CS378 - Autonomous Vehicles in Traffic II

Week 3a - Probability
(Based on slides by Andrew Moore)

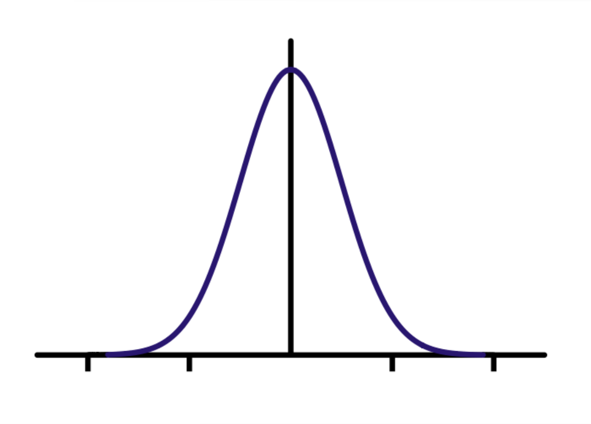
Real-Valued Random Variable

- Boolean
  - A can be {true, false}
  - A: It will rain tomorrow
- Discrete
  - A can take a value from a given set
  - A: number of years it will take for me to graduate
- Continuous
  - A takes all real values
  - A: my distance to the wall

Probability

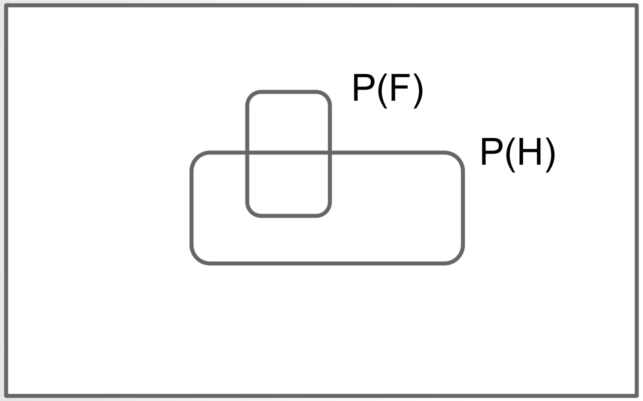
- The probability P(A = x) is the fraction of "worlds" in which A will turn out to be x.
- For boolean and discrete random variables, we define explicit probability values
- For continuous random variable, we define a probability density function (pdf)

For instance, the pdf of me being a certain distance from the wall could be a gaussian with a mean of 5 meters



Conditional Probability

- P(A = x|B = y) - The fraction of worlds (where B is y) in which A is x

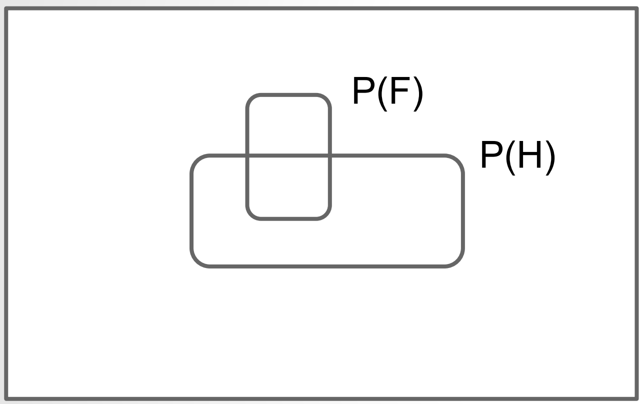


P(F): probability of waking up with the flu = 1/40
P(H): probability of waking up with a headache = 1/10
P(H|F) = 1/2

- If A and B are independent boolean random variables, what is the conditional probability P(A|B)?

Inference

- What is the probability of having the flu if you wake up with a headache?



P(F): probability of waking up with the flu = 1/40
P(H): probability of waking up with a headache = 1/10
P(H|F) = 1/2

- We need P(F|H) = P(F and H) / P(H)
  = (1/40 * 1/2) / 1/10
  = 1/8

But wait...

