HOMEWORK 8: REINFORCEMENT LEARNING

10-301 / 10-601 INTRODUCTION TO MACHINE LEARNING (SPRING 2022)
http://www.cs.cmu.edu/~mgormley/courses/10601/

OUT: April 12, 2022
DUE: April 21, 2022
TAs: Sana, Chu, Hayden, Tori, Prasoon

Summary In this assignment, you will implement a reinforcement learning algorithm for solving the classic mountain-car environment. As a warmup, the first section will lead you through an on-paper example of how value iteration and Q-learning work. Then, in Section 7, you will implement Q-learning with function approximation to solve the mountain car environment.

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. If your scanned submission misaligns the template, there will be a 5% penalty. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader.
  - **Programming:** You will submit your code for programming questions on the homework to Gradescope (https://gradescope.com). After uploading your code, our grading scripts will autograde your assignment by running your program on a virtual machine (VM). When you are developing, check that the version number of the programming language environment Python 3.9.6) and versions of permitted libraries (`numpy` 1.21.2 and `scipy` 1.7.1) match those used on Gradescope. You have 10 free Gradescope programming submissions. After 10 submissions, you will begin to lose points from your total programming score. We recommend debugging your implementation on your local machine (or the Linux servers) and making sure your code is running correctly first before submitting your code to Gradescope.
- **Materials:** The data that you will need in order to complete this assignment is posted along with the writeup and template on the course website.

7 Programming [68 Points]

Your goal in this assignment is to implement Q-learning with linear function approximation to solve the mountain car environment. You will implement all of the functions needed to initialize, train, evaluate, and obtain the optimal policies and action values with Q-learning. In this assignment we will provide the environment for you. The program you write will be automatically graded using the Gradescope system.

7.1 Specification of Mountain Car

In this assignment, you will be given code that fully defines the Mountain Car environment. In Mountain Car you control a car that starts at the bottom of a valley. Your goal is to reach the flag at the top right, as seen in Figure 4. However, your car is under-powered and cannot climb up the hill by itself. Instead you must learn to leverage gravity and momentum to make your way to the flag. It would also be good to get to this flag as fast as possible.

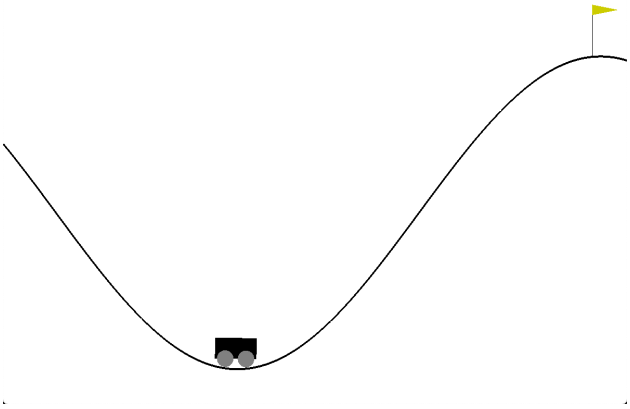


Figure 4: What the Mountain Car environment looks like. The car starts at some point in the valley. The goal is to get to the top right flag.

The state of the environment is represented by two variables, `position` and `velocity`. `position` can be between $[-1.2, 0.6]$ (inclusive) and `velocity` can be between $[-0.07, 0.07]$ (inclusive). These are just measurements along the x-axis.

The actions that you may take at any state are $\{0, 1, 2\}$, where each number corresponds to an action: (0) pushing the car left, (1) doing nothing, and (2) pushing the car right.

7.2 Q-learning with Linear Approximations

The Q-learning algorithm is a model-free reinforcement learning algorithm, where we assume we don't have access to the model of the environment the agent is interacting with. We also don't build a complete model of the environment during the learning process. A learning agent interacts with the environment solely based on calls to **step** and **reset** methods of the environment. Then the Q-learning algorithm updates the q-values based on the values returned by these methods. Analogously, in the approximation setting the algorithm will instead update the parameters of q-value approximator.

Let the learning rate be α and discount factor be γ. Recall that we have the information after one interaction with the environment, (s, a, r, s'). The tabular update rule based on this information is:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right).$$

Instead, for the function approximation setting we use the following update rule derived from the Function Approximation Section (Section 4). Note that we have made the bias term explicit here, where before it was implicitly folded into $\mathbf{w}$:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \left(q(\mathbf{s}, a; \mathbf{w}) - (r + \gamma \max_{a'} q(\mathbf{s}', a'; \mathbf{w}) \right) \nabla_{\mathbf{w}} q(\mathbf{s}, a; \mathbf{w}),$$

