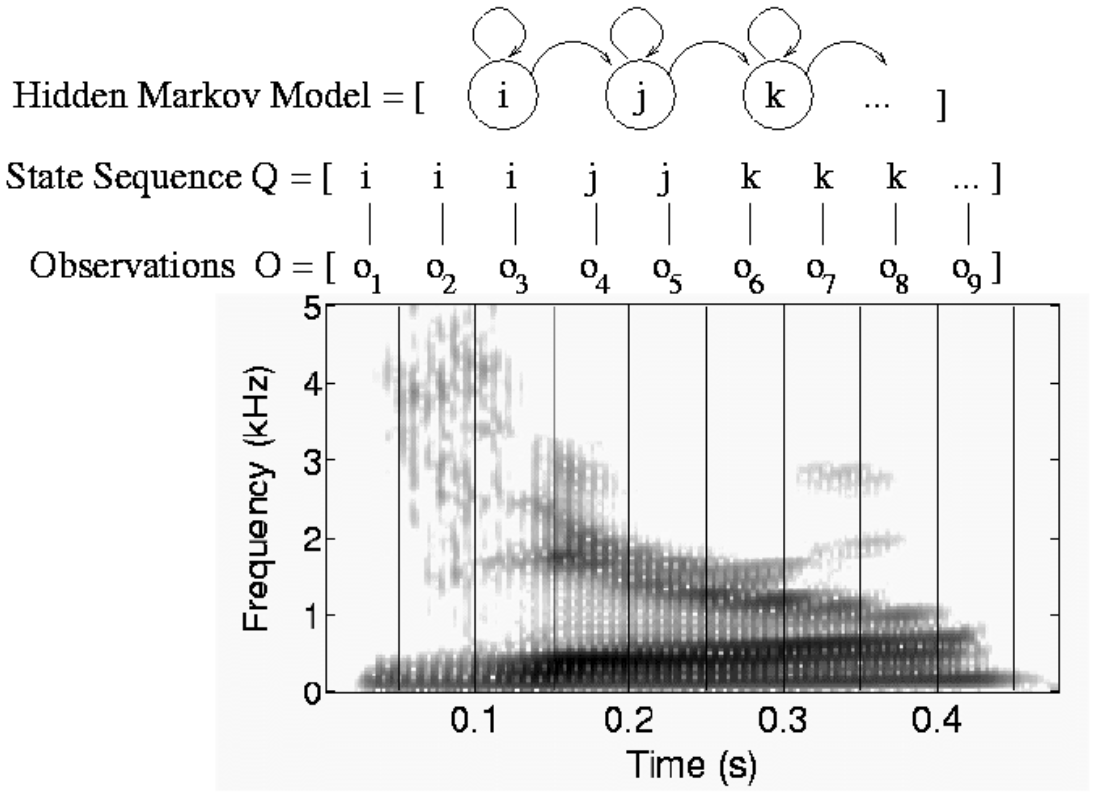
CS440/ECE448 Lecture 15: Hidden Markov Models

Mark Hasegawa-Johnson, 3/2021



Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

Probabilistic reasoning over time

- So far, we've mostly dealt with *episodic* environments
  - Exceptions: games with multiple moves, planning
- In particular, the Bayesian networks we've seen so far describe static situations
  - Each random variable gets a single fixed value in a single problem instance
- Now we consider the problem of describing probabilistic environments that evolve over time
  - Examples: robot localization, human activity detection, tracking, speech recognition, machine translation,

Probabilistic reasoning over time

- At each time slice *t*, the state of the world is described by an unobservable **state variable** Y_t and an observable **observation variable** X_t
- **State Transitions:** in general, the value of Y_t depends on the whole past history:

$$P(Y_t \mid Y_0, \ldots, Y_{t-1}) = P(Y_t \mid \mathbf{Y}_{0:t-1})$$

- **Observation model:** in general, the value of X_t depends on all current and past states and observations:

$$P(X_t \mid Y_0, \ldots, Y_t, X_1, \ldots, X_{t-1}) = P(X_t \mid \mathbf{Y}_{0:t}, \mathbf{X}_{1:t-1})$$

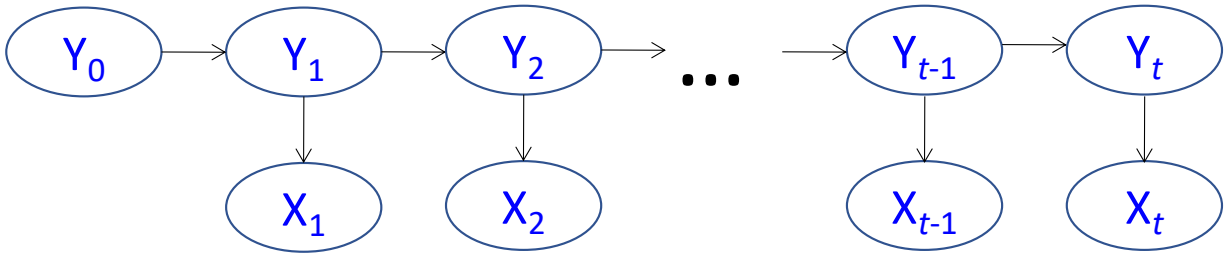
Hidden Markov Model

- A hidden Markov model assumes that both the state and the observation are Markov.
- **State Transitions:** the Markov assumption means that each state variable depends only on the preceding time step:

$$P(Y_t \mid Y_0, \ldots, Y_{t-1}) = P(Y_t \mid Y_{t-1})$$

- **Observation model:** the Markov assumption means that each state variable depends only on the current state:

$$P(X_t \mid Y_0, \ldots, Y_t, X_1, \ldots, X_{t-1}) = P(X_t \mid Y_t)$$



Example Scenario: UmbrellaWorld

Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

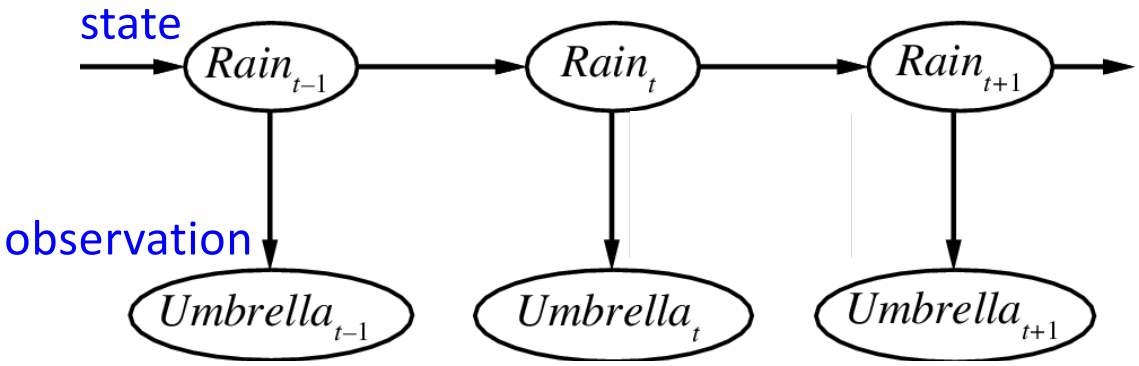
- Elspeth Dunsany is an AI researcher at the Canadian company Unitek.
- Richard Feynman is an AI, named after the famous physicist, whose personality he resembles.
- To keep him from escaping, Richard's workstation is not connected to the internet. He knows about rain but has never seen it.
- He has noticed, however, that Elspeth sometimes brings an umbrella to work. He correctly infers that she is more likely to carry an umbrella on days when it rains.

