Reinforcement Learning 2024/2025 Coursework*

Submission deadline: 12:00 noon on 28 March, 2025

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" by Sutton and Barto, 2020).

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 and 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 via the DRPS page of the RL course.

Note: Please keep in mind that Piazza questions and lab sessions are public for discussions. Given that this coursework is individual work and will be 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 concepts and algorithms (disconnected from their implementation). 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. 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.

*This version is no longer preliminary. If it still needs to be corrected in any way, then this will be announced via Learn. Differences to the first version of the CW description are shown in red.

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 Anaconda.**
   We recommend using Anaconda to manage your Python installation and required packages for this course. First navigate to the Anaconda download page and follow the installation instructions related to your operating system. Anaconda (sometimes shortened to "conda") supports Linux, MacOS, and Windows.

2. **Creating a conda environment.**
   Creating an environment with conda is relatively easy. Within a terminal session, use the conda create command, name your environment with the -n flag, and choose your Python version with `python=v.r` as in the following example:

   ```
   conda create -n rl_course python=3.7
   ```

   Then, enter into your new conda environment with the `conda activate` command as in the following example:

   ```
   conda activate rl_course
   ```

3. **Download the code base to get started**.
   Now you can download the code base as the zip file `rl2025-coursework.zip` from Learn. After unzipping, navigate into the coursework folder with `cd rl2025-coursework`. Within the directory, you should see a file called `setup.py`. This file contains a list of the libraries required to complete your coursework under the name `install_requires`. To install these packages within your conda environment, execute the following command:

   ```
   pip install -e .
   ```

For detailed instructions on Python's library manager `pip`, see the official Python guide.

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 `rl2025` 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 Implement the following DP algorithms for MDPs	**[20 Marks]**
• Value Iteration	[10 Marks]
• Policy Iteration	[10 Marks]
Question 2: Tabular Reinforcement Learning	**[30 Marks]**
• Implement ε-greedy action selection	[3 Marks]
• Implement the following RL algorithms	
– Q-Learning	[9 Marks]
– On-policy every-visit Monte Carlo	[9 Marks]
• Analyse performance of different hyperparameters in FrozenLake8x8-v1	[9 Marks]
Question 3: Deep Reinforcement Learning	**[30 Marks]**
• Implement the following Deep RL algorithms	
– Deep Q-Networks	[10 Marks]
– Equivalent discrete problem	[4 Marks]
• Performance analysis and comparison	[2 Marks]
• DQN performance analysis	
– Implement ε-scheduling strategies	[3 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	**[20 Marks]**
• Implement DDPG for continuous RL	[10 Marks]
• Tune the specified hyperparameters to solve Racetrack	[10 Marks]
Question 5: Extensions of the Algorithms	**[Bonus Marks]**

Table 1: Mark breakage for this assignment.

5 Questions

Question 1 – Dynamic Programming [20 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 $\gamma = 0.85$.

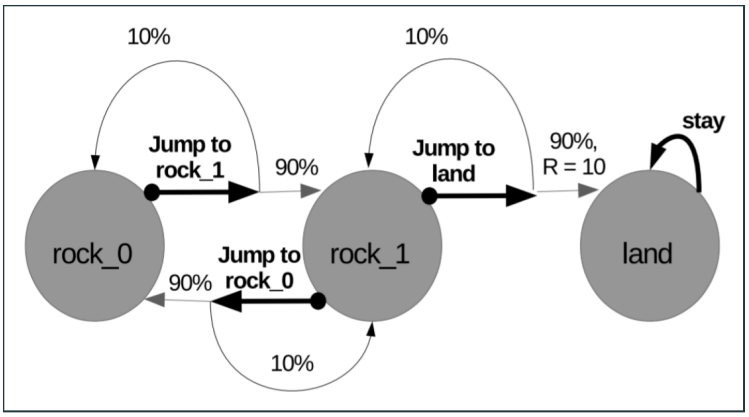

Figure 1: Frog on a Rock example MDP for Exercise 1.

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** [10 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 (here as a table).
   - `_calc_policy`, which must return the greedy deterministic policy given the calculated value function. 2

2. **Policy Iteration** [10 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 [30 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 every-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 FrozenLake8x8-v1 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. Every-visit Monte Carlo (MC): You can find more details including pseudocode for on-policy every-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 Gymnasium FrozenLake8x8-v1 environment. This environment is a simple task where the goal of the agent is to navigate across a frozen lake without falling into any holes in the ice in a grid world.

Figure 2: Rendering of the FrozenLake8x8-v1 environment. The left figure refers to the deterministic and the right figure to the "slippery" case.

