CS471 Homework 4

Yiting Gan gna29@purdue.edu

November 3, 2021

Problem 1

iteration 0:
$V_0^*(-2) = V_0^*(-1) = V_0^*(0) = V_0^*(1) = V_0^*(2) = 0$
iteration 1:
$V_1^*(0) = max(Q(0, a_1), Q(0, a_2))$
$Q(0, a_1) = T(0, a_1, -1)[R(0, a_1, -1) + V(-1)] + T(0, a_1, 1)[R(0, a_1, 1) + V(1)]$
$= 0.8 * (-5) + 0.2 * (-5) = -5$
$Q(0, a_2) = T(0, a_2, -1)[R(0, a_2, -1) + V(-1)] + T(0, a_2, 1)[R(0, a_2, 1) + V(1)]$
$= 0.7 * (-5) + 0.3 * (-5) = -5$
$\rightarrow$V$_1^*(0) = max(-5, -5) = -5$
$V_1^*(-1) = max(Q(-1, a_1), Q(-1, a_2))$
$Q(-1, a_1) = T(-1, a_1, -2)[R(-1, a_1, -2) + V(-2)] + T(-1, a_1, 0)[R(-1, a_1, 0) + V(0)]$
$= 0.8 * 20 + 0 + 0.2 * (-5) = 15$
$Q(-1, a_2) = T(-1, a_2, -2)[R(-1, a_2, -2) + V(-2)] + T(-1, a_2, 0)[R(-1, a_2, 0) + V(0)]$
$= 0.7 * 20 + 0 + 0.3 * (-5) = 12.5$
$\rightarrow$V$_1^*(-1) = max(15, 12.5) = 15$
$V_1^*(1) = max(Q(1, a_1), Q(1, a_2))$
$Q(1, a_1) = T(1, a_1, 2)[R(1, a_1, 1) + V(2)] + T(1, a_1, 0)[R(1, a_1, 0) + V(0)]$
$= 0.2 * 100 + 0 + 0.8 * (-5) = 16$
$Q(1, a_2) = T(1, a_2, 2)[R(1, a_2, 1) + V(2)] + T(1, a_2, 0)[R(1, a_2, 0) + V(0)]$
$= 0.3 * 100 + 0 + 0.7 * (-5) = 26.5$
$\rightarrow$V$_1^*(1) = max(16, 26.5) = 26.5$
$V_1^*(-2) = V_1^*(2) = 0$
iteration 2:
$V_2^*(0) = max(Q(0, a_1), Q(0, a_2))$
$Q(0, a_1) = T(0, a_1, -1)[R(0, a_1, -1) + V(-1)] + T(0, a_1, 1)[R(0, a_1, 1) + V(1)]$
$= 0.8 * (-5 + 15) + 0.2 * (-5 - 5) = 6$
$Q(0, a_2) = T(0, a_2, -1)[R(0, a_2, -1) + V(-1)] + T(0, a_2, 1)[R(0, a_2, 1) + V(1)]$
$= 0.7 * (-5 + 15) + 0.3 * (-5 + 26.5) = 13.45$
$\rightarrow$V$_2^*(0) = max(6, 13.45) = 13.45$
$V_2^*(-1) = max(Q(-1, a_1), Q(-1, a_2))$
$Q(-1, a_1) = T(-1, a_1, -2)[R(-1, a_1, -2) + V(-2)] + T(-1, a_1, 0)[R(-1, a_1, 0) + V(0)]$
$= 0.8 * (20) + 0.2 * (-5 - 5) = 14$
$Q(-1, a_2) = T(-1, a_2, -2)[R(-1, a_2, -2) + V(-2)] + T(-1, a_2, 0)[R(-1, a_2, 0) + V(0)]$
$= 0.7 * 20 + 0.3(-5 - 5) = 11$
$\rightarrow$V$_2^*(-1) = max(11, 14) = 14$
$V_2^*(1) = max(Q(1, a_1), Q(1, a_2))$
$Q(1, a_1) = T(1, a_1, 0)[R(1, a_1, 0) + V(0)] + T(1, a_1, 2)[R(1, a_1, 2) + V(2)]$
$= 0.8 * (-5 - 5) + 0.2 * (100) = 12$

