COMS W4701: Artificial Intelligence, Spring 2025
Homework #3

Peter Driscoll (pvd2112)

March 11, 2025

Problem 1

1.a)

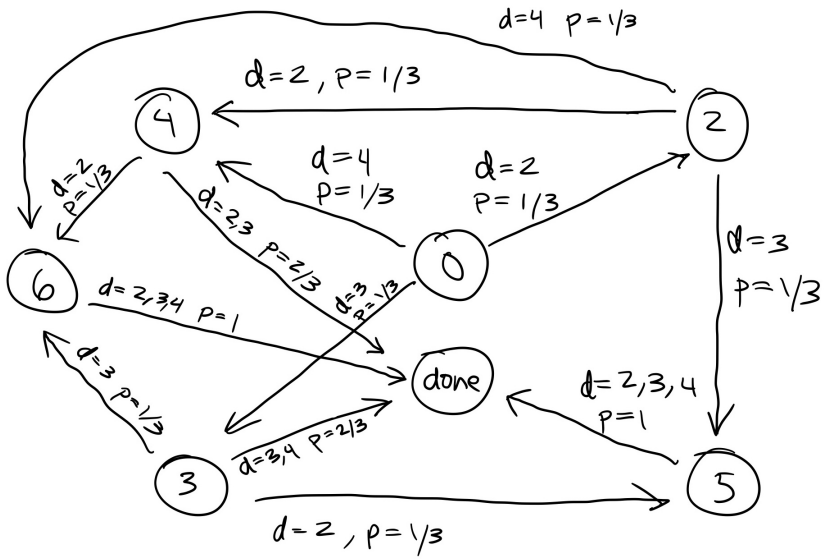

Figure 1: Mini-Blackjack State Transition Diagram

1.b)

The optimal action for states 5 and 6 is the *stop* action. The stop action at these states yields $V^*(\text{stop}) = 5$ and $V^*(\text{stop}) = 6$, respectively. This is because the draw action from these states will always result in a sum greater than 6, which yields $V^*(\text{done}) = 0$.

Table 1: Optimal Actions and Values for Remaining States

State	**Optimal Value Calculation**
4	$V(4) = \max\left\{\frac{1}{3} \cdot V(6), 4\right\} \Rightarrow \max\{2, 4\} \Rightarrow 4$
3	$V(3) = \max\left\{\frac{1}{3} \cdot (V(5) + V(6)), 3\right\} \Rightarrow \max\left\{\frac{11}{3}, 3\right\} \Rightarrow \frac{11}{3}$
2	$V(2) = \max\left\{\frac{1}{3} \cdot (V(4) + V(5) + V(6)), 2\right\} \Rightarrow \max\{5, 2\} \Rightarrow 5$
0	$V(0) = \max\left\{\frac{1}{3}V(2) + \frac{1}{3}V(3) + \frac{1}{3}V(4), 0\right\} \Rightarrow \frac{38}{9}$

Table 2: Summary Optimal Actions and Values

State	$\pi^*_{t=1}$(state)	V^*(state)
0	draw	$\frac{32}{9}$
2	draw	5
3	draw	$\frac{11}{3}$
4	stop	4
5	stop	5
6	stop	6

Dynamic programming is not required for this problem because there is a chain of dependencies amongst the different states that obviates the need for iterative value updates. Specifically, the states of 5 and 6 have trivial optimal values equal to their starting states, while state 4 only depends on comparing its starting value to a reduced value of state 6, and so on and so forth for states 0, 2, and 3.

1.c)

$$\begin{aligned} V^*(2) \Rightarrow & \text{ Need to find } \gamma \text{ s.t.} \\ & \Rightarrow \gamma \cdot \frac{11}{3} < 3 \\ & \Rightarrow \gamma \leq \frac{9}{11} \end{aligned}$$

$$\begin{aligned} V^*(3) \Rightarrow & \text{ Need to find } \gamma \text{ s.t.} \\ & \Rightarrow \gamma \cdot 5 < 3 \\ & \Rightarrow \gamma \leq \frac{2}{5} \end{aligned}$$

Thus,

$$\gamma = \min\left\{\frac{2}{5}, \frac{9}{11}\right\} = \frac{2}{5}$$

Substituting $\gamma = \frac{2}{5}$ into the computation of $V^*(3, 2, 0)$ results in the following:

Table 3: γ-min Optimal Actions and Values for Remaining States

State	**Optimal Value Calculation**
3	$V_1(3) = \max\left\{\frac{2}{5} \cdot \frac{5+6}{3}, 3\right\} \Rightarrow \max\{22/15, 3\} \Rightarrow 3$
2	$V_1(2) = \max\left\{\frac{2}{5} \cdot \frac{4+5+6}{3}, 2\right\} \Rightarrow \max\{2, 2\} \Rightarrow 2$
0	$V_1(0) = \max\left\{\frac{2}{5} \cdot (V(2) + V(3) + V(4)), 0\right\} \Rightarrow \frac{6}{5}$

A lower discount factor γ reduces the values of states 0, 2 and 3 because these states have lower starting values and depend on the draw action, to reach their optimal terminal value, making them more sensitive to a decrease in γ. In contrast, states 4, 5, and 6 have such high starting values that their optimal values are secured by choosing the *done* action at the outset, making them fully insensitive to downward changes in γ.

Problem 2

2.a)

The π^* and V^* for the updated mini-blackjack game will now bias exclusively toward drawing cards until reaching the state of 6 and then stopping, thus the π^* will return the draw actions for states less than 6, and the V^* will be 6 for all states.

2.b)

$$V^{\pi}(s) = \sum_{s'} T\big(s, \pi(s), s'\big) \Big[R\big(s, \pi(s), s'\big) + \gamma\, V^{\pi}(s') \Big].$$

Performing value iteration will uncover π^* and V^*. Starting with V^*:

Table 4: Value Iteration for V^* and π^*

State	**Value Iteration Calculations**
3	$V_1(3) = \max\left\{(\frac{2}{5})^1 \cdot \frac{5+6}{3}, 3\right\} \Rightarrow \max\{22/15, 3\} \Rightarrow 3$

Problem 2

2.c)

Iteration 1 (V_1)

$$V_1(0) = \max\Big\{0,\ \tfrac{1}{3}\big(0+0\big) + \tfrac{1}{3}\big(0+0\big) + \tfrac{1}{3}\big(0+0\big)\Big\} = 0$$

$$V_1(2) = \max\Big\{2,\ \tfrac{1}{3}\big(0+0\big) + \tfrac{1}{3}\big(0+0\big) + \tfrac{1}{3}\big(0+0\big)\Big\} = 2$$

$$V_1(3) = \max\Big\{3,\ \tfrac{1}{3}(0)\Big\} \ = \ 3$$

$$V_1(4) = \max\Big\{4,\ \tfrac{1}{3}(0)\Big\} \ = \ 4$$

$$V_1(5) \ = \ 5$$

$$V_1(6) \ = \ 6$$

Iteration 2 (V_2)

$$V_2(0) = \max\Big\{0,\ \tfrac{1}{3}\big(0 + 0.9 \times 2\big) + \tfrac{1}{3}\big(0 + 0.9 \times 6\big) + \tfrac{1}{3}\big(0 + 0.9 \times 3\big)\Big\} \ = \ \frac{9}{2} \ = \ 4.5$$

$$V_2(2) = \max\Big\{2,\ \tfrac{1}{3}\big(0 + 0.9 \times 5\big) + \tfrac{1}{3}\big(0 + 0.9 \times 6\big) + \tfrac{1}{3}\big(0 + 0.9 \times V_1(\mathit{done})\big)\Big\} \ = \ \tfrac{9}{2}$$

$$V_2(3) = \max\Big\{3,\ \tfrac{1}{3}\big(0 + 0.9 \times 5\big) + \tfrac{1}{3}\big(0 + 0.9 \times 6\big) + \tfrac{1}{3}\big(0 + 0.9 \times V_1(\mathit{done})\big)\Big\} \ = \ \frac{33}{10} \ = \ 3.3$$

$$V_2(4) = \max\Big\{4,\ \tfrac{1}{3}\big(0 + 0.9 \times 6\big) + \tfrac{2}{3}\big(0 + 0.9 \times V_1(\mathit{done})\big)\Big\} \ = \ \frac{33}{10} \ = \ 3.3$$

$$V_2(5) = \max\Big\{5,\ \tfrac{3}{3}\big(0 + 0.9 \times V_1(\mathit{done})\big)\Big\} \ = \ 5$$

$$V_2(6) = \max\Big\{6,\ \tfrac{3}{3}\big(0 + 0.9 \times V_1(\mathit{done})\big)\Big\} \ = \ 6$$

2.d)

The convergence of value iteration is determined by the value of γ. Each update of V(s) shrinks the gap between $V^*(s)$ and $V(s)$ by a factor of γ. Thus, reducing γ to 0.5 will exponentially reduce the amount of iterations necessary to converge under the threshold. I would expect the $\pi(\gamma = 0.9)$ to incur more draw actions because optimal value will be weighted more heavily toward future values, which would retain more of their value in successive future iterations due to the higher γ.

