RL 2022/2023 Coursework

Submission deadline: 12:00 pm (noon) on 31 March 2023

February 14, 2023

1 Introduction

The goal of this coursework is to implement different reinforcement learning algorithms covered in the lectures. By completing this coursework, you will get first-hand experience on how different algorithms perform in different decision-making problems.

Throughout this coursework, we will refer to lecture slides for your understanding and give page numbers to find more information in the RL textbook (”Reinforcement Learning: An Introduction (2^{nd} edition)” by Sutton and Barto, `http://www.incompleteideas.net/book/RLbook2020.pdf`).

As stated in the course prerequisites, we do expect students to have a good understanding of Python programming, and of course any material covered in the lectures is the core foundation to work on this coursework. Many tutorials on Python can be found online.

We encourage you to start the coursework as early as possible to have sufficient time to ask any questions.

2 Contact

Piazza Please post questions about the coursework in the Piazza forum to allow everyone to view the answers in case they have similar questions. We provide different tags/folders in Piazza for each question in this coursework. Please post your questions using the appropriate tag to allow others to easily read through all the posts regarding a specific question.

Lab sessions There will also be lab sessions in person, during which you can ask questions about the coursework. We highly recommend attending these sessions, especially if you have questions about PyTorch and the code base we use. The lab sessions schedule can be accessed at this link.

Note Please keep in mind that Piazza questions and lab sessions are public for discussions. Given that this coursework is individual work and graded, please do not disclose or discuss any information which could be considered a hint towards or part of the solution to any of the questions. However, you can ask and we encourage any questions about instructions that are unclear to you, questions generally asking about algorithms (disconnected from their implementation) and concepts. Please, always ask yourself prior to posting whether you believe your question in itself discloses implementation details or might provoke answers disclosing such information.

We understand that Piazza is a very valuable place to discuss many matters on this course between students and teaching staff, but also between students. Particularly at these times, where exchange among students is severely limited due to (mostly) remote teaching, Piazza can be one of the few places such exchange can be done. We are committed to make this exchange as simple and effective as possible and hope you keep these boundaries in mind about questions regarding the coursework.

3 Getting Started

To get you started, we provide a repository of code to build upon. Each question specifies which sections of algorithms you are expected to implement and will point you to the respective files.

1. **Installing Python3**
   The code base is fully written in Python and we expect you to use several standard machine learning packages to write your solutions with. Therefore, start by downloading Python to your local machine. We recommend you use at least Python version 3.8.

   Python can be installed using the official installers (`https://www.python.org/downloads/`) or alternatively using a respective package-manager on Linux or Homebrew (`https://brew.sh`) on macOS.

2. **Create a virtual environment**
   After installing Python, we highly recommend creating a virtual environment (below we provide instructions for `virtualenv`, another common alternative is conda) to install the required packages. This allows you to neatly organise the required packages for different projects and avoid potential issues caused by insufficient access permissions on your machines. On Linux or macOS machines, type the following command in your terminal:

   ```
   python3 -m venv <environment name>
   ```

   You should now see a new folder with the same name as the environment name you provided in the previous command. In your current directory, you can then execute the following command to activate your virtual environment on Linux or macOS machines:

   ```
   source <environment name>/bin/activate
   ```

   If you are using Windows, please refer to the official Python guide for detailed instructions.

3. **Download the code base to get started**
   Finally, execute the following command to download the code base:

   ```
   git clone https://github.com/uoe-agents/uoe-rl2023-
       coursework.git
   ```

   Navigate to <**Coursework directory with setup**> and execute the following command to install the code base and the required dependencies:

   ```
   pip3 install -e .
   ```

   Note that you may encounter problems during the installation of the above packages on macOS Ventura. If that happens, please try updating your macOS to Ventura 13.1 and Xcode to 14.2.

For detailed instructions on Python's library manager `pip` and virtual environments, see the official Python guide and this guide to Python's virtual environments.

4 Overview

The coursework contains a total of **100 marks** and counts towards **50% of the course grade**. Below you can find an overview of the coursework questions and their respective marks. More details on required algorithms, environments and required tasks can be found in Section 5. Submissions will be marked based on correctness and performance as specified for each question. In Questions 2, 3 and 5, some marks are given based on a short write-up or an answer to a multiple-choice question. When relevant, you will be instructed to provide these answers as the output of a dedicated function in the `answer_sheet.py` script located at the root of the `rl2023` directory (refer to Figure 6 for a breakdown of the folder structure). Details on marking can be found in Section 6 and Section 7 presents instructions on how to submit the required assignment files.

Question 1 – Dynamic Programming **[15 Marks]**

- Implement the following DP algorithms for MDPs
  - Value Iteration [7.5 Marks]
  - Policy Iteration [7.5 Marks]

Question 2 – Tabular Reinforcement Learning **[20 Marks]**

- Implement ϵ-greedy action selection [2 Marks]
- Implement the following RL algorithms
  - Q-Learning [7 Marks]
  - On-policy first-visit Monte Carlo [7 Marks]
- Analyse performance of different hyperparameters in Taxi-v3 [4 Marks]

Question 3 – Deep Reinforcement Learning **[32 Marks]**

- Implement the following Deep RL algorithms
  - Deep Q-Networks [6 Marks]
  - REINFORCE [9 Marks]
- Reinforce performance analysis [2 Marks]
- DQN performance analysis
  - Implement ϵ-scheduling strategies [4 Marks]
  - Select best hyperparameter profiles [2 Marks]
  - Answer questions on ϵ-scheduling [4 Marks]
- Answer questions related to the DQN loss during training [5 Marks]

Question 4 – Continuous Deep Reinforcement Learning **[18 Marks]**

- Implement DDPG for continuous RL [13 Marks]
- Tune the specified hyperparameters to solve Bipedal Walker [5 Marks]

Question 5 – Fine-tuning the Algorithms **[15 Marks]**

- Tune all hyperparameters to maximise score on Bipedal Walker [10 Marks]
- Explain how the above hyperparameter are selected [5 Marks]

5 Questions

Question 1 – Dynamic Programming [15 Marks]

Description

The aim of this question is to provide you with better understanding of dynamic programming approaches to find optimal policies for Markov Decision Processes (MDPs). Specifically, you are required to implement the Policy Iteration (PI) and Value Iteration (VI) algorithms.

For this question, **you are only required to provide implementation of the necessary functions**. For each algorithm, you can find the functions that you need to implement under Tasks below. Make sure to carefully read the code documentation to understand the input and required outputs of these functions. We will mark your submission only based on the correctness of the outputs of these functions.

Algorithms

1. Policy Iteration (PI):
   You can find more details including pseudocode in the RL textbook on page 80. Also see Lecture 4 on dynamic programming (pseudocode on slide 17).

