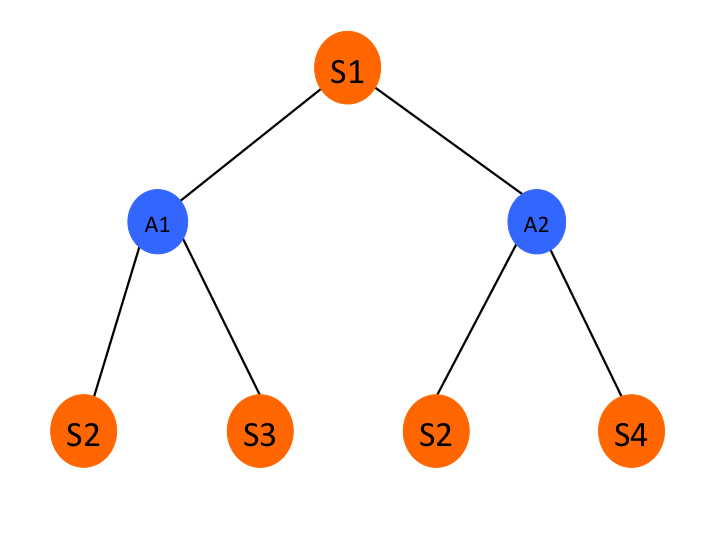
1) Suppose you have the very simple MDP shown below, where S2, S3, and S4 are terminal states.



The policy for S1 is 0.5 A1 and 0.5 A2. The rewards and transition probabilities are given in the table below.

|  |  |  |
| --- | --- | --- |
| Transition (s,a,s’) | Reward | P(s’|s,a) |
| S1 A1 S2 | 10 | 0.2 |
| S1 A1 S3 | 5 | 0.8 |
| S1 A2 S2 | 20 | 0.6 |
| S1 A2 S4 | 10 | 0.4 |

1) Compute Q(S1, A1):

P(S2|S1,A1)\*[r(S1,S2)] + P(S3|S1,A1)\*[r(S1,S3)] => 0.2\*10 + 0.8\*5 => 6

2) Compute Q(S1, A2):

P(S2|S1,A2)\*[r(S1,S2)] + P(S4|S1,A2)\*[r(S1,S4)] => 0.6\*20 + 0.4\*10 => 16

3) Compute V(S1): 0.5(6)+0.5(16)=> 3+8 => 11

2) Suppose you have the very simple MDP shown below, where S2 has value 5, S3 has value 2, and S4 has value 8.

A picture containing orange, small, man, holding

Description automatically generated

**5 2 5 8**

The policy for S1 is 0.25 A1 and 0.75 A2. The rewards and transition probabilities are given in the table below. The discount rate is 1.

|  |  |  |
| --- | --- | --- |
| Transition (s,a,s’) | Reward | P(s’|s,a) |
| S1 A1 S2 | 4 | 0.5 |
| S1 A1 S3 | 5 | 0.5 |
| S1 A2 S2 | 6 | 0.1 |
| S1 A2 S4 | 10 | 0.9 |

1) Compute Q(S1, A1):

P(S2|S1,A1)\*[r(S2,S1)+(1)5] + P(S3|S1,A1)\*[r(S3,S1)+(1)2]

0.5\*[4+5] + 0.5\*[5+2] => 4.5+3.5 => 8

2) Compute Q(S1, A2):

P(S2|S1,A2)\*[r(S2,S1)+(1)6] + P(S4|S1,A1)\*[r(S4,S1)+(1)8]

0.1\*[6+5] + 0.9\*[10+8] => 1.1+16.2=> 17.3

3) Compute V(S1) : 0.25\*8 + 0.75\*17.3 => 14.975

3) Consider the *MDP* defined below with four states (S1-S4) and two actions (A1 and A2). The tables below define the transition function and reward function.

Transition probabilities for A1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| S1 | 0 | 0.2 | 0.4 | 0.4 |
| S2 | 0.5 | 0.5 | 0 | 0 |
| S3 | 0 | 0.1 | 0 | 0.9 |
| S4 | 0.4 | 0.4 | 0.2 | 0 |

Transition probabilities for A2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| S1 | 0.5 | 0.3 | 0.2 | 0 |
| S2 | 0 | 0 | 0.5 | 0.5 |
| S3 | 0.6 | 0.2 | 0.2 | 0 |
| S4 | 0 | 0.5 | 0.4 | 0.1 |

Reward function

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | S1 | S2 | S3 | S4 |
| S1 | 0 | 1 | 0 | 4 |
| S2 | 0 | 0 | 0 | 2 |
| S3 | 0 | 1 | 0 | 1 |
| S4 | 0 | 0 | 2 | 0 |

The initial policy is uniform random, so A1 and A2 are both selected with probability 0.5 in every state. The discount rate (lambda) is 0.9. The initial state values are all set to 0.

