Expectation-Maximization (EM) Algorithm with example

Mehul Gupta · Follow

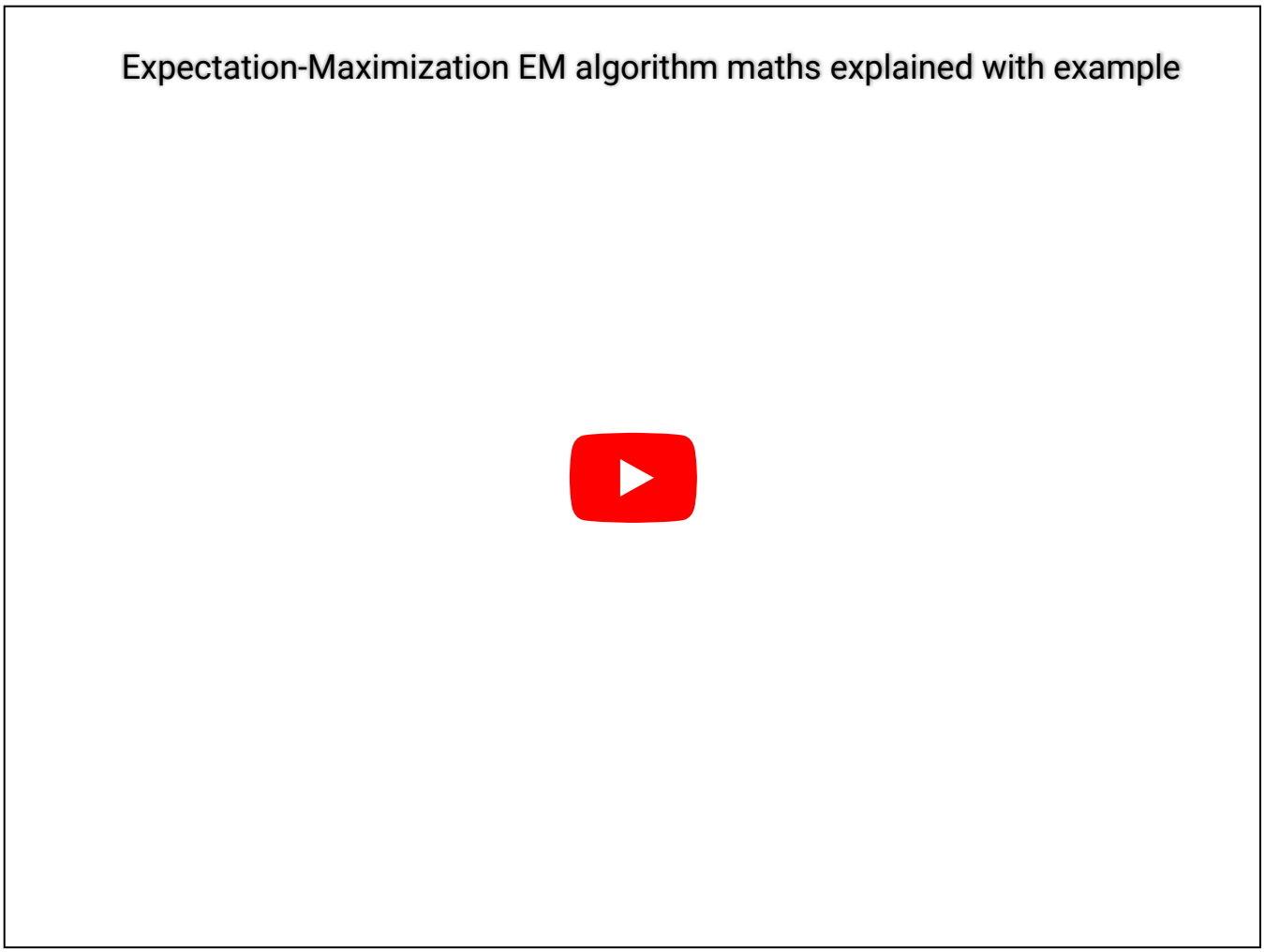
Published in Data Science in your pocket

6 min read · Apr 27, 2020

Real-life Data Science problems are way far away from what we see in Kaggle competitions or in various online hackathons. Before being a professional, what I used to think of Data Science is that I would be given some data initially. Then I need to clean it up a bit (some regular steps), engineer some features, pick up several models from Sklearn or Keras & train.