2. Value Iteration (VI):
   You can find more details including pseudocode in the RL textbook on page 83. Also see Lecture 4 on dynamic programming (pseudocode on slide 22).

Domain

In this exercise, we train dynamic programming algorithms on MDPs. We provide you with functionality which enables you to define your own MDPs for testing. For an example on how to use these functions, see the main function at the end of `exercise1/mdp_solver.py` where the ”Frog on a Rock“ MDP from the tutorials shown in Figure 1 is defined and given as input to the training function with $\gamma = 0.8$.

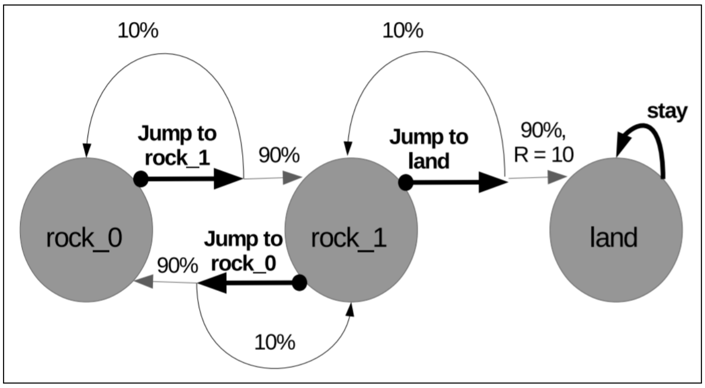


Figure 1: Frog on a Rock example MDP for Exercise 1

As a side note, our interface for defining custom MDPs requires all actions to be valid over all states in the state space. Therefore, remember to include a probability distribution over next states for every possible state-action pair to avoid any errors from the interface.

Tasks

Use the code base provided in the directory `exercise1` and implement the following functions.

1. **Value Iteration** [7.5 Marks]
   To implement the Value Iteration algorithm, you must implement the following functions in the `ValueIteration` class:

   - `_calc_value_func`, which must calculate the value function (table).
   - `_calc_policy`, which must return the greedy deterministic policy given the calculated value function.

2. **Policy Iteration** [7.5 Marks]
   To implement the Policy Iteration algorithm, you must implement the following functions in the `PolicyIteration` class:

   - `_policy_eval`, which must calculate the value function of the current policy.
   - `_policy_improvement`, which must return an improved policy and terminate if the policy is stable (hint: this function will need to call `_policy_eval`).

Aside from the aforementioned functions, the rest of the code base for this question **must be left unchanged**. A good starting point for this question would be to read the code base and the documentations to get a better grasp how the entire training process works.

Directly run the file `mdp_solver.py` to print the calculated policies for VI and PI for a test MDP. Feel free to tweak or change the MDP and make sure it works consistently.

This question does not require a lot of effort to complete and you can provide a correct implementation with less than 50 lines of code. Additionally, training the method should require less than a minute of running time.

Question 2 – Tabular Reinforcement Learning [20 Marks]

Description

The aim of the second question is to provide you with practical experience on implementing model-free reinforcement learning algorithms with tabular Q-functions. Specifically, you are required to implement the **Q-Learning** and **on-policy first-visit Monte Carlo** algorithms.

For all algorithms, **you are required to provide implementations of the necessary functions**. You can find the functions that you need to implement below. Make sure to carefully read the documentation of these functions to understand their input and required outputs. We will mark your submission based on the **correctness of the outputs of the required functions**, the **performance of your learning agents measured by the average returns on the Taxi-v3 environment**, and the **answers** you've provided in `answer_sheet.py`.

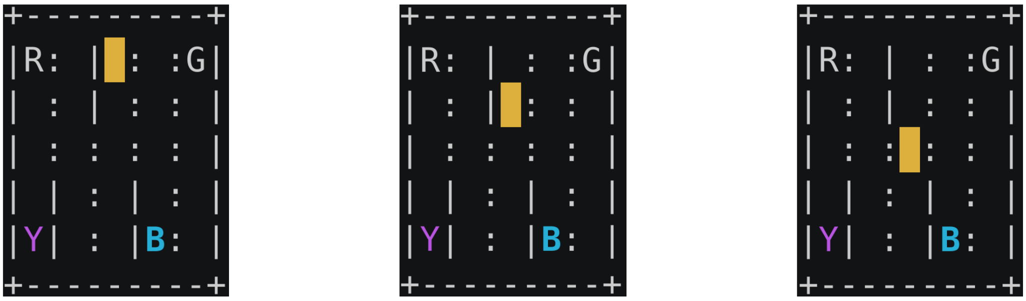
Algorithms

1. Q-Learning (QL):
   You can find more details including pseudocode for QL in the RL textbook on page 131. Also see Lecture 6 on Temporal Difference learning (slide 19).

2. First-visit Monte Carlo (MC):
   You can find more details including pseudocode for on-policy first-visit MC with ϵ-soft policies in the RL textbook on page 101. Also see Lecture 5 on MC methods (slide 17).

Domain

In this question, we train agents on the OpenAI Gym Taxi-v3 environment. This environment is a simple task where the goal of the agent is to navigate a taxi (yellow box - empty taxi; green box - taxi with passenger) to a passenger (blue location), pick it up and drop it off at the destination (purple location) in a grid-world.



```
+---------+
|R: |█: :G|
| : | : : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
```

```
+---------+
|R: | : :G|
| : |█: : |
| : : : : |
| | : | : |
|Y| : |B: |
+---------+
```

```
+---------+
|R: | : :G|
| : | : : |
| : :█: : |
| | : | : |
|Y| : |B: |
+---------+
```

Figure 2: Rendering of two Taxi-v3 environment steps

The episode terminates once the passenger is dropped off at its destination or at a maximum episode length. The agent will be given a reward of -1 at each timestep, a reward of +20 for successfully delivering the passenger to its destination and -10 for executing actions pickup or dropoff illegally, i.e. trying to pickup a passenger at a location where no passenger is located or attempting to drop off without having a passenger in the taxi. Hence, the task consists of learning to navigate the grid-world and bringing the passenger as quickly to its target destination as possible.

A good hyperparameter scheduling for both algorithms should enable the agent to solve the Taxi-v3 environment. **We consider the environment to be solved when the agent can consistently achieve an average return of** ≥ 7.

Tasks

For this exercise, you are required to implement the functions listed below. Besides the correctness of these functions, we will also mark the performance achieved by your agents with the hyperparameters we provide in the Taxi-v3 environment. See each paragraph below for more details on required functions and respective marks.

