An Introduction to Statistical Machine Learning - EM for GMMs -

Samy Bengio

bengio@idiap.ch

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)

CP 592, rue du Simplon 4

1920 Martigny, Switzerland

http://www.idiap.ch/~bengio

Gaussian Mixture Models and EM

1. Reminder: Basics on Probabilities
2. What is a GMM
3. Basics of EM
4. Convergence of EM
5. EM for GMMs
6. Initialization

Reminder: Basics on Probabilities

A few basic equalities that are often used:

1. (conditional probabilities)

$$P(A, B) = P(A|B) \cdot P(B)$$

2. (Bayes rule)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

3. If $(\bigcup B_i = \Omega)$ and $\forall i, j \neq i \ \ (B_i \bigcap B_j = \emptyset)$ then

$$P(A) = \sum_i P(A, B_i)$$

What is a Gaussian Mixture Model

- A Gaussian Mixture Model (GMM) is a distribution
- The likelihood given a Gaussian distribution is

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{|x|}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

  where μ is the mean and Σ is the covariance matrix of the Gaussian. Σ is often diagonal.

- The likelihood given a GMM is

$$p(x) = \sum_{i=1}^{N} w_i \cdot \mathcal{N}(x;\mu,\Sigma)$$

  where N is the number of Gaussians and w_i is the weight of Gaussian i, with

$$\sum_i w_i = 1 \text{ and } \forall i : w_i \geq 0$$

Characteristics of a GMM

- While ANNs are universal approximators of functions,
- GMMs are universal approximators of densities.

  (as long as there are enough Gaussians of course)
- Even diagonal GMMs are universal approximators.
- Full rank GMMs are not easy to handle: number of parameters is the square of the number of dimensions.
- GMMs can be trained by maximum likelihood using an efficient algorithm: Expectation-Maximization.

Basics of Expectation-Maximization

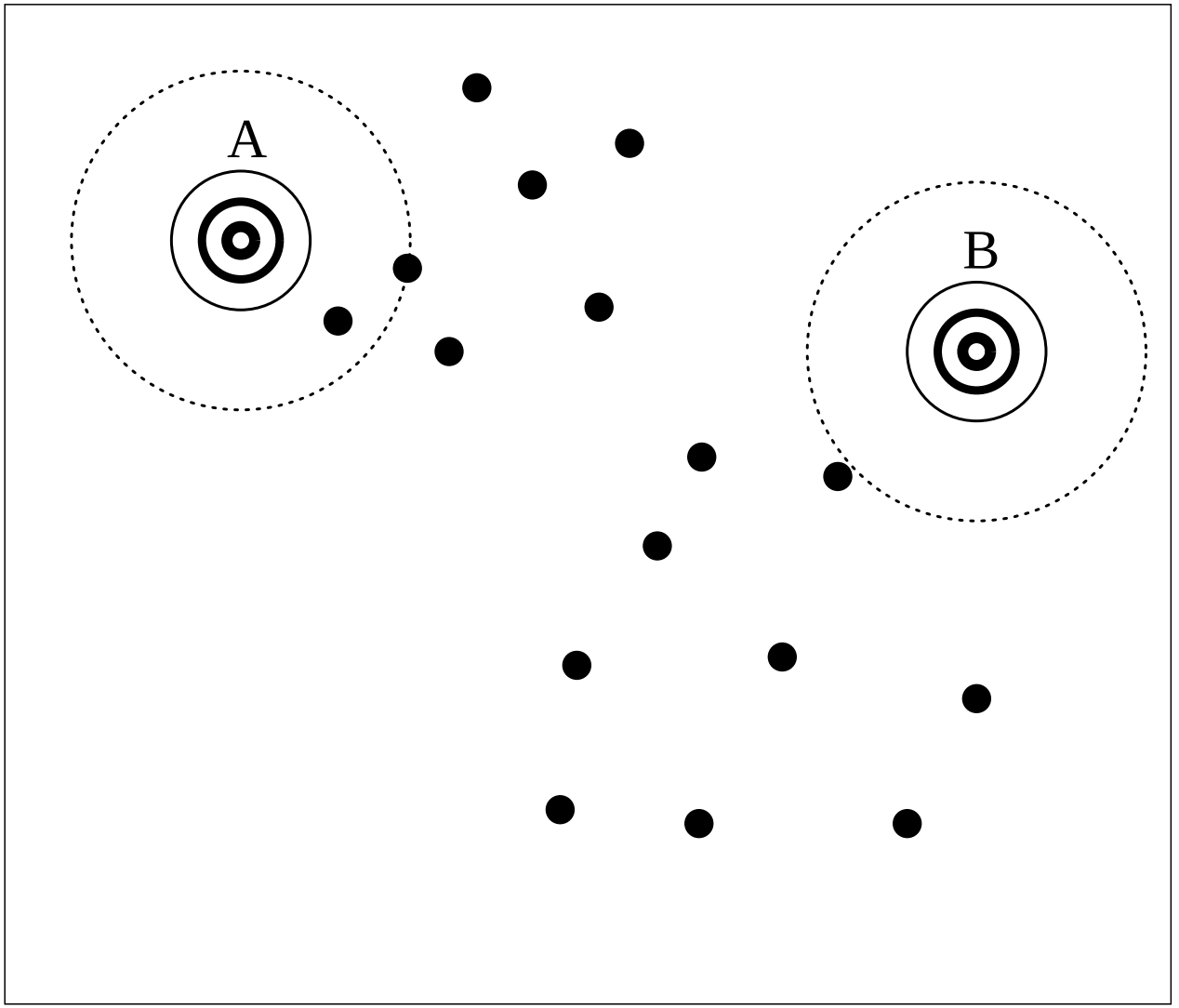
- Objective: maximize the likelihood $p(X;\theta)$ of the data X drawn from an unknown distribution, given the model parameterized by θ:

