COS 484/584

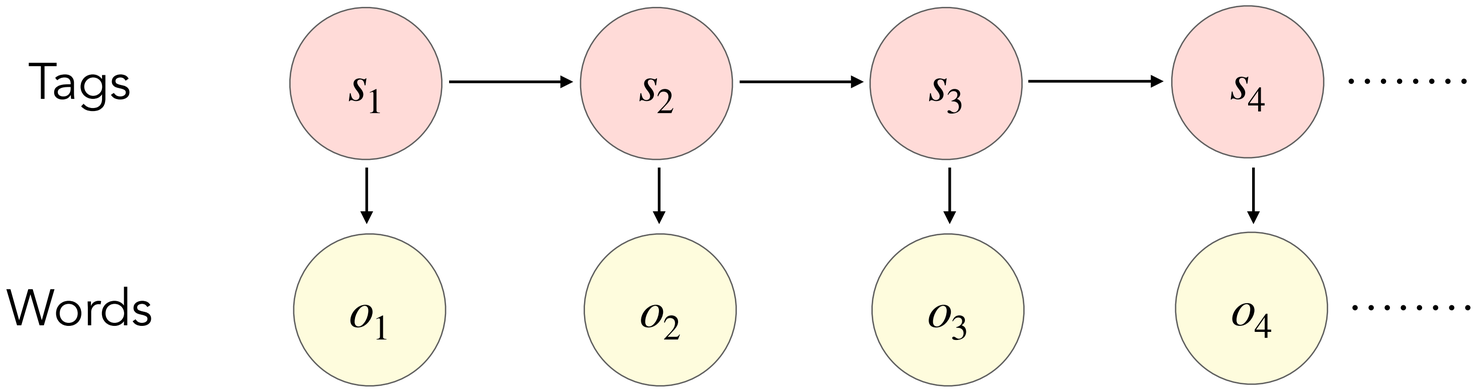
Sequence Models - 2

Spring 2021

Midterm

- **Topics:** Lectures up to March 8 (RNNs, Neural LMs)
- Logistics of the exam will be announced on Canvas tomorrow
- Last year's midterm will be made available
  - Not all topics are relevant for this year, but you will get a sense for format and types of questions
- Midterm review: COS 484 precept this week (March 5)

Recap: Hidden Markov Models



1. Set of states S = {1, 2, ..., K} and set of observations O
2. Initial state probability distribution $\pi(s_1)$
3. Transition probabilities $P(s_{t+1} \mid s_t)$ (OR $\theta_{s_t \rightarrow s_{t+1}}$)
4. Emission probabilities $P(o_t \mid s_t)$ (OR $\phi_{s_t \rightarrow o_t}$)