Implementation **[14 Marks]**

Use the code base provided in the directory `exercise2` and implement the following functions. All the functions that you need to implement for the three algorithms are located in the `agents.py` file. Both algorithms to implement extend the `Agent` class provided in the script.

1. **Base class** [2 Marks]
   In the `Agent` class, implement the following function:
   - `act`, where you must implement the ϵ-greedy exploration policy used by the QL and MC algorithms.
2. **Q-Learning** [6 Marks]
   To implement QL, you must implement the following functions in the `QLearningAgent` class:
   - `learn`, where you must implement Q-value updates.
3. **On-policy first-visit Monte Carlo** [6 Marks]
   To implement the MC with ϵ-soft policy algorithm, you must implement the following functions in the `MonteCarloAgent` class:
   - `learn`, where you must implement the first-visit MC Q-value updates.

Note: All other functions apart from the aforementioned ones **should not be changed**. All functions could be implemented with around 20 lines of code or less. We implemented a hyperparameter scheduler for ϵ in the file `exercise2/agents.py`, **do not change** the `schedule_hyperparameters` functions.

Testing

You can find the training script for QL and MC on Taxi-v3 in `train_q_learning.py` and `train_monte_carlo.py` respectively. These execute training and evaluation using your implemented agents.

Hyperparameters and Performance **[6 Marks]**

Besides correctness of the action selection and learning functions, we also ask you to tune different hyperparameters of your QL and MC agents. As you will see, the performance of RL algorithms is highly dependent on the choices of hyperparameter values, and we hope the following questions help you build some intuition for selecting them. For this question, we will only ask you to collect and analyse the evaluation returns of the two algorithms with different hyperparameter combinations. In the following Table 1, we provide two hyperparameter profiles for each algorithm. You can set the values of these hyperparameters through the `CONFIG` **in** `train_q_learning.py` **and** `train_monte_carlo.py`**.** In `util/result_processing.py` we have provided the class `Run` that may be used to log data across runs. You are welcome to use it during your experiments (or to expand it or replace it by any method or framework you see fit). Please run your implementation with the hyperparameter profiles we provide, and record the corresponding evaluation returns.

Note that **the best evaluation return of a correct implementation will be** ≥ 7 with one of the hyperparameter profiles provided in Table 1 and correct implementations, for both algorithms.

Algorithm	α	ϵ	γ	Algorithm	ϵ	γ
Q-Learning	0.05	0.6	0.99	First-visit Monte Carlo	0.6	0.99
	0.05	0.6	0.8		0.6	0.8

Table 1: The given **hyperparameter profiles** for QL and MC in the Taxi-v3 environment.

Analyse the evaluation returns obtained by the above hyperparameter profiles, and answer the following questions in `answer_sheet.py`:

i) `question2_1` for the QL algorithm, which value of γ leads to the best average evaluation return?
[1 Marks]

ii) `question2_2` for the first-visit MC algorithm, which value of γ leads to the best average evaluation return?
[1 Marks]

iii) `question2_3` between the two algorithms (QL / MC), whose average evaluation return is impacted by the above factor in a greater way? [1 Marks]

iv) `question2_4` provide a short explanation ($<$ 100 words) as to why the value of γ affects more the evaluation returns achieved by [Q-learning / First-Visit Monte Carlo] when compared to the other algorithm. [3 Marks]

Note: there exist hyperparameter combinations that achieve higher scores than the ones provided, and we encourage keen students to search for better ones as an exercise. However, you will not get extra marks for doing so in this question or in Question 3. You will get **no marks** for reporting a hyperparameter profile that is not among the ones proposed. Likewise, make sure the other hyperparameters are set to their **default values** for that environment, which are provided in `EX2_CONSTANTS` in `constants.py`. During our evaluation, we will use the original `constants.py` to overwrite the same file in your submission. Therefore, any change in `constants.py` will be ineffective.

Question 3 – Deep Reinforcement Learning [32 Marks]

Description

In this question you are required to implement two Deep Reinforcement Learning algorithms: **DQN** [2] and **REINFORCE** [4] with function approximation.

In this task, you are **required to implement functions associated with the training process, action selection along with gradient-based updates done by each agent**. Aside from these functions, many components of the training process, along with the primary training setup have already been implemented in our code base. Below, you can find a list of functions that need to be implemented. Make sure to carefully read the documentation of functions you must implement to understand the inputs and required outputs of each component. We will mark your submission based on **the correctness of the functions you've implemented**, along with the **answers associated with this question** you've provided in `answer_sheet.py`.

Algorithms

Before you start implementing your solutions, we recommend reading the original papers and looking at lectures and textbooks to provide you with better understanding of the details of both algorithms.

1. Deep Q-Networks (DQN):
   DQN is one of the earliest Deep RL algorithms, which replaces the usual Q-table used in Q-Learning with a neural network to scale Q-Learning to problems with large or continuous state spaces. You can find more details including pseudocode for DQN in the Nature publication [2]. Also see Lecture 12 on deep RL (pseudocode on slide 17).

2. REINFORCE:
   REINFORCE is an on-policy algorithm which learns a stochastic policy with gradient updates being derived by the policy gradient theorem (see Lecture 11, slide 11). You can find more details in the publication [4] and for pseudocode refer to Algorithm 1 provided below.

Domains

In this question, we train agents on the OpenAI Gym CartPole and Acrobot environments. CartPole is a well-known control task where the agent can move a cart left or right to balance a pole. The goal is to learn balancing the pole for as long as possible. Episodes are limited in length and terminate early whenever the pole tilts beyond a certain degree. The agent is rewarded for each timestep it achieves to maintain the pole in balance.

Acrobot is another control task in which two links are connected by an actuatable joint to form a chain, with the top end of the chain fixed in place. The agent may apply a positive, negative or null torque on the actuatable joint, and its goal is to get the free end of the chain to reach a given height. The agent receives a negative -1 reward at each timestep, and therefore is encouraged to reach the target height as quickly as possible. Episodes have a maximum length and are terminated early if the goal is reached.

