Intelligent Systems (AI-2)

Computer Science cpsc422, Lecture 15

Oct, 10, 2019

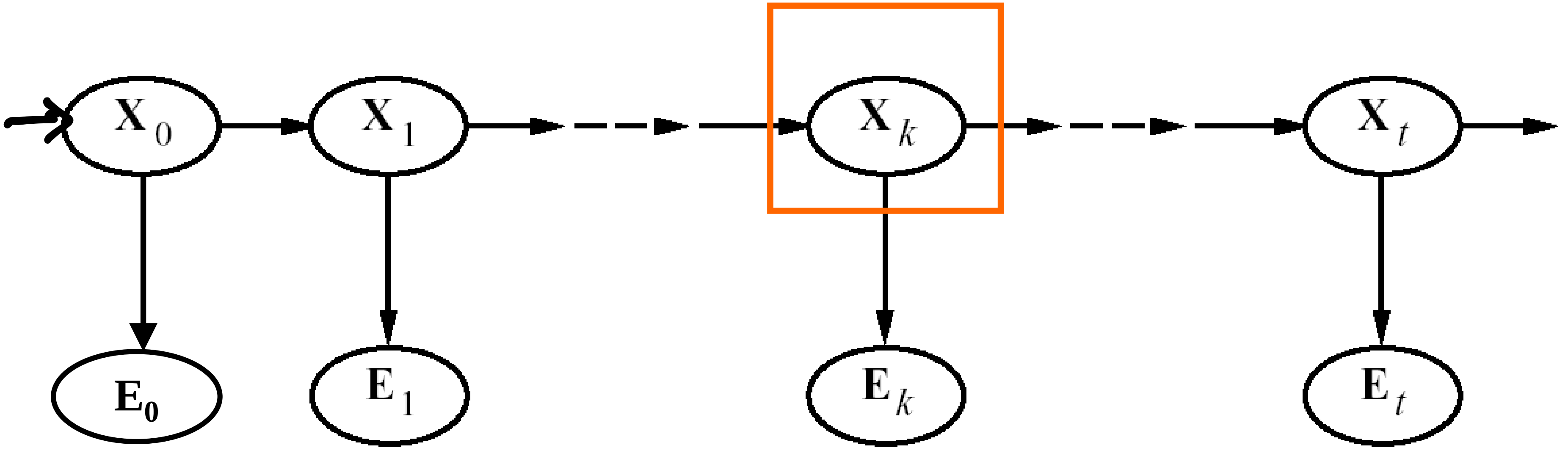
Lecture Overview

Probabilistic temporal Inferences

- Filtering
- Prediction
- **Smoothing (forward-backward)**
- **Most Likely Sequence of States (Viterbi)**

Smoothing

- ***Smoothing***: Compute the posterior distribution over a *past* state given all evidence to date
  - $\boldsymbol{P}(\boldsymbol{X}_k \mid \boldsymbol{e}_{0:t})$ for $1 \leq k < t$



- ***To revise your estimates in the past based on more recent evidence***

Smoothing

- $\boldsymbol{P(X}_k\,|\,\boldsymbol{e}_{0:t}) = \boldsymbol{P(X}_k\,|\,\boldsymbol{e}_{0:k},\boldsymbol{e}_{k+1:t}\,)$ dividing up the evidence

$= \alpha\, \boldsymbol{P(X}_k\,|\,\boldsymbol{e}_{0:k}\,)\, \boldsymbol{P(e}_{k+1:t}\,|\,\boldsymbol{X}_k,\, \boldsymbol{e}_{0:k}\,)$ using…

$= \alpha\, \boldsymbol{P(X}_k\,|\,\boldsymbol{e}_{0:k}\,)\, \boldsymbol{P(e}_{k+1:t}\,|\,\boldsymbol{X}_k)$ using…

iclicker.

