Biostatistics-Lecture 4

Ruibin Xi
Peking University
School of Mathematical Sciences

Markov Models

- A discrete (finite) system:
  - N distinct states.
  - Begins (at time t=1) in some initial state(s).
  - At each time step (t=1,2,…) the system moves from current to next state (possibly the same as the current state) according to transition probabilities associated with current state.
- This kind of system is called a finite, or discrete Markov model

- **After [Andrei Andreyevich Markov](#) (1856 -1922)**

Outline

- Markov Chains (Markov Models)
- Hidden Markov Chains (HMMs)
- Algorithmic Questions
- Biological Relevance

Discrete Markov Model: Example

- Discrete Markov Model with 5 states.
- Each a_{ij} represents the probability of moving from state i to state j
- The a_{ij} are given in a matrix $A = \{a_{ij}\}$
- The probability to start in a given state i is π_i, The vector π repre-sents these start probabilities.

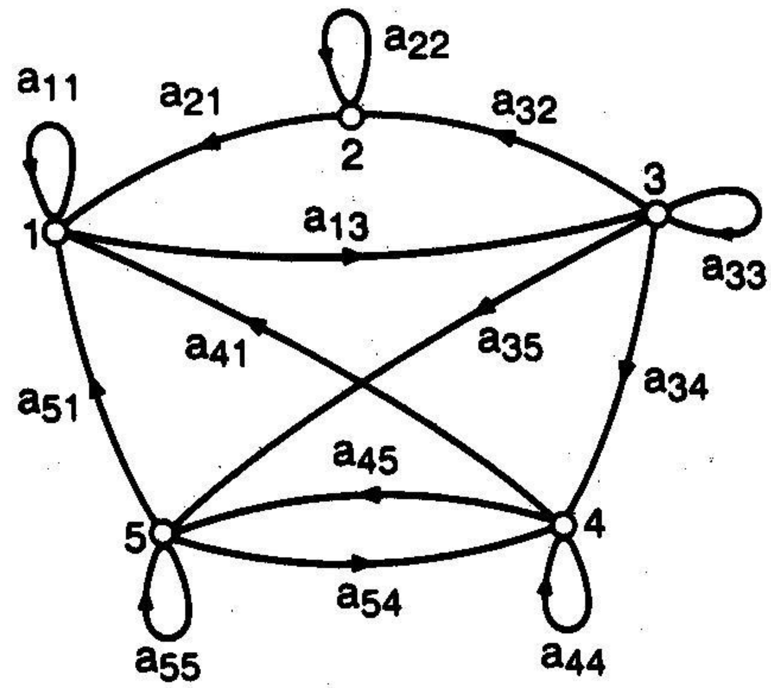
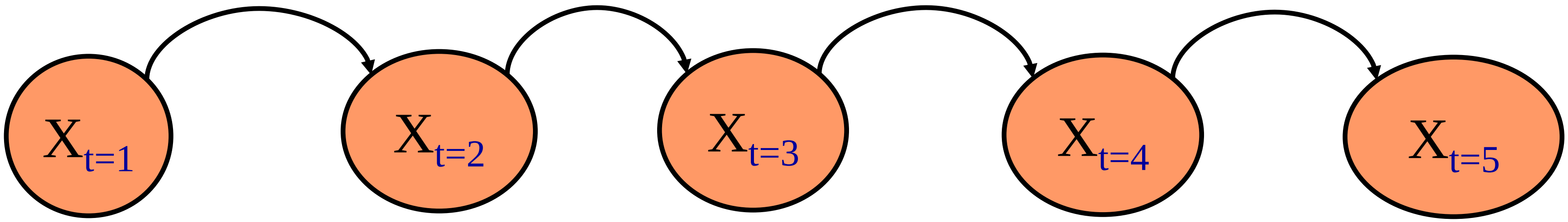


Figure 6.1 A Markov chain with five states (labeled 1 to 5) with selected state transitions.

Markov Property

- **Markov Property: The state of the system at time t+1 depends only on the state of the system at time t**

$$P[X_{t+1} = x_{t+1} \mid X_t = x_t , X_{t-1} = x_{t-1} , \ldots , X_1 = x_1 , X_0 = x_0]$$
$$= P[X_{t+1} = x_{t+1} \mid X_t = x_t]$$



Markov Chains

Stationarity Assumption

Probabilities independent of t when process is "stationary"

So, $$\text{for all } t,\ P[X_{t+1} = x_j | X_t = x_i] = p_{ij}$$

This means that if system is in state i, the probability that the system will next move to state j is p_{ij} , no matter what the value of t is

Simple Minded Weather Example

- **raining today** ⇨ **rain tomorrow** **$p_{rr} = 0.4$**
- **raining today** ⇨ **no rain tomorrow** **$p_{rn} = 0.6$**
- **no raining today** ⇨ **rain tomorrow** **$p_{nr} = 0.2$**
- **no raining today** ⇨ **no rain tomorrow** **$p_{rr} = 0.8$**

Simple Minded Weather Example

Transition matrix for our example

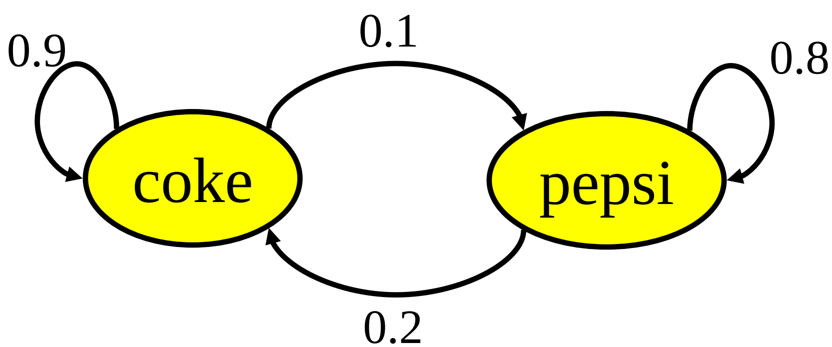
$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}$$

- **Note that rows sum to 1**
- **Such a matrix is called a Stochastic Matrix**

Coke vs. Pepsi

Given that a person's last cola purchase was Coke ™, there is a 90% chance that her next cola purchase will also be Coke ™.

If that person's last cola purchase was Pepsi™, there is an 80% chance that her next cola purchase will also be Pepsi™.



Coke vs. Pepsi

Given that a person is currently a Pepsi purchaser, what is the probability that she will purchase Coke two purchases from now?

The transition matrices are:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

(corresponding to one purchase ahead)

$$P^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

Coke vs. Pepsi

Given that a person is currently a Coke drinker, what is the probability that she will purchase Pepsi three purchases from now?

$$P^3 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix} = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

Coke vs. Pepsi

Assume each person makes one cola purchase per week. Suppose 60% of all people now drink Coke, and 40% drink Pepsi.

What fraction of people will be drinking Coke three weeks from now?

Let $(Q_0,Q_1)=(0.6,0.4)$ be the initial probabilities.