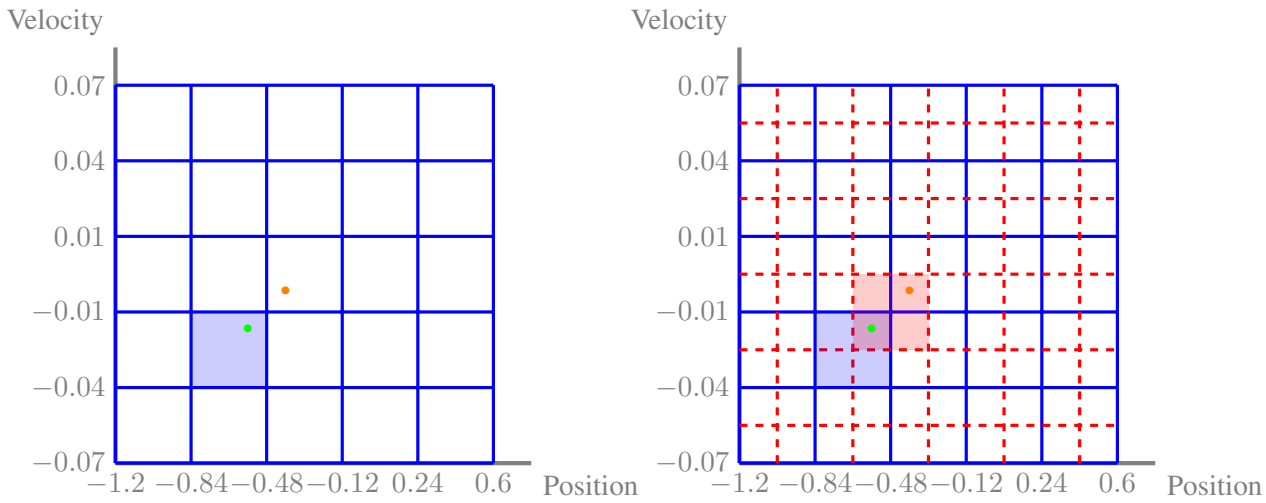
where

$$q(\mathbf{s}, a; \mathbf{w}) = \mathbf{s}^T \mathbf{w}_a + b.$$

The epsilon-greedy action selection method selects the optimal action with probability $1 - \epsilon$ and selects uniformly at random from one of the 3 actions (0, 1, 2) with probability ϵ. The reason that we use an epsilon-greedy action selection is we would like the agent to do explorations by stochastically selecting random actions with small probability. For the purpose of testing, we will test two cases: $\epsilon = 0$ and $0 < \epsilon < 1$. When $\epsilon = 0$ (no exploration), the program becomes deterministic and your output have to match our reference output accurately. In this case, **pick the action represented by the smallest number if there is a draw in the greedy action selection process**. For example, if we are at state s and $Q(s, 0) = Q(s, 2)$, then take action 0. When $0 < \epsilon < 1$, your output will need to fall in a certain range within the reference determined by running exhaustive experiments on the input parameters.

7.3 Feature Engineering

Linear approximations are great in their ease of use and implementations. However, there sometimes is a downside; they're *linear*. This can pose a problem when we think the value function itself is nonlinear with respect to the state. For example, we may want the value function to be symmetric about 0 velocity. To combat this issue we could throw a more complex approximator at this problem, like a neural network. But we want to maintain simplicity in this assignment, so instead we will look at a nonlinear transformation of the "raw" state.



(a) A discretization of the state space of Mountain Car

(b) A tiling of the state space of Mountain Car

Figure 5: State representations for the states of Mountain Car

For the Mountain Car environment, we know that `position` and `velocity` are both bounded. What we can do is draw a grid over the possible `position-velocity` combinations as seen in Figure 5a. We then enumerate the grid from bottom left to top right, row by row. Then we map all states that fall into a grid

square with the corresponding one-hot encoding of the grid number. For efficiency reasons we will just use the index that is non-zero. For example the green point would be mapped to $\{6\}$ and the orange point to $\{12\}$. This is called a *discretization* of the state space.

The downside to the above approach is that although observing the green point will let us learn parameters that generalize to other points in the shaded blue region, we will not be able to generalize to the orange point even though it is nearby. We can instead draw two grids over the state space, each offset slightly from each other as in Figure 5b. Now we can map the green point to two indices, one for each grid, and get $\{6, 39\}$ (note the index for orange grid starts from the end of blue index, i.e. 25). Now the green point has parameters that generalize to points that map to $\{6\}$ (the blue shaded region) in the first discretization and parameters that generalize to points that map to $\{39\}$ (the red shaded region) in the second. We can generalize this to multiple grids, which is what we do in practice. This is called a *tiling* or a *coarse-coding* of the state space.

7.4 Implementation Details

Here we describe the API to interact with the Mountain Car environment available to you.

- `__init__(mode, debug)`: Initializes the environment to the a mode specified by the value of `mode`. This can be a string of either "raw" or "tile".

  "raw" mode tells the environment to give you the state representation of raw features encoded in a sparse format: $\{0 \to \texttt{position}, 1 \to \texttt{velocity}\}$.

  In "tile" mode you are given indices of the tiles which are active in a sparse format: $\{T_1 \to 1, T_2 \to 1, \ldots T_n \to 1\}$ where T_i is the tile index for the ith tiling. All other tile indices are assumed to map to 0. For example the state representation of the example in Figure 5b would become $\{6 \to 1, 39 \to 1\}$.

  The dimension of the state space of the "raw" mode is 2. The dimension of the state space of the "tile" mode is 2048. These values can be accessed from the environment through the `state_space` property, and similarly for other languages.

