Adan Constanzo

MDPs - The Value Function

HW 04

**Exercise 3.7**

Imagine that you are designing a robot to run a maze. You decide to give it a

reward of +1 for escaping from the maze and a reward of zero at all other times.The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward (3.7). After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?

Since there has not been any form of improvement through an episodic task, this means that the decision value process needs to be adjusted to include the use of discounting that than allows to maximize the expected return.

**Exercise 3.8**

Suppose γ = 0.5 and the following sequence of rewards is received R1 = −1, R2 = 2, R3 = 6, R4 = 3, and R5 = 2, with T = 5. What are G0, G1, . . ., G5?

Hint: Work backwards.

Since we know the Terminating state, T= 5, We can calculate GT = 0, so

G5 = 0

G4  = R4+1 + 0.5(G4+1) = 2 + 0.5(0) = 2

G3 = R3+1 + 0.5(G3+1) = 3 + 0.5(2) = 4

G2 = R2+1 + 0.5(G2+1) = 6 + 0.5(4) = 8

G1 = R1+1 + 0.5(G1+1) = 2 + 0.5(8) = 6

G0 = R0+1 + 0.5(G0+1) = -1 + 0.5(6) = 2

**Exercise 3.9**

Suppose γ = 0. 9 and the reward sequence is R1 = 2 followed by an infinite sequence of 7s. What are G1 and G0 ?

= 10

**Exercise 3.10**

Prove the second equality in (3.10).

Let x =

Since it’s a recursive function.

x = 0 + x

x - x = 0

x ( 1 - ) = 1

x =

**Exercise 3.14**

The Bellman equation (3.14) must hold for each state for the value function

vπ shown in Figure 3.2 (right) of Example 3.5. Show numerically that this equation holds

for the center state, valued at +0.7, with respect to its four neighboring states, valued at

+2.3, +0.4, −0.4, and +0.7. (These numbers are accurate only to one decimal place.)