IEOR@IITB

IE 708 Markov Decision Processes

MidSem Exam (19/Sep/2024)

Attempt all questions. If you feel the need to use results not discussed in the class, please state them clearly. State all assumptions and arguments clearly.

---

1. We saw in much detail 4 computational algos for discounted criteria models. For at least 3 of them, write one pro and one con for each algo.

   Soln. (3 = 1 x 3)

   Value Iteration: Iterate update is simple and has a simple stopping criteria; convergence can be slow.

   Policy Iteration: Optimal policy along with the optimal value is obtained; convergence can be slow for models with large state and action spaces.

   Modified Policy Iteration: Faster convergence and is the most popular algo. Like the other two, can't handle functional constraints.

   Linear Programming: Can easily incorporate functional constraints; can be slow, but, present day faster LP solvers can mitigate this.

2. A new job/customer arrives at a service facility each day with probability p, while the facility serves just one job in a day with probability q; if a job is serviced in a day, a new job even if available/waiting for service, is not taken up for service on that day. The facility earns ₹R for each admitted job, while it costs the facility ₹$h/job/day$ as holding charges. The facility manager can either admit or deny admission to each new job. Write an algo that maximizes the manager's expected total discounted revenue.

   Soln. (5) Let the states be $\mathcal{S} = \{0, 1, 2, \ldots\} \times \{0, 1\}$, with $(j, k) \in \mathcal{S}$ denoting j jobs in queue and k job waiting for admission. Let $\mathcal{A}_{(j,1)} = \{0, 1\}$ denote the manager's action to deny or admit a new job. When no new jobs arrive, $\mathcal{A}_{(j,0)} = \{0\}$

   Transition probabilities:

$$p((j',k')|(j,k),a) = \begin{cases} (1-p)(1-q), & \text{if } j' = j+1, k' = 0, k = 1, a = 1 \\ p(1-q), & \text{if } j' = j+1, k' = 1, k = 1, a = 1 \\ (1-p)q, & \text{if } j' = j > 0, k' = 0, k = 1, a = 1 \\ pq, & \text{if } j' = j > 0, k' = 1, k = 1, a = 1 \\ (1-p)(1-q), & \text{if } j' = j > 0, k' = 0, k \in \{0,1\}, a = 0 \\ p(1-q), & \text{if } j' = j > 0, k' = 1, k \in \{0,1\}, a = 0 \\ (1-p), & \text{if } j' = j = 0, k' = 0, k = 0, a = 0 \\ p, & \text{if } j' = j = 0, k' = 1, k = 0, a = 0 \\ (1-p)q, & \text{if } j' = j-1 \geq 0, k' = 0, k \in \{0,1\}, a = 0 \\ pq, & \text{if } j' = j-1 \geq 0, k' = 1, k \in \{0,1\}, a = 0 \\ 0, & \textit{otherwise} \end{cases}$$

   Rewards, $r((j,k),a) = R \times a \times k - h \times j$

Any of the algorithms from question 1. can be used with a discount rate $\alpha \in (0, 1)$ and a truncated state space with N states to get an approximate solution.

3. A system in working state can move to a failed state in the next day with probability 0.2 or continues to be working state with probability 0.8; the revenue in working state is ₹$10/day$. Option A restores failed system in a day with probability 0.6 and costs ₹$5/day$, but needs one extra day of repair with probability 0.4. Option B repairs the failed system in a day with probability 0.7 and costs ₹$6/day$, but needs one more day of repair with probability 0.3.

   (a) Determine the optimal expected total discounted revenue option using a provably correct and computationally efficient algo when the discount factor $\alpha = 0.95$.

   (b) Repeat the above for discount factor $\alpha = 0.3$.

   (c) Comment on the above.

   Don't forget to write key assumptions made.

   Soln. (5 = 1+1.5+1.5+1)

   States $\mathcal{S} = \{W, F1, F2A, F2B\}$ corresponding to working state, failed day-1 state and failed day-2 state. Actions $\mathcal{A}_{F1} = \{A, B\}$, $\mathcal{A}_{F2A} = \mathcal{A}_{F2B} = \mathcal{A}_W = \{\emptyset\}$. Transition probabilities,

$$
\begin{aligned}
p(W|W, \cdot) &= 0.8 \\
p(F1|W, \cdot) &= 0.2 \\
p(W|F2A, \cdot) &= 1 \\
p(W|F2B, \cdot) &= 1 \\
p(W|F1, a = A) &= 0.6 \\
p(F2A|F1, a = A) &= 0.4 \\
p(W|F1, a = B) &= 0.7 \\
p(F2B|F1, a = B) &= 0.3
\end{aligned}
$$

   Rewards,

$$
\begin{aligned}
r(W, \cdot) &= 10 \\
r(F1, A) &= -5 \\
r(F2A, \cdot) &= -5 \\
r(F1, B) &= -6 \\
r(F2B, \cdot) &= -6
\end{aligned}
$$

   Since, there will be deterministic stationary optimal policy we need to only check and compare the optimal value for two such policies $d_1(F1) = A$ and $d_2(F1) = B$.