  `debug` is an optional argument for debugging. See Section 7.5 for more details.

- `reset()`: Reset the environment to starting conditions.

- `step(action)`: Take a step in the environment with the given action. `action` must be either 0, 1 or 2. This will return a tuple of (`state`, `reward`, `done`) which is the next state, the reward observed, and a boolean indicating if you reached the goal or not, ending the episode. The `state` will be either a raw' or tile representation, as defined above, depending on how you initialized Mountain Car. If you observe `done = True` then you should `reset` the environment and end the episode. Failure to do so will result in undefined behavior.

- `render()`: Visualize the environment (not graded). Requires the installation of `pyglet`[4]. We highly recommend you to use this only after you implement everything. Do *not* use this as a tool for debugging—this should rather be used as a tool for understanding Q-learning better. It is computationally intensive to render graphics, so only call the function once every 100 or 1000 episodes. This will be a no-op in Gradescope.

You should now implement your Q-learning algorithm with linear approximations in `q_learning.py`. The program will assume access to a given environment file(s) which contains the Mountain Car environment which we have given you. **Initialize the parameters of the linear model with all 0 (and don't forget**

[4] You can install it by typing `pip install pyglet` in your shell.

to include a bias!) and use the epsilon-greedy strategy for action selection.

Your program should write a output file containing the total rewards (the returns) for every episode after running Q-learning algorithm. There should be one return per line.

Your program should also write an output file containing the weights of the linear model. The first line should be the value of the bias. Then the following $|\mathcal{S}| \times |\mathcal{A}|$ lines should be the values of weights, outputted in row major order[5], assuming your weights are stored in a $|\mathcal{S}| \times |\mathcal{A}|$ matrix.

The autograder will use the following commands to call your function:

```
$ python q_learning.py [args...]
```

where above `[args...]` is a placeholder for command-line arguments: `<env>` `<mode>` `<weight_out>` `<returns_out>` `<episodes>` `<max_iterations>` `<epsilon>` `<gamma>` `<learning_rate>`. These arguments are described in detail below:

1. `<env>`: the environment that you are running, either `mc` for Mountain Car or `gw` for Grid World.
2. `<mode>`: mode to run the environment in. Should be either `raw` or `tile`. Note that Grid World operates only in `tile` mode.
3. `<weight_out>`: path to output the weights of the linear model.
4. `<returns_out>`: path to output the returns of the agent.
5. `<episodes>`: the number of episodes your program should train the agent for. One episode is a sequence of states, actions and rewards, which ends with terminal state or ends when the maximum episode length has been reached.
6. `<max_iterations>`: the maximum of the length of an episode. When this is reached, we terminate the current episode.
7. `<epsilon>`: the value ϵ for the epsilon-greedy strategy.
8. `<gamma>`: the discount factor γ.
9. `<learning_rate>`: the learning rate α of the Q-learning algorithm.

Example command:

```
$ python q_learning.py mc raw mc_raw_weight.out mc_raw_returns.out \
 4 200 0.05 0.99 0.01
```

Example output from the above command (may not be exactly the same, but should be close up to 0.01):

`<weight_out>`

```
-7.66116708660012
1.3411763263964611
1.3419332653944924
1.3370748857368524
-0.0013201697867872468
0.0010668243394517697
0.0012565450062079566
```

[5] https://en.wikipedia.org/wiki/Row-_and_column-major_order

<returns_out>

```
-200.0
-200.0
-200.0
-200.0
```

7.5 Debugging Tips

To help with debugging, we have provided the option for printing each step of the Q-learning train function based on the reference output for the Grid World environment. We created this output by adding the `debug=True` argument when initializing the Grid World environment. You may do the same to compare your output against ours.

We recommend first checking your outputs based on a run with extremely simple parameters. Remember to set `<epsilon>=0` so the program is run without the epsilon-greedy strategy.

We have provided output on the Grid World for the following simple command:

```
$ python q_learning.py gw tile gw_simple_weight.out \
  gw_simple_returns.out 1 1 0.0 1 1
```

Once this works, you can change the parameters to be slightly more complex (such as the ones we have below), and check with our calculations again:

```
$ python q_learning.py gw tile gw_weight.out gw_returns.out \
  3 5 0.0 0.9 0.01
```

The logs for both of the above commands should be in `reference_output/gw_simple.log` and `reference_output/gw.log`, respectively.

In addition, we have provided `mc_weight.out` and `mc_returns.out` in the handout, which are generated using the following parameters:

- `<env>`: `mc`
- `<mode>`: `tile`
- `<episodes>`: `25`
- `<max_iterations>`: `200`
- `<epsilon>`: `0.0`
- `<gamma>`: `0.99`
- `<learning_rate>`: `0.005`

Example command:

```
$ python q_learning.py mc tile mc_tile_weight.out \
 mc_tile_returns.out 25 200 0.0 0.99 0.005
```

7.6 Gradescope Submission

You should submit your `q_learning.py` to Gradescope. **Any other files uploaded will be discarded or reverted back to the original version provided in the handout.** Do *not* use other file names.