Notes from Blackboard
Hidden Markov Models
(Material from 3rd edition of “AI: A Modern Approach”
by S. Russell and P. Norvig; in particular Chapter 15.2)

First things first: Two useful equations

1. Bayes’ rule with background evidence:
   \[ P(Y | X, e) = \frac{P(X | Y, e) P(Y | e)}{P(X | e)} \]
   where \( e \) is background evidence.
2. Conditioning:
   \[ P(Y) = \sum_x P(Y, x) = \sum_x P(Y | x) P(x) \]

1 Recapitulation of Model

1.1 Markov Chains

States \( X_1, ..., X_t \)

1. Markov assumption:
   \[ P(X_t | X_{0:t-1}) = P(X_t | X_{t-1}) \]
   • \( \rightarrow \) current state depends only on the previous state and not on any earlier states
   • \( \rightarrow \) "state provides enough information to make the future conditionally independent of the past"

2. Stationarity assumption:
   \[ P(X_{t+1} | X_t) = P(X_t | X_{t-1}) \text{ f.a. } t \]
   • \( \rightarrow \) process of change that is governed by laws that do not themselves change over time

1.2 Hidden Markov Models (HMM)

Additionally: Observations \( Y_1, ..., Y_t \)

• Sensor Markov assumption:
  \[ P(Y_t | X_{1:t}, Y_{1:t-1}) = P(Y_t | X_t) \]
  \( \rightarrow \) observation in \( t \) fully explained by \( X_t \)

2 Basic Inference Tasks

• Filtering: \( P(X_t | y_{1:t}) \)
• Prediction: \( P(X_{t+k} | y_{1:t}) \) for some \( k > 0 \)
• Smoothing: \( P(X_t | y_{1:T}) \) with \( 0 \leq t < T \)
• MPE: \( \arg \max_{x_{1:T}} P(x_{1:T} | y_{1:T}) \)
3 Example: Umbrella World

3.1 Filtering
Reminder of class content (see slides for details)
Assume we have $P(X_t|y_{1:t-1})$, then:
- Conditioning step: $P(X_t|y_{1:t}) = \frac{1}{Z} \cdot P(X_t|y_{1:t-1}) \cdot P(y_t|X_t)$
- Prediction step: $P(X_{t+1}|y_{1:t}) = \sum_{x_t} P(X_{t+1}|x_t) \cdot P(x_t|y_{1:t})$

Numerical Example
- $P(R_2|u_{1:2})$ with $u_1 = u_2 = true$ and $P(R_1) = \langle 0.5, 0.5 \rangle$
  "Forward" algorithm (because we start with $t = 1$ and work our way forward to $t = 2$):

  $P(R_1|u_1) = \frac{1}{Z} \cdot P(R_1) \cdot P(u_1|R_1)$
  $= \frac{1}{Z} \langle 0.5, 0.5 \rangle \langle 0.9, 0.2 \rangle$
  $= \frac{1}{Z} \langle 0.45, 0.1 \rangle$
  $\simeq \langle 0.818, 0.182 \rangle$

  $P(R_2|u_1) = \sum_{r_1} P(R_2|r_1)P(r_1|u_1)$
  $= \langle 0.7, 0.3 \rangle \cdot 0.818 + \langle 0.3, 0.7 \rangle \cdot 0.182$
  $\simeq \langle 0.627, 0.373 \rangle$
\[ P(R_2|u_1, u_2) = \frac{1}{Z} \cdot P(R_2|u_1) \cdot P(u_2|R_2) \]
\[ = \frac{1}{Z} (0.627, 0.373)(0.9, 0.2) = (0.883, 0.117) \]

3.2 Prediction

\[ P(X_{t+k+1}|y_{1:t}) = \sum_{x_{t+k}} P(X_{t+k+1}|x_{t+k}) \cdot P(x_{t+k}|y_{1:t}) \]

Question

- What if forecasting further and further into the future, given just the first two umbrella observations? What is \( \lim_{t \to \infty} P(R_t|u_1, u_2) \)? Answer: The Markov chain converges to the stationary distribution, which depends only on transition model \( P(X_t|X_{t-1}) \). That is, it “forgets where it is coming from.” (In particular, it does not depend on \( P(Y_t|X_t) \) or \( P(X_1) \).)

