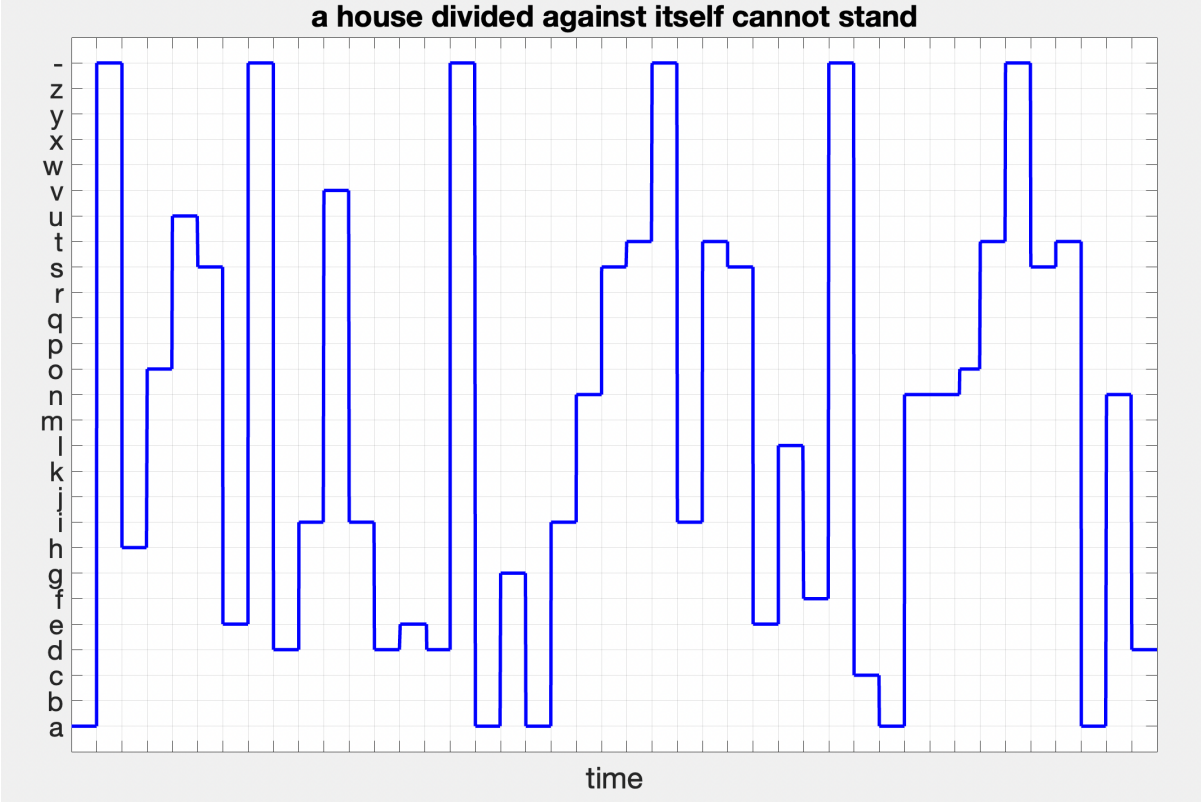
CSE 150A/250A. Assignment 6 *(42 pts total)*

6.1 Viterbi algorithm *(14 pts)*

Source code: *10 pts*

Plot with correct message: *4 pts*



6.2 Inference in HMMs *(13 pts)*

Consider a discrete HMM with hidden states S_t, observations O_t, transition matrix $a_{ij} = P(S_{t+1}=j \mid S_t=i)$ and emission matrix $b_{ik} = P(O_t=k \mid S_t=i)$. In class, we defined the forward-backward probabilities:

$$\alpha_{it} = P(o_1, o_2, \ldots, o_t, S_t=i)$$
$$\beta_{it} = P(o_{t+1}, o_{t+2}, \ldots, o_T \mid S_t=i)$$

for a particular observation sequence $\{o_1, o_2, \ldots, o_T\}$ of length T. In terms of these probabilities, which you may assume to be given, as well as the transition and emission matrices of the HMM, show how to (efficiently) compute the following quantities:

