CS 188
Summer 2023

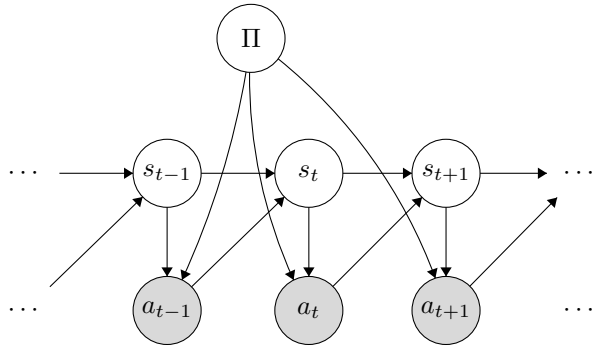
Discussion 4C

1 Particle Filtering Apprenticeship

We are observing an agent's actions in an MDP and are trying to determine which out of a set $\{\pi_1, \ldots, \pi_n\}$ the agent is following. Let the random variable Π take values in that set and represent the policy that the agent is acting under. We consider only *stochastic* policies, so that A_t is a random variable with a distribution conditioned on S_t and Π. As in a typical MDP, S_t is a random variable with a distribution conditioned on S_{t-1} and A_{t-1}. The full Bayes net is shown below.

The agent acting in the environment knows what state it is currently in (as is typical in the MDP setting). Unfortunately, however, we, the observer, cannot see the states S_t. Thus we are forced to use an adapted particle filtering algorithm to solve this problem. Concretely, we will develop an efficient algorithm to estimate $P(\Pi \mid a_{1:t})$.

(a) The Bayes net for part (a) is



(i) Select all of the following that are guaranteed to be true in this model for $t > 3$:

- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid S_{t-1}$
- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid S_{t-1}, A_{1:t-1}$
- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid \Pi$
- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid \Pi, A_{1:t-1}$
- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid \Pi, S_{t-1}$
- ☐ $S_t \perp\!\!\!\perp S_{t-2} \mid \Pi, S_{t-1}, A_{1:t-1}$
- ☐ None of the above

We will compute our estimate for $P(\Pi \mid a_{1:t})$ by coming up with a recursive algorithm for computing $P(\Pi, S_t \mid a_{1:t})$. (We can then sum out S_t to get the desired distribution; in this problem we ignore that step.)

(ii) Write a recursive expression for $P(\Pi, S_t \mid a_{1:t})$ in terms of the CPTs in the Bayes net above.

$P(\Pi, S_t \mid a_{1:t}) \propto$ _____

We now try to adapt particle filtering to approximate this value. Each particle will contain a single state s_t and a potential policy π_i.

(iii) The following is pseudocode for the body of the loop in our adapted particle filtering algorithm. Fill in the boxes with the correct values so that the algorithm will approximate $P(\Pi, S_t \mid a_{1:t})$.

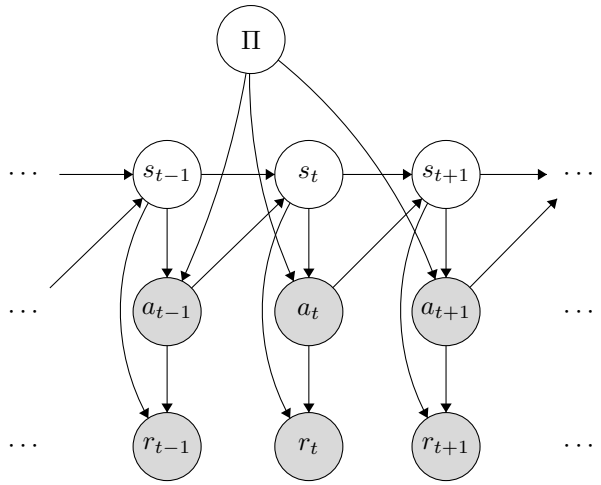
1. Elapse time: for each particle (s_t, π_i), sample a successor s_{t+1} from _____.

   The policy π' in the new particle is _____ .

