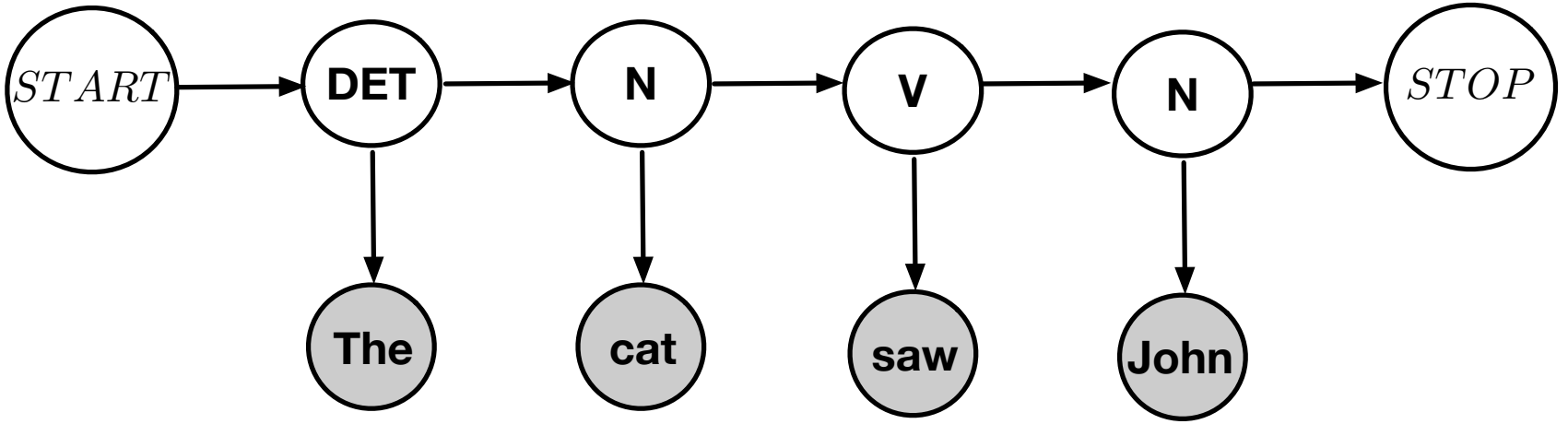
Special Topics in Natural Language Processing

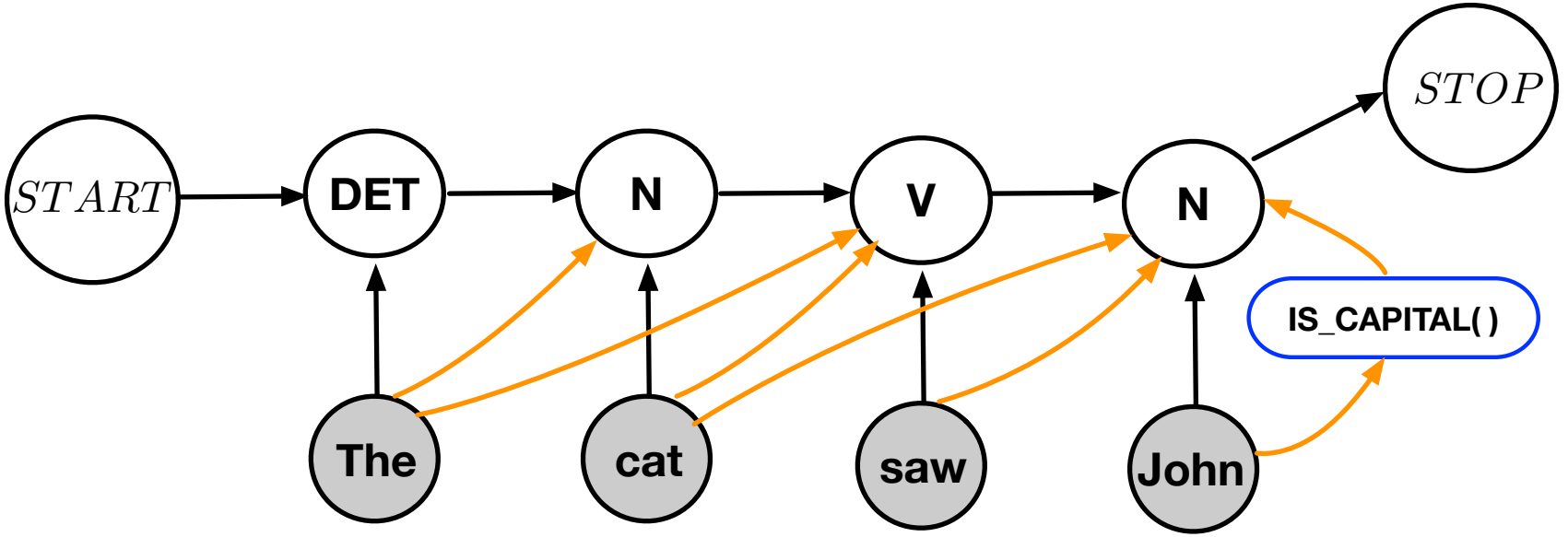
CS698O

Ashutosh Modi
CSE Department, IIT Kanpur

Lecture 14: Sequence Prediction 5
Feb 5, 2020

HMM and MEMM


START
DET
N
V
N
STOP
The
cat
saw
John


START
DET
N
V
N
STOP
IS_CAPITAL( )
The
cat
saw
John

Maximum Entropy Markov Model (MEMM)

- Allow highly flexible representations, allowing features to be easily integrated into the model.
- Also called as ***Log-Linear Tagging Model***
- Discriminative Model

HMM

$$\hat{Y}_{1:N} = \underset{y_{1:N} \in \Lambda}{\operatorname{argmax}} P(X_{1:N}, Y_{1:N} = y_{1:N})$$

$$= \underset{y_{1:N} \in \Lambda}{\operatorname{argmax}} \prod_{i=1}^{N+1} P(Y_i \mid Y_{i-1}) \times P(X_i \mid Y_i)$$

MEMM

$$\hat{Y}_{1:N} = \underset{y_{1:N} \in \Lambda}{\operatorname{argmax}} P(Y_{1:N} = y_{1:N} \mid X_{1:N})$$

$$= \underset{y_{1:N} \in \Lambda}{\operatorname{argmax}} \prod_{i=1}^{N+1} P(Y_i \mid Y_{i-1}, X_{1:N})$$

Maximum Entropy Markov Model (MEMM)

$$P(Y_{1:N} = y_{1:N} \mid X_{1:N} = x_{1:N}) = \prod_{i=1}^{N} P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N})$$

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N}) = \frac{exp(\theta^T f(h_i, y_i))}{\sum_{y' \in K} exp(\theta^T f(h_i, y'))}$$

$$h_i = < y_{i-1}, x_1 \ldots x_N, i >$$

DECODING IN MEMM

DECODING IN MEMM

$$\underset{y_1,\ldots,y_N}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i \mid h_i; \theta)$$

DECODING IN HMM

$$\underset{y_1,\ldots,y_N}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i \mid y_{i-1}) \times p(x_i \mid y_i)$$

DECODING IN MEMM: VITERBI

$$\underset{y_1,\ldots,y_N}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i \mid h_i; \theta)$$

$$\text{viterbi}(k, c_k) = \max_{y_1,\ldots,y_k} \prod_{i=1}^{k} p(y_i \mid h_i; \theta)$$

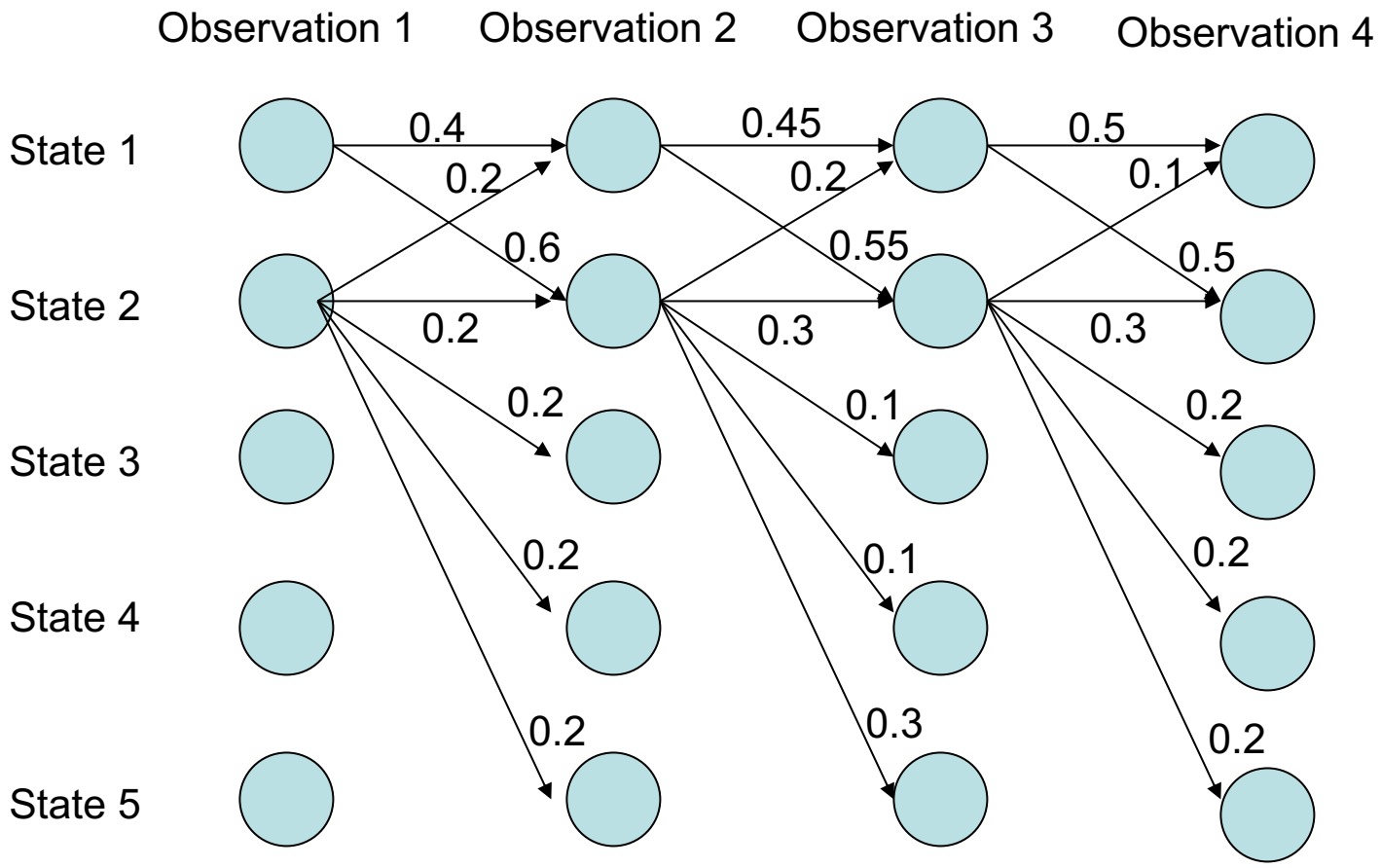
$$\text{viterbi}(k, c_k) = \max_{c_l} \text{viterbi}(k-1, c_l) \times p(Y_k = c_k \mid h_k; \theta)$$