The episode terminates once the agent reaches the goal location, the agent falls in a hole or at a maximum episode length. The agent will be given a reward of +1 for successfully reaching the goal, and a reward of 0 otherwise. The frozen lake can be slippery (controlled by the `is_slippery` flag which is `True` by default. To compare to the non-slippery case use: `env = gym.make(CONFIG["env"], is_slippery=False)`), in this the agent will move in the intended direction with probability $\frac{1}{3}$, otherwise it will move perpendicular to the intended direction (with equal probability of $\frac{1}{3}$ in both directions). Hence, the task consists of learning to navigate the slippery grid world to reach the goal location without falling into a hole. A good hyperparameter scheduling for both algorithms should enable the agent to solve the FrozenLake8x8-v1 environment. **We consider the environment to**

be solved when the agent can consistently achieve an average return of ≥0.6. for the slippery version, while in the deteministic case better results are easily achievable.

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 FrozenLake8x8-v1 environment. See each paragraph below for more details on required functions and respective marks.

Implementation **[23 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**
   In the Agent class, implement the following function: [3 Marks]
   - `act`, where you must implement the ε-greedy exploration policy used by the QL and MC algorithms.
2. **Q-Learning** [10 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 every-visit Monte Carlo** [10 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 every-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 FrozenLake8x8-v1 in `train_q_learning.py` and train `monte_carlo.py` respectively. These execute training and evaluation using your implemented agents.

Hyperparameters and Performance [7 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 2, 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.

We recommend running at **least 10 seeds per hyperparameter configuration** for statistical consistency.

Note that the **best evaluation return of a correct implementation will be** ≥ 0.6 with one of the hyperparameter profiles provided in Table 2and correct implementations, for both algorithms.

Algorithm	α	ε	γ	Algorithm	ε	γ
Q-Learning	0.05	0.9	0.99	Every-visit Monte Carlo	0.9	0.99
	0.05	0.9	0.8		0.9	0.8

Table 2: The given **hyperparameter profiles** for QL and MC in the FrozenLake8×8-v1 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 Mark]

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

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

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

v) `question2_5` provide a short explanation ($\lessapprox$ 100 words) on the differences between the non-slippery and the slippery varian of the problem. [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 a Deep Reinforcement Learning algorithm: DQN [2] 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. **Comparison to a discrete solution of the problem**
   The problem(s) considered here are not diffcult, and can in fact be solved also with a tabular reinforcement learning algorithms as considered in the previous question. You are expected to use the algorithm that you have conveniently already studied in the previous questions (or another tabular algorithm) and to compare its performance with the DQN network considered above. You will describe the result of this comparison in the `answer_sheet.py`, see below for details.

Domains

In this question, we train agents on the Gymnasium MountainCar-v0 environment. In MountainCar, the agent controls a car that spawns in a random location at the bottom of a valley. The agent can either accelerate the car left or right. Rewards in MountainCar are based on how close the car is to the goal flag (see right side of Figure 3).

You may find it useful to work also with the CartPole example for some of the tasks, but marks will be available here only for your results on the MountainCar problem.

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 representation and hyperparameters for the discrete version and for the DQN agent in the MountainCar environment. To simplify the hyperparameter search, we will provide you with a range of hyperparameter profiles to pick from.

Implementation [14 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.

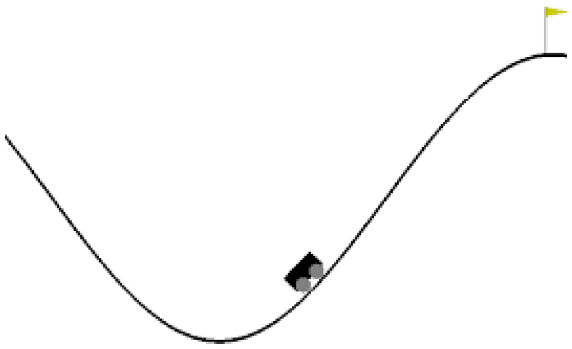

Figure 3: Rendering MountainCar (right) environment

1. DQN [10 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 $\langle s, a, r, d, s\prime \rangle$ gathered from the replay buffer, update the parameters of the value network to minimize the mean squared error:

$$\mathbb{L}_\theta = \frac{1}{N} \sum_{i=1}^{N} \Big(r + \gamma \left(1 - d_t\right) \max_a \mathcal{Q}\left(a | s_i; \phi', s_{t+1}; \theta'\right) - \mathcal{Q}\left(a_t, s_t; \theta\right) \Big)^2$$

  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. Discrete version [4 Marks]

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

- `__init__`, which creates the learning agent. You can set additional hyperparameters and values required for training the agent here.
- `act`, which implements the action selection based on a policy function.
- `update`, which updates the policy based on the sequence of experience $\langle s_t, a_t, r_t, d_t, s_{t+1} \rangle_{t=1}^{T}$ received by the agent during an episode. You must then complete the implementation of the update rule.

