EE 126 Probability and Random Processes
April 26, 2018 EM Algorithm

1 Lecture

EM Algorithm

The **EM (expectation-maximization) algorithm** arises in problems that involve parameter estimation. Perhaps we have incomplete (or insufficient) observation data and want to estimate parameters that capture the data we see. The EM algorithm can be used for this. It has applications in RNA sequencing, clustering and classification, speaker identification... (the list goes on).

Recall the toy casino problem, which had a "fair" die and a "loaded" die. Last time, we used the Viterbi algorithm to find the best sequence of states given our observations, but while doing so we made a crucial assumption: that all parameters were known.

- Certainly if we know the F and the L parameters (e.g. $P(6) = 1/2$ for L), then finding the MLSE using the Viterbi algorithm is easy. But what if we don't? Then we need something like the EM algorithm to *estimate* the parameters.

The best way to understand EM is through an example.

Example: coin flipping experiment (easy setting). We have a pair of coins A and B with unknown biases θ_A and θ_B. The goal is to estimate θ_A and θ_B. Our experiment will be to

(1) choose one of the two coins (either A or B) at random with equal probability,

(2) perform 10 independent coin flips with that coin, and then

(3) repeat (1) and (2) five times.

In total we perform 50 flips (5 sets of 10 flips each). Note that a sufficient statistic for our estimation is the number of heads in each round of tosses. We'll first assume that we get to observe which coin is used in each round.

Let's say the data is

	Coin A	**Coin B**
		5H, 5T
	9H, 1T	
	8H, 2T	
		4H, 6T
	7H, 3T	
Total	**24H, 6T**	**9H, 11T**

By MLE in this complete data setting, $\hat{\theta}_A = \frac{24}{24+6} = 0.8$ and $\hat{\theta}_B = \frac{9}{9+11} = 0.45$.

- "Complete": we see not only the results of the coin tosses, but also the labels of the coins before we toss them.

As an aside,

$$\begin{aligned}\text{MLE}[\theta_A \mid H_A \text{ heads and } T_A \text{ tails}] = \theta_A^* &= \arg\max_\theta P(H_A \text{ heads}, T_A \text{ tails} \mid \theta) \\ &= \arg\max_\theta \theta^{H_A}(1-\theta)^{T_A} \\ &= \arg\max_\theta f(\theta)\end{aligned}$$

Then, to maximize $\log f(\theta) = H_A \log\theta + T_A \log(1-\theta)$ we can take the derivative and solve:

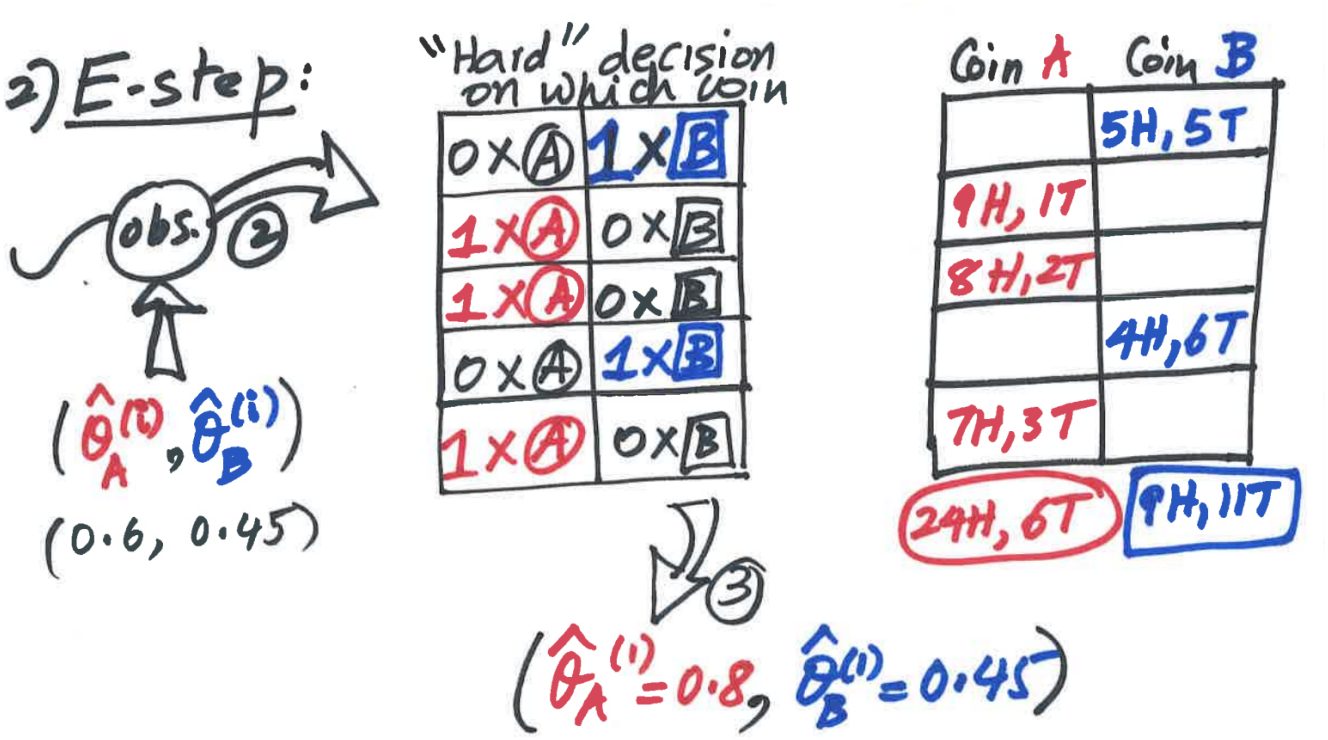
$$\begin{aligned}\frac{d}{d\theta}\log f(\theta) = \frac{H_A}{\theta} - \frac{T_A}{1-\theta} &= 0 \\ \implies \theta^* &= \frac{H_A}{H_A + T_A}\end{aligned}$$

Now comes the interesting part: imagine we don't know the labels of the coins we're flipping. We don't know that, say, coin B produced these 10 observations; we only know the 10 observations. The EM algorithm deals with this setting. It'll do its own clustering of the observations into labels.

Specifically, the EM algorithm can be used to estimate $\hat{\theta}_A$ and $\hat{\theta}_B$ when coin label information is missing. It is *iterative*; with every iteration it improves its estimation quality (although there is no guarantee that it will conclude at the global optimum).

"Hard-Decision" EM (HEM)

1. Start with initial estimate $(\hat{\theta}_A^{(0)}, \hat{\theta}_B^{(0)})$, e.g. $(0.6, 0.45)$.
2. **E-step:** try to allocate/align each observation to a label. We'll perform an expectation in which we put all of the mass into the label that we believe in.