$$\theta^* = \arg\max_\theta p(X|\theta) = \arg\max_\theta \prod_{p=1}^{n} p(x_p|\theta)$$

- Basic ideas of EM:
  - Introduce a hidden variable such that *its knowledge would simplify the maximization of* $p(X;\theta)$
  - At each iteration of the algorithm:
    - E-Step: estimate the distribution of the hidden variable given the data and the current value of the parameters
    - M-Step: modify the parameters in order to maximize the joint distribution of the data and the hidden variable

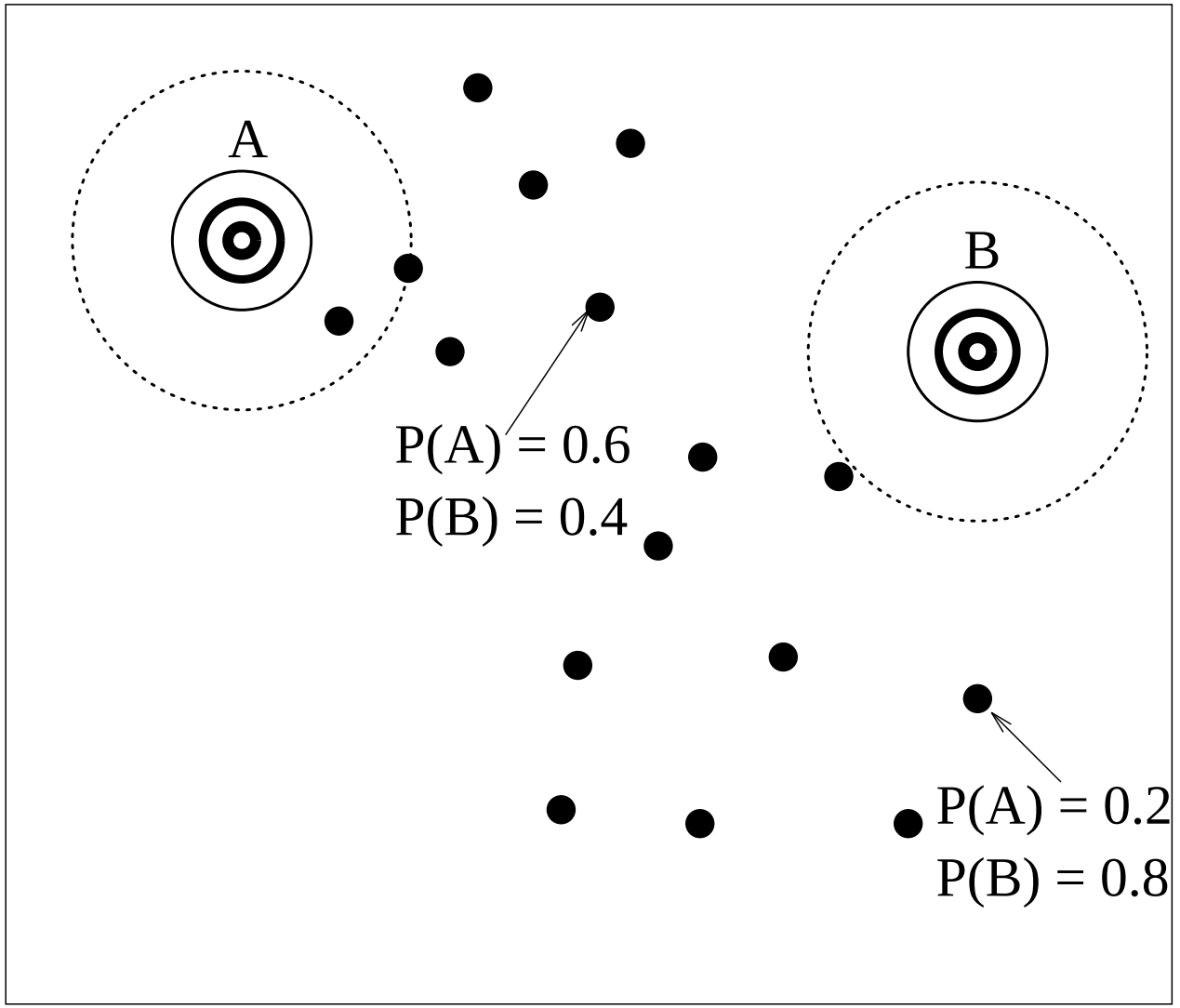
EM for GMM (Graphical View, 1)