Algorithm	`learning_rate`
Discrete algorithm	0.02
	0.002
	0.0002

Table 3: Provided hyperparameters for tuning the learning rate for the discrete algorithm.

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.

Testing

To test your implementation, we provide you with two scripts which execute your DQN implementation. You can find the scripts inside `train_dqn.py` to train DQN, respectively. Inside these scripts, we provide you with configurations that enable you to train the DQN in the MountainCar environment. To better understand how your implemented functions are used in the training process, read the code and documentation provided in these scripts.

We also provide a configuration for DQN for the well-known CartPole problem to allow you to more easily test your DQN implementation since the CartPole environment is easier and quicker to train than MountainCar, but **only MountainCar should be used to complete the questions on DQN performance**. For a correct implementation, the training process requires less than 2 minutes to train DQN in CartPole, and less than 30 minutes to train DQN in MountainCar.

Hyperparameter tuning [11 Marks]

Besides correctness of the aforementioned algorithms, we also ask you to tune different hyperparameters of your DQN 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 MOUNTAINCAR_CONFIG in `train_dqn.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.

1. [2 Marks] Here, we simply ask you to tune the learning rate in the discretised version of state space representing the MountainCar environment. You can find the possible values to pick from for the learning rate in Table 2. 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 an average score of at least 180 for the best performing profile.

2. DQN [9 Marks]

We ask you to implement different epsilon scheduling strategies for DQN and tune them in the MountainCar environment.

ε	exploration fraction	ε decay strategy	`epsilon_decay`
Linear	0.99	Exponential	1.0
	0.75		0.5
	0.01		0.00001

Table 4: Provided hyperparameters for tuning epsilon scheduling for DQN in the MountainCar 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 [$\varepsilon_{\text{start}}$, ε_{min}, `exploration_fraction`]) decays ε linearly from some starting value εstart to a minimum value ε_{min}. After reaching ε_{min} , ε remains constant. The parameter ε should reach ε_{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 [$\varepsilon_{\text{start}}$, ε_{min}, `epsilon_decay`]) decays ε exponentially such that $\varepsilon_{t+1} \leftarrow r^{t/t_{\text{max}}}\varepsilon_t$, where r is the decay rate set by epsilon decay. The parameter ε decays from some starting value $\varepsilon_{\text{start}}$ to a minimum value ε_{min} . After reaching ε_{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 MountainCar (these are also reported in Table 4). Try out the different exploration fraction values in MOUNTAINCAR_HPARAMS_LINEAR_DECAY and the epsilon decay values in MOUNTAINCAR_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 an average score of at least -125 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 0.95?

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. As an example consider Figure 4 which provides you with a plot of the DQN loss during training within a single run of MountainCar with the x-axis and y-axis corresponding to "timesteps trained" and the DQN loss, respectively.

You may prefer to consider instead the data that you have obtained yourself for the DQN loss on the MountainCar problem, but if you encounter any problems with this you can solve this question also based on Figure 4. If you are working here with your own data, use the provided functionality to collect and plot the DQN loss.

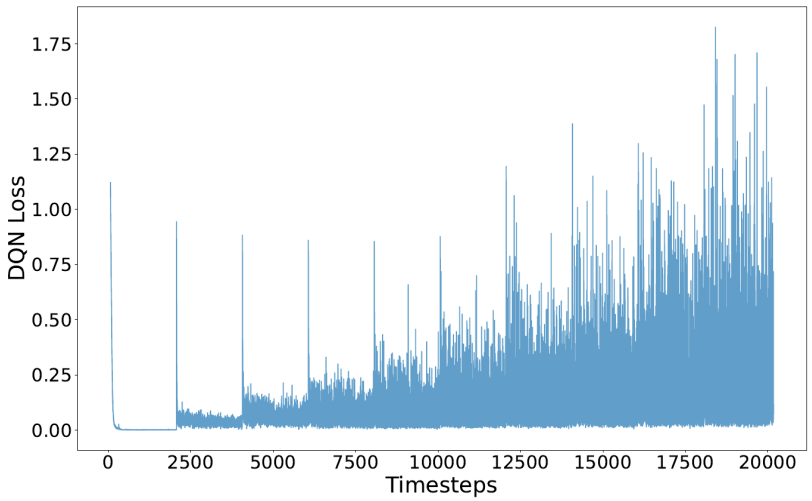

Figure 4: DQN loss during training in the CartPole environment. Generated with the following hyperparameters: learning rate of 0.001, a single hidden layer Q-network with 64 hidden units, batch size of 64, target update frequency of 2000, and a buffer capacity of one million experiences.