A. Bayes Rule

B. Cond. Independence

C. Product Rule

forward message from filtering up to state k, $\boldsymbol{f}_{0:k}$

backward message, $\boldsymbol{b}_{k+1:t}$ computed by a recursive process that runs backwards from *t*

Smoothing

- $\boldsymbol{P}(\boldsymbol{X}_k \mid \boldsymbol{e}_{0:t}) = \boldsymbol{P}(\boldsymbol{X}_k \mid \boldsymbol{e}_{0:k}, \boldsymbol{e}_{k+1:t})$ dividing up the evidence

  $= \alpha\, \boldsymbol{P}(\boldsymbol{X}_k \mid \boldsymbol{e}_{0:k})\, \boldsymbol{P}(\boldsymbol{e}_{k+1:t} \mid \boldsymbol{X}_k, \boldsymbol{e}_{0:k})$ using Bayes Rule

  $= \alpha\, \boldsymbol{P}(\boldsymbol{X}_k \mid \boldsymbol{e}_{0:k})\, \boldsymbol{P}(\boldsymbol{e}_{k+1:t} \mid \boldsymbol{X}_k)$ By Markov assumption on evidence

forward message from filtering up to state k, $\boldsymbol{f}_{0:k}$

backward message, $\boldsymbol{b}_{k+1:t}$ computed by a recursive process that runs backwards from *t*

Backward Message

Product Rule

$\boldsymbol{P(e_{k+1:t} \mid X_k)} = \sum_{x_{k+1}} \boldsymbol{P(e_{k+1:t}, x_{k+1} \mid X_k)} = \sum_{x_{k+1}} \boldsymbol{P(e_{k+1:t} \mid x_{k+1}, X_k)\, P(x_{k+1} \mid X_k)} =$

$= \sum_{x_{k+1}} \boldsymbol{P(e_{k+1:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)}$ by Markov assumption on evidence

Product Rule

$= \sum_{x_{k+1}} \boldsymbol{P(e_{k+1}, e_{k+2:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)}$

$= \sum_{x_{k+1}} \boldsymbol{P(e_{k+1} \mid x_{k+1}, e_{k+2:t})\, P(e_{k+2:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)}$

because $\boldsymbol{e}_{k+1}$ and $\boldsymbol{e}_{k+2:t}$, are conditionally independent given $\boldsymbol{x}_{k+1}$

$= \sum_{x_{k+1}} \boldsymbol{P(e_{k+1} \mid x_{k+1})\, P(e_{k+2:t} \mid x_{k+1})\, P(x_{k+1} \mid X_k)}$

sensor model

recursive call

transition model

- In message notation

$\boldsymbol{b}_{k+1:t} = \text{BACKWARD}\,(\boldsymbol{b}_{k+2:t}, \boldsymbol{e}_{k+1})$

Proof of equivalent statements

X and Y are conditional independent given Z

(1) If $\boxed{P(X \mid Y Z) = P(X \mid Z)} \Rightarrow$

$\Rightarrow$ (A) $\dfrac{P(X,Y,Z)}{P(Y,Z)} = \dfrac{P(X,Z)}{P(Z)} \Rightarrow$ (2) $\boxed{P(Y \mid X,Z) = P(Y \mid Z)}$

$\Rightarrow \dfrac{P(X,Y,Z)}{P(X,Z)} = \dfrac{P(Y,Z)}{P(Z)} \Rightarrow$

(3) $P(X,Y \mid Z) = \dfrac{P(X,Y,Z)}{P(Z)} \overset{\text{from A}}{\Rightarrow} \dfrac{P(Y,Z)\,P(X,Z)}{P(Z)} \cdot \dfrac{1}{P(Z)}$

$= \dfrac{P(Y,Z)}{P(Z)} \cdot \dfrac{P(X,Z)}{P(Z)} = \boxed{P(Y \mid Z) \cdot P(X \mid Z)}$

More Intuitive Interpretation (Example with three states)

$$P(e_{k+1:t} \mid X_k) = \sum_{x_{k+1}} P(x_{k+1} \mid X_k) P(e_{k+1} \mid x_{k+1}) P(e_{k+2:t} \mid x_{k+1})$$

$X = \{s_1, s_2, s_3\}$

e_k $\quad$ e_{k+1} $\quad$ e_{k+2} .... e_t

X_k $\quad$ X_{k+1}

S_1 .... $P(e_{k+2:t} \mid s_1)$

S_2 $P(e_{k+1:t} \mid S_2)$ $\quad$ $P(e_{k+2:t} \mid s_2)$

S_3 .... $P(e_{k+2:t} \mid S_3)$

Forward-Backward Procedure

- To summarize, we showed
  - $P(X_k \mid e_{0:t}) = \alpha\, P(X_k \mid e_{0:k})\, P(e_{k+1:t} \mid X_k)$

- Thus,
  - $\boldsymbol{P(X_k \mid e_{0:t})} = \alpha\, \boldsymbol{f}_{0:k}\, \boldsymbol{b}_{k+1:t}$

and this value can be computed by recursion through time, running forward from *0* to *k* and backwards from *t* to *k+1*

$f_{0:1} \quad f_{0:2} \quad \cdots \quad f_{0:k} * b_{k+1:t} \quad b_{k+2:t} \quad \cdots \quad b_{t+1:t}$

direction of computation

How is it Backward initialized?

