# EM algorithm (Expectation-Maximization)

- arises in many apps. involving parameter estimation involving prob. models with incomplete obs. data

ex.: RNA-seq. (comp. bio lab), clustering & classification (ML), speaker ID, ...

Ex.: Recall our toy casino prob. with a "Fair" die & a "Loaded" die. If we know the "F" & "L" parameters (e.g. $P(6) = \frac{1}{2}$ for "L"), then finding MLSE using the Viterbi Alg. is "easy."

But what if you don't?

Need to est. or learn the parameters: the EM alg. is popularly used.

# Toy Ex.: Coin-flipping Experiment

- Given a pair of coins A and B having unknown biases $\theta_A$ and $\theta_B$
- Goal: estimate $(\theta_A, \theta_B)$.

- Experiment ① Choose one of the two coins at random with equal prob.
② Perform 10 indep. coin flips with that coin.
③ Repeat ① & ② 5 times

- ML: when you get to observe which coin was used in Step ①: "Easy"

| | |
|---|---|
| B | H T T T H H T H T H |
| A | H H H H T H H H H H |
| A | ... |
| B | ... |
| A | ... |

| Coin | Coin |
|---|---|
| | 5H, 5T |
| 9H, 1T | |
| 8H, 2T | |
| | 4H, 4T |
| 7H, 3T | |
| 24H, 6T | 9H, 11T |

MLE: $\theta_A = \frac{24}{30} = 0.8$ $\theta_B = \frac{9}{20} = 0.45$

Aside:

$$\text{MLE}(\theta_A \mid H_A \text{ heads } \& \ T_A \text{ tails}) = ?$$

$$\theta_A^* = \underset{\theta}{\operatorname{argmax}} \ \mathbb{P}(H_A \text{ heads}, T_A \text{ tails} \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \ \theta^{H_A} \cdot (1-\theta)^{T_A}$$

$$f(\theta) = \theta^{H_A} \cdot (1-\theta)^{T_A}$$

$$\log f(\theta) = H_A \log \theta + T_A \log(1-\theta)$$

$$\frac{\partial \log f(\theta)}{\partial \theta} = \frac{H_A}{\theta} - \frac{T_A}{1-\theta} = 0$$

$$\Rightarrow \boxed{\theta_A^* = \frac{H_A}{H_A + T_A}}$$

b) $\boxed{\text{"Hard-decision" EM (HEM)}}$

1) Start with initial est. $\left(\hat{\theta}_A^{(0)}, \hat{\theta}_B^{(0)}\right) \overset{\text{e.g.}}{=} (0.6, 0.45)$

2) E-step:


$\left(\hat{\theta}_A^{(i)}, \hat{\theta}_B^{(i)}\right)$

$(0.6, 0.45)$

"Hard" decision on which coin

| | |
|---|---|
| 0 × A | 1 × B |
| 1 × A | 0 × B |
| 1 × A | 0 × B |
| 0 × A | 1 × B |
| 1 × A | 0 × B |

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

3

$\left(\hat{\theta}_A^{(1)} = 0.8, \hat{\theta}_B^{(1)} = 0.45\right)$

3) M-step: Re-estimate $\left(\hat{\theta}_A^{(i+1)}, \hat{\theta}_B^{(i+1)}\right)$ based on latest E-step.

4) Go back to E-step & iterate until convergence.

EM algorithm used when you do not get to observe which coin was used in step 1 (i.e. coin **label** info is "hidden")

c) "Soft-Decision" EM (SEM)

"Soft-decision" on coin label

| | | |
|---|---|---|
| (5H, 5T) | 0.45 × A | 0.55 × B |
| (9H, 1T) | 0.8 × A | 0.2 × B |
| (8H, 2T) | 0.73 × A | 0.27 × B |
| (4H, 6T) | 0.35 × A | 0.65 × B |
| (7H, 3T) | 0.65 × A | 0.35 × B |

| Coin A | Coin B |
|---|---|
| 2.2H, 2.2T | 2.8H, 2.8T |
| 7.2H, 0.8T | 1.8H, 0.2T |
| 5.9H, 1.5T | 2.1H, 0.5T |
| 1.4H, 2.1T | 2.6H, 3.9T |
| 4.5H, 1.9T | 2.5H, 1.1T |
| 21.3H, 8.6T | 11.7H, 8.4T |

$$P_{\theta^{(0)}}(A \mid 5H, 5T) = \frac{P(5H, 5T \mid A)\,P(A)}{P(5H, 5T \mid A)P(A) + P(5H, 5T \mid B)\cdot P(B)}$$