Example Scenario: UmbrellaWorld

Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

Since he has read a lot about rain, Richard proposes a hidden Markov model:

- Rain on day t-1 (R_{t-1}) makes rain on day t (R_t) more likely.
- Elspeth usually brings her umbrella (U_t) on days when it rains (R_t), but not always.



Example Scenario: UmbrellaWorld

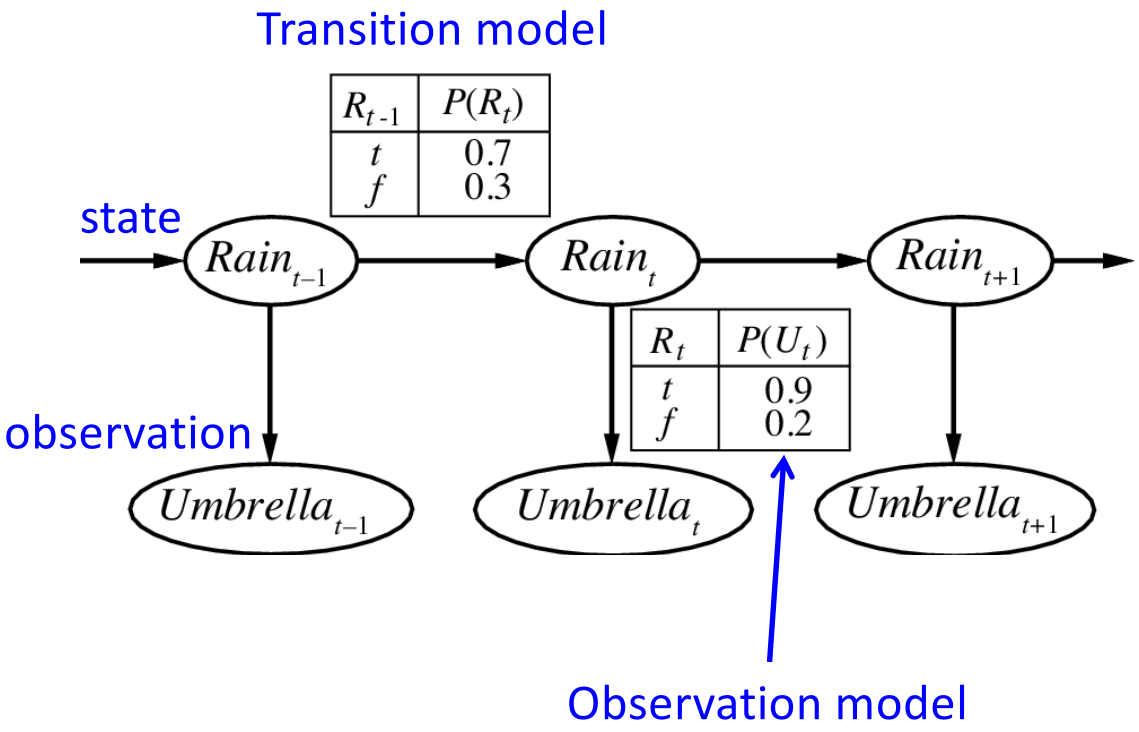
Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

- Richard learns that the weather changes on 3 out of 10 days, thus

$$P(R_t|R_{t-1}) = 0.7$$
$$P(R_t|\neg R_{t-1}) = 0.3$$

- He also learns that Elspeth sometimes forgets her umbrella when it's raining, and that she sometimes brings an umbrella when it's not raining. Specifically,

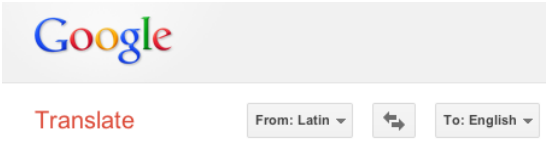
$$P(U_t|R_t) = 0.9$$
$$P(U_t|\neg R_t) = 0.2$$



Applications of HMMs

- Speech recognition HMMs:
  - Observations are acoustic signals (continuous valued)
  - States are specific positions in specific words (so, tens of thousands)

- Machine translation HMMs:
  - Observations are words (tens of thousands)
  - States are translation options



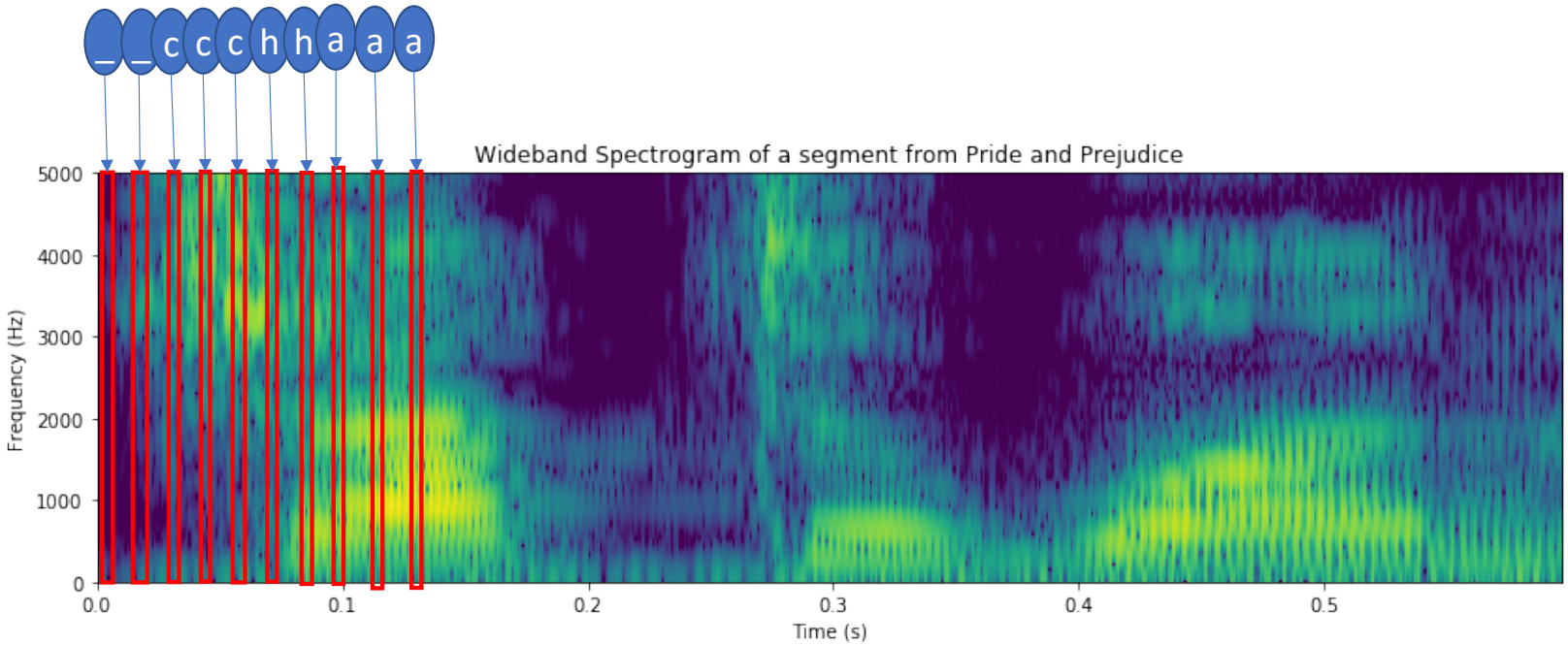
- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