> Tadaaaa, I am done.

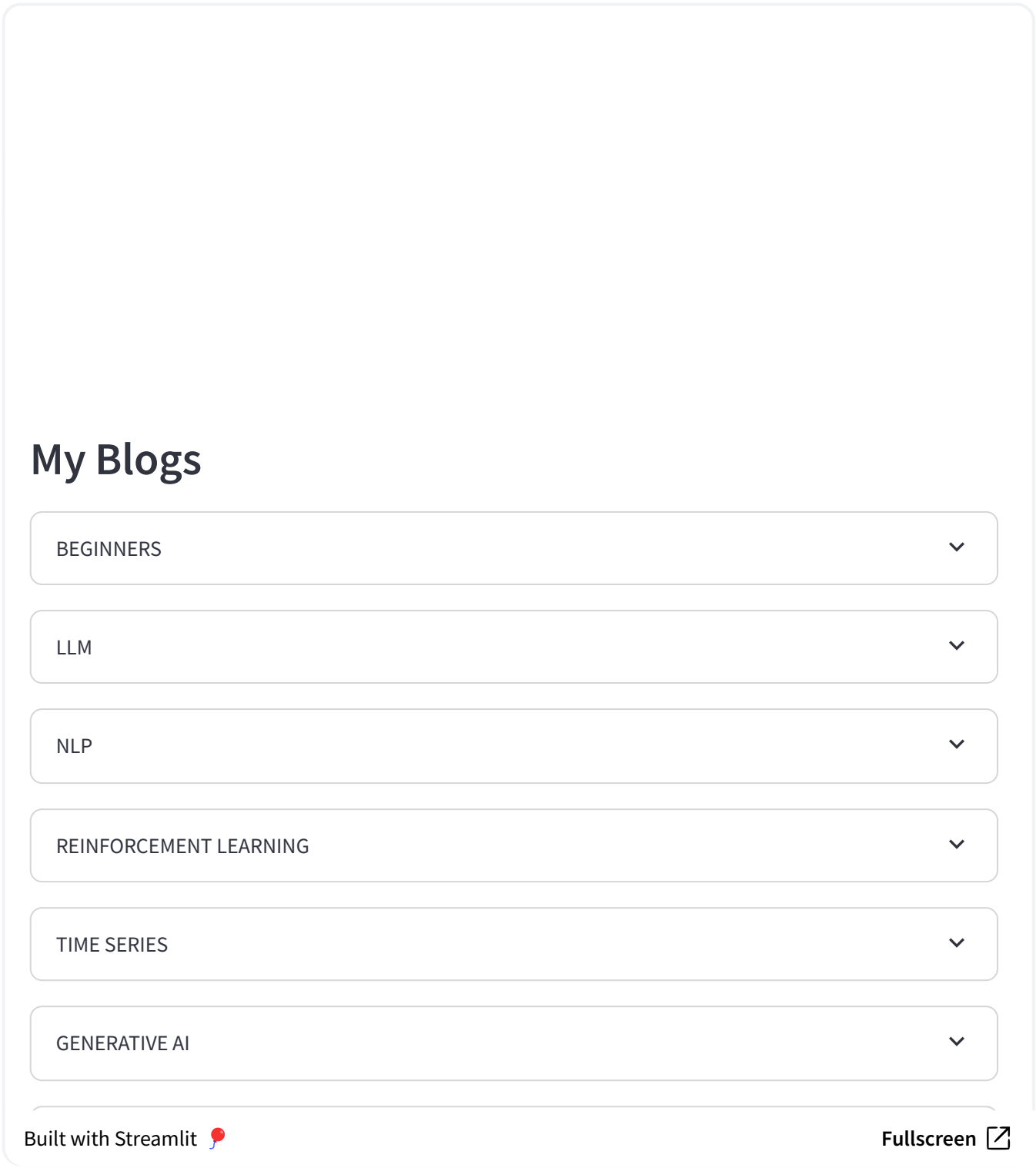
But things aren't that easy. To get perfect data, that initial step, is where it is decided whether your model will be giving good results or not. Most of the time, **there exist some features that are observable for some cases, not available for others (which we take NaN very easily)**. And if we can determine these missing features, our predictions would be way better rather than substituting them with NaNs or mean or some other means.



