Exercise 5
Gregory Attra
03.6.2022
CS 7180 - Prof. Amato
Code Setup:

1. Unzipped the 'ex5.zip' file
2. 'cd' into the 'code' directory
3. run 'source ./init' to setup the pyenv

Questions:

1. TD-learning is biased on the initial estimates as $Q(s', a')$ is not a true sample like it is in MC Control. On our first episode, we haven't sampled any states yet, however, we are using our initial estimates of those states to influence our updates. Regardless, TD-learning has less variance than MC. TD learning is typically more efficient and converges to a lower RMS error. TD combines Monte Carlo sampling with DP bootstrapping. Since TD can be applied to any problem where MC can be applied, it is generally better than MC.

2.
   - The TD error is computed using a greedy target policy by maxing over A to choose the next action (greedy selection) at the next state s'. While on-policy TD-learning uses the same ϵ-soft policy to pick the next action at s'.
   - Yes, it is the same algorithm as $\max_a Q(s, a)$ is by definition a greedy policy / choice.

3. (a)
   - I do not think the conclusion would be changed by a wider range of α values. We see that, with small α values, MC has less noise than at higher values, but learns more slowly. With slightly higher values of α, MC learns faster but the noise is significant, even at $\alpha = 0.04$. Whereas, with TD(0), even relatively high values for α yield faster learning and less variance.
   - No. If we had used a higher α for MC, perhaps it would have learned faster, but the variance would be so great, it would not be consistently better than TD. If we used a higher value of α for TD, we would have seen the variance increase and the performance degrade overtime. Whereas, if we'd used a smaller α, the learning would have been slower but the variance less. The α values chosen reflect an accurate representation of how these algorithms perform at any α.

   (b) I don't know.

   (c)
   - n-Step TD is beneficial when the change in state at each timestep is not significant with respect to reaching the goal state. In other

words, moving one step towards the goal when the walk is long has a smaller influence on the probability of reaching the goal state than with a shorter walk. n-Step TD enables us to evaluate longer trajectories, or more meaningful changes in state between timestep t and $t + n$.

- With a smaller walk, having a large n would cause higher variance in the value estimate for a given state. For example, assume we are one state to the right of the left terminal state: with a sufficiently large n, we may sample a trajectory in which we end up in the goal state, thus dramatically improving the value of being in a relatively bad state.
- Changing the reward at LEFT to be -1 made the estimates more cautious / conservative. In other words: the risk of moving closer to the left-side terminal state was estimated to be much greater and so the values of states closer to the left-side terminal were estimated to be much smaller than if the reward were still 0. At a higher n, we would see greater variance as the gradient in value-estimates between states would be greater.

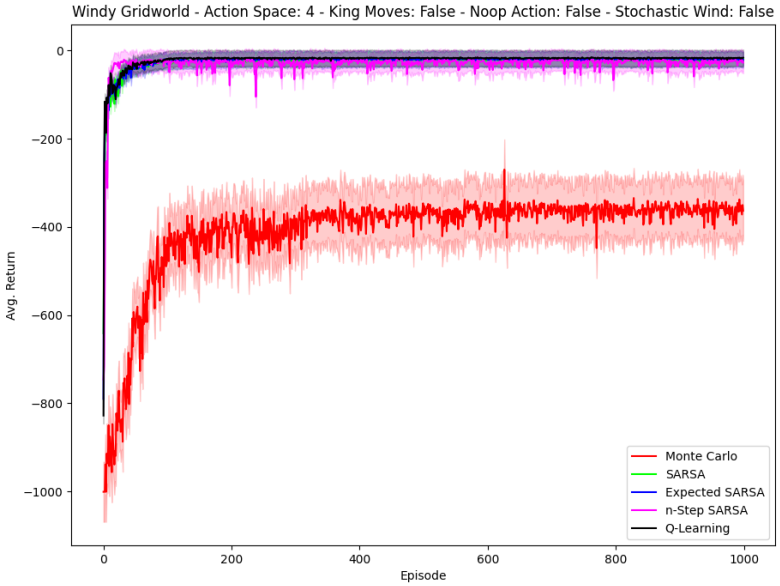
4. Windy Gridworld
   Note: to set the max number of timesteps per episode: add '–max-t ¡int¿' to the commands below.
   To run all the experiments, add '–run-all' to the commands below.

