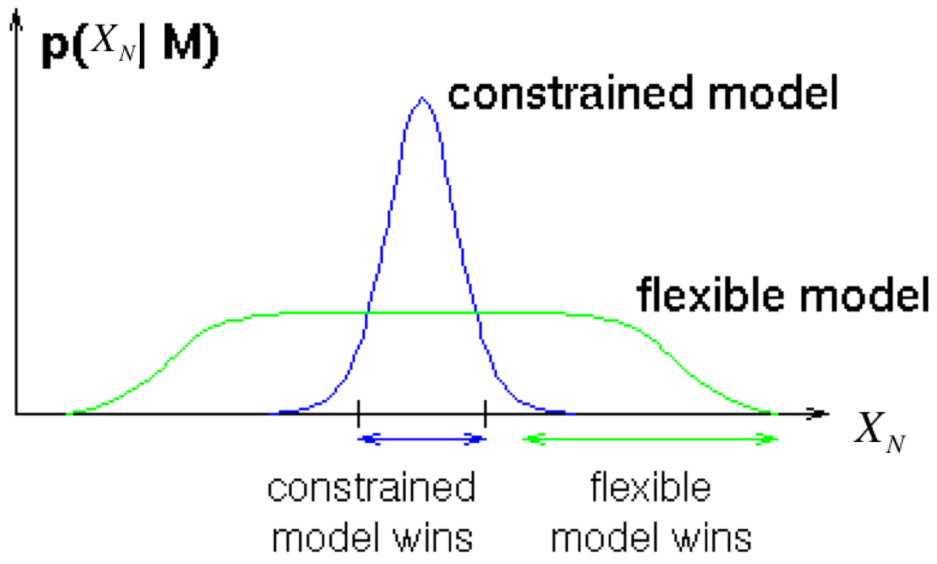
# Model selection

Probabilistic model

$$p(X_N \mid \Theta_K)$$

Candidate models:

$$\Theta_1 \subseteq \Theta_2 \subseteq \cdots \subseteq \Theta_K \subseteq \cdots$$



# Using Occam's Razor to Learn Model Structure

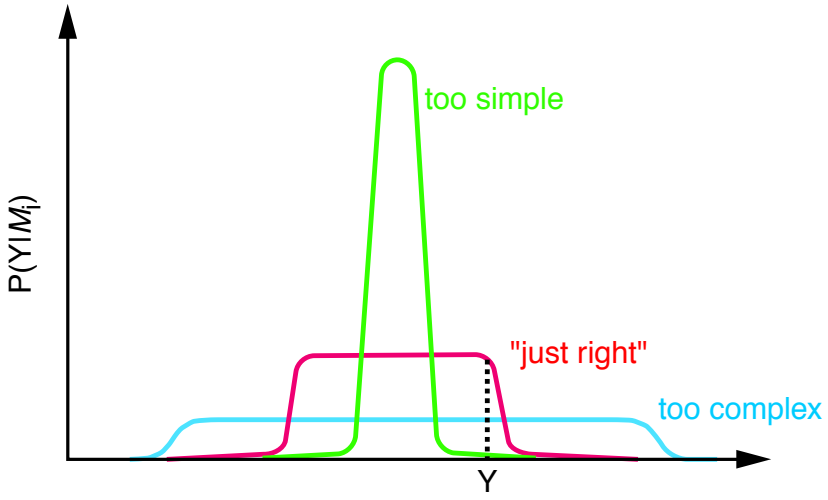
Compare model classes $m$ using their posterior probability given the data:

$$P(m|\mathbf{y}) = \frac{P(\mathbf{y}|m)P(m)}{P(\mathbf{y})}, \qquad P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\boldsymbol{\theta}_m, m)P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

**Interpretation of $P(\mathbf{y}|m)$:** The probability that *randomly selected* parameter values from the model class would generate data set $\mathbf{y}$.

Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



# Bayesian model selection

- A model class $m$ is a set of models parameterised by $\boldsymbol{\theta}_m$, e.g. the set of all possible mixtures of $m$ Gaussians.

- The marginal likelihood of model class $m$:

$$P(\mathbf{y}|m) = \int_{\Theta_m} P(\mathbf{y}|\boldsymbol{\theta}_m, m) P(\boldsymbol{\theta}_m|m) \, d\boldsymbol{\theta}_m$$

  is also known as the Bayesian evidence for model $m$.