Perform two rounds of value iteration; in each round you should loop through S1 through S4, updating each value once, so you will do a total of two updates for each state.

A picture containing text

Description automatically generated

V\_1(S1) = 0.5(A1) + 0.5(A2)

A1 = 0 + 0.2(1+0.9\*(0)) + 0.4(0+0.9(0)) +0.4(4+0.9(0)) => 0.2+1.6 => 1.8

A2 = 0.5(0+0.9\*(0)) + 0.3(1+0.9(0)) +0.2(0+0.9(0)) + 0=> 0.3

V\_1 = 0.9+0.15=>1.05

V\_2(S1) =

A1 = 0 + 0.2(1+0.9\*(0)) + 0.4(0+0.9(0)) +0.4(4+0.9(0)) => 0.2+1.6 => 1.8

A2 = 0.5(0+0.9\*(1.05)) + 0.3(1+0.9(0)) +0.2(0+0.9(0)) + 0=> 0.4725+ 0.3=> 0.7725

V\_2(S1) = 2.1005

V\_1(S4) = 0.5(A1) + 0.5(A2)

A1 = 0.4(0+0.9(0))+0.4(0+0.9(0))+0.2(2+0.9(0))+0 => 0.4

A2 = 0+0.5(0+0.9(0))+0.4(2+0.9(0))+0.1(0+0.9(0)) => 0.8

V\_1(S4) => 0.2+0.4 => 0.6

V\_2(S4) =

A1 = 0.4(0+0.9(0)) + 0.4(0+0.9(0)) + 0.2(2+0.9(0)) + 0 => 0.4

A2 = 0+0.5(0+0.9(0))+0.4(2+0.9(0))+0.1(0+0.9(0.6)) => 0.854

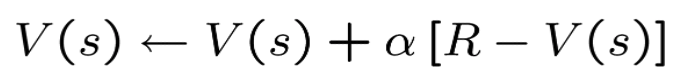
V\_2(S4) = 0.627

4) Consider a simple MDP with four states and the initial value estimates of 0 for all four states. You run two episodes in the environment and use the Monte Carlo method to update the state values after each episode. Use a learning rate (alpha) of 0.2 and **first visit** Monte Carlo updates. The transitions for the four episodes are shown below. There is only one reward for each episode given at the end of the episode.

S1 -> A1 -> S2 -> A2 -> S3 Reward: 10

S4 -> A2 -> S1 -> A2 -> S2 Reward: 20

|  |  |  |  |
| --- | --- | --- | --- |
| State | Initial Value | Episode 1 | Episode 2 |
| S1 | 0 | 2 | 5.6 |
| S2 | 0 | 2 | 5.6 |
| S3 | 0 | 2 | 2 |
| S4 | 0 | 0 | 4 |



1)

S1 = S1 + 0.2(10-0) => 2

S2 = S2 +0.2(10-0) => 2

S3 = S3 +0.2(10-0) =>2

2)

S4 = S4 + 0.2(20-0) => 4

S1 = 2 + 0.2(20-2) => 2+3.6 => 5.6

S2 = 2 + 0.2(20-2) => 2 + 3.6 => 5.6

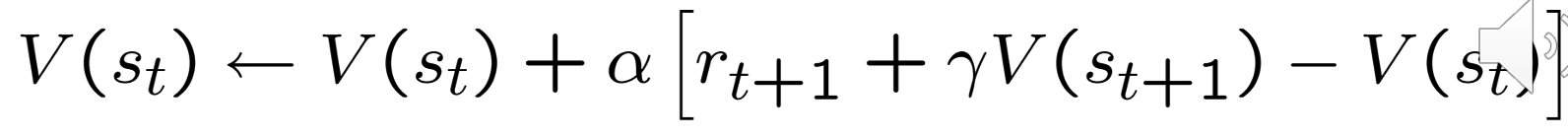
5) Consider a simple MDP with three states and the initial value estimates shown in the table below. Let the discount rate λ = 1.0 and learning rate α= 0.1. Perform the temporal differences (TD) learning updates for the transitions shown and record the updated values for all states in the columns provided in the table.

Transition 1: S3 –> A0 –> S1 with reward rt = 6

Transition 2: S1 –> A1 –> S2 with reward rt = 3

Transition 3: S2 –> A1 –> S1 with reward rt = 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Initial Value | After Transition 1 | After Transition 2 | After Transition 3 |
| S1 | 4 | 4 | 3.9 | 3.9 |
| S2 | 0 | 0 | 0 | 0.89 |
| S3 | 1 | 1.9 | 1.9 | 1.9 |



S1->S2

V(S1) = 4 + 0.1\*(3+1.0(0)-4) => 3.9

S2->S1

V(S2) = 0 + 0.1\*(5+1.0\*3.9-0) => 0.89

S3->S1

V(S3) = V(S3) + 0.1\*(6 +1.0\*4-1) => 1 + 0.1\*(6+4-1) => 1.9