P_{00}

We will regard Coke as 0 and Pepsi as 1

$$P=\begin{bmatrix}0.9 & 0.1\\ 0.2 & 0.8\end{bmatrix}$$

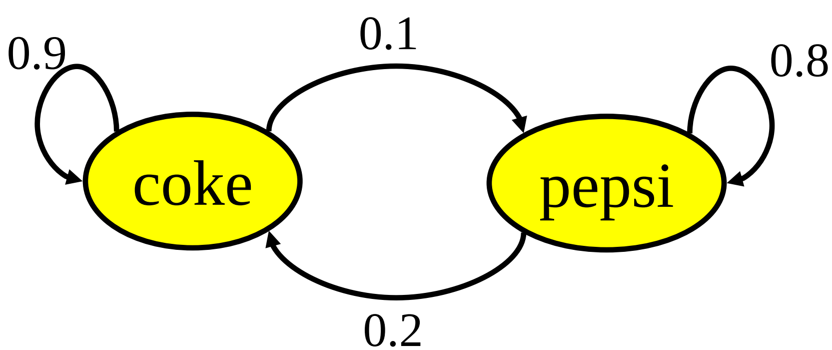
We want to find $P(X_3=0)$

$$P(X_3=0)=\sum_{i=0}^{1}Q_i p_{i0}^{(3)}=Q_0 p_{00}^{(3)}+Q_1 p_{10}^{(3)}=0.6\cdot 0.781+0.4\cdot 0.438=0.6438$$

Equilibrium (Stationary) Distribution

- Suppose 60% of all people now drink Coke, and 40% drink Pepsi. What fraction will be drinking Coke 10,100,1000,10000 … weeks from now?
- For each week, probability is well defined. But does it converge to some equilibrium distribution $[p_0,p_1]$?
- If it does, then eqs. : $.9p_0+.2p_1 = p_0$, $.8p_1+.1p_0 = p_1$

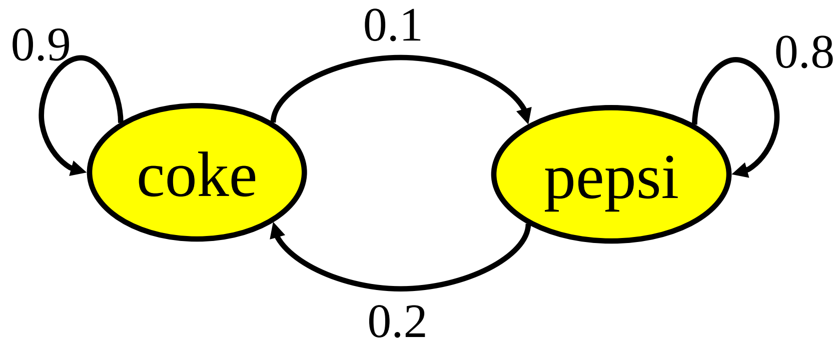
must hold, yielding $p_0 = 2/3$, $p_1 = 1/3$.



Equilibrium (Stationary) Distribution

Whether or not there is a stationary distribution, and whether or not it is unique if it does exist, are determined by certain properties of the process.

- *Irreducible* means that every state is accessible from every other state.
- *Aperiodic* means that any state return to itself can occur at irregular times.
- *Positive recurrent* means that the expected return time is finite for every state.



Equilibrium (Stationary) Distribution

- If the Markov chain is positive recurrent, there exists a stationary distribution. If it is positive recurrent and irreducible, there exists a unique stationary distribution, and furthermore the process constructed by taking the stationary distribution as the initial distribution is ergodic. Then the average of a function *f* over samples of the Markov chain is equal to the average with respect to the stationary distribution,

Equilibrium (Stationary) Distribution

- Writing P for the transition matrix, a stationary distribution is a vector π which satisfies the equation
  - $P\pi = \pi$.
- In this case, the stationary distribution π is an eigenvector of the transition matrix, associated with the eigenvalue 1.

Discrete Markov Model - Example

- States – Rainy:1, Cloudy:2, Sunny:3
- Matrix A – $$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}.$$
- Problem – given that the weather on day 1 (t=1) is sunny(3), what is the probability for the observation O:

$$\begin{array}{lcccccccccc} \mathbf{O} & = & (\text{sunny}, & \text{sunny}, & \text{sunny}, & \text{rain}, & \text{rain}, & \text{sunny}, & \text{cloudy}, & \text{sunny} &) \\ & = & (3, & 3, & 3, & 1, & 1, & 3, & 2, & 3 &) \\ \text{day} & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & \end{array}$$

Discrete Markov Model – Example (cont.)