(a) $P(S_{t+1}=j \mid S_t=i, o_1, o_2, \ldots, o_T)$ *(3 pts)*

$$P(S_{t+1}=j \mid S_t=i, o_1, o_2, \ldots, o_T)$$
$$= \frac{P(S_t=i, S_{t+1}=j, o_1, \ldots, o_T)}{P(S_t=i, o_1, o_2, \ldots, o_T)} \quad \boxed{\textbf{product rule}}$$
$$= \frac{P(S_t=i, o_1, \ldots, o_t)\, P(S_{t+1}=j \mid S_t=i)\, P(o_{t+1} \mid S_{t+1})\, P(o_{t+2}, \ldots, o_T \mid S_{t+1}=j)}{P(S_t=i, o_1, \ldots, o_t)\, P(o_{t+1}, \ldots, o_T \mid S_t=i)} \quad \boxed{\textbf{CI}}$$
$$= \frac{\alpha_{it}\, a_{ij}\, b_j(o_{t+1})\, \beta_{j,t+1}}{\alpha_{it}\, \beta_{it}} \quad \boxed{\textbf{definitions}}$$
$$= \frac{a_{ij}\, b_j(o_{t+1})\, \beta_{j,t+1}}{\beta_{it}} \quad \boxed{\textbf{cancel common factors}}$$

(b) $P(S_t=i \mid S_{t+1}=j, o_1, o_2, \ldots, o_T)$ *(3 pts)*

$$P(S_t=i \mid S_{t+1}=j, o_1, o_2, \ldots, o_T)$$
$$= \frac{P(S_t=i, S_{t+1}=j, o_1, \ldots, o_T)}{P(S_{t+1}=j, o_1, o_2, \ldots, o_T)} \quad \boxed{\textbf{product rule}}$$
$$= \frac{P(S_t=i, o_1, \ldots, o_t)\, P(S_{t+1}=j \mid S_t=i)\, P(o_{t+1} \mid S_{t+1}=j)\, P(o_{t+2}, \ldots, o_T \mid S_{t+1}=j)}{P(S_{t+1}=j, o_1, \ldots, o_{t+1})\, P(o_{t+2}, \ldots, o_T \mid S_{t+1}=j)} \quad \boxed{\textbf{CI}}$$
$$= \frac{\alpha_{it}\, a_{ij}\, b_j(o_{t+1})\, \beta_{j,t+1}}{\alpha_{j,t+1}\, \beta_{j,t+1}} \quad \boxed{\textbf{definitions}}$$
$$= \frac{\alpha_{it}\, a_{ij}\, b_j(o_{t+1})}{\alpha_{j,t+1}} \quad \boxed{\textbf{cancel common factors}}$$

(c) $P(S_{t-1}=i, S_t=j, S_{t+1}=k \mid o_1, o_2, \ldots, o_T)$ *(4 pts)*

$$
\begin{aligned}
&P(S_{t-1}=i, S_t=j, S_{t+1}=k \mid o_1, o_2, \ldots, o_T) \\
&\quad = \frac{P(S_{t-1}=i, S_t=j, S_{t+1}=k, o_1, o_2, \ldots, o_T)}{P(o_1, o_2, \ldots, o_T)} \quad \boxed{\textbf{product rule}} \\
&\quad = \frac{P(S_{t-1}=i, o_1, \ldots, o_{t-1})\, P(S_t=j \mid S_{t-1}=i)\, P(o_t \mid S_t=j)}{P(S_{t+1}=k \mid S_t=j)\, P(o_{t+1} \mid S_{t+1}=k)\, P(o_{t+2}, \ldots, o_T \mid S_{t+1}=k)} \quad \boxed{\textbf{CI}} \\
&\quad = \frac{\alpha_{i,t-1}\, a_{ik}\, b_j(o_t)\, a_{kj}\, b_j(o_{t+1})\, \beta_{j,t+1}}{P(o_1, o_2, \ldots, o_T)} \quad \boxed{\textbf{definitions}} \\
&\quad = \frac{\alpha_{i,t-1}\, a_{ik}\, b_k(o_t)\, a_{kj}\, b_j(o_{t+1})\, \beta_{j,t+1}}{\sum_i P(S_t=i, o_1, o_2, \ldots, o_T)} \quad \boxed{\textbf{marginalization}} \\
&\quad = \frac{\alpha_{i,t-1}\, a_{ik}\, b_k(o_t)\, a_{kj}\, b_j(o_{t+1})\, \beta_{j,t+1}}{\sum_\ell P(S_t=\ell, o_1, \ldots, o_t)\, P(o_{t+1}, \ldots, o_T \mid S_t=\ell)} \quad \boxed{\textbf{CI}} \\
&\quad = \frac{\alpha_{i,t-1}\, a_{ik}\, b_k(o_t)\, a_{kj} b_j(o_{t+1})\, \beta_{j,t+1}}{\sum_\ell \alpha_{\ell t}\, \beta_{\ell t}} \quad \boxed{\textbf{definitions}}
\end{aligned}
$$

