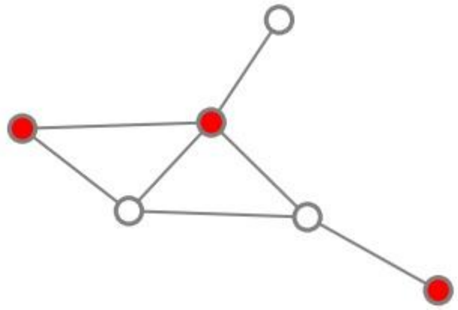
Belief Propagation

Csaba Both and Cory Glover

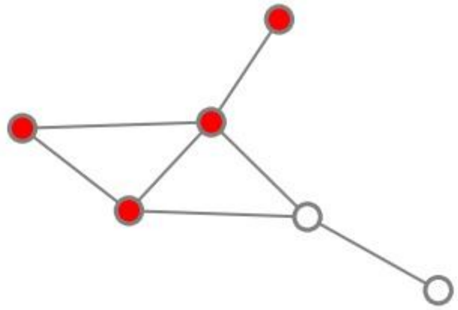
Belief Propagation

- Nodes pass information to each other to reach a consensus on their state.
- We may not have direct information of the states of the node and need to infer them based on some probability distribution.

Observed:



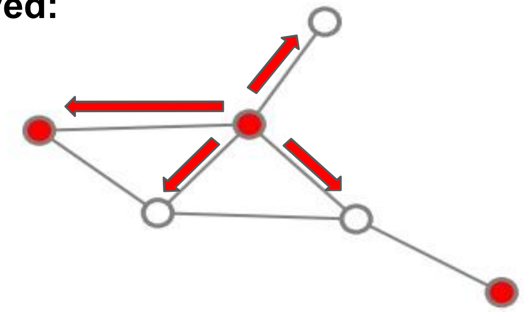
True States:



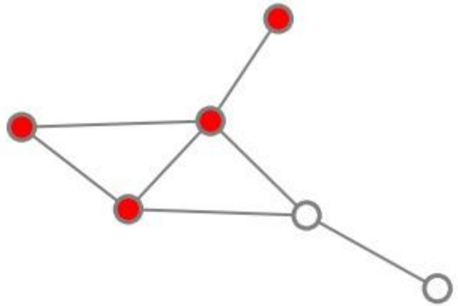
Belief Propagation

- Nodes pass information to each other to reach a consensus on their state.
- Based on observing the deliberation in the network, we want to identify the true state of each node.

Observed:



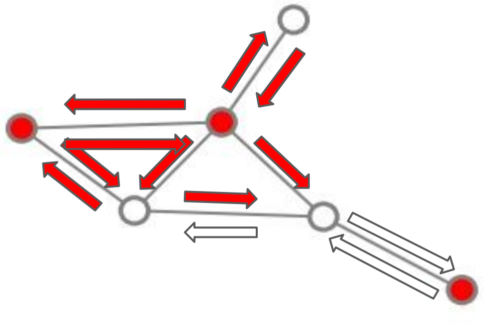
True States:



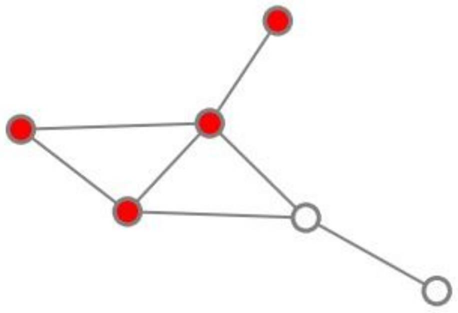
Belief Propagation

- Nodes pass information to each other to reach a consensus on their state.
- Based on observing the deliberation in the network, we want to identify the true state of each node.

Observed:



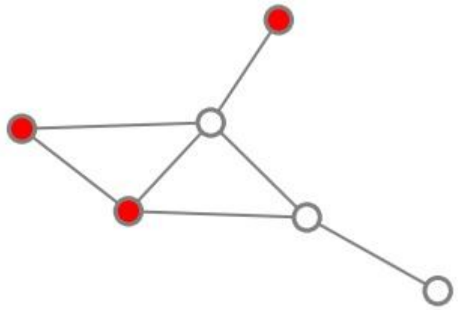
True States:



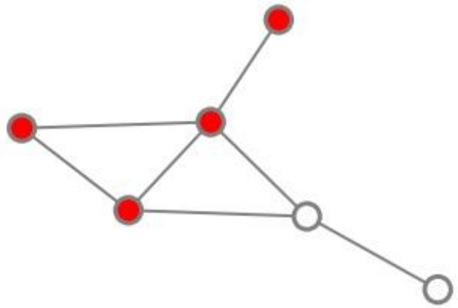
Belief Propagation

- Nodes pass information to each other to reach a consensus on their state.
- Based on observing the deliberation in the network, we want to identify the true state of each node.

Observed:

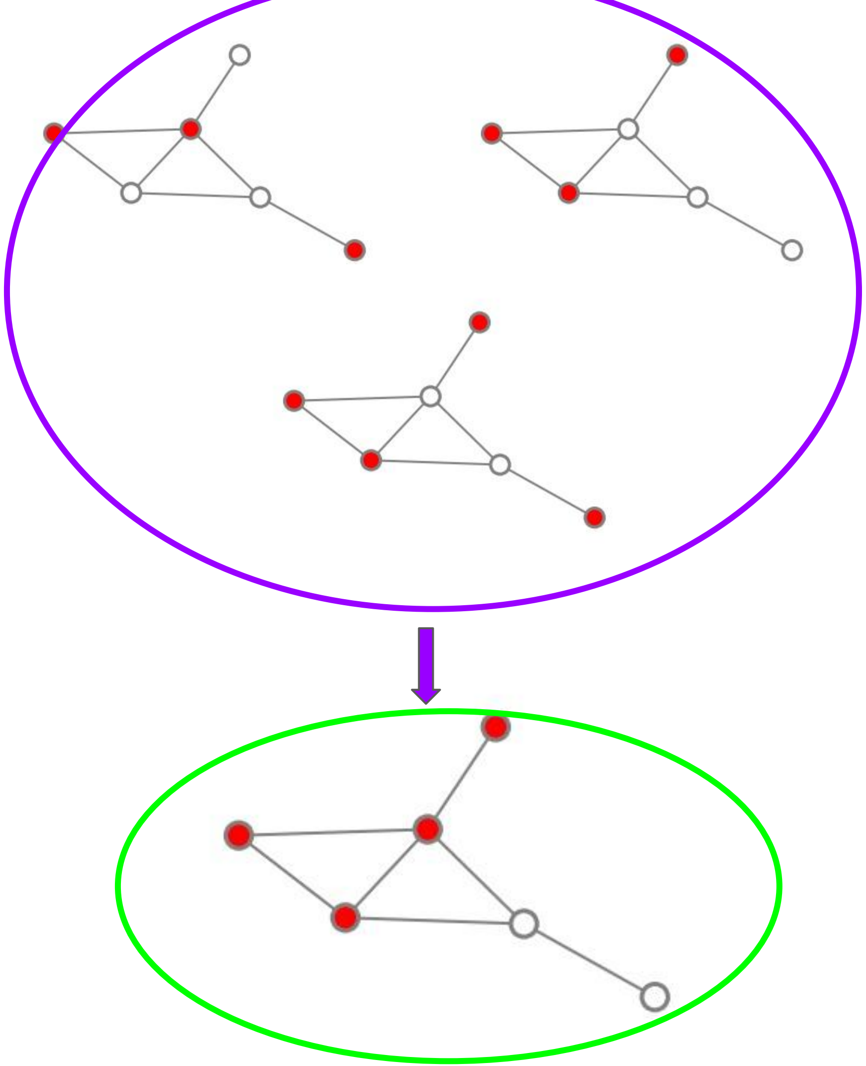


True States:

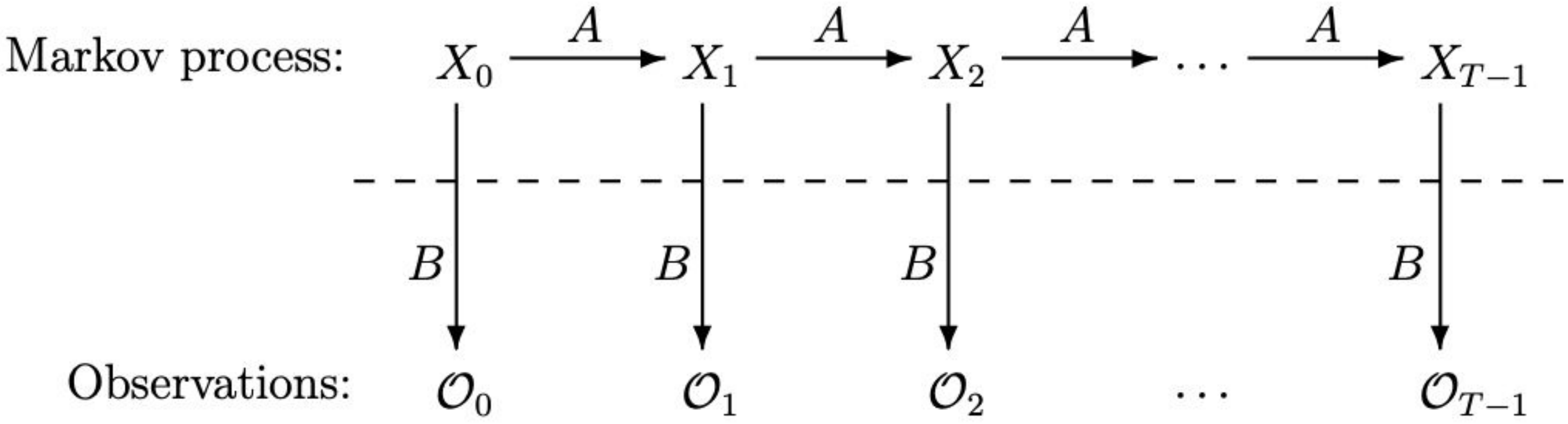


Belief Propagation

- Nodes pass information to each other to reach a consensus on their state.
- Based on observing the deliberation in the network, we want to identify the true state of each node.



Hidden Markov Model



Mark Stamp Lecture notes https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf

Weather Example

Markov Process

Observations

States:

- Hot
- Cold

$$
\begin{array}{c} \\ H \\ C \end{array}
\begin{array}{c} \begin{array}{cc} H & C \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}
$$

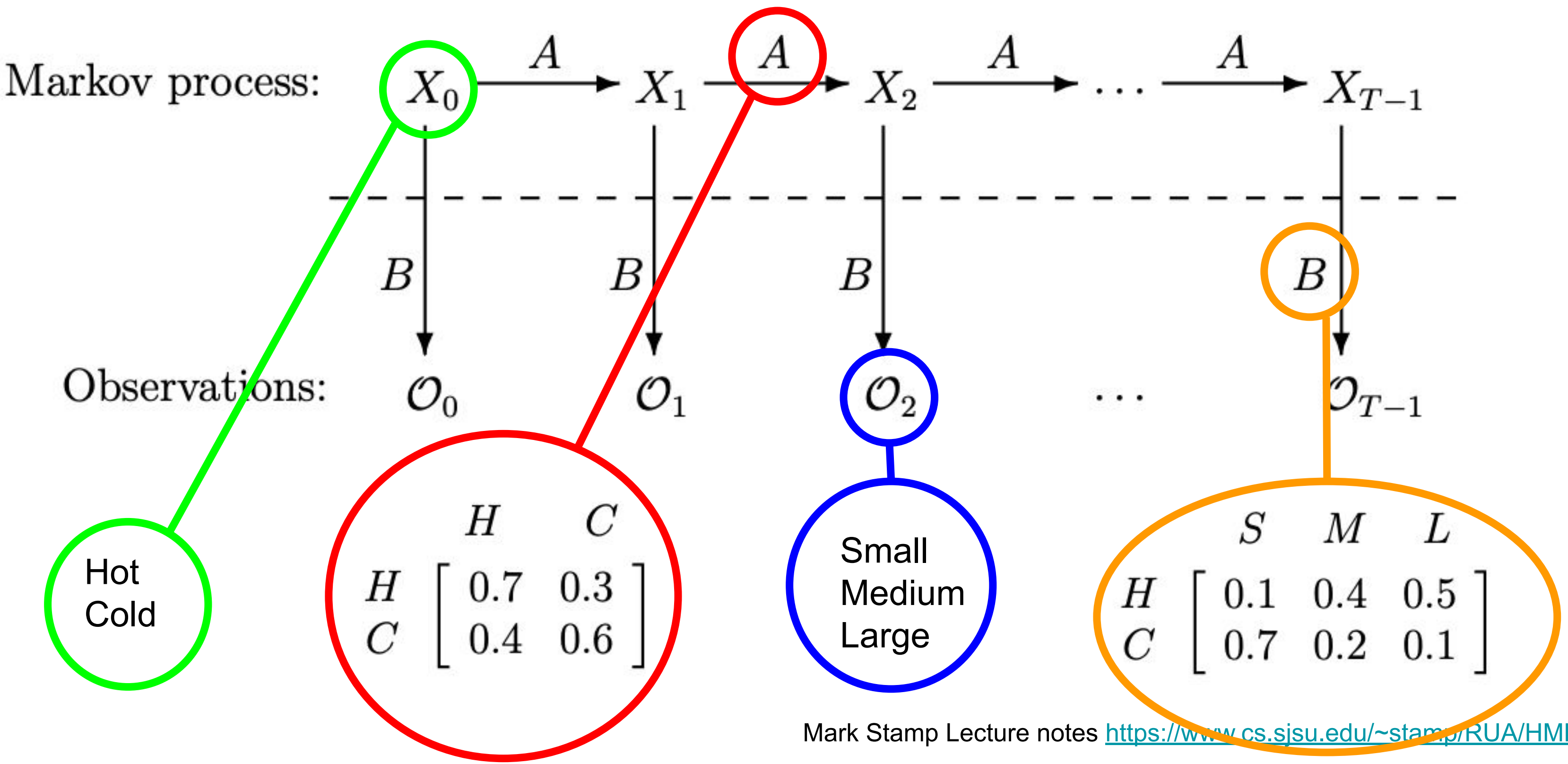
States:

- Small
- Medium
- Large

$$
\begin{array}{c} \\ H \\ C \end{array}
\begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}
$$

Mark Stamp Lecture notes https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf

Hidden Markov Model



Mark Stamp Lecture notes https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf

Three Problems to Solve with HMMs

Problem 1

Given a model and a sequence of observations, what is the probability of getting those observations with the given model?

$$\mathcal{O} + \lambda$$

$$\Downarrow$$

$$P(\mathcal{O} \mid \lambda)$$

Problem 2

Given a model and a sequence of observations, what is the optimal state sequence to obtain the observations?

$$\mathcal{O} + \lambda$$

$$\Downarrow$$

$$\mathrm{X}$$

Problem 3

Given a sequence of observations and the number of markov and observation states, what model maximizes the probability of obtaining the sequence of observations?

$$\mathcal{O} + \mathrm{N,M}$$

$$\Downarrow$$

$$\lambda$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$\mathcal{O} = (\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_{T-1})$ - Observation sequence

π - Initial distribution

$X = (x_0, x_1, \cdots, x_{T-1})$ - Given state sequence

B - Observation state transition matrix

A - State space transition matrix

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0) B_{x_1}(\mathcal{O}_1) \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0) B_{x_1}(\mathcal{O}_1) \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

Example

$$X = (H, H, C)$$
$$\mathcal{O} = (M, L, S)$$

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}$$

$$P(\mathcal{O} \mid X, \lambda) = (0.4)(0.5)(0.7) = 0.14$$

Problem 1 - $P(\mathcal{O} \,|\, \lambda)$

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0) B_{x_1}(\mathcal{O}_1) \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

$$P(X \mid \lambda) = \pi_{x_0} A_{x_0,x_1} A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}}$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0) B_{x_1}(\mathcal{O}_1) \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

$$P(X \mid \lambda) = \pi_{x_0} A_{x_0,x_1} A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}}$$

$$P(\mathcal{O}, X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

$$P(\mathcal{O} \mid X, \lambda) P(X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(X \cap \lambda)} \cdot \frac{P(X \cap \lambda)}{P(\lambda}) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0)B_{x_0}(\mathcal{O}_1)B_{x_1} \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

$$P(X \mid \lambda) = \pi_{x_0} A_{x_0,x_1} A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}}$$

$$P(\mathcal{O}, X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

$$P(\mathcal{O} \mid X, \lambda)P(X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(X \cap \lambda)} \cdot \frac{P(X \cap \lambda)}{P(\lambda}) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

$$P(\mathcal{O} \mid \lambda) = \sum_X P(\mathcal{O}, X \mid \lambda)$$

$$= \sum_X \pi_{x_0} B_{x_0}(\mathcal{O}_o)A_{x_0,x_1} B_{x_1}(\mathcal{O}_1)A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}} B_{x_{T-1}}(\mathcal{O}_{T-1})$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$$P(\mathcal{O} \mid X, \lambda) = B_{x_0}(\mathcal{O}_0) B_{x_0}(\mathcal{O}_1) B_{x_1} \cdots B_{x_{T-1}}(\mathcal{O}_{T-1})$$

$$P(X \mid \lambda) = \pi_{x_0} A_{x_0,x_1} A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}}$$

$$P(\mathcal{O}, X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

$$P(\mathcal{O} \mid X, \lambda) P(X \mid \lambda) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(X \cap \lambda)} \cdot \frac{P(X \cap \lambda)}{P(\lambda}) = \frac{P(\mathcal{O} \cap X \cap \lambda)}{P(\lambda)}$$

$$2TN^T$$

$$P(\mathcal{O} \mid \lambda) = \sum_X P(\mathcal{O}, X \mid \lambda)$$

$$= \sum_X \pi_{x_0} B_{x_0}(\mathcal{O}_o) A_{x_0,x_1} B_{x_1}(\mathcal{O}_1) A_{x_1,x_2} \cdots A_{x_{T-2},x_{T-1}} B_{x_{T-1}}(\mathcal{O}_{T-1})$$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

Forward Algorithm

- Recursive algorithm
- N^2T

$$\alpha_t(i) = P(\mathcal{O}_0, \mathcal{O}_1, \cdots, \mathcal{O}_t \mid x_t = q_i, \lambda)$$

1) $\alpha_0(i) = \pi_i B_i(\mathcal{O}_i), \forall i \in [0, N-1]$
2) $\alpha_t(i) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j) a_{ji}\right] b_i(\mathcal{O}_t), \forall i \in [0, N-1]$
3) $P(\mathcal{O} \mid \lambda) = \sum_{i=0}^{N-1} \alpha_{T-1}(i)$

Problem 1 - $P(\mathcal{O} \mid \lambda)$

$$\alpha_0(i) = \pi_i B_i(\mathcal{O}_i), \forall i \in [0, N-1]$$

$$\alpha_t(i) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j) a_{ji}\right] b_i(\mathcal{O}_t), \forall i \in [0, N-1]$$

$$P(\mathcal{O} \mid \lambda) = \sum_{i=0}^{N-1} \alpha_{T-1}(i)$$

$$X = (H, H, C)$$

$$\mathcal{O} = (M, L, S)$$

$$\begin{array}{c|cc} & H & C \\ \hline H & 0.7 & 0.3 \\ C & 0.4 & 0.6 \end{array}$$

$$\begin{array}{c|ccc} & S & M & L \\ \hline H & 0.1 & 0.4 & 0.5 \\ C & 0.7 & 0.2 & 0.1 \end{array} \qquad \pi = \begin{pmatrix} .4 \\ .6 \end{pmatrix}$$

$$\alpha_0(H) = (.4)(.4) = .16$$

$$\alpha_0(C) = (.6)(.2) = .12$$

$$\alpha_1(H) = [(0.16)(0.7) + (0.12)(0.4)](0.4) = 0.064$$

$$\alpha_1(C) = [(0.16)(0.3) + (0.12)(0.6)](0.5) = 0.06$$

$$\alpha_2(H) = [(0.064)(0.7) + (0.06)(0.4)](0.1) = 0.00688$$

$$\alpha_2(C) = [(0.064)(0.3) + (0.06)(0.6)](0.7) = 0.03864$$

$$P(\mathcal{O} \mid \lambda) = \alpha_2(H) + \alpha_2(C) = 0.04552$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