- The answer is -

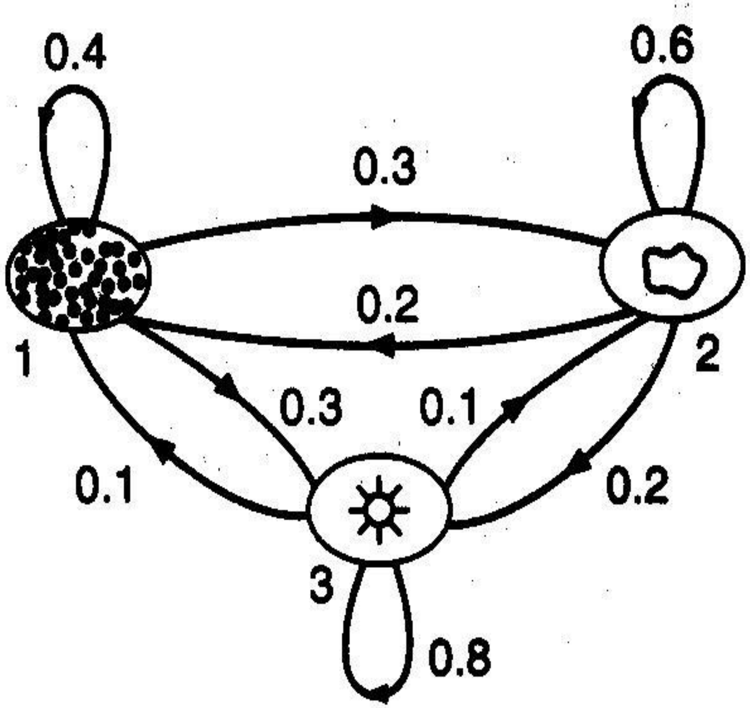


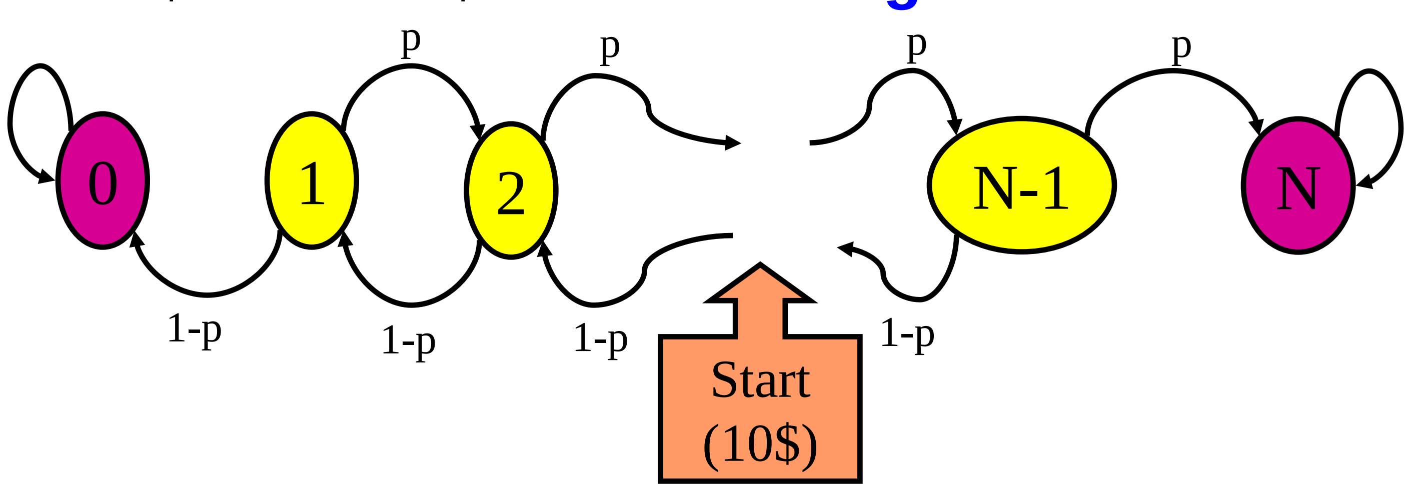
Figure 6.2 Markov model of the weather.

$$
\begin{aligned}
P(\mathbf{O}|\text{Model}) &= P[3, 3, 3, 1, 1, 3, 2, 3|\text{Model}] \\
&= P[3]\,P[3|3]^2 P[1|3]\,P[1|1] \\
&\quad P[3|1]\,P[2|3]\,P[3|2] \\
&= \pi_3 \cdot (a_{33})^2\, a_{31}\, a_{11}\, a_{13}\, a_{32}\, a_{23} \\
&= (1.0)(0.8)^2(0.1)(0.4)(0.3)(0.1)(0.2) \\
&= 1.536 \times 10^{-4}
\end{aligned}
$$

Third Example: A Friendly Gambler

Game starts with 10$ in gambler's pocket

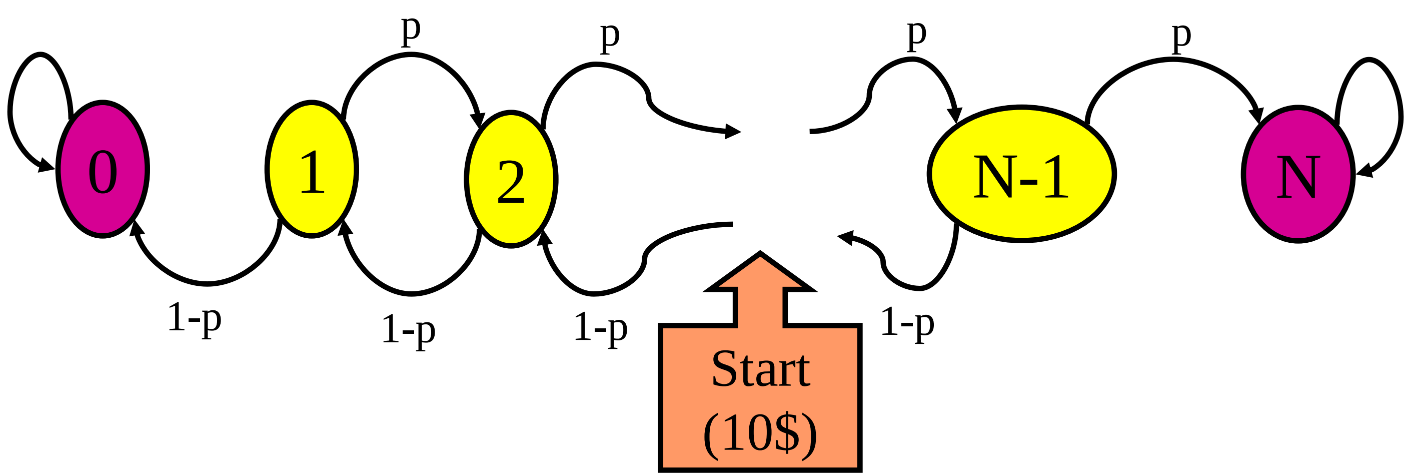
- **At each round we have the following:**
  - or
    - **Gambler wins 1$ with probability p**
    - **Gambler loses 1$ with probability 1-p**
- **Game ends when gambler goes broke (no sister in bank), or accumulates a capital of 100$ (including initial capital)**
- **Both 0$ and 100$ are absorbing states**



Fourth Example: A Friendly Gambler

Irreducible means that every state is accessible from every other state.
Aperiodic means that any state return to itself can occur at irregular times.
Positive recurrent means that the expected return time is finite for every state. If the Markov chain is positive recurrent, there exists a stationary distribution.

Is the gambler's chain positive recurrent? Does it have a stationary distribution (independent upon initial distribution)?



Markov Chains

See
http://www.statslab.cam.ac.uk/~james/Markov/
Or
http://probability.ca/MT/
for more detail

Markov Chain A Simple Example

- States: fair coin *F*, unfair (biased) coin *B*
- Discrete times: flip 1, 2, 3, …
- Initial probability: $\pi_F = 0.6$, $\pi_B = 0.4$
- Transition probability

- Prob(FFBBFFFB)

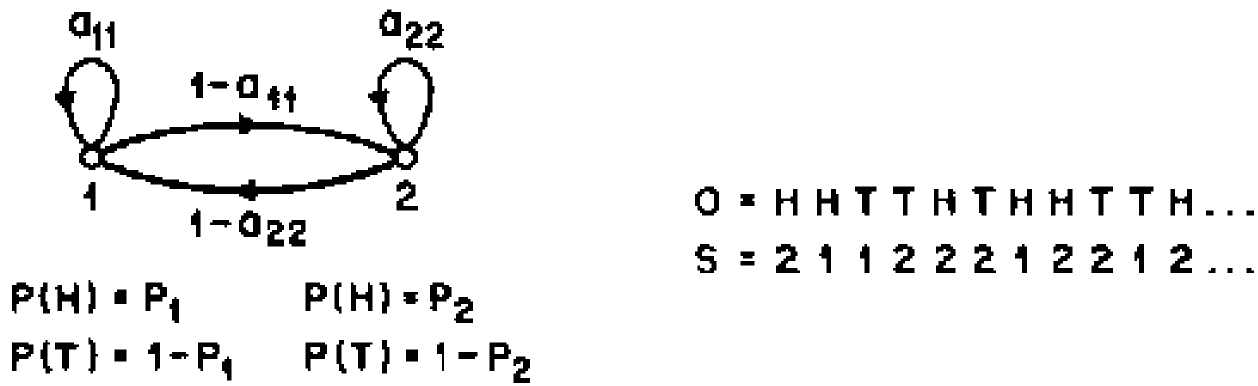
$$P(O = \{FFBBFFFB\} \mid A, \pi)$$

$$= \pi_F \cdot a_{FF} \cdot a_{FB} \cdot a_{BB} \cdot a_{BF} \cdot a_{FF} \cdot a_{FF} \cdot a_{FB}$$

$$= 0.6 \times 0.9 \times 0.1 \times 0.7 \times 0.3 \times 0.9 \times 0.9 \times 0.1 = 9.19 \times 10^{-4}$$

