XCS234 Assignment 2

Due Sunday, April 10 at 11:59pm PT.

Guidelines

1. If you have a question about this homework, we encourage you to post your question on our Slack channel, at http://xcs234-scpd.slack.com/
2. Familiarize yourself with the collaboration and honor code policy before starting work.
3. For the coding problems, you must use the packages specified in the provided environment description. Since the autograder uses this environment, we will not be able to grade any submissions which import unexpected libraries.

Submission Instructions

Written Submission: Some questions in this assignment require a written response. For these questions, you should submit a PDF with your solutions online in the online student portal. As long as the PDF is legible and organized, the course staff has no preference between a handwritten and a typeset LaTeX submission. If you wish to typeset your submission and are new to LaTeX, you can get started with the following:

- Type responses only in `submission.tex`.
- Submit the compiled PDF, **not** `submission.tex`.
- Use the commented instructions within the `Makefile` and `README.md` to get started.

Coding Submission: Some questions in this assignment require a coding response. For these questions, you should submit **all files indicated in the question** to the online student portal. For further details, see Writing Code and Running the Autograder below.

Online Submission: Some questions in this assignment require responses to be submitted interactively within a Gradescope online assessment. For these questions, you should consult your Gradescope dashboard for the availability of this assessment.

Honor code

We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solutions independently, and without referring to written notes from the joint session. In other words, each student must understand the solution well enough in order to reconstruct it by him/herself. In addition, each student should write on the problem set the set of people with whom s/he collaborated. Further, because we occasionally reuse problem set questions from previous years, we expect students not to copy, refer to, or look at the solutions in preparing their answers. It is an honor code violation to intentionally refer to a previous year's solutions. More information regarding the Stanford honor code can be found at https://communitystandards.stanford.edu/policies-and-guidance/honor-code.

Writing Code and Running the Autograder

All your code should be entered into the `src/submission/` directory. When editing files in `src/submission/`, please only make changes between the lines containing `### START_CODE_HERE ###` and `### END_CODE_HERE ###`. Do not make changes to files outside the `src/submission/` directory.

The unit tests in `src/grader.py` (the autograder) will be used to verify a correct submission. Run the autograder locally using the following terminal command within the `src/` subdirectory:

```
$ python grader.py
```

There are two types of unit tests used by the autograder:

- `basic`: These tests are provided to make sure that your inputs and outputs are on the right track, and that the hidden evaluation tests will be able to execute.
- `hidden`: These unit tests are the evaluated elements of the assignment, and run your code with more complex inputs and corner cases. Just because your code passed the basic local tests does not necessarily mean that they will pass all of the hidden tests. These evaluative hidden tests will be run when you submit your code to the Gradescope autograder via the online student portal, and will provide feedback on how many points you have earned.

For debugging purposes, you can run a single unit test locally. For example, you can run the test case `3a-0-basic` using the following terminal command within the `src/` subdirectory:

```
$ python grader.py 3a-0-basic
```

Before beginning this course, please walk through the Anaconda Setup for XCS Courses to familiarize yourself with the coding environment. Use the env defined in `src/environment.yml` to run your code. This is the same environment used by the online autograder.

Test Cases

The autograder is a thin wrapper over the python `unittest` framework. It can be run either locally (on your computer) or remotely (on SCPD servers). The following description demonstrates what test results will look like for both local and remote execution. For the sake of example, we will consider two generic tests: `1a-0-basic` and `1a-1-hidden`.

