8. Expectation-Maximization

SNU
BIOINFORMATICS
INSTITUTE

Table of Contents

- Expectation-Maxmization
  - Parameter estimation
  - Clustering

Parameter estimation with ML

- Two coins are present
- One is randomly selected and tossed 10 times
- Coins are selected 5 times
  - So a total of 5*10 tosses are performed

Probability of coin A tossing a Head

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

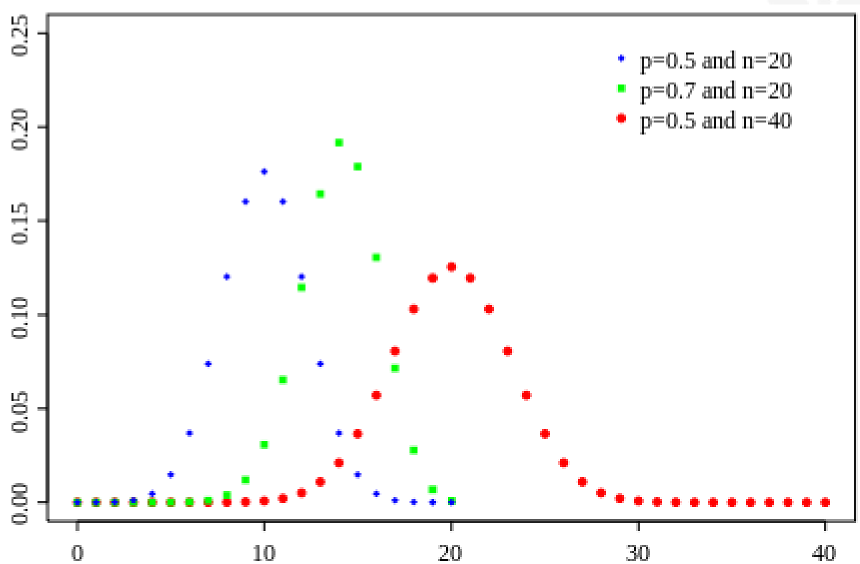
$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

Probability of coin B tossing a Head

SNU BIOINFORMATICS INSTITUTE

Parameter estimation with EM

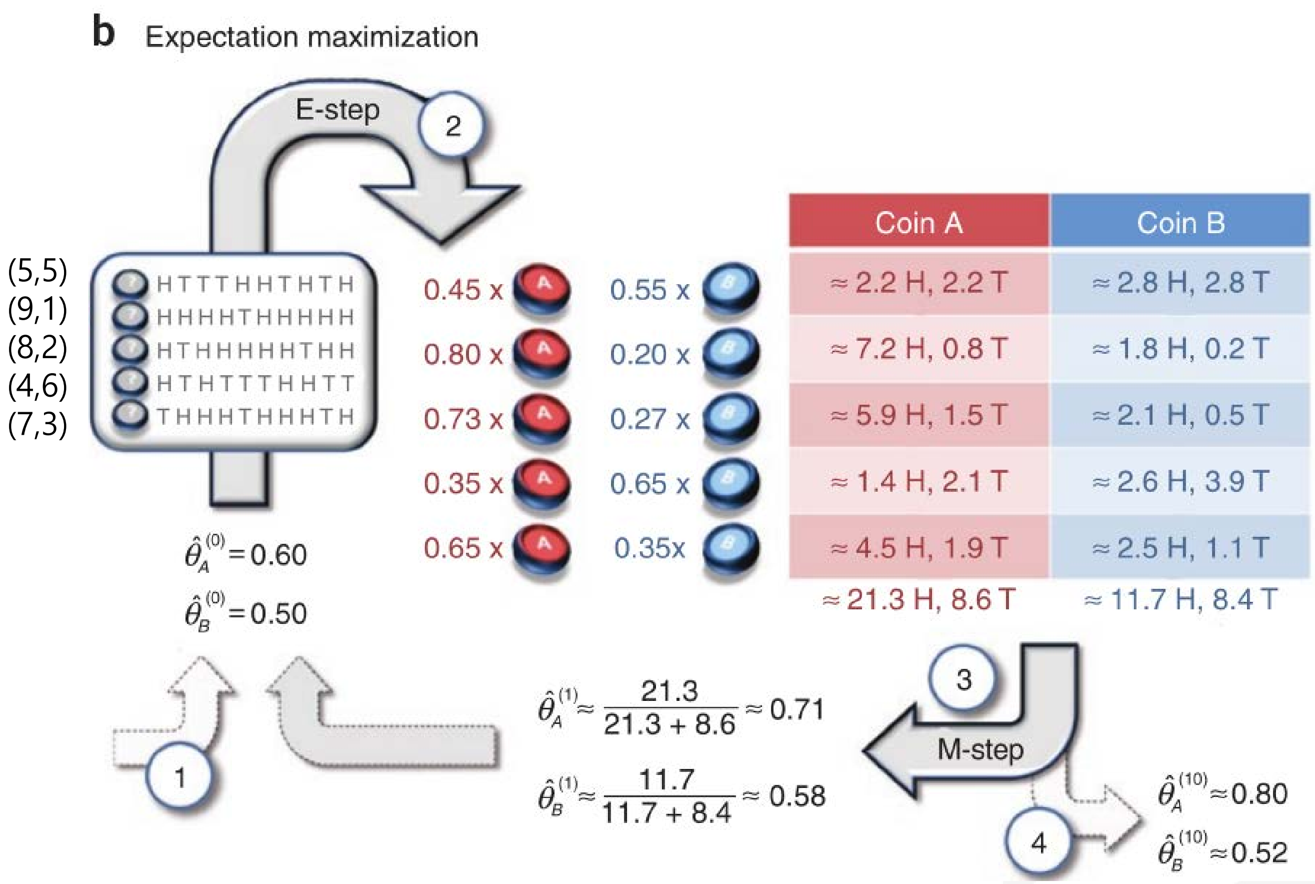
- What if we don't know which coin has been selected?
- The selected coin is the hidden variable (or latent variable)
- EM can be used to estimate $\hat{\theta}_A$ and $\hat{\theta}_B$
- The coin tossing can be modelled by the binomial distribution
  - **Binomial distribution** is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a binary question (yes/no, Head/Tail)