Here comes the Expectation-Maximization algorithm.

I myself heard it a few days back when I was going through some papers on Tokenization algos in NLP.

The EM algorithm helps us to infer(conclude) those hidden variables using the ones that are observable in the dataset Hence making our predictions even better.



Didn't get it?

Examples always help

Consider 2 biased coins.

Suppose bias for 1st coin is 'Θ_A' & for 2nd is 'Θ_B' where Θ_A & Θ_B lies between 0<x<1. By bias 'Θ_A' & 'Θ_B', I mean that the probability of Heads with 1st coin isn't 0.5 (for unbiased coin) but 'Θ_A' & similarly for 2nd coin, this probability is 'Θ_B'.

Can we estimate 'Θ_A' & 'Θ_B' if we are given some trials(flip results) of these coins?

Simple!!

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

> *Consider Blue rows as 2nd coin trials & Red rows as 1st coin trials.*

What I can do is count the number of Heads for the total number of samples for the coin & simply calculate an average. This can give us the value for 'Θ_A' & 'Θ_B' pretty easily. We can simply average the number of heads for the total number of flips done for a particular coin as shown below

> *Θ_A = 24/30=0.8*
>
> *Θ_B=9/20=0.45*

But what if I give you the below condition:

Here, we can't differentiate between the samples that which row belongs to which coin.

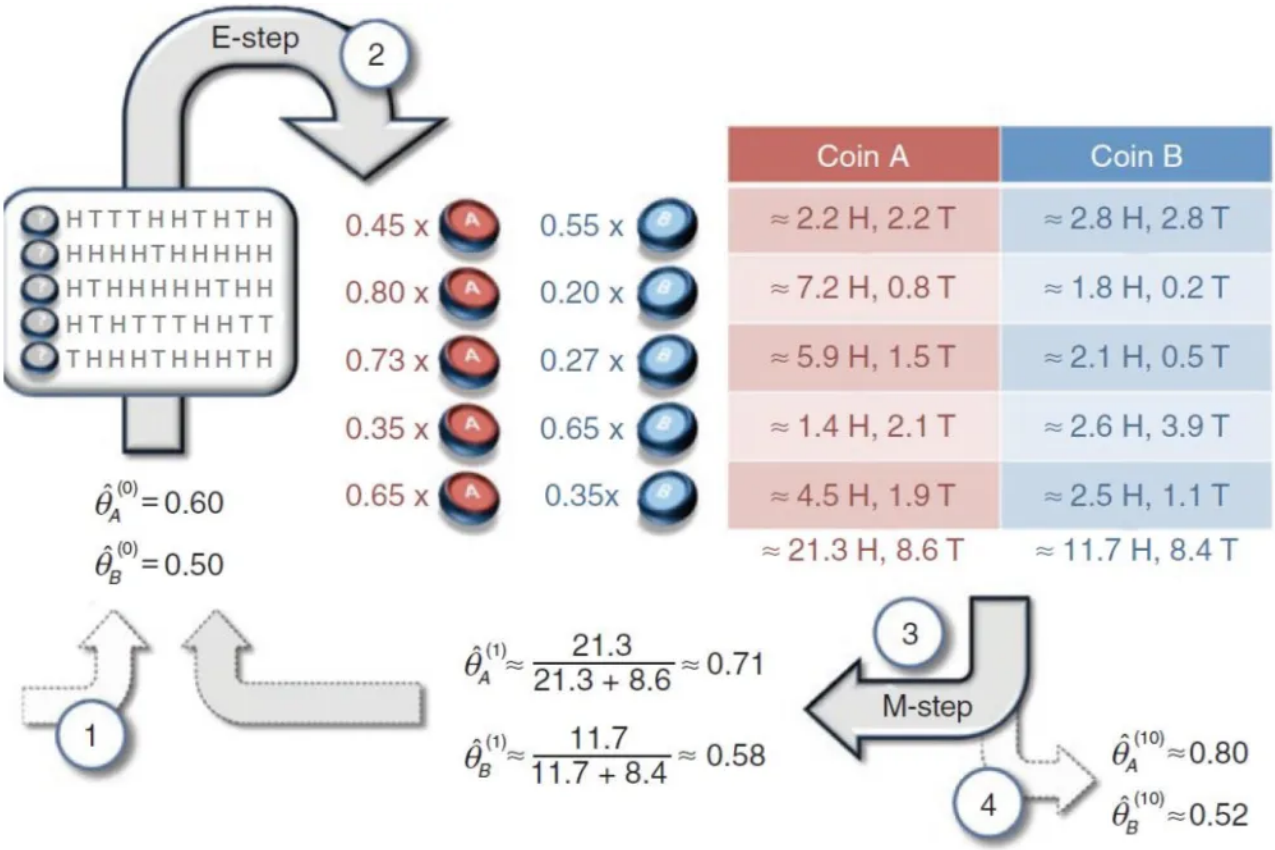
How can I calculate 'Θ_A' & 'Θ_B' then?

