Practice Problem Set 3

IE 708 Markov Decision Processes
IEOR, IIT Bombay

Monsoon Sem '24

Exercise 1 Existence of Blackwell Optimal Policies Recall for a Markov policy

$$v_{\alpha,\pi} = (I - \alpha P_\pi)^{-1} r_\pi$$

The matrix inverse is ratio of determinants of submatricies (Harald Cramer) and hence each entry is a ratio of polynomials of α. Fix a state s. For policies π and π', these polynomials intersect at finite number of α values. So, for state s, one of them is better for α that is to the right of the last intersection.

As we have finite number of Markov Deterministic policies, there is a discount factor α_s for which one MD policy π_s is optimal for all $\alpha \in (\alpha_s, 1)$ for some $\alpha_s \in (0,1)$. This policy π_s satisfies the OE, state/component wise.

$$v_\alpha(s) = \min[r(s,a) + \alpha \sum_j p_{sj}(a) v_\alpha(j)]$$

Now observe that π_s satisfies OE for all discount factors α over the shrunk interval $(\max_s \alpha_s, 1)$ and hence is optimal for discount factors over the interval $(\max \alpha_s, 1)$. Thus, for each state s there is an optimal policy π_s that is optimal over the interval $(\max \alpha_s, 1)$.

Next, for each state s define a policy $\bar{\pi}(s) = \pi_s(s)$. Show that policy $\bar{\pi}$ attains the minima in the OE; *i.e., satisfies* OE. Couple of comments: Think of policy as a matrix with states as rows and columns as actions. For policy $\bar{\pi}$, actions are picked 'diagonally' – this is called diagonal argument. To argue that it satisfies the OE, what is the value vector for this OE? What is the intuition to pick actions diagonally? Remember discounted model takes actions 'greedily' and $v_\alpha(i)$ is discounted cost when initial state is i.

Since $\bar{\pi}$ satisfies OE for all $\alpha \in (\max \alpha_s, 1)$, it is Blackwell optimal; the interval being $(\max \alpha_s, 1)$ which is common across states.

Exercise 2 Here is an interpretation of the relation between average and discounted rewards/costs of a Markov Reward Process, MRP. This is basically Laurent series expansion of the discounted cost for the first terms.

Consider a terminating reward/cost processes where a random time/variable τ determines the termination/stopping time of the Markov Chain with τ being independent of the MC. And on termination, cost $c(X_\tau)$ is incurred (with no running cost). Let τ be geometric: $P(\tau = n) = \alpha^{n-1}\alpha$, $n = 1, 2, \cdots$.

- Then, verify that $E_s[c(X_\tau] = (1 - \alpha)v_\alpha(s)$.
- Now verify that $\lim_{\alpha \uparrow 1} E_s[c(X_\tau] = E_s[c(X_\infty)] = g$.

  You can assume that the interchange of limiting and expectation operations in the above is valid.
- Hence $g = \lim_{\alpha \uparrow 1}(1 - \alpha)v_\alpha(s)$.
- We also considered power series expansion of discounted cost v_α in the class that has $\cdots$. Complete this sentence.

Exercise 3 Consider the familiar 2-state MC where the state flips deterministically. Let the cost/reward vector be $(1, -1)$.

- Write the limiting matrix $P^\star$.
- What is $v_n(s_1)$, for $n = 1, 2, \cdots$?
- Write the gain g.
- Write the fundamental matrix H_P.
- And the bias $h = H_P r$.
- Verify that $v_n(s_1)$ converges in Cesaro sense, but not in the usual sense.
- Identify the above limit.
- Verify $h(s_1) - h(s_2) = \lim \frac{1}{N} \sum_1^N [v_n(s_1) - v_n(s_2])$
- Explain the need for the expression on RHS (why not a simpler one?)

Exercise 4 Consider a two state MDP. State A has two actions. One action makes the system continue in the same state in the next period and cost is 1 unit. The other action moves the system to the other state at cost of -10 units. The single action in the state B makes the system to continue in this state for one more period and the cost is 2 units.

- Find the Blackwell optimal policies.
- Find the associated discount interval.
- Check if the action set in the second optimization problem has to be restricted.
- Find the average optimal policy.
- Check if the optimal policies for α values lower than $\alpha_{\max}$ are optimal over intervals.

- If so, identify these intervals.

Exercise 5 Overtaking optimality

Exercise 6 Optimal machine replacement

Exercise 7 Sensitive discount optimality

This is a family of criteria that uses comparison of limiting behavior the (expected total) discounted cost as the discount rate approaches to 1.

For a given $n \in \{-1, 0, 1, \cdots\}$ a policy $\pi^\star$ is called *n-discount optimal*

$$\liminf_{\alpha \uparrow 1}(1-\alpha)^{-n}[v_\alpha^{\pi^\star}(s) - v_\alpha^{\pi}(s)] \geq 0 \quad \text{if for each state s} \in \mathcal{S},$$

for all policies $\pi \in HR$.

- Show that (-1)-discount optimality is same as gain or average optimality (see above).
- Show 0-discount optimality is *bias* optimality.
- Show and interpret this monotonicity property: n-discount optimality implies m-discount optimality for all $m < n$. This means that larger n is more selective.

  Recall the nested/coupled optimization problems while solving average cost OE equations. And hence, those give us gain and bias optimal policies. So, in case of multiple average optimal polices they give us that average optimal policy which is bias optimal as well. This is one way to break the ties when there are more than one (average) optimal.

  Thus, n-discount optimality, and hence, Blackwell optimality, is a selection criteria.
- A policy is called ∞-optimal if it is n-discount optimal for all $n \geq -1$.
- A fundamental theorem is that Blackwell optimality is same as ∞-optimality.

  Putting two things together, this means that existence of a $\bar{\alpha} < 1$ is same as the existence of a policy that is n-discount optimal for all $n \geq -1$.

Exercise 8 Blackwell interval Read up an example of identifying Blackwell interval; we did one in the class.