Source: Tamara Berg

Example: Speech Recognition

- Observations: X_t = FFT of 25ms frame of the speech signal.
- State: Y_t = phoneme or letter being currently produced

Example utterance: “chapter one,” from a Librivox recording of *Pride and Prejudice.*



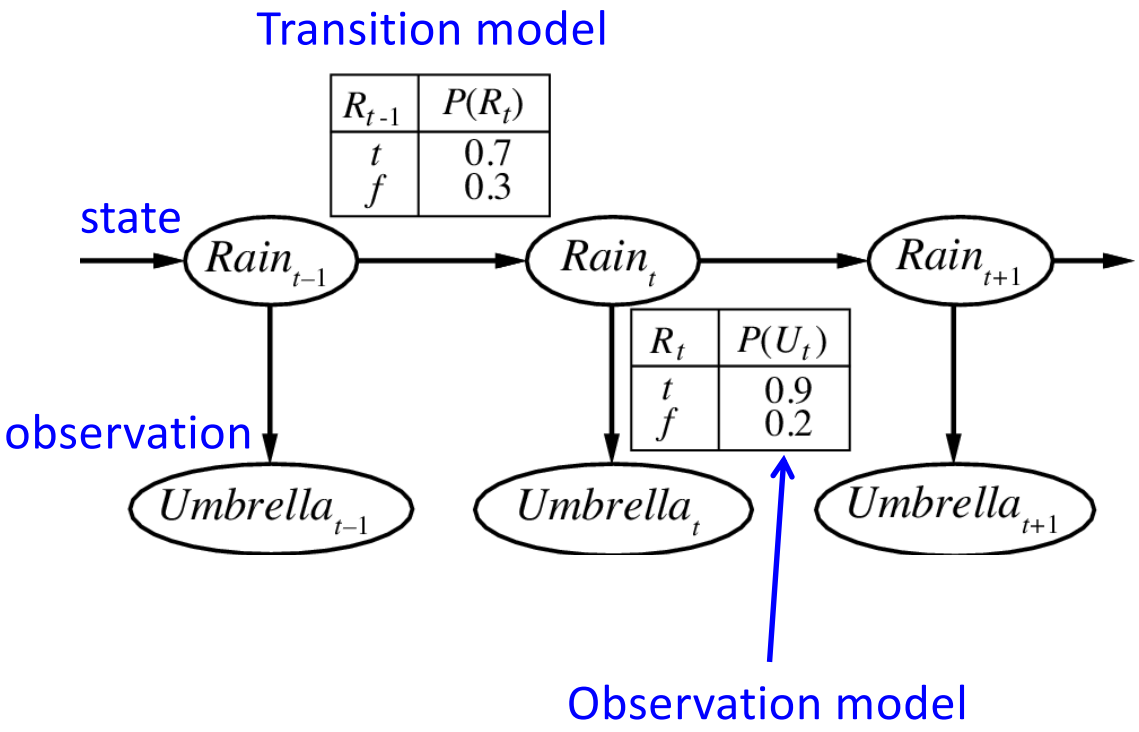
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

HMM as a Bayes Net

This slide shows an HMM as a Bayes Net. You should remember the graph semantics of a Bayes net:

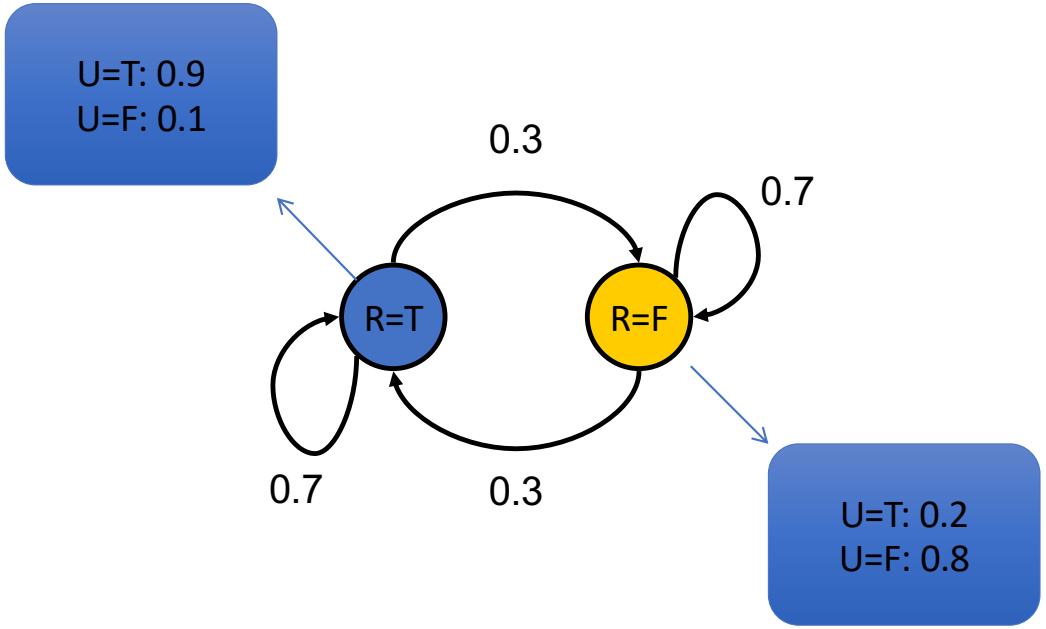
- Nodes are random variables.
- Edges denote stochastic dependence.



HMM as a Finite State Machine

This slide shows ***exactly the same HMM,*** viewed in a totally different way. Here, we show it as a finite state machine:

- Nodes denote states.
- Edges denote possible transitions between the states.
- Observation probabilities must be written using little table thingies, hanging from each state.