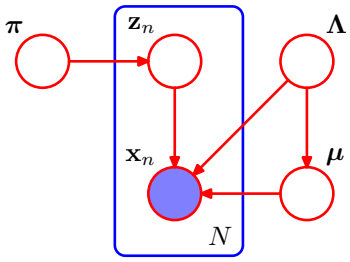
- The ratio of two marginal likelihoods is known as the Bayes factor:

$$\frac{P(\mathbf{y}|m)}{P(\mathbf{y}|m')}$$

- The Occam's Razor principle is, roughly speaking, that one should prefer simpler explanations than more complex explanations.

- Bayesian inference formalises and *automatically* implements the Occam's Razor principle.

# VBEM for GMM




- Model descriptions:

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}} \qquad p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

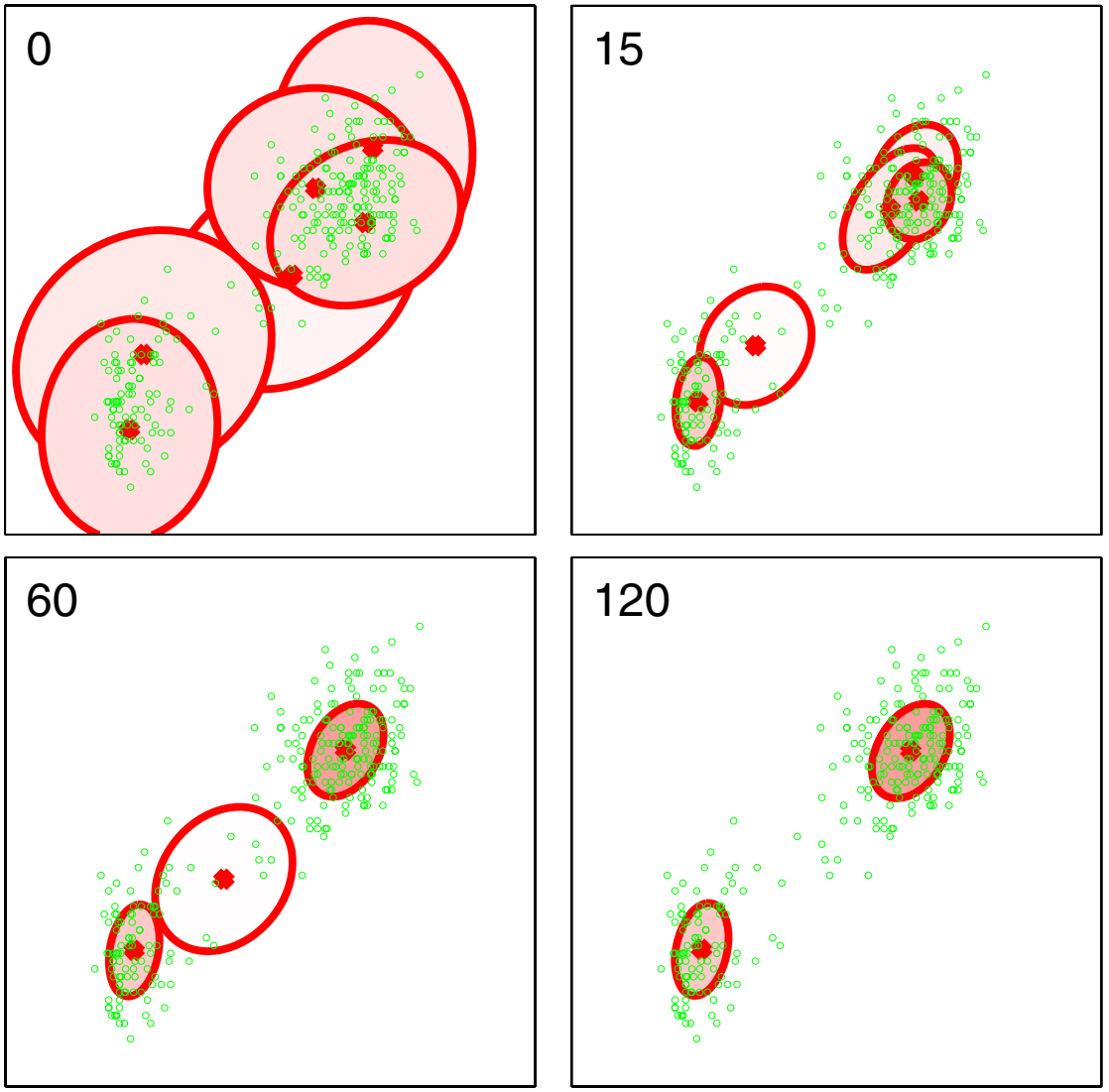
- Prior distributions over parameters:

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

$$\begin{aligned} p(\boldsymbol{\mu},\boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0,\nu_0) \end{aligned}$$

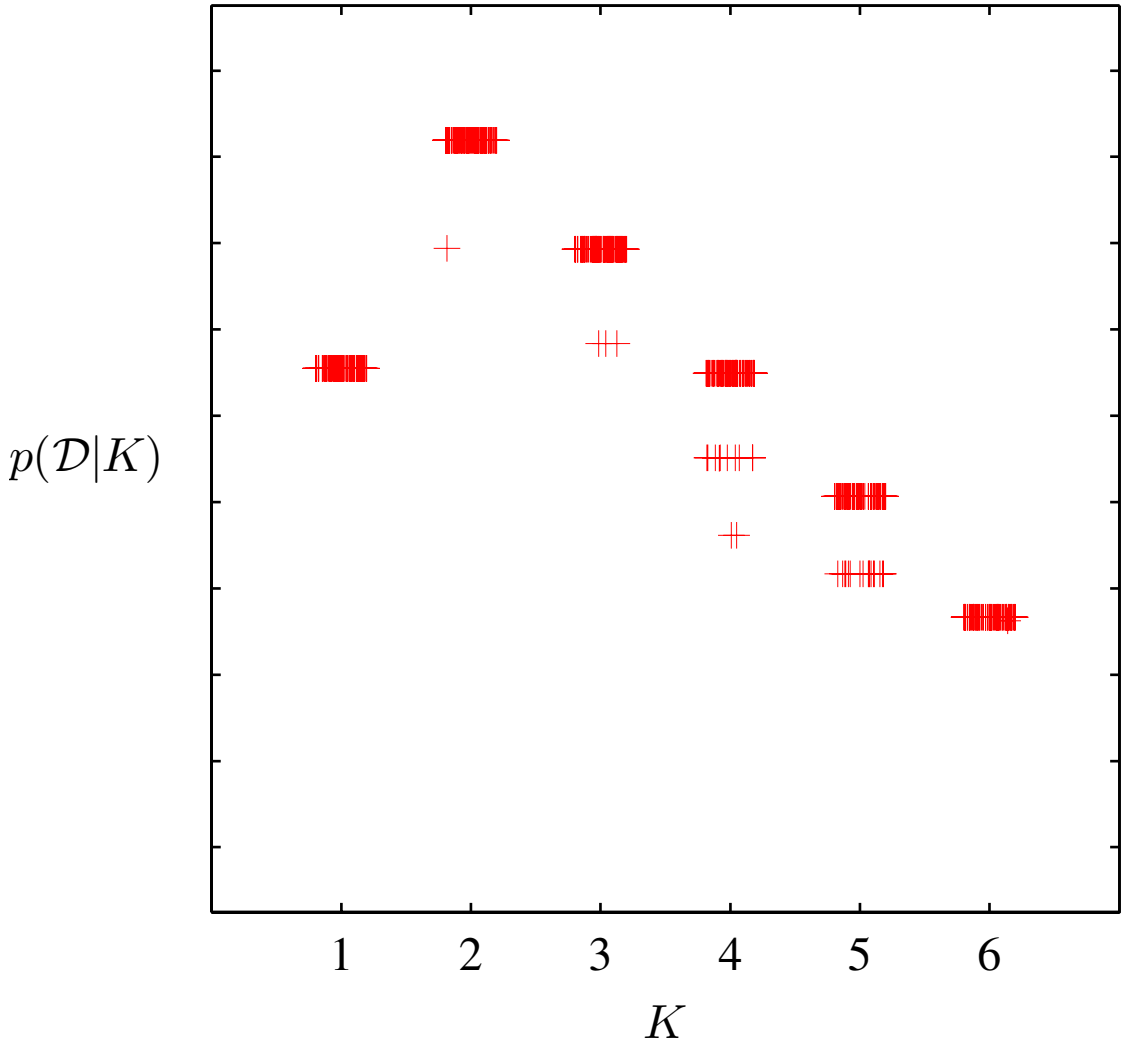
$$p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

# How VBEM for GMM works




http://www.cs.ubc.ca/~murphyk/Software/VBEMGMM/index.html
http://scikit-learn.org/stable/modules/mixture.html

# Determine K by the variational lower bound (free energy)



Thank you!