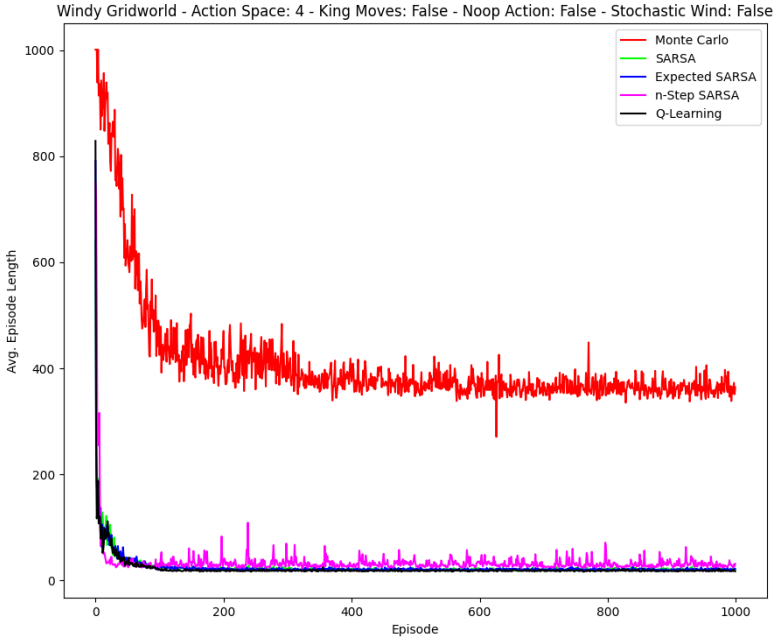
   (a) **Standard Windy Gridworld**: 'python src/temporal_difference/run_windy_gridworld.py'

   (b) **Standard Windy Gridworld with King Moves and no-move Action**: 'python src/temporal_difference/run_windy_gridworld.py –king-moves –noop-action'

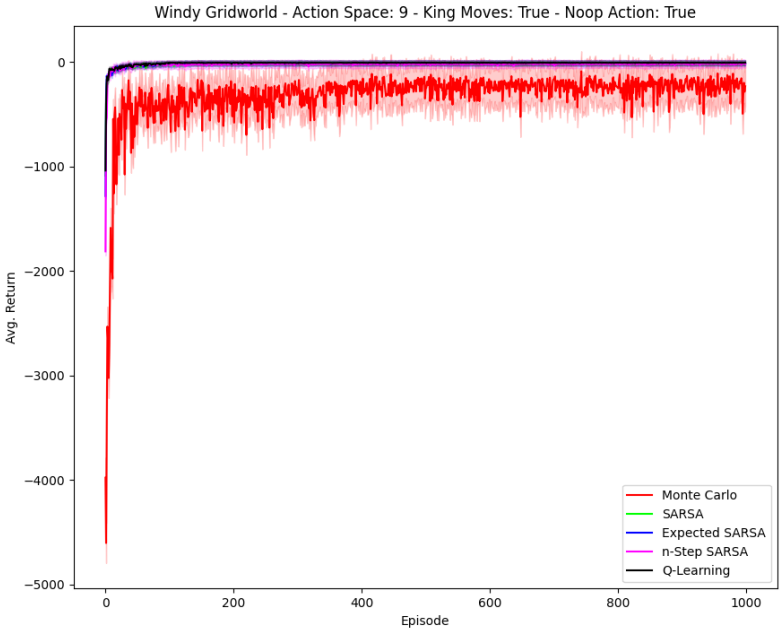
   (c) **Standard Windy Gridworld with stochastic wind**: 'python src/temporal_difference/run_windy_gridworld.py –stochastic-wind'

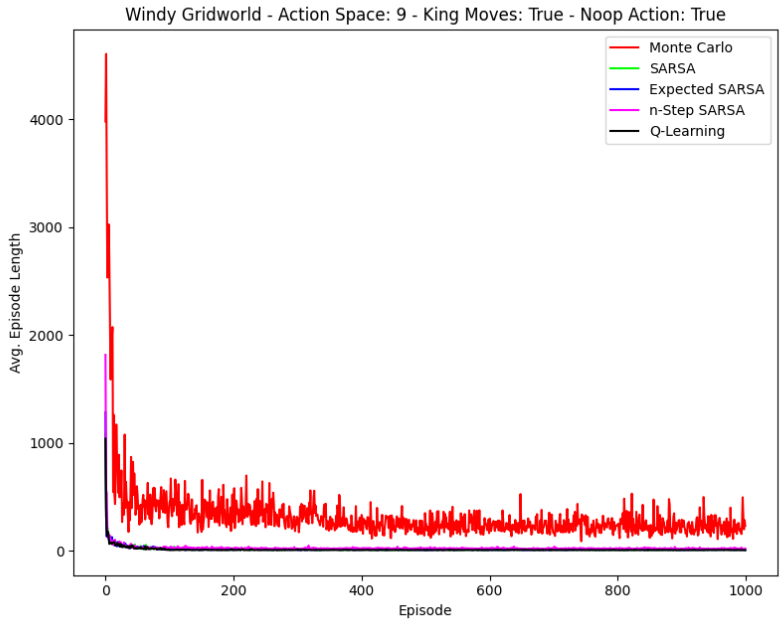
   Adding extra actions enabled MC control to peform better but still with significant variance. Adding stochasticity to the wind delayed convergence and increased variance for all algorithms.



