Expectation Maximization

Thanks to

Dr.Srinivas Battula

Associate Professor

VIT-AP University

Expectation Maximization

- **EM algorithm provides a general approach to learning in presence of unobserved variables.**

- **In many practical learning settings, only a subset of relevant features or variables might be observable. – Eg: Hidden Markov, Bayesian Belief Networks**

- **Estimation: Estimate the expectation from some random data**

- **Maximization: Whatever is estimated should be maximized to find the best result.**

- **From given data EM learn a theory which tells that how each example to be classified and how to predict the feature value of each class.**

Expectation Maximization

Suppose you have 2 coins, A and B, each with a certain bias of landing heads, θ_A , θ_B.

Given data sets $X_A = \{x_{1,A}, \ldots, x_{m_A,A}\}$ and $X_B = \{x_{1,B}, \ldots, x_{m_B,B}\}$

Where $x_{i,j} = \begin{cases} 1\,; if\ heads \\ 0\,; otherwise \end{cases}$

No hidden variables – easy solution. $\theta_j = \frac{1}{m_j}\sum_{i=1}^{m_j} x_{i,j}$; sample mean

Example

- Assume that we have two coins, C1 and C2
- Assume the bias of C1 is θ_1

  (i.e., probability of getting heads with C1)
- Assume the bias of C2 is θ_2

  (i.e., probability of getting heads with C2)

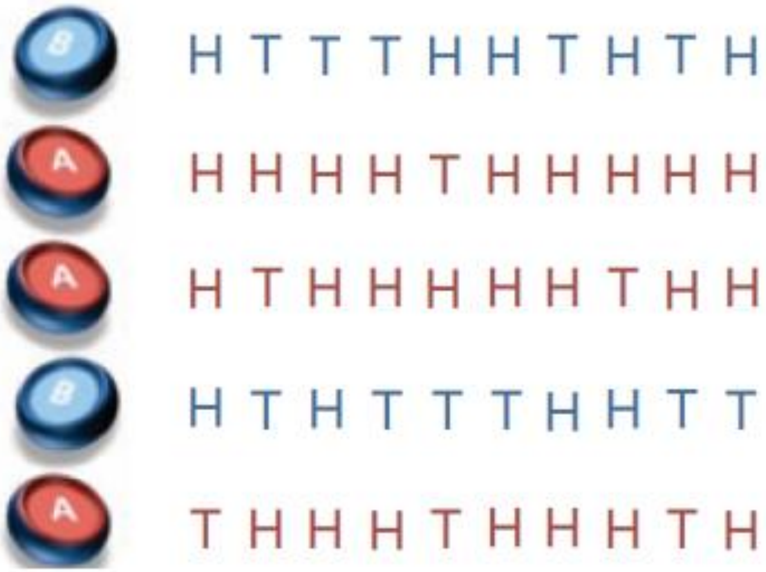
- We want to find θ_1, θ_2 by performing a number of trials

  (i.e., coin tosses)

Example

First experiment

- We choose 5 times one of the coins.
- We toss the chosen coin 10 times



$$\theta_1 = \frac{number\ of\ heads\ using\ C1}{total\ number\ of\ flips\ using\ C1}$$

$$\theta_2 = \frac{number\ of\ heads\ using\ C2}{total\ number\ of\ flips\ using\ C2}$$

Maximum likelyhood:

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

$$\theta_1 = \frac{24}{24+6} = 0.8$$

$$\theta_2 = \frac{9}{9+11} = 0.45$$

Example with Hidden Variable

Assume a more challenging problem

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

•We do not know the identities of the coins used for each set of tosses (we treat them as hidden variables).

Example with Hidden Variable

- What if you were given the same dataset of coin flip results, but no coin identities defining the datasets?