Local Execution - Hidden Tests

All hidden tests rely on files that are not provided to students. Therefore, the tests can only be run remotely. When a hidden test like `1a-1-hidden` is executed locally, it will produce the following result:

```
----- START 1a-1-hidden:  Test multiple instances of the same word in a sentence.
----- END 1a-1-hidden [took 0:00:00.011989 (max allowed 1 seconds), ???/3 points] (hidden test ungraded)
```

Local Execution - Basic Tests

When a basic test like `1a-0-basic` passes locally, the autograder will indicate success:

```
----- START 1a-0-basic:  Basic test case.
----- END 1a-0-basic [took 0:00:00.000062 (max allowed 1 seconds), 2/2 points]
```

When a basic test like `1a-0-basic` fails locally, the error is printed to the terminal, along with a stack trace indicating where the error occurred:

```
----- START 1a-0-basic:  Basic test case.
<class 'AssertionError'>
{'a': 2, 'b': 1} != None
  File "/Users/grinch/Local_Documents/Software/anaconda3/envs/XCS221/lib/python3.6/unittest/case.py", line 59, in testPartExecutor
    yield
  File "/Users/grinch/Local_Documents/Software/anaconda3/envs/XCS221/lib/python3.6/unittest/case.py", line 605, in run
    testMethod()
  File "/Users/grinch/Local_Documents/SCPD/XCS221/A1/src/graderUtil.py", line 54, in wrapper
    result = func(*args, **kwargs)
  File "/Users/grinch/Local_Documents/SCPD/XCS221/A1/src/graderUtil.py", line 83, in wrapper
    result = func(*args, **kwargs)
  File "/Users/grinch/Local_Documents/SCPD/XCS221/A1/src/grader.py", line 23, in test_0
    submission.extractWordFeatures("a b a"))
  File "/Users/grinch/Local_Documents/Software/anaconda3/envs/XCS221/lib/python3.6/unittest/case.py", line 829, in assertEqual
    assertion_func(first, second, msg=msg)
  File "/Users/grinch/Local_Documents/Software/anaconda3/envs/XCS221/lib/python3.6/unittest/case.py", line 822, in _baseAssertEqual
    raise self.failureException(msg)
----- END 1a-0-basic [took 0:00:00.003809 (max allowed 1 seconds), 0/2 points]
```

This error caused the test to fail.

In this case, start your debugging in line 23 of grader.py.

Remote Execution

Basic and hidden tests are treated the same by the remote autograder. Here are screenshots of failed basic and hidden tests. Notice that the same information (error and stack trace) is provided as the in local autograder, now for both basic and hidden tests.

1a-0-basic) Basic test case. (0.0/2.0)

```
<class 'AssertionError'>:  {'a': 2, 'b': 1} != None
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 59, in testPartExecutor
    yield
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 605, in run
    testMethod()
  File "/autograder/source/graderUtil.py", line 54, in wrapper
    result = func(*args, **kwargs)
  File "/autograder/source/graderUtil.py", line 83, in wrapper
    result = func(*args, **kwargs)
  File "/autograder/source/grader.py", line 23, in test_0
    submission.extractWordFeatures("a b a"))
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 829, in assertEqual
    assertion_func(first, second, msg=msg)
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 822, in _baseAssertEqual
    raise self.failureException(msg)
```

Just like in the local autograder, this error caused the test to fail.

Just like in the local autograder, start your debugging in line 23 of grader.py.

1a-1-hidden) Test multiple instances of the same word in a sentence. (0.0/3.0)

```
<class 'AssertionError'>:  {'a': 23, 'ab': 22, 'aa': 24, 'c': 16, 'b': 15} != None
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 59, in testPartExecutor
    yield
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 605, in run
    testMethod()
  File "/autograder/source/graderUtil.py", line 54, in wrapper
    result = func(*args, **kwargs)
  File "/autograder/source/graderUtil.py", line 83, in wrapper
    result = func(*args, **kwargs)
  File "/autograder/source/grader.py", line 31, in test_1
    self.compare_with_solution_or_wait(submission, 'extractWordFeatures', lambda f: f(sentence))
  File "/autograder/source/graderUtil.py", line 183, in compare_with_solution_or_wait
    self.assertEqual(ans1, ans2)
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 829, in assertEqual
    assertion_func(first, second, msg=msg)
  File "/autograder/source/miniconda/envs/XCS221/lib/python3.6/unittest/case.py", line 822, in _baseAssertEqual
    raise self.failureException(msg)
```

This error caused the test to fail.