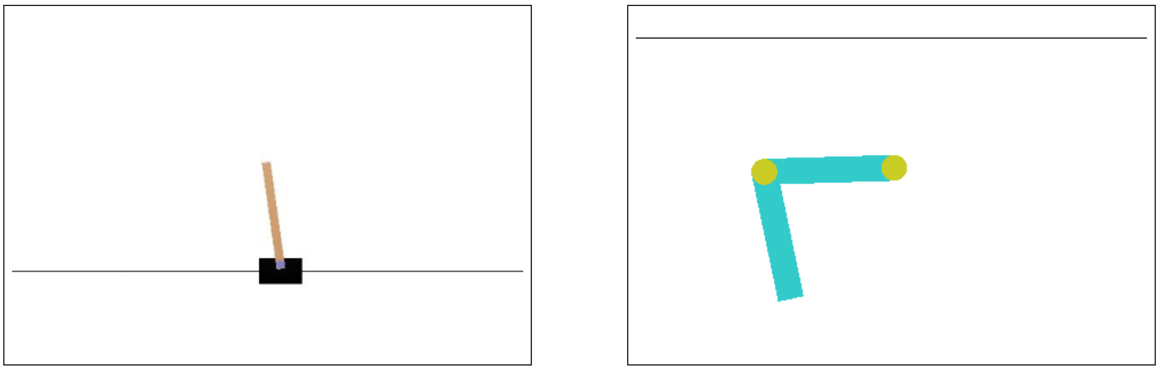


Figure 3: Rendering of the CartPole and Acrobot environments

Tasks

For this exercise, you are required to implement the functions listed below. Besides the correctness of these functions, we will also evaluate your choice of hyperparameters for the REINFORCE agent in the Acrobot environment and for the DQN agent in the CartPole environment. To simplify the hyperparameter search, we will provide you with a range of hyperparameter profiles to pick from.

Implementation **[15 Marks]**

Use the code base provided in the directory `exercise3` and implement the following functions. All of the functions which you need to implement for both algorithms are located in the `agents.py` file. Both algorithms to implement extend the `Agent` class provided in the script.

1. **DQN** [6 Marks]

   In `agents.py`, you will find the `DQN` class which you need to complete. For this class, implement the following functions:

   - `__init__`, which creates a DQN agent. Here, you can set any hyperparameters and initialise any values for the class you need.
   - `act`, which implements a ϵ-greedy action selection. Aside from the observation, this function also receives a boolean flag as input. When the value of this boolean flag is `True`, agents should follow the ϵ-greedy policy. Otherwise, agents should follow the greedy policy. This flag is useful when we interchange between training and evaluation.
   - `update`, which receives a batch of N (batch size) experience samples from the replay buffer. Using experiences, which are tuples in the form of $< s, a, r, d, s' >$ gathered from the replay buffer, update the parameters of the value network to minimize the mean squared error:

     $$\mathbb{L}_\theta = \frac{1}{N} \sum_{i=1}^{N} \left(r_i + \gamma(1-d_i) max_a Q(a|s'_i;\theta') - Q(a_i|s_i;\theta)\right)^2,$$

     where θ and θ' are the parameters of the value and target network, respectively. Also, this function is required to update the target network parameters at the stated update frequency by overwriting it with the current Q-network parameters $\theta' \leftarrow \theta$ (hard update).

2. **REINFORCE** [9 Marks]

   The functions that you need to implement for REINFORCE are also located inside the `agents.py` file under the `Reinforce` class. For this class, provide the implementation of the following functions:

   - `__init__`, which creates the REINFORCE agent. You can set additional hyperparameters and values required for training the agent here.
   - `act`, which implements the action selection based on the stochastic policy produced by the policy network.
   - `update`, which updates the policy based on the sequence of experience

     $$\{< s_t, a_t, r_t, d_t, s_{t+1} >\}_{t=1}^{T}$$

     received by the agent during an episode. You must then implement a process that updates the policy parameters to minimize the following function:

     $$\mathbf{L}_\theta = \frac{1}{T} \sum_{t=1}^{T} -\log(\pi(a_t|s_t;\theta))(G_t)$$

     where θ are the parameters of the policy network, and G_t is the discounted reward-to-go calculated starting from timestep t.

   You can find the pseudocode for REINFORCE below in Algorithm 1.

All other functions apart from the aforementioned ones **should not be changed**. In general, all of the required functions can be implemented with less than 20 lines of code.

```
Algorithm 1: REINFORCE: Monte-Carlo Policy Gradient
  Output:
      $\pi(a|s,\theta^*)$ : optimised parameterised policy
  Input:
      $\alpha$ : Learning rate
      $\gamma$ : Discount factor
  Initialise:
      $\pi(a|s,\theta)$ : Randomly initialise policy parameters $\theta$
  ------------------------------------------------------------------
  Loop forever (for each episode):
      Generate an episode $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$
      $L_\theta \leftarrow 0$                              // Initialise loss to 0
      $G \leftarrow 0$                                     // Initialise the returns to 0
      Loop backward in the episode $t = T-1, ..., 0$ :
          $G \leftarrow R_{t+1} + \gamma G$
          $L_\theta \leftarrow L_\theta - G \log \pi(A_t|S_t,\theta)$
      $L_\theta \leftarrow L_\theta / T$
      Perform a gradient step with learning rate $\alpha$ on $L_\theta$ with respect to $\theta$
```

Testing

To test your implementation, we provide you with two scripts which execute your DQN and REINFORCE implementations. You can find the scripts inside `train_dqn.py` and `train_reinforce.py` to train DQN and REINFORCE, respectively. Inside these scripts, we provide you with configurations that enable you to train the algorithms in the CartPole (for DQN) and Acrobot (for REINFORCE) environments. To better understand how your implemented functions are used in the training process, read the code and documentation provided in these scripts.

For a correct implementation, the training process requires less than 2 minutes to train DQN in CartPole and less than 10 minutes to train REINFORCE in Acrobot.

Hyperparameter tuning [12 Marks]

Besides correctness of the aforementioned algorithms, we also ask you to tune different hyperparameters of your DQN and REINFORCE agents. For this question, we will only ask you to tune one hyperparameter at a time, and you will be provided with a number of profiles to choose from for each parameter. To get full marks, you only need to select the best performing hyperparameter value among the ones proposed. We will give you hints in the form of the score to expect with the right hyperparameter choice.

There exists hyperparameter combinations that achieve higher scores than the ones provided, and we encourage keen students to search for better ones as an exercise. However you will not get extra marks for doing so in this question.

You will get **no marks** for reporting an hyperparameter value that is not among the ones proposed. Likewise, make sure the other hyperparameters are set to their **default values** for that environment, which are provided in `CARTPOLE_CONFIG` in `train_dqn.py` and in `ACROBOT_CONFIG` in `train_reinforce.py`. We recommend running at least **10 seeds per hyperparameter configuration** for statistical consistency.

In `util/result_processing.py` we have provided the class `Run` and some helper functions that may be used to log and process your results. You are welcome to use it during your experiments and to expand it or replace it by any method or framework you see fit.

Algorithm	`learning_rate`
Reinforce	$6e-1$
	$6e-2$
	$6e-3$

Table 2: Provided hyperparameters for tuning the learning rate for Reinforce in the Acrobot environment.

Epsilon decay strategy	`exploration_fraction`	Epsilon decay strategy	`epsilon_decay`
Linear	0.75	Exponential	1.0
	0.25		0.75
	0.01		0.001

Table 3: Provided hyperparameters for tuning epsilon scheduling for DQN in the CartPole environment.

