CS 4376/5376

Homework 3, Part I

Solving MDPs

DUE: Wed, Oct 14 at 11:59 PM

NOTE: This part of the assignment must be completed and submitted INDIVIDUALLY by every student; only the second part is a group assignment.

The assignment should be turned in using blackboard as either a word document or pdf file. Scanned work is acceptable, as long as it is legible.

regret for the decision tree instead.

1) Suppose you have the MDP shown below, where S2 has value 10, S3 has value 4, and S4 has value 8.

A picture containing orange, small, man, holding

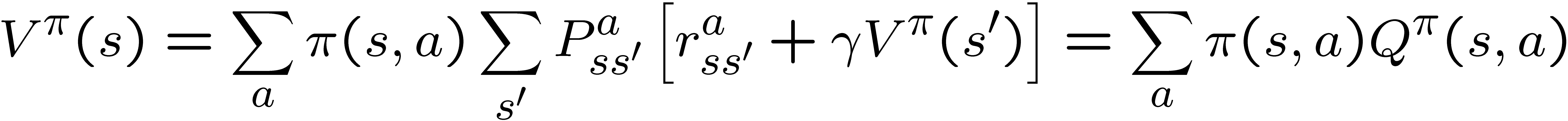
Description automatically generated

**10 4 10 8**

The policy for A1 is 0.8 A1 and 0.2 A2. The transition probabilities are given in the table below. The discount rate is 1.

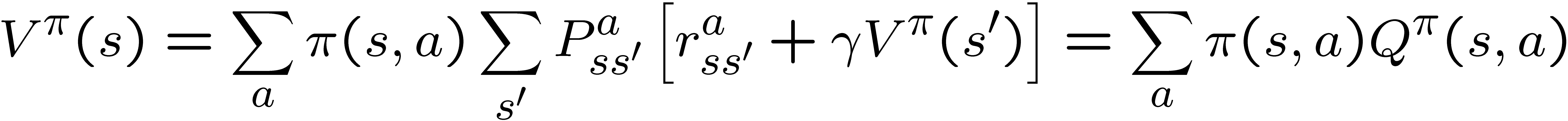
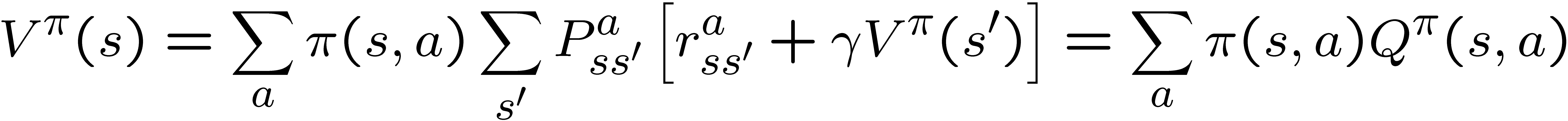
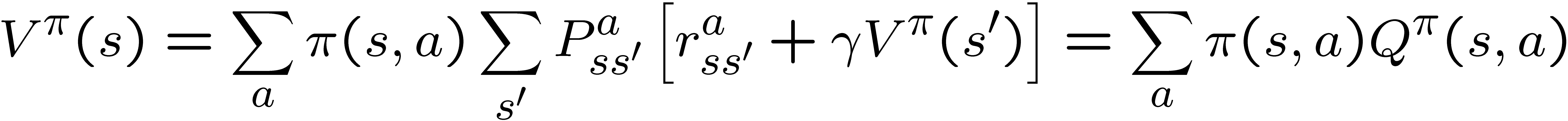
|  |  |
| --- | --- |
| Transition (s,a,s’) | P(s’|s,a) |
| S1 A1 S2 | 0.3 |
| S1 A1 S3 | 0.7 |
| S1 A2 S2 | 0.6 |
| S1 A2 S4 | 0.4 |

***SHOW YOUR CALCULATIONS***

******

***\*Transitions do not have rewards***

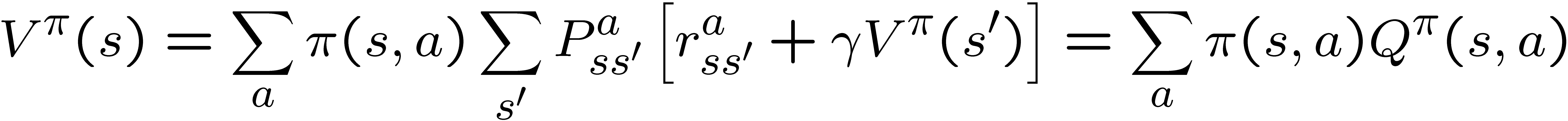
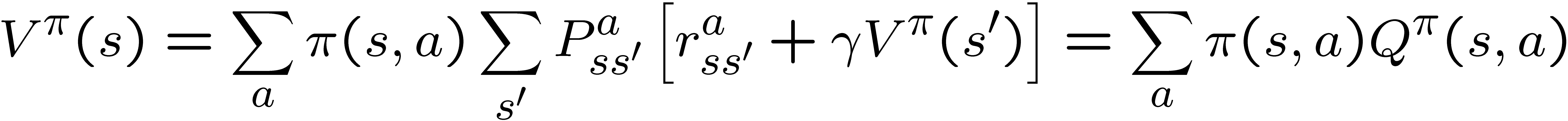
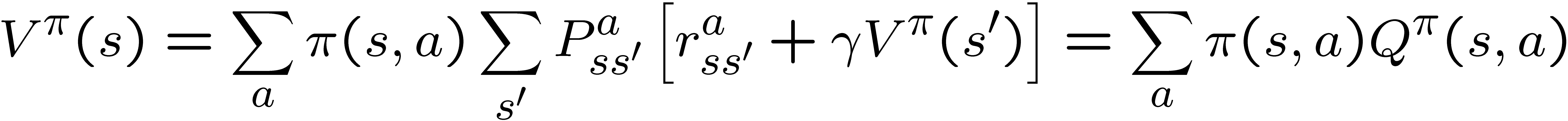
1. Compute Q(S1, A1)