Start your debugging in line 31 of grader.py.

Finally, here is what it looks like when basic and hidden tests pass in the remote autograder.

1a-0-basic) Basic test case. (2.0/2.0)

1a-1-hidden) Test multiple instances of the same word in a sentence. (3.0/3.0)

0 Introduction

In this assignment we will be exploring deep Q-learning. In particular, we will explore the application of deep Q-learning in training an agent to outperform the average human in an Atari game known as Pong! The ultimate goal of this assignment is to demonstrate the effectiveness of combining the reinforcement learning concepts we have learnt thus far with the effectiveness of neural networks as function approximators. In addition, we will convey the importance of some of the techniques used in practice to stabilize training and achieve better performance.

Some of the latter questions in this assignment make reference to research papers by DeepMind. Below you may find links to two papers which are closely related to this assignment (Note: questions whose answer requires knowledge from a given paper will contain a link to the paper within the corresponding question):

1. Human Level Control Through Deep Reinforcement Learning
2. Playing Atari with Deep Reinforcement Learning

It is important to note that this assignment contains 3 unique environment configuration files. Below we provide an explanation of each file as well as an outline of when you should use each file:

1. `environment.yml`: this is the default conda environment file which can be used for locally testing code associated to Q1-Q6 in the assignment. This is the same environment that is used by our autograder to test your code.
2. `environment_gpu.yml`: this conda environment is identical to `environment.yml` except that it ensures that the version of Pytorch which is installed is compatible with CUDA. This environment will be used when training models on the Azure vm instance or your local host using a Nvidia GPU.

Advice

- It takes approximately **8 hours** to train the DQN used in this assignment on the Atari enviroment (Q7). Be sure to allocate enough time at the end of the assignment to account for this.
- In this assignment you will make use your own vm. Please make sure to terminate any vm instance you are no longer using to avoid the loss of your allocated GPU time which is limited. Please consult the following guide to get started with setting up your vm on Azure.
- Throughout the assignment we will be using Pytorch to train our neural networks. It is strongly recommended you become familiar with the basics of Pytorch before starting the coding exercises from Q5 onwards. Please consult our Pytorch tutorial for a full review of Pytorch essentials.
- When running `run.py` with the dqn model in Q7, please ensure that the submission folder has write permissions for your user so that model weights can be saved there. This can be accomplished on the Azure vm through running `sudo chmod -R a+rw submission` from the assignment `src` directory.

Coding Deliverables

For this assignment, please submit the following files to gradescope to recieve points for coding questions:

- `src/submission/__init__.py`
- `src/submission/q3_schedule.py`
- `src/submission/q5_linear_torch.py`
- `src/submission/q6_dqn_torch.py`
- `src/submission/model.weights`

1 Distributions Induced by a Policy

In this problem, we'll work with an infinite-horizon MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ and consider stochastic policies of the form $\pi : \mathcal{S} \to \Delta(\mathcal{A})$[1]. Additionally, we'll assume that $\mathcal{M}$ has a single, fixed starting state $s_0 \in \mathcal{S}$ for simplicity.

(a) **[3 points (Written)]**

Consider a fixed stochastic policy and imagine running several rollouts of this policy within the environment. Naturally, depending on the stochasticity of the MDP $\mathcal{M}$ and the policy itself, some trajectories are more likely than others. Write down an expression for $\rho^\pi(\tau)$, the likelihood of sampling a trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ by running π in $\mathcal{M}$.

Note: Having an expression for this likelihood is very useful in practice. For further context consider the following equation which can be used to calculate the value of particular state s_0 for a policy π,

$$V^\pi(s_0) = \mathbb{E}_{\tau \sim \rho^\pi} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid s_0 \right]$$

In practice, we require the distribution of trajectories to evaluate the above expectation. The likelihood expression we derive in this question is useful in describing this distribution.