For intuition, think about the Umbrella World: if you saw an umbrella yesterday and today, it will not tell you anything about rain in 2 years from now.

- Intuitively, how is this different from \( \lim_{t \to \infty} P(R_t) \)? Answer: it is not.

3.3 Smoothing

Inference task is \( P(X_t|y_{1:T}) \). For example, in the Umbrella World, you could use smoothing if you have observations up until time step \( T = 5 \) and you want to know whether it rained at \( t = 2 \).

Forward-backward algorithm

\[ P(X_t|y_{1:T}) = P(X_t|y_{1:t}, y_{t+1:T}) \]
\[ = \frac{1}{Z} P(X_t|y_{1:t}) P(y_{t+1:T}|X_t, y_{1:t}) \quad \text{(Bayes’ with background evidence, where } y_{1:t} \text{ is } “e”) \]
\[ = \frac{1}{Z} P(X_t|y_{1:t}) P(y_{t+1:T}|X_t) \quad \text{(Conditional Independence)} \]

Observe that first factor is simply filtering (using the “forward” algorithm).
The second factor is obtained using the “backward” algorithm:

\[
P(y_{t+1:T}|x_t) = \sum_{x_{t+1}} P(y_{t+1:T}|x_t, x_{t+1}) P(x_{t+1}|x_t) \quad \text{(conditioning on } X_{t+1})
\]

\[
= \sum_{x_{t+1}} P(y_{t+1:T}|x_{t+1}) P(x_{t+1}|x_t) \quad \text{(conditional independence)}
\]

\[
= \sum_{x_{t+1}} P(y_{t+1}, y_{t+2:T}|x_{t+1}) P(x_{t+1}|x_t) \quad \text{(simple expansion)}
\]

\[
= \sum_{x_{t+1}} P(y_{t+1}|x_{t+1}) P(y_{t+2:T}|x_{t+1}) P(x_{t+1}|x_t) \quad \text{(C.I. of } y_{t+1} \text{ and } y_{t+2:T} \text{ given } X_{t+1})
\]

Observe that the first and third factor are given by model. And the second factor is the recursive step.

**Numerical example**

\[
P(R_1|u_1:2) \text{ with } u_1 = u_2 = true \text{ and } P(R_1) = \langle 0.5, 0.5 \rangle
\]

\[
P(R_1|u_1, u_2) = \frac{1}{Z} P(R_1|u_1) P(u_2|R_1)
\]

\[
= \frac{1}{Z} (0.818, 0.182) P(u_2|R_1)
\]

\[
P(u_2|R_1) = \sum_{r_2} P(u_2|r_2) P(r_2|R_1)
\]

\[
= 0.9 \cdot 1 \cdot (0.7, 0.3) + 0.2 \cdot 1 \cdot (0.3, 0.7)
\]

\[
= \langle 0.69, 0.41 \rangle
\]

Plugging this into (1), we obtain:

\[
P(R_1|u_1, u_2) = \frac{1}{Z} (0.818, 0.182) P(u_2|R_1)
\]

\[
= \frac{1}{Z} (0.818, 0.182) (0.69, 0.41)
\]

\[
\simeq (0.883, 0.117)
\]

Compare this with \( P(r_1|u_1) = 0.818 \): the smoothed estimate for rain on day 1 is higher than filtered estimate. Why? Because an umbrella on day 2 means that rain on day 2 is more likely and since rain tends to persist (see transfer model), rain on day 1 is more likely.

### 3.4 MPE

Smoothing vs MPE: can we just say use smoothing for MPE directly? No! One can even construct examples, where the argmax of (separately) smoothed \( X_1, \ldots, X_T \) return impossible sequence. To do MPE inference, one can use the Viterbi algorithm.