$\theta^{(0)}$: (0.6, 0.45)

$$= \frac{(0.6)^5 (0.4)^5 \frac{1}{2}}{(0.6)^5 (0.4)^5 \frac{1}{2} + (0.45)^5 (0.55)^5 \frac{1}{2}}$$

(Bayes')

$$= 0.45$$

(3) M-step: Coin A: $\hat{\theta}_A^{(1)} = \frac{21.3}{21.3 + 8.6} \cong 0.71$

Coin B: $\hat{\theta}_B^{(1)} = \frac{11.7}{11.7 + 8.4} \cong 0.58$

Continue ... $\hat{\theta}_A^{(10)} = 0.8$ ; $\hat{\theta}_B^{(10)} = 0.52$ CONVERGED!

# Hard-EM

$$* \boxed{P(A \mid \underline{\theta}^{(i)}, \underline{Y}) \underset{B}{\overset{A}{\gtrless}} P(B \mid \underline{\theta}^{(i)}, \underline{Y})}$$

# Soft-EM

$$(\dagger) \boxed{\begin{aligned} P(A \mid \underline{\theta}^{(i)}, \underline{Y}) &= \frac{P(\underline{Y} \mid A; \theta^{(i)}) P(A)}{P(\underline{Y}; \underline{\theta}^{(i)})} \\ P(B \mid \underline{\theta}^{(i)}, \underline{Y}) &= \frac{P(\underline{Y} \mid B; \theta^{(i)}) P(B)}{P(\underline{Y}; \underline{\theta}^{(i)})} \\ &\text{(BAYES' THM.)} \end{aligned}}$$

(*) Ex. $\underline{Y} = (5H, 5T)$ (first expt.)

$$P(A \mid \hat{\theta}_A^{(0)} = 0.6, \hat{\theta}_B^{(0)} = 0.45, \text{5 Heads, 5 Tails})$$
$$< P(B \mid \hat{\theta}_A^{(0)} = 0.6, \hat{\theta}_B^{(0)} = 0.45, 5H, 5T)$$

(†) Ex. $P_{\theta^{(0)}}(A \mid 5H, 5T) = \dfrac{P(5H, 5T \mid A) P(A)}{P(5H, 5T \mid A) P(A) + P(5H, 5T \mid B) P(B)}$

$$= \frac{(0.6)^5 (0.4)^5 \frac{1}{2}}{(0.6)^5 (0.4)^5 \frac{1}{2} + (0.45)^5 (0.55)^5} = \boxed{0.45}$$

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

$Y = (X, Z)$

"complete" data

"hidden" or "missing" or "latent" data

"incomplete" data

In our coin-flipping ex.

$\underbrace{\text{Complete data}}_{Y} \equiv (\underbrace{\text{COIN TOSSES}}_{X}, \underbrace{\text{COIN LABELS}}_{Z})$

e.g. $((5H, 5T), \ldots)$

e.g. (BABBA)

---

$\operatorname{argmax}_{\theta} P(X \mid \theta)$ is "hard"

but $\operatorname{argmax}_{\theta} P(X, Z \mid \theta)$ is "easy"

e.g. in our coin-flipping ex., MLE$\left(\theta \,\middle|\, \begin{array}{c}\text{coin tosses} \\ \text{AND} \\ \text{coin labels}\end{array}\right)$

is "easy" $\left(\theta^* = \frac{\#H}{\#H + \#T}\right)$

$$\boxed{\theta^* = \arg\max_{\theta} \log\left[\sum_z f(x \mid z;\theta)\, p(z;\theta)\right]}$$ ← GOAL (COMPUTATIONALLY INFEASIBLE)

H-EM: Replace $\sum_z$ by a single-term $z^*$

where

(1) "E-step" ← $$\boxed{z^* = \arg\max_z p(z \mid x; \underbrace{\theta^{(m)}}_{\text{current est. of } \theta})}$$

then improve guess

(2) "M-step" ← $$\boxed{\theta^{(m+1)} = \arg\max_{\theta} f(x \mid z^*; \theta)\, p(z^*; \theta)}$$

where $z^*$ is from the previous step

Iterate between (1) & (2) until convergence.

