CSE 250a. Assignment 9

Out: *Wed Nov 30*
Due: *Wed Dec 07* (by 11:59 PM, Pacific Time, via gradescope)
Grace period: *none*.
Reading: *Sutton & Barto*, Chapters 3.1-4.4

9.1 Effective horizon time

Consider a Markov decision process (MDP) whose rewards $r_t \in [0, 1]$ are bounded between zero and one. Let $h = (1-\gamma)^{-1}$ define an *effective* horizon time in terms of the discount factor $0 \le \gamma < 1$. Consider the approximation to the (infinite horizon) discounted return,

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4 + \ldots,$$

obtained by neglecting rewards from some time t and beyond. Recalling that $\log\gamma \le \gamma - 1$, show that the error from such an approximation decays exponentially as:

$$\sum_{n\ge t} \gamma^n r_n \;\le\; h e^{-t/h}.$$

Thus, we can view MDPs with discounted returns as similar to MDPs with finite horizons, where the finite horizon $h = (1-\gamma)^{-1}$ grows as $\gamma \to 1$. This is a useful intuition for proving the convergence of many algorithms in reinforcement learning.

9.2 Three-state, two-action MDP

Consider the Markov decision process (MDP) with three states $s \in \{1, 2, 3\}$, two actions $a \in \{\uparrow, \downarrow\}$, discount factor $\gamma = \frac{2}{3}$, and rewards and transition matrices as shown below:

s	$R(s)$
1	-15
2	30
3	-25

s	s'	$P(s'\|s, a=\uparrow)$
1	1	$\frac{3}{4}$
1	2	$\frac{1}{4}$
1	3	0
2	1	$\frac{1}{2}$
2	2	$\frac{1}{2}$
2	3	0
3	1	0
3	2	$\frac{3}{4}$
3	3	$\frac{1}{4}$

s	s'	$P(s'\|s, a=\downarrow)$
1	1	$\frac{1}{4}$
1	2	$\frac{3}{4}$
1	3	0
2	1	0
2	2	$\frac{1}{2}$
2	3	$\frac{1}{2}$
3	1	0
3	2	$\frac{1}{4}$
3	3	$\frac{3}{4}$

(a) **Policy evaluation**

Consider the policy π that chooses the action shown in each state. For this policy, solve the linear system of Bellman equations (by hand) to compute the state-value function $V^\pi(s)$ for $s \in \{1, 2, 3\}$. Your answers should complete the following table. (*Hint:* the missing entries are integers.) **Show your work for full credit.**

s	$\pi(s)$	$V^\pi(s)$
1	$\uparrow$	-18
2	$\uparrow$	
3	$\downarrow$	

(b) **Policy improvement**

Compute the greedy policy $\pi'(s)$ with respect to the state-value function $V^\pi(s)$ from part (a). Your answers should complete the following table. **Show your work for full credit.**

s	$\pi(s)$	$\pi'(s)$
1	$\uparrow$	
2	$\uparrow$	
3	$\downarrow$	

9.3 Value function for a random walk

Consider a Markov decision process (MDP) with discrete states $s \in \{0, 1, 2, \ldots, \infty\}$ and rewards $R(s) = s$ that grow linearly as a function of the state. Also, consider a policy π whose action in each state either leaves the state unchanged or yields a transition to the next highest state:

$$P(s'|s, \pi(s)) \;=\; \begin{cases} \frac{2}{3} & \text{if} \quad s' = s \\[2ex] \frac{1}{3} & \text{if} \quad s' = s+1 \\ 0 & \text{otherwise} \end{cases}$$

Intuitively, this policy can be viewed as a right-drifting random walk. As usual, the value function for this policy, $V^\pi(s) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle|\, s_0 = s\right]$, is defined as the expected sum of discounted rewards starting from state s, with discount factor $0 \leq \gamma < 1$.

(a) Assume that the value function $V^\pi(s)$ satisfies a Bellman equation analogous to the one in MDPs with finite state spaces. Write down the Bellman equation satisfied by $V^\pi(s)$.

(b) Show that one possible solution to the Bellman equation in part (a) is given by the linear form $V^\pi(s) = as + b$, where a and b are coefficients that you should express in terms of the discount factor γ. (*Hint*: substitute this solution into both sides of the Bellman equation, and solve for a and b by requiring that both sides are equal for all values of s.)

(*) *Challenge (optional, no credit):* justify that the value function $V^\pi(s)$ has this linear form. In other words, rule out other possible solutions to the Bellman equation for this policy.