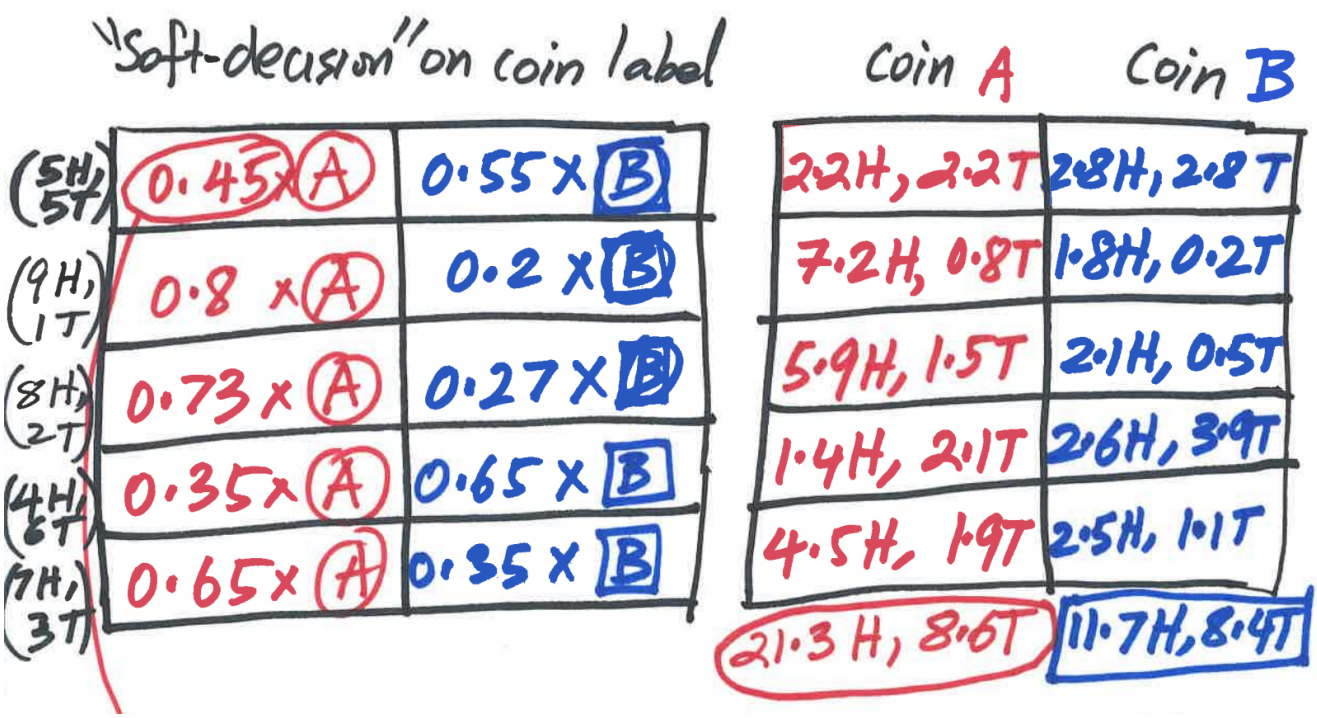
source: notes from Kannan Ramchandran's EE 126 lecture

3. **M-step:** re-estimate $\hat{\theta}_A$ and $\hat{\theta}_B$ from the latest E-step.

4. Start again from the E-step and iterate until convergence.

"Soft-Decision" EM (SEM)

In this case, we assign a probability to each label (the decision is no longer a binary "this label or *this* label," but rather "0.55 chance of being this label, 0.45 chance of being *this* label..."). Then, in tracking statistics we multiply probabilities by the total counts – assigning 55% of the mass here, 45% of the mass here, etc.



source: notes from Kannan Ramchandran's EE 126 lecture

As an example, let's perform the E-step for the first round of flips. Take our initial estimate $\hat{\theta}^{(0)}$ to be $(0.6, 0.45)$.

$$\begin{aligned} P_{\hat{\theta}^{(0)}}(A \mid 5H, 5T) &= \frac{P(5H, 5T \mid A)P(A)}{P(5H, 5T \mid A)P(A) + P(5H, 5T \mid B)P(B)} \quad \text{(Bayes' rule)} \\ &= \frac{0.6^5 \cdot 0.4^5 \cdot \frac{1}{2}}{0.6^5 \cdot 0.4^5 \cdot \frac{1}{2} + 0.45^5 \cdot 0.55^5 \cdot \frac{1}{2}} \\ &= 0.45 \end{aligned}$$

Then, as the M-step, we have

$$\begin{aligned} \text{Coin A:} \quad \hat{\theta}_A^{(1)} &= \frac{21.3}{21.3 + 8.6} \approx 0.71 \\ \text{Coin B:} \quad \hat{\theta}_B^{(1)} &= \frac{11.7}{11.7 + 8.4} \approx 0.58 \end{aligned}$$

If we continue, we will end up with $\hat{\theta}_A^{(10)} = 0.8$ and $\hat{\theta}_B^{(10)} = 0.52$ at convergence.

