

# HW 3

EECE 571N – Sequential Decision-Making (EECE 571N)

**Instructor:** Cyrus Neary

**Due:** 2025-11-10 at 23:59 PT

**Name:** Dominic Klukas

**Student Number:** 64348378

## Instructions

Submit a single PDF to Canvas. Please include your name and student number at the top of that PDF. For all questions, show your work and clearly justify your steps and thinking. Please feel free to include any code as an attachment at the end of the PDF. State any assumptions. Unless otherwise specified, you may collaborate conceptually but must write up your own solutions independently.

## Grading

Points for each part are indicated. The total number of achievable points is 100. Partial credit is available for incorrect answers with clear reasoning.

## Starter Code

This assignment asks you to implement certain ideas and algorithms in Python. Please use the starter code provided at:

```
https:  
//github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3
```

Fill in the missing TODO comments to complete the assignment. Insert your figures as images to the relevant questions of your submission PDF document, and snippets of the relevant portions of your code.

## Questions

1. [5] **Distinguishing on-policy and off-policy RL algorithms.** In your own words, explain the difference between on-policy and off-policy learning. Give one example of an algorithm from each category, and describe briefly how the data they use to update their value estimates differs.

**Solution.** On policy learning both involve improving a policy's decisions with respect to the actions it takes at a given state from data about the predicted gains in value that a policy will have if it changes its choices at a given state, but on-policy learning makes these predictions based off of data collected by running the policy that is being improved in the real world, whereas off-policy learning makes the predictions based off of data collected from a different policy, evaluating its decisions instead to determine if the policy that is learning should change its actions.

For the two examples, the traditional Monte-Carlo general policy improvement algorithm is on policy, and the Monte-Carlo policy improvement algorithm that uses importance sampling

is an example of an off-policy algorithm. I will go through my understanding of how they work below:

Traditional monte-carlo method: for each state  $s \in S$  and action  $a \in A(s)$ , runs rollouts directly on the policy, and uses them to approximate  $Q(s, a)$ . Then, the policy updates with  $\pi_{\text{next}}(s) =_a Q(s, a)$ , and then runs rollouts again to improve this next policy.

The off-policy version of the Monte Carlo uses importance sampling. It modifies the estimation of the expectation of trajectories to match the probability that the learning policy, rather than the policy that the trajectory was actually sampled from. Essentially, the rollouts estimate the quantity  $Q_\pi(s, a) = \sum_\tau G_t \Pr(\tau | (s_0, a_0) = (s, a), \pi) = (s, a), \pi$ . We can compute the learning expectation, then, with

$$\begin{aligned} Q_{\pi_{\text{learn}}}(s, a) &= \sum_\tau G \cdot \Pr(\tau | (s_0, a_0) = (s, a), \pi_{\text{learn}}) \\ &= \sum_\tau G \cdot \frac{\Pr(\tau | (s_0, a_0) = (s, a), \pi_{\text{learn}})}{\Pr(\tau | (s_0, a_0) = (s, a), \pi_{\text{sampled}})} \Pr(\tau | (s_0, a_0) = (s, a), \pi_{\text{sampled}}). \end{aligned}$$

However, the expressions  $\Pr(\dots)$  are products of the transition probabilities and the probabilities that the policies take an action at each state, so when we take the ratio of two for the same trajectory but with different policies, we just can can compute this ratio without knowledge of the dynamics of the function, and we call it  $\rho_\tau = \prod_{t=1}^{T_{\text{final}}} \frac{\pi_{\text{learning}}(a_{\tau_t} | s_{\tau_t})}{\pi_{\text{sampled}}(a_{\tau_t} | s_{\tau_t})}$ . With sloppy notation I have denoted  $a_{\tau_t}$  and  $s_{\tau_t}$  to be the action and state respectively taken in trajectory  $\tau$  at time  $t$ . In practice, this true expectation is sampled by doing lots of rollouts and taking the average:  $Q(s, a) \approx \frac{1}{N} \sum_{n=1}^N G_n$  for  $N$  trajectories. So then, to compute the importance sampled-estimate, we get:  $Q(s, a) \approx \frac{1}{N} \sum_{n=1}^N \rho_{\tau_n} G_n$

2. [5] **Temporal difference learning.** In your own words, what makes Temporal-Difference (TD) learn- ing different from Monte Carlo methods? Why is TD learning considered important in reinforcement learning?