Here: $X = \{x_1, \ldots x_m\}$; the observed variable

$$Z = \begin{Bmatrix} z_{1,1} & \cdots & z_{m,1} \\ \cdots & z_{i,j} & \cdots \\ z_{1,k} & \cdots & z_{m,k} \end{Bmatrix} \text{ where } z_{i,j} = \begin{Bmatrix} 1\,; if\ x_i\ is\ from\ j^{th}\ coin \\ 0; otherwise \end{Bmatrix}$$

But Z is not known. (Ie: ‘hidden’ / ‘latent’ variable)

EM Algorithm

One way to think about this is:

1. Assign random averages to both coins
2. For each of the 5 rounds of 10 coin tosses
   - Check the percentage of heads
   - Find the probability of it coming from each coin
   - Compute the expected number of heads: using that probability as a weight, multiply it by the number of heads
   - Record those numbers
   - Re-Compute new means for coin A and B.
3. With these new means go back to step 2.

Example with Hidden Variable

0) Initialize some arbitrary hypothesis of parameter values (θ):

$\theta = \{ \theta_1, \ldots, \theta_k \}$ coin flip example: $\theta = \{\theta_A, \theta_B\}$ = {0.6, 0.5}

1) Expectation (E-step)

$$E[z_{i,j}] = \frac{p(x = x_i | \theta = \theta_j)}{\sum_{n=1}^{k} p(x = x_i | \theta = \theta_n)}$$

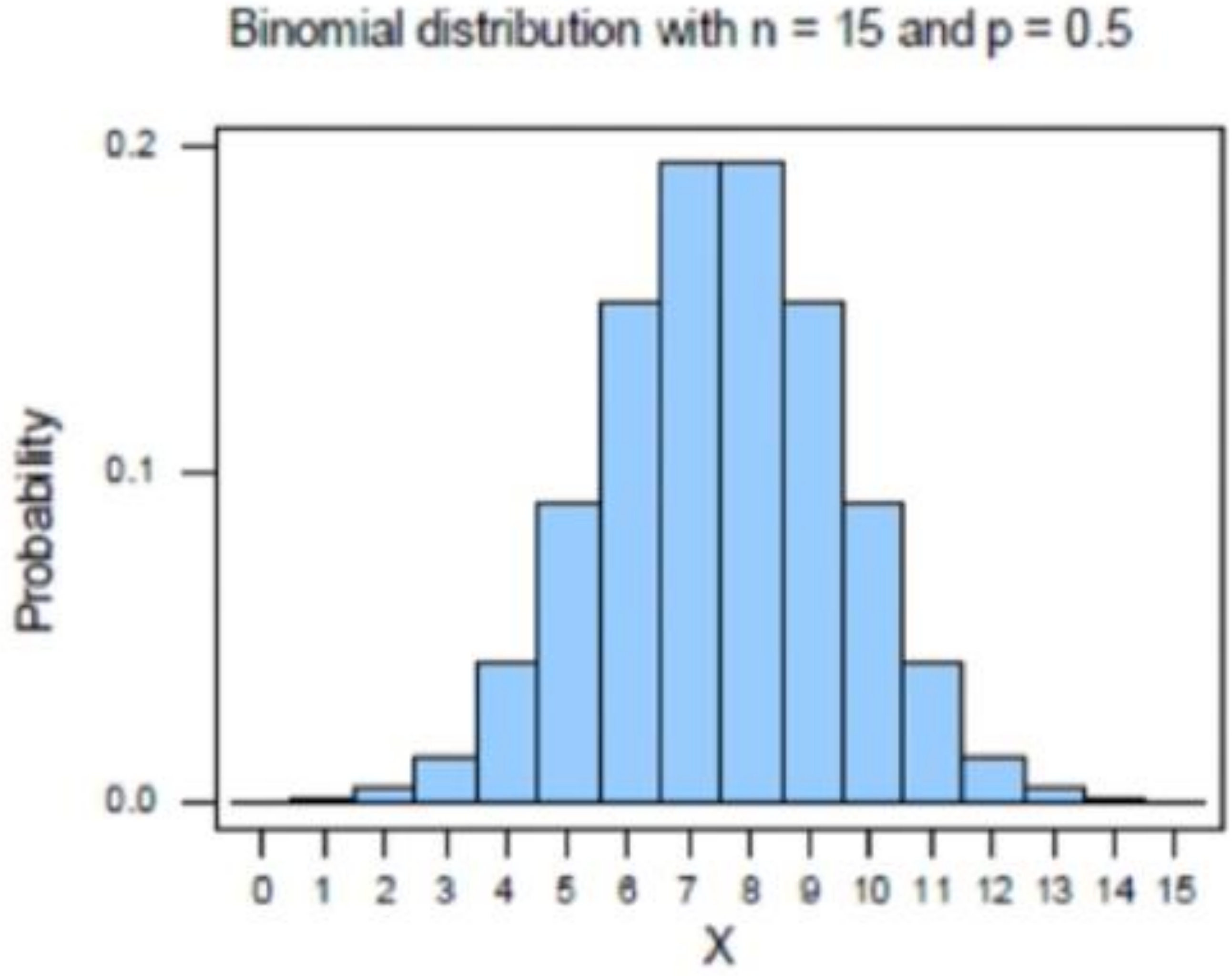
2) Maximization (M-step)