Transition probabilities

	R_t = T	R_t = F
R_{t-1} = T	0.7	0.3
R_{t-1} = F	0.3	0.7

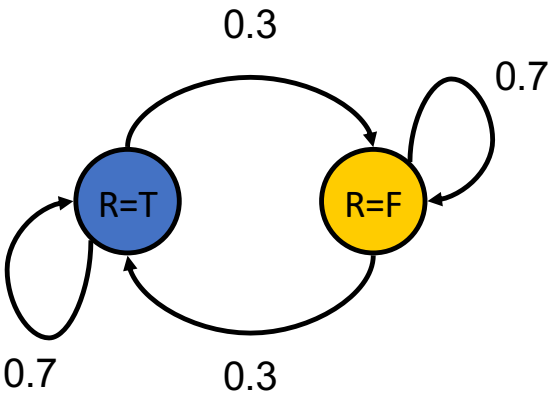
Observation probabilities

	U_t = T	U_t = F
R_t = T	0.9	0.1
R_t = F	0.2	0.8

Bayes Net vs. Finite State Machine

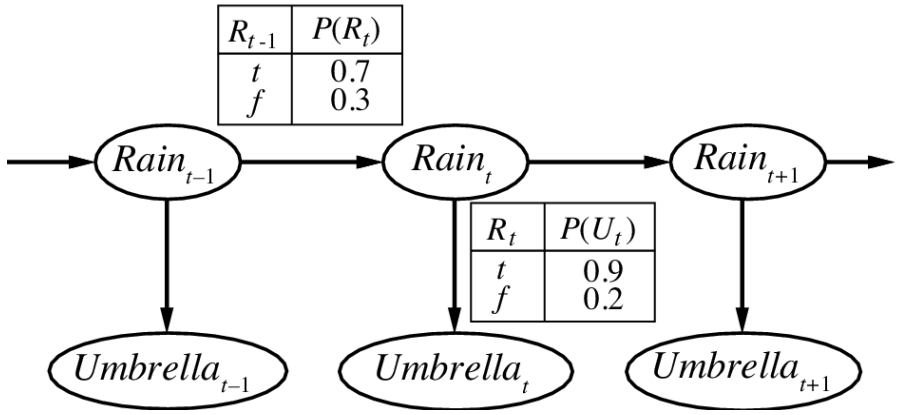
Finite State Machine:

- Lists the different possible states that the world can be in, at one particular time.
- Evolution over time is not shown.



Bayes Net:

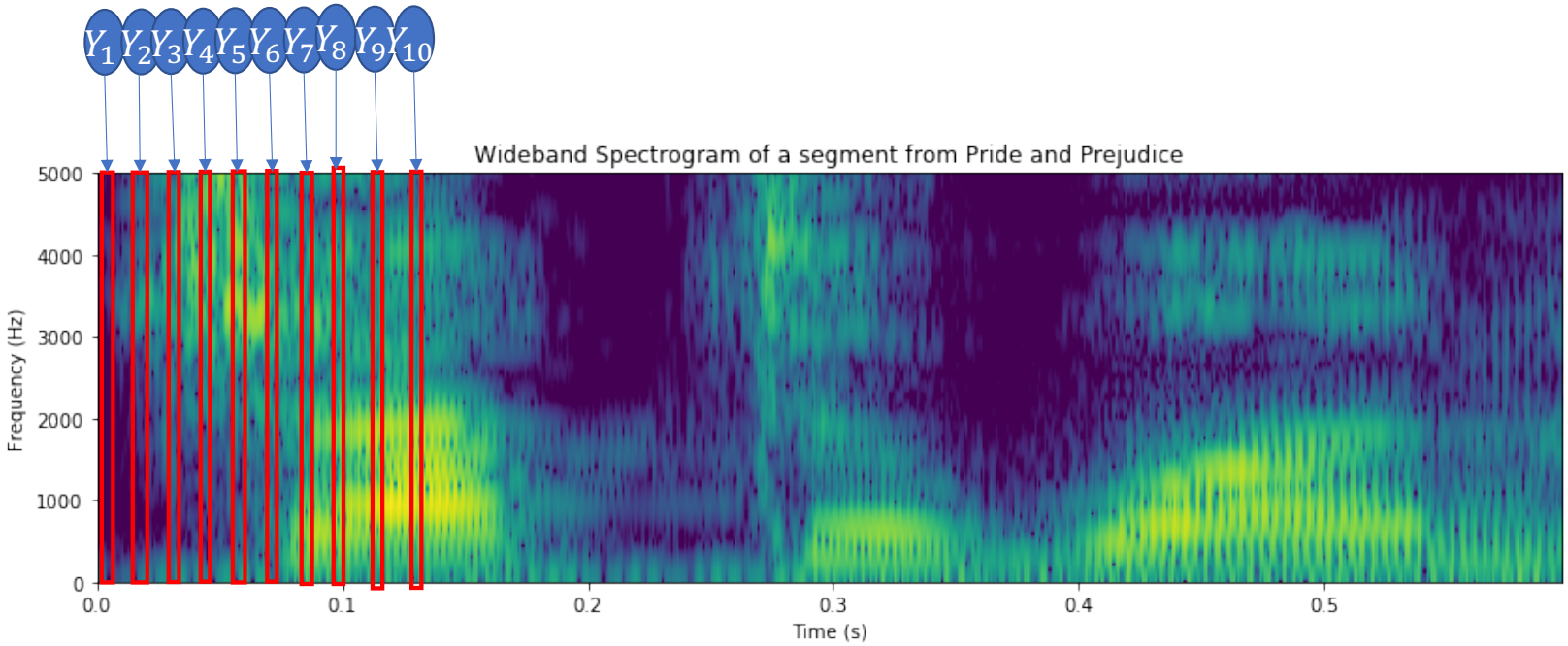
- Lists the different time slices.
- The various possible settings of the state variable are not shown.



Speech Recognition as a Bayes Net

- Observations: X_t = FFT of 25ms frame of the speech signal.
- State: Y_t = phoneme or letter being currently produced

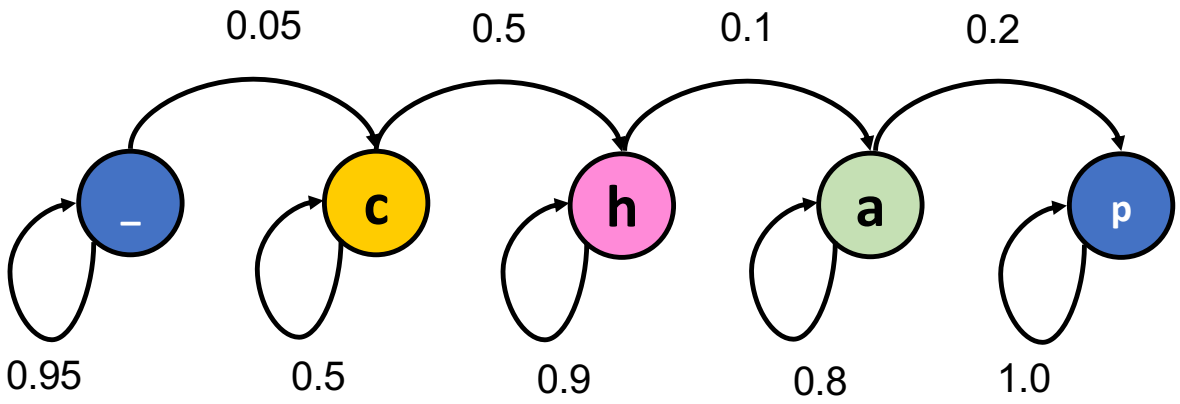
Example utterance: “chapter one,” from a Librivox recording of *Pride and Prejudice.*



Speech Recognition as a Finite State Machine

- Observations: X_t = FFT of 10ms "frame" of the speech signal.
- States: Y_t = letter or phoneme.

