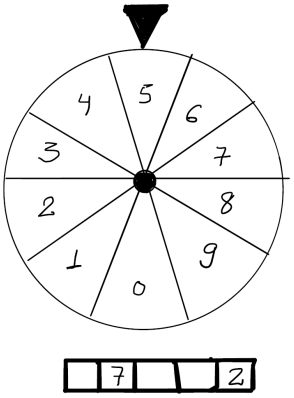
ORCS 4529: Reinforcement Learning

Homework 1

Instructor: Shipra Agrawal

Problem 1. (25 points) Consider the following single-player version of the two-player game "So who's counting" that originally appeared in a television program on the U.S. public broadcasting network. At each of the five consecutive rounds of this game, a spinner produces a number between 0 and 9, each with equal probability. After each spin, the player selects an available digit of a five digit number to place the number produced by the spinner. The score of the player is equal to the value of the five digit number generated at the end of the five rounds.



Formulate the problem of maximizing the expected five-digit number in this game as a Markov Decision Process (MDP). This invovles formulating the state space, the action space, the reward model, the transition model, and the goal of the MDP.

Problem 2. (25 points) Consider a customer shopping for three rounds. In each round, there are two products available for recommendation, at price \$1 and price \$2, respectively. For each product, the customer can be either "eager" or "uninterested" to buy that product. In the beginning of round one, the customer is "eager" for both products.

In the beginning of each round, the seller observes the customer's excitement (eager or uninterested) for each product, and then decides which (exactly one) of the two products to recommend. The customer reacts to the recommendation in the following way: if eager for the recommended product, the customer buys the product and pays its price, and if uninterested, the customer rejects the product. The customer cannot buy the product that wasn't recommended.

At the end of the round, the customer' excitement ("eager" or "uninterested") for the each product changes in the following way. If product #1 was recommended in this round: then the customer's excitement for product #1 flips with probability 0.1 and stays the same with probability 0.9. if product #2 was recommended in this round, then the customer excitement for product #2 flips with probability 0.5 and stays the same with probability 0.5. The excitement for the product that wasn't recommended in this round remains unaffected.

The goal of the seller is to maximize the total expected revenue over the three rounds.

Formulate the above recommendation problem as an MDP. Use dynamic programming to find an optimal (possibly non-stationary) policy that maximizes the seller's total expected revenue over the three periods.

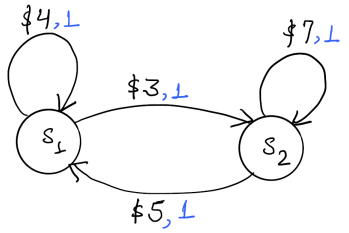
Problem 3. (25 points) In this problem, we consider a family of sensitive optimality criteria that generalize the two criteria of average and discounted reward optimality discussed in class. We say that a policy π^* is n-discount optimal for $n = -1, 0, 1, 2, \ldots$, if for each $s \in S$,

$$\lim_{\gamma \to 1} (1-\gamma)^{-n} \left(V_\gamma^{\pi^*}(s) - V_\gamma^{\pi}(s) \right) \geq 0, \text{ for all policies } \pi$$

Observe that (-1)-discount optimality is equivalent to average reward optimality; and (0)-discount optimality is equivalent to bias-optimality. Further, we say a policy is ∞-discount optimal if it is n-discount optimal for all $n \geq -1$. And a policy π^* is Blackwell optimal if for each $s \in S$, there exists a $0 \leq \gamma^*(s) < 1$ such that

$$V_\gamma^{\pi^*}(s) - V_\gamma^{\pi}(s) \geq 0, \text{ for all } \pi, \text{ for all } \gamma^*(s) \leq \gamma < 1$$

Consider the following two-state MDP with two states and two actions available in each state. All the transitions are deterministic. The numbers with the dollar sign on the arrows indicate the rewards on taking the respective action, and the numbers in blue indicate the probabilities of transition. That is, $\Pr(s_1|a_1, s_1) = 1, r(s_1, a_1) = 4$, $\Pr(s_2|a_2, s_1) = 1, r(s_1, a_2) = 3$, $\Pr(s_2|a_1, s_2) = 1, r(s_2, a_1) = 7$, $\Pr(s_1|a_2, s_2) = 1, r(s_2, a_2) = 5$.



(a) Compute the expressions for infinite horizon expected total discounted rewards $V_\gamma^\pi(s)$, (in terms of γ) for each deterministic stationary policy π and state $s \in \{s_1, s_2\}$. Also, compute the infinite horizon expected average reward for each deterministic stationary policy π and each state. (Note that there are four possible deterministic stationary policies in the given MDP).

(b) Use the expressions computed in part (a) to find all deterministic stationary policies which are

(1) (-1)-discount optimal

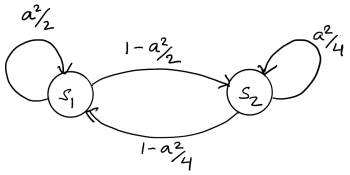
(2) 0-discount optimal

(3) 1-discount optimal

(4) ∞-discount optimal

(c) Find a Blackwell optimal policy. State $\gamma^*(s)$ for each state s.

Problem 4. (25 points) Consider the following MDP with two states $S = \{s_1, s_2\}$ and three actions $A = \{0, \frac{1}{2}, 1\}$. The expressions on the arrows indicate the probability of corresponding transition on taking an action $a \in A$. That is, $\Pr(s_1|s_1, a) = a^2/2, \Pr(s_2|s_2, a) = a^2/4$. Rewards are given by $r(s_1, a) = -a$, $r(s_2, a) = -1 + \frac{a}{12}$.



Solve this MDP (i.e., find a stationary policy that maximizes expected discounted reward) for $\gamma = 0.5$, using policy iteration and value iteration. For policy iteration, start with $\pi^0(s_1) = 0, \pi^0(s_2) = 0$. For value iteration, you may

start with $v^0(s_1) = 0, v^0(s_2) = 0$. You may execute these algorithms either by hand or using a computer program. (Approximate answers rounded to two decimal places will be accepted).

In your solution, copy the code, and provide the value vector v^k for at least 4 iterations of value iteration, and policy π^k for at least 4 iterations of policy iteration.

You are required to implement value iteration and policy iteration in your code, and not use a built in tool like MDP toolbox.