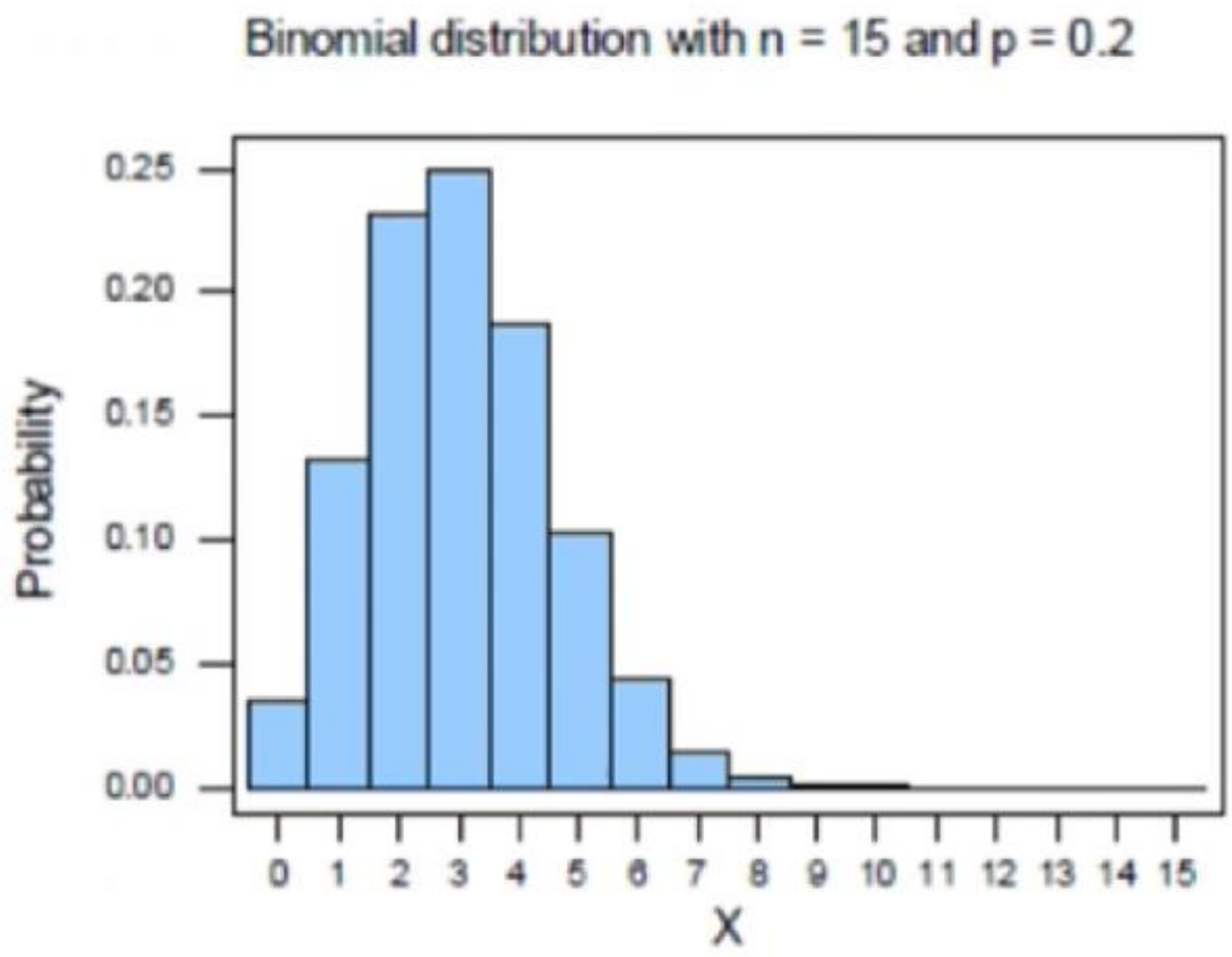
$$\theta_j = \frac{\sum_{i=1}^{m} E[z_{i,j}] x_i}{\sum_{i=1}^{m} E[z_{i,j}]}$$

