



**CSE 4392 SPECIAL TOPICS**  
**NATURAL LANGUAGE PROCESSING**

# **Expectation Maximization**

1

2024 Spring

# INTUITION OF EM

- Let's say I have 3 coins in my pocket,
  - Coin 0 has probability  $\lambda$  of heads
  - Coin 1 has probability  $p_1$  of heads
  - Coin 2 has probability  $p_2$  of heads
- **For each trial:**
  - First, I toss Coin 0
  - If coin 0 turns up **heads**, I toss coin 1 three times
  - If coin 0 turns up **tails**, I toss coin 2 three times
  - I don't tell you the results of the coin 0 toss, or whether coin 1 or coin 2 was tossed, but I tell you how many heads/tails are seen after each trial
- • You see the following sequence:  
 $\langle H, H, H \rangle, \langle T, T, T \rangle, \langle H, H, H \rangle, \langle T, T, T \rangle, \langle H, H, H \rangle$

*Quiz: Guess what are the estimated values of  $\lambda, p_1, p_2$ ?*

# MAXIMAL LIKELIHOOD ESTIMATE

- Data points  $x_1, x_2, \dots, x_n$  from (finite or countable) set  $\mathcal{X}$  ( $x_i$  is a triplet of three tosses)
- Parameter vector  $\theta$
- Parameter space  $\Omega$
- We have a distribution  $P(x | \theta)$  for any  $\theta \in \Omega$ , such that

$$\sum_{x \in \mathcal{X}} P(x | \theta) = 1$$
$$P(x | \theta) \geq 0, \forall x$$

- Assume data points are drawn independently and identically distributed from a distribution  $P(x | \theta^*)$  for some  $\theta^* \in \Omega$

# LOG LIKELIHOOD

- Probability distribution  $P(x | \theta)$  for any  $\theta \in \Omega$
- Likelihood of  $\theta$ :

$$\text{Likelihood}(\theta) = P(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n P(x_i | \theta)$$

- Log likelihood of  $\theta$ :

$$L(\theta) = \sum_{i=1}^n \log P(x_i | \theta)$$

## EXAMPLE 1: COIN TOSSING

- $\mathcal{X} = \{H, T\}$ . Our data set  $x_1, x_2, \dots, x_n$  is a sequence of heads and tails, e.g.,

HTHTHHHHTTT

- Parameter vector  $\theta$  is a single parameter, i.e. probability of coin showing heads
- Parameter space  $\Omega = [0, 1]$

- Distribution 
$$P(x|\theta) = \begin{cases} \theta & \text{if } x = H \\ 1 - \theta & \text{if } x = T \end{cases}$$

## EXAMPLE 2: MARKOV CHAINS

- $\mathcal{X}$  is the set of all possible state (or tag) sequences generated by an underlying generative process. Our sample is  $n$  sequences  $X_1, X_2, \dots, X_n$ , where  $X_i \in \mathcal{X}$ .
- $\theta_T$  is the vector of all transition ( $s_i \rightarrow s_j$ ) parameters. W.L.O.G., we assume there is a dummy start state  $\phi$  and initial transition  $\phi \rightarrow s_1$
- Let  $T(\alpha) \subset T$  be all transition of the form  $\alpha \rightarrow \beta$
- $\Omega$  is the set of  $\theta \in [0,1]^{|S+1||S|}$  where  $S$  is the set of all states (tags), such that:

$$\forall \alpha \in S, \sum_{t \in T(\alpha)} \theta_t = 1$$

## EXAMPLE 2: MARKOV CHAINS

- Since  $\theta_T$  is the vector of all transtion parameters
- We have:

$$P(X | \theta_T) = \prod_{t \in T} \theta_t^{\text{Count}(X, t)}$$

where  $\text{Count}(X, t)$  is the number of times transition  $t$  occures in sequence  $X$ .