$f_{0:1} \quad f_{0:2} \cdots f_{0:k} * b_{k+1:t} \quad b_{k+2:t} \cdots b_{t+1:t}$

direction of computation

- The backwards phase is initialized with making an *unspecified* observation $\boldsymbol{e}_{t+1}$ at $t+1$……

$$\boldsymbol{b}_{t+1:t} = \mathbf{P}(\boldsymbol{e}_{t+1} | \boldsymbol{X}_t) = \mathbf{P}(\, unspecified \mid \boldsymbol{X}_t) = ?$$

A. 0 **B.** 0.5 **C.** 1

iclicker.

How is it Backward initialized?

- The backwards phase is initialized with making an unspecified observation $\boldsymbol{e}_{t+1}$ at t+ 1……

$$\boldsymbol{b}_{t+1:t} = \mathbf{P}(\boldsymbol{e}_{t+1} | \boldsymbol{X}_t) = \mathbf{P}(\textit{unspecified} \mid \boldsymbol{X}_t) = 1$$

- You will observe something for sure! It is only when you put some constraints on the observations that the probability becomes less than 1

Rain Example

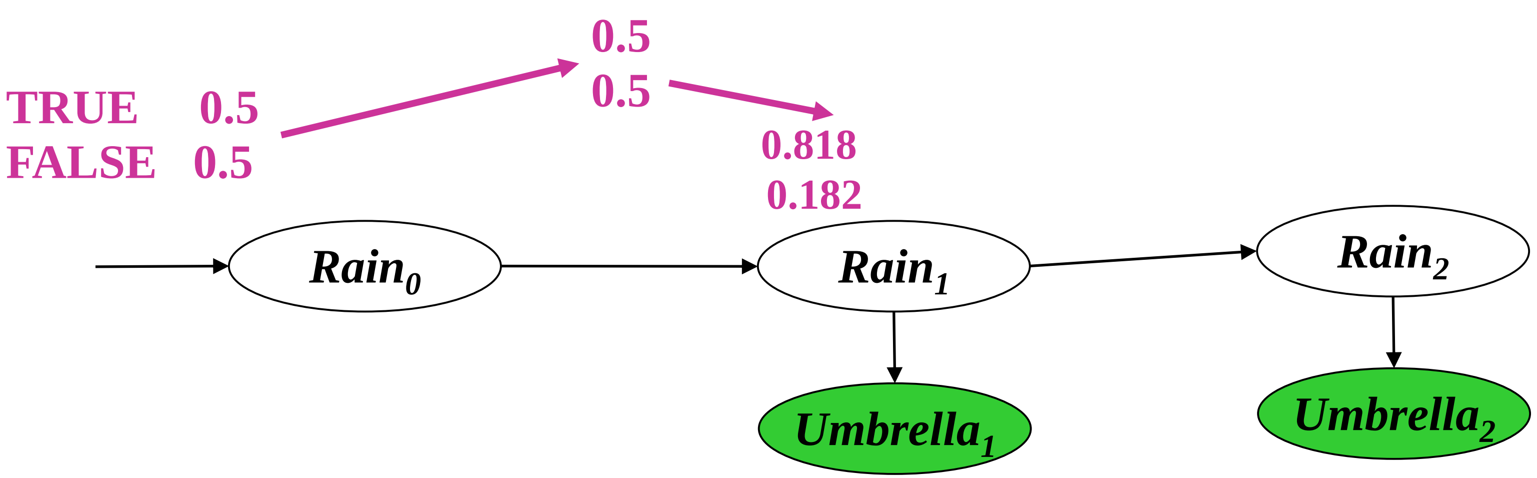
- Let's compute the probability of rain at t = 1, given umbrella observations at t=1 and t =2
- From $\mathbf{P}(\mathbf{X}_k \mid \mathbf{e}_{1:t}) = \alpha\, \mathbf{P}(\mathbf{X}_k \mid \mathbf{e}_{1:k})\, \mathbf{P}(\mathbf{e}_{k+1:t} \mid \mathbf{X}_k)$ we have