- Want to find “most likely” state sequence, i.e. maximize the expected number of correct states

Backwards Algorithm

$$\beta_t(i) = P(\mathcal{O}_{t+1}, \mathcal{O}_{t+2}, \cdots, \mathcal{O}_{T-1} \mid x_t = q_i, \lambda)$$

1) $\beta_{T-1}(i) = 1, \forall i \in [0, N-1]$

2) $\beta_t(i) = \sum_{j=0}^{N-1} A_{ij} B_j(\mathcal{O}_{t+1}) \beta_{t+1}(j), \forall i \in [0, N-1]$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

$$\beta_{T-1}(i) = 1, \forall i \in [0, N-1]$$

$$\beta_t(i) = \sum_{j=0}^{N-1} A_{ij} B_j(\mathcal{O}_{t+1}) \beta_{t+1}(j), \forall i \in [0, N-1]$$

$$\beta_2(H) = 1$$
$$\beta_2(C) = 1$$

$$\beta_1(H) = (0.7)(0.1)(1) + (0.3)(0.7)(1) = 0.28$$
$$\beta_1(C) = (0.4)(0.1)(1) + (0.6)(0.7)(1) = 0.46$$

$$\beta_0(H) = (0.7)(0.5)(0.28) + (0.3)(0.1)(0.46) = .1118$$
$$\beta_0(C) = (0.4)(0.5)(0.28) + (0.6)(0.1)(0.46) = 0.0836$$

$$X = (H, H, C)$$
$$\mathcal{O} = (M, L, S)$$

$$\begin{array}{c c} & \begin{array}{cc} H & C \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

$$\begin{array}{c c} & \begin{array}{ccc} S & M & L \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array} \quad \pi = \begin{pmatrix} .4 \\ .6 \end{pmatrix}$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow$ X

GOAL - $\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda)$

$$\gamma_t(i) = P(x_t = q_i \mid \mathcal{O}, \lambda) = \frac{P(x_t = q_i \cap \mathcal{O} | \lambda)}{P(\mathcal{O} | \lambda)}$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

GOAL - $\mathrm{argmax}_i \, P(x_t = q_i \mid \mathcal{O}, \lambda)$

$$\gamma_t(i) = P(x_t = q_i \mid \mathcal{O}, \lambda) = \frac{P(x_t = q_i \cap \mathcal{O} | \lambda)}{P(\mathcal{O}|\lambda)}$$

$$P(x_t = q_i \cap \mathcal{O} \mid \lambda) = P(\mathcal{O}_1, \cdots, \mathcal{O}_t \mid x_t = q_i, \lambda) P(\mathcal{O}_{t+1}, \cdots, \mathcal{O}_{T-1} \mid x_t = q_i, \lambda) P(x_t = q_i)$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow X$

GOAL - $\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda)$

$$\gamma_t(i) = P(x_t = q_i \mid \mathcal{O}, \lambda) = \frac{P(x_t = q_i \cap \mathcal{O} \mid \lambda)}{P(\mathcal{O} \mid \lambda)}$$

$$P(x_t = q_i \cap \mathcal{O} \mid \lambda) = \underbrace{P(\mathcal{O}_1, \cdots, \mathcal{O}_t \mid x_t = q_i, \lambda)}_{\alpha_t(i)} \underbrace{P(\mathcal{O}_{t+1}, \cdots, \mathcal{O}_{T-1} \mid x_t = q_i, \lambda)}_{\beta_t(i)} \underbrace{P(x_t = q_i)}_{1}$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

GOAL - $\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda)$

$$\gamma_t(i) = P(x_t = q_i \mid \mathcal{O}, \lambda) = \frac{P(x_t = q_i \cap \mathcal{O} \mid \lambda)}{\cancel{P(\mathcal{O} \mid \lambda)}}$$

$$P(x_t = q_i \cap \mathcal{O} \mid \lambda) = \underbrace{P(\mathcal{O}_1, \cdots, \mathcal{O}_t \mid x_t = q_i, \lambda)}_{\alpha_t(i)} \underbrace{P(\mathcal{O}_{t+1}, \cdots, \mathcal{O}_{T-1} \mid x_t = q_i, \lambda)}_{\beta_t(i)} \underbrace{P(x_t = q_i)}_{1}$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

GOAL - $\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda)$

$$\gamma_t(i) = P(x_t = q_i \mid \mathcal{O}, \lambda) = \frac{P(x_t = q_i \cap \mathcal{O} \mid \lambda)}{\cancel{P(\mathcal{O} \mid \lambda)}}$$

$$P(x_t = q_i \cap \mathcal{O} \mid \lambda) = \underbrace{P(\mathcal{O}_1, \cdots, \mathcal{O}_t \mid x_t = q_i, \lambda)}_{\alpha_t(i)} \underbrace{P(\mathcal{O}_{t+1}, \cdots, \mathcal{O}_{T-1} \mid x_t = q_i, \lambda)}_{\beta_t(i)} \underbrace{P(x_t = q_i)}_{1}$$

$$\boxed{\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda) = \operatorname{argmax}_i \alpha_t(i)\beta_t(i)}$$

$$X = (\operatorname{argmax}_i \gamma_1(i), \operatorname{argmax}_i \gamma_2(i), \ldots, \operatorname{argmax}_i \gamma_{T-1}(i))$$

Problem 2 - $\mathcal{O} + \lambda \longrightarrow \mathrm{X}$

$$X = (\operatorname{argmax}_i \gamma_1(i), \operatorname{argmax}_i \gamma_2(i), \ldots, \operatorname{argmax}_i \gamma_{T-1}(i))$$

$$\operatorname{argmax}_i P(x_t = q_i \mid \mathcal{O}, \lambda) = \operatorname{argmax}_i \alpha_t(i)\beta_t(i)$$

$$\alpha_0(H) = 0.16 \quad \beta_0(H) = 0.1118 \rightarrow \gamma_0(H) = 0.017888$$
$$\alpha_0(C) = 0.12 \quad \beta_0(C) = 0.0836 \rightarrow \gamma_0(C) = 0.010032$$

$$\alpha_1(H) = 0.064 \quad \beta_1(H) = 0.28 \rightarrow \gamma_1(H) = 0.01792$$
$$\alpha_1(C) = 0.06 \quad \beta_1(C) = 0.46 \rightarrow \gamma_1(C) = 0.0276$$

$$\alpha_2(H) = 0.00688 \quad \beta_2(H) = 1 \rightarrow \gamma_2(H) = 0.00688$$
$$\alpha_2(C) = 0.03864 \quad \beta_2(C) = 1 \rightarrow \gamma_2(C) = 0.03864$$

$$\rightarrow X = (H, C, C)$$

Problem 3 - $\mathcal{O} + N, M \to \lambda$

- Goal: Adjust the model parameters to best fit the observation
- Efficiently re-estimate the model

$$\gamma_t(i,j) = P(x_t = q_i, x_{t+1} = q_j \,|\, \mathcal{O}, \lambda)$$

di-gammas

$$\gamma_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(\mathcal{O}_{t+1}) \beta_{t+1}(j)}{P(\mathcal{O} \,|\, \lambda)}$$