Hidden variable: for each point, which Gaussian generated it?



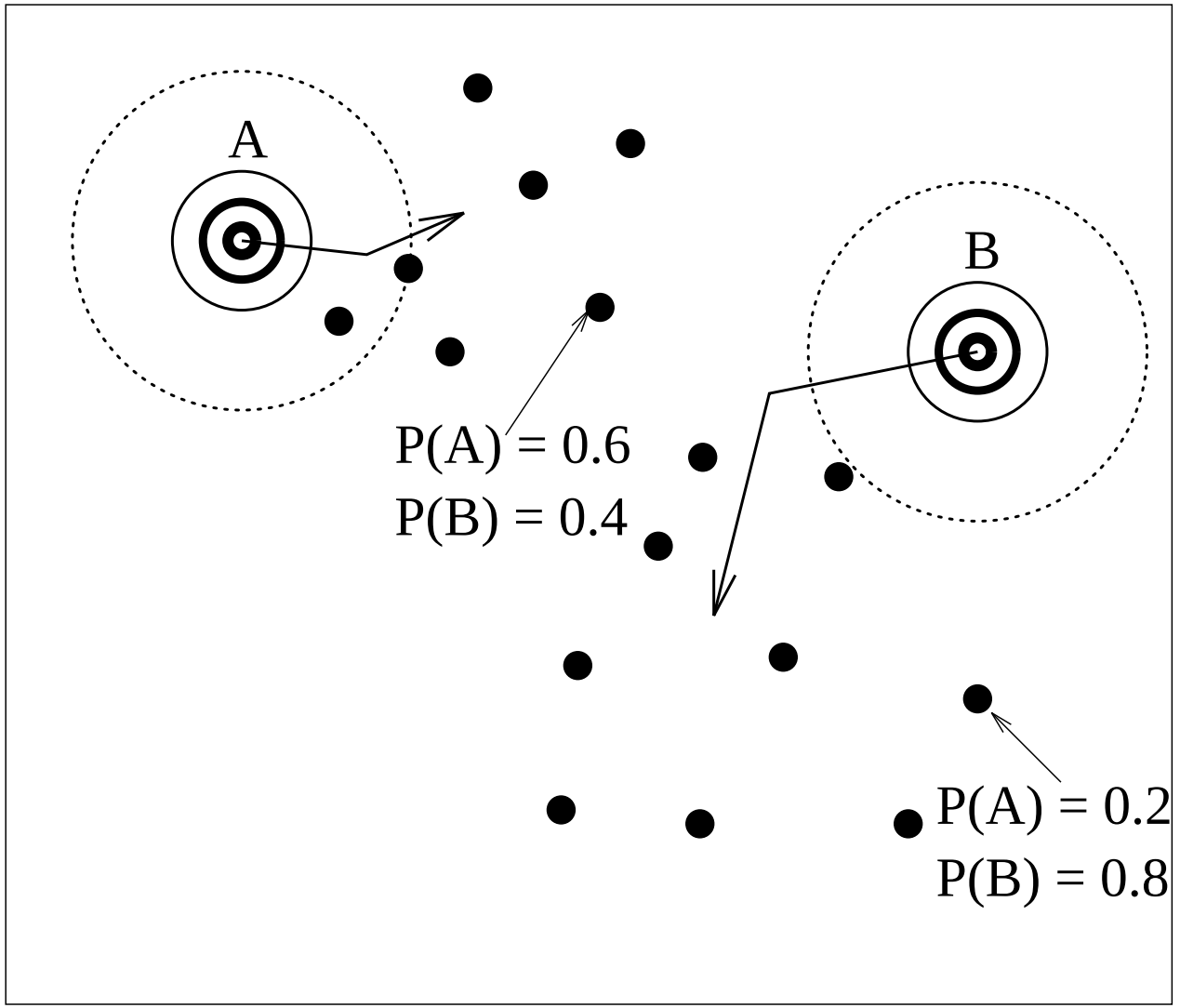
EM for GMM (Graphical View, 2)