(b) **[5 points (Written)]**

Just as ρ^π captures the distribution over trajectories induced by π, we can also examine the distribution over states induced by π. In particular, define the *discounted, stationary state distribution* of a policy π as,

$$d^\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p(s_t = s),$$

where $p(s_t = s)$ denotes the probability of being in state s at timestep t while following policy π. Consider an arbitrary function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Prove the following identity:

$$\mathbb{E}_{\tau \sim \rho^\pi} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] = \frac{1}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[\mathbb{E}_{a \sim \pi(s)} \left[f(s, a) \right] \right]$$

Hint 1: Using your answer from part (a), try to understand the following identity $\mathbb{E}_{\tau \sim \rho^\pi}[\mathbb{1}[s_t = s]] = p(s_t = s)$.
Hint 2: Recall the linearity property of the expectation operator.

(c) **[5 points (Written)]**

For any policy π, we define the following function

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$$

Prove the following statement holds for all policies π, π':

$$V^\pi(s_0) - V^{\pi'}(s_0) = \frac{1}{(1-\gamma)} \mathbb{E}_{s \sim d^\pi} \left[\mathbb{E}_{a \sim \pi(s)} \left[A^{\pi'}(s, a) \right] \right]$$

Hint: The result we have proven in part (b) will help simplify your derivation.

Note: We have provided you with scaffolding for your derivation. In the provided scaffold we have substituted in our expression for $V^\pi(s_0)$ from part (a). In addition, we have reexpressed $V^{\pi'}(s_0)$ as a telescoping sum.

[1] For a finite set $\mathcal{X}$, $\Delta(\mathcal{X})$ refers to the set of categorical distributions with support on $\mathcal{X}$ or, equivalently, the $\Delta^{|\mathcal{X}|-1}$ probability simplex.

$$
\begin{aligned}
V^{\pi}(s_0) - V^{\pi'}(s_0) &= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\mathcal{R}(s_t,a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)\right)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\mathcal{R}(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right]
\end{aligned}
$$

2 Test Environment

In this question we introduce the test environment which we will be using throughout the rest of the assignment. In particular, we will use this environment to test our code locally before running our code on an Azure instance with a GPU.

To start, we will reason about optimality in the provided test environment by hand. Later in the assignment, to sanity-check your code, you will verify that your implementation is able to achieve the optimal return for the test environment. In general, you should be able to run your models in the test enviroment on your machine's CPU in a reasonable amount of time. Below you may find a full specification of the test environment,

Feature	**Specification**
States	There are four states in total: {0,1,2,3}
Actions	There are five actions in total: {0,1,2,3,4}. Actions $0 \leq i \leq 3$ makes the agent go to state i. Action 4 makes the agent stay in the same state.
Rewards	Going to state i from states 0, 1 and 3 gives a reward $R(i)$, where $R(0) = 0.2$ $R(1) = -0.1$ $R(2) = 0.0$ $R(3) = -0.3$ If we start in state 2, then the rewards defined above are multiplied by -10.
Episodes	One episode lasts 5 time steps (for a total of 5 actions) and always starts in state 0 (no rewards at the initial state).

Table 1: Specification of features for the test environment

State (s)	**Action (a)**	**Next State (s')**	**Reward (R)**
0	0	0	0.2
0	1	1	-0.1
0	2	2	0.0
0	3	3	-0.3
0	4	0	0.2
1	0	0	0.2
1	1	1	-0.1
1	2	2	0.0
1	3	3	-0.3
1	4	1	-0.1
2	0	0	-2.0
2	1	1	1.0
2	2	2	0.0
2	3	3	3.0
2	4	2	0.0
3	0	0	0.2
3	1	1	-0.1
3	2	2	0.0
3	3	3	-0.3
3	4	3	-0.3

Table 2: Transition table for the test environment

An example of a trajectory (or episode) in the test environment is shown in Figure 1, and the trajectory can be represented in terms of s_t, a_t, R_t as:

$$s_0 = 0, a_0 = 1, R_0 = -0.1, s_1 = 1, a_1 = 2, R_1 = 0.0, s_2 = 2, a_2 = 4, R_2 = 0.0, s_3 = 2, a_3 = 3, R_3 = 3.0,$$
$$s_4 = 3, a_4 = 0, R_4 = 0.2, s_5 = 0$$