- This gives:

$$\log(P(X | \theta_T)) = \sum_{t \in T} \text{Count}(X, t) \log \theta_t$$

$$L(\theta_T) = \sum_i \log P(X_i | \theta_T) = \sum_i \sum_{t \in T} \text{Count}(X, t) \log \theta_t$$

# MLE FOR MARKOV CHAINS

- We use  $\theta$  for  $\theta_T$  for simplicity
- To solve for  $\theta_{MLE} = \arg \max_{\theta \in \Omega} L(\theta)$
- We solve  $\theta$  in

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

with appropriate probability constraints

- Therefore:

$$\theta_t = \frac{\sum_i \text{Count}(X_i, t)}{\sum_i \sum_{t' \in T(\alpha)} \text{Count}(X_i, t')}$$

where  $t$  is of the form  $\alpha \rightarrow \beta$  for some  $\beta$ ,  $T(\alpha)$  is all the transitions originating from  $\alpha$ .



# MODELS WITH HIDDEN VARIABLES

- Suppose we have two sets  $\mathcal{X}$  and  $\mathcal{Y}$ , and a joint distribution  $P(x, y \mid \theta)$

- If we have **fully-observable data**,  $(x_i, y_i)$  pairs, then

$$L(\theta) = \sum_i \log P(x_i, y_i \mid \theta)$$

- If we have **partially-observable data**,  $x_i$  examples only, then

$$\begin{aligned} L(\theta) &= \sum_i \log P(x_i \mid \theta) \\ &= \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta) \end{aligned}$$

- This is **unsupervised learning**, very similar to **clustering**.
- We will use *an iterative algorithm* to infer  $\theta$  like k-means

# EXPECTATION-MAXIMILATION

- If we have **partially-observable data**,  $x_i$  examples only, then

$$L(\theta) = \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta)$$

- The EM (Expectation Maximization) algorithm is a method for finding

$$\theta_{MLE} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \theta)$$

# THREE COINS EXAMPLE

- In the three-coin example:
  - $\mathcal{Y} = \{H, T\}$  (possible outcomes of coin 0)
  - $\mathcal{X} = \{HHH, TTT, HTT, THH, HHT, TTH, HTH, THT\}$
  - $\theta = \{\lambda, p_1, p_2\}$
- And  $P(x, y \mid \theta) = P(y \mid \theta) P(x \mid y, \theta)$

where

$$P(y \mid \theta) = \begin{cases} \lambda & \text{if } y = H \\ 1 - \lambda & \text{if } y = T \end{cases}$$

and

$$P(x \mid y, \theta) = \begin{cases} p_1^h (1 - p_1)^t & \text{if } y = H \\ p_2^h (1 - p_2)^t & \text{if } y = T \end{cases}$$

$h$  is num of heads in  $x$   
 $t$  is num of tails in  $x$

## THREE COINS EXAMPLE

- Calculate various probabilities:

one H and two T  
from THT

$$P(x = THT, y = H | \theta) = \lambda p_1 (1 - p_1)^2$$

$$P(x = THT, y = T | \theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

$$\begin{aligned} P(x = THT | \theta) &= P(x = THT, y = H | \theta) + P(x = THT, y = T | \theta) \\ &= \lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2 \end{aligned}$$

$$\begin{aligned} P(y = H | x = THT, \theta) &= \frac{P(x = THT, y = H | \theta)}{P(x = THT | \theta)} \quad (\text{Bayes rule}) \\ &= \frac{\lambda p_1 (1 - p_1)^2}{\lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2} \end{aligned}$$

## THREE COINS EXAMPLE

- Suppose fully observed data looks like:

$(\langle HHH \rangle, H), (\langle TTT \rangle, T), (\langle HHH \rangle, H), (\langle TTT \rangle, T), (\langle HHH \rangle, H)$