# S-EM

(a) Replace $\log\left\{ \mathbb{E}_{z;\theta}\left[ \frac{f(x|z;\theta)}{p(z;\theta)} \right] \right\}$ by

$$\mathbb{E}_{z;\theta}\left\{ \log\left[ \frac{f(x|z;\theta)}{p(z;\theta)} \right] \right\}$$

"actual" A

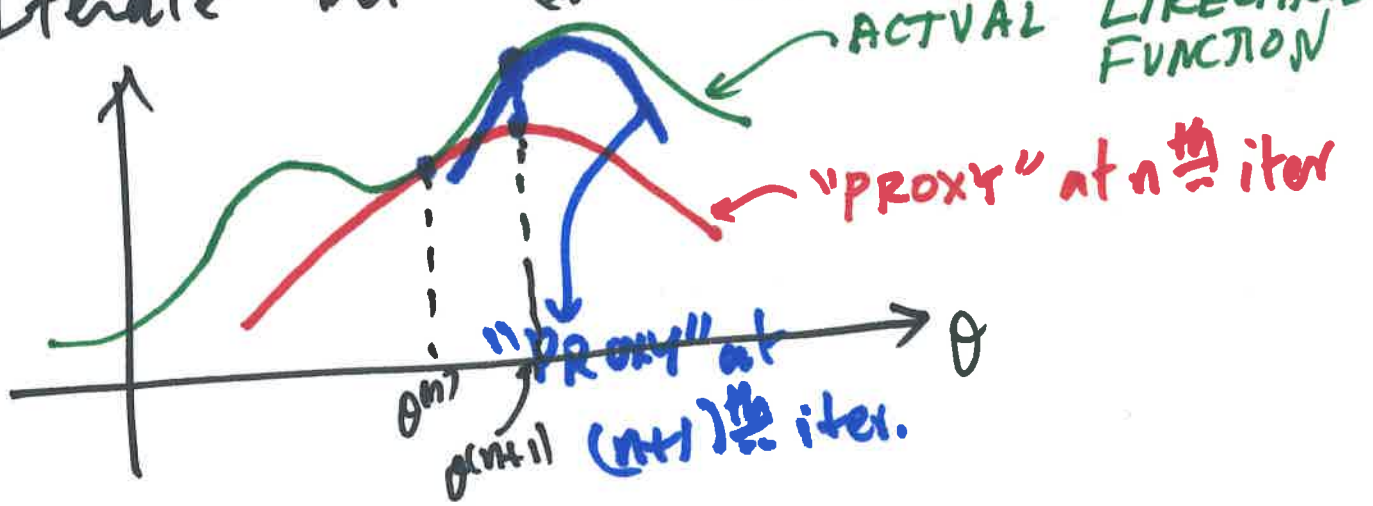
B "proxy"

(b) Replace $\mathbb{E}_{z;\theta}\{\cdot\}$ by $\mathbb{E}_{z|x;\theta^{(m)}}\{\cdot\}$

(3) $$\boxed{\text{E-step: } Q(\theta|\theta^{(m)}) = \mathbb{E}_{z|x;\theta^{(m)}}\left[ \log \frac{f(x|z;\theta)}{p(z;\theta)} \right]}$$

(4) $$\boxed{\text{M-step: } \theta^{(m+1)} = \underset{\theta}{\operatorname{argmax}}\ Q(\theta|\theta^{(m)})}$$

Iterate bet. (3) & (4) till convergence.



$$L(X;\theta) \triangleq p(X|\theta) = \sum_Z p(X,Z|\theta)$$

$$\boxed{\begin{array}{l} \underline{\text{E-step}}: \; Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}\{\log L(X,Z;\theta)\} \\ \\ \underline{\text{M-step}}: \; \theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)}) \end{array}}$$

(E-M alg)

Pf. of correctness: EM iterates and improves on $Q(\theta|\theta^{(t)})$ rather than $\log p(X|\theta)$. Improvements to the former $\Rightarrow$ improvements to the latter.

