This slides considered a scenario where a coin $Z$ is tossed to choose between one of two other coins $A$ and $B$ and then the chosen coin is then tossed $N$ times. The whole procedure is repeated $D$ times.

The hidden variable variant is where as data you just know on each trial what the $N$ coin tosses yielded with the chosen coin, but you don’t know which coin was being tossed: the outcome on $Z$ is hidden.

In this lab you can do some experiments with a program implementing the EM algorithm to do parameter estimation in this scenario.

1 Background Info

Location To use the program you should be logged on to one of the Linux machines. The code is located in

```
/shared/teaching/CSLL/4CSLL5_18_19/CoinEMLab
```

and you should start a terminal/shell program

Execution As example execution would be

```
./coin_em_tester -file coin_data_notes_order \n   -sup no -unsup yes \n   -max_it 10 \n   -coin_chce_probs_string 0.5/0.5 \n   -ht_probs_string 0.4/0.6/0.3/0.7
```

you could write all of the above on one line, or use the \ to allow it be written on several lines. You may well find it handy to edit the instruction with an editor of some kind and then paste it into the shell.

```
processing coin_data_notes_order
read all data
Total amount of extracted data is: 9
CHOICE: A
TOSSES: HHHHHHHHHTT H:8 T:2
CHOICE: B
TOSSES: THTTTTHTTT H:2 T:8
CHOICE: A
TOSSES: HTHTHTHTHT H:7 T:3
CHOICE: A
TOSSES: HTHHHTHHHH H:8 T:2
CHOICE: B
TOSSES: TTTTTTTTTTT H:1 T:9
CHOICE: A
```
TOSSES: HHTHHHHHH H:9 T:1
CHOICE: A
TOSSES: THHTHHHHHT H:7 T:3
CHOICE: A
TOSSES: HHHHHHTHHH H:9 T:1
CHOICE: B
TOSSES: HTTTTTHTT H:3 T:7

******doing rounds of unsup training ******

| p(A)  | p(H|A) | p(H|B) | logprob | prob  |
|-------|-------|-------|---------|-------|
| 0.5   | 0.4   | 0.3   | -101    | 3.86e-31 |
| 0.699 | 0.707 | 0.352 | -77.4   | 5.19e-24 |
| 0.667 | 0.793 | 0.213 | -73.3   | 8.9e-23  |
| 0.667 | 0.799 | 0.201 | -73.2   | 9.09e-23 |
| 0.667 | 0.799 | 0.2   | -73.2   | 9.09e-23 |
| 0.667 | 0.799 | 0.2   | -73.2   | 9.09e-23 |
| 0.667 | 0.799 | 0.2   | -73.2   | 9.09e-23 |
| 0.667 | 0.799 | 0.2   | -73.2   | 9.09e-23 |
| 0.667 | 0.799 | 0.2   | -73.2   | 9.09e-23 |

output format: the output shows the parameter values on 10 successive iterations.

The first column shows the probs for Z=A – what was referred to as $\theta_a$ in the slides.

The second column shows the probs for a coin toss being H if the tossed coin is A – what was referred to as $\theta_{h|a}$ in the slides.

The third column shows H given B – $\theta_{h|b}$ in the slides.

The fourth and fifth columns show the log probability and probability of the entire data set.

program parameters

-file FileName

FileName contains the data. The directory contains two example data files. Note that although the data file contains the values of the hidden variable, the estimation program does not actually use it.

-coin_chce_probs_string $A/B$

$A$ is some number representing the probability $P(Z = A)$ and this is the initial value of that parameter, which EM then alters. $B$ is $P(Z = B)$.

-ht_probs_string $HA/TA/HB/TB$

4 slash separated numbers. The first two are $P(H|A)$ and $P(T|A)$, the second two are $P(H|B)$ and $P(T|B)$.

-max_it $N$

$N$ is the number of iterations EM to run.

-sup no -unsup yes

The above combination specifies that it should treat Z as hidden and do unsupervised estimation. The other possibility is -sup yes -unsup no.

2 Task 1

Here's a table of possible start values for the parameters:
start trying them out and see what solution results in each case
one of the things which you should find is that not all the solutions are the same. You should also find that there is a relationship among the alternative solutions. Try to figure out why the alternatives must exist.

3 Task 2

Try out the other data set test\_data\_two. Again try out some different start values (the slides give one example, and note also that this is the data set of the assignment)