HMM:

$$\text{viterbi}(i, c_k) = \left(\max_{c_l} \text{viterbi}(i-1, c_l) \times P\left(Y_i = c_k \mid Y_{i-1} = c_l\right) \right) \times P\left(X_i \mid Y_i = c_k\right)$$

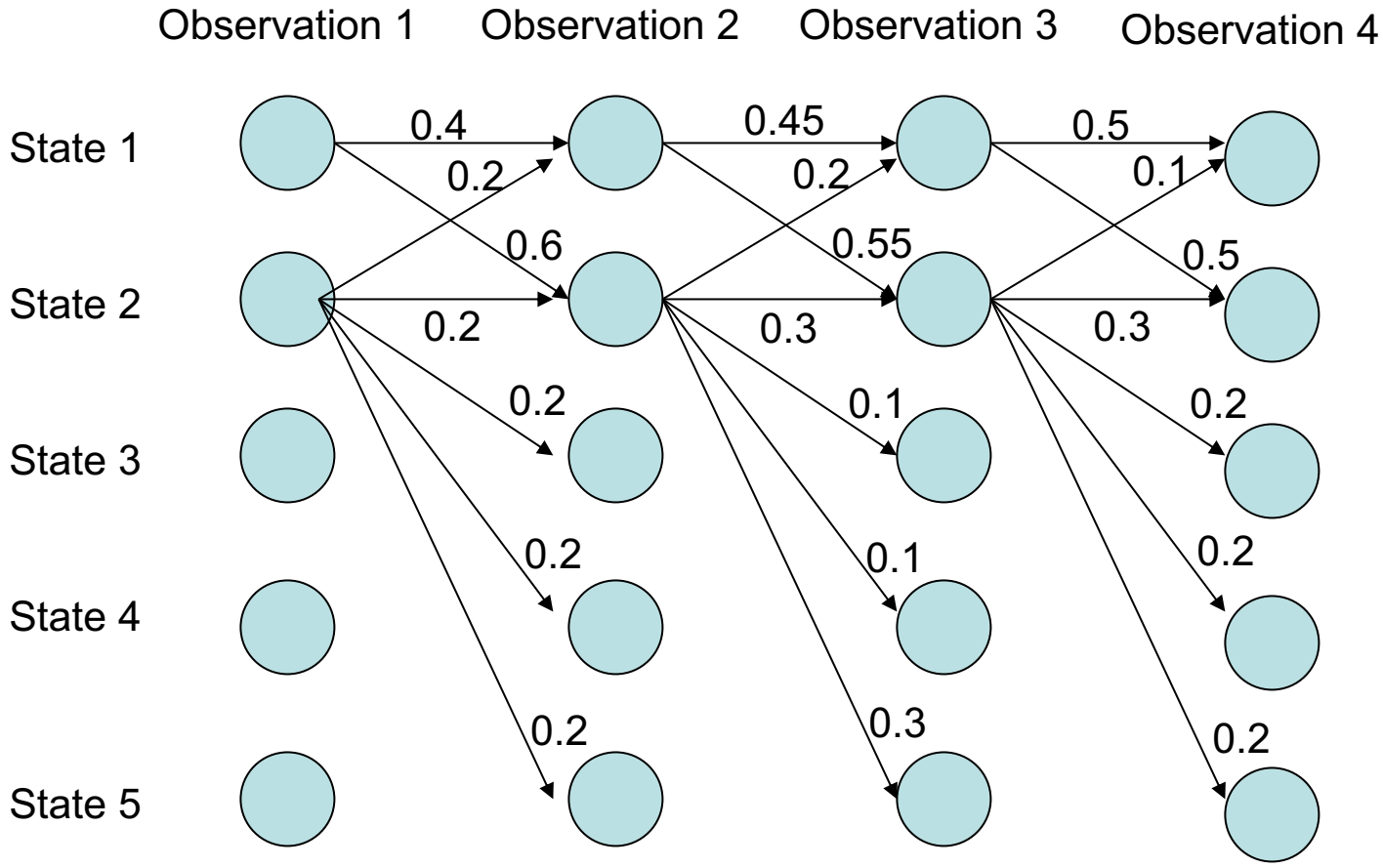
MEMMs Limitations: Label Bias Problem

MEMMs Limitations: Label Bias Problem



Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

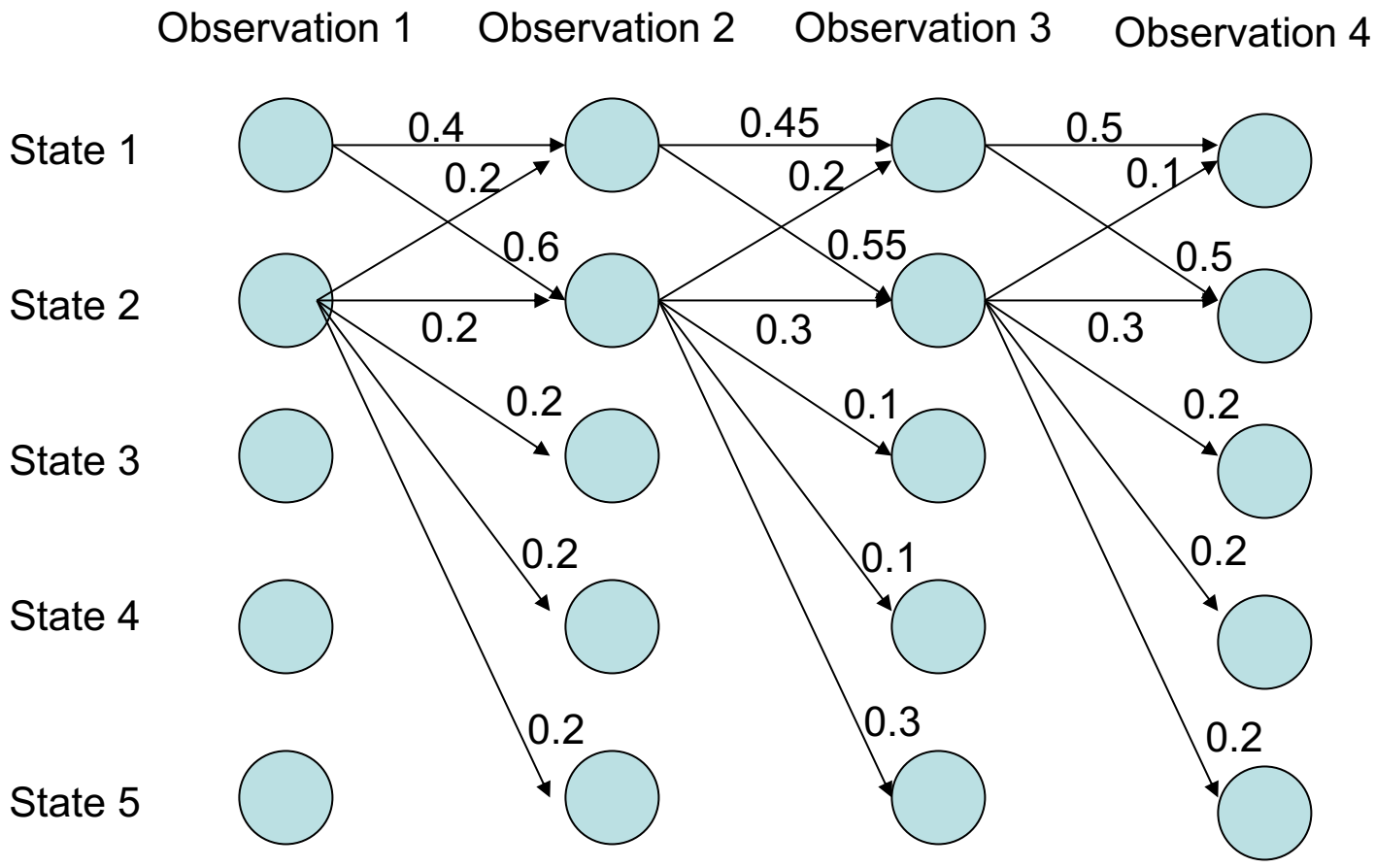
MEMMs Limitations: Label Bias Problem



What the local transition probabilities say:

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem

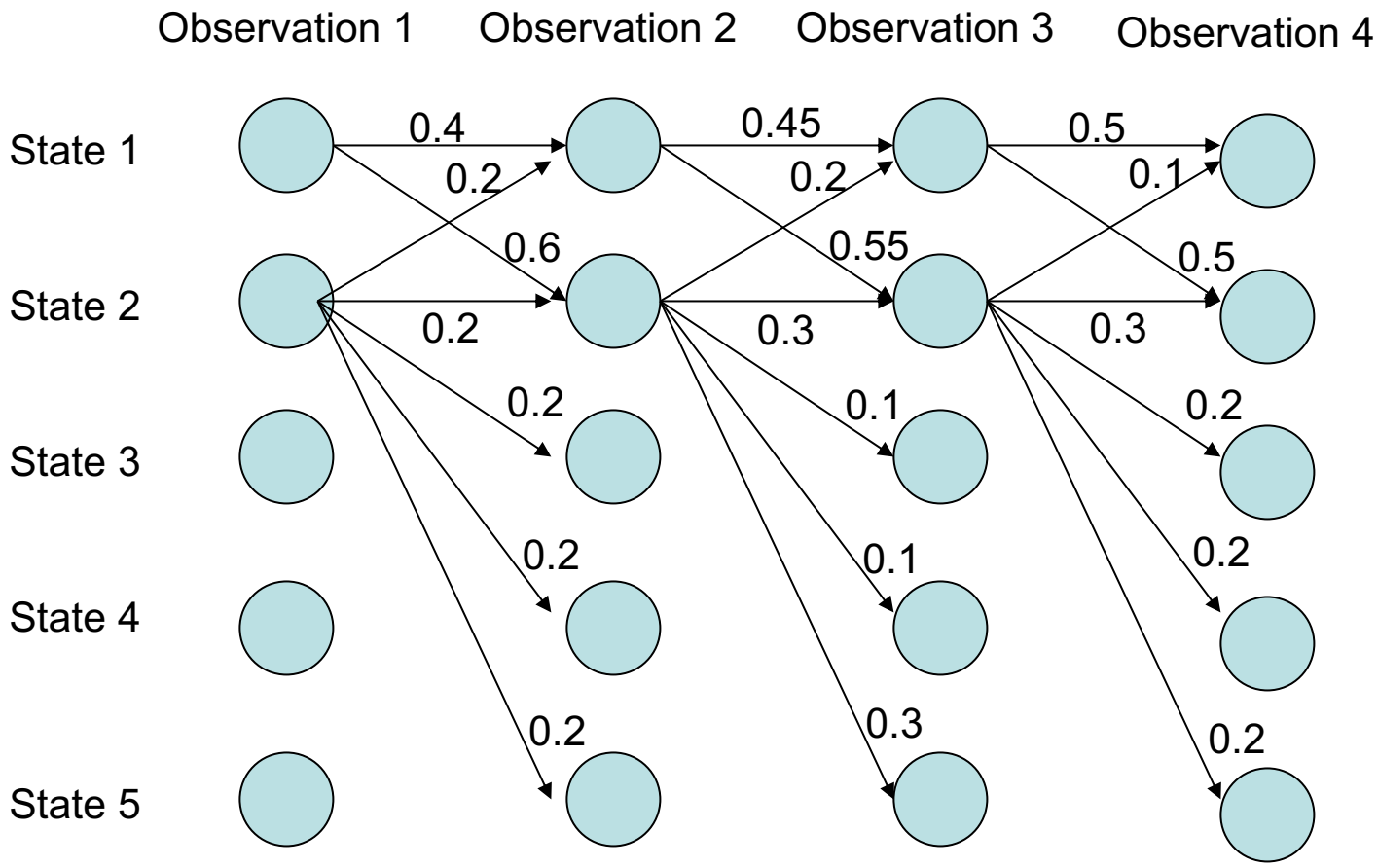


What the local transition probabilities say:

- State 1 almost always prefers to go to state 2
- State 2 almost always prefer to stay in state 2