$Q(1,a_2) = T(1,a_2,0)[R(1,a_2,0)+V(0)] + T(1,a_2,2)[R(1,a_2,2)+V(2)]$
$= 0.7*(-5+15) + 0.3*(-5+26.5) = 13.45$
$\rightarrow$V$_2^*(1) = max(12, 13.45) = 13.45$
$V_2^*(-2) = V_2^*(2) = 0$

Problem 2

$pi^*(-1) = -1$
$pi^*(0) = +1$
$pi^*(1) = +1$
$pi^*(-2) = pi^*(2)$ = no action

Problem 3

Counterexample:
A$\leftarrow$ B $\rightarrow$ C, (MDP with three states)
Let B be the start state and A,C be terminal states. Assume there is an action that takes us from state B to state A with reward 0 and probability 0.9 and state C with reward 10 and probability 0.1. If we add noise, we are more likely to end at C and received reward 10. Thus, the optimal value goes up.

Problem 4

We have to iterate MDP more than once because we do not have an ordering over the state, but if we have an acyclic MDP, we could compute V(s) by dynamic programming recurrence, which will only go through all (s, a, s') once.

Problem 5

limit $T'(s,a,s') = \gamma T(s,a,s')$
limit $T'(s,a,o) = 1-\gamma$
Case 1:
limit $R'(s,a,s') = R(s,a,s')$
limit $R'(s,a,o) = \sum_{s'\in State} T(s,a,s')R(s,a,s')$
Case 2:
limit $R'(s,a,s') = \frac{1}{\gamma}R(s,a,s')$
limit $R'(s,a,o) = 0$
Case 3: limit $R'(s,a,s') = R(s,a,s')$
limit $R'(s,a,o) = 0$
$\rightarrow V^{*\prime}(s) = \max_{a\in Action(s)} \sum_{s'\in State} T'(s,a,s')[R'(s,a,s') + V^{*\prime}(s')]$
$= \max_{a\in Action(s)\backslash o} \sum_{s'\in State} T'(s,a,s')[R'(s,a,s') + V^{*\prime}(s')] + T'(s,a,o)[R'(s,a,o) + V^{*\prime}(o)]$
Then, we plug in the above three cases.
Case 1:
$V^{*\prime}(s) = \max_{a\in Action(s)} \sum_{s'\in State} \gamma T(s,a,s')[R(s,a,s') + V^{*\prime}(s')] + (1-\gamma)\sum_{s'\in State}(s,a,s')R(s,a,s')$
$= \max_{a\in Action(s)} \sum_{s'\in State} T(s,a,s')[R(s,a,s') + \gamma V^{*\prime}(s')]$
$\rightarrow$ same as the old MDP.
Case 2:
$V^{*\prime}(s) = \max_{a\in Action(s)} \sum_{s'\in State} \gamma T(s,a,s')[\frac{1}{\gamma}R(s,a,s') + V^{*\prime}(s')] + (1-\gamma)*0$
$=\max_{a\in Action(s)} \sum_{s'\in State} T(s,a,s')[R(s,a,s') + \gamma V^{*\prime}(s')]$

$\rightarrow$ same as the old MDP.
Case 3:
$V^{*\prime}(s) = \max_{a \in Action(s)} \sum_{s' \in State} \gamma T(s,a,s')[R(s,a,s') + V^{*\prime}(s')] + (1-\gamma) * 0$
$= \max_{a \in Action(s)} \sum_{s' \in State} T(s,a,s')[\gamma R(s,a,s') + \gamma V^{*\prime}(s')]$
Thus, except the third case, the new MDP has the same optimal values as the old MDP.