- What we did is an example of Bayes' rule
- P(F|H) = P(F and H) / P(H)
- i.e. P(F|H) = P(H|F) * P(F) / P(H)

CS378 - Autonomous Vehicles in Traffic II

Week 3a - Expectation Maximization

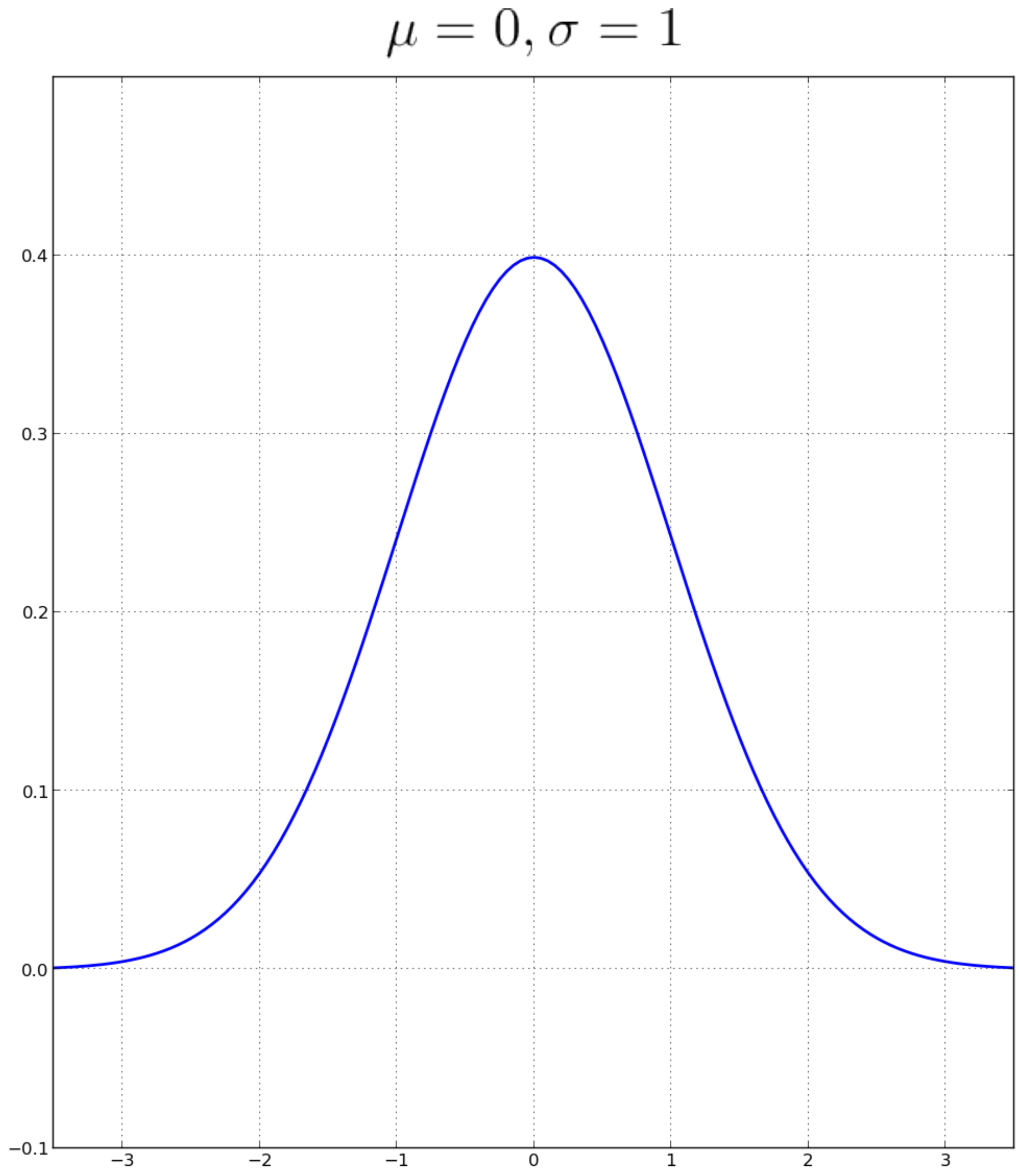
Probability Density Function

- A probability density function gives an estimate of the distribution of output values *given* the input parameters.
- In the case of a normal distribution (i.e. gaussian), the pdf looks something like:

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- We can calculate the probability by taking the area under the curve:

$$P(X = x; \mu, \sigma) = f_{\mu,\sigma^2}(x)\Delta x$$



$\mu = 0, \sigma = 1$
0.4
0.3
0.2
0.1
0.0
−0.1
−3
−2
−1
0
1
2
3

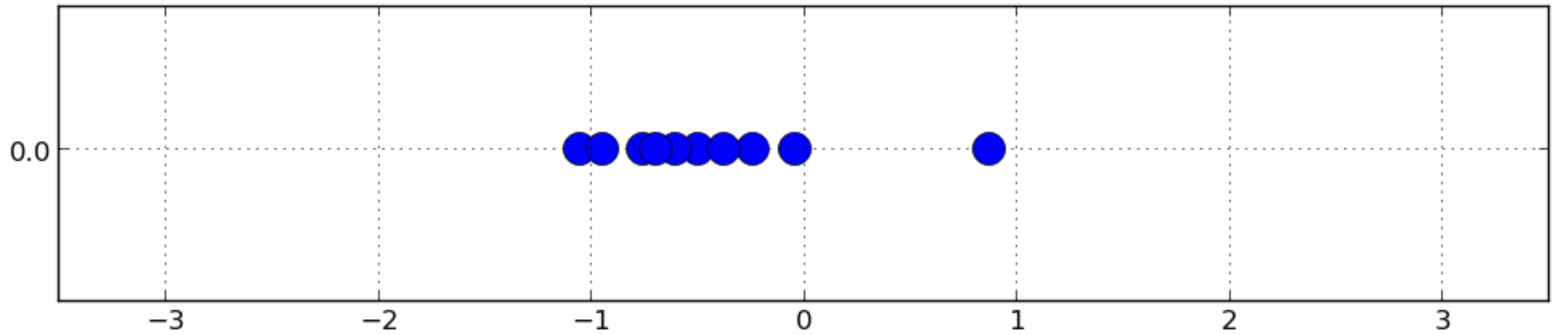
Samples

- Now, the *pdf* here defines how likely a given observation *x* is.
- Using this pdf, you can draw a number of samples from this distribution

- Aside: to get samples from an arbitrary pdf, use the cummulative pdf trick.

What do 10 samples look like?

[-1.05, -0.94, -0.49, -0.60, -0.24, -0.04, -0.37, -0.76, -0.69, 0.87]



Likelihood

- Now let's take the reverse scenario. I give you a distribution, and tell you that it is from a gaussian. What can you say about the input parameters that generated this data?

- Likelihood is defined as the probability some set of input parameters generated the given output:

$$P(\mu, \sigma | X) \quad \text{or} \quad P(\theta | X)$$

Likelihood

- We can define the likelihood of the *same pdf* by changing the arguments of the pdf:

$$P(\mu, \sigma | x) \propto f_x(\mu, \sigma^2)$$

Maximum Likelihood Estimation

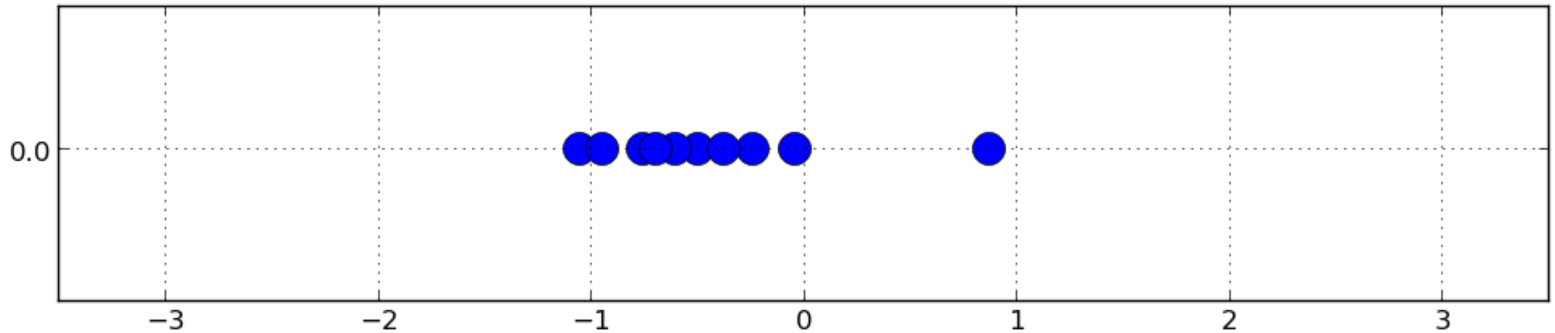
- Maximum Likelihood Estimation is the process by which we can determine the parameters that *most likely* explain the data.
- So what we are trying to do is find the theta which produces the maximum $P(\theta|X)$
- Since we just inferred that: $P(\theta|x) \propto P(x|\theta)$
- This means that MLE boils down to:

$$argmax_{\theta}(P(x|\theta))$$

$$argmax_{\theta}(logP(x|\theta))$$

Let's take an example of MLE

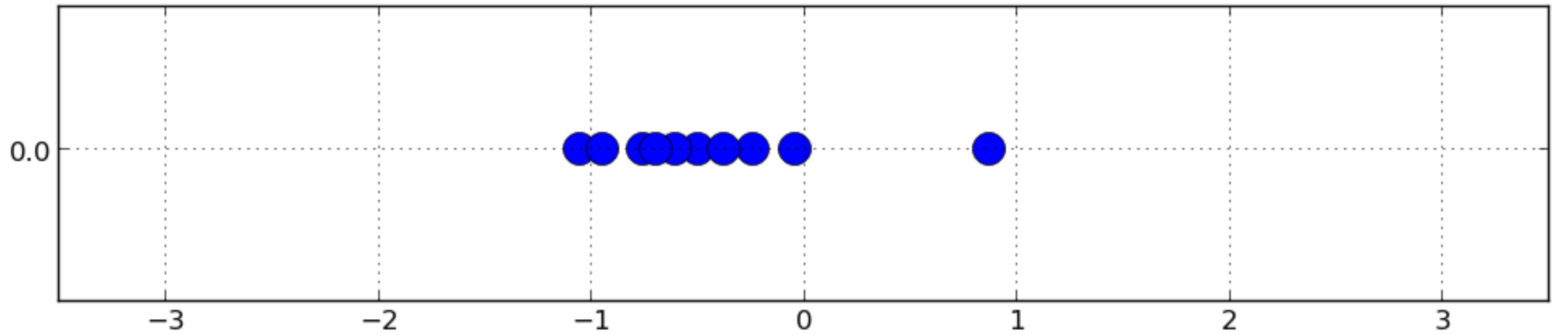
[-1.05, -0.94, -0.49, -0.60, -0.24, -0.04, -0.37, -0.76, -0.69, 0.87]



- What is the maximum likelihood of this distribution?

A closed form solution perhaps?

[-1.05, -0.94, -0.49, -0.60, -0.24, -0.04, -0.37, -0.76, -0.69, 0.87]



$$\hat{\mu} = \frac{1}{n}\sum_{1}^{n} x_i = -0.436$$

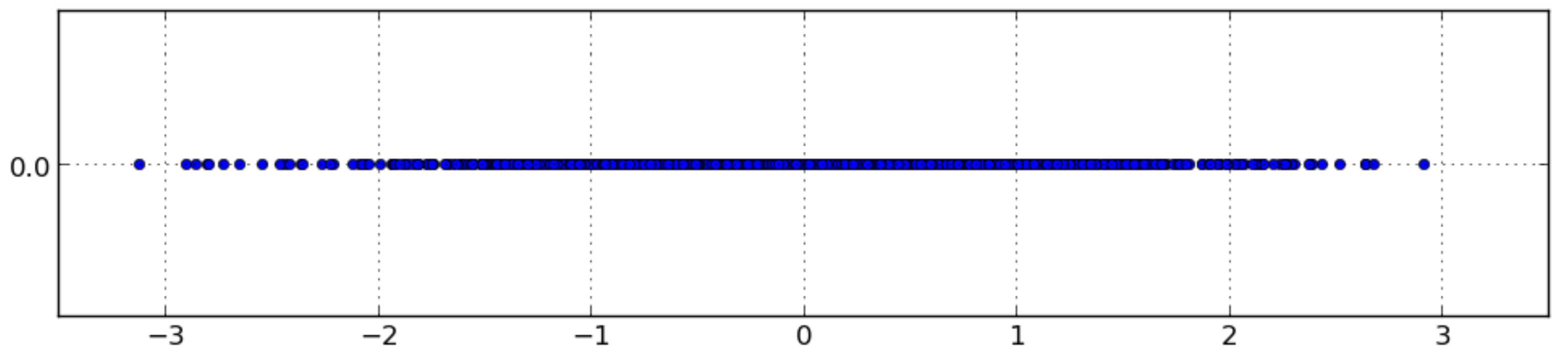
$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{1}^{n}(x_i - \hat{\mu})^2} = 0.525$$

How did we do?