$$\mathbf{P}(R_1 \mid \mathbf{e}_{1:2}) = \mathbf{P}(R_1 \mid u_1, u_2) = \alpha\, \mathbf{P}(R_1 \mid u_1)\, \mathbf{P}(u_2 \mid R_1)$$

forward message from filtering up to state 1

backward message for propagating evidence backward from time 2

- $\mathbf{P}(R_1 \mid u_1) = \langle 0.818, 0.182 \rangle$ as it is the filtering to $t = 1$ that we did in lecture 14



Rain Example

- From $\mathbf{P}(\mathbf{e}_{k+1:t} \mid \mathbf{X}_k) = \sum_{x_{k+1}} P(\mathbf{e}_{k+1}|\mathbf{x}_{k+1}) \, P(\mathbf{e}_{k+2:t} |\mathbf{x}_{k+1}) \, \mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k)$
- $\mathbf{P}(u_2 \mid R_1) = \sum_{r \in r_2, \neg r_2} P(u_2|r) \, P(|r) \, P(r \mid R_1) =$

Term corresponding to the Fictitious unspecified observation sequence $e_{3:2}$

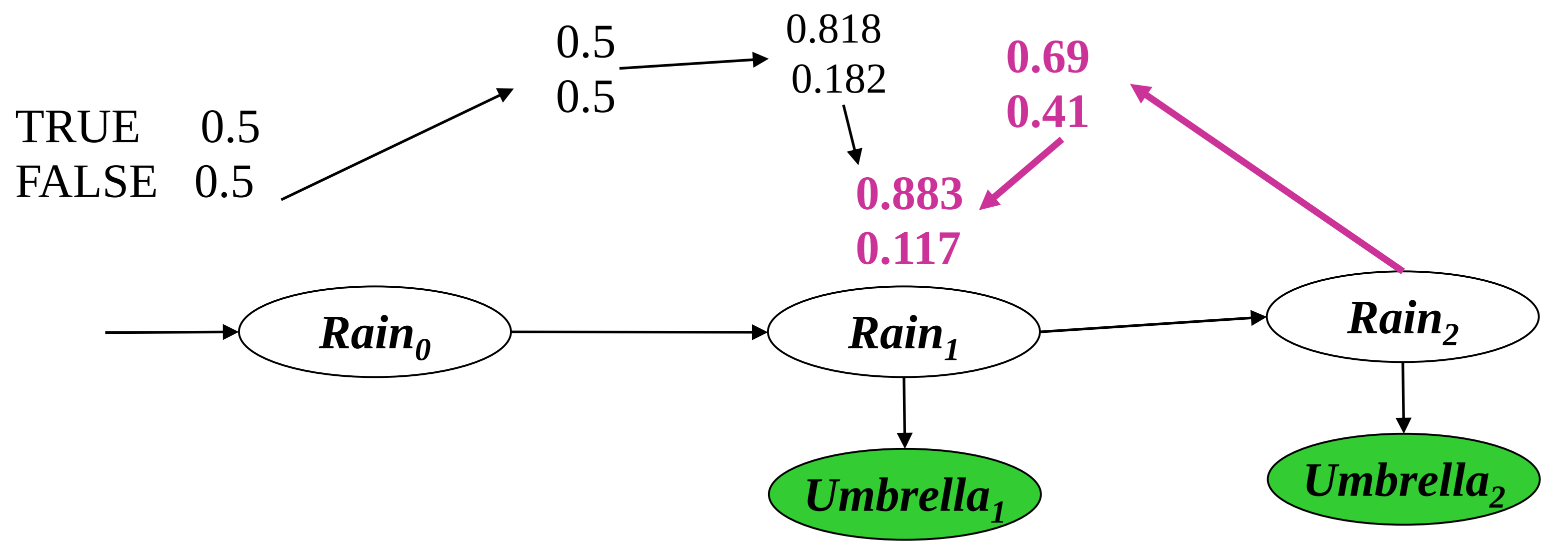
- $P(u_2|r_2) \, P(|r_2) \, \langle P(r_2 \mid r_1), P(r_2 \mid \neg r_1) \rangle +$