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

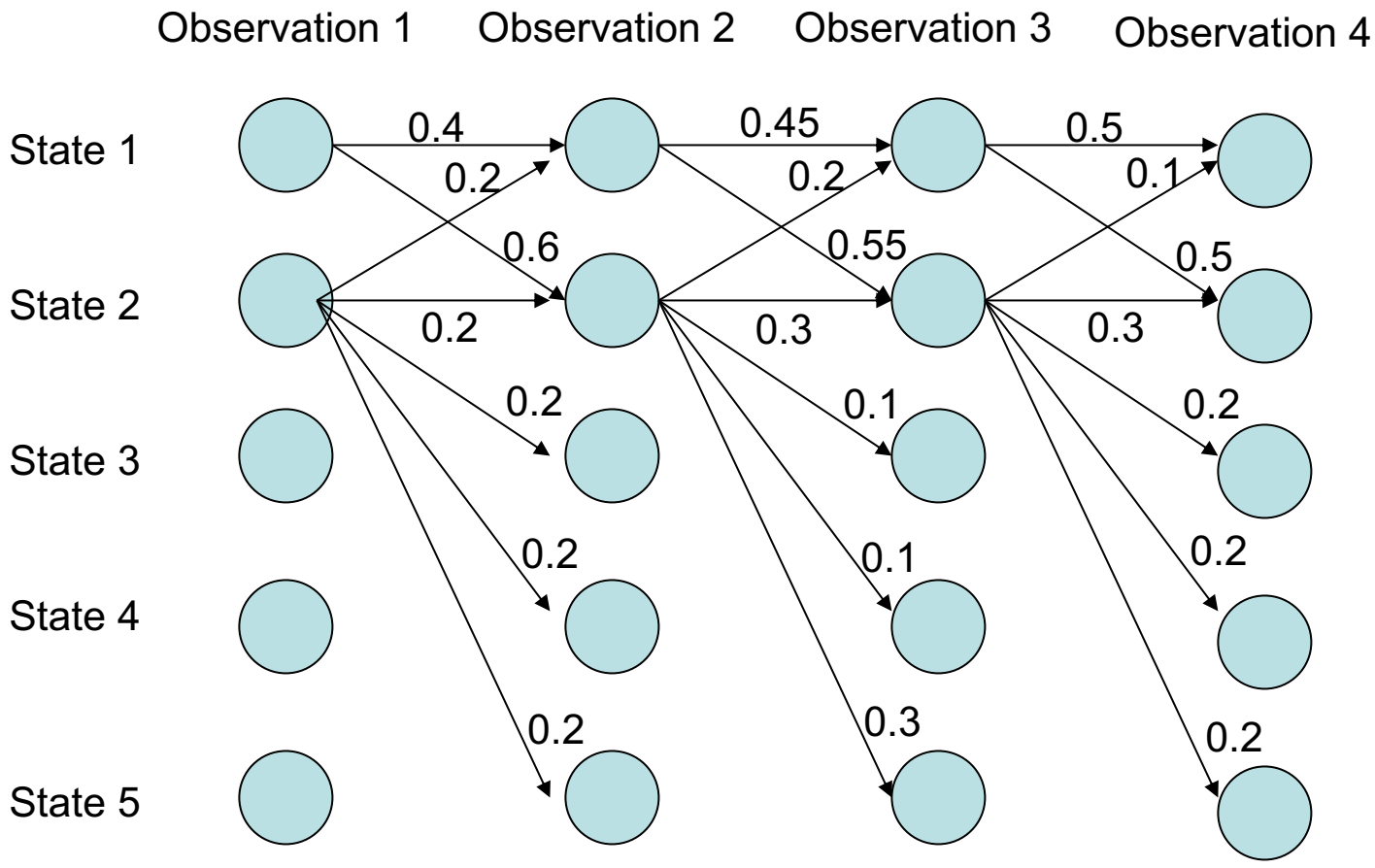
MEMMs Limitations: Label Bias Problem



Probability of path 1-> 1-> 1-> 1

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem

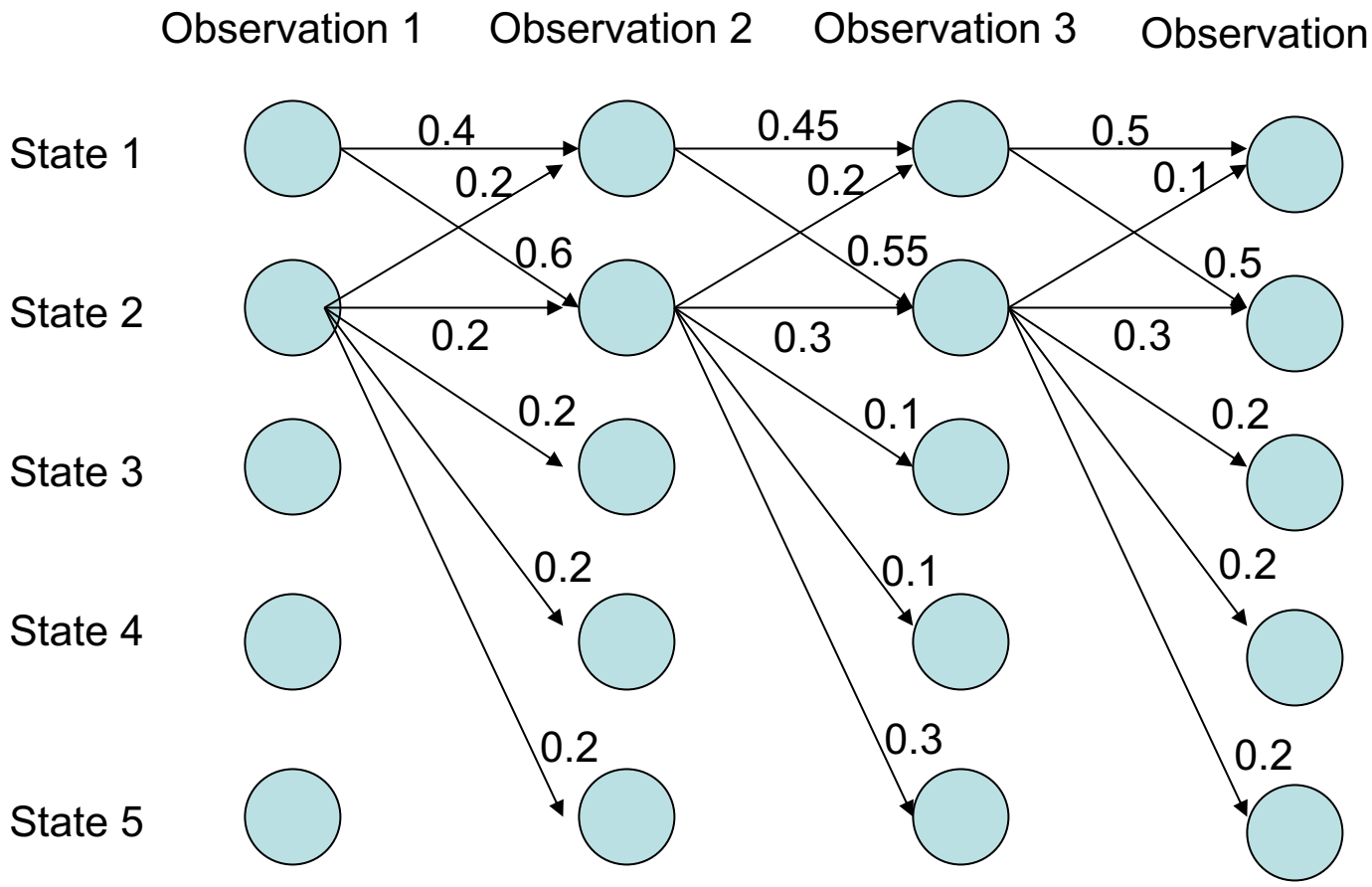


Probability of path 1-> 1-> 1-> 1

0.4 x 0.45 x 0.5 = 0.09

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem

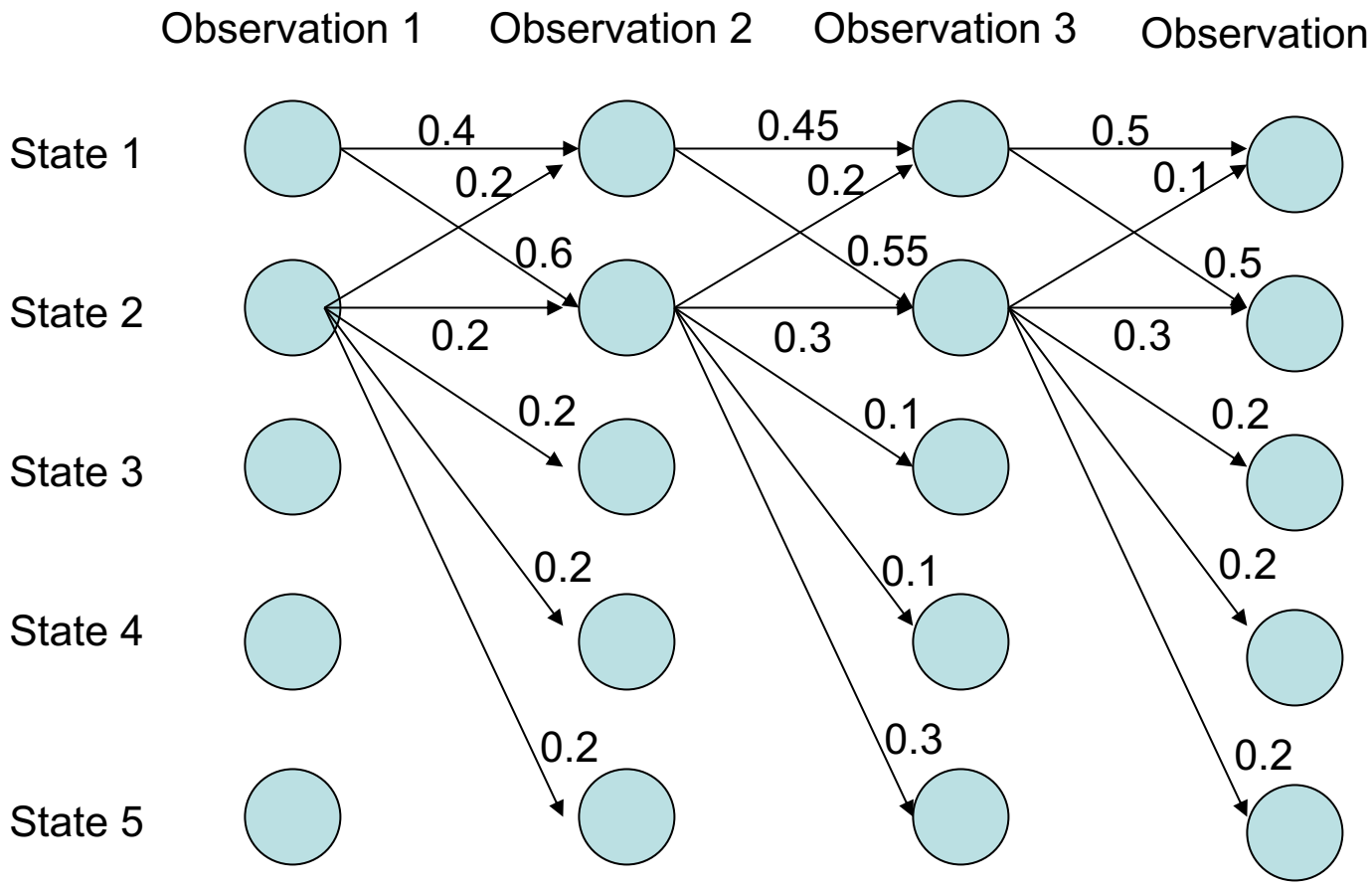


Probability of path 2->2->2->2

1-> 1-> 1-> 1: 0.09

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



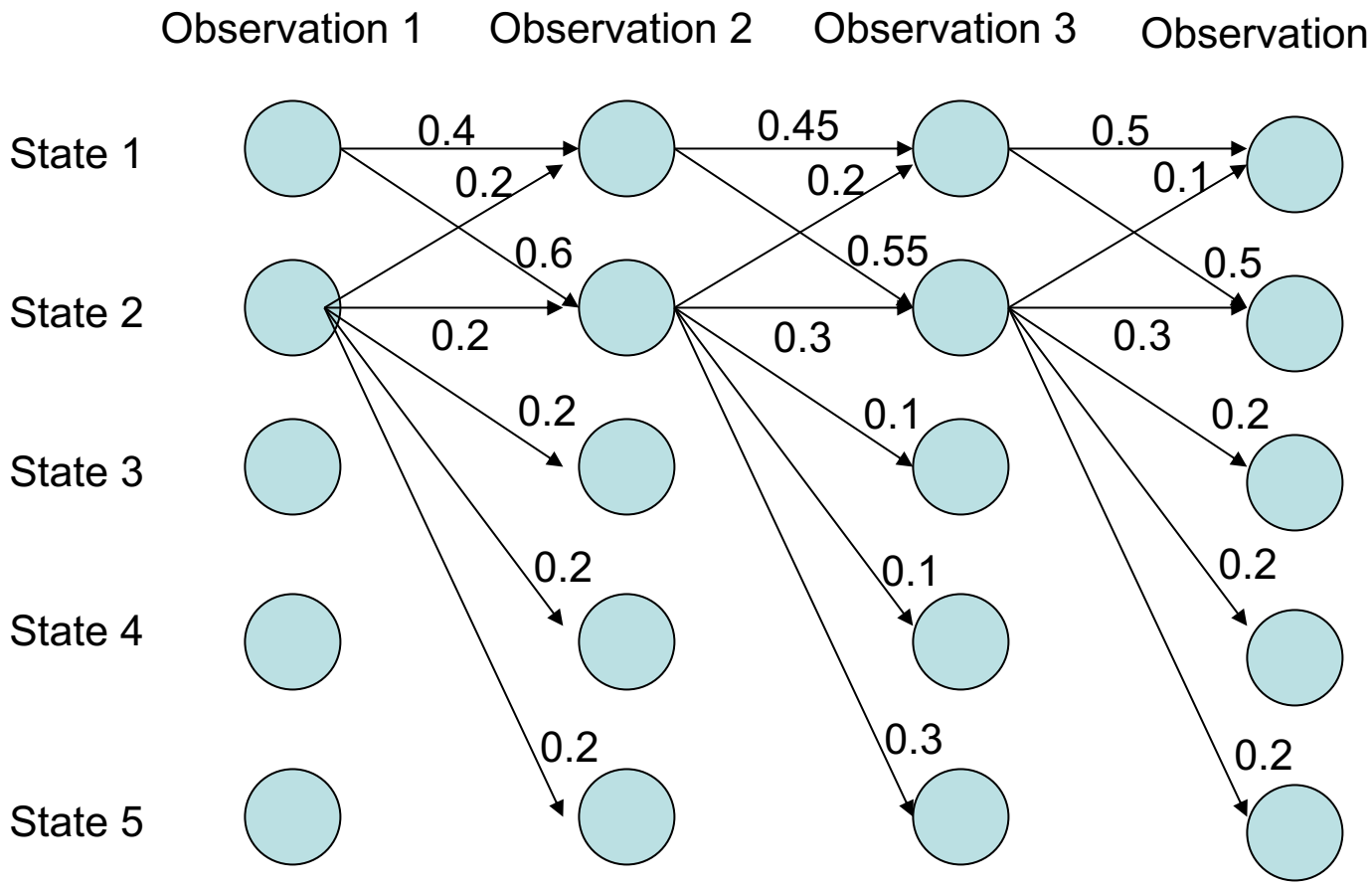
Probability of path 2->2->2->2 :

0.2 X 0.3 X 0.3 = 0.018

1-> 1-> 1-> 1: 0.09

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



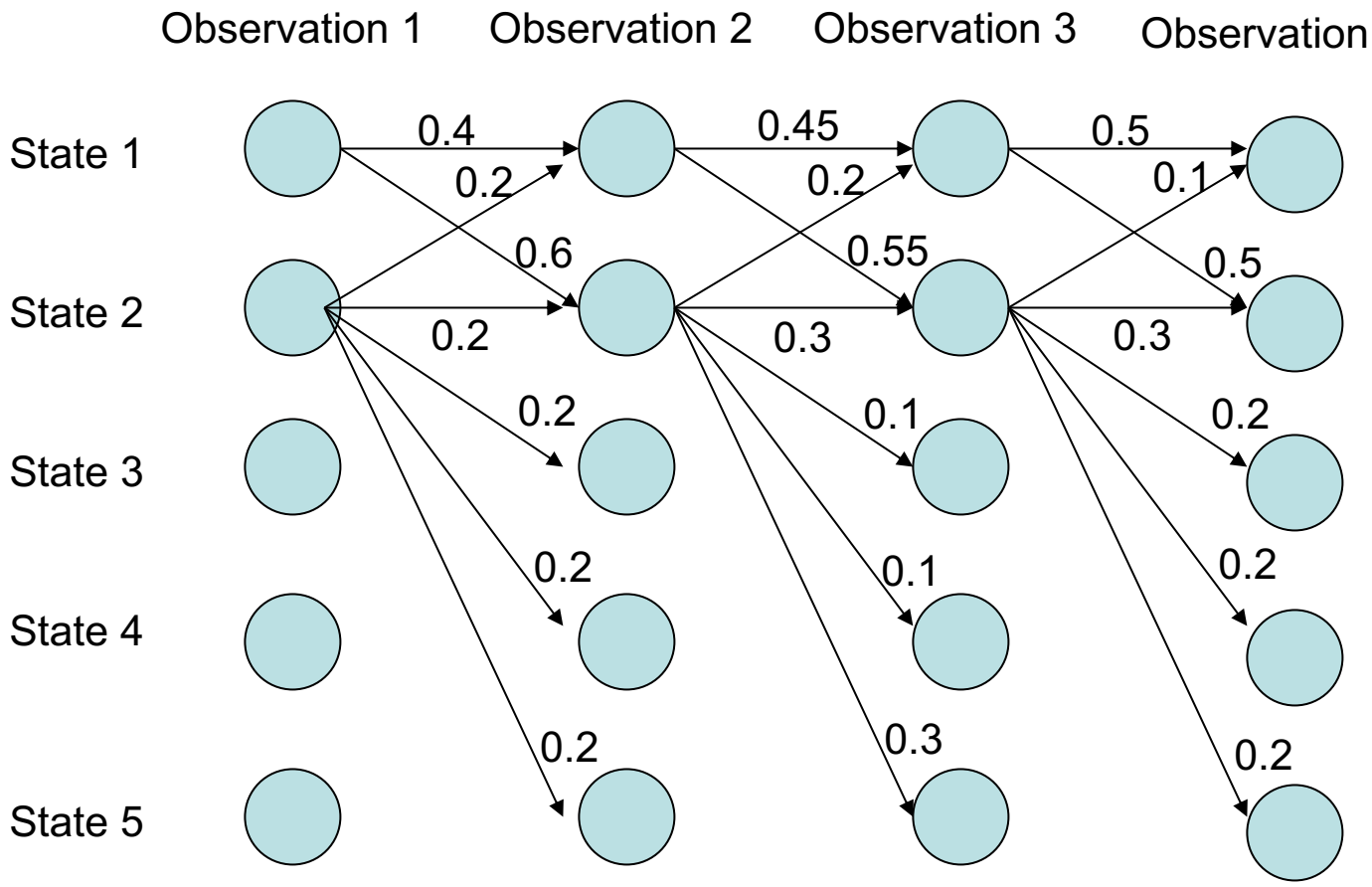
Probability of path 1->2->1->2 :

