Problem Set 3

Important:

- Write your name as well as your NU ID on your assignment. Please number your problems.
- Submit both results and your code.
- Give complete answers. Do not just give the final answer; instead show steps you went through to get there and explain what you are doing. Do not leave out critical intermediate steps.
- This assignment must be submitted electronically through Gradescope by February 30th 2025 (Sunday) by 11:59 PM.

1 Value Iteration

In this problem, you will perform the value iteration updates manually on a very basic game just to solidify your intuitions about solving MDPs. The set of possible states in this game is $\{-2, -1, 0, 1, 2\}$. You start at state 0, and if you reach either -2 or 2, the game ends. At each state, you can take one of two actions: $\{a_1, a_2\}$.

If you're in state s and choose the action a_1:

- You have an 80% chance of reaching the state $s - 1$.
- You have a 20% chance of reaching the state $s + 1$.

If you're in state s and choose the action a_2:

- You have an 70% chance of reaching the state $s - 1$.
- You have a 30% chance of reaching the state $s + 1$.

If your action results in transitioning to state -2, then you receive a reward of 20. If your action results in transitioning to state 2, then your reward is 100. Otherwise, your reward is -5. Assume the discount factor $\gamma = 1$.

1. We denote $V^i_{opt}(s)$ the value for state s after i iterations of value iteration. Iteration 0 just initializes all the values of V to 0, that is $V^0_{opt}(s) = 0$ for all states $s \in \{-2, -1, 0, 1, 2\}$. Terminal states do not have any optimal policies and take on a value of 0, i.e., $V^i_{opt}(2) = V^i_{opt}(-2) = 0$ for all i. Give the value of $V_{opt}(s)$ for each state s after $0, 1$, and 2 iterations of value iteration.

Recall that the value iteration update rule is given by,

$$V^{i+1}_{opt}(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P(s, a, s') \big[R(s, a, s') + \gamma V^i_{opt}(s') \big].$$

2. Based on $V^1_{opt}(s)$ after the first iteration, what is the corresponding optimal policy π_{opt} for all non-terminal states?

2 Transforming MDPs

Given one basic algorithm for computing optimal value functions in MDPs, we will develop some optimization methods to other MDPs with special structures.

1. Consider an MDP with states S and a discount factor $\gamma < 1$, denoted by M. Assume you have access to an MDP solver that can only solve MDPs with discount factor of 1. How can you utilize the given MDP solver to solve the MDP with a discount factor $\gamma < 1$?

We guide you in solving this problem. We need to define a new MDP M' such that, if we were to apply the provided solver, we would get the solutions of the MDP M.

- We use the same set actions in M' as in M.
- For the MDP M', the set of states S' is given by $S' = S \cup \tilde{s}$ where $\tilde{s}$ is a new state.
- We use a discount factor $\gamma = 1$ in M'.

You will need to define the new transition probabilities $P'(s, a, s')$ and rewards $R'(s, a, s')$ in terms of $P(s, a, s')$ and $R(s, a, s')$, respectively, such that the optimal values of $V^i_{opt}(s)$ are equal under both the MDPs M and M'.

Hints:

- Recall that,
$$V^{i+1}_{opt}(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P(s, a, s') \big[R(s, a, s') + \gamma V^i_{opt}(s') \big].$$
Your goal is to modify the above expression to obtain,
$$V^{i+1}_{opt}(s) = \max_{a \in \{a_1, a_2\}} \sum_{s'} P'(s, a, s') \big[R'(s, a, s') + V^i_{opt}(s') \big],$$
so that the MDP is transformed to a problem with a discount factor of 1. This will help you find the expressions of $P'(s, a, s')$ and $R'(s, a, s')$.
- The sum of the transition probabilities $P'(s, a, s')$ you currently found is not 1. How can you fix that? This is where the new state $\tilde{s}$ in S' will come in handy.

2. In real-world problems, the data gathered can have some noise. We consider an MDP M with a reward function $R(s, a, s')$, set of states S, and transition probabilities $P(s, a, s')$. We define a new MDP M' with the same reward function $R(s, a, s')$ and set of states S. We include some noise in the transition probabilities $P(s, a, s')$ yielding some new transition probabilities $\tilde{P}(s, a, s')$. Let $V^i_{opt}(s)$ and $\tilde{V}^i_{opt} i(s)$ be the optimal values for state s after i iterations of value iteration for the MDPs M and M', respectively. Is it always the case that $V^i_{opt}(s) \leq \tilde{V}^i_{opt}(s)$? If so, prove it and put return None for each of the code blocks in the class *CounterexampleMDP* in the file *submission.py*. Otherwise, construct a counterexample by filling out the methods in *CounterexampleMDP* in *submission.py*.