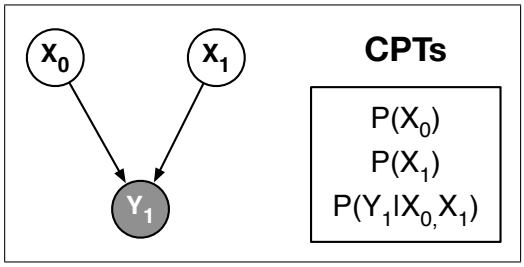
(d) $\operatorname{argmax}_i P(S_t=i \mid o_1, o_2, \ldots, o_T)$ *(3 pts)*

$$
\begin{aligned}
&\operatorname*{argmax}_i P(S_t=i \mid o_1, o_2, \ldots, o_T) \\
&\quad = \operatorname*{argmax}_i \left[\frac{P(S_t=i, o_1, o_2, \ldots, o_T)}{P(o_1, o_2, \ldots, o_T)} \right] \quad \boxed{\textbf{product rule}} \\
&\quad = \operatorname*{argmax}_i P(S_t=i, o_1, o_2, \ldots, o_T) \quad \boxed{\textbf{denominator does not depend on } \boldsymbol{i}} \\
&\quad = \operatorname*{argmax}_i \Big[P(S_t=i, o_1, o_2, \ldots, o_t)\, P(o_{t+1}, o_{t+2}, \ldots, o_T \mid S_t=i, o_1, o_2, \ldots, o_t) \Big] \quad \boxed{\textbf{product rule}} \\
&\quad = \operatorname*{argmax}_i \Big[P(S_t=i, o_1, o_2, \ldots, o_t)\, P(o_{t+1}, o_{t+2}, \ldots, o_T \mid S_t=i) \Big] \quad \boxed{\textbf{conditional independence}} \\
&\quad = \operatorname*{argmax}_i \Big[\alpha_{it}\, \beta_{it} \Big] \quad \boxed{\textbf{substitution}}
\end{aligned}
$$

6.3 Belief updating *(9 pts)*

Consider the simple belief network on the right with nodes X_0, X_1, and Y_1. To compute the posterior probability $P(X_1 \mid Y_1)$, we can use Bayes rule:

$$P(X_1 \mid Y_1) = \frac{P(Y_1 \mid X_1)\, P(X_1)}{P(Y_1)}$$



(a) **Numerator** *(2 pts)*

$$\begin{aligned} P(Y_1 \mid X_1) &= \sum_x P(Y_1, X_0{=}x \mid X_1) && \boxed{\textbf{marginalization}} \\ &= \sum_x P(X_0{=}x \mid X_1)\, P(Y_1 \mid X_0{=}x, X_1) && \boxed{\textbf{product rule}} \\ &= \sum_x P(X_0{=}x)\, P(Y_1 \mid X_0{=}x, X_1) && \boxed{\textbf{conditional independence}} \end{aligned}$$

