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

In the smallMDP, Q-learning almost matched the policy from value iteration, with only 3 differences in actions across states. This suggests Q-learning performed well on a simpler MDP after 30,000 trials.

For the largeMDP, there were 880 different actions between the Q-learning and value iteration policies, indicating Q-learning faced challenges in learning an optimal policy for a more complex MDP within the same number of trials. Possible reasons include the complexity of the state space, suboptimal feature extraction, and insufficient exploration.

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

The higher expected reward from Q-learning (12.0) compared to using the original policy on the new MDP (6.868) indicates that Q-learning adapted effectively to the new threshold in 'newThresholdMDP'. The original policy was optimal for a different set of rules and didn't perform as well when those rules changed. Q-learning's ability to explore and learn from the new environment allowed it to develop a strategy better suited to the new conditions, hence the improved reward.

5.a

For the 40-year horizon:

- For the 40-year horizon MDP, the economically preferable action is to wait and only invest sparingly, as the initial infrastructure level of 12 provides a sufficient buffer against the projected sea level rise within this period.
- With an initial infrastructure level of 12, and considering the sea level rise projections of 1 cm, 2 cm, or 3 cm at each 10-year interval, it is improbable that the sea level will surpass the infrastructure height within 40 years. This contributes to a policy favoring fewer immediate investments, as reflected by the 0.0032 ratio of "invest" to "wait" states in the MDP results. The economic plan thus suggests conserving the budget, potentially allocating funds to other priorities until such investments become essential.

For the 100-year horizon:

- For the 100-year horizon MDP, the economic policy shifts towards proactive and frequent investments in infrastructure, due to the increased likelihood of sea level rise exceeding the initial infrastructure level over the longer term.
- Over the course of a century, the likelihood of the sea level rising beyond the initial infrastructure level of 12 increases significantly. This risk necessitates a more proactive investment strategy to build infrastructure resilience against potential flooding. The MDP results, with a ratio of "invest" to "wait" states at 0.4387, recommend a shift toward frequent and early investments to ensure infrastructure can handle or exceed the projected increases in sea level over the long term.

In conclusion, the initial infrastructure level of 12 acts as a buffer against short-term sea level rise, influencing the optimal policy towards reduced immediate infrastructure investment for a 40-year time frame. However, for the 100-year scenario, this buffer is inadequate against the projected long-term risks, necessitating a more aggressive investment strategy.

5.b

Using the principle of the Symmetry of Future Generations, I advocate for the 100-year horizon MDP's approach, which suggests a proactive and consistent investment in infrastructure. This policy not only aligns with the optimal economic strategy identified by the MDP but also ensures that we do not arbitrarily favor the current generation's economic conditions over the future generations' right to a secure and prosperous life. It respects the equal moral importance of all generations by investing in mitigation efforts that will benefit both present and future residents of Los Angeles.

5.c

The discounted MDP suggests a more balanced approach to infrastructure investment with a ratio of 0.3844 of invest to wait states, indicating a significant but not overwhelming preference for investment. This approach recognizes the importance of immediate action due to the threat of hurricanes, but it also takes into account the reduced value of future rewards, making some degree of budget conservation reasonable.

In contrast, the non-discounted MDP, with a ratio of 1.0712 of invest to wait states, recommends a more aggressive investment in infrastructure. This model treats the value of future rewards as equal to current rewards, leading to a policy that prioritizes long-term safety and preparedness over immediate economic concerns.

The different strategies are recommended because the discounted MDP accounts for the decreasing value of future benefits, leading to a more conservative approach, whereas the non-discounted MDP operates on the principle that future benefits are just as valuable as immediate ones, encouraging more upfront investment in infrastructure to mitigate the long-term risks of sea-level rise and hurricanes.

5.d

Given the substantial difference in expected reward when adjusting the MDP to account for a more realistic flooding cost, the -$10 million flooding cost MDP is inadequate for infrastructure economic decision-making. The grader output indicates that underestimating the cost of flooding leads to an expected reward on the modified MDP that is significantly negative, suggesting severe under-preparation for potential disasters.

Therefore, the city council should not rely on the model calibrated with the underestimated cost. Instead, a recalibration of the MDP with the flooding cost of -$10 billion is advisable. This will ensure that the model accurately reflects the high stakes of climate change impacts and guides investments in a way that safeguards the city's future.