(a) (1.5) $\alpha = 0.95$. With d_1,

$$\begin{aligned}
v^{d_1}(W) &= r(W,\cdot) + \alpha p(W|W,\cdot)v^{d_1}(W) + \alpha p(F1|W,\cdot)v^{d_1}(F1) \\
&= 10 + 0.76v^{d_1}(W) + 0.19v^{d_1}(F1) \\
0.24v^{d_1}(W) - 0.19v^{d_1}(F1) &= 10 \\
v^{d_1}(F1) &= r(F1,A) + \alpha p(W|F1,A)v^{d_1}(W) + \alpha p(F2A|F1,A)v^{d_1}(F2A) \\
&= -5 + 0.57v^{d_1}(W) + 0.38v^{d_1}(F2A) \\
0.57v^{d_1}(W) + 0.38v^{d_1}(F2A) - v^{d_1}(F1) &= 5 \\
v^{d_1}(F2A) &= r(F2A,\cdot) + \alpha p(W|F2A,\cdot)v^{d_1}(W) \\
&= -5 + 0.95v^{d_1}(W) \\
0.95v^{d_1}(W) - v^{d_1}(F2A) &= 5
\end{aligned}$$

$$v^{d_1}(W) \approx 137.68, v^{d_1}(F1) \approx 121.28, v^{d_1}(F2A) \approx 125.8$$

With d_2,

$$\begin{aligned}
v^{d_2}(W) &= r(W,\cdot) + \alpha p(W|W,\cdot)v^{d_2}(W) + \alpha p(F1|W,\cdot)v^{d_2}(F1) \\
&= 10 + 0.76v^{d_2}(W) + 0.19v^{d_2}(F1) \\
0.24v^{d_2}(W) - 0.19v^{d_2}(F1) &= 10 \\
v^{d_2}(F1) &= r(F1,B) + \alpha p(W|F1,B)v^{d_2}(W) + \alpha p(F2B|F1,B)v^{d_2}(F2 \\
&= -6 + 0.665v^{d_2}(W) + 0.285v^{d_2}(F2B) \\
0.665v^{d_2}(W) - v^{d_2}(F1) + 0.285v^{d_2}(F2B) &= 6 \\
v^{d_2}(F2B) &= r(F2B,\cdot) + \alpha p(W|F2B,\cdot)v^{d_2}(W) \\
&= -6 + 0.95v^{d_2}(W) \\
0.95v^{d_2}(W) - v^{d_2}(F2B) &= 6
\end{aligned}$$

$$v^{d_2}(W) \approx 137.20, v^{d_2}(F1) \approx 120.69, v^{d_2}(F2B) \approx 124.34$$

It makes sense to start from the working state. $v^{d_1}(W) > v^{d_2}(W)$. So, option A is optimal from W.

(b) (1.5) $\alpha = 0.3$

With d_1,

$$\begin{aligned}
v^{d_1}(W) &= 10 + 0.24v^{d_1}(W) + 0.06v^{d_1}(F1) \\
0.76v^{d_1}(W) - 0.06v^{d_1}(F1) &= 10 \\
v^{d_1}(F1) &= -5 + 0.18v^{d_1}(W) + 0.12v^{d_1}(F2) \\
0.18v^{d_1}(W) - v^{d_1}(F) + 0.12v^{d_1}(F2) &= 5 \\
v^{d_1}(F2A) &= -5 + 0.3v^{d_1}(W) \\
0.3v^{d_1}(W) - v^{d_1}(F2A) &= 5
\end{aligned}$$

Thus,

$$v^{d_1}(W) \approx 12.94, v^{d_1}(F1) \approx -2.81, v^{d_1}(F2A) \approx -1.12$$

With d_2,

$$\begin{aligned} v^{d_2}(W) &= 10 + 0.24v^{d_2}(W) + 0.06v^{d_2}(F1) \\ 0.76v^{d_2}(W) - 0.06v^{d_2}(F1) &= 10 \\ v^{d_2}(F1) &= -6 + 0.21v^{d_2}(W) + 0.09v^{d_2}(F2B) \\ 0.21v^{d_2}(W) - v^{d_2}(F1) + 0.09v^{d_2}(F2B) &= 6 \\ v^{d_2}(F2B) &= -6 + 0.3v^{d_2}(W) \\ 0.3v^{d_2}(W) - v^{d_1}(F2B) &= 6 \end{aligned}$$

The above give,

$$v^{d_2}(W) \approx 12.88, v^{d_2}(F1) \approx -3.49, v^{d_2}(F2B) \approx -2.14$$

$v^{d_1}(W) > v^{d_2}(W)$. So, option A is optimal from W.