Why did we not do well?

- Unfortunately 10 samples can sometimes be insufficient to capture the distribution!
- Maximum likelihood estimation just gave us the *most likely* answer that explained this data.
- What would have happened if we had more data points from the true distribution?

With a 1000 samples



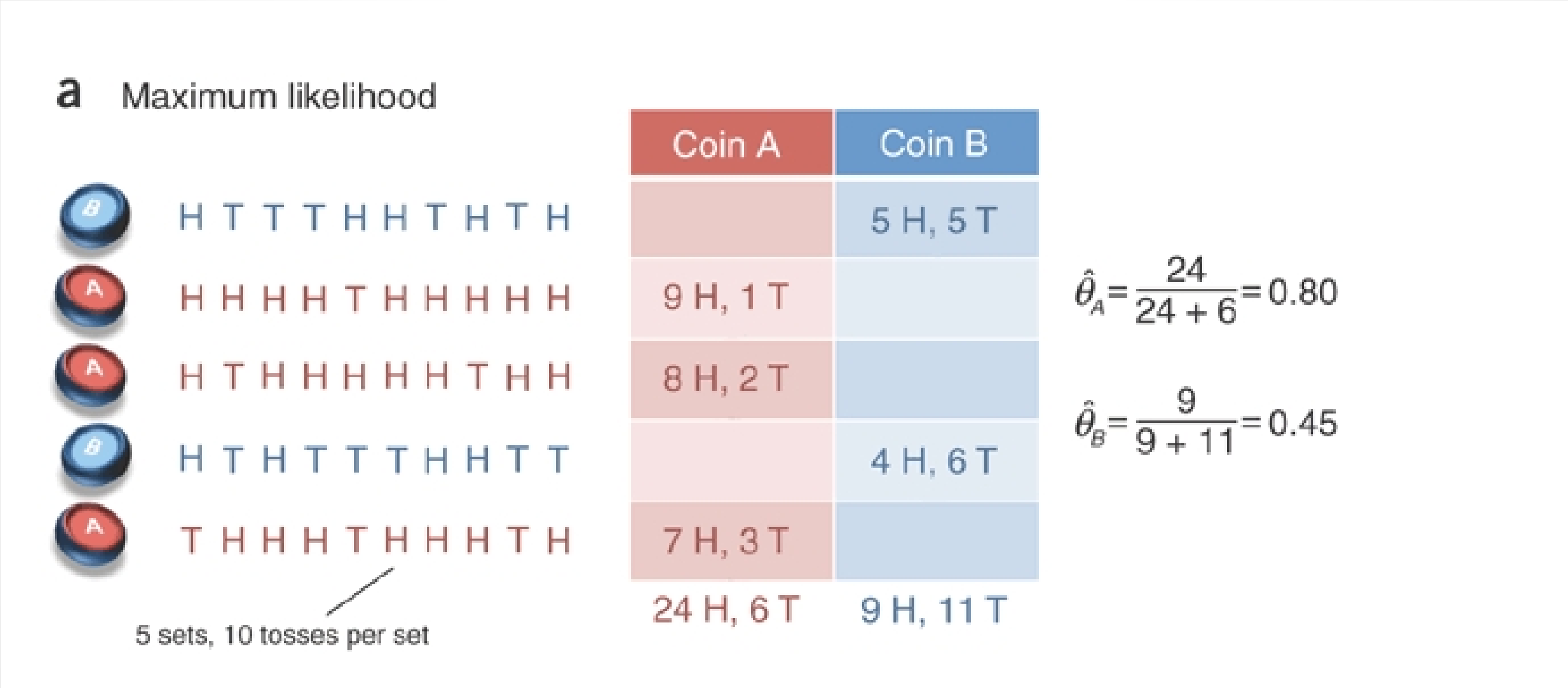
- Mean: 0.044
- Standard Deviation: 1.003

MLE Summary

- Likelihood explains some a set of given data using different input parameters
- Likelihood values only mean something when compared against other such values
- Maximum likelihood estimation is producing parameters that *most likely* produced the data.
- Depending on the domain, we can sometimes do closed form analysis to obtain the MLE parameters.