Strong assumptions

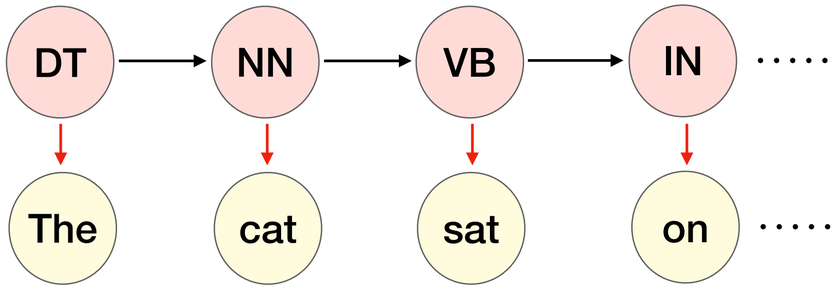
Maximum Entropy Markov Models

Generative vs Discriminative

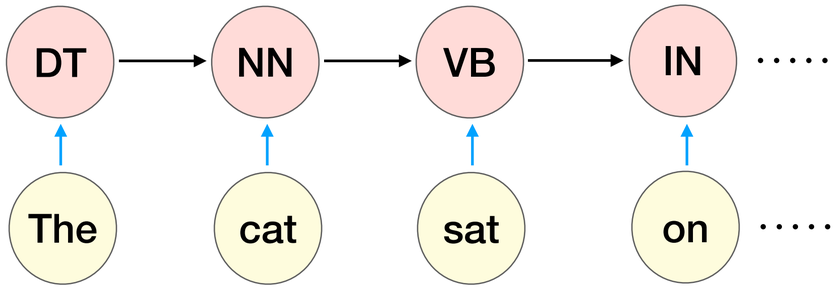
- HMM is a *generative* model
- Can we model $P(s_1, \ldots, s_n \mid o_1, \ldots, o_n)$ directly?

	Generative	**Discriminative**
Text classification	Naive Bayes: $P(c)P(d \mid c)$	Logistic Regression: $P(c \mid d)$
Sequence prediction	HMM: $P(s_1, \ldots, s_n)P(o_1, \ldots, o_n \mid s_1, \ldots, s_n)$	MEMM: $P(s_1, \ldots, s_n \mid o_1, \ldots, o_n)$ (No factorization)

Maximum Entropy Markov Model



HMM



MEMM

$$O = \langle o_1, o_2, \ldots, o_n \rangle$$

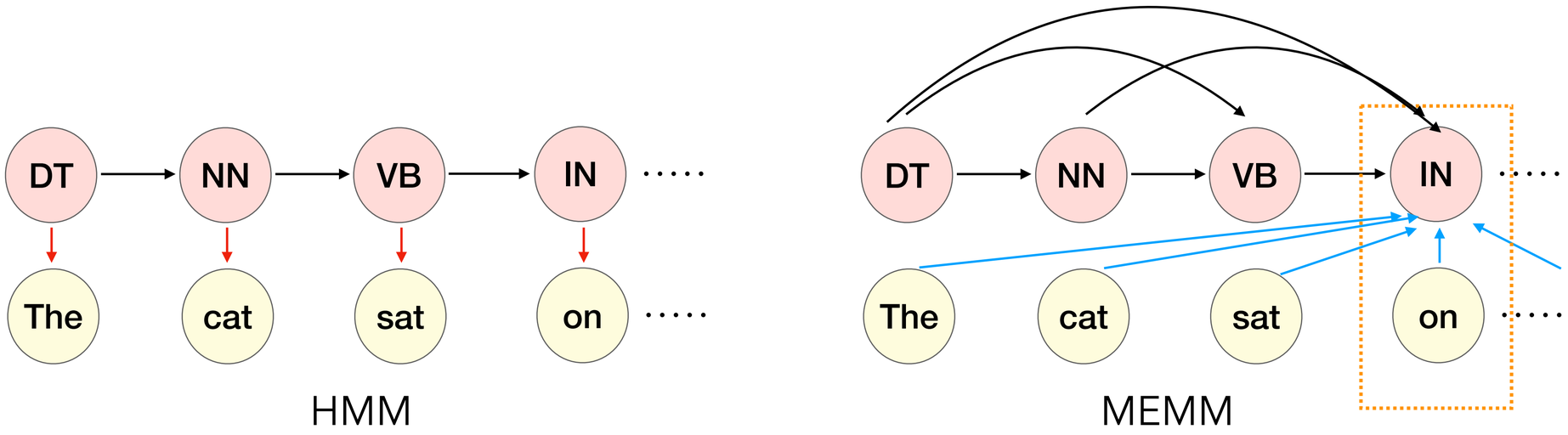
- Compute the posterior directly:
  - $\hat{S} = \arg\max_S P(S \mid O) \approx \arg\max_S \prod_i P(s_i \mid o_i, s_{i-1})$ ← No factorization into transition, emission
- Use features: $P(s_i \mid o_i, s_{i-1}) \propto \exp(w \cdot f(s_i, o_i, s_{i-1}))$

Features

weights

(Bigram MEMM)

MEMM



- In general, we can use all observations and all previous states:

$$\hat{S} = \arg\max_{S} P(S \mid O) = \arg\max_{S} \prod_{i} P(s_i \mid o_n, o_{i-1}, \ldots, o_1, s_{i-1}, \ldots, s_1)$$

$$P(s_i \mid s_{i-1}, \ldots, s_1, O) \propto \exp(w \cdot f(s_i, s_{i-1}, \ldots, s_1, O)$$

Why couldn't we do this with HMMs?