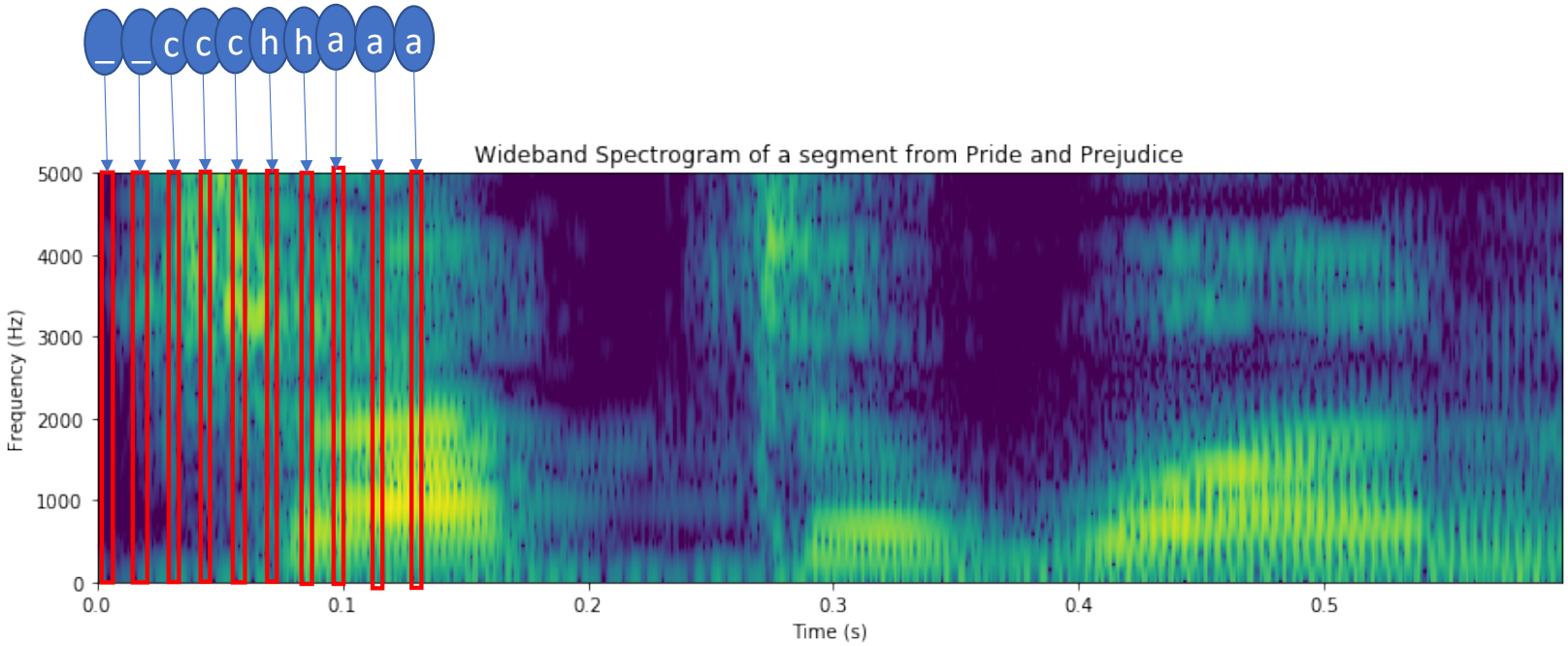
Finite State Machine model of the first part of the word "chapter"



Example: Speech Recognition

- Observations: X_t = FFT of 25ms frame of the speech signal.
- State: Y_t = phoneme or letter being currently produced

Example utterance: “chapter one,” from a Librivox recording of *Pride and Prejudice.*



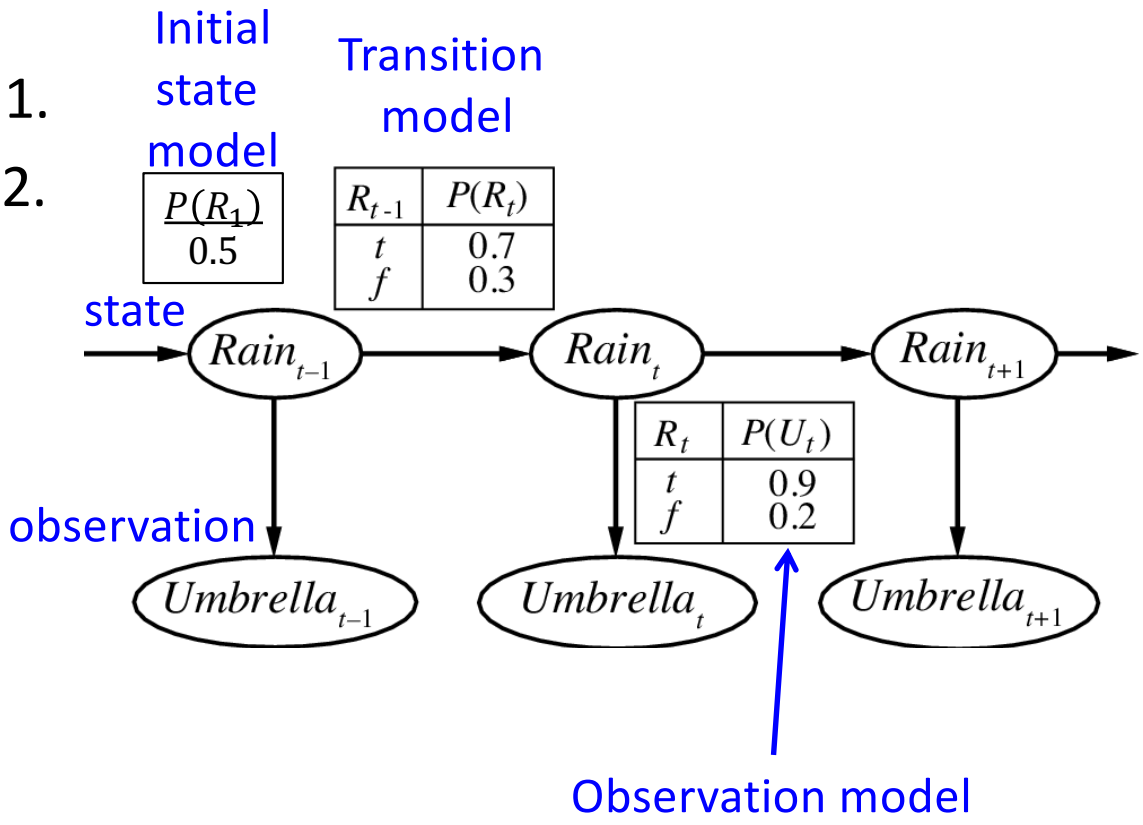
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

Belief propagation in an HMM: Example

- Elspeth has no umbrella on day 1.
- Elspeth has an umbrella on day 2.
- Assume $P(R_1) = 0.5$
- What is the probability that it's raining on day 2?