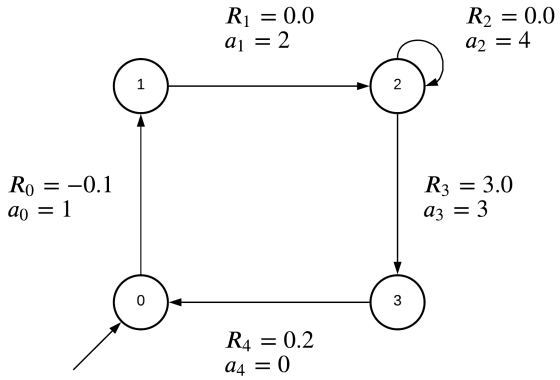


Figure 1: Example of a trajectory in the Test Environment

(a) **[4 points (Written)]**

What is the maximum sum of rewards that can be achieved in a single trajectory in the test environment, assuming $\gamma = 1$? Show first that this value is attainable in a single trajectory, and then briefly argue why no other trajectory can achieve greater cumulative reward. Please make reference to the properties outlined in the tables [1, 2] accompanying this question and note that the episode length is 5 timesteps.

3 Tabular Q-Learning

If the state and action spaces are sufficiently small, we can simply maintain a table containing the value of $Q(s, a)$, an estimate of $Q^*(s, a)$, for every (s, a) pair. In this *tabular setting*, given an experience sample (s, a, r, s'), the update rule is

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a' \in \mathcal{A}} Q(s', a') - Q(s, a) \right)$$

where $\alpha > 0$ is the learning rate, $\gamma \in [0, 1)$ the discount factor.

In addition, to formalizing our update rule for Q-learning in the tabular setting we must also consider strategies for exploration. In this question, we will be considering an ϵ-greedy exploration strategy. This strategy means that each time we look to choose an action A, we will do so as follows,

$$A \leftarrow \begin{cases} argmax_{a \in \mathcal{A}} Q(s, a) & \text{with probability } 1 - \epsilon \\ a \in \mathcal{A} \text{ chosen uniformly at random} & \text{with probability } \epsilon \end{cases}$$

In the following questions, you will need to implement both the update rule and ϵ-greedy exploration strategy for Q-learning in the tabular setting.

(a) **[4 points (Coding)]**

Implement the `get_action` and `update` functions in `q3_schedule.py`. Test your implementation by running:

```
$ python q3_schedule.py
```

(b) **[5 points (Written)]**

We will now examine the issue of overestimation bias in Q-learning. The crux of the problem is that, since we take a max over actions, errors which cause Q to overestimate will tend to be amplified when computing the target value, while errors which cause Q to underestimate will tend to be suppressed.

Assume for simplicity that our Q function is an unbiased estimator of Q^*, meaning that $\mathbb{E}[Q(s, a)] = Q^*(s, a)$ for all states s and actions a. Show that, even in this seemingly benign case, the estimator overestimates the real target in the following sense:

$$\forall s, \quad \mathbb{E}\left[\max_a Q(s, a) \right] \geq \max_a Q^*(s, a)$$

Note: The expectation $\mathbb{E}[Q(s, a)]$ is over the randomness in Q resulting from the stochasticity of the exploration process.

4 Q-Learning with Function Approximation

Please refer to question 1 of the Gradescope online assessment `A2 (Quiz)`.

5 Linear Approximation

In the following question, we will implement linear approximation in PyTorch. If you feel you need a refresher of Pytorch concepts and functionality please consult the Pytorch Tutorial we have made available. It is worth noting that the functions you implement in this section will be used in future questions of this assignment. Therefore, in order to avoid the loss of points, please ensure you that your implementation passes the basic test cases we have provided in `grader.py` for this question.