2. Incorporate evidence: To each new particle (s_{t+1}, π'), assign weight _____.

3. Resample particles from the weighted particle distribution.

(b) We now observe the acting agent's actions *and* rewards at each time step (but we still don't know the states). Unlike the MDPs in lecture, here we use a stochastic reward function, so that R_t is a random variable with a distribution conditioned on S_t and A_t. The new Bayes net is given by



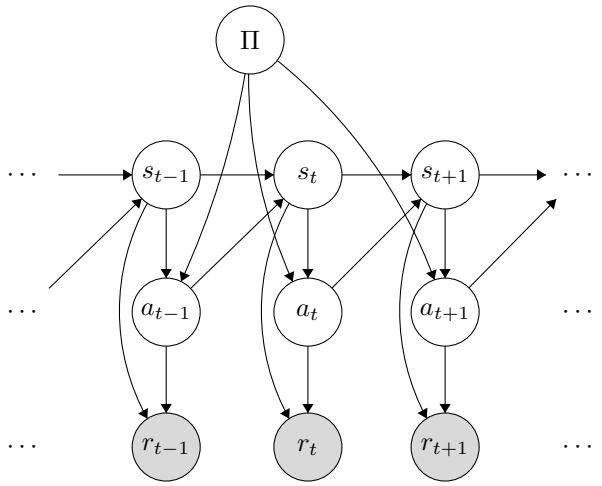
Notice that the observed rewards do in fact give useful information since d-separation does not give that $R_t \perp\!\!\!\perp \Pi \mid A_{1:t}$.

(i) Give an active path connecting R_t and Π when $A_{1:t}$ are observed. Your answer should be an ordered list of nodes in the graph, for example "$S_t, S_{t+1}, A_t, \Pi, A_{t-1}, R_{t-1}$".

(ii) Write a recursive expression for $P(\Pi, S_t \mid a_{1:t}, r_{1:t})$ in terms of the CPTs in the Bayes net above.

$P(\Pi, S_t \mid a_{1:t}, r_{1:t}) \propto$ _____

(c) We now observe *only* the sequence of rewards and no longer observe the sequence of actions. The new Bayes net is shown on the right.



(i) Write a recursive expression for $P(\Pi, S_t, A_t \mid r_{1:t})$ in terms of the CPTs in the Bayes net above.

$P(\Pi, S_t, A_t \mid r_{1:t}) \propto$ _____

We now try to adapt particle filtering to approximate this value. Each particle will contain a single state s_t, a single action a_t, and a potential policy π_i.

(ii) The following is pseudocode for the body of the loop in our adapted particle filtering algorithm. Fill in the boxes with the correct values so that the algorithm will approximate $P(\Pi, S_t, A_t \mid r_{1:t})$.

1. Elapse time: for each particle (s_t, a_t, π_i), sample a successor state s_{t+1} from _____.

   Then, sample a successor action a_{t+1} from _____.

   The policy π' in the new particle is _____.

2. Incorporate evidence: To each new particle (s_{t+1}, a_{t+1}, π'), assign weight _____.
3. Resample particles from the weighted particle distribution.

Q2. HMMs

Consider a process where there are transitions among a finite set of states $s_1, \cdots, s_k$ over time steps $i = 1, \cdots, N$. Let the random variables $X_1, \cdots, X_N$ represent the state of the system at each time step and be generated as follows:

- Sample the initial state s from an initial distribution $P_1(X_1)$, and set $i = 1$
- Repeat the following:
  1. Sample a duration d from a duration distribution P_D over the integers $\{1, \cdots, M\}$, where M is the maximum duration.
  2. Remain in the current state s for the next d time steps, i.e., set
  $$x_i = x_{i+1} = \cdots = x_{i+d-1} = s \tag{1}$$
  3. Sample a successor state s' from a transition distribution $P_T(X_t|X_{t-1} = s)$ over the other states $s' \neq s$ (so there are no self transitions)
  4. Assign $i = i + d$ and $s = s'$.

This process continues indefinitely, but we only observe the first N time steps.

(a) Assuming that all three states s_1, s_2, s_3 are different, what is the probability of the sample sequence $s_1, s_1, s_2, s_2, s_2, s_3, s_3$? Write an algebraic expression. Assume $M \geq 3$.