Features in an MEMM

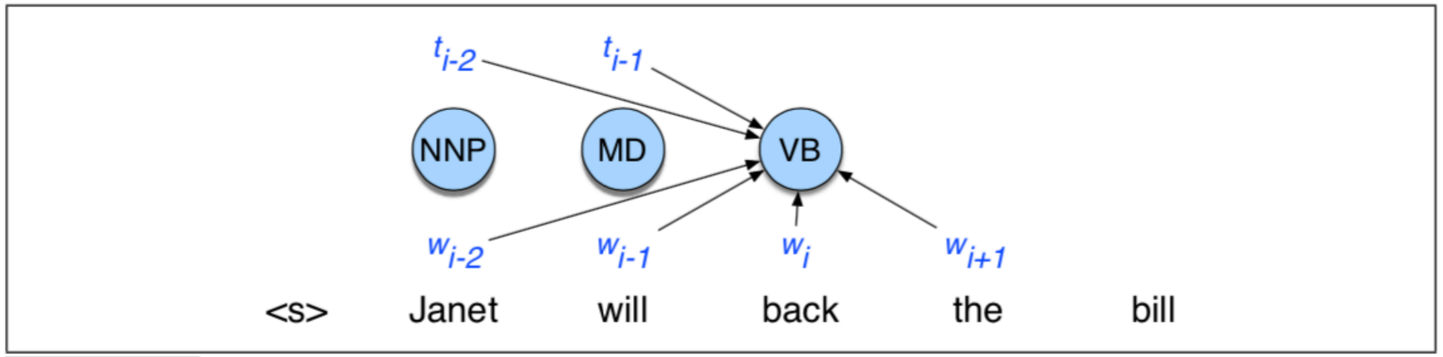


Figure 8.13 An MEMM for part-of-speech tagging showing the ability to condition on more features.

t = tags (states)

w = words (observations)

$$\langle t_i, w_{i-2}\rangle, \langle t_i, w_{i-1}\rangle, \langle t_i, w_i\rangle, \langle t_i, w_{i+1}\rangle, \langle t_i, w_{i+2}\rangle$$
$$\langle t_i, t_{i-1}\rangle, \langle t_i, t_{i-2}, t_{i-1}\rangle,$$
$$\langle t_i, t_{i-1}, w_i\rangle, \langle t_i, w_{i-1}, w_i\rangle \langle t_i, w_i, w_{i+1}\rangle,$$

Feature templates

$t_i = \text{VB}$ and $w_{i-2} = \text{Janet}$
$t_i = \text{VB}$ and $w_{i-1} = \text{will}$
$t_i = \text{VB}$ and $w_i = \text{back}$
$t_i = \text{VB}$ and $w_{i+1} = \text{the}$
$t_i = \text{VB}$ and $w_{i+2} = \text{bill}$
$t_i = \text{VB}$ and $t_{i-1} = \text{MD}$
$t_i = \text{VB}$ and $t_{i-1} = \text{MD}$ and $t_{i-2} = \text{NNP}$
$t_i = \text{VB}$ and $w_i = \text{back}$ and $w_{i+1} = \text{the}$

Features

Features in an MEMM

Incorrect	DT	JJ	NN	DT	NN
Correct	DT	NN	VB	DT	NN
	The	old	man	the	boat

Which of these feature templates would help most to tag 'old' correctly?

