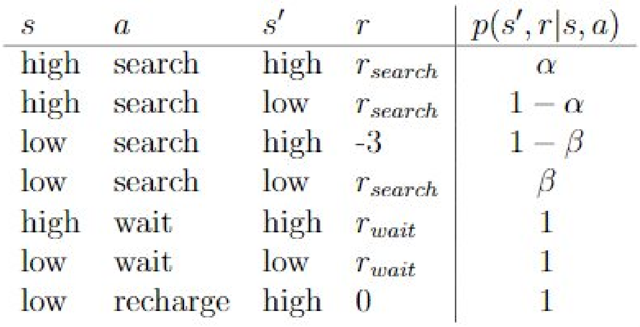
Exercise 3.4

Give a table analogous to that in Example 3.3, but for p(s0, r|s, a). It should have columns for s, a, s0, r, and p(s0, r|s, a), and a row for every 4-tuple for which p(s0, r|s, a) > 0.

Answer



s	a	s'	r	$p(s', r \mid s, a)$
high	search	high	r_{search}	α
high	search	low	r_{search}	$1-\alpha$
low	search	high	-3	$1-\beta$
low	search	low	r_{search}	β
high	wait	high	r_{wait}	1
low	wait	low	r_{wait}	1
low	recharge	high	0	1

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.

Answer

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma \left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots \right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

G5 = 0.
G4 = R5 + gamma*G5 = R5 = 2.
G3 = R4 + 0.5*2 = 4
G2 = R3 + 0.5*4 = 8
G1 = R2 + 0.5*8 = 6
G0 = R1 + 0.5*6 = 4

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?

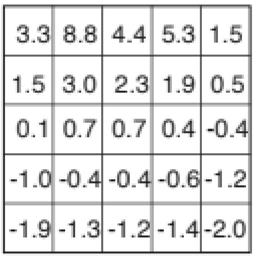
Answer

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}.$$

G1 = Rk * (1 / 0.1) = 7*10 = 70.
G0 = 2 + 0.9*70 = 65.

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 centre state, valued at +0.7, with respect to its four neighbouring states, valued at +2.3, +0.4, -0.4, and +0.7. (These numbers are accurate only to one decimal place.). Gamma = 0.9



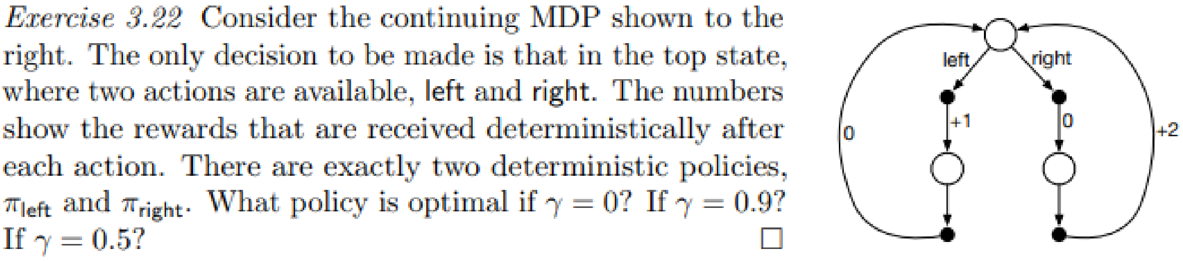
3.3	8.8	4.4	5.3	1.5
1.5	3.0	2.3	1.9	0.5
0.1	0.7	0.7	0.4	-0.4
-1.0	-0.4	-0.4	-0.6	-1.2
-1.9	-1.3	-1.2	-1.4	-2.0

Answer

0.25 * (0 + 2.3*0.9) + 0.25 (0+0.4*0.9) + 0.25*(0 + (-0.4)*0.9) + 0.25*(0 + 0.7*0.9) =

= 0.5175 + 0.09 + (-0.09) + 0.1575 = 0.675 ~= 0.7. (holds for the centre state with reward 0.7).

Exercise 3.22



Exercise 3.22 Consider the continuing MDP shown to the right. The only decision to be made is that in the top state, where two actions are available, left and right. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, π_{left} and π_{right}. What policy is optimal if $\gamma = 0$? If $\gamma = 0.9$? If $\gamma = 0.5$? □

Answer

With gamma = 0, the agent is greedy, and does not take into account future rewards. Therefore, an optimal policy will be *$pi_{left.}$*

With gamma = 0.9, the agent is very considerate about potential rewards in the long run. Hence, policy pi$_{right}$ is best.

With gamma = 0.5, the agent is neither too greedy, nor too cautious about the long-term rewards, therefore both policies are optimal.