- In this case, the maximum likelihood estimates of the parameters are:

$$\lambda = \frac{3}{5}$$
$$p_1 = \frac{9}{9} = 1$$
$$p_2 = \frac{0}{6} = 0$$

## THREE COINS EXAMPLE

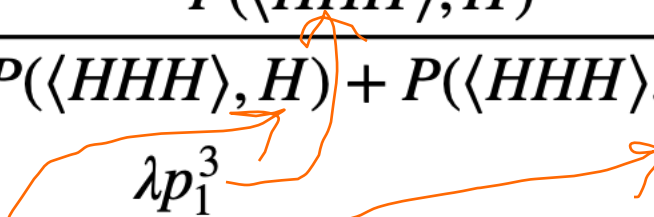
- Partial observed data might look like:  
 $\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$
- How do you estimate the MLE parameters?

## THREE COINS EXAMPLE

- Partial observed data might look like:

$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$

- If the current parameters are  $\lambda, p_1, p_2$

$$\begin{aligned} P(y = H | x = \langle HHH \rangle) &= \frac{P(\langle HHH \rangle, H)}{P(\langle HHH \rangle, H) + P(\langle HHH \rangle, T)} \\ &= \frac{\lambda p_1^3}{\lambda p_1^3 + (1 - \lambda) p_2^3} \end{aligned}$$


$$\begin{aligned} P(y = H | x = \langle TTT \rangle) &= \frac{P(\langle TTT \rangle, H)}{P(\langle TTT \rangle, H) + P(\langle TTT \rangle, T)} \\ &= \frac{\lambda (1 - p_1)^3}{\lambda (1 - p_1)^3 + (1 - \lambda) (1 - p_2)^3} \end{aligned}$$

## THREE COINS EXAMPLE

- If the current parameters are  $\lambda, p_1, p_2$

$$\begin{aligned}P(y = H | x = \langle HHH \rangle) &= \frac{P(\langle HHH \rangle, H)}{P(\langle HHH \rangle, H) + P(\langle HHH \rangle, T)} \\&= \frac{\lambda p_1^3}{\lambda p_1^3 + (1 - \lambda) p_2^3}\end{aligned}$$

$$\begin{aligned}P(y = H | x = \langle TTT \rangle) &= \frac{P(\langle TTT \rangle, H)}{P(\langle TTT \rangle, H) + P(\langle TTT \rangle, T)} \\&= \frac{\lambda(1 - p_1)^3}{\lambda(1 - p_1)^3 + (1 - \lambda)(1 - p_2)^3}\end{aligned}$$

- If  $\lambda=0.3, p_1 = 0.3, p_2 = 0.6$

$$P(y = H | x = \langle HHH \rangle) = 0.0508$$

$$P(y = H | x = \langle TTT \rangle) = 0.6967$$



# THREE COINS EXAMPLE

- After filling in hidden variables for each example, the partially observed data looks like this:

$(\langle \text{HHH} \rangle, H)$	$P(y = H \mid \text{HHH}) = 0.0508$	}	sum to 1
$(\langle \text{HHH} \rangle, T)$	$P(y = T \mid \text{HHH}) = 0.9492$		
$(\langle \text{TTT} \rangle, H)$	$P(y = H \mid \text{TTT}) = 0.6967$	}	sum to 1
$(\langle \text{TTT} \rangle, T)$	$P(y = T \mid \text{TTT}) = 0.3033$		
$(\langle \text{HHH} \rangle, H)$	$P(y = H \mid \text{HHH}) = 0.0508$	}	sum to 1
$(\langle \text{HHH} \rangle, T)$	$P(y = T \mid \text{HHH}) = 0.9492$		
$(\langle \text{TTT} \rangle, H)$	$P(y = H \mid \text{TTT}) = 0.6967$	}	sum to 1
$(\langle \text{TTT} \rangle, T)$	$P(y = T \mid \text{TTT}) = 0.3033$		
$(\langle \text{HHH} \rangle, H)$	$P(y = H \mid \text{HHH}) = 0.0508$	}	sum to 1
$(\langle \text{HHH} \rangle, T)$	$P(y = T \mid \text{HHH}) = 0.9492$		