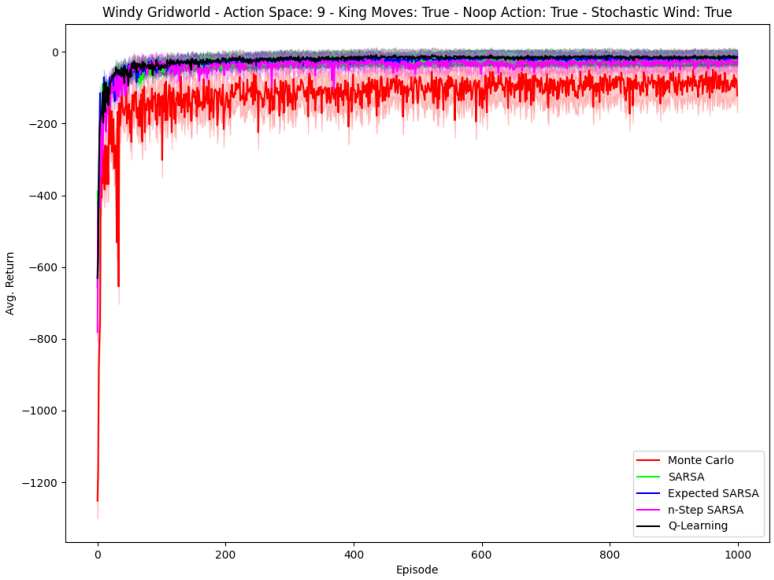


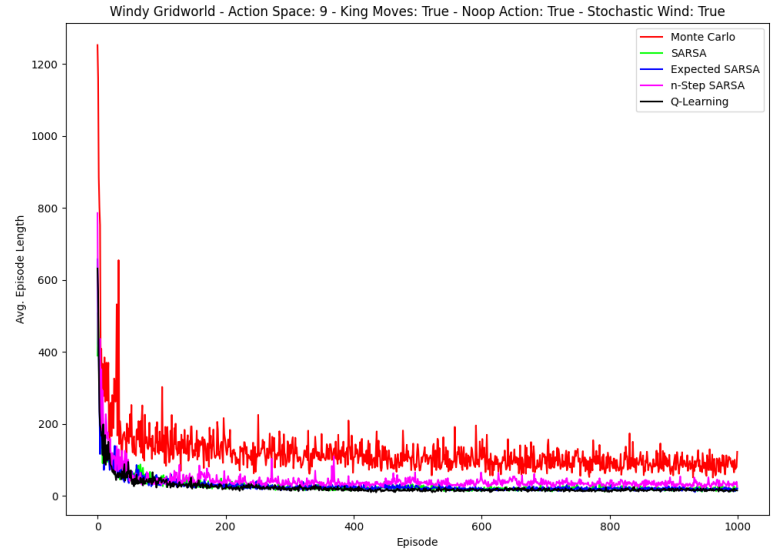
Windy Gridworld Plots





King-Moves Windy Gridworld Plots



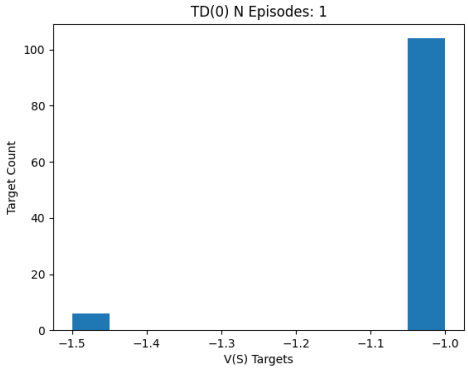


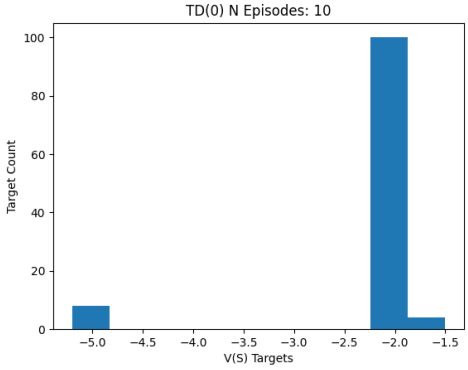
Stochastic Windy Gridworld Plots

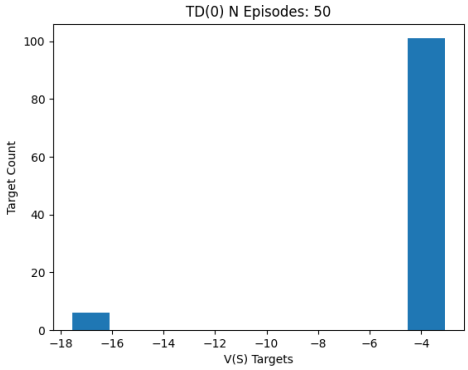
5. Bias / Variance Evaluation
   Plots below