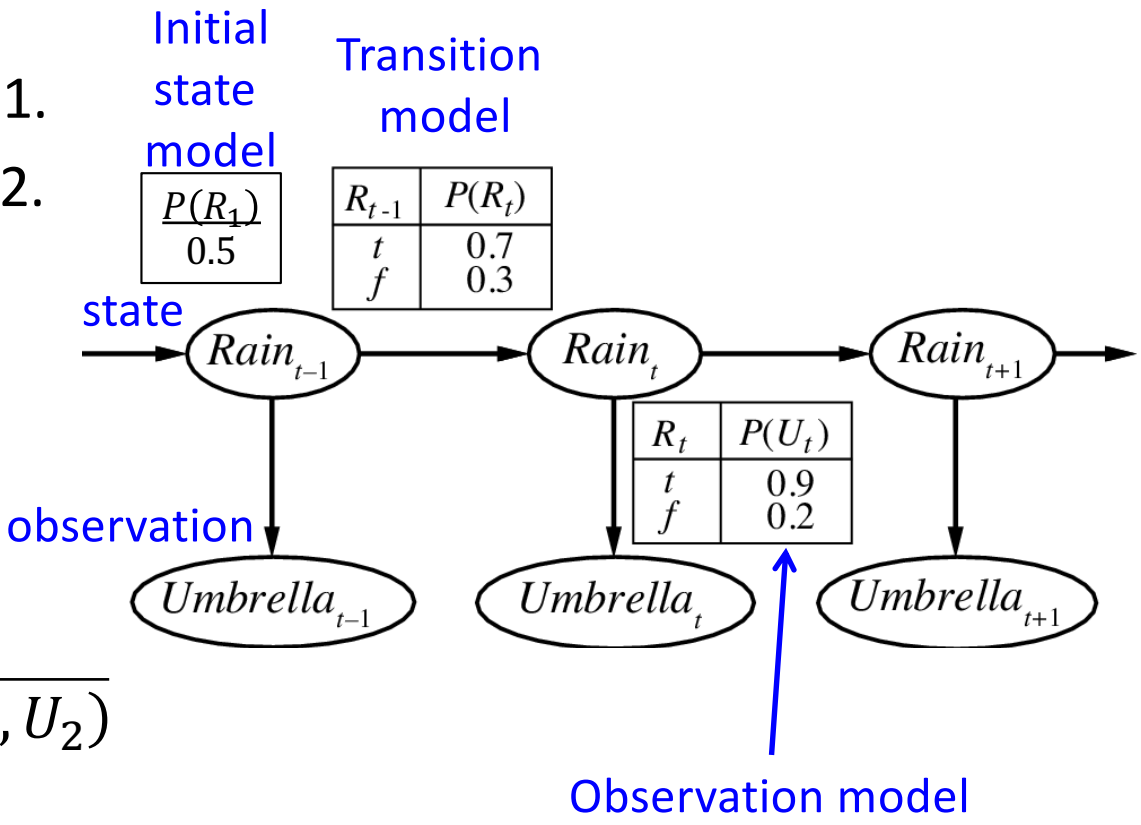
$$P(R_2|\neg U_1, U_2)?$$



Belief propagation in an HMM: Example

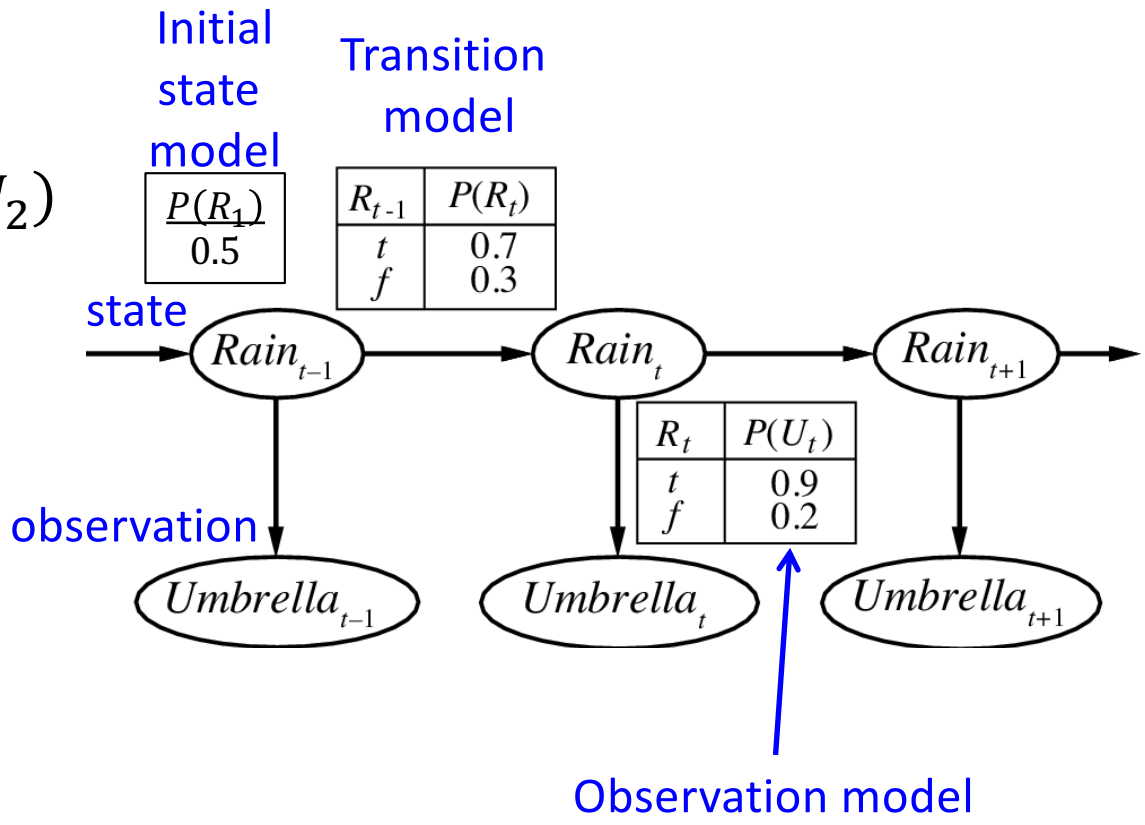
- Elspeth has no umbrella on day 1.
- Elspeth has an umbrella on day 2.
- Assume $P(R_1) = 0.5$
- What is the probability that it's raining on day 2?

$$P(R_2|\neg U_1, U_2)$$
$$= \frac{P(R_2, \neg U_1, U_2)}{P(R_2, \neg U_1, U_2) + P(\neg R_2, \neg U_1, U_2)}$$



Belief propagation in an HMM: Example

$$P(R_2, \neg U_1, U_2)$$
$$= \sum_{r_1 \in \{f,t\}} P(R_1 = r_1, R_2, \neg U_1, U_2)$$

