Practice Problem Set 2

IE 708 Markov Decision Processes
IEOR, IIT Bombay

Monsoon Sem '24

Exercise 1 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 revenue over the next N days.

Exercise 2 Does discount factor affect any theoretical results or algorithms in the finite-horizon case? If not, what exactly did it change?

Exercise 3 Why are utility functions used to compare policies rather than vector stochastic partial orders? What is the alternative of expected utility criteria?

Exercise 4 Prove the "Principle of Optimality" in finite horizon MDPs. If a policy is optimal for a subproblem, than show that when combined with the optimal policy of remaining time steps, it is the optimal policy for the whole problem.

Exercise 5 Show that $u_t(h_t)$ depends on state s_t alone, for all $t = 1, \cdots, N$.

Exercise 6 Each quarter the marketing manager of a retail store divides customers into two classes based on their purchase behavior in the previous quarter. Denote the classes as L for low and H for high. The manager wishes to determine to which classes of customers he should send quarterly catalogs. The cost of sending a catalog is ₹15 per customer and the expected purchase depends on the customer's class and the manager's action. If a customer is in class L and receives a catalog, then the expected purchase in the current quarter is ₹20, and if a class L customer does not receive a catalog his expected purchase is ₹10. If a customer is in class H and receives a catalog, then his expected purchase is ₹50, and if a class H customer does not receive a catalog his expected purchase is ₹25.

The decision whether or not to send a catalog to a customer also affects the customer's classification in the subsequent quarter. If a customer is class L at

the start of the present quarter, then the probability he is in class L at the subsequent quarter is 0.3 if he receives a catalog and 0.5 if he does not. If a customer is class H in the current period, then the probability that he remains in class H in the subsequent period is 0.8 if he receives a catalog and 0.4 if he does not. Assume a discount rate of 0.9 and an objective of maximizing expected total discounted reward.

(a) Formulate this as an infinite-horizon discounted Markov decision problem.

(b) Write all the states, actions available at these states, transition probabilities and the associated rewards appropriately.

(c) Find an optimal policy using policy iteration starting with the stationary policy which has greatest one-step reward.

Exercise 7 The Gulmohar restaurant in our campus has started showing a digital menu to its customers, based on their previous behavior. Tushar is one such customer. The menu shows him N items, out of which he will purchase at most one item. Tushar starts browsing the menu from top to bottom. If he likes an item i (which happens with probability β_i), he buys it for the price p_i and stops scanning the menu. Otherwise, that is, when he does not like the item, he either leaves the restaurant with probability $1-\gamma$, or he continues to the next item $(i+1)$ with probability γ. If he does not like the last item in the menu, he stops browsing the menu and leaves the restaurant without purchasing anything.

Help the restaurant want to optimally place the items in the menu such that their expected revenue is maximized. Towards this,

(a) For a given permutation $\pi = \{i_1, i_2, \cdots, i_N\}$ of the menu of N items, write down the expected revenue that can be acquired by Gulmohar from Tushar.

(b) Identify whether this problem can be modeled as a discounted Markov decision problem.. If yes, what is the discount factor?

(c) Formulate this as a finite-horizon Markov decision problem.

(d) Write all the states, actions available at these states, transition probabilities and the associated rewards appropriately.

(e) If π^* is the optimal permutation of arranging the menu to derive the maximum revenue, then show that if item i is placed before item j in π^*, then

$$\frac{\beta_i p_i}{1-\gamma\beta_i} \geq \frac{\beta_j p_j}{1-\gamma\beta_j}.$$

Exercise 8 Solve the example **6.2.3.** from the book by Martin L. Puterman.

Exercise 9 Recall the fundamental result that for any HR policy π, there is a MR policy π' such that both have the same state-action frequencies a result in the Chapter, 'Infinite horizon models: Foundations', of the classic by Martin Puterman.

There are a couple of typos in the proof of this result; identify them.

Exercise 10 For probability transition matrix P for finite MDP, argue that $(I - \lambda P)^{-1}$ exists for $\lambda \in (0, 1)$.

Exercise 11 Recall this policy improvement step in PI and MPI algos:
$d_{n+1} \in \arg\max \{r_d + \lambda P_d v^n\}$
setting $d_{n+1} = d_n$, if possible.

- First, justify this step
- Now, write a pseudo code snippet for the above

Exercise 12 Interpret modified policy itetation, MPI, algo in terms of value iteration, VI, algo and policy itereation, PI, algo.

Exercise 13 Write the relationship between PI and linear programming, LP, algo. (*this seems to have been first identified by Peter Whittle, Stats Lab, Cambridge*)