Hidden Markov Model

- Coin toss example



- Coin transition is a Markov chain
- Probability of H/T depends on the coin used
- Observation of H/T is a hidden Markov chain (coin state is hidden)

Hidden Markov Model

- Elements of an HMM (coin toss)
  - N, the number of states (F / B)
  - M, the number of distinct observation (H / T)
  - A = {a_{ij}} state transition probability
    $$A = \begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{bmatrix}$$
  - B = {$b_i(k)$} emission probability
    $$b_F(H) = 0.5, b_F(T) = 0.5$$
    $$b_B(H) = 0.8, b_B(T) = 0.2$$
  - π ={π_i} initial state distribution
    - $\pi_F = 0.4$, $\pi_B = 0.6$

λ

HMM Applications

- Stock market: bull/bear market hidden Markov chain, stock daily up/down observed, depends on big market trend
- Speech recognition: sentences & words hidden Markov chain, spoken sound observed (heard), depends on the words
- Digital signal processing: source signal (0/1) hidden Markov chain, arrival signal fluctuation observed, depends on source
- Bioinformatics: sequence motif finding, gene prediction, genome copy number change, protein structure prediction, protein-DNA interaction prediction

Basic Problems for HMM

1. Given λ, how to compute $P(O|\lambda)$ observing sequence $O = O_1O_2 \ldots O_T$
   - Probability of observing HTTHHHT …
   - Forward procedure, backward procedure
2. Given observation sequence $O = O_1O_2 \ldots O_T$ and λ, how to choose state sequence $Q = q_1q_2 \ldots q_t$
   - What is the hidden coin behind each flip
   - Forward-backward, Viterbi
3. How to estimate $\lambda = (A,B,\pi)$ so as to maximize $P(O|\lambda)$
   - How to estimate coin parameters λ
   - Baum-Welch (Expectation maximization)

Problem 1: P(O|λ)

- Suppose we know the state sequence Q

$$P(O \mid Q, \lambda) = b_{q1}(O_1) b_{q2}(O_2) \ldots b_{qT}(O_T)$$

  - O = H T T H H H T
  - Q = F F B F F B B

$$P(O \mid Q, \lambda) = b_F(H) b_F(T) b_B(T) b_F(H) b_F(H) b_B(H) b_B(T)$$

$$= 0.5 \times 0.5 \times 0.2 \times 0.5 \times 0.5 \times 0.8 \times 0.2$$

  - Q = B F B F B B B

$$P(O \mid Q, \lambda) = b_B(H) b_F(T) b_B(T) b_F(H) b_B(H) b_B(H) b_B(T)$$

$$= 0.8 \times 0.5 \times 0.2 \times 0.5 \times 0.8 \times 0.8 \times 0.2$$

- Each given path Q has a probability for O

Problem 1: P(O|λ)

- What is the prob of this path Q?

  $$P(Q \mid \lambda)$$

  – Q = F F B F F B B

  $$P(Q \mid \lambda) = \pi_F \; a_{FF} \; a_{FB} \; a_{BF} \; a_{FF} \; a_{FB} \; a_{BB}$$
  $$= 0.6 \times 0.9 \times 0.1 \times 0.3 \times 0.9 \times 0.1 \times 0.7$$

  – Q = B F B F B B B

  $$P(Q \mid \lambda) = \pi_B \; a_{BF} \; a_{FB} \; a_{BF} \; a_{FB} \; a_{BB} \; a_{BB}$$
  $$= 0.4 \times 0.3 \times 0.1 \times 0.3 \times 0.1 \times 0.7 \times 0.7$$

- Each given path Q has its own probability

Problem 1: $P(O|\lambda)$

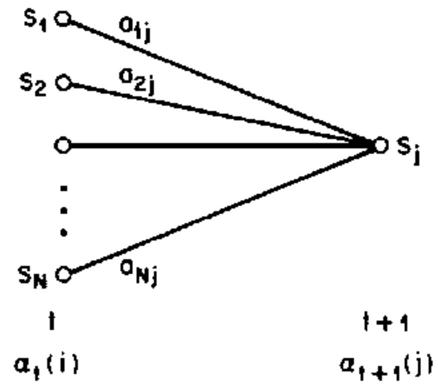
- Therefore, total pb of O = HTTHHHT
- Sum over all possible paths Q: each Q with its own pb multiplied by the pb of O given Q

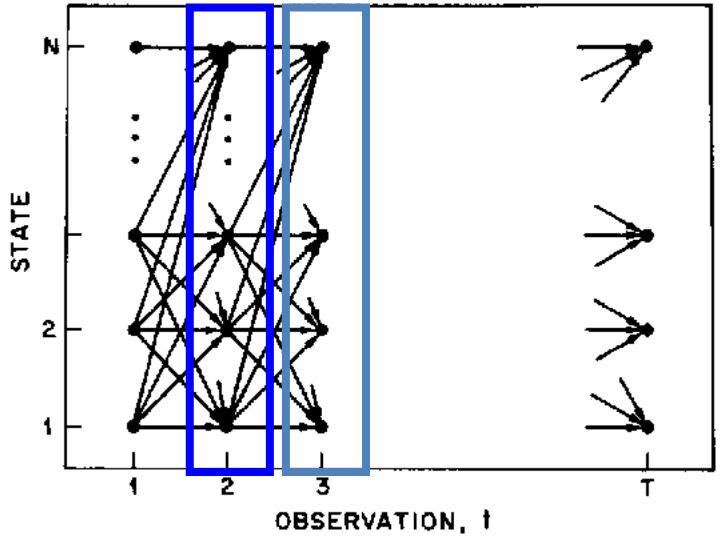
$$P(O \mid \lambda) = \sum_{all\,Q} P(O, Q \mid \lambda) = \sum_{all\,Q} P(Q \mid \lambda) P(O \mid Q, \lambda)$$

- For path of T long and N hidden states, there are N^T paths, unfeasible calculation

Solution to Prob1: Forward Procedure

- Use dynamic programming
- Summing at every time point
- Keep previous subproblem solution to speed up current calculation





Forward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\alpha_1(i) = \pi_i b_i(O_1)$

$$\alpha_1(F) = \pi_F b_F(H) = 0.4 \times 0.5 = 0.2$$

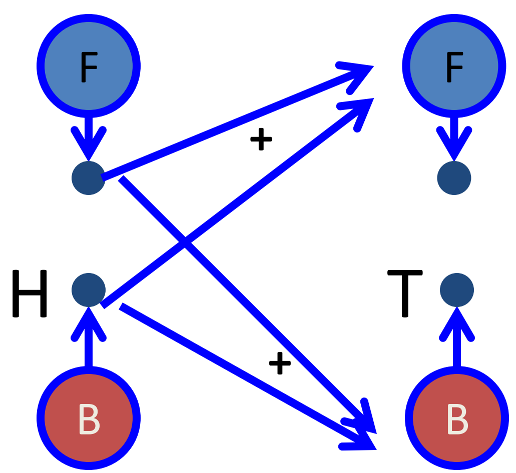
$$\alpha_1(B) = \pi_B b_B(H) = 0.6 \times 0.8 = 0.48$$