And on to the tutorial...

- When all data is given, we can do MLE to obtain parameters.



Expectation Maximization

- When some of the data is hidden, it is no longer possible to calculate the MLE parameters directly
- EM is a maximum likelihood estimation technique when there is hidden data.
- These hidden variables are called latent variables.
- In the paper:
  - What data is hidden?
  - Why can't we do parameter estimation without this data?

How does EM work?

- Assume arbitrary values for the input parameters.
- Compute *soft* assignments for latent variables
- Calculate parameters using MLE now that you have all the observation data.
- Repeat till parameters no longer change.

How to compute assignments?

- For any given point of data (HTTTHHTHTH), we need to find the missing information (i.e. which coin did this data point come from?)

$$\begin{aligned} x_i &: HTTTHHTHTH \\ z_a &: x_i \in coin_a \\ z_b &: x_i \in coin_b \end{aligned}$$

- Essentially we need to compute the probabilities of this data point belonging to each coin

$$p(z_a|x_i) \qquad p(z_b|x_i)$$

How to compute assignments?

- We'll use Bayes' rule!!

$$
\begin{aligned}
P(z_a|x_i) &= \frac{P(x_i|z_a)P(z_a)}{P(x_i)} \\
P(z_b|x_i) &= \frac{P(x_i|z_b)P(z_b)}{P(x_i)} \\
P(z_a|x_i) + P(z_b|x_i) &= 1
\end{aligned}
$$

- Now since the coins were selected randomly

$$\implies P(z_a) = P(z_b) = 0.5$$

- This gives us:

$$P(z_a|x_i) = \frac{P(x_i|z_a)}{P(x_i|z_a) + P(x_i|z_b)}$$

$$P(z_b|x_i) = \frac{P(x_i|z_b)}{P(x_i|z_a) + P(x_i|z_b)}$$

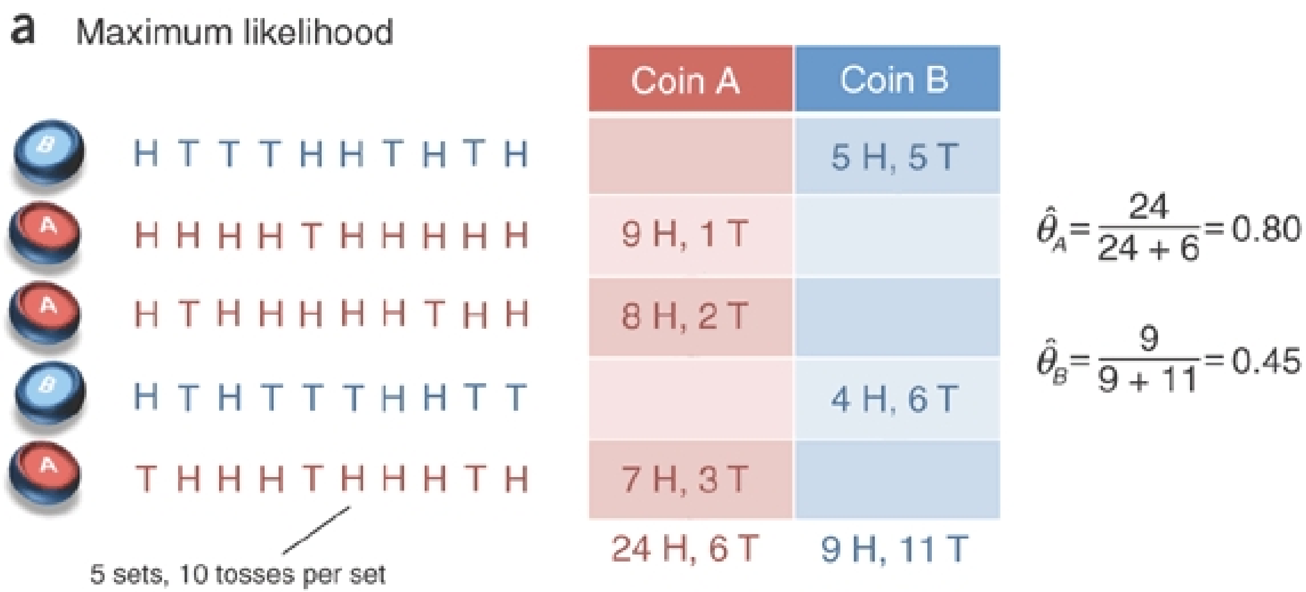
How to compute $P(x_i|z_a)$

- $P(x_i|z_a)$ is the probability of seeing a particular set of coin tosses given a particular coin
- If you roll an unbiased coin 10 times, are you more likely to get 10 heads in a row, or 5 heads in a row and then 5 tails in a row?
- If you roll an unbiased coin 10 times, are you more likely to see a total of 10 heads? or 5 heads and 5 tails?

How to compute P(z|a) and P(z|b)

- Let's take a look at the link I emailed you
- http://math.stackexchange.com/questions/25111/how-does-expectation-maximization-work

Hard Assignments



Soft Assignments

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

Hard vs Soft Assignments

- Hard assignments mean a "greedy" strategy
- Soft assignments are more forgiving. Take the first assignment here:

- The probability of this belonging to either coin is fairly close (i.e 0.45 and 0.55). It might not make sense to assign this to a single coin alone
- In practice both may work decently well.

So how did we do?

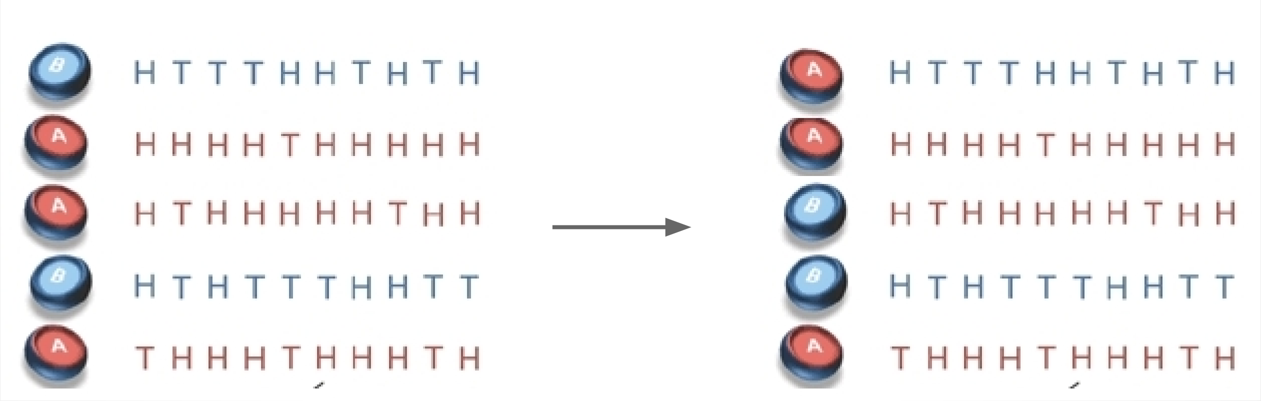
- The true values (without missing data) produced a result of 0.8 and 0.45
- Using EM, we converged at 0.8 and 0.45.
- Not bad!

Food for thought #1

- Would the output be different if we started with a different initial guess?

Food for thought #2

- What if the true assignments were like this? What would $\hat{\theta}_a$ and $\hat{\theta}_b$ be



- How would the output of EM change? Would you say we still did well?

Food for thought #3

- Let's say if we repeated the experiment a 1000 times instead of 5 times. Would the output of EM always be closer to the true values or not?
    - Think about this one (or even better, code it up!). I'll discuss it on Wednesday

k-means

- k-means is a data clustering algorithm that is (almost) an example of EM. soft k-means is an example of EM
- I give you a data set of 1000 points in x,y space, and tell you to give me 5 clusters centers. How would you go about this problem?
- k-means approximates probabilities with distances

k-means

- Choose arbitrary cluster centers
- Assign each point to the closest cluster center (E step)
- Now that you have individual clusters, calculated the mean of each cluster (M step)
- Repeat this process until the cluster centers no longer change

soft k-means

- Choose initial cluster centers randomly
- Instead of hard assigning a point to a cluster center, *soft assign* it to all the cluster centers (E step)
- Use a *weighted mean* for each cluster to calculate the cluster center (M step)
- Repeat until convergence