Outside of checking the tests within `grader.py` you can also test the performance of your implementation locally on the test environment **using your computer's CPU** by running:

```
$ python run.py --config_filename=q5_linear
```

(a) **[3 points (Written)]**

Suppose we represent the Q function as $Q_\theta(s,a) = \theta^\top \delta(s,a)$, where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\delta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with

$$[\delta(s,a)]_{s',a'} = \begin{cases} 1 & \text{if } s' = s, a' = a \\ 0 & \text{otherwise} \end{cases}$$

Compute $\nabla_\theta Q_\theta(s,a)$ and write the update rule for θ. Argue that equation for the tabular q-learning update rule we saw before:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a' \in \mathcal{A}} Q(s',a') - Q(s,a) \right)$$

and the following equation:

$$\theta \leftarrow \theta + \alpha \left(r + \gamma \max_{a' \in \mathcal{A}} Q_\theta(s',a') - Q_\theta(s,a) \right) \nabla_\theta Q_\theta(s,a)$$

are exactly the same when this form of linear approximation is used.

(b) **[3 points (Coding)]** Implement the `initialize_models` function in `submission/q5_linear_torch.py` which will define our linear model for both the Q-network and the target network.

(c) **[3 points (Coding)]** Implement the `get_q_values` function in `submission/q5_linear_torch.py` to output Q-values for a particular state and network.

(d) **[3 points (Coding)]** Implement the `update_target` function in `submission/q5_linear_torch.py` in order to update the target network with the Q-network weights.

(e) **[3 points (Coding)]** Implement the `calc_loss` function in `submission/q5_linear_torch.py` which will define our loss function.

(f) **[3 points (Coding)]** Implement the `add_optimizer` function in `submission/q5_linear_torch.py` which will define our optimizer.

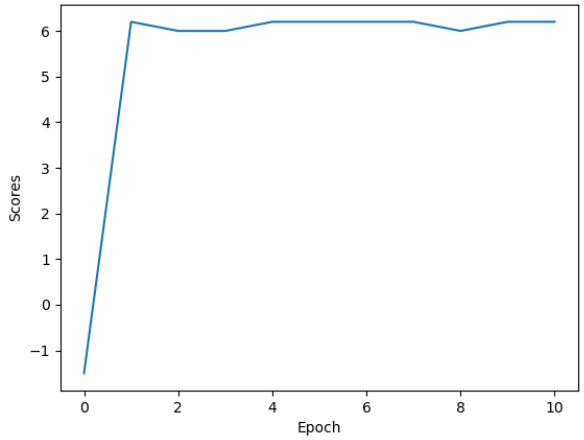
(g) **[3 points (Coding)]**

In this question, our grader will evaluate the performance of your linear implementation on the test environment based on the code you have already developed in previous questions in this section (No additional code needs to be written for this question).

If you would like, you can observe the performance metrics of your model through running the following command:

```
$ python run.py --config_filename=q5_linear
```

This should train your linear model on the test environment with the configuration defined in `config/q5_linear.yml`. You may view the evaluation scores from your training run under the following directory `results/q5_linear/`. We expect your implementation to achieve the optimal return on the test environment. Below we have provided a plot of scores which we expect the scores generated by your implementation to closely resemble:



Note: You will be need these results to provide responses to future questions which are made available online via Gradescope.

6 Implementing DeepMind's DQN

We are now ready to implement Deep Mind's DQN using Pytorch.

In the following questions, you will implement the deep Q-network as described in Human Level Control Through Deep Reinforcement Learning by implementing `initialize_models` and `get_q_values` in `q6_nature_torch.py`. The rest of the code inherits from what you wrote for linear approximation.

You will be able to test your implementation locally on the test environment **using your computer's CPU** by running:

```
$ python run.py --config_filename=q6_dqn
```

(a) **[3 points (Coding)]** Implement the `initialize_models` function in `submission/q6_dqn_torch.py` which will define our network architecture for both the Q-network and the target network.

(b) **[3 points (Coding)]** Implement the `get_q_values` function in `submission/q6_dqn_torch.py` to output Q-values for a particular state and network.