  – Pb of seeing H_1 from F_1 or B_1

Forward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\alpha_1(i) = \pi_i b_i(O_1)$
- Induction $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1})$

  – Pb of seeing T_2 from F_2 or B_2



F_2 could come from F_1 or B_1

Each has its pb, add them up

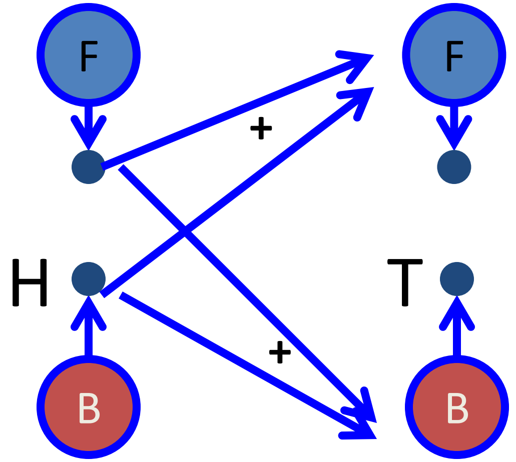
T H …

Forward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\alpha_1(i) = \pi_i b_i(O_1)$
- Induction $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}]\, b_j(O_{t+1})$

$$\alpha_2(F) = (\alpha_1(F)a_{FF} + \alpha_1(B)a_{BF})b_F(T) = (0.2 \times 0.9 + 0.48 \times 0.3) \times 0.5 = 0.162$$

$$\alpha_2(B) = (\alpha_1(F)a_{FB} + \alpha_1(B)a_{BB})b_B(T) = (0.2 \times 0.1 + 0.48 \times 0.7) \times 0.2 = 0.0712$$



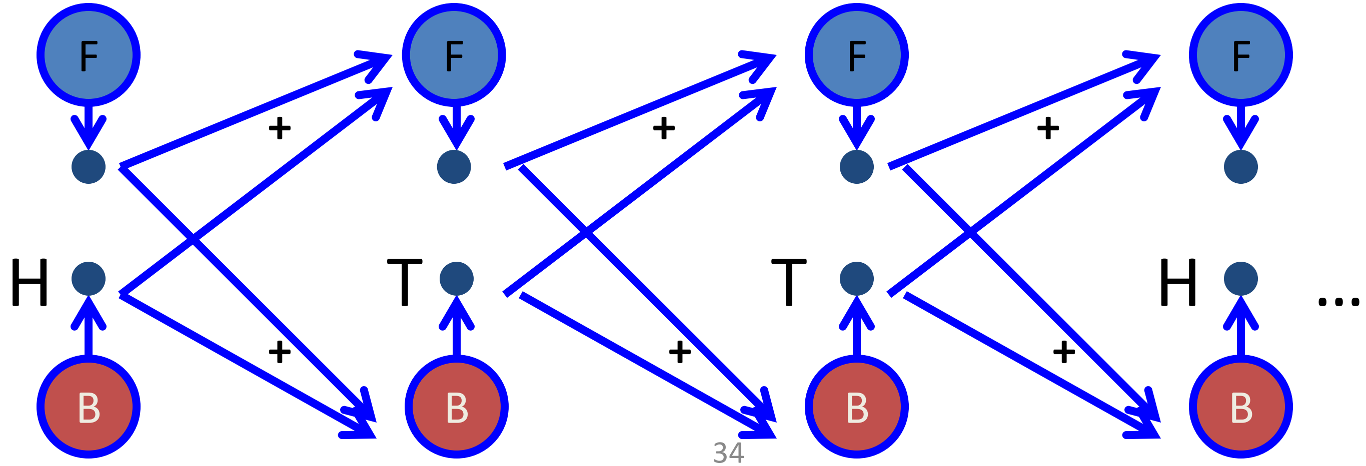
T H ...

Forward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\alpha_1(i) = \pi_i b_i(O_1)$
- Induction $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}]\, b_j(O_{t+1})$

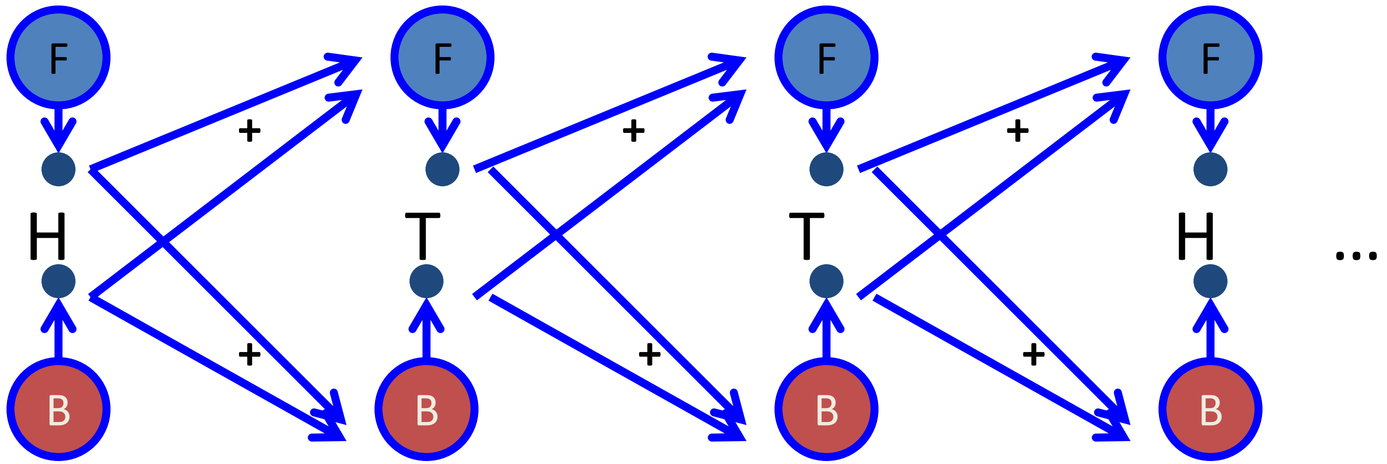
$$\alpha_3(F) = [0.162 \times 0.9 + 0.0712 \times 0.3] \times 0.5 = 0.08358$$

$$\alpha_3(B) = [0.162 \times 0.1 + 0.0712 \times 0.7] \times 0.2 = 0.013208$$



Forward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\alpha_1(i) = \pi_i b_i(O_1)$
- Induction $\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1})$
- Termination $P(O|\lambda) = \sum_{i=1}^{N} \alpha_t(i) \qquad = \alpha_4(F) + \alpha_4(B)$