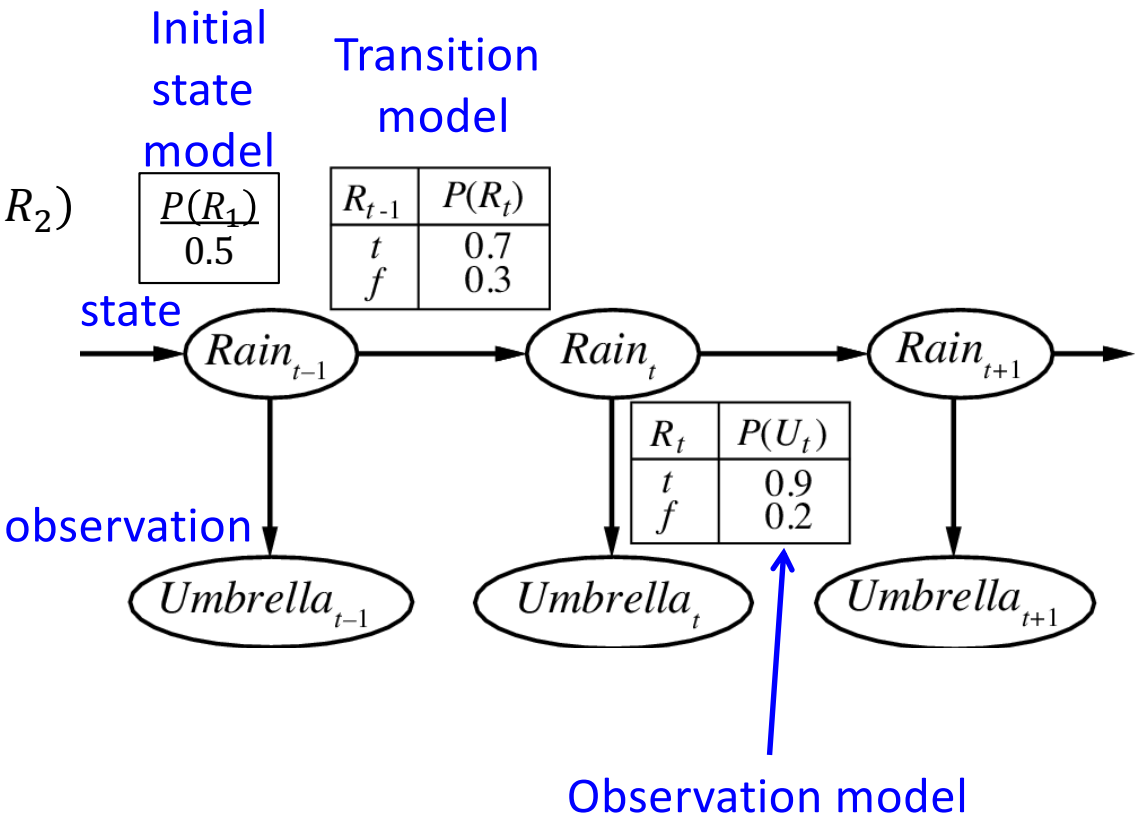


Belief propagation in an HMM: Example

$$P(R_1 = f, R_2, \neg U_1, U_2)$$

$$= P(\neg R_1)P(\neg U_1|\neg R_1)P(R_2|\neg R_1)P(U_2|R_2)$$

$$= (0.5)(0.8)(0.3)(0.9)$$

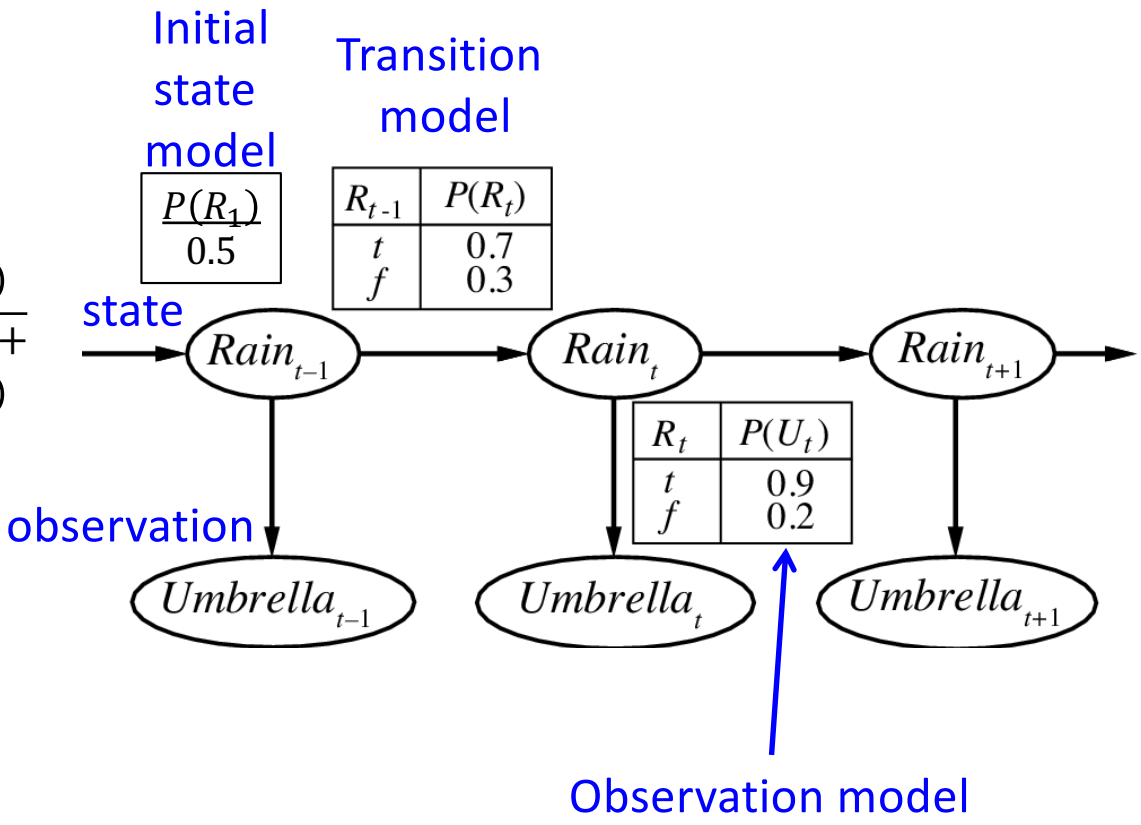


Belief propagation in an HMM: Example

Putting it all together:

$$P(R_2|\neg U_1, U_2)$$

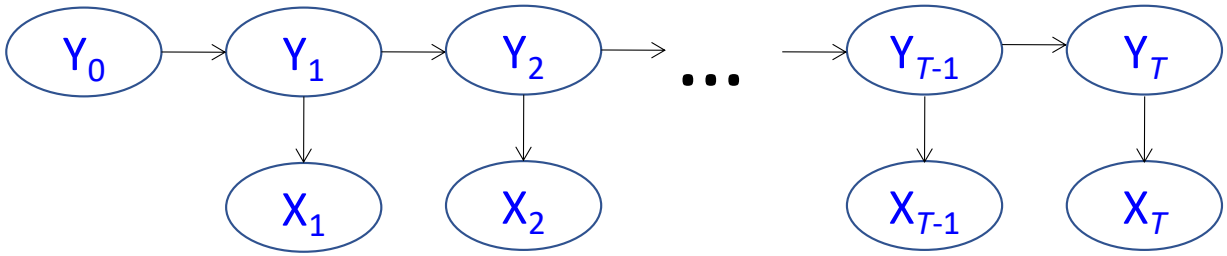
$$= \frac{(0.5)(0.8)(0.3)(0.9) + (0.5)(0.1)(0.7)(0.9)}{(0.5)(0.8)(0.3)(0.9) + (0.5)(0.1)(0.7)(0.9) + (0.5)(0.8)(0.7)(0.2) + (0.5)(0.1)(0.3)(0.2)}$$