At each time step i we observe a noisy version of the state X_i that we denote Y_i and is produced via a conditional distribution $P_E(Y_i|X_i)$.

(b) Only in this subquestion assume that $N > M$. Let $X_1, \cdots, X_N$ and $Y_1, \cdots, Y_N$ random variables defined as above. What is the maximum index $i \leq N - 1$ so that $X_1 \perp\!\!\!\perp X_N | X_i, X_{i+1}, \cdots, X_{N-1}$ is guaranteed?

(c) Only in this subquestion, assume the max duration $M = 2$, and P_D uniform over $\{1, 2\}$ and each x_i is in an alphabet $\{a, b\}$. For $(X_1, X_2, X_3, X_4, X_5, Y_1, Y_2, Y_3, Y_4, Y_5)$ draw a Bayes Net over these 10 random variables with the property that removing any of the edges would yield a Bayes net inconsistent with the given distribution.

(d) In this part we will explore how to write the described process as an HMM with an extended state space. Write the states $z = (s, t)$ where s is a state of the original system and t represents the time elapsed in that state. For example, the state sequence $s_1, s_1, s_1, s_2, s_3, s_3$ would be represented as $(s_1, 1), (s_1, 2), (s_1, 3), (s_2, 1), (s_3, 1), (s_3, 2)$.

Answer all of the following in terms of the parameters $P_1(X_1), P_D(d), P_T(X_{j+1}|X_j), P_E(Y_i|X_i), k$ (total number of possible states), N and M (max duration).

(i) What is $P(Z_1)$?

$P(x_1, t_1) =$

(ii) What is $P(Z_{i+1}|Z_i)$? Hint: You will need to break this into cases where the transition function will behave differently.

$P(X_{i+1}, t_{i+1} \mid X_i, t_i) =$

(iii) What is $P(Y_i|Z_i)$?

$P(Y_i \mid X_i, t_i) =$

(e) In this question we explore how to write an algorithm to compute $P(X_N|y_1, \cdots, y_N)$ using the particular structure of this process.

Write $P(X_t|y_1, \cdots, y_{t-1})$ in terms of other factors. Construct an answer by checking the correct boxes below:

$P(X_t|y_1, \cdots, y_{t-1}) =$ **(i)** **(ii)** **(iii)**

(i)
- ◯ $\sum_{i=1}^{k} \sum_{d=1}^{M} \sum_{d'=1}^{M}$
- ◯ $\sum_{i=1}^{k} \sum_{d=1}^{M}$
- ◯ $\sum_{i=1}^{k}$
- ◯ $\sum_{d=1}^{M}$

(ii)
- ◯ $P(Z_t = (X_t, d)|Z_{t-1} = (s_i, d))$
- ◯ $P(X_t|X_{t-1} = s_i)$
- ◯ $P(X_t|X_{t-1} = s_d)$
- ◯ $P(Z_t = (X_t, d')|Z_{t-1} = (s_i, d))$

(iii)
- ◯ $P(Z_{t-1} = (s_d, i)|y_1, \cdots, y_{t-1})$
- ◯ $P(X_{t-1} = s_d|y_1, \cdots, y_{t-1})$
- ◯ $P(Z_{t-1} = (s_i, d)|y_1, \cdots, y_{t-1})$
- ◯ $P(X_{t-1} = s_i|y_1, \cdots, y_{t-1})$

(iv) Now we would like to include the evidence y_t in the picture. What would be the running time of each update of the **whole table** $P(X_t|y_1, \cdots, y_t)$?. Assume tables corresponding to any factors used in **(i)**, **(ii)**, **(iii)** have already been computed.

- ◯ $O(k^2)$
- ◯ $O(k^2M)$
- ◯ $O(k^2M^2)$
- ◯ $O(kM)$

Note: Computing $P(X_N|y_1, \cdots, y_N)$ will take time $N\times$ your answer in **(iv)**.

(v) Describe an update rule to compute $P(X_t|y_1, \cdots, y_{t-1})$ that is faster than the one you discovered in parts **(i)**, **(ii)**, **(iii)**. **Specify its running time**. Hint: Use the structure of the transitions $Z_{t-1} \to Z_t$.