******

P(S2|A1,S1) \* [r(S1,S2) + (1)(10)] = 0.3(0+(1)(10)) => 3

P(S3|A1,S1) \* [r(S1,S3) + (1)(4)] = 0.7(0+(1)(4)) => 2.8

1. Compute Q(S1, A2)



P(S2|A2,S1) \* [r(S1,S2) + (1)(10)] = 0.6(0+(1)(10)) => 6

P(S4|A2,S1) \* [r(S1,S3) + (1)(8)] = 0.4(0+(1)(8)) => 3.2

1. Compute V(S1):

S1 with probability 0.8

S2 with probability 0.2

V(S1) = 0.8(3+2.8) + 0.2(6+3.2) => 0.8\*5.8 + 0.2\*(9.2) => 4.64+1.84 => 6.48

2) Consider a simple MDP with four states and the initial value estimates of 0 for all four states. You run three episodes in the environment and use the Monte Carlo method to update the state values after each episode. Use a learning rate (alpha) of 0.1 and **first visit** Monte Carlo updates. The transitions for the three episodes are shown below. There is only one reward for each episode given at the end of the episode.

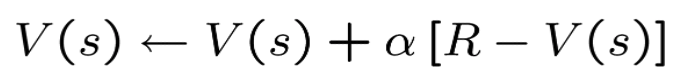
S2 -> A1 -> S4 -> A2 -> S3 Reward: 5

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

S3 -> A2 -> S2 -> A2 -> S4 Reward: 10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Initial Value | Episode 1 | Episode 2 | Episode 3 |
| S1 | 0 | 0 | 2 | 2 |
| S2 | 0 | 0.5 | 2.45 | 3.205 |
| S3 | 0 | 0.5 | 0.5 | 1.45 |
| S4 | 0 | 0.5 | 0.5 | 1.45 |

***SHOW YOUR CALCULATIONS***



1. Episode 1

V(S2) = V(S2) + 0.1\*[5-0] = 0.5

V(S4) = V(S4) + 0.1\*[5-0] = 0.5

V(S3) = V(S3) + 0.1\*[5-0] = 0.5

1. Episode 2

V(S1) = V(S1) + 0.1\*[20-0] => 0 + 2 => 2

V(S2) = V(S2) + 0.1\*[20-0.5] => 0.5 + 1.95 => 2.45

V(S1) = V(S1) + 0.1\*[20-0] => 0 + 2 => 2

1. Episode 3

V(S3) = V(S3) + 0.1\*[10-0.5] => 0.5 + 0.1\*9.5 => 1.45

V(S2) = V(S2) + 0.1\*[10-2.45] => 2.45 + 0.755=> 4.4

V(S1) = V(S1) + 0.1\*[10-0.5] => 0.5 + 0.95 => 1.45

3) 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: S2 –> A0 –> S3 with reward rt = 1

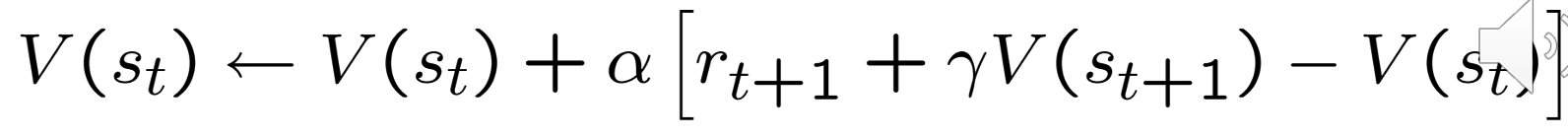
Transition 2: S3 –> A3 –> S1 with reward rt = 4

Transition 3: S1 –> A2 –> S2 with reward rt = 3

Transition 4: S2 –> A1 –> S4 with reward rt = 5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| State | Initial Value | After Transition 1 | After Transition 2 | After Transition 3 | After Transition 4 |
| S1 | 2 | 2 | 2 | 3.59 | 3.59 |
| S2 | 5 | 4.9 | 4.9 | 4.9 | 4.91 |
| S3 | 3 | 3 | 3.3 | 3.3 | 3.3 |

***SHOW YOUR CALCULATIONS***



Transition 1:

V(S2) = V(S2) + 0.1\*[1+1.0(3)-5] => 5 + 0.1\*[1+3-5] => 5-0.1 => 4.9

Transition 2:

V(S3) = V(S3) + 0.1\*[4+1.0(2)-3] => 3 + 0.1\*[6-3] => 3+0.3 => 3.3

Transition 3:

V(S1) = V(S1) + 0.1\*[3+1.0(4.9)-2] => 3 + 0.1\*[5.9] => 3+0.59 => 3.59

Transition 4:

V(S2) = V(S2) + 0.1\*[5+1.0(0)-4.9] => 4.9 + 0.1\*[5-4.9] => 4.9+0.01 => 4.91