- The probability of being in state q_i at time t and transiting to state q_j at time $t+1$

$$\gamma_t(i) = \sum_{j=0}^{N-1} \gamma_t(i,j) \qquad \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(\mathcal{O} \,|\, \lambda)}$$

- The most likely state at time t

Problem 3 - $\mathcal{O} + N, M \rightarrow \lambda$

Given: $$\gamma_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(\mathcal{O}_{t+1}) \beta_{t+1}(j)}{P(\mathcal{O} \mid \lambda)}$$ $$\gamma_t(i) = \sum_{j=0}^{N-1} \gamma_t(i,j).$$ $$\lambda(A, B, \Pi) = ?$$

$$\pi_i = \gamma_0(i)$$

$$a_{ij} = \sum_{t=0}^{T-2} \gamma_t(i,j) \Bigg/ \sum_{t=0}^{T-2} \gamma_t(i)$$

expected number of transitions from q_i to q_j /
expected number of transitions from q_i to any states

$$b_j(k) = \sum_{\substack{t \in \{0,1,\ldots,T-1\} \\ \mathcal{O}_t = k}} \gamma_t(j) \Bigg/ \sum_{t=0}^{T-1} \gamma_t(j)$$

expected number that the model is in q_j state with observation k /
expected number that the model is in q_j state

Problem 3 - $\mathcal{O} + N, M \rightarrow \lambda$

$$\gamma_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(\mathcal{O}_{t+1})\beta_{t+1}(j)}{P(\mathcal{O} \mid \lambda)}$$

$$\gamma_0(i,j) = \begin{pmatrix} (0.16)(.7)(0.5)(.28)/0.04552 & (0.16)(0.3)(0.1)(0.46)/0.04552 \\ (0.12)(0.4)0.5)(0.46)/0.04552 & (0.12)(0.6)(0.1)(0.46)/0.04552 \end{pmatrix}$$

$$\mathcal{O} = (M, L, S)$$

$$\alpha_0(H) = 0.16 \qquad \beta_0(H) = 0.1118$$
$$\alpha_0(C) = 0.12 \qquad \beta_0(C) = 0.0836$$

$$\alpha_1(H) = 0.064 \qquad \beta_1(H) = 0.28$$
$$\alpha_1(C) = 0.06 \qquad \beta_1(C) = 0.46$$

$$\alpha_2(H) = 0.00688 \qquad \beta_2(H) = 1$$
$$\alpha_2(C) = 0.03864 \qquad \beta_2(C) = 1$$

$$\begin{array}{cc} & \begin{array}{cc} H & C \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & \begin{array}{ccc} S & M & L \end{array} \\ \begin{array}{c} H \\ C \end{array} & \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array}$$

$$P(\mathcal{O} \mid \lambda) = \alpha_2(H) + \alpha_2(C) = 0.04552$$

$$\gamma_0(i,j) = \begin{pmatrix} 0.34 & 0.05 \\ 0.24 & 0.07 \end{pmatrix}$$

Problem 3 - $\mathcal{O} + N, M \rightarrow \lambda$

Re-estimation of the model is an iterative process: $\lambda(A, B, \Pi) =?$

1. Initialize $\lambda(A, B, \Pi)$
2. Compute $\alpha_t(i)$ $\beta_t(i)$ $\gamma_t(i, j)$ $\gamma_t(i)$
3. Re-estimate $\lambda(A, B, \Pi)$
4. If $P(\mathcal{O}|\lambda)$ increases, goto 2

- Stop if $P(\mathcal{O}|\lambda)$ doesn’t increases

HMM for speech processing

1913, Markov analyzed chains of vowels and consonants in Pushkin's poem Eugene Onegin - Markov chain

The language is the result of long complex process.

- Relation of Markov chains to the English language.
- Try to understand the manner in which letters are put together.
- View language as a sequence of symbols from a 27-letter of the alphabet (space + letters)

HMM: model for generation of English language, which preserves Markov's division

- Separate the alphabet into classes (e.g. vowels & consonants)
    - e.g. english language = vcccvcc cvccvvcv

N = 2 A 2x2 matrix - probabilities of four sequence (v-v, v-c, c-v, c-c)
M = 27 B 2x27 matrix - probability that each letter is consonant or a vowel.
Q = {v,c}

What A, B matrix maximize the probability of observing the text?

Problem 3

S State HMM for speech processing

Different set of parameters: s = 2 (vowel + consonants), s = 5, s =11

TABLE I-1. HIDDEN MARKOV MODEL*

S = 2

Transition Probabilities

1	.275	.725
2	.780	.220

Output Probabilities

	1	2
A	—	.133
B	.022	—
C	.063	—
D	.056	—
E	—	.218
F	.037	—
G	.015	.010
H	.074	—
I	—	.150
J	—	—
K	.009	—
L	.060	—
M	.041	—
N	.140	—
O	—	.136
P	.030	.001
Q	.001	—
R	.087	—
S	.105	—
T	.157	.019
U	—	.045
V	.016	—
W	.020	—
X	.002	—
Y	.004	.018
Z	.001	—
#	.060	.269

Stationary Probabilities

.52	.48

TABLE I-4

S - 5

Transition Probabilities

1	.118	.026	.854	—	—
2	.014	.161	—	—	.824
3	.141	.135	.127	.520	.075
4	.197	.396	.026	.015	.364
5	.732	.004	.263	—	—

Output Probabilities

	1	2	3	4	5
A	—	—	.236	—	—
B	.036	—	—	.006	—
C	.114	—	—	.016	.005
D	.053	.182	—	.020	—
E	—	.260	.237	—	—
F	.040	.003	—	.043	—
G	.029	.054	—	.007	.006
H	.048	—	.088	—	—
I	—	—	.150	.004	.072
J	.004	—	—	—	—
K	.012	.007	—	.001	—
L	.047	.047	—	.137	—
M	.052	—	—	.034	—
N	.034	.004	—	.326	—
O	—	—	.216	—	—
P	.077	.004	—	.011	—
Q	.002	—	—	—	—
R	.093	—	—	.213	—
S	.082	.192	—	.108	—
T	.192	.167	—	.022	—
U	—	—	.068	.003	.002
V	.024	.001	—	.005	—
W	.038	—	—	.007	—
X	—	—	—	.016	—
Y	.012	.073	.001	.010	—
Z	.002	—	—	—	—
#	—	—	—	—	.912