1-> 1-> 1-> 1: 0.09

2->2->2->2 : 0.018

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



Probability of path 1->2->1->2 :

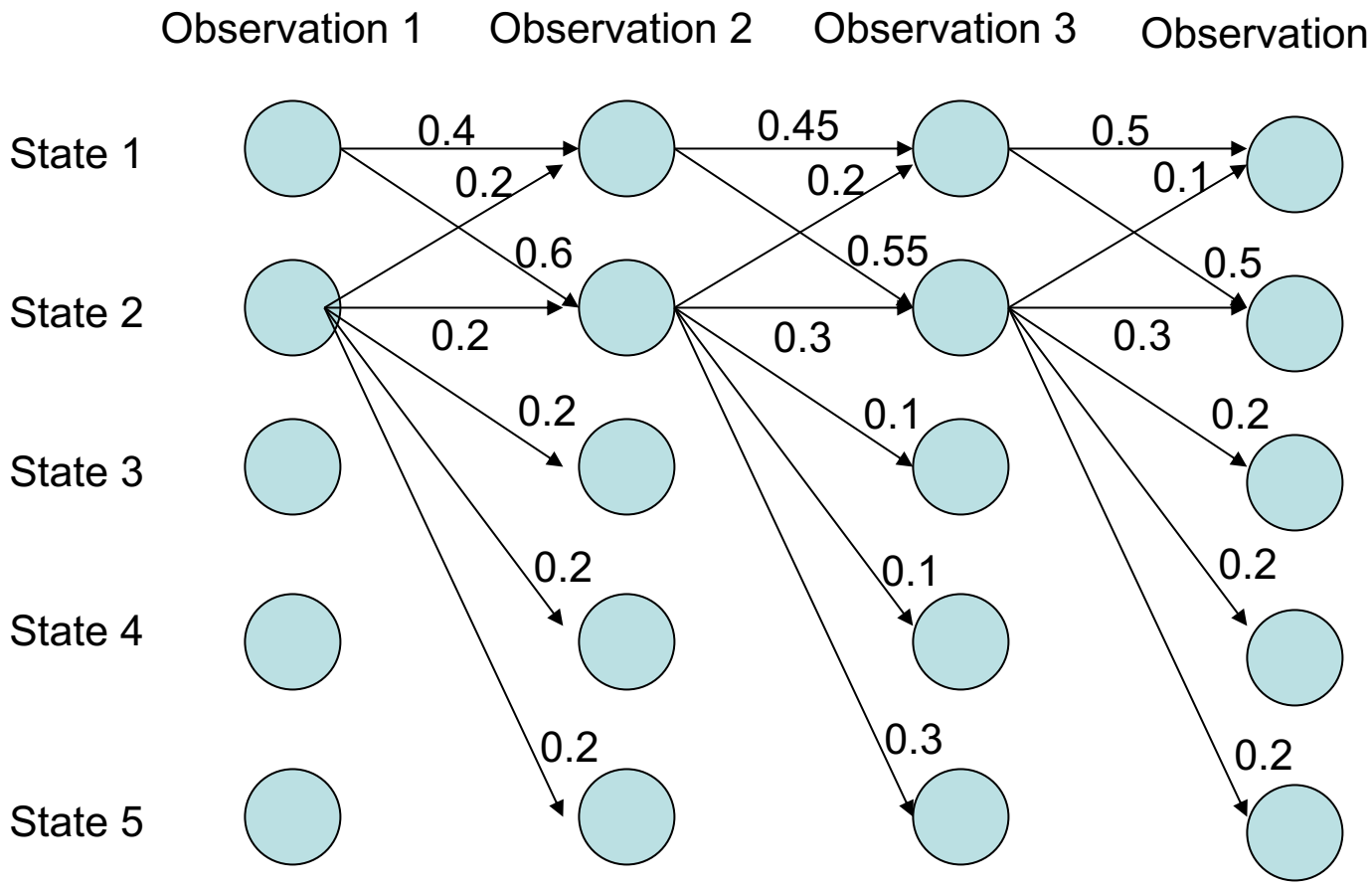
0.6 X 0.2 X 0.5 = 0.06

1-> 1-> 1-> 1: 0.09

2->2->2->2 : 0.018

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



Probability of path 1->1->2->2 :

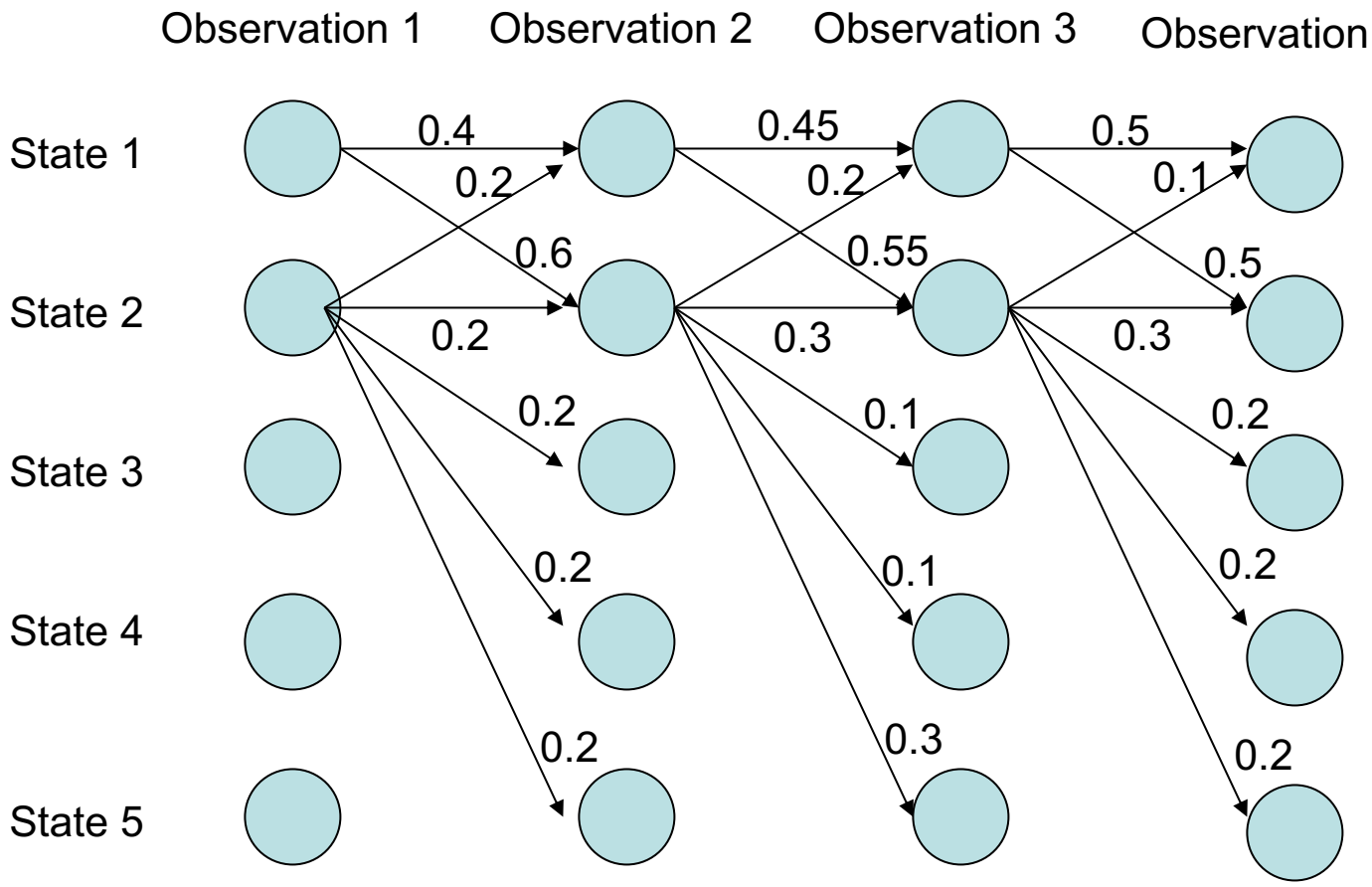
1-> 1-> 1-> 1: 0.09

2->2->2->2 : 0.018

1->2->1->2 : 0.06

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



Probability of path 1->1->2->2 :