1. **REINFORCE** [2 Marks]
   For REINFORCE, we simply ask you to tune the learning rate in the Acrobot environment. You are not required to perform any hyperparameter tuning in CartPole. You can find the possible values to pick from for the learning rate in Table 2 and in `train.py`, under the variable `ACROBOT_HPARAMS`. In `question3_1` of `answer_sheet.py`, report which learning rate achieves the highest mean returns at the end of training.

   Hint: You should expect a score of at least -400 for the best performing profile.

2. **DQN** [10 Marks]
   We ask you to implement different *epsilon scheduling* strategies for DQN and tune them in the CartPole environment.

   (a) **Implementing an ϵ-scheduling strategy:** When following an ϵ-greedy policy, it can be beneficial to not keep ϵ constant but instead gradually decay it over the course of training. In this question, you will experiment with two different decay strategies and select hyperparameters for them. In the `DQN` class of `agents.py`, you are asked to implement the following inner functions within the `schedule_hyperparameters` function.

      i. `epsilon_linear_decay` - hyperparameters [$\epsilon_{\texttt{start}}$, $\epsilon_{\texttt{min}}$, `exploration_fraction`]: decays ϵ linearly from some starting value $\epsilon_{\texttt{start}}$ to a minimum value $\epsilon_{\min}$. After reaching $\epsilon_{\texttt{min}}$, ϵ remains constant. ϵ should reach $\epsilon_{\texttt{min}}$ when the ratio between the current train timestep and the maximum number of train timesteps ${}^{t}/_{t_{\max}}$ reaches the value set by `exploration_fraction`.

      ii. `epsilon_exponential_decay` - hyperparameters [$\epsilon_{\texttt{start}}$, $\epsilon_{\texttt{min}}$, `epsilon_decay`]: decays ϵ exponentially such that $\epsilon_{t+1} \leftarrow r^{1/t_{\max}}\epsilon_t$, where $r \in (0,1)$ is the decay rate set by `epsilon_decay`. ϵ decays from some starting value $\epsilon_{\texttt{start}}$ to a minimum value $\epsilon_{\texttt{min}}$. After reaching $\epsilon_{\texttt{min}}$, ϵ remains constant.

   (b) **Tuning the ϵ-scheduling strategy:** In `train_dqn.py`, we have provided you with a range of possible values for ϵ-scheduling in CartPole (these are also reported in Table 3). Try out the different `exploration_fraction` values in `CARTPOLE_HPARAMS_LINEAR_DECAY` and the `epsilon_decay` values in `CARTPOLE_HPARAMS_EXP_DECAY`, and report which profile achieves the highest mean returns achieved at the end of training for each scheme in `question3_2` and `question3_3` of `answer_sheet.py`.

      Hint: You should expect a score of at least 390 for the best performing profile.

   (c) In `answer_sheet.py`, answer the following questions:

      i) `question3_4`: What would the value of epsilon be at the end of training when employing an exponential decay strategy with `epsilon_decay` set to 1.0?

      ii) `question3_5`: What would the value of epsilon be at the end of training when employing an exponential decay strategy with `epsilon_decay` set to 0.990?

      iii) `question3_6`: Based on your answer to (c) ii), briefly explain why a decay strategy based on an `exploration_fraction` parameter may be more generally applicable across different environments than a decay strategy based on a `epsilon_decay` parameter.

Understanding the Loss **[5 Marks]**

This part of the exercise will attempt to further your understanding of the loss function in DQN. Figure 4 provides you with a plot of the DQN loss during training within a single run of CartPole with the x-axis and y-axis corresponding to "timesteps trained" and the DQN loss, respectively.

You can also plot the DQN loss yourself using the provided functionality to collect and plot the DQN loss. Simply set the `"plot_loss"` value within the `CARTPOLE_CONFIG` in `train_dqn.py` to `True` and you should receive a plot as stated at the end of training.

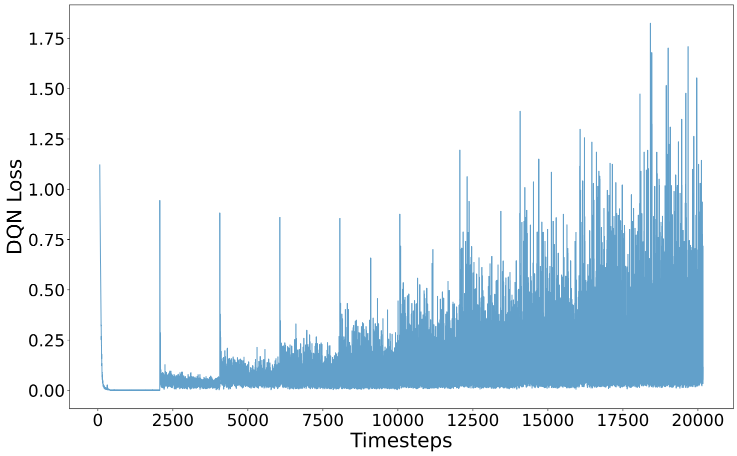


Figure 4: DQN loss during training in the CartPole environment. Generated with the following hyperparameters: learning rate of 0.001, a single hidden layer Q-network with 64 hidden units, batch size of 64, target update frequency of 2000, and a buffer capacity of 1 million experiences.

In machine learning, it is often expected for the value of the loss to drop during training. However, Figure 4 shows that this does not occur in DQN! To demonstrate your understanding, we ask you to answer the following questions in `answer_sheet.py`.

i) `question3_7`: Explain why the loss is not behaving as in typical supervised learning approaches (where we usually see a fairly steady decrease of the loss throughout training).

ii) `question3_8`: Provide an explanation for the spikes which can be observed at regular intervals throughout the training.

Question 4 – Continuous Deep Reinforcement Learning [18 Marks]

Description

So far, we implemented algorithms such as DQN and REINFORCE which define value functions and policies, respectively, for discrete actions, i.e. each action in a state is assigned a specific value or action selection probability. However, in some problems such as control in robotics there might be continuous actions, e.g. representing force which is applied by a motor. To be able to learn policies for such continuous action spaces, we need different RL techniques. The goal of this question is to provide you with experience on (deep) RL algorithms which can be applied in such continuous action spaces. To achieve this aim, you are required to implement the **Deep Deterministic Policy Gradient** (DDPG) [1] algorithm and train it to solve the **Bipedal Walker control task**.