  $P(u_2|\neg r_2) \, P(|\neg r_2) \, \langle P(\neg r_2 \mid r_1), P(\neg r_2 \mid \neg r_1) \rangle$

$= (0.9 * 1 * \langle 0.7, 0.3 \rangle) + (0.2 * 1 * \langle 0.3, 0.7 \rangle) = \langle 0.69, 0.41 \rangle$

Thus

- $\alpha \, \mathbf{P}(R_1 | u_1) \, \mathbf{P}(u_2 \mid R_1) = \alpha \langle 0.818, 0.182 \rangle * \langle 0.69, 0.41 \rangle \sim \langle 0.883, 0.117 \rangle$



Lecture Overview

Probabilistic temporal Inferences

- Filtering
- Prediction
- Smoothing (forward-backward)
- **Most Likely Sequence of States (Viterbi)**

Most Likely Sequence

- Suppose that in the *rain* example we have the following *umbrella* observation sequence

  [true, true, false, true, true]

- Is the most likely state sequence?

  [rain, rain, no-rain, rain, rain]

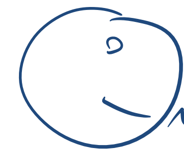
- In this case you may have guessed right… but if you have more states and/or more observations, with complex transition and observation models…..

HMMs : most likely sequence (from 322)

Natural Language Processing: e.g., Speech Recognition

- *States:* phoneme \ word
- *Observations*: acoustic signal \ phoneme

HMM 1 HMM 2



mm ⇒ My name is

Bioinformatics: Gene Finding

- *States:* coding / non-coding region
- *Observations:* DNA Sequences

x x ✓ ✓ ✓ x x

→ A T C G G A A

For these problems the critical inference is:

find the most likely sequence of states given a sequence of observations

Viterbi Algo

Part-of-Speech (PoS) Tagging

- Given a text in natural language, label (*tag*) each word with its syntactic category
  - E.g, Noun, verb, pronoun, preposition, adjective, adverb, article, conjunction
- ***Input***
  - Brainpower not physical plant is now a firm's chief asset.
- ***Output***
  - Brainpower_NN not_RB physical_JJ plant_NN is_VBZ now_RB a_DT firm_NN 's_POS chief_JJ asset_NN ._.

Tag meanings

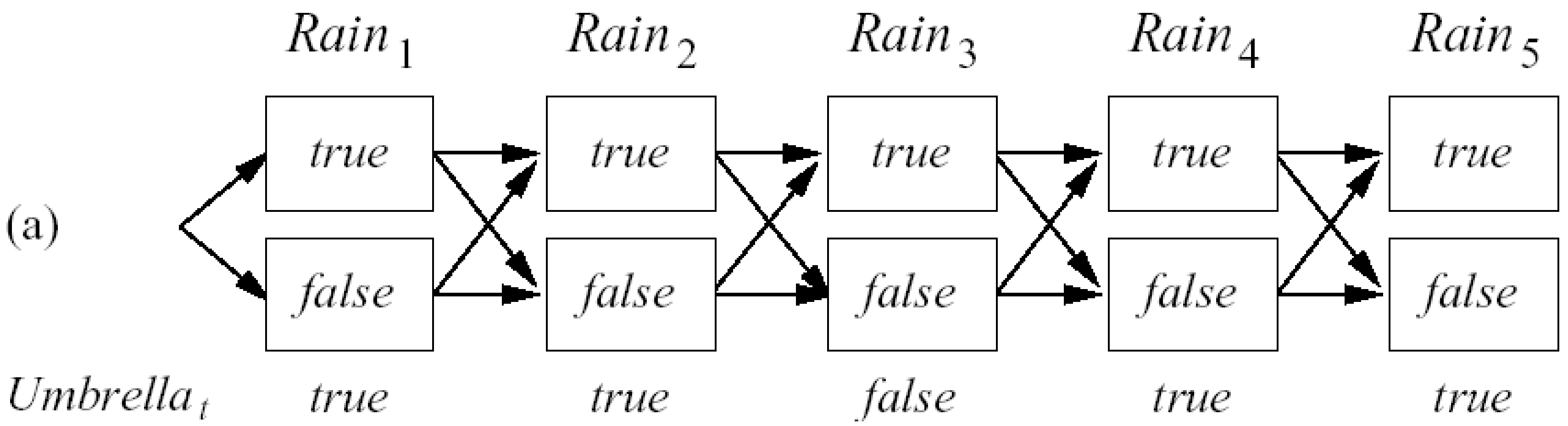
- NNP (Proper Noun singular), RB (Adverb), JJ (Adjective), NN (Noun sing. or mass), VBZ (Verb, 3 person singular present), DT (Determiner), POS (Possessive ending), . (sentence-final punctuation)