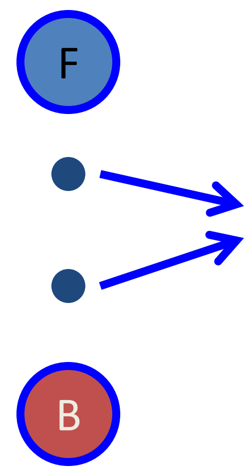
Solution to Prob1: Backward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\beta_{T*}(i) = 1$

$$\beta_{T*}(F) = \beta_{T*}(B) = 1$$

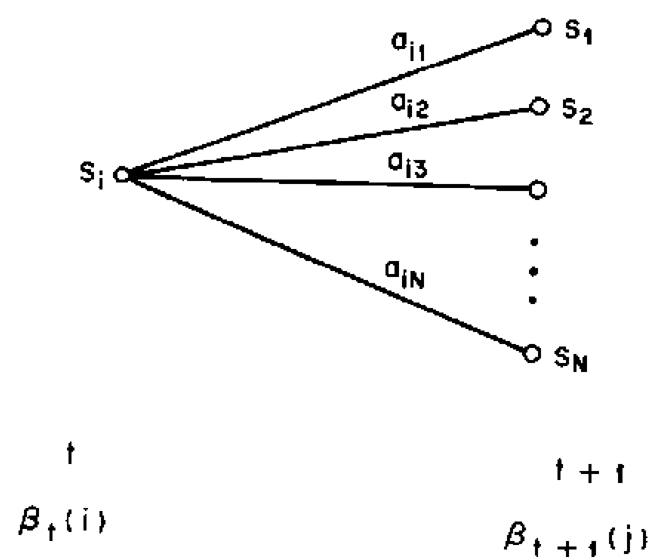
- Pb of coin to see certain flip after it

...H H H T



Backward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\beta_{T*}(i) = 1$
- Induction $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$
- Pb of coin to see certain flip after it

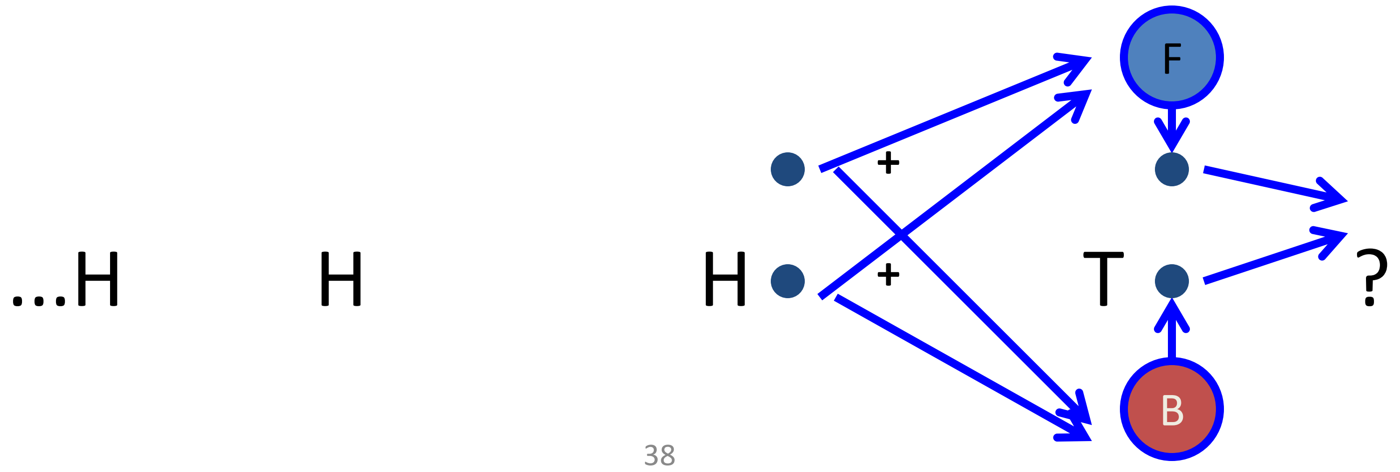


Backward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\beta_{T*}(i) = 1$
- Induction $\beta_t(i) = \sum_{i=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$

$$\beta_{T-1}(F) = a_{FF} b_F(T) \times 1 + a_{FB} b_B(T) \times 1 = 0.9 \times 0.5 + 0.1 \times 0.2 = 0.47$$

$$\beta_{T-1}(B) = a_{BF} b_F(T) \times 1 + a_{BB} b_B(T) \times 1 = 0.3 \times 0.5 + 0.7 \times 0.2 = 0.29$$

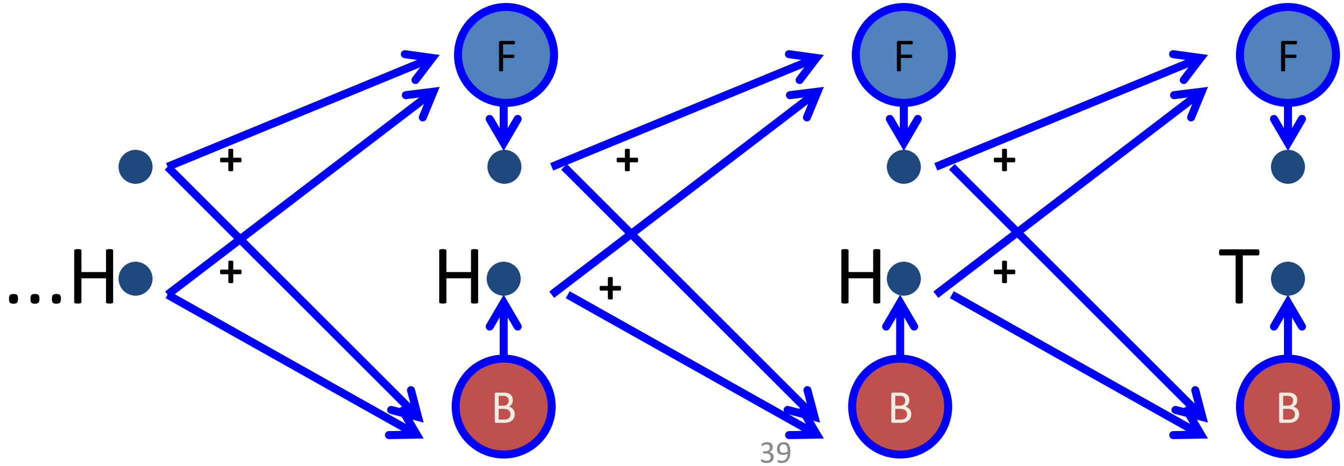


Backward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\beta_{T*}(i) = 1$
- Induction $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$

$$\beta_{T-2}(F) = a_{FF} b_F(H) \times \beta_{T-1}(F) + a_{FB} b_B(H) \times \beta_{T-1}(B) = 0.9 \times 0.5 \times 0.47 + 0.1 \times 0.8 \times 0.29 = 0.2347$$

$$\beta_{T-2}(B) = a_{BF} b_F(H) \times \beta_{T-1}(F) + a_{BB} b_B(H) \times \beta_{T-1}(B) = 0.3 \times 0.5 \times 0.47 + 0.7 \times 0.8 \times 0.29 = 0.2329$$



Backward Procedure

- Coin toss, O = HTTHHHT
- Initialization $\beta_{T*}(i) = 1$
- Induction $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$
- Termination $\beta_0(*) = \pi_F b_F(H) \beta_1(F) + \pi_B b_B(H) \beta_1(B)$
- Both forward and backward could be used to solve problem 1, which should give identical results

