**Review - Markov models**

* Random variables \( S_t = \{1, 2, \ldots, n\} \) state at time \( t \).
* Belief network

\[
\begin{array}{cccccc}
S_1 & \rightarrow & S_2 & \rightarrow & \cdots & \rightarrow S_t \\
& & & & & \\
\end{array}
\]

* Assumptions
  - finite context \( P(S_t | S_1, S_2, \ldots, S_{t-1}) = P(S_t | S_{t-1}) \)
  - shared CPTs \( P(S_t = s' | S_{t-1} = s) = P(S_{t+1} = s' | S_t = s) \)
* Weaknesses
  - modeling \( k \)-th order correlations requires CPTs with \( O(n^k) \) elements.
  - assumes that the true state of the world can be observed.

**Hidden Markov models (HMMs)**

* Random variables

\[ S_t \in \{1, 2, \ldots, n\} \] state at time \( t \).
\[ O_t \in \{1, 2, \ldots, m\} \] observation at time \( t \).

Observations \( O_t \) are noisy, partial reflections of states \( S_t \).

Ex: toilet training

- \( S = \{ \text{have-to-go}, \text{don't-need-to-go}, \text{wee} \} \)
- \( O = \{ \text{neutral}, \text{funny walk}, \text{intense concentration}, \text{squat} \} \)

Ex: speech recognition

\[ \text{time t} \]

\[ \text{window} \]

- \( O_t \): acoustic measurements on windowed waveform at time \( t \)
- \( S_t \): unit of language (word, syllable, phoneme)

Ex: robotics

- \( O_t \): sensor readings
- \( S_t \): location, orientation.

* Belief network

\[
\begin{array}{cccccc}
S_1 & \rightarrow & S_2 & \rightarrow & \cdots & \rightarrow S_t \\
& & & & & \\
\end{array}
\]

* Assumptions
  - finite context \( P(S_t | S_1, S_2, \ldots, S_{t-1}) = P(S_t | S_{t-1}) \)
  \[ P(O_t | S_1, S_2, \ldots, S_t) = P(O_t | S_t) \]
- Shared CPTs

\[ P(S_{t+1} = s' | S_t = s) = P(S_t = s | S_{t-1} = s) \]
\[ P(O_t = o | S_t = s) = P(O_{t+i} = o | S_{t+i} = s) \]

* Joint distribution

\[ P(S_1, S_2, ..., S_T, O_1, O_2, ..., O_T) = P(S_1) \prod_{t=2}^{T} P(S_t | S_{t-1}) \prod_{t=1}^{T} P(O_t | S_t) \]

* Parameters

\[ T_{S_i} = P(S_i = i) \quad \text{initial state distribution} \]
\[ a_{ij} = P(S_t = j | S_{t-1} = i) \quad \text{transition matrix} \]
\[ b_{ik} = P(O_t = k | S_t = i) \quad \text{emission matrix} \]

For clarity: \( b_{ik} = b(\text{c} | k) \).

* Key computations / questions in HMMs.

1) How to compute likelihood \( P(O_1, O_2, ..., O_T) \)?
   - Ex: isolated word recognition.

2) How to compute most likely (hidden) state sequence
   \[ \arg\max_{S_1} P(S_1, S_2, ..., S_T | O_1, O_2, ..., O_T) \]
   - Ex: continuous speech recognition.