The Joint Distribution

- Transition model: $P(Y_t \mid \mathbf{Y}_{0:t-1}) = P(Y_t \mid Y_{t-1})$
- Observation model: $P(X_t \mid \mathbf{Y}_{0:t}, \mathbf{X}_{1:t-1}) = P(X_t \mid Y_t)$
- How do we compute the full joint probability table $P(\mathbf{Y}_{0:T}, \mathbf{X}_{1:T})$?

$$P(Y_{0:T}, X_{1:T}) = P(Y_0) \prod_{t=1}^{T} P(Y_t|Y_{t-1})\, P(X_t|Y_t)$$



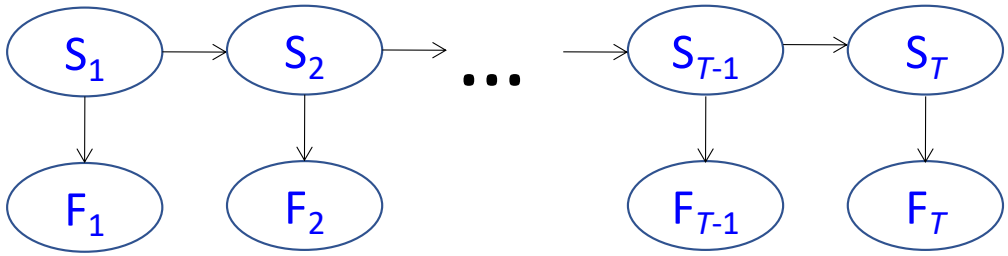
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

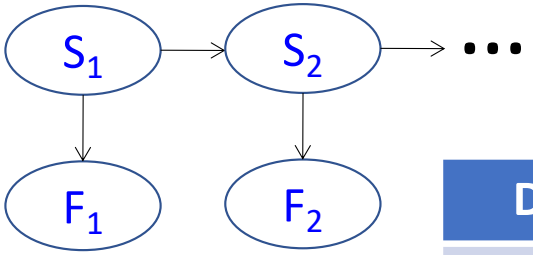
Flying Cows

The University of Illinois Vaccavolatology Department has a new model of the way in which cows learn to fly.

- If a smart cow arrives in the pasture, it tends to remain for more than one day. There is a transition probability, $P(S_t|S_{t-1})$.
- If there is smart cow present, then on that day, it is likely that one or more cows will fly away: $P(F_t|S_t)$.



Flying cows

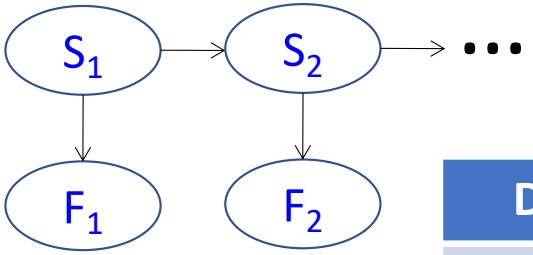


The Vaccavolatologists went out to watch a nearby pasture for ten days.

- Their results are shown in the table at left (True is marked as "T"; False is shown with a blank).

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Maximum Likelihood



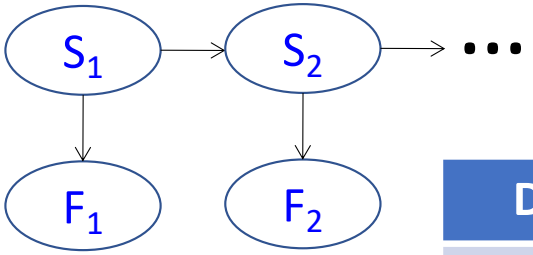
The transition probabilities can be estimated as:

$$P(S_t|S_{t-1}) = \frac{\#\text{ days }(S_t, S_{t-1})}{\#\text{ days }(S_{t-1})} = \frac{4}{5}$$

$$P(S_t|\neg S_{t-1}) = \frac{\#\text{ days }(S_t, \neg S_{t-1})}{\#\text{ days }(\neg S_{t-1})} = \frac{1}{4}$$

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Maximum Likelihood



The observation probabilities can be estimated as:

$$P(F_t|S_t) = \frac{\#\text{ days }(F_t, S_t)}{\#\text{ days }(S_t)} = \frac{3}{5}$$

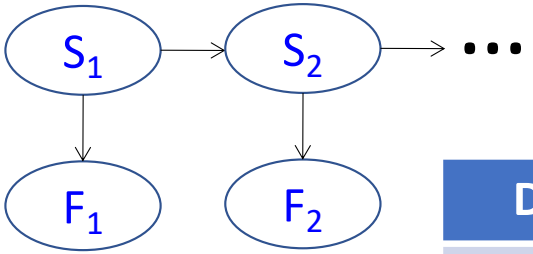
$$P(F_t|\neg S_t) = \frac{\#\text{ days }(F_t, \neg S_t)}{\#\text{ days }(\neg S_t)} = \frac{1}{5}$$

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

Missing data



What can we do if some of the observations are missing?

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	?	T
8	?	
9	?	T
10	?	

Missing data

What can we do if some of the observations are missing?

- Answer: we can use EM, or hard EM, just like any other Bayes Net.

$$P(S_t|S_{t-1}) = \frac{E[\#\text{ days }(S_t, S_{t-1})]}{\mathrm{E}[\#\text{ days }(S_{t-1})]} = \frac{\sum_{t=1}^{T} P(S_t, S_{t-1}|\text{ observations})}{\sum_{t=1}^{T} P(S_{t-1}|\text{ observations})}$$

What we will discuss next week

- EM and hard EM work just fine for an HMM, but …
- They have a computational problem. If we compute $P(S_t, S_{t-1}|F_1, \ldots, F_T)$ in a naïve way, as I showed you at the beginning of this lecture, the computational cost will be $\mathcal{O}\{2^T\}$.
- Solution: pay more attention to the "divide-and-conquer" strategy of belief propagation.
  - For soft EM, the resulting algorithm is called the Baum-Welch algorithm. We will not learn it in this class, but you can learn it in ECE 417 if you wish.
  - For hard EM, the resulting algorithm is called the Viterbi algorithm. We'll learn that in Monday's lecture.

Outline

- HMM: Probabilistic reasoning over time

$$P(Y_{0:T}, X_{1:T}) = P(Y_0) \prod_{t=1}^{T} P(Y_t|Y_{t-1})\, P(X_t|Y_t)$$

- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood

$$P(S_t|S_{t-1}) = \frac{\#\text{ days } (S_t, S_{t-1})}{\#\text{ days } (S_{t-1})}$$

- Parameter learning: EM and Hard EM

$$P(S_t|S_{t-1}) = \frac{E[\#\text{ days } (S_t, S_{t-1})]}{\mathrm{E}[\#\text{ days } (S_{t-1})]}$$