# THREE COINS EXAMPLE

- New estimates:

$$(\langle \text{HHH} \rangle, H) \quad P(y = H \mid \text{HHH}) = 0.0508$$

$$(\langle \text{HHH} \rangle, T) \quad P(y = T \mid \text{HHH}) = 0.9492$$

$$(\langle \text{TTT} \rangle, H) \quad P(y = H \mid \text{TTT}) = 0.6967$$

$$(\langle \text{TTT} \rangle, T) \quad P(y = T \mid \text{TTT}) = 0.3033$$

...

$$\lambda = \frac{3 \times 0.0508 + 2 \times 0.6967}{5} = 0.3092$$

how many  
heads in  $X_i$ ?

out of 5 coin 0 tosses  
how many are heads?

$$p_1 = \frac{3 \times 3 \times 0.0508 + 0 \times 2 \times 0.6967}{3 \times 3 \times 0.0508 + 3 \times 2 \times 0.6967} = 0.0987$$

$$p_2 = \frac{3 \times 3 \times 0.9492 + 0 \times 2 \times 0.3033}{3 \times 3 \times 0.9492 + 3 \times 2 \times 0.3033} = 0.8244$$

# SUMMARY OF THREE COINS EXAMPLE

- Begins with  $\lambda=0.3$ ,  $p_1 = 0.3$ ,  $p_2 = 0.6$

- Fill in hidden variables using:

$$P(y = H \mid x = \langle HHH \rangle) = 0.0508$$


$$P(y = H \mid x = \langle TTT \rangle) = 0.6967$$

- Re-estimate parameters to be

$$\lambda=0.3092, p_1 = 0.0987, p_2 = 0.8244$$

# EM INTERACTIONS

$$P(y = H | X_i)$$



Iteration	$\lambda$	$p_1$	$p_2$	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$	$\bar{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

- Coin example for  $\{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle\}$
- $\lambda$  is now 0.4, indicating that coin 0 has a probability 0.4 of selecting the tail-biased coin (coin 1)
- $\theta$  (parameters) are like the **cluster centers** in k-means

# EM INTERACTIONS

Iteration	$\lambda$	$p_1$	$p_2$	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

- Coin example for  $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$ .
- This solution of  $\lambda = 0.5, p_1 = 0$ , and  $p_2 = 1$  is intuitively correct: the coin tosser has two coins, one which always shows heads, and another which always shows tails, and is picking between them with equal probability .
- Posterior probabilities  $\bar{p}_i$  show that we are certain that coin 1 (tail-biased) generate  $x_2$  and  $x_4$ , whereas coin 2 generated  $x_1$  and  $x_3$ .



# INITIALIZATION MATTERS

Iteration	$\lambda$	$p_1$	$p_2$	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000
1	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
2	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
3	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
4	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
5	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
6	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000

- Coin example for  $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$ .
- In this case, EM is stuck in a “**saddle point**”, or local optimal.

# INITIALIZATION MATTERS

Iteration	$\lambda$	$p_1$	$p_2$	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7001	0.7000	0.3001	0.2998	0.3001	0.2998
1	0.2999	0.5003	0.4999	0.3004	0.2995	0.3004	0.2995
2	0.2999	0.5008	0.4997	0.3013	0.2986	0.3013	0.2986
3	0.2999	0.5023	0.4990	0.3040	0.2959	0.3040	0.2959
4	0.3000	0.5068	0.4971	0.3122	0.2879	0.3122	0.2879
5	0.3000	0.5202	0.4913	0.3373	0.2645	0.3373	0.2645
6	0.3009	0.5605	0.4740	0.4157	0.2007	0.4157	0.2007
7	0.3082	0.6744	0.4223	0.6447	0.0739	0.6447	0.0739
8	0.3593	0.8972	0.2773	0.9500	0.0016	0.9500	0.0016
9	0.4758	0.9983	0.0477	0.9999	0.0000	0.9999	0.0000
10	0.4999	1.0000	0.0001	1.0000	0.0000	1.0000	0.0000
11	0.5000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

Coin example for  $x = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$ .