We can still have an estimate of 'Θ_A' & 'Θ_B' using the EM algorithm!!

The algorithm follows 2 steps iteratively: **Expectation & Maximization**

- **Expect:** Estimate the expected value for the hidden variable
- **Maximize:** Optimize parameters using Maximum likelihood

Observe the below image:



EM Algorithm Steps:

1. **Assume some random values for your hidden variables:**

Θ_A = 0.6 & Θ_B = 0.5 in our example

By the way, Do you remember the **binomial distribution** somewhere in your school life?

If not, let's have a recapitulation for that as well.

The binomial distribution is used to model the **probability** of a system with only 2 possible outcomes(binary) where we perform 'K' number of trials & wish to know the probability for a certain combination of success & failure using the formula

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

Where

- n= total number of trials

- X= Total number of successful events
- p= Probability of a successful event
- q=Probability of an unsuccessful event

> *For refreshing your concepts on Binomial Distribution, check here.*

Now, if you have a good memory, you might remember why we multiply the Combination (n!/(n-X)! * X!) constant?

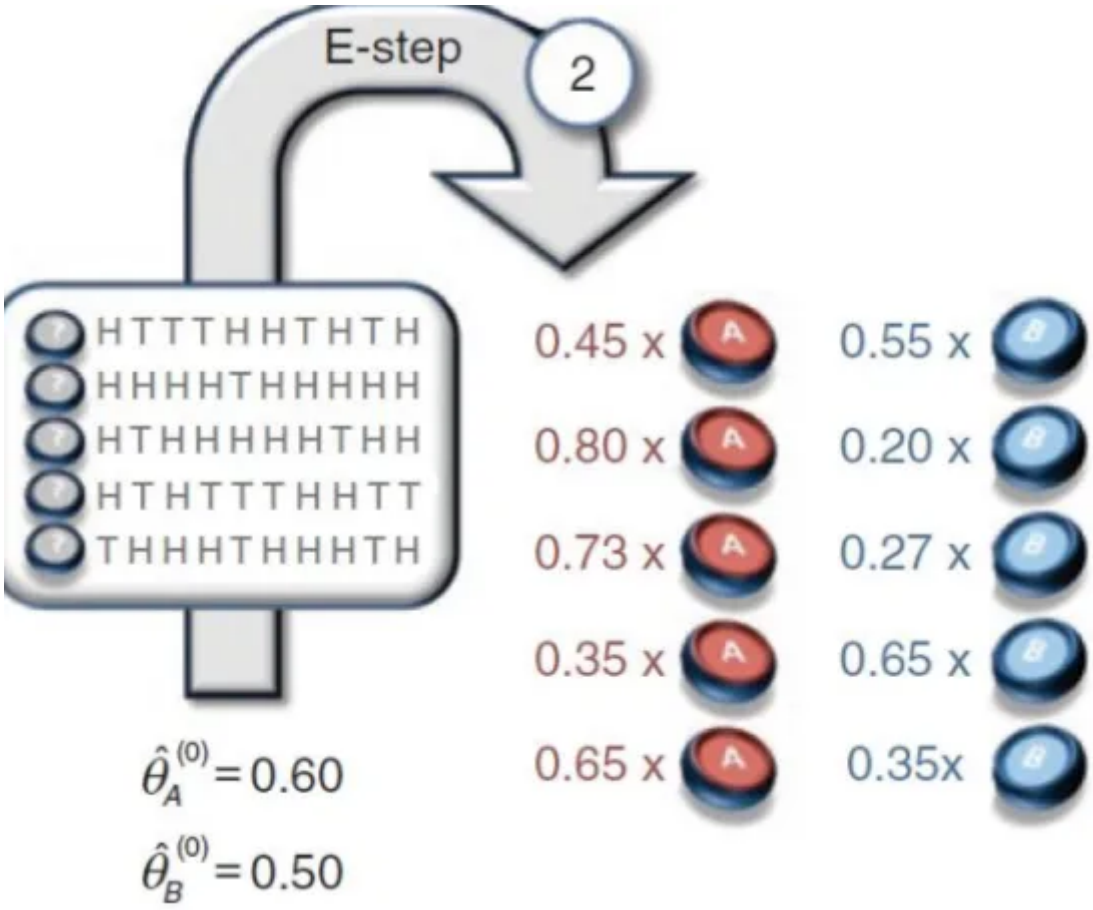
This term is taken when we aren't aware of the sequence of events taking place. Like,