   - To run the experiment: 'python src/temporal_difference/run_bias_variance.py'

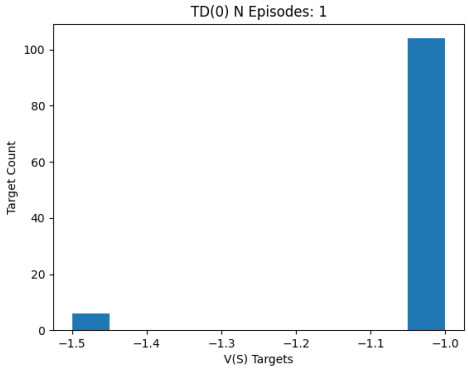
   In MC control we see a greater diversity of targets for $V(S)$, which indicates higher variance as we would have to update $V(S)$ to move towards each target. In TD we see mostly 2-3 target values and the vast majority of each target in the set falls under one target value. This indicates less variance as we don't need to update $V(S)$ as frequently.

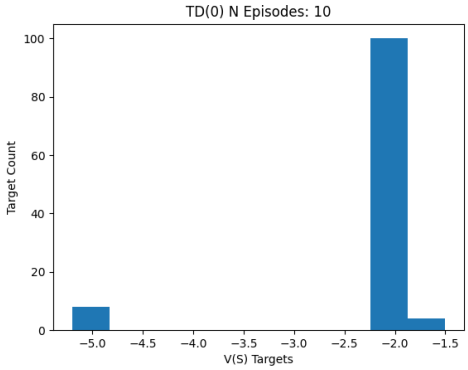


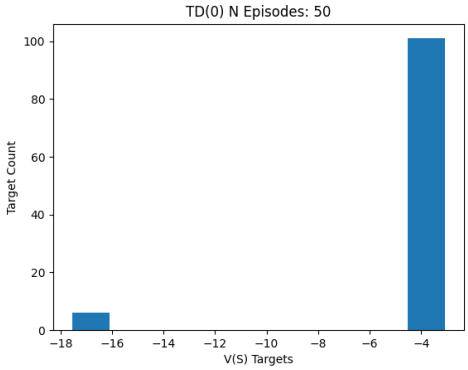




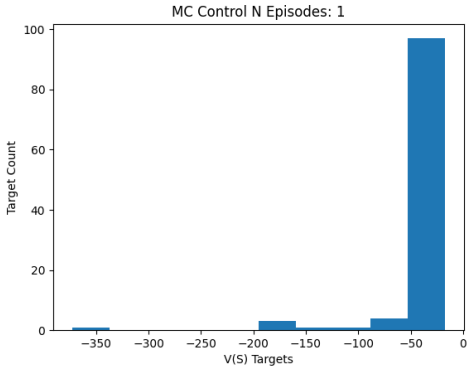
TD(0) Evaluation

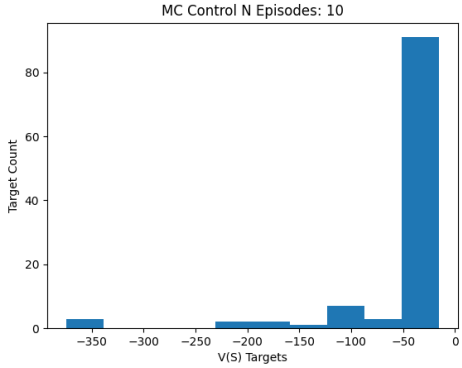


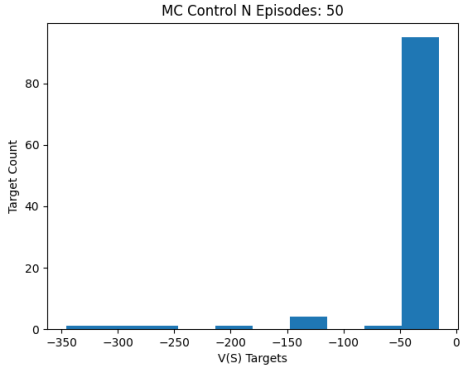




n-Step TD Evaluation







MC Evaluation