Solution to Problem 2
Forward-Backward Procedure

- First run forward and backward separately
- Keep track of the scores at every point
- Coin toss
  - α: pb of this coin for seeing all the flips now and before
  - β: pb of this coin for seeing all the flips after

H	T	T	H	H	H	T
$\alpha_1(F)$	$\alpha_2(F)$	$\alpha_3(F)$	$\alpha_4(F)$	$\alpha_5(F)$	$\alpha_6(F)$	$\alpha_7(F)$
$\alpha_1(B)$	$\alpha_2(B)$	$\alpha_3(B)$	$\alpha_4(B)$	$\alpha_5(B)$	$\alpha_6(B)$	$\alpha_7(B)$
$\beta_1(F)$	$\beta_2(F)$	$\beta_3(F)$	$\beta_4(F)$	$\beta_5(F)$	$\beta_6(F)$	$\beta_7(F)$
$\beta_1(B)$	$\beta_2(B)$	$\beta_3(B)$	$\beta_4(B)$	$\beta_5(B)$	$\beta_6(B)$	$\beta_7(B)$

Solution to Problem 2
Forward-Backward Procedure

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

- Coin toss

$$\gamma_3(F) = \frac{\alpha_3(F)\beta_3(F)}{\alpha_3(F)\beta_3(F) + \alpha_3(B)\beta_3(B)}$$

$$e.g. = \frac{0.0025 \times 0.0047}{0.0025 \times 0.0047 + 0.0016 \times 0.0038} = 0.659$$

- Gives probabilistic prediction at every time point
- Forward-backward maximizes the expected number of correctly predicted states (coins)

Solution to Problem 2
Viterbi Algorithm

- Report the path that is most likely to give the observations

- Initiation
$$\delta_1(i) = \pi_i b_i(O_1)$$
$$\psi_1(i) = 0$$

- Recursion
$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$$
$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]$$

- Termination
$$P^* = \max_{1\le i\le N}[\delta_T(i)]$$
$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)]$$

- Path (state sequence) backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, \ldots, 1$$

Viterbi Algorithm

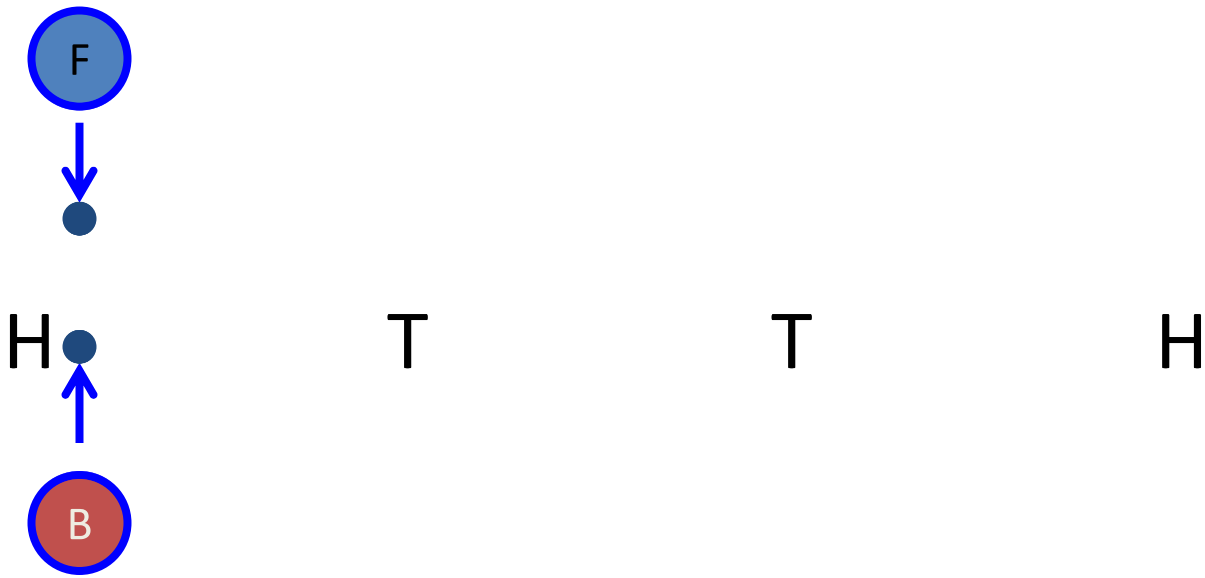
- Observe: HTTHHHT
- Initiation $\delta_1(i) = \pi_i b_i(O_1)$

  $\psi_1(i) = 0$

$$\delta_1(F) = 0.4 \times 0.5 = 0.2$$

$$\delta_1(B) = 0.6 \times 0.8 = 0.48$$

Viterbi Algorithm



Viterbi Algorithm

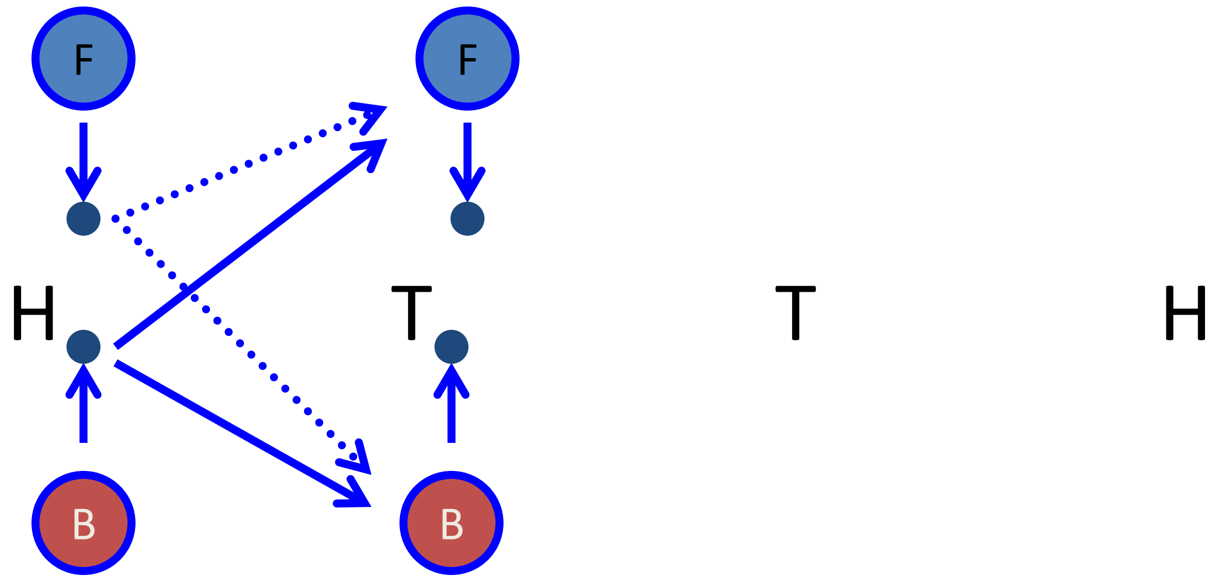
- Observe: HTTHHHT
- Initiation
$$\delta_1(i) = \pi_i b_i(O_1)$$
$$\psi_1(i) = 0$$
- Recursion
$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$$
$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]$$