(c) (1.5) With $\alpha = 0.95$, $v^d_{0.95}(W), v^d_{0.95}(F1) > 120$. While with $\alpha = 0.3$, $v^d_{0.3}(W) > 12$ and $v^d_{0.3}(F) < 0$, i.e., the initial state has significant impact.

With $\alpha = 0.3$, the future rewards are heavily discounted, while with $\alpha = 0.95$ this is not the case.

4. Show that $g(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ is superadditive when

$$g(s+1, a+1) - g(s+1, a) \geq g(s, a+1) - g(s, a)$$

where $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, \cdots\}$.

Soln. (2) Proof is straightforward, can be easily checked.

5. The tree diagram in Figure 1 shows the transition probabilities for a model with two states s_1, s_2 along with the probabilities for each action $a_{i,j}$ available at state s_i according to a randomized history dependent policy $\pi = (d_1, d_2)$.

   (a) If policy $\pi' = (d'_1, d'_2) \in \Pi^{MR}$ is such that it satisfies,

$$P^{\pi'}\{X_t = j, Y_t = a | X_1 = s_1\} = P^{\pi}\{X_t = j, Y_t = a | X_1 = s_1\} \tag{1}$$

   compute $q_{d'_2(s_1)}(a_{1,2})$.

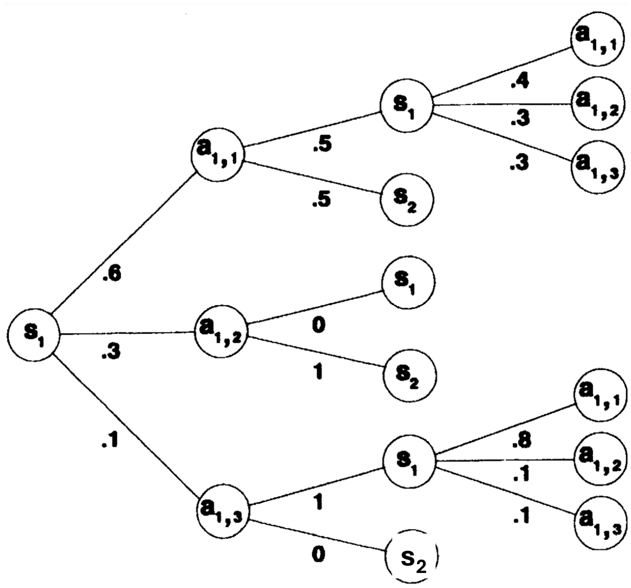

Figure 1

Soln. (2)

$$\begin{aligned}
q_{d'_2(s_1)}(a_{1,2}) &= P^{\pi'}\{Y_2 = a_{1,2}|X_2 = s_1\} \\
&= P^{\pi'}\{Y_2 = a_{1,2}|X_2 = s_1, X_1 = s_1\} \\
&= \frac{P^{\pi'}\{Y_2 = a_{1,2}, X_2 = s_1|X_1 = s_1\}}{P^{\pi'}\{X_2 = s_1|X_1 = s_1\}} \\
&= \frac{P^{\pi}\{Y_2 = a_{1,2}, X_2 = s_1|X_1 = s_1\}}{P^{\pi}\{X_2 = s_1|X_1 = s_1\}} \qquad \text{(check as exercise)} \\
&= \frac{P^{\pi}\{Y_2 = a_{1,2}, X_2 = s_1, X_1 = s_1\}}{P^{\pi}\{X_2 = s_1, X_1 = s_1\}} \\
&= \frac{(0.6)(0.5)(0.3) + (0.1)(1)(0.1)}{(0.6)(0.5) + (0.1)(1)} \\
&= 0.25
\end{aligned}$$

(b) Interpret the above equality (1)

Soln. (1) The state-action frequencies of the constructed Markov Randomized policy equals those of the given History Randomized policy.

(c) Write a consequence of the above; outline your arguments

Soln. (2)

As a consequence of the above state-action frequencies being the same, the discounted cost of the *given/arbitrary* HR policy equals that of the *constructed* MR policy.

Further consequence is more important: First, for the optimal policy, we just need to search MR class and the *optimal policy* would be Markov Randomized one.

Now, combine this with an earlier argument that one need not randomise for this MDP as this is unconstrained MDP; *optimal policy would be Markov deterministic*; the argument was simpler.

This conclusion of optimal policy being Markov Deterministic also holds for *average cost* and *total cost* models also, as these costs depend just on state-action frequencies.