Reinforcement Learning CAP 6629, Homework 1

Professor Xiangnan Zhong

Matthew Acs, Fall 2023

Question 1 (10 points):

Suppose $\gamma = 0.5$ and the following sequence of rewards is received $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, with $T = 5$. What are $G_0, G_1, \ldots, G_5$?
Hint: Work backwards.

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots$
$G_t = R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots)$
$G_t = R_{t+1} + \gamma (G_{t+1})$
$\gamma = 0.5$

$G_5 = 0$
$G_4 = R_5 + \gamma(G_5) = 2 + 0.5(0) = 2 + 0 = 2$
$G_3 = R_4 + \gamma(G_4) = 3 + 0.5(2) = 3 + 1 = 4$
$G_2 = R_3 + \gamma(G_3) = 6 + 0.5(4) = 6 + 2 = 8$
$G_1 = R_2 + \gamma(G_2) = 2 + 0.5(8) = 2 + 4 = 6$
$G_0 = R_1 + \gamma(G_1) = -1 + 0.5(6) = -1 + 3 = 2$

$G_5 = 0$ ***$G_4 = 2$*** ***$G_3 = 4$*** ***$G_2 = 8$*** ***$G_1 = 6$*** ***$G_0 = 2$***

Question 2 (15 points):

Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots$
$\gamma = 0.9$

$G_1 = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots = 7 + \gamma(7) + \gamma^2(7) + \gamma^3(7) + \ldots = 7(1 + \gamma(1) + \gamma^2(1) + \gamma^3(1) + \ldots) = 7 (1 / (1- \gamma)) = 7 / (1- \gamma)$

$G_1 = 7 / (1- \gamma) = 7 / (1-0.9) = 7 / 0.1 = 70$
$G_0 = R_1 + \gamma(G_1) = 2 + 0.9(70) = 2 + 63 = 65$

$G_1 = 70$
$G_0 = 65$

Question 3 (20 points):

In the gridworld example, rewards are positive for goals, negative for running into the edge of the world, and zero the rest of the time. Are the signs of these rewards important, or only the intervals between them? Prove, that adding a constant c to all the rewards adds a constant, v_c, to the values of all states, and thus does not affect the relative values of any states under any policies. What is v_c in terms of c and γ?

Hint: Consider the return as the discounted sums of future rewards: $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

The signs for the rewards are not important, only the interval between them. This is because adding a constant c to each reward results in adding a constant v_c to the values of all states, which does not affect the relative values of any state under any policy. The constant V_c is equal to c/(1- γ), as shown below.

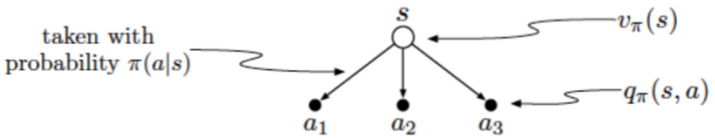
$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots$

$G'_t = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k c = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \frac{c}{1-\gamma} = G_t + \frac{c}{1-\gamma}$

$\boldsymbol{v_c = \frac{c}{1-\gamma}}$

Question 4 (15 points):

The value of a state depends on the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action:



Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$. This equation should include an expectation conditioned on following the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation.

---

1. $\mathbf{v_\pi(s) = E_\pi[\, q_\pi(s, a_1)\,] + E_\pi[\, q_\pi(s, a_2)\,] + E_\pi[\, q_\pi(s, a_3)\,]}$

2. $\mathbf{v_\pi(s) = \pi(a_1|s)\, q_\pi(s, a_1) + \pi(a_2|s)\, q_\pi(s, a_2) + \pi(a_3|s)\, q_\pi(s, a_3)}$