

# HOMEWORK 7

>>NAME HERE<<  
>>ID HERE<<

**Instructions:** Use this latex file as a template to develop your homework. Please submit a single pdf to Canvas. Late submissions may not be accepted. You can choose any programming language (i.e. python, R, or MATLAB). Please check Piazza for updates about the homework.

## 1 Getting Started

Before you can complete the exercises, you will need to setup the code. In the zip file given with the assignment, there is all of the starter code you will need to complete it. You will need to install the requirements.txt where the typical method is through python's virtual environments. Example commands to do this on Linux/Mac are:

```
python -m venv .venv
source .venv/bin/activate
pip install -r requirements.txt
```

For Windows or more explanation see here: <https://docs.python.org/3/tutorial/venv.html>

## 2 Value Iteration [40 pts]

The ValueIteration class in solvers/Value\_Iteration.py contains the implementation for the value iteration algorithm. Complete the train\_episode and create\_greedy\_policy methods.

### Submission [6 pts each + 10 pts for code submission]

Submit a screenshot containing your train\_episode and create\_greedy\_policy methods (10 points).

For these 5 commands. Report the episode it converges at and the reward it achieves. See examples for what we expect. An example is:

```
python run.py -s vi -d Gridworld -e 200 -g 0.2
```

Converges to a reward of \_\_\_\_ in \_\_\_\_ episodes.

**Note: For FrozenLake the rewards go to many decimal places. Report convergence to the nearest 0.0001.**

Submission Commands:

1. python run.py -s vi -d Gridworld -e 200 -g 0.05
2. python run.py -s vi -d Gridworld -e 200 -g 0.2
3. python run.py -s vi -d FrozenLake-v0 -e 500 -g 0.5
4. python run.py -s vi -d FrozenLake-v0 -e 500 -g 0.9
5. python run.py -s vi -d FrozenLake-v0 -e 500 -g 0.75

### Examples

For each of these commands. The expected reward is given for a correct solution. If your solution gives the same reward it doesn't guarantee correctness on the test cases that you report results on – you're encouraged to develop your own test cases to supplement the provided ones.

```
python run.py -s vi -d Gridworld -e 100 -g 0.9
```

Converges in 3 episodes with reward of -26.24.

```
python run.py -s vi -d Gridworld -e 100 -g 0.4
```

Converges in 3 episodes with reward of -18.64.

```
python run.py -s vi -d FrozenLake-v0 -e 100 -g 0.9
```

Achieves a reward of 2.176 after 53 episodes.

## 3 Q-learning [40 pts]

The `QLearning` class in `solvers\Q_Learning.py` contains the implementation for the Q-learning algorithm. Complete the `train_episode`, `create_greedy_policy`, and `make_epsilon_greedy_policy` methods.

### Submission [10 pts each + 10 pts for code submission]

Submit a screenshot containing your `train_episode`, `create_greedy_policy` and `make_epsilon_greedy_policy` methods (10 points).

Report the reward for these 3 commands with your implementation (10 points each) by submitting the "Episode Reward over Time" plot for each command:

1. `python run.py -s ql -d CliffWalking -e 100 -a 0.2 -g 0.9 -p 0.1`
2. `python run.py -s ql -d CliffWalking -e 100 -a 0.8 -g 0.5 -p 0.1`
3. `python run.py -s ql -d CliffWalking -e 500 -a 0.6 -g 0.8 -p 0.1`

For reference, command 1 should end with a reward around -60, command 2 should end with a reward around -25 and command 3 should end with a reward around -40.

### Example

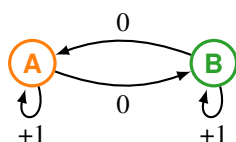
Again for this command, the expected reward is given for a correct solution. If your solution gives the same reward it doesn't guarantee correctness on the test cases.

```
python run.py -s ql -d CliffWalking -e 500 -a 0.5 -g 1.0 -p 0.1
```

Achieves a best performing policy with -13 reward.

## 4 Q-learning [20 pts]

For this question you can either reimplement your Q-learning code or use your previous implementation. You will be using a custom made MDP for analysis. Consider the following Markov Decision Process. It has two states  $s$ . It has two actions  $a$ : move and stay. The state transition is deterministic: "move" moves to the other state, while "stay" stays at the current state. The reward  $r$  is 0 for move, 1 for stay. There is a discounting factor  $\gamma = 0.8$ .



The reinforcement learning agent performs Q-learning. Recall the  $Q$  table has entries  $Q(s, a)$ . The  $Q$  table is initialized with all zeros. The agent starts in state  $s_1 = A$ . In any state  $s_t$ , the agent chooses the action  $a_t$  according to a behavior policy  $a_t = \pi_B(s_t)$ . Upon experiencing the next state and reward  $s_{t+1}, r_t$  the update is:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left( r_t + \gamma \max_{a'} Q(s_{t+1}, a') \right).$$

Let the step size parameter  $\alpha = 0.5$ .

1. (5 pts) Run Q-learning for 200 steps with a deterministic greedy behavior policy: at each state  $s_t$  use the best action  $a_t \in \arg \max_a Q(s_t, a)$  indicated by the current action-value table. If there is a tie, prefer move. Show the action-value table at the end.
2. (5 pts) Reset and repeat the above, but with an  $\epsilon$ -greedy behavior policy: at each state  $s_t$ , with probability  $1 - \epsilon$  choose what the current Q table says is the best action:  $\arg \max_a Q(s_t, a)$ ; Break ties arbitrarily. Otherwise, (with probability  $\epsilon$ ) uniformly chooses between move and stay (move or stay both with  $1/2$  probability). Use  $\epsilon = 0.5$ .
3. (5 pts) Without doing simulation, use Bellman equation to derive the true action-value table induced by the MDP. That is, calculate the true optimal action-values by hand.
4. (5 pts) To the extent that you obtain different solutions for each question, explain why the action-values differ.

## 5 A2C (Extra credit)

### 5.1 Implementation

You will implement a function for the A2C algorithm in `solvers/A2C.py`. Skeleton code for the algorithm is already provided in the relevant python files. Specifically, you will need to complete `train` for A2C. To test your implementation, run:

```
python run.py -s a2c -t 1000 -d CartPole-v1 -G 200
-e 2000 -a\ 0.001 -g 0.95 -l [32]
```

This command will train a neural network policy with A2C on the CartPole domain for 2000 episodes. The policy has a single hidden layer with 32 hidden units in that layer.

### Submission

For submission, plot the final reward/episode for 5 different values of either alpha or gamma. Then include a short (<5 sentence) analysis on the impact that alpha/gamma had for the reward in this domain.