Algorithm

Deep Deterministic Policy Gradient (DDPG) [1] is building on top of Deterministic Policy Gradient (DPG) [3] and extending this RL algorithm for continuous action spaces with function approximators. We highly recommend reading the DDPG paper in addition to lecture materials to familiarise yourself with the algorithm. In contrast to discrete action environments, where an action is a scalar integer, the action in continuous action environments is an N-dimensional vector where, N is the dimension of the action space. Therefore, the Q-network in DDPG outputs a value estimate given a state and action, in contrast to just receiving a state in DQN. Additionally, the action space usually has an upper and a lower bound.

For example, imagine a car with two-dimensional action space, throttle and turn, where throttle takes values in $[-1, 1]$, and turn takes values in $[-45, 45]$. At each time step, the controlled agent should return a two-dimensional action, where the first element represents the throttle and should be in the range of $[-1, 1]$, and the second element represents the turn and therefore should be in the range of $[-45, 45]$.

Please note that an epsilon-greedy policy, which was applied in DQN, cannot be applied in continuous action environments, because the number of possible actions are infinite. Instead, we add Gaussian noise $\mathcal{N}$ to actions chosen by the deterministic policy μ to explore.

$$a = \mu(s) + \eta$$

$$\eta \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\sigma})$$

For this exercise, we consider that the noise is a Gaussian function with mean $\boldsymbol{m} = \mathbf{0}$ and standard deviation $\boldsymbol{\sigma} = 0.1\boldsymbol{I}$ for identity matrix $\boldsymbol{I}$.

Using a batch of N experiences, which are tuples in the form of $< s, a, r, d, s' >$ gathered from the replay buffer, update the parameters of the critic network to minimize the mean squared error:

$$\mathbb{L}_\theta = \frac{1}{N} \sum_{i=1}^{N} \left(r_i + \gamma(1 - d_i) Q\left(s'_i, \mu(s'_i; \phi'); \theta'\right) - Q(s_i, a_i; \theta)\right)^2,$$

where θ and θ' are the parameters of the critic and target critic network, respectively, and ϕ' are the parameters of the target actor network. Using the same batch, implement and minimise the mean squared deterministic policy gradient error to update the parameters of the actor:

$$\mathbb{L}_\phi = \frac{1}{N} \sum_{i=1}^{N} -Q(s_i, \mu(s_i; \phi); \theta)$$

where ϕ are the parameters of the actor's network. The gradient flows through the critic network back to the parameters of the actor. **Please note that during the update of the actor's parameters, the parameters of the critic network should remain fixed and not be updated.**

Domain

In this question, we ask you to train agents in the OpenAI Gym Bipedal Walker and Pendulum environments.

Pendulum is a control task where an agent can apply force to balance a pendulum upwards. The goal is to learn to bring and keep the pendulum in an upward position. The agent observes the angle of the pendulum and chooses an action representing the torque applied to the pendulum. The agent is rewarded for keeping the pendulum in an upward position. A well-tuned implementation should achieve an average return higher than -300.

Bipedal Walker is a control task where the agent (a robot) needs to walk forward while ensuring its balance. The agent receives rewards for moving forward, but gets a penalty for falling and exerting motor torque.

You will only be marked for your agents performance in *Bipedal Walker*! However, we strongly recommend training your algorithm first in the Pendulum task. Due to its simplicity, training will be completed quicker compared to the Bipedal Walker task and therefore allows you to test and ensure the correctness of your implementation. To simplify this process, we already provide you with hyperparameters for both tasks, which should solve this task given a correct implementation.

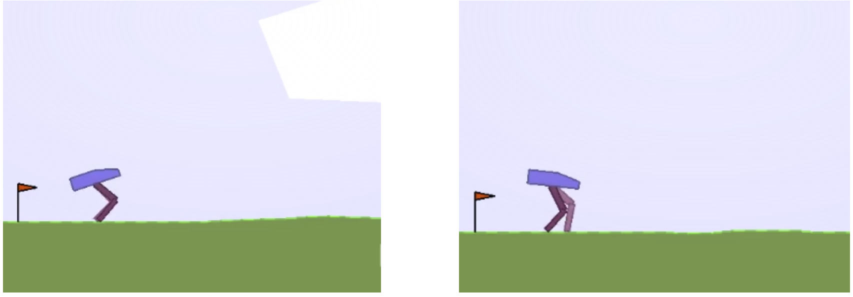


Figure 5: Rendering of two Bipedal Walker environment steps

Tasks

For this exercise, you are required to implement the functions listed below. Besides the correctness of these DDPG functions, we will also mark the performance achieved by your agents with the hyperparameters we provide **only in the Bipedal Walker** environment. **To get full marks, the evaluation average return achieved by your DDPG agent needs to be** ≥ -100**.** See each paragraph below for more details on required functions and respective marks.

Implementation [13 Marks]

Use the code base provided in the directory `exercise4` and implement the following functions. In `agents.py`, you will find the `DDPG` class which you need to complete. For this class, implement the following functions:

- `__init__`, which creates a DDPG agent. Here, you have to initialise the Gaussian noise. Use the imported class from `torch.distributions`, `Normal`, to define a noise variable. During exploration you should call the function `sample()` from the `Normal` instance. Also, you can set any additional hyperparameters and initialise any values for the class you need.

- `act`, which implements the action selection method of DDPG. Aside from the observation, this function also receives a boolean flag as input. When the value of this boolean flag is `True`, agents should follow an exploratory policy using noise as specified above. Otherwise, agents should follow the deterministic policy without any noise. This flag is useful when we interchange between training and evaluation.

  Hint: Remember to clip the action between the upper and lower bound of the action space before returning the action.

- `update`, which receives a batch of experience from the replay buffer. Using a batch of experiences, which are tuples in the form of $< s_t, a_t, r_t, d_t, s_{t+1} >$ gathered from the replay buffer, update the parameters of the critic network to minimize the mean squared error:

  $$\mathbb{L}_\theta = (r + \gamma(1 - d_t)Q(\mu(s_{t+1}; \phi'), s_{t+1}; \theta') - Q(a_t, s_t; \theta))^2,$$

  where θ and θ' are the parameters of the critic and target critic network respectively, and ϕ' are the parameters of the target actor network. Using the same batch implement and minimise the deterministic policy gradient error to update the parameters of the actor:

  $$\mathbb{L}_\phi = \frac{1}{N} \sum_{i=1}^{N} -Q(s_i, \mu(s_i; \phi); \theta)$$

where ϕ are the parameters of the actor's network. The gradient flows through the critic network back to the parameters of the actor. Please note, that during the update of the actor's parameters, the parameters of the critic network should remain fixed and not be updated.