(c) **[3 points (Online)]**

Please refer to question 2.1 of the Gradescope online assessment `A2 (Quiz)`.

7 DQN on Atari

Now that you have completed your implementation of DQN it is time to train an agent to play Pong using your code. Before we start this section we would also like to provide a kind reminder to **terminate any unused Azure instances in order to avoid running out of credits**.

A Brief History of Pong:

Pong is a table tennis themed arcade game which was officially released by Atari in 1972. Bizarrely, the creation of the arcade game was originally assigned as a training exercise for staff and only later transitioned to being an officially released arcade game. Despite these humble beginings, Pong quickly became a major success and is often touted as being responsible for the rise in popularity of arcade video games.

Fortunately, we will have the opportunity to test the performance of our reinforcement learning agent on this historic game. What's more, we can expect our implementation of DQN to perform at a super human level. Let's get started!

Simulating Atari Games:

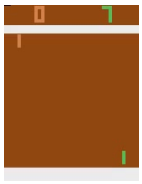
We will leverage OpenAI's gym python package to simulate the arcade game Pong. In general, OpenAI's gym package provides a suite of environments within which you may test reinforcement learning algorithms. For details of all the available environments and usage of the gym python package, please consult OpenAI's official docs.

Data Preprocessing:

The Atari environment from OpenAI gym returns observations (or original frames) of size $(210 \times 160 \times 3)$, the last dimension corresponds to the RGB channels filled with values between 0 and 255 (`uint8`). Following DeepMind's paper, we will apply the following preprocessing to the observations:

1. To encode a single frame, we take the maximum value for each pixel color value over the frame being encoded and the previous frame. In other words, we return a pixel-wise max-pooling of the last 2 observations.
2. Convert the encoded frame to grey scale; crop and rescale it to $(80 \times 80 \times 1)$. (See Figure 2)

The above preprocessing is applied to the 4 most recent observations and these encoded frames are stacked together to produce the input (of shape $(80 \times 80 \times 4)$) to the Q-function. Also, for each time we decide on an action, we perform that action for 4 time steps. This reduces the frequency of decisions without impacting the performance too much and enables us to play 4 times as many games while training. You can refer to the *Methods* section of DeepMind's paper for more details.



(a) Original input $(210 \times 160 \times 3)$

(b) Input after preprocessing $(80 \times 80 \times 1)$

Figure 2: `Pong-v0` environment

Setting up your Virtual Machine

Follow the instructions in the Azure Guide in order to create your vm instance and deploy the assignment code to your vm. Though you will need the GPU to train your model, we strongly advise that you first develop the code locally and ensure that it runs, before attempting to train it on your vm (passing the grader tests and debugging on the test environment should be sufficient). GPU time is expensive and limited. It takes approximately **8 hours** to train our DQN model. We don't want you to accidentally use all your GPU time for the assignment, debugging

your model rather than training and evaluating it. Finally, **make sure that your VM is turned off whenever you are not using it.**

In order to run the model code on your vm, please run the following command to create the proper virtual environment:

```
$ conda update -n base conda
$ conda env create --file environment_gpu.yml
```

Note: we are using the gpu version of the conda environment file.

For local development and testing, see the Introduction for instructions on which environment you should use.

If you wish to monitor your model training in real-time, you may optionally activate tensorboard and port forward the tensorboard outputs on your vm to an available port on your local machine (we strongly recommend that you do). This will enable you to view the tensorboard dashboard and your model training metrics on your local machine in real-time. To get this setup you should first start training your implementation of DQN in order to generate training metrics for tensorboard to plot (or optionally have already trained your implementation and simply wish to visualize the results).