Parameter estimation with EM

- Initial step: Initialize $\hat{\theta}_A$ and $\hat{\theta}_B$
- **E-step:**
  - Using the binomial distribution, calculate the **number of expected Heads and tails** for each Coin ($\hat{\theta}_A$, $\hat{\theta}_B$) by observing the data. Here, the log **likelihoods** are computed using the pdf of binomial distribution.
- **M-step:**
  - Update $\hat{\theta}_A$ and $\hat{\theta}_B$ by the ratio of heads from the expectation values
- Iteratively perform E-M until $\hat{\theta}_A$ and $\hat{\theta}_B$ converges to some threshold value

1st iteration of Coin E-M parameter estimation



[1] Do, Chuong B., and Serafim Batzoglou. "What is the expectation maximization algorithm?." *Nature biotechnology* 26.8 (2008): 897-899.

Practice 1 – Calculating log likelihood of coins

$\hat{\theta}_A$	Probability of coin A tossing a Head

- For $\hat{\theta}_A$, the binomial distribution is as follows
  - $\binom{N}{k} p^k q^{N-k} = \frac{N!}{k!(N-k)!} p^k q^{N-k}$
  - Here, N is the number of trials, k the number of heads
  - p and q are $\hat{\theta}_A$ and 1- $\hat{\theta}_A$
  - Eg) H=5, and $\hat{\theta}_A$=0.6 (initialized value)

$$\mathrm{p}(5|\hat{\theta}_A) = \binom{10}{5} 0.6^5 0.4^{10-5} = \frac{10!}{5!\,(10-5)!} 0.6^5 0.4^5$$

  - Due to many multiplications, we calculate the log-likelikhood
    - $ln(\binom{N}{k} p^k q^{N-k}) = \binom{N}{k} + k * ln(p) + (n-k) * ln(1-p)$

- Similarly, calculate the log-likelihood of $\hat{\theta}_B$

SNU BIOINFORMATICS INSTITUTE

Practice 2 – Calculating the expectations of Head and tails

- Expectation of heads and tails are calculated based on the ratio of log likelihoods of coin A and coin B
- $E(H|\hat{\theta}_A) = W_A \times Observation\ (5,5)$ 5 heads, 5 tails
- $E(H|\hat{\theta}_B) = W_B \times Observation\ (5,5)$
- $W_A = \frac{LL(\hat{\theta}_A)}{LL(\hat{\theta}_A)LL(\hat{\theta}_B)}$ $W_B = \frac{LL(\hat{\theta}_B)}{LL(\hat{\theta}_A)LL(\hat{\theta}_B)}$

- What are the values of W_A and W_b for observation (5, 5)?
- What are the expectation of H and T for $\hat{\theta}_A$ and $\hat{\theta}_B$ for observation (5,5)?