0.4 X 0.55 X 0.3 = 0.066

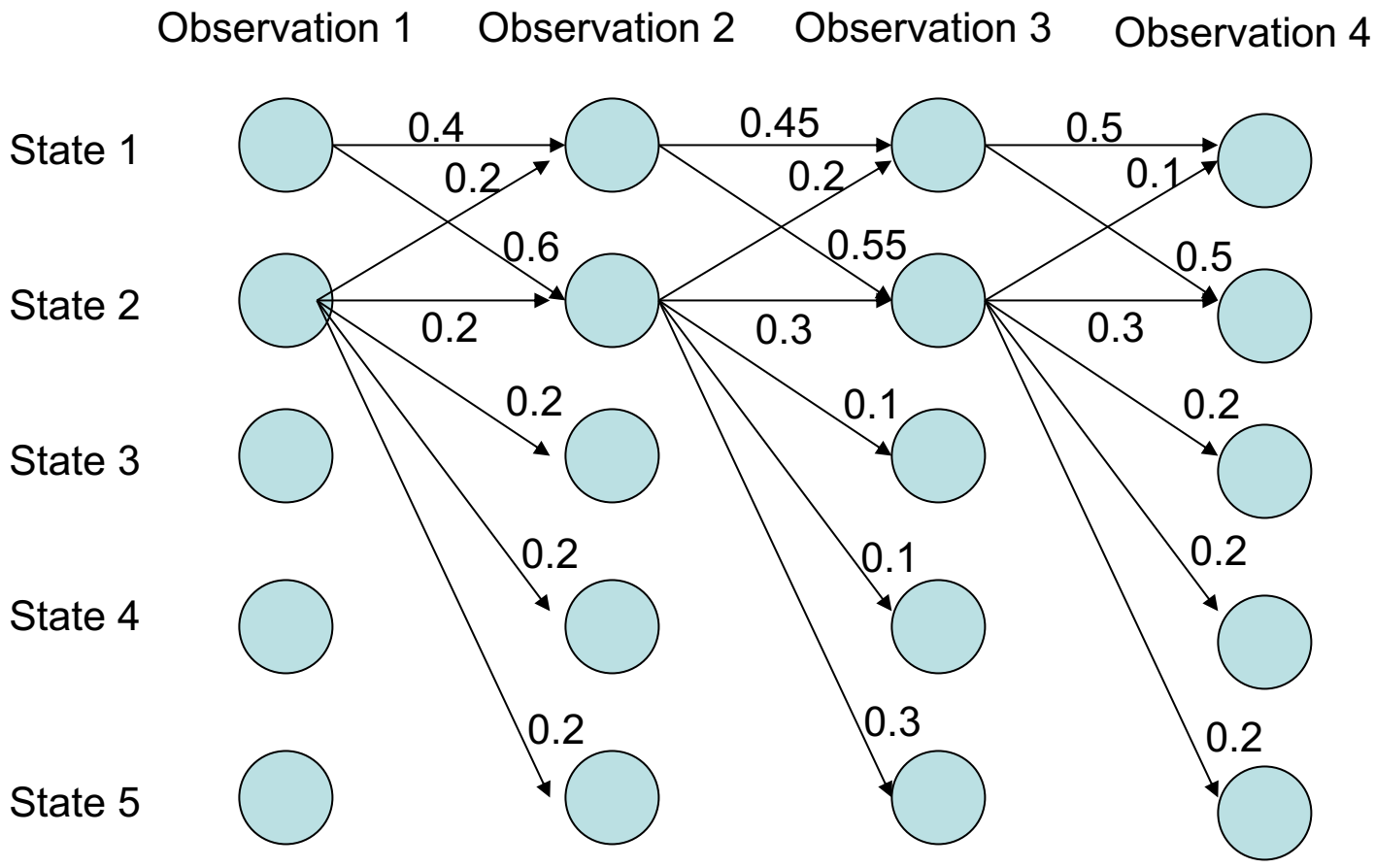
1-> 1-> 1-> 1: 0.09

2->2->2->2 : 0.018

1->2->1->2 : 0.06

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



1-> 1-> 1-> 1: 0.09

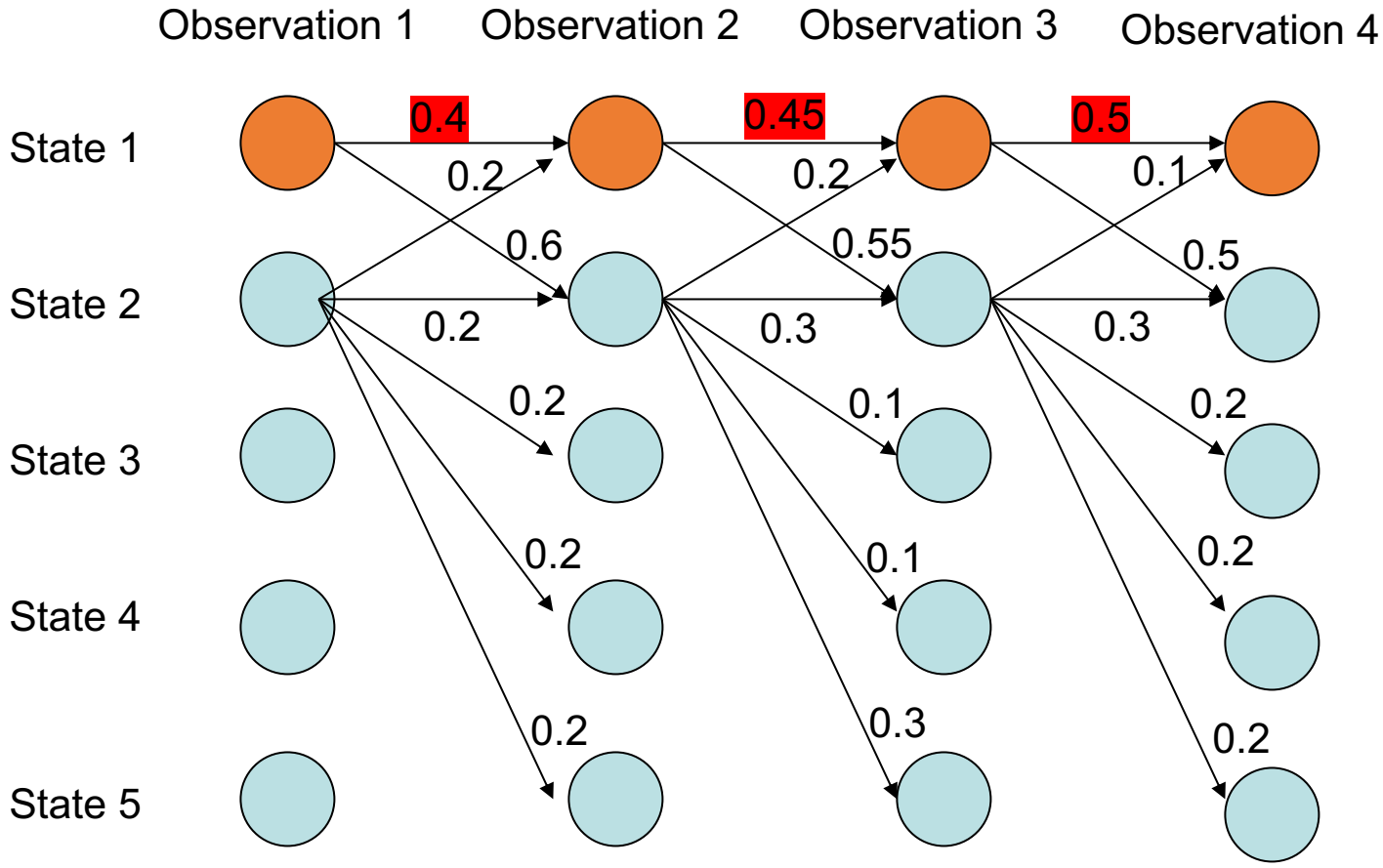
2->2->2->2 : 0.018

1->2->1->2 : 0.06

1->1->2->2 : 0.066

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem

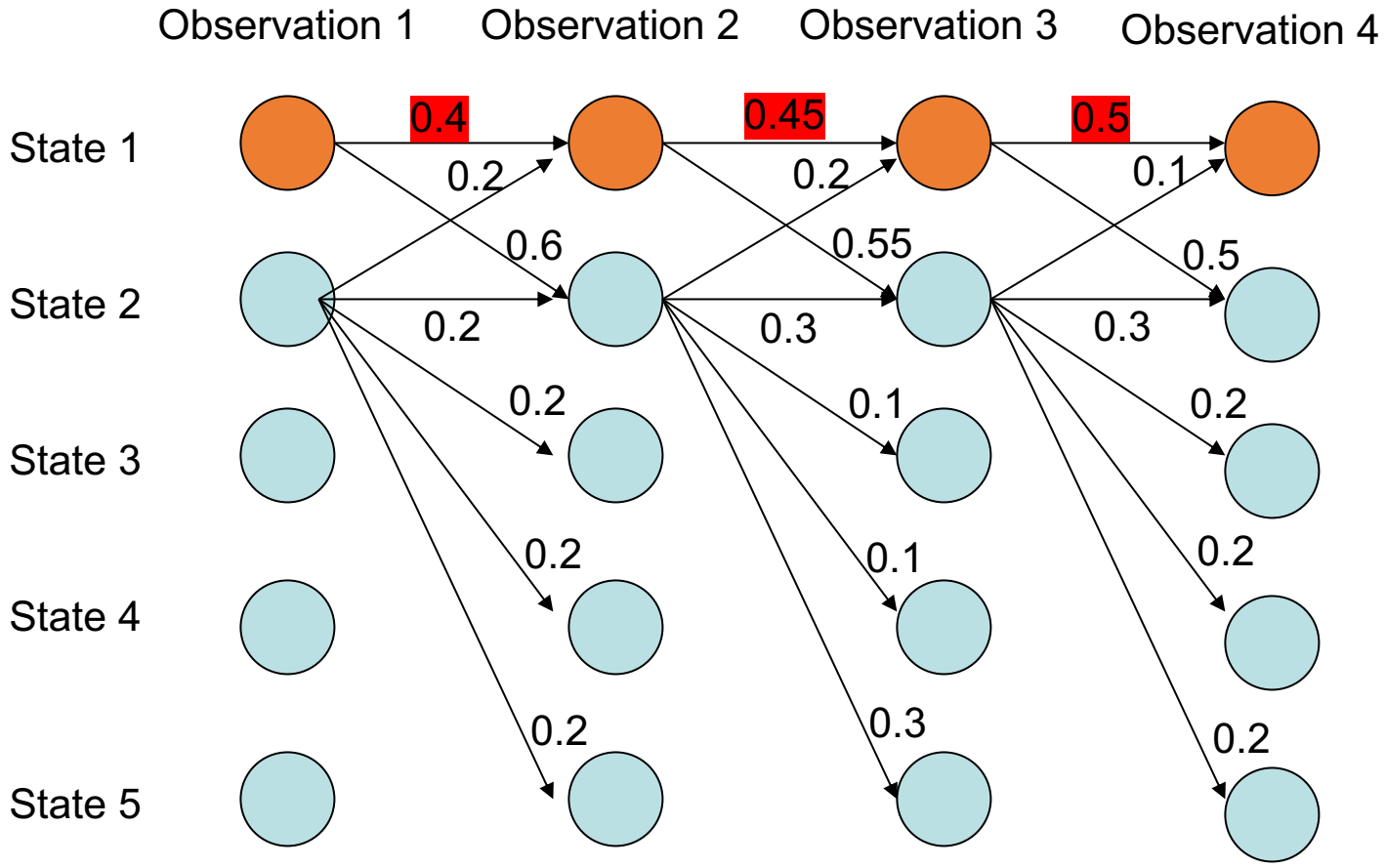


Most Likely path:

1-> 1-> 1-> 1

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



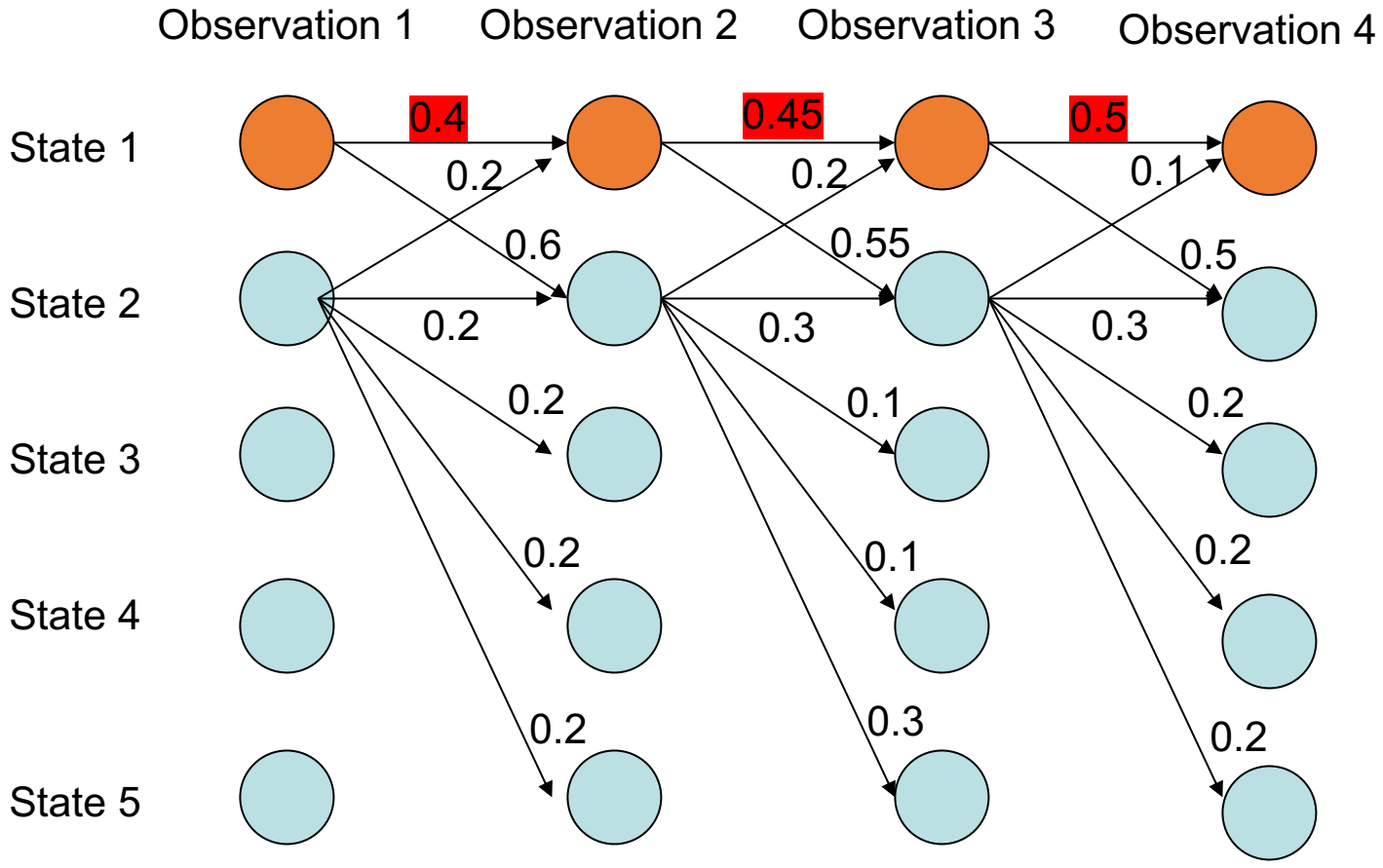
Most Likely path:

1-> 1-> 1-> 1

Although locally it seems state 1 wants to go to state 2 and state 2 wants to remain in state 2.

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



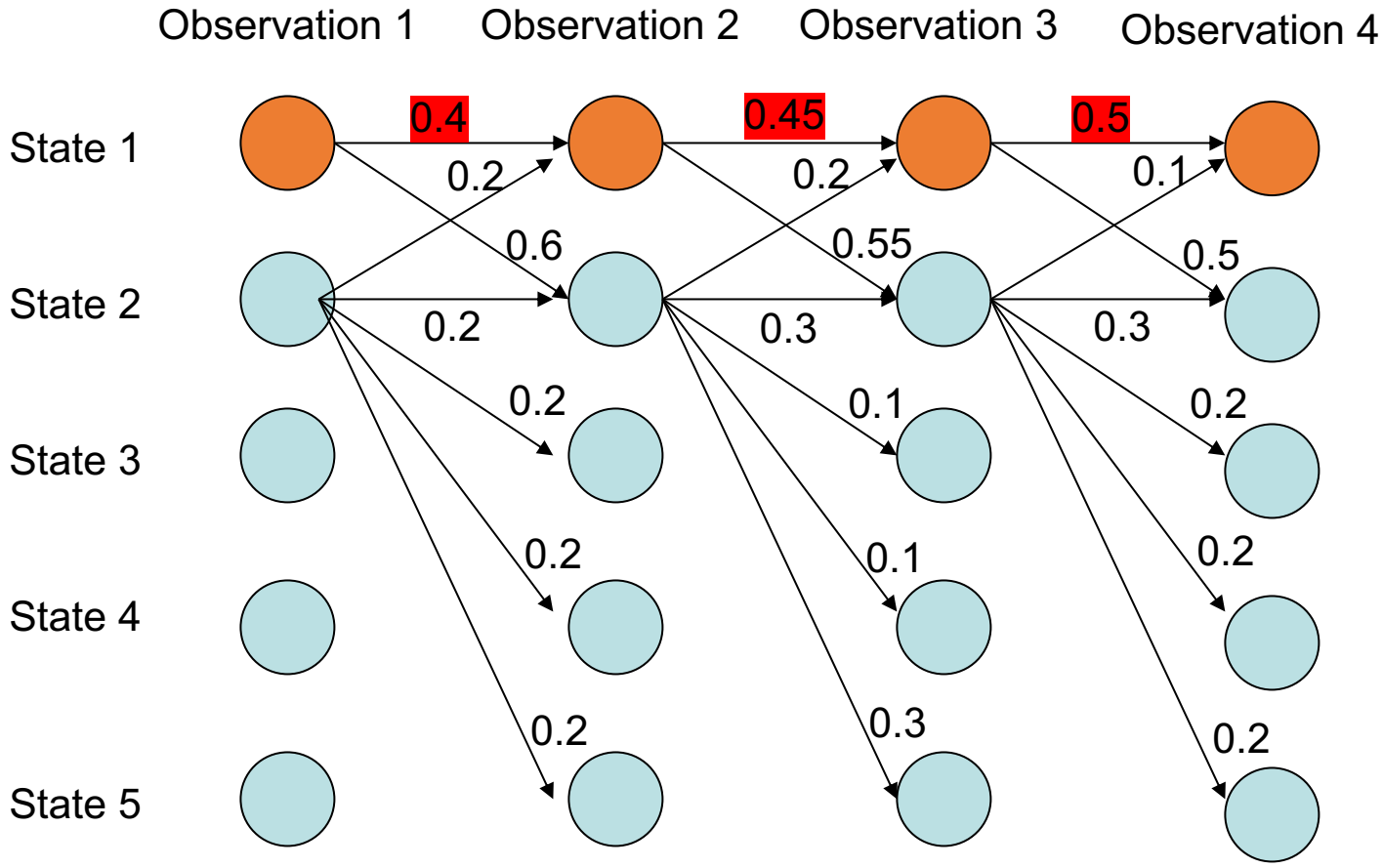
Most Likely path:

1-> 1-> 1-> 1

- State 1 has only two transitions but state 2 has 5
- Average transition probability from state 2 is lower

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem



Label bias problem in MEMM:

Preference of states with lower number of transitions over others

Based on slides by Ramesh Nallapati: https://slideplayer.com/slide/5078532/

MEMMs Limitations: Label Bias Problem

- Transition scores are the conditional probabilities of the possible next states given the current state and the observation sequence.
- Transitions leaving a given state only compete only against each other rather than against all transitions in the model.

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Lafferty, et al., ICML 2001

MEMMs Limitations: Label Bias Problem

- Transition scores are the conditional probabilities of the possible next states given the current state and the observation sequence.
- Transitions leaving a given state only compete only against each other rather than against all transitions in the model.
- *Conservation of Score Mass*: All the probability mass that arrives at a state must be distributed among the possible successor states.
- An observation can effect which destination states get the mass but not how much total mass to pass on.

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Lafferty, et al., ICML 2001

MEMMs Limitations: Label Bias Problem

- Transition scores are the conditional probabilities of the possible next states given the current state and the observation sequence.
- Transitions leaving a given state only compete only against each other rather than against all transitions in the model.
- *Conservation of Score Mass*: All the probability mass that arrives at a state must be distributed among the possible successor states.
- An observation can effect which destination states get the mass but not how much total mass to pass on.
- This causes a bias toward states with fewer outgoing transitions. In the extreme case, a state with a single outgoing transition effectively ignores the observation.
- The Markovian assumptions in MEMMs and similar state-conditional models insulate decisions at one state from future decisions in a way that does not match the actual dependencies between consecutive states.

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Lafferty, et al., ICML 2001

MEMMs Limitations: Label Bias Problem

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N}) = \frac{exp(\theta^T f(h_i, y_i))}{\sum_{y' \in K} exp(\theta^T f(h_i, y'))}$$

- MEMMs are locally normalized

MEMMs Limitations: Label Bias Problem

$$P(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N}) = \frac{exp(\theta^T f(h_i, y_i))}{\sum_{y' \in K} exp(\theta^T f(h_i, y'))}$$

- MEMMs are locally normalized
- Do not do local normalization but global normalization

Label Bias Problem

- Generative Models do not suffer from label bias problem
  - HMM do not have the label bias problem

Label Bias Problem

- Generative Models do not suffer from label bias problem
  - HMM do not have the label bias problem

- Discriminative Models suffer from label bias problem
  - Neural Sequence to Sequence models have the label bias problem.

Label Bias Problem

- Generative Models do not suffer from label bias problem
  - HMM do not have the label bias problem

- Discriminative Models suffer from label bias problem
  - Neural Sequence to Sequence models have the label bias problem.

- The first recorded observation of the label bias problem was in Léon Bottou's PhD thesis (1991) in context of neural network for speech recognition.
- The term "label bias" was coined in the seminal work of Lafferty, McCallum and Pereira introducing conditional random fields.

From MEMMS to CRF (Conditional Random Field)

The critical difference between CRFs and MEMMs is that a MEMM uses per-state exponential models for the conditional probabilities of next states given the current state, while a CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. Therefore, the weights of different features at different states can be traded off against each other.