- Just modify  $p_1$  a bit, EM is able to skip the saddle point and reach global optimum.

# THE EM ALGORITHM

- $\theta^t$  is the parameter vector at the  $t^{\text{th}}$  iteration.
- Choose  $\theta^0$  at random (or using some smart heuristics)
- Iterative procedure defined as:

$$\theta^t = \arg \max_{\theta} Q(\theta, \theta^{t-1})$$

where

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y | x_i, \theta^{t-1}) \log P(x_i, y | \theta)$$



# THE EM ALGORITHM

- (E-step): Compute *expected* counts.

$$\overline{Count}(r) = \sum_{i=1}^n \sum_y P(y | x_i, \theta^{t-1}) Count(x_i, y, r)$$

for every parameter  $\theta_r$ , e.g.,

$$\overline{Count}(DT \rightarrow NN) = \sum_i \sum_y P(S | O_i, \theta^{t-1}) Count(O_i, S, \theta_{DT \rightarrow NN})$$

- (M-step): Re-estimate parameters using expected counts to *maximize* likelihood.

e.g.,

$$\theta_{DT \rightarrow NN} = \frac{\overline{Count}(DT \rightarrow NN)}{\sum_{\beta} \overline{Count}(DT \rightarrow \beta)}$$

# THE EM ALGORITHM

- Intuition: Fill in hidden variables according to  $P(y \mid x_i, \theta)$
- EM is guaranteed to converge to a local maximum, or saddle-point, of the likelihood function
- In general, if

$$\arg \max_{\theta} \sum_i \log P(x_i, y_i \mid \theta)$$

has a simple analytic solution, then

$$\arg \max_{\theta} \sum_i \sum_y P(y \mid x_i, \theta) \log P(x_i, y \mid \theta)$$

also has a simple solution.

## EXAMPLE: EM FOR HMM

- We observe only word sequences  $X_1, X_2, \dots, X_n$  (no tags)
- $\theta$  is the vector of all transition parameters (include initial state distribution as a special case,  $\phi \rightarrow s$ )
- $\phi$  is the vector of all emission parameters
- Initialize parameters  $\theta^0$  and  $\phi^0$

## EXAMPLE: EM FOR HMM

- Initialize parameters  $\theta^0$  and  $\phi^0$
- E-step:

$$\overline{Count}(\theta_k) = \sum_{i=1}^n \sum_Y P(Y|X_i, \theta^{t-1}, \phi^{t-1}) Count(X_i, Y, \theta_k)$$

$$= \sum_{i=1}^n \sum_Y P(Y|X_i, \theta^{t-1}, \phi^{t-1}) Count(Y, \theta_k)$$

$$\overline{Count}(\phi_k) = \sum_{i=1}^n \sum_Y P(Y|X_i, \theta^{t-1}, \phi^{t-1}) Count(X_i, Y, \phi_k)$$

## EXAMPLE: EM FOR HMM

### ◦ M-step:

$$\theta_k^t = \frac{\overline{Count}(\theta_k)}{\sum_{\theta' \in M(\theta_k)} \overline{Count}(\theta')}$$

where  $M(\theta_k)$  is the set of all transitions ( $a \rightarrow b$ , for all  $b$ ) that share the same previous state as the  $k^{\text{th}}$  transition ( $a \rightarrow c$ , for some  $c$ )

$$\phi_k^t = \frac{\overline{Count}(\phi_k)}{\sum_{\phi' \in M'(\phi_k)} \overline{Count}(\phi')}$$

where  $M'(\phi_k)$  is the set of all emissions ( $a \rightarrow x$ , for all  $x$ ) that share the same previous state as the  $k^{\text{th}}$  emission ( $a \rightarrow x'$ , for some  $x'$ ).