A) $\langle t_i, w_i \rangle$

B) $\langle t_i, w_i, w_{i-1} \rangle$

C) $\langle t_i, w_i, w_{i-1}, w_{i+1} \rangle$

D) $\langle t_i, w_i, w_{i-1}, w_{i+1}, w_{i+2} \rangle$

t = tags

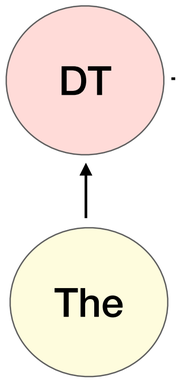
w = words

MEMMs: Decoding

$$\hat{S} = \arg\max_{S} P(S \mid O) = \arg\max_{S} \Pi_i P(s_i \mid o_i, s_{i-1})$$

(assume features only on previous time step and current obs)

- Greedy decoding:



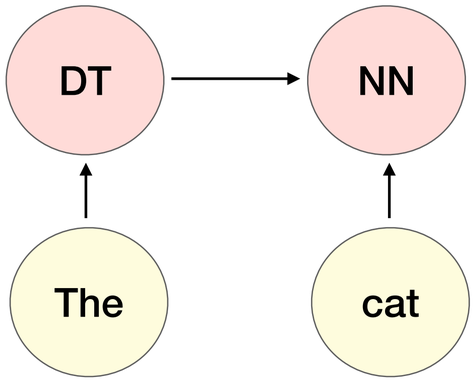
$$\hat{S}_1 = \underset{S}{\text{argmax}}\ P(S \mid \text{The})$$

$$= DT$$

MEMMs: Decoding

$$\hat{S} = \arg\max_{S} P(S \mid O) = \arg\max_{S} \Pi_i P(s_i \mid o_i, s_{i-1})$$

- Greedy decoding:



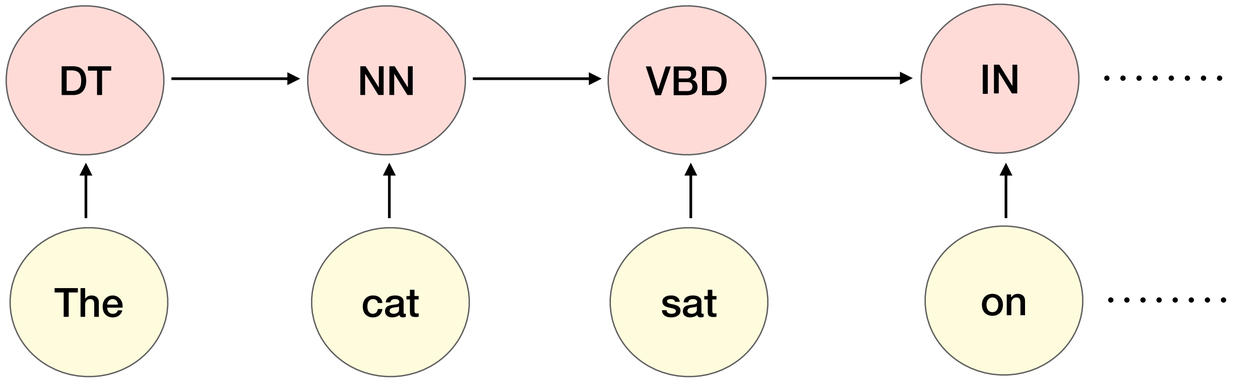
$$\hat{S}_2 = \arg\max_{S} P(S \mid \text{cat}, \text{DT})$$

$$= \text{NN}$$

MEMMs: Decoding

$$\hat{S} = \arg\max_{S} P(S \mid O) = \arg\max_{S} \Pi_i P(s_i \mid o_i, s_{i-1})$$

- Greedy decoding:



$$\forall t, \quad \hat{s}_{t+1} = \underset{s}{\text{argmax}}\ P(s \mid o_{t+1}, \hat{s}_t)$$

MEMMs: Decoding

$$\hat{S} = \arg\max_S P(S \mid O) = \arg\max_S \Pi_i P(s_i \mid o_i, s_{i-1})$$

- Greedy decoding

- Viterbi decoding:

$$M[i,j] = \max_k M[i-1,k]\ P(s_j \mid o_i, s_k) \quad 1 \le k \le K \quad 1 \le i \le n$$

(or equivalent log form)

DP Lattice

states

timesteps

(Best sequence ending in s_j)

MEMMs: Decoding

What do you think of the computational complexity of Viterbi decoding for bigram MEMMs compared to decoding for bigram HMMs?