Stationary Probabilities

.238	.129	.293	.155	.185

TABLE I-10

S - 11

Transition Probabilities

1	—	.049	—	—	.068	—	—	.105	.535	—	.240
2	—	.040	.008	—	—	.089	.437	.001	—	.049	.372
3	—	.036	—	.698	.221	.001	—	—	.042	—	—
4	.013	.285	—	.007	.501	—	—	—	.190	.001	—
5	—	.033	.007	—	.011	.166	.159	.012	—	.608	—
6	—	—	—	—	.019	—	—	.015	—	.965	—
7	.010	.054	.029	—	.072	.107	.001	.078	—	.276	.370
8	—	.313	—	—	—	—	—	.008	.675	—	.002
9	—	.025	.004	—	.004	—	.665	.116	.076	.005	.101
10	—	.178	.253	.044	—	—	—	.409	.114	—	—
11	.192	.105	.002	—	.218	.124	—	.013	.076	.236	.029

Output Probabilities

	1	2	3	4	5	6	7	8	9	10	11
A	—	.596		—	—	—	—	—	.035	—	—
B	—	—	.077	—	—	—	.015	.063	—	—	.001
C	—	—	.041	—	—	—	.043	.110	—	—	.071
D	—	—	.002	—	—	.333	.011	.085	—	—	.081
E	—	—	—	—	.854	—	—	—	.340	—	—
F	—	—	.014	—	—	—	.058	.081	—	—	.012
G	—	—	.024	—	—	.030	—	.031	—	—	.067
H	.147	—	—	.700	—	.068	—	.021	—	—	—
I	.679	.359	—	—	—	—	.003	—	.033	—	—
J	—	—	—	—	—	—	—	.025	—	—	—
K	—	—	—	—	—	.001	—	.001	—	—	.030
L	—	—	—	.047	—	—	.098	.054	—	—	.126
M	.068	—	—	—	—	.002	.041	.081	—	—	.027
N	—	—	—	—	—	.013	.319	.038	—	—	.036
O	—	.002	—	.003	.116	—	—	—	.443	—	—
P	—	—	.115	—	—	—	.013	.105	—	—	.001
Q	—	—	—	—	—	—	—	.006	—	—	—
R	—	—	—	.212	—	—	.248	.073	—	—	.008
S	—	.041	.027	—	—	.361	.080	.095	—	—	.065
T	.039	—	.635	—	—	—	.009	—	—	—	.413
U	.065	—	—	.003	—	—	.028	—	.147	—	—
V	—	—	—	—	—	—	—	.037	—	—	.035
W	—	—	.059	.009	—	—	.012	.084	—	—	—
X	—	—	—	—	—	—	.015	—	—	—	—
Y	—	—	—	.023	.029	.188	—	—	—	—	.015
Z	—	—	—	—	—	—	—	—	—	—	.004
#	—	—	—	—	—	—	—	—	—	1.00	—

Stationary Probabilities

.024	.106	.049	.042	.071	.051	.144	.101	.129	.166	.116

S State HMM for speech processing

TABLE I-4

S - 5

Transition Probabilities

1	.118	.026	.854	—	—
2	.014	.161	—	—	.824
3	.141	.135	.127	.520	.075
4	.197	.396	.026	.015	.364
5	.732	.004	.263	—	—

Output Probabilities

	1	2	3	4	5
A	—	—	.236	—	—
B	.036	—	—	.006	—
C	.114	—	—	.016	.005
D	.053	.182	—	.020	—
E	—	.260	.237	—	—
F	.040	.003	—	.043	—
G	.029	.054	—	.007	.006
H	.049	—	.088	—	—
I	—	—	.150	.004	.072
J	.004	—	—	—	—
K	.012	.007	—	.001	—
L	.047	.047	—	.137	—
M	.052	—	—	.034	—
N	.034	.004	—	.326	—
O	—	—	.216	—	—
P	.077	.004	—	.011	—
Q	.002	—	—	—	—
R	.093	—	—	.213	—
S	.082	.192	—	.108	—
T	.192	.167	—	.022	—
U	—	—	.068	.003	.002
V	.024	.001	—	.005	—
W	.038	—	—	.007	—
X	—	—	—	.016	—
Y	.012	.073	.001	.010	—
Z	.002	—	—	—	—
#	—	—	—	—	.912

Stationary Probabilities

.238	.129	.293	.155	.185

5 States

Vowel or h	<—	(3)
Space	—>	(5)
Consonant	<—	(1,4)
Final Letter	—>	(2,3,4)
Initial Letter	—>	(1,3)

S State HMM for speech processing

TABLE I-10

S - 11

Transition Probabilities

	1	2	3	4	5	6	7	8	9	10	11
1	—	.049	—	—	.068	—	—	.105	.535	—	.240
2	—	.040	.008	—	—	.089	.437	.001	—	.049	.372
3	—	.036	—	.698	.221	.001	—	—	.042	—	—
4	.013	.285	—	.007	.501	—	—	—	.190	.001	—
5	—	.033	.007	—	.011	.166	.159	.012	—	.608	—
6	—	—	—	—	.019	—	—	.015	—	.965	—
7	.010	.054	.029	—	.072	.107	.001	.078	—	.276	.370
8	—	.313	—	—	—	—	—	.008	.675	—	.002
9	—	.025	.004	—	.004	—	.665	.116	.076	.005	.101
10	—	.178	.253	.044	—	—	—	.409	.114	—	—
11	.192	.105	.002	—	.218	.124	—	.013	.076	.236	.029

Output Probabilities

	1	2	3	4	5	6	7	8	9	10	11
A	—	.596		—	—	—	—	—	.035	—	—
B	—	—	.077	—	—	—	.015	.063	—	—	.001
C	—	—	.041	—	—	—	.043	.110	—	—	.071
D	—	—	.002	—	—	.333	.011	.085	—	—	.081
E	—	—	—	—	.854	—	—	—	.340	—	—
F	—	—	.014	—	—	—	.058	.081	—	—	.012
G	—	—	.024	—	—	.030	—	.031	—	—	.067
H	.147	—	—	.700	—	.068	—	.021	—	—	—
I	.679	.359	—	—	—	—	.003	—	.033	—	—
J	—	—	—	—	—	—	—	.025	—	—	—
K	—	—	—	—	—	.001	—	.001	—	—	.030
L	—	—	—	.047	—	—	.098	.054	—	—	.126
M	.068	—	—	—	—	.002	.041	.081	—	—	.027
N	—	—	—	—	—	.013	.319	.038	—	—	.036
O	—	.002	—	.003	.116	—	—	—	.443	—	—
P	—	—	.115	—	—	—	.013	.105	—	—	.001
Q	—	—	—	—	—	—	—	.006	—	—	—
R	—	—	—	.212	—	—	.248	.073	—	—	.008
S	—	.041	.027	—	—	.361	.080	.095	—	—	.065
T	.039	—	.635	—	—	—	.009	—	—	—	.413
U	.065	—	—	.003	—	—	.028	—	.147	—	—
V	—	—	—	—	—	—	—	.037	—	—	.035
W	—	—	.059	.009	—	—	.012	.084	—	—	—
X	—	—	—	—	—	—	.015	—	—	—	—
Y	—	—	—	.023	.029	.188	—	—	—	—	.015
Z	—	—	—	—	—	—	—	—	—	—	.004
#	—	—	—	—	—	—	—	—	—	1.00	—