Also, this function is required to update the target critic and actor parameters using soft updates at every update with step size τ.

$$\theta' \leftarrow (1-\tau)\theta' + \tau\theta \qquad\qquad \phi' \leftarrow (1-\tau)\phi' + \tau\phi$$

Hyperparameter Tuning **[5 Marks]**

Besides correctness of the action selection and learning functions, we will also mark the performance of your agents **in the Bipedal Walker environment**. As mentioned in the previous questions, the performance of DRL algorithms is highly dependent on the choices of hyperparameters. For this question, we will only ask you to tune the size of hidden layers of both the critic and policy networks. That said, we won't tell you which size of hidden layers to try, and you have to search yourself. The default values of all hyperparameters are provided in the in the `BIPEDAL_CONFIG` in `train_ddpg.py`, and you can set your own values of `critic_hidden_size` and `policy_hidden_size`. Please keep the other hyperparameters as they are during your fine-tuning in this question. Given a correct implementation and well selected hidden size of the two neural networks, it shall be not hard for your agents to achieve ≥ -100 evaluation returns.

You will also need to provide us with saved [parameters/weights] of the critic and policy neural networks for DDPG in Bipedal Walker so that we can verify the performance[1]. The saved [parameters/weights] of the neural networks shall be named as `'bipedal_q4_latest.pt'` which is specified by the `EX4_BIPEDAL_CONSTANTS` in `constants.py`. Make sure that the performance achieved by your saved parameters (saved at the end of training in `train_ddpg.py`) are reliable by using the `evaluate_ddpg.py` script.

Note: Make sure the other hyperparameters are set to their **default values** in this exercise, which are provided in `EX4_CONSTANTS` in `constants.py`. During our evaluation, we will use the original `constants.py` to overwrite the same file in your submission. Therefore, any change in `constants.py` will be ineffective.

[1] The saved parameters/weights of a model is also known as a "checkpoint".

Question 5 – Fine-tuning the Algorithms [15 Marks]

Description

We mentioned several times in the pervious question descriptions that the selection of hyperparameter values greatly impact the performance of (deep) RL algorithms. In this question, you are required to **implement a hyperparameter tuning method**. The goal of this question is to provide you with experience on fine-tuning the hyperparameters for DRL algorithms. Below, you can find a brief description of the two hyperparameter search methods, and the functions you need to implement. Make sure to carefully read the documentation of these functions to understand their input and required outputs. We will mark your submission based on the **performance** of your learning agents measured by the average evaluation returns (10 marks) as well as **how you select** the hyperparameter values used to train your agents (5 marks), in the Bipedal-Walker environment.

Algorithm

For this question, we use the **DDPG** algorithm introduced in Question 4. Please read the **Algorithm** section of Question 4 for more details about the DDPG algorithm.

Domain

In this question, we also ask you to train agents in the OpenAI Gym Bipedal Walker environment, as in Question 4. For a short description of the environment, please read the **Domain** section of Question 4 again.

Tasks

For this question, you are required to achieve a much higher reward than required in Question 4. Achieving the scores listed in Table 4 will require an extensive search of the hyperparameter space, and therefore we highly recommend you to use/implement a systematic hyperparameter search method. You are free to use any hyperparameter searching technique you see fit, and we **won't** mark its implementation. Instead, we will only mark your submission in the Bipedal-Walker environment based on: i) the performance of your learning agents measured by the average evaluation returns of the model you submit (10 marks); ii) how you select the hyperparameters used to train your agents (5 marks).

To help you establish a rough idea about how to sweep the hyperparameters, we briefly illustrate two common hyperparameter sweeping methods below:

- **grid search** iterates over all combinations of the hyperparameter values. Suppose there are two hyperparameters $a \in \{1, 2\}$ and $b \in \{2, 3\}$, then grid search will iterate over the set $\{1, 2\} \times \{2, 3\} = \{(1, 2), (1, 3), (2, 2), (2, 3)\}$. This method is computationally infeasible if a hyperparameter has infinitely many possible values without discretising the parameter value domains.
- **random search**, as its name suggests, randomly picks up a combination of hyperparameters at each iteration. For different types of hyperparameters, you can specify different types of distributions. For example, for a discrete value, you can specify arbitrary categorical distributions for the sweeper to sample from. If a hyperparameter has infinitely many values, you can then specify a continuous distribution for the sweeper to sample from.
  Hint 1: you may prefer to search some hyperparameters in log space, e.g. learning rate. You may prefer to search the learning rates in a set like $\{10^{-1}, 2 \times 10^{-1}, 10^{-2}, 2 \times 10^{-2}, \dots\}$.
  Hint 2: you may want to work iteratively and start by a coarse sweep over a wide range of values for the hyperparameters, and carry-on with finer sweeps that explore hyperparameters regions close to a well-performing run.

We provide skeleton functions `grid_search` and `random_search` in `util/hparam_sweeping.py` for implementing grid search and random search functions. As per the previous questions, you are recommended to use the provided class `Run` in `util/result_processing.py` to log and process your results. You are also advised to train at least **10 seeds per hyperparameter configuration** for statistical consistency.

You can also implement hyperparameter scheduling within the `schedule_hyperparameters` function of the `DDPG` class in `rl2023/exercise4/agents.py` for a better performance of your agent. You were asked to do this in Section 5.3 for the exploration probability ϵ, you may decide to implement some scheduling for other hyperparameters here.

A difference between hyperparameter sweeping and scheduling is that the value of the hyperparameter might be **changed** by your scheduler **during the training**, whereas they **keep identical** during the training procedure under the hyperparameter sweeping. As you saw in Section 5.3, the hyperparameter scheduling routines may have hyperparameters themselves (for example, the `epsilon_decay` hyperparameter when using ϵ scheduling in DQN).

Note: we **won't** mark the correctness of your hyperparameter sweeping and scheduling implementations. You can use any hyperparameter turning method you'd like, and you **are not required** to implement all `search/scheduling` functions, although we recommend you to do so for better performance. But, you are required to briefly describe your hyperparameter sweeping and scheduling methods to answer the questions listed below.

Hyperparameter Tuning and Performance **[15 Marks]**

Hyperparameter tuning (adjusting hyperparameters in the `config` in `exercise5/train_ddpg.py`) and scheduling (through `schedule_hyperparameters` in the `DDPG` class in `exercise4/agents.py`) will be required to achieve full performance marks. You will need to provide us with the saved [parameters/weights] of the `DDPG` model so that we can verify the performance of your trained agents. The saved [parameters/weights] of the model shall be named as **`'bipedal_q5_latest.pt'`** which is specified by the `EX5_BIPEDAL_CONSTANTS` in `constants.py`. **Make sure that your saved model for this question is different from the one for Question 4, i.e. `'bipedal_q5_latest.pt'` differs from `'bipedal_q4_latest.pt'`. If the two saved models are identical, you will get** 0 **mark for this question.** In the meantime, make sure that the performance achieved by your saved parameters (saved at the end of training in `exercise5/train_ddpg.py`) are reliable by using the `exercise5/evaluate_ddpg.py` script. [10 Marks]

Performance marks	**0/10**	**5/10**	**10/10**
DDPG	< 150	< 280	≥ 280

Table 4: Average (evaluation) returns required for given **performance marks** for DDPG in the Bipedal Walker environment.