Suppose I say I had 10 tosses out of which 5 were heads & rest tails. Here, we will be multiplying that constant as we aren't aware of in which sequence this happened(HHHHHTTTTT or HTHTHTHTHT or some other sequence, there exist a number of sequences in which this could have happened). But if I am given the sequence of events, we can drop this constant value.

> ***Remember this!!***

Coming back to the EM algorithm, what we have done so far is assumed two values for 'Θ_A' & 'Θ_B'

2. Expectation Step:



> *It must be assumed that any experiment/trial (experiment: each row with a sequence of Heads & Tails in the grey box in the image) has been performed using only a specific coin (whether 1st or 2nd but not both). The grey box contains 5 experiments*

Look at the first experiment with 5 Heads & 5 Tails (1st row, grey block)

Now using the binomial distribution, we will try to estimate what is the probability of 1st experiment carried on with 1st coin that has a bias 'Θ_A'(where Θ_A=0.6 in the 1st step).

As we already know the sequence of events, I will be dropping the constant part of the equation. Hence Probability of such results, if the 1st experiment belonged to 1st coin, is

$(0.6)^5 x (0.4)^5 = 0.00079$ (As p(Success i.e Head)=0.6, p(Failure i.e Tails)=0.4)

Similarly, If the 1st experiment belonged to 2nd coin with Bias 'Θ_B'(where Θ_B=0.5 for the 1st step), the probability for such results will be:

$0.5^5 x 0.5^5 = 0.0009$ (As p(Success)=0.5; p(Failure)=0.5)

On normalizing these 2 probabilities, we get

- P(1st coin used for 1st experiment)=0.45
- P(2nd coin used for 1st experiment)=0.55

Similarly, for the 2nd experiment, we have 9 Heads & 1 Tail.