Problem 3

3.a)

Episode 1:

$$V(0) = V(2) = V(4) = 0$$

Episode 2:

$$V(0) = V(3) = 0$$

Episode 3:

$$V(2) = \gamma^2(5) = 5 \quad \text{and} \quad V(5) = \gamma^2(5) = 5.$$
$$V(2) = \frac{0+5}{2} = 2.5$$

Episode 4:

$$V(3) = \gamma^2(5) = 5 \quad \text{and} \quad V(5) = \gamma^2(5) = 5.$$
$$V(3) = \frac{0+5}{2} = 2.5$$

Episode 5:

$$V(4) = 6 \quad \text{and} \quad V(6) = 6.$$
$$V(4) = \frac{0+6}{2} = 3$$

The order in which the episodes are observed does not affect the estimated state values. This is because the value of each state is estimated independently, only using the states and rewards witnessed within the episode. The estimated state values of each episode are incremently averaged togther, which also is not affected by the order, being a commutative operation.

3.b)

TD(0) Updates for Episodes 3, 4, and 5 In this mini-blackjack scenario, we use temporal-difference learning with $\alpha = 0.5$ and $\gamma = 1$. After the first two episodes (both yielding reward 0), all state values remain 0:

$$V(0) = V(2) = V(3) = V(4) = V(5) = V(6) = 0.$$

*Episode 3: $0 \to 2 \to 5 \to$ done, Reward $= 5$ For each transition $(s \to s', r)$, we apply:

$$V(s) \leftarrow V(s) + \alpha\big[r + V(s') - V(s)\big].$$

1. $0 \to 2$, $r = 0$:
$$V(0) = 0 + 0.5\big[0 + V(2) - 0\big] = 0.$$

2. $2 \to 5$, $r = 0$:
$$V(2) = 0 + 0.5\big[0 + V(5) - 0\big] = 0.$$

3. $5 \to \text{done}$, $r = 5$:

$$V(5) = 0 + 0.5\big[5 + V(\text{done}) - 0\big] = 0.5 \times 5 = 2.5.$$

Thus, after Episode 3: $V(5) = 2.5$, and all other states remain 0.

*Episode 4: $0 \to 3 \to 5 \to \text{done}$, Reward $= 5$

1. $0 \to 3$, $r = 0$:

$$V(0) = 0 + 0.5\big[0 + V(3) - 0\big] = 0.$$

2. $3 \to 5$, $r = 0$:

$$V(3) = 0 + 0.5\big[0 + V(5) - 0\big] = 0.5 \times 2.5 = 1.25.$$

3. $5 \to \text{done}$, $r = 5$:

$$V(5) = 2.5 + 0.5\big[5 + V(\text{done}) - 2.5\big] = 2.5 + 0.5 \times 2.5 = 3.75.$$

Hence, after Episode 4: $V(3) = 1.25$, $V(5) = 3.75$, and the rest remain 0.

*Episode 5: $0 \to 4 \to 6 \to \text{done}$, Reward $= 6$

1. $0 \to 4$, $r = 0$:

$$V(0) = 0 + 0.5\big[0 + V(4) - 0\big] = 0.$$

2. $4 \to 6$, $r = 0$:

$$V(4) = 0 + 0.5\big[0 + V(6) - 0\big] = 0.$$

3. $6 \to \text{done}$, $r = 6$:

$$V(6) = 0 + 0.5\big[6 + V(\text{done}) - 0\big] = 3.$$

So, after Episode 5: $V(6) = 3$, $V(5) = 3.75$, $V(3) = 1.25$, and $V(4) = V(2) = V(0) = 0$.

Conclusion

These updates match the TD(0) learning rule with $\alpha = 0.5$ and $\gamma = 1$ for the given episodes. Notably, $V(5)$ increases from 0 to 2.5, then to 3.75, $V(3)$ becomes 1.25, and $V(6)$ becomes 3 by the end of Episode 5.