SNU BIOINFORMATICS INSTITUTE

Practice 3 – Maximizing $\hat{\theta}_A$ and $\hat{\theta}_B$

- If you have calculated all the expectation values in the table, we can calculate a new $\hat{\theta}_A$ and $\hat{\theta}_B$ using the same technique as in MLE **(generating the data for the right table is the key concept here)**

When the coins are known

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

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

When the coins are hidden

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

$$\hat{\theta}_A = \frac{21.3}{21.3+8.6}$$

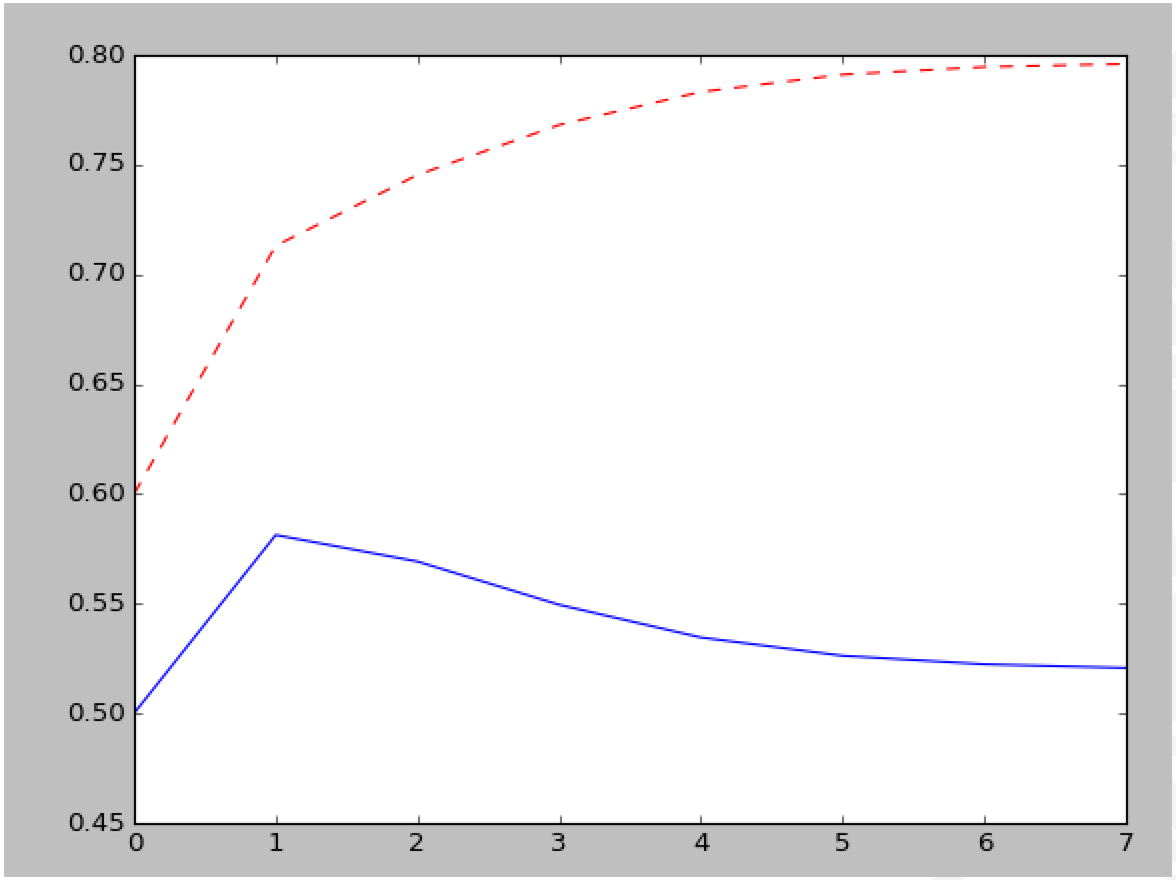
$$\hat{\theta}_B = \frac{11.7}{11.7+8.4}$$

Practice 4 – Improvement of $\hat{\theta}_A$ and $\hat{\theta}_B$

- The difference of $\hat{\theta}_A$ and $\hat{\theta}_B$ at each iteration is measured
- Until the difference becomes smaller than some threshold (i.e., converges), the E-M algorithm stops

- How many iterations are needed to converge with a threshold of 0.001?
- What are the final values of $\hat{\theta}_A$ and $\hat{\theta}_B$?