E-Step: for each point, estimate the probability the each Gaussian generated it



EM for GMM (Graphical View, 3)

M-Step: modify the parameters according to the hidden variable to maximize the likelihood of the data (and the hidden variable)



EM: More Formally

- Let us call the hidden variable Q.
- Let us consider the following auxiliary function:

$$A(\theta, \theta^s) = E_Q[\log p(X, Q|\theta)|X, \theta^s]$$

- It can be shown that maximizing A

$$\theta^{s+1} = \arg\max_{\theta} A(\theta, \theta^s)$$

always increases the likelihood of the data $p(X|\theta^{s+1})$, and a maximum of A corresponds to a maximum of the likelihood.

EM: Proof of Convergence

- First let us develop the auxiliary function:

$$
\begin{aligned}
A(\theta, \theta^s) &= E_Q[\log p(X, Q|\theta)|X, \theta^s] \\
&= \sum_Q P(Q|X, \theta^s) \log p(X, Q|\theta) \\
&= \sum_Q P(Q|X, \theta^s) \log(P(Q|X, \theta) \cdot p(X|\theta)) \\
&= \left[\sum_Q P(Q|X, \theta^s) \log P(Q|X, \theta)\right] + \left[\sum_Q P(Q|X, \theta^s) \log p(X|\theta)\right] \\
&= \left[\sum_Q P(Q|X, \theta^s) \log P(Q|X, \theta)\right] + \log p(X|\theta)
\end{aligned}
$$

EM: Proof of Convergence

- then if we evaluate it at θ^s

$$A(\theta^s, \theta^s) = \left[\sum_Q P(Q|X, \theta^s) \log P(Q|X, \theta^s) \right] + \log p(X|\theta^s)$$

- the difference between two consecutive log likelihoods of the data can be written as

$$\log p(X|\theta) - \log p(X|\theta^s) = A(\theta, \theta^s) - A(\theta^s, \theta^s) + \sum_Q P(Q|X, \theta^s) \log \frac{P(Q|X, \theta^s)}{P(Q|X, \theta)}$$

- hence,
  - since the last part of the equation is a Kullback-Leibler divergence which is always positive or null,
  - if A increases, the log likelihood of the data also increases
  - Moreover, one can show that when A is maximum, the likelihood of the data is also at a maximum.

EM for GMM: Hidden Variable

- For GMM, the hidden variable Q will describe which Gaussian generated each example.
- If Q was observed, then it would be simple to maximize the likelihood of the data: simply estimate the parameters Gaussian by Gaussian
- Moreover, we will see that we can easily estimate Q
- Let us first write the mixture of Gaussian model for one x_i:

$$p(x_i|\theta) = \sum_{j=1}^{N} P(j|\theta)p(x_i|j,\theta)$$

- Let us now introduce the following indicator variable:

$$z_{i,j} = \begin{cases} 1 & \text{if Gaussian } j \text{ emitted } x_i \\ 0 & \text{otherwise} \end{cases}$$

EM for GMM: Auxiliary Function

- We can now write the joint likelihood of all the X and Q:

$$p(X, Q|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{N} P(j|\theta)^{z_{i,j}} p(x_i|j, \theta)^{z_{i,j}}$$

- which in log gives

$$\log p(X, Q|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{N} z_{i,j} \log P(j|\theta) + z_{i,j} \log p(x_i|j, \theta)$$

EM for GMM: Auxiliary Function

- Let us now write the corresponding auxiliary function:

$$
\begin{aligned}
A(\theta, \theta^s) &= E_Q[\log p(X, Q|\theta)|X, \theta^s] \\
&= E_Q\left[\sum_{i=1}^{n}\sum_{j=1}^{N} z_{i,j} \log P(j|x_i, \theta) + z_{i,j} \log p(x_i|\theta)|X, \theta^s\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{N} E_Q[z_{i,j}|X, \theta^s] \log P(j|x_i, \theta) + E[z_{i,j}|X, \theta^s] \log p(x_i|\theta)
\end{aligned}
$$