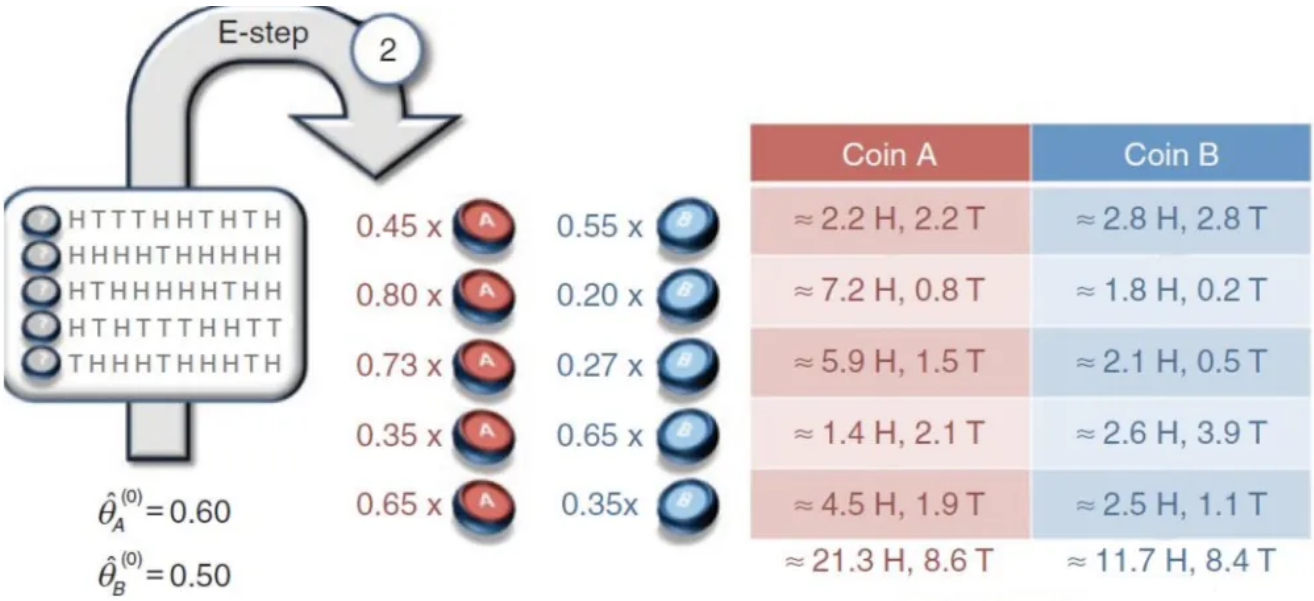
Hence,

- P(1st coin used for 2nd experiment) = 0.6^9x0.4^1=0.004
- P(2nd coin used for 2nd experiment) = 0.5^9x0.5 = 0.0009

On Normalizing, the values we get are approximately 0.8 & 0.2 respectively

> *Do check the same calculation for other experiments as well*

Moving to 2nd step:



Now, we will be multiplying the Probability of the experiment belonging to the specific coin(calculated above) by the number of Heads & Tails in the experiment i.e

> *0.45 * 5 Heads, 0.45* 5 Tails= 2.2 Heads, 2.2 Tails for 1st Coin (Bias 'Θ_A')*

Similarly,

> *0.55 * 5 Heads, 0.55* 5 Tails = 2.8 Heads, 2.8 Tails for 2nd coin*

We can calculate other values as well to fill up the table on the right.

Now once we are done, Calculate the total number of Heads & Tails for respective coins.

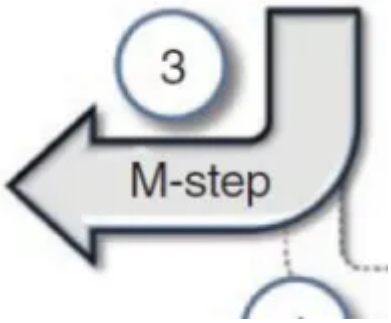
> *For 1st coin, we have 21.3 H & 8.6 T;*
>
> *For 2nd coin, 11.7 H & 8.4 T*

Maximization step:

Now, what we want to do is to converge to the correct values of 'Θ_A' & 'Θ_B'.

$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3 + 8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7 + 8.4} \approx 0.58$$



As the bias represented the probability of a Head, we will calculate the revised bias:

> *'Θ_A'= Heads due to 1st coin/ All Heads observed= 21.3/21.3+8.6=0.71*

Similarly,

Θ_B = 0.58 shown in the above equation. Now we will again switch back to the Expectation step using the revised biases.

On 10 such iterations, we will get Θ_A=0.8 & Θ_B=0.52

Have you observed one thing!!

These values are quite close to the values we calculated when we knew the identity of coins used for each experiment was Θ_A=0.8 & Θ_B=0.45 (taking the average at the very beginning of the post)

Hence, the algorithm works!!!