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Lafferty, et al., ICML 2001

CRF (Conditional Random Field)

GLOBALLY LOG-LINEAR MODEL

$$\underline{s} =< y_1 \ldots y_N >=< s_1 \ldots s_N >$$

CRF (Conditional Random Field)

GLOBALLY LOG-LINEAR MODEL

$$\underline{s} =< y_1 \dots y_N >=< s_1 \dots s_N >$$

$$p(Y_1 = y_1, \cdots, Y_N = y_N \mid X_1 = x_1, \cdots, X_N = x_N; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y_{1:N}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, \underline{s}'))}$$

CRF (Conditional Random Field)

GLOBALLY LOG-LINEAR MODEL

$$p(Y_1 = y_1, \cdots, Y_N = y_N \mid X_1 = x_1, \cdots, X_N = x_N; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y_{1:N}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, \underline{s}'))}$$

VS

LOCALLY LOG-LINEAR MODEL

$$p(Y_1 = y_1, \cdots, Y_N = y_N \mid X_1 = x_1, \cdots, X_N = x_N; \underline{\theta}) = \prod_{i=1}^{N} p(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N}; \underline{\theta})$$

$$p(Y_i = y_i \mid Y_{i-1} = y_{i-1}, X_{1:N}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y_i))}{\sum_{y' \in K} exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y'))}$$

CRF (Conditional Random Field)

GLOBALLY LOG-LINEAR MODEL

$$p(Y_1 = y_1, \cdots, Y_N = y_N \mid X_1 = x_1, \cdots, X_N = x_N; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y_{1:N}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, \underline{s}'))}$$

$$\underline{x} =< x_1 \ldots x_N >$$

$$\underline{s} =< y_1 \ldots y_N >=< s_1 \ldots s_N >$$

CRF (Conditional Random Field)

GLOBALLY LOG-LINEAR MODEL

$$p(Y_1 = y_1, \cdots, Y_N = y_N \mid X_1 = x_1, \cdots, X_N = x_N; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, y_{1:N}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(x_{1:N}, \underline{s}'))}$$

$$\underline{x} =< x_1 \ldots x_N >$$

$$\underline{s} =< y_1 \ldots y_N >=< s_1 \ldots s_N >$$

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

CRF (Conditional Random Field)

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

Three Questions:

1. ***Feature Vectors***: How do we define feature vector?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Learning Problem***: How do we learn parameters?

CRF (Conditional Random Field)

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

Three Questions:

1. ***Feature Vectors***: How do we define feature vector?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Learning Problem***: How do we learn parameters?

CRF Feature Vector

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

How do we define $\underline{\Phi}(\underline{x}, \underline{s})$?

CRF Feature Vector

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

How do we define $\underline{\Phi}(\underline{x}, \underline{s})$?

$$\underline{\Phi}(\underline{x}, \underline{s}) = \sum_{j=1}^{N} \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

CRF Feature Vector

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$\underline{\Phi}(\underline{x}, \underline{s}) = \sum_{j=1}^{N} \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

$$\Phi_k(\underline{x}, \underline{s}) = \sum_{j=1}^{N} \phi_k(\underline{x}, j, y_{j-1}, y_j)$$

CRF (Conditional Random Field)

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

Three Questions:

1. ***Feature Vectors***: How do we define feature vector?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Learning Problem***: How do we learn parameters?

CRF DECODING

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\operatorname{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\operatorname{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

$$= \underset{\underline{s} \in K^N}{\operatorname{argmax}} \; \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s})$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \underline{\theta}^T \left(\sum_{j=1}^{N} \underline{\phi}(\underline{x}, j, y_{j-1}, y_j) \right)$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s})$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \underline{\theta}^T \left(\sum_{j=1}^{N} \underline{\phi}(\underline{x}, j, y_{j-1}, y_j) \right)$$

$$= \underset{\underline{s} \in K^N}{\text{argmax}} \; \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\operatorname{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\operatorname{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\operatorname{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\operatorname{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\text{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

Initialization: $For \; k = 1, \ldots, K$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\text{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

Initialization: $For \; k = 1, \ldots, K$

$$\text{viterbi}(1, c_k) = \underline{\theta}^T \underline{\phi}(\underline{x}, 1, START, c_k)$$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\text{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

Initialization: $For\ k = 1, \ldots, K$

$$\text{viterbi}(1, c_k) = \underline{\theta}^T \underline{\phi}(\underline{x}, 1, START, c_k)$$

$For\ j = 2, \ldots, N$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\text{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

Initialization: $For\ k = 1, \ldots, K$

$$\text{viterbi}(1, c_k) = \underline{\theta}^T \underline{\phi}(\underline{x}, 1, START, c_k)$$

$For\ j = 2, \ldots, N$

$For\ k = 1, \ldots, K$

CRF DECODING

$$\underset{\underline{s} \in K^N}{\text{argmax}} \; p(\underline{s} \mid \underline{x}; \underline{\theta}) = \underset{\underline{s} \in K^N}{\text{argmax}} \sum_{j=1}^{N} \underline{\theta}^T \underline{\phi}(\underline{x}, j, y_{j-1}, y_j)$$

Initialization: $For\ k = 1, \ldots, K$

$$\text{viterbi}(1, c_k) = \underline{\theta}^T \underline{\phi}(\underline{x}, 1, START, c_k)$$

$For\ j = 2, \ldots, N$

$For\ k = 1, \ldots, K$

$$\text{viterbi}(j, c_k) = \max_{c_l} \Big(\text{viterbi}(j-1, c_l) + \underline{\theta}^T \underline{\phi}(\underline{x}, j, c_l, c_k) \Big)$$

$end\ for$

$end\ for$

CRF DECODING

$$\max_{\underline{s} \in K^N} p(\underline{s} \mid \underline{x}; \underline{\theta}) = \max_{c_l} \text{viterbi}(N, c_l)$$

CRF (Conditional Random Field)

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

Three Questions:

1. ***Feature Vectors***: How do we define feature vector?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Learning Problem***: How do we learn parameters?

CRF Learning the Parameters

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

CRF Learning the Parameters

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

$$\underline{\theta}^* = \underset{\underline{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

CRF Learning the Parameters

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

$$\mathcal{L}(\underline{\theta}) = \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

$$\underline{\theta}^* = \underset{\underline{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

$$\frac{d\mathcal{L}}{d\theta_k} = ?$$

CRF Learning the Parameters

$$\underline{\theta}^* = \underset{\underline{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

$$\frac{d\mathcal{L}}{d\theta_k} = \sum_i \Phi_k \left(\underline{x}^i, \underline{s}^i \right) - \sum_i \sum_{\underline{s} \in K^N} p \left(\underline{s} | \underline{x}^i; \underline{\theta} \right) \Phi_k \left(\underline{x}^i, \underline{s} \right) - \lambda \theta_k$$

CRF Learning the Parameters

$$\underline{\theta}^* = \underset{\underline{\theta} \in \mathbb{R}^d}{\operatorname{argmax}} \sum_{i=1}^{m} \log \; p(\underline{s}^i \mid \underline{x}^i; \underline{\theta}) - \frac{\lambda}{2} \mid\mid \underline{\theta} \mid\mid^2$$

$$\frac{d\mathcal{L}}{d\theta_k} = \sum_i \Phi_k \left(\underline{x}^i, \underline{s}^i \right) - \sum_i \sum_{\underline{s} \in K^N} p \left(\underline{s} | \underline{x}^i; \underline{\theta} \right) \Phi_k \left(\underline{x}^i, \underline{s} \right) - \lambda \theta_k$$

$$\sum_i \Phi_k \left(\underline{x}^i, \underline{s}^i \right) = \sum_i \sum_{j=1}^{N} \phi_k \left(\underline{x}^i, j, s^i_{j-1}, s^i_j \right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s} | \underline{x}^i; \underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \sum_{j=1}^{m} \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i,\underline{s}\right) = \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \sum_{j=1}^{m} \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \sum_{j=1}^{m} \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \sum_{j=1}^{m} \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} \phi_k\left(\underline{x}^i, j, a, b\right) \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \sum_{j=1}^{m} \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{\underline{s} \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} \sum_{\substack{s \in K^N: \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \phi_k\left(\underline{x}^i, j, s_{j-1}, s_j\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} \phi_k\left(\underline{x}^i, j, a, b\right) \sum_{\substack{s \in K^N: \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

$$= \sum_{j=1}^{N} \sum_{a \in K, b \in K} q_j^i(a,b) \phi_k\left(\underline{x}^i, j, a, b\right) \qquad \text{where,} \quad q_j^i(a,b) = \sum_{\substack{s \in K^N: \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{j=1}^{N} \sum_{a \in K, b \in K} q_j^i(a,b) \phi_k\left(\underline{x}^i, j, a, b\right)$$

$$\text{where,} \quad q_j^i(a,b) = \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{j=1}^{N} \sum_{a \in \mathcal{S}, b \in \mathcal{S}} q_j^i(a,b) \phi_k\left(\underline{x}^i, j, a, b\right)$$

$$\text{where, } \; q_j^i(a,b) = \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

$q_j^i(a,b) =$ Probability of the i^{th} training example x^i having state a at position $j-1$ and state b at position j under the distribution $p\left(\underline{s}|\underline{x};\underline{\theta}\right)$

CRF Learning the Parameters

$$\sum_{s \in K^N} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right) \Phi_k\left(\underline{x}^i, \underline{s}\right) = \sum_{j=1}^{N} \sum_{a \in \mathcal{S}, b \in \mathcal{S}} q_j^i(a,b) \phi_k\left(\underline{x}^i, j, a, b\right)$$

$$\text{where, } q_j^i(a,b) = \sum_{\substack{s \in K^N : \\ s_{j-1}=a, s_j=b}} p\left(\underline{s}|\underline{x}^i;\underline{\theta}\right)$$

$q_j^i(a,b) =$ Probability of the i^{th} training example x^i having state a at position $j-1$ and state b at position j under the distribution $p\left(\underline{s}|\underline{x};\underline{\theta}\right)$

$q_j^i(a,b)$ can be calculated using **Forward-Backward** Algorithm

CRF (Conditional Random Field)

$$p(\underline{s} \mid \underline{x}; \underline{\theta}) = \frac{exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}))}{\sum_{\underline{s}' \in K^N} exp(\underline{\theta}^T \underline{\Phi}(\underline{x}, \underline{s}'))}$$

Three Questions:

1. ***Feature Vectors***: How do we define feature vector?
2. ***Decoding Problem***: Given the observed text what is the hidden POS sequence that best explains the observation?
3. ***Learning Problem***: How do we learn parameters?

Summary

1. MEMMs make it possible to include rich set of features in sequence prediction
2. However, MEMMs are locally normalized and suffer from label-bias problem
3. Global normalization can help
4. CRF perform global normalization and predict the entire sequence
5. Decoding in CRF can be done efficiently using Viterbi
6. Parameter Estimation in CRF involves forward-backward algorithm, hence training in CRF is slow

References

1. Michael Collin's NLP Lecture Notes: http://www.cs.columbia.edu/~mcollins/crf.pdf
2. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Lafferty, et al., ICML 2001: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers
3. Label Bias Problem explained well: https://awni.github.io/label-bias/