3) How to estimate parameters \( T_{S_i}, a_{ij}, b_{ik} \) that maximize \( P(O_1, O_2, ..., O_T) \)?
   (or maybe multiple observation sequences)

* Computing likelihood

\[ P(O_1, O_2, ..., O_T) = \sum_{S_1} P(S_1, S_2, ..., S_T | O_1, ..., O_T) \quad \text{marginalization} \]
\[ = \sum_{S_1} P(S_1) \prod_{t=2}^{T} P(S_t | S_{t-1}) \prod_{t=1}^{T} P(O_t | S_t) \]

* Efficient recursion

\[ P(O_1, O_2, ..., O_T, S_{t+1} = j) = \sum_{i=1}^{T} P(O_1, O_2, ..., O_{t+i}, S_{t+i} = j, S_t = i) \quad \text{marginalization} \]
\[ = \sum_{i=1}^{T} P(O_1, O_2, ..., O_{t+i}, S_{t+i} = j) P(S_{t+i} = j | S_t = i, O_1, ..., O_{t+i}) \quad \text{conditional independence} \]
\[ = \sum_{i=1}^{T} \left( P(O_1, O_2, ..., O_t, S_t = i) P(S_{t+i} = j | S_t = i) P(O_{t+i} | S_{t+i} = j) \right) \quad \text{recursion} \]

* Shorthand notation

\[ a_{it} = P(O_1, O_2, ..., O_t, S_t = i) \quad \text{[n x T matrix]} \]
\[ a_{j+i} = \sum_{i=1}^{N} a_{it} a_{ij} b_{ij}(O_{t+i}) \quad \text{"forward algorithm"} \]

sum last column to get likelihood.
base case: \( \alpha_{i1} = P(O_1, S_1 = i) = P(S_1 = i) P(O_1 | S_1 = i) = \pi_i b_1(O_1) \)

* likelihood

\[
P(O_1, \ldots, O_T) = \sum_{S_1 S_2 \ldots S_T} P(O_1, O_2, \ldots, O_T, S_1 = i) = \sum_{S_1 S_2 \ldots S_T} \alpha_{iT}
\]

* Warning: for long sequence, watch out for underflow.

(2) Computing most likely state sequence

\[
S^* = \{S_1^*, S_2^*, \ldots, S_T^* \}
\]

\[
= \arg \max_S \left[ \log P(S_1, S_2, \ldots, S_T | O_1, O_2, \ldots, O_T) \right] \quad \text{constant with respect to hidden states.}
\]

\[
= \arg \max_S \left[ \log \left( \frac{P(S_1, S_2, \ldots, S_T, O_1, O_2, \ldots, O_T)}{P(O_1, O_2, \ldots, O_T)} \right) \right] \quad \text{to hidden states.}
\]

\[
= \arg \max_S \left[ \log P(S_1, S_2, \ldots, S_T, O_1, O_2, \ldots, O_T) \right] \quad \text{to hidden states.}
\]

Define \( l_{ik}^* = \max_{s_1 s_2 \ldots s_{i-1}} \log P(O_1, O_2, \ldots, O_i, S_i, S_2, \ldots, S_{i-1}, S = i) \)

\[
= \log \text{probability of most likely } i \text{-step "path" that ends in state } i \text{ at time } t \text{ for observations } O_1, O_2, \ldots, O_t
\]

* Form recursion

(i) base case \( t = 1 \)

\[
l_{i1}^* = \log P(S_1 = i, O_1) = \log \left[ P(S_1 = i) P(O_1 | S_1 = i) \right]
\]

\[
= \log \pi_i + \log b_{i1}(O_1)
\]

(ii) from time \( t \) to time \( t+1 \)

\[
l_{ij}^* = \max_{s_1 \ldots s_t} \log P(S_1, S_2, \ldots, S_{t+1} | S_{t+1} = j, O_1, O_2, \ldots, O_{t+1})
\]

\[
= \max_{s_1 \ldots s_t} \max_i \left[ \log P(S_1, \ldots, S_t, S_{t+1} = i, O_1, \ldots, O_t) P(S_{t+1} = j | S_{t+1} = i) \right] \quad \text{representing State } S_{t+1}
\]

\[
= \max_i \left[ \log P(S_1, \ldots, S_{t+1} = i, O_1, \ldots, O_t) + \log P(S_{t+1} = j | S_{t+1} = i) \right] + \log P(O_{t+1} | S_{t+1} = j)
\]

* How to derive \( S^* \) from \( l^* \)?