This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

Initialization (Iteration 0)

- $V(-2) = 0$ (Terminal State)
- $V(-1) = 0$
- $V(0) = 0$
- $V(1) = 0$
- $V(2) = 0$ (Terminal State)

Iteration 1

Using the Bellman equation for value iteration, $V(s)$ values are calculated as:

For state -1:

$$V(-1) = \max\left(0.2 \times (-5+0) + 0.8 \times (20+0), 0.3 \times (-5+0) + 0.7 \times (20+0)\right)$$
$$V(-1) = \max(16, 13.5)$$
$$V(-1) = 16$$

For state 0:

$$V(0) = \max\left(0.2 \times (-5+0) + 0.8 \times (-5+0), 0.3 \times (-5+0) + 0.7 \times (-5+0)\right)$$
$$V(0) = \max(-5, -5)$$
$$V(0) = -5$$

For state 1:

$$V(1) = \max\left(0.2 \times (100+0) + 0.8 \times (-5+0), 0.3 \times (100+0) + 0.7 \times (-5+0)\right)$$
$$V(1) = \max(16, 26.5)$$
$$V(1) = 26.5$$

Iteration 2

For state -1:

$$V(-1) = \max\left(0.2 \times (-5+-5) + 0.8 \times (20+0), 0.3 \times (-5+-5) + 0.7 \times (20+0)\right)$$
$$V(-1) = \max(14, 11)$$
$$V(-1) = 14$$

For state 0:

$$V(0) = \max\left(0.2 \times (-5+26.5) + 0.8 \times (-5+16), 0.3 \times (-5+26.5) + 0.7 \times (-5+16)\right)$$
$$V(0) = \max(13.1, 14.15)$$

$$V(0) = 14.15$$

For state 1:

$$V(1) = \max\left(0.2 \times (100 + 0) + 0.8 \times (-5 + -5), 0.3 \times (100 + 0) + 0.7 \times (-5 + -5)\right)$$

$$V(1) = \max(12, 23)$$

$$V(1) = 23$$

Summary

After Iteration 0:

- $V(-2) = 0$
- $V(-1) = 0$
- $V(0) = 0$
- $V(1) = 0$
- $V(2) = 0$

After Iteration 1:

- $V(-2) = 0$
- $V(-1) = 16$
- $V(0) = -5$
- $V(1) = 26.5$
- $V(2) = 0$

After Iteration 2:

- $V(-2) = 0$
- $V(-1) = 14$
- $V(0) = 14.15$
- $V(1) = 23$
- $V(2) = 0$

1.b

- $S(-1)$: the best policy is take $A(-1)$, which will have $V_{\text{opt}}(-1) = 14$
- $S(0)$: the best policy is take $A(1)$, which will have $V_{\text{opt}}(0) = 14.15$
- $S(1)$: the best policy is take $A(1)$, which will have $V_{\text{opt}}(0) = 23$

2.a

Extend the state space by adding an artificial terminal state S(term)

Redifine the transition actions

- for the artificial state S(term), define its transition probabilities to be $1 - \lambda$
- for the original states, update its transition probabilities $T'(s, a, s') = \lambda \times T(s, a, s')$

Redifine the rewards

- for the artificial state S(term), define its rewards 0
- for the original states, keep its rewards as original rewards

4.b

Comparing Q-learning and Value Iteration for smallMDP:

- With state (1, 1, (1, 2)): Differing actions between VI (Take) and QL (Quit)
- With state (5, 1, (2, 1)): Differing actions between VI (Take) and QL (Quit)
- With state (6, 0, (1, 1)): Differing actions between VI (Take) and QL (Quit)

Differing actions between VI and QL: 3

Comparing Q-learning and Value Iteration for largeMDP:

Differing actions between VI and QL: 880

4.d

Comparing Q-learning and Value Iteration for newThresholdMDP: ValueIteration: 5 iterations The expected reward from simulating the original policy on the newThresholdMDP is: 6.868 The expected reward under the new Q-learning policy is: 12.0

5.a

5.b

5.c

5.d