**Solution.** TD learning is different from Monte Carlo methods because, instead of rolling out a trajectory or multiple trajectories in order to get an estimate for  $V(s)$ , or  $Q(s, a)$  at a state  $s \in S$ , it rolls out the result of one action/state pair to get the reward  $r$  and future state  $s'$ , and then combines this with the current value estimate  $V(s')$ , rather than continuing to roll out  $\sum_{t=1}^{t_{\text{final}}} \gamma^t r_t$  for a whole trajectory.

It is considered important in reinforcement learning because it combines the benefits of DP in having backpropagation without the requirement of knowing a transition function, which is the benefit of the Monte-Carlo method.

3. [10] **Creating a gridworld environment for RL training.** Implement an RL environment class representing a gridworld that follows the standard Gymnasium API for RL environments: <https://gymnasium.farama.org/index.html>. In Homework 1, you already implemented an MDP class modeling a gridworld environment. To help you get started with this question I have included a complete implementation of this GridworldMDP class in the starter code repository. To answer this question, please simply create a wrapper around this GridworldMDP class, by completing the starter code available here: [https://github.com/cyrusneary/UBC\\_EECE571N\\_fall\\_2025/blob/main/Homework-starter-code/HW3/gridworld\\_gym\\_env.py](https://github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3/gridworld_gym_env.py). Include a snippet of the code implement- ing this Gymnasium environment wrapper in your submission file.

The initialization function of your GridworldEnv class should take the height and width of the grid as inputs, as well as the initial state, goal location, sink location(s), wall positions, slip probability, discount factor  $\gamma$ , and reward function parameters. Figure 1 illustrates an example gridworld environment that we will be using in several of the following questions.

4. [20] **Monte Carlo estimation of  $V^\pi(s)$ .** Using the gridworld environment from the previous question, implement the "First-Visit Monte Carlo Prediction" algorithm that we saw in class. Use the starter code file available at : [https://github.com/cyrusneary/UBC\\_EECE571N\\_fall\\_2025/blob/main/Homework-starter-code/HW3/monte\\_carlo\\_code.py](https://github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3/monte_carlo_code.py).

To test your implementation, use it to predict the value function corresponding to the optimal policy for the gridworld illustrated in Figure 1. Run the algorithm for  $\text{num\_episodes} = 10000$  episodes. The gridworld parameters (e.g.,  $\gamma = 0.9$ , slip probability  $p = 0.05$ ), as well as the optimal policy for the gridworld, are already available at the bottom of the provided starter code file.

- (a) Include a plot of the RMSE of the estimated value function  $\hat{V}_n(s)$ , as a function of the number episodes  $n$  sampled in the environment. Here, the RMSE should be defined as
- $$\text{RMSE}(n) = \sqrt{\sum_{s \in S} (\hat{V}_n(s) - V^*(s))^2},$$
- where  $V^*(s)$  is the value of the optimal policy in state  $s$ .
- (b) Use the `mdp.plot_values()` function (provided in the GridworldMDP starter code) to visualize the error  $\text{err}(s) = |\hat{V}_n(s) - V^*(s)|$  of the estimated value function after  $n = 10000$  episodes separately for each state in the gridworld. Also use it to visualize the number of visits  $N_n(s)$  to each state  $s$  after  $n = 10000$  episodes.
  - (c) In your own words, describe what you notice about the plots from part b). In which states is  $\text{err}(s)$  the highest? What do you notice about the values of  $N_n(s)$  in the corresponding states?
  - (d) Include snippets of your Python code implementing the Monte Carlo prediction algorithm, and the generation/plotting of results.
5. [20] **SARSA, On-Policy Control.** Using the same gridworld environment as the previous questions, implement the "SARSA: On-Policy Control" algorithm that we saw in class. Use the starter code file available at: [https://github.com/cyrusneary/UBC\\_EECE571N\\_fall\\_2025/blob/main/Homework-starter-code/HW3/SARSA\\_code.py](https://github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3/SARSA_code.py).

To test your code file, use it to learn an optimal policy and state-action value function in the gridworld environment from Figure 1. Once again, run the algorithm for  $\text{num\_episodes} = 10000$  episodes. Try each of the following learning rates:  $\alpha = 0.01$ ,  $\alpha = 0.05$ ,  $\alpha = 0.15$ .