If $z_{i,j}$ is known:

$$\theta_j = \frac{\sum_{i=1}^{m_j} x_i}{m_j}$$

How do Coin Tosses Behave




Binomial distribution with n = 15 and p = 0.2
Probability
0.25
0.20
0.15
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
X

EM Algorithm: Example

The 5 rounds of 10 coin tosses with $\theta_A = 0.6;\ \theta_B =$ 0.5

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

Let's take the first round: $\frac{5}{10}$ heads and $\frac{5}{10}$ tails.

compute the likelihood that it was coin "A" and coin "B" using the binomial distribution with mean probability θ on n trials with k successes. $p(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

---

[5] θ_i is the average number of heads for coin i. Initially it is randomly assigned

Example with Hidden Variable

$$L(C)=\Theta^{k}(1-\Theta)^{n-k}$$

Likelihood For first coin Flips

$$L(A)=0.6^{5}(1-0.6)^{10-5}=0.0007963$$

$$L(B)=0.5^{5}(1-0.5)^{10-5}=0.0009766$$

$$P(A)=L(A)/L(A)+L(B)=0.0007963/(0.0007963+0.0009766)=0.45$$

$$P(B)=L(B)/L(A)+L(B)=0.0009766/(0.0007963+0.0009766)=0.55$$

EM Algorithm: M-Step

So, We have:

$\theta_A = 0.6;\ \theta_B = 0.5$

1	H	T	T	T	H	H	T	H	T	H
2	H	H	H	H	H	T	H	H	H	H
3	H	T	H	H	H	H	H	T	H	H
4	H	T	H	T	T	T	H	H	T	T
5	T	H	H	H	T	H	H	H	T	H

Recap: $P(Coin = A) = 0.45;\ P(Coin = B) = 0.55$

Estimating likely number of heads and tails from:

- "A": $H = 0.45 \times 5\ heads = 2.2\ heads;\ \ T = 0.45 \times 5\ tails = 2.2\ tails$
- "B": $H = 0.55 \times 5\ heads = 2.8\ heads;\ \ T = 0.55 \times 5\ tails = 2.8\ tails$