Recap

In HEM, the E-step boils down to

$$P(A \mid \theta^{(i)}, Y) \underset{B}{\overset{A}{\lessgtr}} P(B \mid \theta^{(i)}, Y)$$

(i.e. if $P(A \mid \theta^{(i)}, Y) > P(B \mid \theta^{(i)}, Y)$, declare A, otherwise declare B).

In SEM, the E-step boils down to Bayes' rule:

$$P(A|\theta^{(i)}, Y) = \frac{P(Y|A; \theta^{(i)})P(A)}{P(Y; \theta^{(i)})}$$

$$P(B|\theta^{(i)}, Y) = \frac{P(Y|B; \theta^{(i)})P(B)}{P(Y; \theta^{(i)})}$$

Note: $P(X; \theta)$ is equivalent to $P(X \mid \theta)$; it's just a difference of notation. θ is a parameter we want to estimate (not a random variable), so the former is the preferred way to write it.

Analysis

In general, we would like to maximize

$$P(\underbrace{\text{observed data}}_{=X} \mid \theta)$$

with respect to θ. We adhere to the model $Y = (X, Z)$, where Y is the "complete" data, X is the "incomplete" data, and Z is the "hidden" data. In the coin-flipping example,

$$\underbrace{\text{complete data}}_{Y} = (\underbrace{\text{coin tosses}}_{X}, \underbrace{\text{coin labels}}_{Z})$$

(15H, 5T)... BABB...

Note that $\arg\max_\theta P(X \mid \theta)$ is hard, while $\arg\max_\theta P(X, Z \mid \theta)$ is easy. In the coin-flipping context, then, the MLE of θ given coin tosses *and* coin labels is easy (recall that the estimate is simply θ^* = number of heads / number of flips).

$P(X \mid \theta)$ is hard to compute because $P(X \mid \theta) = \sum_Z P(X, Z \mid \theta) = \sum_Z P(X \mid Z, \theta)P(Z \mid \theta)$ which is computationally infeasible. We would have to sum over `|Z| ** num_experiments` values, e.g. 2^n in the coin-flipping case.

This would lead to

$$\theta^* = \arg\max_\theta \log \left[\sum_Z f(X|Z; \theta)P(Z; \theta) \right]$$

Again: computationally infeasible!

Accordingly, HEM replaces the offending sum $\sum_Z$ with a single dominant term Z^*, where

$$Z^* = \arg\max_Z P(X|X; \theta^{(m)})$$

for $\theta^{(m)}$ our current estimate of θ. This is the E-step. It then improves its guess with the M-step:

$$\theta^{(m+1)} = \arg\max_\theta f(X|Z^*; \theta)P(Z^*; \theta)$$

where Z^* is from the previous step. Finally it iterates between the E and M-steps until convergence.

On the other hand, SEM

- replaces $\log \mathbb{E}_{Z;\theta}[f(X|Z; \theta)P(Z; \theta)]$ with $\mathbb{E}_{Z;\theta}[\log f(X|Z; \theta)P(Z; \theta)]$, and
- replaces $\mathbb{E}_{Z;\theta}[...]$ with $\mathbb{E}_{Z|X;\theta^{(m)}}[...]$.

This leads to an E-step of

$$Q(\theta \mid \theta^{(m)}) = \mathbb{E}_{Z|X;\theta^{(m)}}[\log f(X|Z;\theta)P(Z;\theta)]$$

and an M-step of

$$\theta^{(m+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(m)})$$

Again we iterate between the E and M-steps until convergence.

This is the EM algorithm.

Note: if we let $L(X;\theta) = P(X \mid \theta) = \sum_Z P(X, Z \mid \theta)$, then the E-step becomes

$$Q(\theta \mid \theta^{(m)}) = \mathbb{E}_{Z|X;\theta^{(m)}}[\log L(X, Z;\theta)]$$

(The EM algorithm iteratively improves on $Q(\theta \mid \theta^{(m)})$, and by extension improves on $\log P(X \mid \theta)$.)