- (a) Compare the plots of RMSE vs. number of episodes for each of the tested values of  $\alpha$ . More specifically, let  $\hat{Q}_n^\alpha(s, a)$  be the estimated value function after  $n$  episodes when the algorithm is using learning rate  $\alpha$ . Define  $\hat{V}_n^\alpha(s) = \arg \max_{a \in A} \hat{Q}_n^\alpha(s, a)$ , and

$$\text{RMSE}(n, \alpha) = \sqrt{\sum_{s \in S} (\hat{V}_n^\alpha(s) - V^*(s))^2}.$$

Plot  $\text{RMSE}(n, \alpha)$  as a function of  $n$  for  $\alpha = 0.01$ ,  $\alpha = 0.05$ , and  $\alpha = 0.15$ .

- (b) In your own words, what do you observe? What happens to the algorithm as you change  $\alpha$ ? Why is this the case?
  - (c) Include snippets of your Python code implementing the SARSA algorithm, and the generation/plotting of results.
6. [20] **Q-Learning.** Implement the tabular Q-learning algorithm we discussed in class. Use the starter code available at: [https://github.com/cyrusneary/UBC\\_EECE571N\\_fall\\_2025/blob/main/Homework-starter-code/HW3/Q\\_learning\\_code.py](https://github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3/Q_learning_code.py). As a part of your implementation, please have the algorithm return both `Q_list` and `N_visit_list`. `Q_list` should be a list of numpy arrays, representing the estimated Q-function after each learning episode. Meanwhile, `N_visit_list` should also be a list of numpy arrays, each of these arrays should contain the visit counts for each state  $s$  in the gridworld, after each learning episode.
- To test your code file, use it to learn an optimal policy and state-action value function in the gridworld environment from Figure 1. Once again, run the algorithm for `num_episodes = 10000` episodes. Try each of the following learning rates:  $\alpha = 0.01$ ,  $\alpha = 0.05$ ,  $\alpha = 0.15$ .
- (a) As described in part a) of the previous question, compare plots of the RMSE vs. number of episodes for each of the tested values of  $\alpha$ .
  - (b) Use the results with learning rate  $\alpha = 0.15$  to create two plots: the first should illustrate the error  $\text{err}(s) = |\hat{V}_n^\alpha(s) - V^*(s)|$  of the estimated optimal value function after  $n = 10000$  episodes; the second plot should illustrate the number of visits  $N_n^\alpha(s)$  to each state  $s$  after  $n = 10000$  episodes. Use the implementation of `mdp.plot_values()` provided in the GridworldMDP starter code to create these plots.
  - (c) Remove the wall at position  $(0,7)$  of the gridworld, re-run the Q-learning algorithm on this modified environment, and re-produce the plots of  $\text{err}(s)$  and  $N_n^\alpha(s)$  described in b).
  - (d) In your own words, what's the difference between the plots in parts b) and c)? In which states does the algorithm's value estimates have high error? Which states are visited more frequently by the algorithm? Why does removing the wall at  $(0, 7)$  change these properties, and what does this demonstrate about the algorithm's exploration process?
  - (e) Include snippets of your Python code implementing the Q-learning algorithm, and the generation/plotting of results.

7. [20] **Model-Based RL: Dyna-Q.** Implement the Dyna-Q algorithm that we discussed in class. Use the starter code available at: [https://github.com/cyrusneary/UBC\\_EECE571N\\_fall\\_2025/blob/main/Homework-starter-code/HW3/dyna\\_q\\_code.py](https://github.com/cyrusneary/UBC_EECE571N_fall_2025/blob/main/Homework-starter-code/HW3/dyna_q_code.py).

To test your code file, use it to learn an optimal policy and state-action value function in the gridworld environment from Figure 1. Run the algorithm for `num_episodes = 10000` episodes. Use a single fixed learning rate of  $\alpha = 0.15$ . Try each of the following values for `nplan`, the number of "planning" steps to take per environment step: `nplan = 0`, `nplan = 5`, `nplan = 50`.

- (a) Similarly to as in part a) of the previous two questions, compare plots of the RMSE vs. number of episodes for each of the tested values of `nplan`.
- (b) Change the environment slip probability to  $p = 0.3$ , re-run the algorithm on this modified environment, and re-produce the RMSE plots described in part a).
- (c) In your own words, how does the value of `nplan` affect the algorithm performance? What effect does increasing the slip probability in the environment have on this performance?

Why is this the case?

- (d) Include snippets of your Python code implementing the SARSA algorithm, and the generation/plotting of results.