By default tensorboard runs on port 6006. Therefore, we will look to port forward 6006 on your vm to an available port on your local machine using ssh. For this you will need to identify an available port on your local machine. To test if a process is already running on a particular port simply enter the following command in your bash session `lsof -i : <port_number>`. If this returns a process, you will need to either kill this process or find a new port (don't kill any processes without understanding its purposes and the reprecusions of stopping the process). Once you have established an available port, let's say for example 12345, then run the following command to port forward 6006 from your vm:

```
$ ssh -L 12345:localhost:6006 -p xxxxx student@ml-lab-xxxxxxx.eastus.cloudapp.azure.com
```

Where, in the above command, you will need to fill in the sections marked with "x". These values should align with those you regularly use to ssh onto your vm (see Azure Guide for further details). You should now be prompted to enter the password you have already setup for your virtual machine. Enter your password and run the following command to start tensorboard:

```
$ tensorboard --logdir=<results_folder> --host 0.0.0.0
```

Finally, open up a browser on your local machine and visit localhost:12345 to view the tensorboard dashboard.

(a) **[3 points (Online)]**

Now we're ready to train on the Atari `Pong-v0` environment. First, launch linear approximation on pong by running the following command:

```
$ python run.py --config_filename=q7_linear
```

This will train the model for $500,000$ steps and should take approximately an hour. Once training is complete observe the contents of `results/q7_linear` and in particular the plot `scores.png` (the Azure Guide contains details of moving files from the vm to your local machine).

Using this plot please refer to and answer question 3.1 of the Gradescope online assessment `A2 (Quiz)`.

(b) **[5 points (Coding)]**

In this question, we'll train the agent with DeepMind's architecture on the Atari `Pong-v0` environment. Run the following command to start the training process:

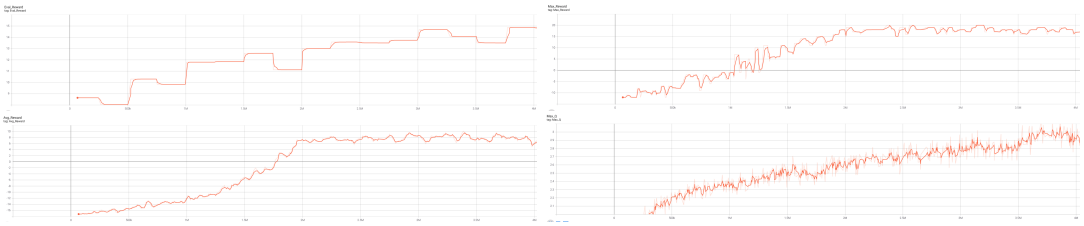
```
$ python run.py --config_filename=q7_dqn
```

To speed up training, we have trained the model for 5 million steps (these pretrained weights will be automatically loaded once you run the above command). You are responsible for training it to completion, which should take **8 hours**. You should get a score of around 12-15 after 4 million total time steps. As stated previously, the DeepMind paper claims average human performance is -3. Once your model has fully trained download the following file to your local machine `src/submission/model.weights` and include these weights with your code submission to Gradescope.

Note: the weights file needs to be in the submission folder for the autograder to read them when you run the grader on your local machine.

As the training time is roughly 8 hours, you may want to check after a few epochs that your network is making progress. The following are some training tips:

- If you terminate your terminal session, any training processes which are running within this session will terminate. In order to avoid this, you can start a session with Tmux which will persist even if you lose your connection to the vm. See the Azure guide for further details).
- The evaluation score printed on terminal should start at -21 and increase.
- The max of the q values should also be increasing.
- The standard deviation of q shouldn't be too small. Otherwise it means that all states have similar q values.
- Please find our Tensorboard graphs from one training session below.



(c) **[1 point (Online)]**

Please refer to question 3.2 of the Gradescope online assessment `A2 (Quiz)`.

(d) **[1 point (Online)]**

Please refer to questions 3.3 of the Gradescope online assessment `A2 (Quiz)`.

(e) **[1 point (Online)]**

Please refer to questions 3.4 of the Gradescope online assessment `A2 (Quiz)`.

This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the | README.md | for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

1.b

1.c

$$
\begin{aligned}
V^{\pi}(s_0) - V^{\pi'}(s_0) &= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\mathcal{R}(s_t,a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)\right)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(\mathcal{R}(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right]
\end{aligned}
$$

2.a

3.b

5.a