Plot



SNU BIOINFORMATICS INSTITUTE

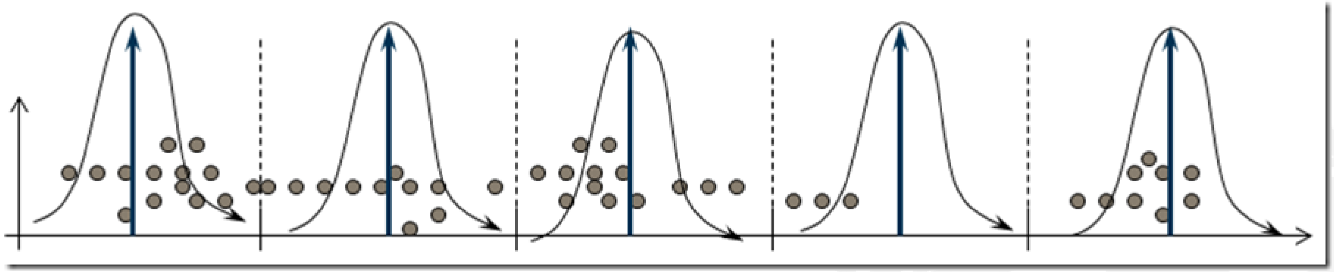
EM-clustering

- As we have seen, EM performs parameter estimation based on a statistical model
  - Binomial distributions (or mixtures)
  - Gaussian distributions
  - Poisson distributions
  - Etc.

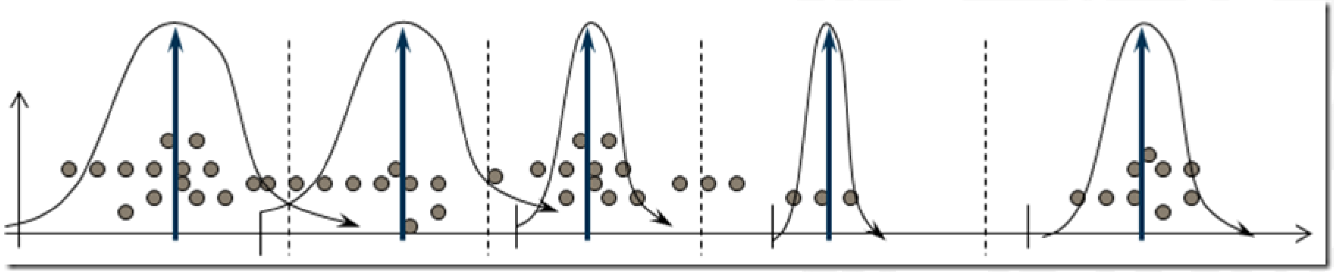
- Calculating centroid of clusters (i.e., parameters) to maximize the probability of statistical models on the data is very popular

EM-clustering example

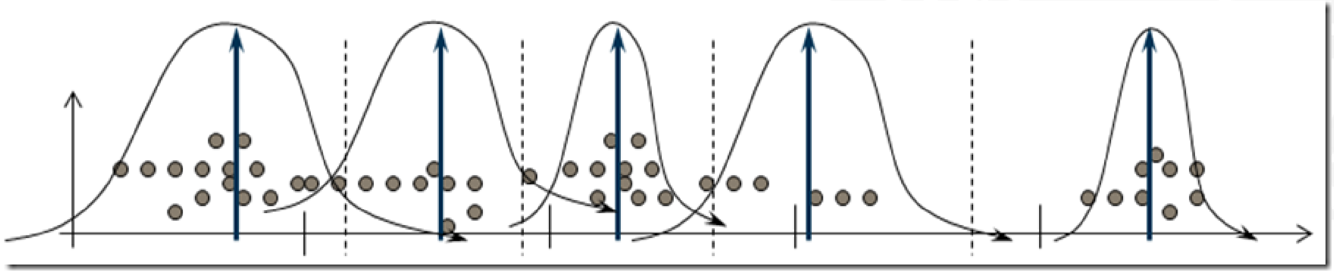
Initialize centroids of 5 clusters



1st iteration of EM



Final iteration of EM



SNU BIOINFORMATICS INSTITUTE

EM-clustering with Iris data

- Clustering is usually performed on unlabeled data
- With labeled data, we can take advantage of label information for improved clustering quality
    - How?
- How well does EM-clustering perform on the iris data?