EM for GMM: E-Step and M-Step

$$A(\theta, \theta^s) = \sum_{i=1}^{n} \sum_{j=1}^{N} E_Q[z_{i,j}|X, \theta^s] \log P(j|x_i, \theta) + E[z_{i,j}|X, \theta^s] \log p(x_i|\theta)$$

- Hence, the E-Step estimates the posterior:

$$\begin{aligned} E_Q[z_{i,j}|X, \theta^s] &= 1 \cdot P(z_{i,j} = 1|X, \theta^s) + 0 \cdot P(z_{i,j} = 0|X, \theta^s) \\ &= P(j|x_i, \theta^s) = \frac{p(x_i|j, \theta^s) P(j|\theta^s)}{p(x_i|\theta^s)} \end{aligned}$$

- and the M-step finds the parameters θ that maximizes A, hence searching for

$$\frac{\partial A}{\partial \theta} = 0$$

for each parameter (μ_j, variances σ_j^2, and weights w_j).

- Note however that for the weights w_j, we need to enforce their sum to 1: add a Lagrange term.

EM for GMM: M-Step for Means

$$A(\theta, \theta^s) = \sum_{i=1}^{n} \sum_{j=1}^{N} E_Q[z_{i,j}|X, \theta^s] \log P(j|x_i, \theta) + E[z_{i,j}|X, \theta^s] \log p(x_i|\theta)$$

Let us develop $\frac{\partial A}{\partial \theta} = 0$ for μ_j

$$\begin{aligned}
\frac{\partial A}{\partial \mu_j} &= \sum_{i=1}^{n} \frac{\partial A}{\partial \log p(x_i|\theta)} \frac{\partial \log p(x_i|\theta)}{\partial \mu_j} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \frac{\partial \log p(x_i|\theta)}{\partial \mu_j} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \frac{\partial \log p(x_i|\theta)}{\partial p(x_i|\theta)} \frac{\partial p(x_i|\theta)}{\partial p(x_i|j, \theta)} \frac{\partial p(x_i|j, \theta)}{\partial \log p(x_i|j, \theta)} \frac{\partial \log p(x_i|j, \theta)}{\partial \mu_j} \\
&= \sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{1}{p(x_i|\theta)} \cdot w_j \cdot p(x_i|j, \theta) \cdot \frac{(x_i - \mu_j)}{\sigma_j^2} = 0
\end{aligned}$$

EM for GMM: M-Step for Means

$$\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{1}{p(x_i|\theta)} \cdot w_j \cdot p(x_i|j, \theta) \cdot \frac{(x_i - \mu_j)}{\sigma_j^2} = 0$$

$\Longrightarrow$ (removing constant terms in the sum)

$$\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j, \theta)}{p(x_i|\theta)} \cdot x_i \quad - \quad \sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j, \theta)}{p(x_i|\theta)} \cdot \hat{\mu}_j = 0$$

$$\frac{\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j, \theta)}{p(x_i|\theta)} \cdot x_i}{\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j, \theta)}{p(x_i|\theta)}} \quad = \quad \hat{\mu}_j$$

EM for GMM: Update Rules

- End results:

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} x_i \cdot P(j|x_i, \theta^s) \cdot \frac{p(x_i|j,\theta)}{p(x_i|\theta)}}{\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j,\theta)}{p(x_i|\theta)}}$$

$$(\hat{\sigma}_j)^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu}_j)^2 \cdot P(j|x_i, \theta^s) \cdot \frac{p(x_i|j,\theta)}{p(x_i|\theta)}}{\sum_{i=1}^{n} P(j|x_i, \theta^s) \cdot \frac{p(x_i|j,\theta)}{p(x_i|\theta)}}$$

$$\hat{w}_j = \frac{\sum_{i=1}^{n} w_j \cdot P(j|x_i, \theta^s) \cdot \frac{p(x_i|j,\theta)}{p(x_i|\theta)}}{\sum_{k=1}^{N} \sum_{i=1}^{n} w_k \cdot P(k|x_i, \theta^s) \cdot \frac{p(x_i|k,\theta)}{p(x_i|\theta)}}$$

Initialization

- EM is an iterative procedure that is very sensitive to initial conditions!
- Start from trash $\rightarrow$ end up with trash.
- Hence, we need a good and fast initialization procedure.
- Often used: K-Means.
- Other options: hierarchical K-Means, Gaussian splitting.