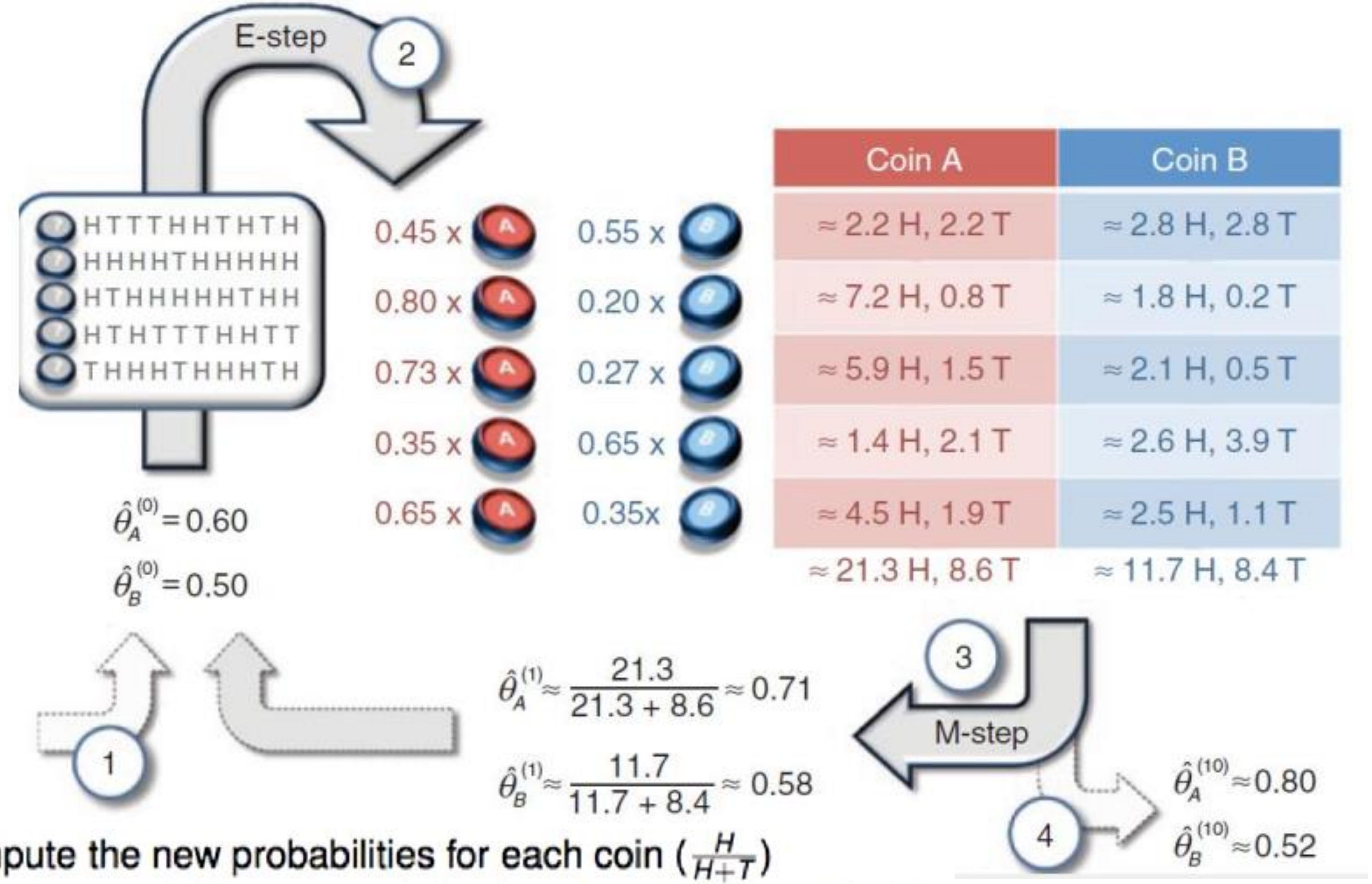
Example with Hidden Variable

In similar fashion find probability of all coins with all flips. It will be as follows:

L(H): Likely no of heads   L(T): Likely no of tails

											Iteration 1->:		Coin A		Coin B	
											P(A)	P(B)	L(H)	L(T)	L(H)	L(T)
B	H	T	T	T	H	H	T	H	T	H	0.45	0.55	2.2	2.2	2.8	2.8
A	H	H	H	H	T	H	H	H	H	H	0.80	0.20	7.2	0.8	1.8	0.2
A	H	T	H	H	H	H	H	T	H	H	0.73	0.27	5.9	1.5	2.1	0.5
B	H	T	H	T	T	T	H	H	T	T	0.35	0.65	1.4	2.1	2.6	3.9
A	T	H	H	H	T	H	H	H	T	H	0.65	0.35	4.5	1.9	2.5	1.1

Example with Hidden Variable



Compute the new probabilities for each coin ($\frac{H}{H+T}$)

That gives you the new maximized parameter θ for each coin

Expectation Maximization

1. Choose starting parameters
2. Estimate probability using these parameters that each data set (x_i) came from j^{th} coin ($E[z_{i,j}]$)
3. Use these probability values ($E[z_{i,j}]$) as weights on each data point when computing a new θ_j to describe each distribution
4. Summate these expected values, use maximum likelihood estimation to derive new parameter values to repeat process