Stationary Probabilities

1	2	3	4	5	6	7	8	9	10	11
.024	.106	.049	.042	.071	.051	.144	.101	.129	.166	.116

11 States

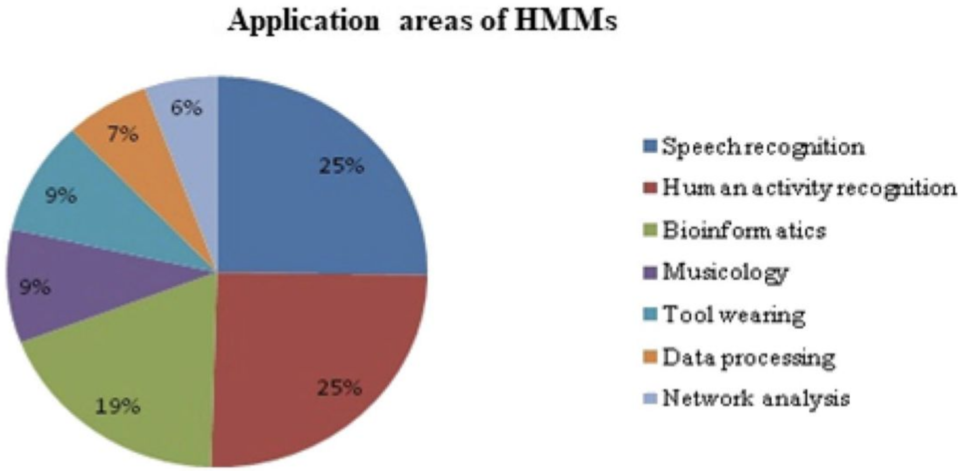
Vowel	<—	{2,5,9}
Space	<—>	{10}
Consonant	<—	{3,4,6,7,8,11}
Final Letter	—>	{2,5,6,7,11}
Initial Letter	—>	{2,3,4,8,9}
Vowel Follower	<—	{7}
Vowel Preceder	<—	{4,8}
Final Letter	<—	{6}
Post-Consonant	<—	{1}

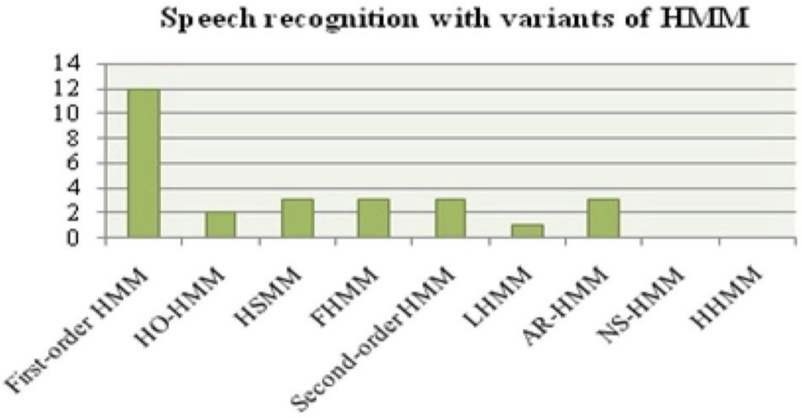
The transition 3 —> 4 is strong. t dominates 3, h dominates 4, i dominates 1, and 1 is entered 90% of the time from 3 (t dominated).

- Disjoint states
- As the number of state increases the uncertainty of letter produced by a state decreases uniformly.

Applications of HMM

- speech recognition
- facial expression recognition
- gene prediction
- gesture recognition
- musical composition
- bio-informatics

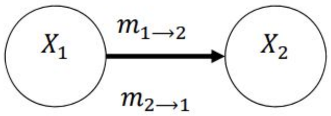




HO-HMM: higher order HMM extends the dependency from the previous state to n state

Belief propagation

- Message passing for performing inference on graphical-models.
- Each node has a marginal, $P(X_i)$ - belief
- Exact on trees graphs, but not exact on general graphs (loops)
- Node's belief affected by neighbors (message)



- The higher the value of the message, the more likely the nodes change their belief respect to the message
- At convergence the belief of the node is its marginal probability

Belief propagation

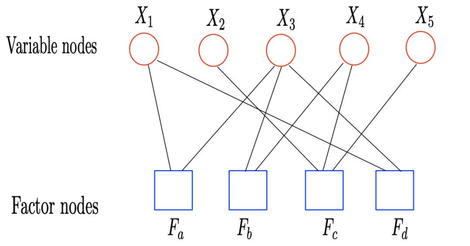
- Given finite set of random discrete variable (nodes), the goal is to calculate the marginal probability for each individual variable X_i

$$p_{X_i}(x_i) = \sum_{\mathbf{x}' : x'_i = x_i} p(\mathbf{x}')$$

,where $\mathbf{x}'$ possible values (features, states) for X_i

- Expensive calculation, most of the time it is intractable
- Efficient calculation: factor graphs (factorization of probability distribution function)
- The joint mass function

$$p(\mathbf{x}) = \prod_{a \in F} f_a(\mathbf{x}_a)$$

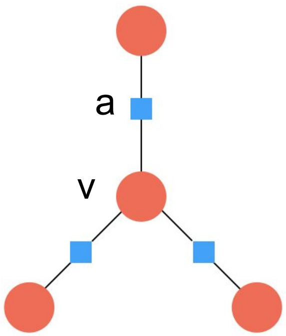


$$p(\mathbf{x}) = f_a(x_1, x_3) f_b(x_3, x_4) f_c(x_2, x_4, x_5) f_d(x_1, x_3)$$

Belief propagation

- sum-product algorithm - sending messages along the edges between nodes
- the message is a set of values, that can be taken by a random variable

- Two different node types: variable (v) and factor (a)
- Two different messages



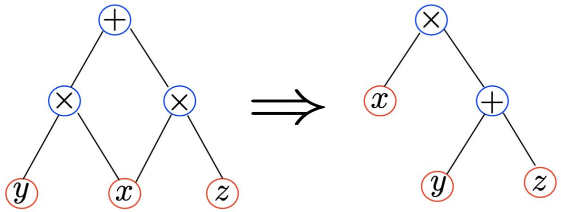
- Variable to factor: $$\mu_{v \to a}(x_v) = \prod_{a^* \in N(v) \setminus \{a\}} \mu_{a^* \to v}(x_v)$$

