Reinforcement Learning: Assignment 3

Alexander Y. Shestopaloff

June, 2024

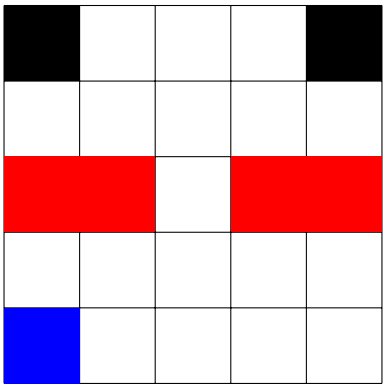
This assignment will act as the final evaluation for the course. Your goal will be to put together some of the methods that we have covered in the class and apply them to a few reinforcement learning problems.

You should provide a report including plots describing your results and a link to an online repository with commented and reproducible code. You should also provide a readme file that can make it easy for anyone to replicate the results in this assignment. Note that modern machine learning conferences e.g., NeurIPS, require submission of documented code for reproducibility of reported results.

Each part is worth 25 marks. You will be graded on correctness, clarity and reproducibility of your results. The assignment is due on August 10th. Please note that due to the proximity to the grades submission deadline no extensions will be granted. It may be done individually or in pairs. Both participants in a pair will receive the same grade and no preference will be given to people working individually or in pairs.

Part 1

. Consider the following grid world problem.



The agent starts at the blue square and moves to a neighbouring state with equal probability. If the agent moves to a red state, it receives a reward of -20 and goes back to the start, i.e., the blue square. A move between any two other states receives a reward of -1. A move that attempts to move outside of the grid receives a reward of -1. The black squares serve as a terminal states. Intuitively, you can see how the goal here is to pass through the opening in the red "wall" and get to one of the black squares and hence terminate the episode.

Use the Sarsa and Q-learning algorithms to learn the optimal policy for this task. Plot a trajectory of an agent utlizing the policy learned by each of the methods. Are they different or similar? Why or why not? You may assume to use ϵ-greedy action selection for this task. How does the sum of rewards over an episode behaves for each of these two methods.

Part 2

. Consider a scenario where we have a random walk on a 7×7 grid. That is, we are equally likely to move up, down, left, or right. Suppose that we start the random walk at the precise center of the grid.

We assume that the lower left and upper right corners are terminal states, with, respectively, rewards of -1 and 1. Rewards for transitions between two states are 0, if an attempt to transition outside the wall is made, the agent stays in the same spot and receives a reward of 0. Compute the value function for this "random walk" policy using (1) gradient Monte Carlo method and (2) the semi-gradient TD(0) method with an affine function approximation. How does it compare to the exact value function?