In addition to the performance marks, we will also mark your submission based on how you select the hyperparameters to get the best evaluation return. Please provide a short description ($<$ 200 words) about how you did the hyperparameter turning and scheduling to get the best performance by filling the `question5_1` in `answer_sheet.py`. [5 Marks]

Note: make sure the hyperparameters provided in `EX5_BIPEDAL_CONSTANTS` in `constants.py` are set to their **default values**. During our evaluation, we will use the original `constants.py` to overwrite the same file in your submission. Therefore, any change in `constants.py` will be ineffective.

6 Marking

Academic Conduct Please note that any assessed work is subject to University regulations and students are expected to follow any such regulations on academic conduct:
`http://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct`

Correctness Marking As mentioned for most questions, we partly mark your submissions based on the correctness of the implemented functions. For pre-defined functions we ask you to implement, including most functions stated across all questions, we use unit testing scripts. In these scripts, we pass the same input into both your and our reference implementation and assign you marks according to whether the output of your function matches the expected output provided by our reference implementation. For functions which are evaluated for correctness, you must read the documentation to ensure that your implementation follows the expected format. **Only change files and functions specified for Questions 1–5 and ensure that the implementations match the specifications provided in the instructions! Any deviations might cause automated marking to fail which could lead to a deduction in marks. This includes optimisations and implementation tricks which could improve performance!**

Performance Marking For performance evaluation in Questions 4 and 5, we will evaluate your models against the default training scripts of the code base to ensure that your agent solves the environments we used for training measured by the achieved average returns, and we will only import the agents and their respective configuration dictionaries from the files you submitted. Therefore, **make sure that the hyperparameters of your algorithms have been appropriately tuned and are set in the configurations of the respective training scripts to achieve the required thresholds**. Also, for Questions 4 and 5, **make sure to provide saved model parameters for DDPG trained on Bipedal Walker** as instructed in the respective Questions. In particular, make sure to save your model for Question 4 as `bipedal_q4_latest.pt` in the `exercise4` folder and your model for Question 5 as `bipedal_q5_latest.pt` in the `exercise5` folder.

7 Submission Instructions

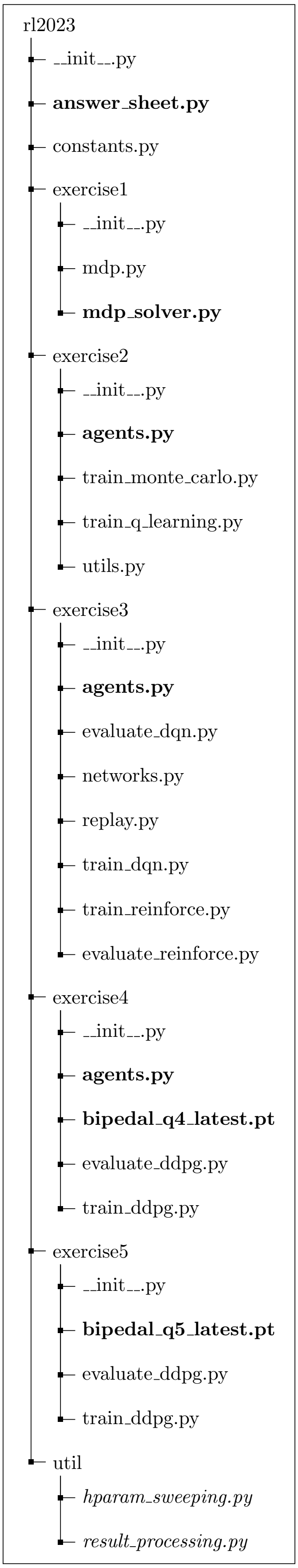
Before you submit your implementations, make sure that you have organised your files according to the structure indicated in Figure 6.

Finally, compress the `rl2023` folder into a **zip** file and submit the compressed file through Learn. In your Learn page, go to the **Assessment** panel and find the **Coursework** page. For general guidance on submitting files through Learn, you can find further information through the blog post linked below:
https://blogs.ed.ac.uk/ilts/2019/09/27/assignment-hand-ins-for-learn-guidance-for-students/.

You may also refer to the link below for instructions specific to the CodeGrade submission platform https://docs.codegra.de/guides/use-codegrade-as-a-student.html.

Late Submissions All submissions are timestamped automatically and **we will mark the latest submission**. If you submit your work after the deadline a late penalty will be applied to this submission unless you have received an approved extension. Please be aware that marking for late submissions may be delayed and marks may not be returned within the same timeframe as for on-time submissions.

For additional information or any queries regarding late penalties and extension requests, follow the instructions stated on the School web page below:
web.inf.ed.ac.uk/infweb/student-services/ito/admin/coursework-projects/late-coursework-extension-requests



- rl2023
  - __init__.py
  - **answer_sheet.py**
  - constants.py
  - exercise1
    - __init__.py
    - mdp.py
    - **mdp_solver.py**
  - exercise2
    - __init__.py
    - **agents.py**
    - train_monte_carlo.py
    - train_q_learning.py
    - utils.py
  - exercise3
    - __init__.py
    - **agents.py**
    - evaluate_dqn.py
    - networks.py
    - replay.py
    - train_dqn.py
    - train_reinforce.py
    - evaluate_reinforce.py
  - exercise4
    - __init__.py
    - **agents.py**
    - **bipedal_q4_latest.pt**
    - evaluate_ddpg.py
    - train_ddpg.py
  - exercise5
    - __init__.py
    - **bipedal_q5_latest.pt**
    - evaluate_ddpg.py
    - train_ddpg.py
  - util
    - *hparam_sweeping.py*
    - *result_processing.py*

Figure 6: Required folder structure for submission. Files which need to be modified or created for this coursework are marked in **bold**. Files which may optionally be modified to facilitate completion of the coursework are *italicised*.

References

[1] Timothy P Lillicrap et al. “Continuous control with deep reinforcement learning”. In: *International Conference on Learning Representations* (2015).

[2] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), pp. 529–533.

[3] David Silver et al. “Deterministic policy gradient algorithms”. In: 2014.

[4] Richard S Sutton et al. “Policy gradient methods for reinforcement learning with function approximation”. In: *Advances in Neural Information Processing Systems*. 2000, pp. 1057–1063.