- Factor to variable: $$\mu_{a \to v}(x_v) = \sum_{\mathbf{x}'_a : x'_v = x_v} \left(f_a(\mathbf{x}'_a) \prod_{v^* \in N(a) \setminus \{v\}} \mu_{v^* \to a}(x'_{v^*}) \right)$$

Belief propagation

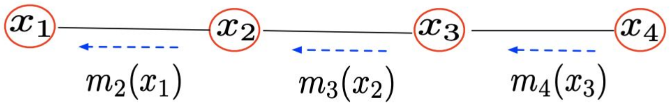
- Key ide, behind the sum-product algorithm:

$$xy + xz = x(y + z)$$



- Sum-product algorithm involves using message passing scheme to change the order of sum and product:

$$p(\mathrm{x}) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)$$



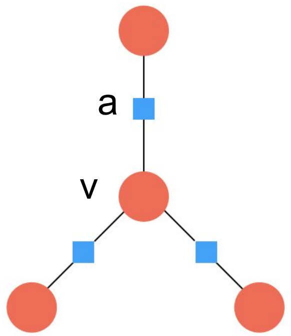
$$\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\sum_{x_3}\sum_{x_4}\psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi_{12}(x_1, x_2)\sum_{x_3}\psi_{23}(x_2, x_3)\sum_{x_4}\psi_{34}(x_3, x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi_{12}(x_1, x_2)\sum_{x_3}\psi_{23}(x_2, x_3)m_4(x_3) \\
&= \frac{1}{Z}\sum_{x_2}\psi_{12}(x_1, x_2)m_3(x_2) = \frac{1}{Z}m_2(x_1)
\end{aligned}$$

Belief propagation

- In each iteration, each message will be updated iteratively
- At convergence, the marginal distribution of each node is proportional to the product of all messages from adjoining factors

$$p_{X_v}(x_v) \propto \prod_{a \in N(v)} \mu_{a \to v}(x_v)$$

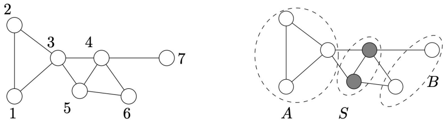
$$p_{X_a}(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod_{v \in N(a)} \mu_{v \to a}(x_v)$$



Belief propagation on trees

- Tree (graph is oriented, one node is root)
- Convergence after two full passes
    - 1. Messages are send inwards: from leaves to root. (the tree structure guarantee that all messages are calculated) (forward algorithm)
    - 2. Messages are passing back from the root to the leaves. (backward algorithm)

- Large graphs: build a clique graph (there exists link between all pair of nodes - mega nodes in the subset) - clique trees



Belief propagation on loopy networks

- It doesn’t converge
- Not well understood

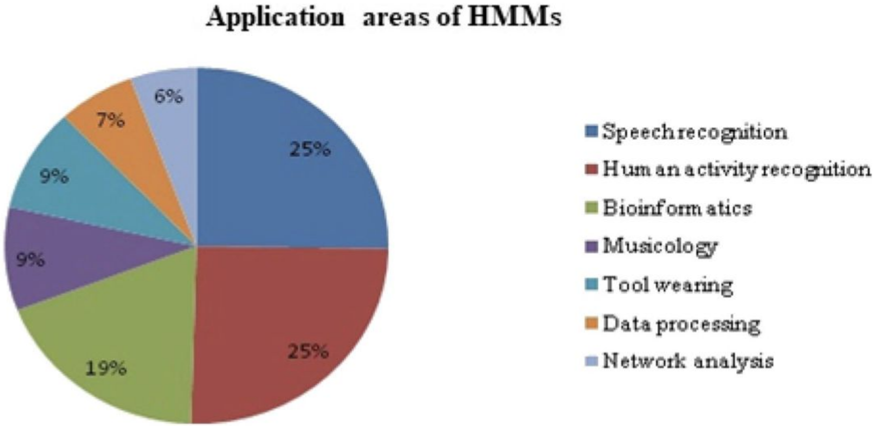
Summary

Belief propagation - message passing: predicting node's state using data and underlying graphical models.

$$\mathcal{O}+\lambda \;\Big\downarrow\; P(\mathcal{O}\,|\,\lambda) \qquad \mathcal{O}+\lambda \;\Big\downarrow\; \mathrm{X} \qquad \mathcal{O}+\mathrm{N,M} \;\Big\downarrow\; \lambda$$

HMM - well understood belief propagation on a simple chain graph

For complex graph topologies - graph factorization



References

Mark Stamp Lecture notes https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf

Bhavya Mor, Sunita Garhwal, Ajay Kumar: A Systematic Review of Hidden Markov Models and Their Applications

Kevin P. Murphy: Probabilistic Machine Learning

R. L. Cave and L. P. Neuwirth, Hidden Markov models for English, in J. D. Ferguson, editor, Hidden Markov Models for Speech, IDA-CRD, Princeton, NJ, October 1980. https://www.cs.sjsu.edu/ ~ stamp/RUA/CaveNeuwirth/index.html

Wikipidia: Belief propagation

Martin Wainwright
https://yosinski.com/mlss12/media/slides/MLSS-2012-Wainwright-Graphical-Models-and-Message-Passing.pdf

Yunshu Liu
https://faculty.coe.drexel.edu/jwalsh/Yunshu_GPnBP.pdf

Antonio Torralba (UCLA)

http://helper.ipam.ucla.edu/publications/gss2013/gss2013_11344.pdf

James Coughlan

https://www.ski.org/sites/default/files/publications/bptutorial.pdf

Francis Ferraro

https://redirect.cs.umbc.edu/~ferraro/teaching/691-s20/slides/11-bp.pdf

Kayhan Batmanghelich

https://www.batman-lab.com/wp-content/uploads/2020/09/scribe_le14.pdf

Frank R. Kschischang (MIT)

http://www.mit.edu/~6.454/www_fall_2002/lizhong/bp.pdf

Jonathan Yedidia

http://people.csail.mit.edu/billf/publications/Understanding_Belief_Propogation.pdf

Yair Weiss

https://courses.cs.washington.edu/courses/cse577/04sp/notes/uaitut.pdf

Robert Collins (Stanford)

https://www.cse.psu.edu/~rtc12/CSE586/lectures/cse586GMplusMP_6pp.pdf

Dan Jurafsky (Stanford)

https://web.stanford.edu/~jurafsky/slp3/A.pdf

Ben Langmead

https://www.cs.jhu.edu/~langmea/resources/lecture_notes/hidden_markov_models.pdf

Venu Govindaraju

https://cse.buffalo.edu/~jcorso/t/CSE555/files/lecture_hmm.pdf

Andrew M. Moore

https://www.cs.cmu.edu/~./awm/tutorials/hmm14.pdf

Sean Eddy

https://www.nature.com/articles/nbt1004-1315

Alperen Degirmenci

https://scholar.harvard.edu/files/adegirmenci/files/hmm_adegirmenci_2014.pdf