Max instead of +, keep track path

$$\delta_2(F) = [\max_i(\delta_1(F)a_{FF}, \delta_1(B)a_{BF})]b_F(T) = \max(\boxed{0.18}, 0.144) \times 0.5 = 0.09$$
$$\delta_2(B) = [\max_i(\delta_1(F)a_{FB}, \delta_1(B)a_{BB})]b_B(T) = \max(0.02, \boxed{0.336}) \times 0.2 = 0.0672$$
$$\varphi_2(F) = F, \varphi_2(B) = B$$

Viterbi Algorithm

- Max instead of +, keep track of path
- Best path (instead of all path) up to here



Viterbi Algorithm

- Observe: HTTHHHT
- Initiation

$$\delta_1(i) = \pi_i b_i(O_1)$$
$$\psi_1(i) = 0$$

- Recursion

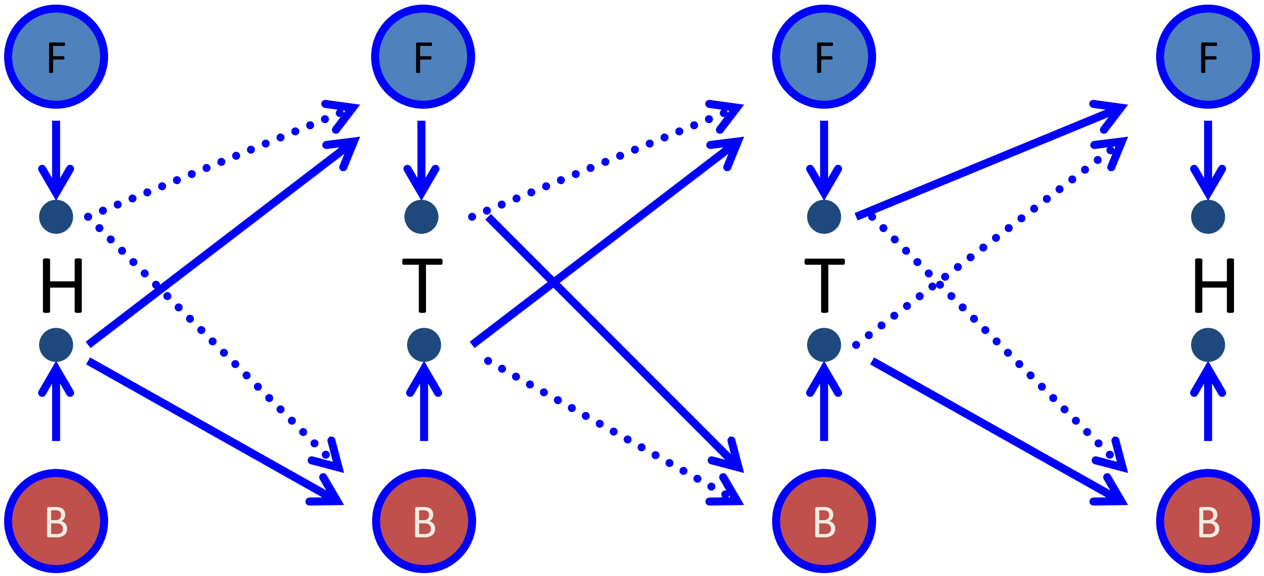
$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t)$$
$$\psi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]$$

Max instead of +, keep track path

$$\delta_3(F) = [\max_i(\delta_2(F)a_{FF}, \delta_2(B)a_{BF})]b_F(T)$$
$$= \max(\boxed{0.09\times 0.9}, 0.0672\times 0.3)\times 0.5 = 0.0405$$
$$\delta_3(B) = [\max_i(\delta_2(F)a_{FB}, \delta_2(B)a_{BB})]b_B(T)$$
$$= \max(0.09\times 0.1, \boxed{0.0672\times 0.7})\times 0.2 = 0.0094$$
$$\varphi_3(F) = F, \varphi_3(B) = B$$

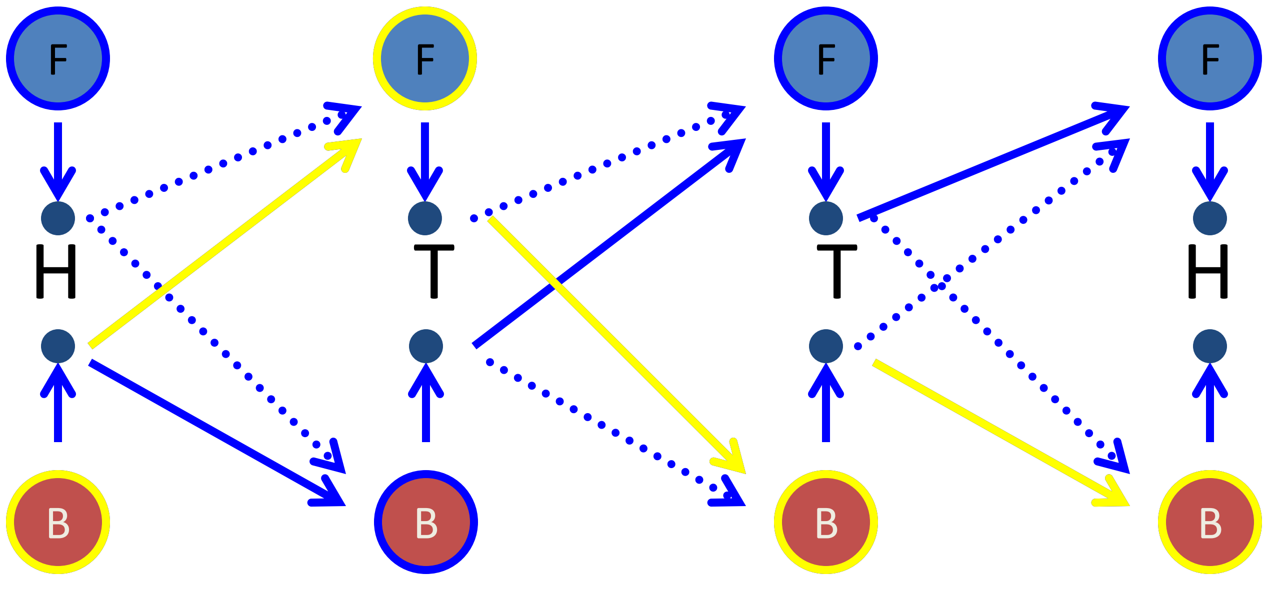
Viterbi Algorithm

- Max instead of +, keep track of path
- Best path (instead of all path) up to here



Viterbi Algorithm

- Terminate, pick state that gives final best δ score, and backtrack to get path



- BFBB most likely to give HTTH

Solution to Problem 3

- No optimal way to do this, so find local maximum
- Baum-Welch algorithm (equivalent to expectation-maximization)
  - Random initialize λ =(A,B,π)

  - Update λ =(A,B,π)
    - π: % of F vs B on Viterbi path
    - A: frequency of F/B transition on Viterbi path
    - B: frequency of H/T emitted by F/B