POS Tagging is very useful

- As a basis for **parsing** in NL understanding
- **Information Retrieval**
  - ✓Quickly finding names or other phrases for information extraction
  - ✓Select important words from documents (e.g., nouns)
- **Word-sense disambiguation**
  - ✓I made her duck (*how many meanings does this sentence have*)?
- **Speech synthesis**: Knowing PoS produce more natural pronunciations
  - ✓E.g,. Content (noun) vs. content (adjective); object (noun) vs. object (verb)

Most Likely Sequence (Explanation)

- ***Most Likely Sequence***: $\text{argmax}_{x_{1:T}} P(X_{1:T} \mid e_{1:T})$
- Idea
  - find the most likely path to each state in X_T
  - As for filtering etc. we will develop a recursive solution

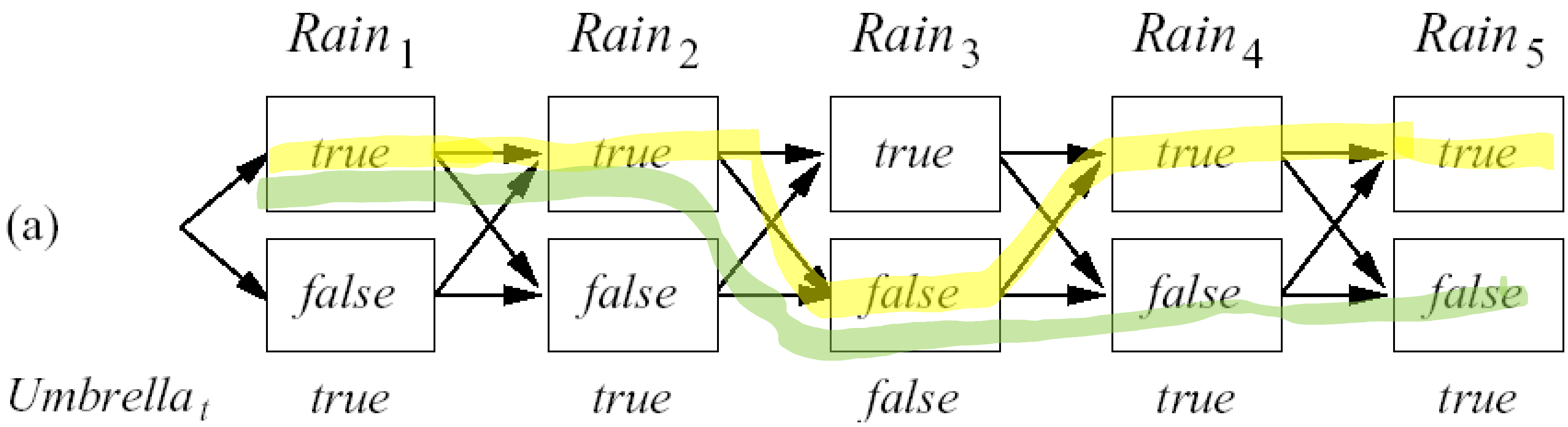


Most Likely Sequence (Explanation)

- ***Most Likely Sequence***: $\text{argmax}_{x_{1:T}} P(X_{1:T} \mid e_{1:T})$
- Idea
  - find the most likely path to each state in X_T
  - As for filtering etc. we will develop a recursive solution

Rain$_5$ = true

Rain$_5$ = false



Learning Goals for today's class

➢You can:

- Describe the **smoothing problem** and derive a solution by manipulating probabilities
- Describe the problem of finding the **most likely sequence of states** (given a sequence of observations)
- Derive recursive solution (if time)

TODO for Wed

- Keep working on Assignment-2: due Wed Oct 18
- Midterm : Fri October 25