A) More operations in MEMM

B) More operations in HMM

C) Equal

D) Depends on number of features in MEMM

$$M[i,j] = \max_k M[i-1,k]\ P(s_j \mid o_i, s_k) \quad 1 \le k \le K \quad 1 \le i \le n$$

DP Lattice

states

timesteps

MEMM: Learning

- **Gradient descent:** similar to logistic regression!

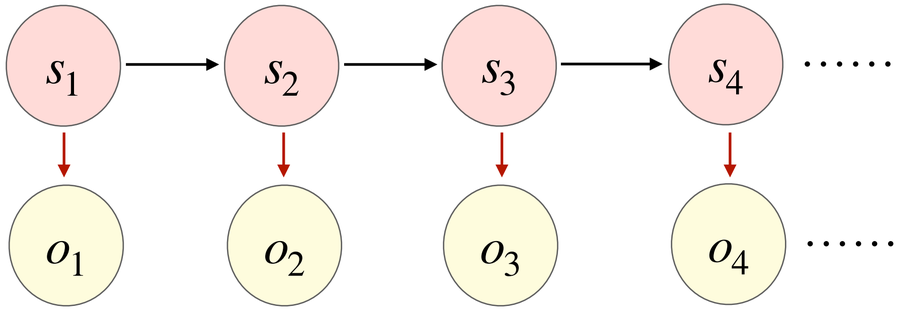
$$P(s_i \mid s_1, \ldots, s_{i-1}, O) = \frac{\exp(w \cdot f(s_1, \ldots, s_{i-1}, s_i, O))}{\sum_{s'} \exp(w \cdot f(s_1, \ldots, s_{i-1}, s', O))}$$

- **Given:** pairs of (S, O) where each $S = \langle s_1, s_2, \ldots, s_n \rangle$

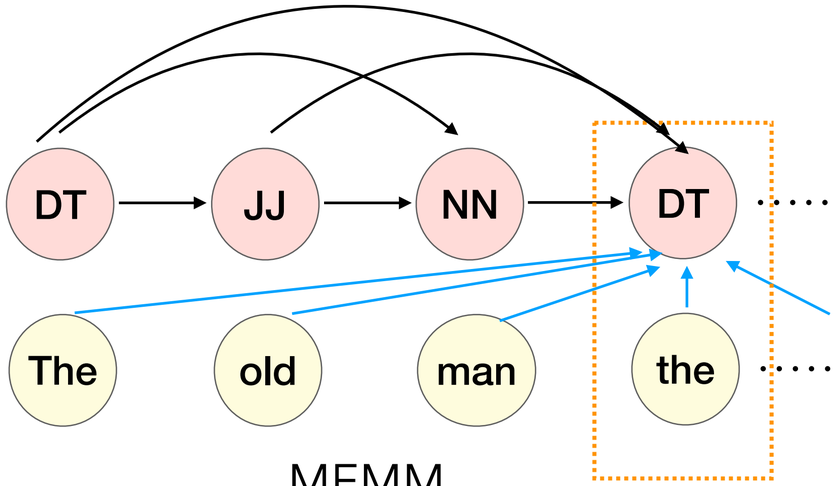
$$\text{Loss for one sequence, } L = -\sum_i \log P(s_i \mid s_1, \ldots, s_{i-1}, O)$$

- Compute gradients with respect to weights w and update

Label bias



HMM



MEMM

The/? old/? man/? the/? boat/?