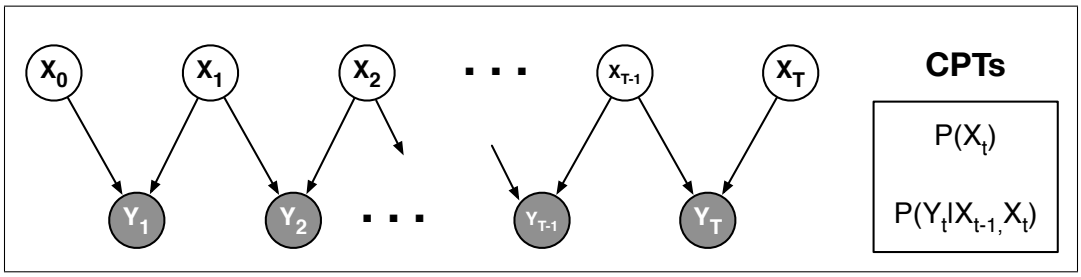
(b) **Denominator** *(2 pts)*

$$\begin{aligned} P(Y_1) &= \sum_{x'} P(Y_1, X_1{=}x') && \boxed{\textbf{marginalization}} \\ &= \sum_{x'} P(X_1{=}x')\, P(Y_1 \mid X_1{=}x') && \boxed{\textbf{product rule}} \\ &= \sum_{x'} \sum_x P(X_0{=}x)\, P(X_1{=}x')\, P(Y_1 \mid X_0{=}x, X_1{=}x') && \boxed{\textbf{substitution from (a)}} \end{aligned}$$

Now consider the belief network shown above. It does not have the same structure as an HMM, but using similar ideas we can derive efficient algorithms for inference. In particular, consider how to compute the posterior probability $P(X_t \mid Y_1, Y_2, \ldots, Y_t)$ that accounts for evidence up to and including time t. We can derive an efficient recursion from Bayes rule:

$$P(X_t \mid Y_1, Y_2, \ldots, Y_t) = \frac{P(Y_t \mid X_t, Y_1, Y_2, \ldots, Y_{t-1})\, P(X_t \mid Y_1, Y_2, \ldots, Y_{t-1})}{P(Y_t \mid Y_1, \ldots, Y_{t-1})},$$

where the nodes $Y_1, Y_2, \ldots, Y_{t-1}$ are treated as background evidence.



In parts (c–e) of this problem you will compute the individual terms that appear in this version of Bayes rule. You should express your answers in terms of the CPTs of the belief network and the probabilities $P(X_{t-1} = x \mid Y_1, Y_2, \ldots, Y_{t-1})$, *which you may assume have been computed in a previous step of the recursion*. Your answers to parts (a) and (b) may be instructive for parts (d) and (e).

(c) **Second term in numerator** *(1 pt)*

$$P(X_t \mid Y_1, Y_2, \ldots, Y_{t-1}) = P(X_t) \quad \boxed{\textbf{conditional independence}}$$

(d) **First term in numerator** *(2 pts)*

$$\begin{aligned}
&P(Y_t \mid X_t, Y_1, Y_2, \ldots, Y_{t-1}) \\
&\quad = \sum_x P(X_{t-1}{=}x, Y_t \mid X_t, Y_1, Y_2, \ldots, Y_{t-1}) \quad \boxed{\textbf{marginalization}} \\
&\quad = \sum_x P(X_{t-1}{=}x \mid X_t, Y_1, Y_2, \ldots, Y_{t-1})\, P(Y_t \mid X_{t-1}{=}x, X_t, Y_1, Y_2, \ldots, Y_{t-1}) \quad \boxed{\textbf{product rule}} \\
&\quad = \sum_x P(X_{t-1}{=}x \mid Y_1, Y_2, \ldots, Y_{t-1})\, P(Y_t \mid X_{t-1}{=}x, X_t) \quad \boxed{\textbf{conditional independence}}
\end{aligned}$$

(e) **Denominator** *(2 pts)*

$$\begin{aligned}
&P(Y_t \mid Y_1, Y_2, \ldots, Y_{t-1}) \\
&\quad = \sum_{x'} P(X_t{=}x', Y_t \mid Y_1, Y_2, \ldots, Y_{t-1}) \quad \boxed{\textbf{marginalization}} \\
&\quad = \sum_{x'} P(X_t{=}x' \mid Y_1, Y_2, \ldots, Y_{t-1})\, P(Y_t \mid X_t{=}x', Y_1, Y_2, \ldots, Y_{t-1}) \quad \boxed{\textbf{product rule}} \\
&\quad = \sum_{x'} \sum_x P(X_{t-1}{=}x \mid Y_1, Y_2, \ldots, Y_{t-1})\, P(X_t{=}x')\, P(Y_t \mid X_{t-1}{=}x, X_t{=}x') \quad \boxed{\textbf{substitution}}
\end{aligned}$$

Note that the steps for (d) and (e) are identical to those for (a) and (b).