$$\log p(X|\theta) = \log p(X,Z|\theta) - \log p(Z|X,\theta)$$

Multiply by $p(Z|X,\theta^{(t)})$ and sum over $Z$

$$\sum_Z p(Z|X,\theta^{(t)}) \log p(X|\theta) = \overbrace{\sum_Z p(Z|X,\theta^{(t)}) \log p(X,Z|\theta)}^{Q(\theta|\theta^{(t)})} \underbrace{- \sum_Z p(Z|X,\theta^{(t)}) \log p(Z|X,\theta)}_{\triangleq H(\theta|\theta^{(t)})}$$

$$\boxed{\log p(X|\theta) = Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)})} \quad (A)$$

(A) is true for all values of $\theta$, incl. $\theta = \theta^{(t)}$

$$\boxed{\log p(X|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)})} \quad (B)$$

Subtract (B) from (A):

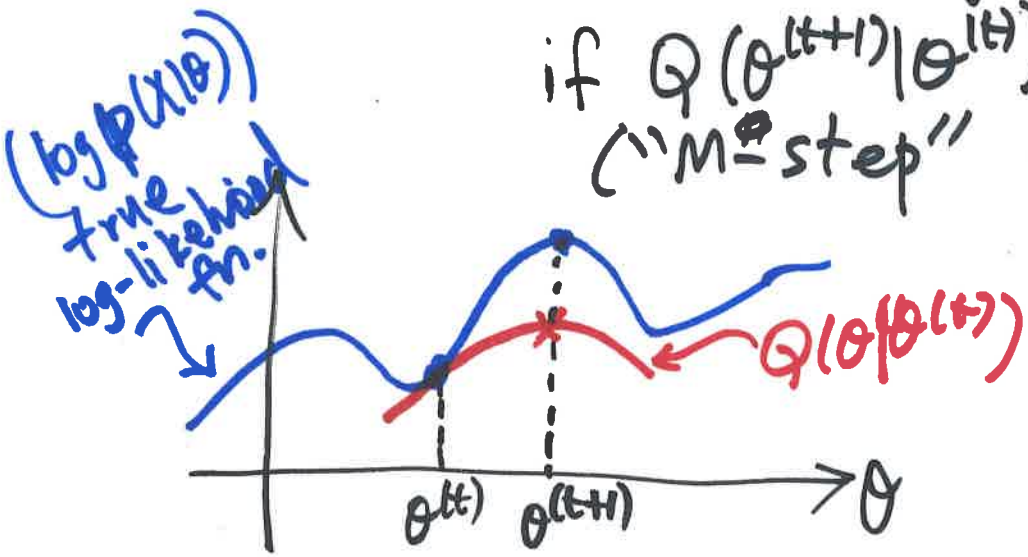
$$\log p(X|\theta) - \log p(X|\theta^{(t)}) = Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \underbrace{H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})}_{\geq 0 \text{ (to be shown)}}$$

$$\Rightarrow \log p(X|\theta) - \log p(X|\theta^{(t)}) \geq \underbrace{Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})}_{\geq 0 \text{ if } Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})}$$

$\Rightarrow$ Thus, $\boxed{\log p(X|\theta^{(t+1)}) \geq \log p(X|\theta^{(t)})}$

if $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)})$

("M-step" in E-M alg.)



Proof of $\boxed{H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \geq 0}$

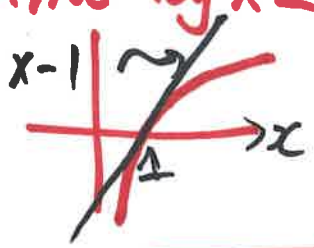
defn. of $H(\cdot)$

$$\text{LHS} = \sum_z p(z|x,\theta^{(t)}) \log\left[\frac{p(z|x,\theta^{(t)})}{p(z|x,\theta)}\right]$$

$$= D\left(p(z|x,\theta^{(t)}) \,\|\, p(z|x,\theta)\right) \geq 0$$

Recall $D(p\|q) \triangleq -\sum_i p_i \log \frac{q_i}{p_i}$

True since $\log x \leq x-1$




$$\geq \sum_i p_i \left[\frac{q_i}{p_i} - 1\right]$$

$$= \sum_i q_i - \sum_i p_i = 0$$

<u>Note</u>: $H(\theta^{(t)}|\theta^{(t)}) = -\sum_z p(z|x,\theta^{(t)}) \log p(z|x,\theta^{(t)})$

= ENTROPY OF THE COND. DIST. $\{z|x,\theta^{(t)}\}$