$$P(JJ \mid DT)\, P(\text{old} \mid JJ)\, P(NN \mid JJ)\, P(\text{man} \mid NN)\, P(DT \mid NN)$$

$$P(NN \mid DT)\, P(\text{old} \mid NN)\, P(VB \mid NN)\, P(\text{man} \mid VB)\, P(DT \mid VB)$$

Low entropy transitions between labels may override the effect of observations

Stanford Parser

Please enter a sentence to be parsed:

The old man the boat

Language: English Sample Sentence Parse

Your query

The old man the boat

Tagging

```
The/DT  old/JJ  man/NN  the/DT  boat/NN
```

Solution?
Conditional Random Fields
(advanced)

Expectation Maximization

Expectation Maximization

- Unsupervised learning method
- Can train a model without any labeled data
- Treat unknowns as “hidden” or “latent” variables
- Maximize (expected) likelihood of observed data

EM: Some intuition

- Let's say I have 3 coins in my pocket,
  - Coin 0 has probability λ 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:
  $\langle H, H, H \rangle, \langle T, T, T \rangle, \langle H, H, H \rangle, \langle T, T, T \rangle, \langle H, H, H \rangle$

What would you estimate as the values for λ, p_1, p_2 ?

A) 1/2, 1, 0

B) 3/5, 1, 0

C) 1/2, 1/2, 1/2

Maximum Likelihood Estimate

- Data points $x_1, x_2, \ldots, x_n$ from (finite or countable) set $\mathcal{X}$
- Parameter vector θ
- Parameter space Ω, i.e. $\theta \in \Omega$
- We have a distribution $P(x \mid \theta)$ for any $\theta \in \Omega$, such that

$$\sum_{x \in \mathcal{X}} P(x \mid \theta) = 1 \text{ and } P(x \mid \theta) \geq 0 \quad \forall x$$

- Assume data points are drawn independently and identically distributed from a distribution $P(x \mid \theta^*)$ for some $\theta^* \in \Omega$
  - This θ^* is the MLE

Log Likelihood

- Data points $x_1, x_2, \ldots, x_n$ from (finite or countable) set $\mathcal{X}$
- Parameter vector θ and a parameter space Ω
- Probability distribution $P(x \mid \theta)$ for any $\theta \in \Omega$
- Data Likelihood$(\theta) = P(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$ (Each x_i is a data point)
- Log-likelihood, $L(\theta) = \sum_{i=1}^{n} \log P(x_i \mid \theta)$

Example 1: Coin Tossing

- $\mathcal{X} = \{H, T\}$. Our data points $x_1, x_2, \ldots, x_n$ are a sequence of heads and tails, e.g.
  - HTHTHHHHTTT
- Parameter vector θ is a single parameter, i.e probability of coin coming up heads
- Parameter space $\Omega = [0,1]$
- Distribution $P(x \mid \theta) = \begin{cases} \theta \text{ if } x = H \\ 1 - \theta \text{ if } x = T \end{cases}$

What distribution is this?

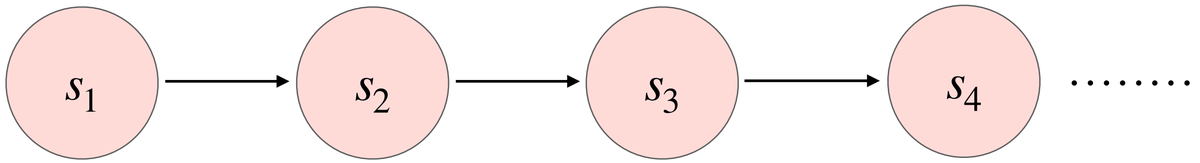
A) Binomial

B) Bernoulli

C) Multinomial

D) Gaussian

Example 2: Markov chains



- $\mathcal{X}$ is the set of all possible state (e.g tag) sequences created by the underlying generative process. Our sample is n sequences $X_1, \ldots, X_n$ such that each $X_i \in \mathcal{X}$, consists of a sequence of states $s_1, s_2, s_3, \ldots$
- θ_T is the vector of all transition $(s_i \rightarrow s_j)$ parameters. Without loss of generality, assume a dummy start state ϕ and initial transition $\phi \rightarrow s_1$ (how many parameters?)
- Let $T(\alpha) \subset T$ be all the transitions of the form $\alpha \rightarrow \beta$ (i.e. all transitions from state α)
- Then, parameter space Ω is the set of $\theta \in [0,1]^{|S+1||S|}$ where S is set of all states (tags), such that:

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

(why?)

Example 2: Markov chains

- Now, if θ_T is the vector of all transition parameters
- Then, we have for a sequence X:

$$P(X \mid \theta) = \prod_{t \in T} \theta_t^{Count(X,t)}$$

where $Count(X, t)$ is the number of times transition t is seen in sequence X

$$\implies \log P(X \mid \theta) = \sum_{t \in T} Count(X, t) \ \log \theta_t$$

MLE for Markov chains

- We have

$$\log P(X \mid \theta) = \sum_{t \in T} Count(X, t) \ \log \theta_t$$

where $Count(X, t)$ is the number of times transition t is seen in sequence X

- And,

$$L(\theta) = \sum_i \log P(X_i \mid \theta) = \sum_i \sum_{t \in T} Count(X_i, t) \ \log \theta_t$$

MLE for Markov chains

- $L(\theta) = \sum_i \log P(X_i \mid \theta) = \sum_i \sum_{t \in T} Count(X_i, t) \ \log \theta_t$
- Solve $\theta_{MLE} = \arg\max_{\theta \in \Omega} L(\theta)$

  $\Longrightarrow$ find θ s. t. $\frac{\partial L(\theta)}{\partial \theta} = 0$ with appropriate probability constraints
- This gives: $\theta_t = \frac{\sum_i Count(X_i, t)}{\sum_i \sum_{t' \in T(\alpha)} Count(X_i, t')}$

  where t is of the form $\alpha \rightarrow \beta$ for some β
- Intuitively, the denominator is simply counting all occurrences of state α at the start of a transition.

Models with hidden variables

- Now say we have two sets $\mathcal{X}$ and $\mathcal{Y}$, and a joint distribution $P(x, y \mid \theta)$
- If we had **fully observable data**, (x_i, y_i) pairs, then log likelihood can be estimated as:

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

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

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

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

(y is hidden)

Unsupervised Learning

Expectation Maximization

- 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

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

Maximization

"Expectation"

The three coins example

- In the three coins example,

  $\mathcal{Y} = \{H, T\}$ (possible outcomes of coin 0)

  $\mathcal{X} = \{HHH, TTT, HTT, THH, HHT, TTH, HTH, THT\}$ (all possible observations of length 3)

  $\theta = \{\lambda, p_1, p_2\}$

- 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

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

The three coins example

- Calculating various probabilities:

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

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

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

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

The three coins example

- Fully observed data might look like:

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

- In this case, maximum likelihood estimates are:

$$\lambda = \frac{3}{5}$$

$$p_1 = \frac{9}{9}$$

$$p_2 = \frac{0}{6}$$

The three coins example

- Partially observed data might look like:

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

- How do we find the MLE parameters? EM!

The three coins example

- Partially observed data might look like:

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

- Assume we guess the current parameters to be some λ, p_1, p_2. Then:

$$P(y = H \,|\, x = \langle HHH \rangle) = \frac{P(\langle HHH \rangle, H)}{P(\langle HHH \rangle, H) + P(\langle HHH \rangle, T)}$$

$$= \frac{\lambda p_1^3}{\lambda p_1^3 + (1 - \lambda) p_2^3}$$

$$P(y = H \,|\, x = \langle TTT \rangle) = \frac{P(\langle TTT \rangle, H)}{P(\langle TTT \rangle, H) + P(\langle TTT \rangle, T)}$$

$$= \frac{\lambda (1 - p_1)^3}{\lambda (1 - p_1)^3 + (1 - \lambda)(1 - p_2)^3}$$

The three coins example

- If the current parameters are λ, p_1, p_2

$$P(y = H \mid x = \langle HHH \rangle) = \frac{P(\langle HHH \rangle, H)}{P(\langle HHH \rangle, H) + P(\langle HHH \rangle, T)}$$

$$= \frac{\lambda p_1^3}{\lambda p_1^3 + (1 - \lambda) p_2^3}$$

$$P(y = H \mid x = \langle TTT \rangle) = \frac{P(\langle HHH \rangle, H)}{P(\langle TTT \rangle, H) + P(\langle TTT \rangle, T)}$$

$$= \frac{\lambda (1 - p_1)^3}{\lambda (1 - p_1)^3 + (1 - \lambda)(1 - p_2)^3}$$

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

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

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

The three coins example

- After filling in hidden variables for each example, partially observed data might look like:

$$
\begin{array}{ll}
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492
\end{array}
$$

each pair sums to 1

Treat this as a pseudo-annotated dataset (with appropriate weights) and use MLE!

The three coins example

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

- New estimates:

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

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

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

The three coins example

$$
\begin{array}{ll}
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492
\end{array}
$$

- New estimates:

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

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

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

The three coins example

$$
\begin{array}{ll}
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492 \\
(\langle \texttt{TTT} \rangle, H) & P(y = \texttt{H} \mid \texttt{TTT}) = 0.6967 \\
(\langle \texttt{TTT} \rangle, T) & P(y = \texttt{T} \mid \texttt{TTT}) = 0.3033 \\
(\langle \texttt{HHH} \rangle, H) & P(y = \texttt{H} \mid \texttt{HHH}) = 0.0508 \\
(\langle \texttt{HHH} \rangle, T) & P(y = \texttt{T} \mid \texttt{HHH}) = 0.9492
\end{array}
$$

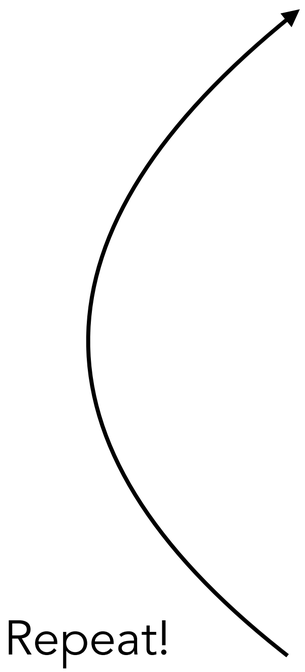
- New estimates:

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

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

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

Summary



- Begin with parameters: $\lambda = 0.3,\ p_1 = 0.3,\ p_2 = 0.6$
- Fill in hidden variables, using

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

  $P(y = H \mid x = \langle TTT \rangle) = 0.6967$
- This gives us a pseudo-annotated dataset with **fractional** counts
- Re-estimate parameters to be

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

EM iterations

Iteration	λ	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

The coin example for $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. The solution that EM reaches 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 ($\lambda = 0.5$).

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

EM iterations

Iteration	λ	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 $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$

λ is now 0.4, indicating that coin 0 has a probability 0.4 of selecting the tail-biased coin

EM iterations

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

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

EM selects a tails-only coin, and a coin which is heavily heads-biased ($p_2 = 0.8284$). It's certain that x_1 and x_3 were generated by coin 2 since they contain heads. x_2 and x_4 could have been generated by either coin but coin 1 (tail-biased) is far more likely.

Initialization matters

Iteration	λ	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 at a **saddle point**.

Iteration	λ	p_1	p_2	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$	$\bar{p}_4$
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 $x = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$.

If we initialize p_1 and p_2 even a small amount away from the saddle point $p_1 = p_2$, EM diverges and eventually reaches the global maximum