6.4 Most likely hidden states *(6 pts)*

The Viterbi algorithm in HMMs computes the most likely *sequence* of hidden states for a particular sequence of observations:

$$\{s_1^*, s_2^*, \ldots, s_T^*\} = \underset{s_1, s_2, \ldots, s_T}{\operatorname{argmax}} P(s_1, s_2, \ldots, s_T \mid o_1, o_2, \ldots, o_T).$$

Compare these *collectively* optimal hidden states s_t^* to the *individually* optimal hidden states $\hat{s}_t$:

$$\hat{s}_t = \underset{s_t}{\operatorname{argmax}} P(S_t = s_t \mid o_1, o_2, \ldots, o_T) \qquad \text{for } t \in \{1, 2, \ldots, T\}.$$

Answer the following True/False questions and explain your reasoning:

(a) It is *possible* that $P(\hat{s}_1, \ldots, \hat{s}_T \mid o_1, \ldots, o_T) > P(s_1^*, \ldots, s_T^* \mid o_1, \ldots, o_T)$. *(2 pts)*

> **False**: By definition, the hidden states s_t^* are those that maximize the probability on the RHS:
>
> $$P(s_1^*, \ldots, s_T^* \mid o_1, \ldots, o_T) = \max_{s_1, \ldots, s_T} P(s_1, \ldots, s_T \mid o_1, \ldots, o_T).$$
>
> If the inequality were true, then there'd have to exist a sequence of hidden states that had a *greater* probability than the maximum, which is a contradiction.

(b) It is *possible* that $\hat{s}_t = s_t^*$ for all t. *(2 pts)*

> **True:** The $\hat{s}_t$ states might happen to coincide with the s_t^* sequence if the HMM's probabilities are defined such that one single sequence has a much higher joint probability than any other sequence. This dominant sequence would be the s_t^* sequence, but it would also be the $\hat{s}_t$ states because it's so unlikely that the HMM's nodes would be in any state outside of this dominant sequence.
>
> (Other explanations/hypotheticals are possible.)
>
> **Example:** Consider a completely uniform HMM where $a_{ij} = \frac{1}{n}$ and $b_{ik} = \frac{1}{n}$ for all states i, j: every state is equally likely to occur, every state is equally likely to transition to any other state, and the observations provide no information about the hidden states. In this scenario, *every* sequence of hidden states $\{s_1, s_2, \ldots, s_T\}$ is equally collectively optimal, and every hidden state s_t is also equally individually optimal.

Do *not* accept explanations of the form "It is possible because the sequences might happen to match exactly". The entire claim is that it is possible for the two sequences to match exactly, so such explanations are essentially just restating the claim. The explanation must explain *why/how* such a scenario could ever occur, either in general terms or with a specific example.

(c) It is *always* true that $\hat{s}_t = s_t^*$ for all t. *(2 pts)*

False: Even though the $\hat{s}_t$ states are individually optimal, transition probabilities between these states may actually be very low, in which case the joint probability of the $\hat{s}_t$ states would also be very low. In contrast, a different sequence with lower individual probabilities might have a higher joint probability because of its transition probabilities.

(Other explanations/hypotheticals are possible.)

Technically valid (but unsatisfying) counterexample: Consider the completely uniform HMM from part (b). Simply choose an arbitrary sequence to be our $\hat{s}_t$ states (which is valid because all states are individually optimal at every t), and then choose any other sequence to be our s_t^* sequence (which is valid because all sequences are collectively optimal).

A fully worked-out counterexample is *not* required for this subproblem. That would require computing by hand the ℓ^* matrix from the Viterbi algorithm *and* the α and β matrices from the forward-backward algorithm. An intuitive (but correct) explanation would suffice.

As before, do *not* accept explanations that essentially just restate the claim.