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 [20 Marks]

Description

So far, we implemented algorithms such as DQN 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 Racetrack 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 ε-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.

$$a = \mu(s) + \eta$$
$$\eta \sim \mathcal{N}(m, \sigma)$$

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

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

$$\mathbf{L}_\theta = \frac{1}{N} \sum_{i=1}^{N} \left(r + \gamma \left(1 - d_t\right) \mathcal{Q}\left(\mu\left(s_{t+1}; \phi'\right), s_{t+1}; \theta'\right) - \mathcal{Q}\left(a_t, s_t; \theta\right)\right)^2$$

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:

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

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.

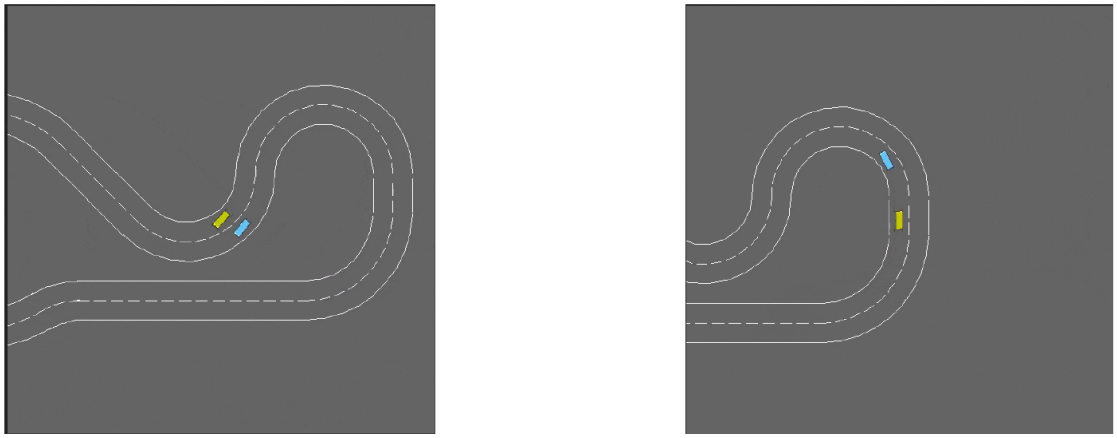

Figure 5: Rendering of two Racetrack environment steps.

Domain

In this question, we ask you to train agents in the Racetrack task from the HighwayEnv environment suite. In Racetrack, the agent steers a vehicle (yellow) around a racetrack while avoiding a collision with a computer-controlled vehicle (blue). The agent receives a positive reward for staying within the racetrack's lanes, and episodes terminate if a collision occurs.

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 DDPG agent in the Racetrack. See each paragraph below for more details on required functions, performance thresholds, and respective marks.

Implementation [10 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 $\langle s_t, a_t, r_t, d_t, s_{t+1} \rangle$ gathered from the replay buffer,

Performance marks	0/10	5/10	10/10
DDPG	< 300	< 500	≥ 500

Table 5: Average (evaluation) returns required for given performance marks for DDPG in the Racetrack environment.

update the parameters of the critic network to minimize the mean squared error:

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

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 deterministic policy gradient error to update the parameters of the actor:

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

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 **[10 Marks]**

Besides correctness of the action selection and learning functions, we will also mark the performance of your agents in the Racetrack 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 `RACETRACK_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. You will also need to provide us with saved parameters/weights of the critic and policy neural networks for DDPG in Racetrack so that we can verify the performance1 . Your mark will depend on the performance of your saved agent. For marking thresholds, see Table 5. The saved parameters/weights[1] of the neural networks should be named as `racetrack_latest.pt` which is specified by the `EX4_RACETRACK_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*: Tricks and improvements **[Bonus Marks]**

We mentioned several times in the previous questions that you are expected to follow prescribed functions and that you should avoid any departures. This may help you to safe time and effort, but in some cases you may have felt the the given template were inconvenient, not optimal, or not particularly interesting. While any suggestions for improvements are welcome via the feedback channels, we can give you a few bonus marks, if you can demonstrate that your work beyond what is prescribed in the template did indeed lead to improvements or to any interesting and interpretable effect. We don't want to limit your creativity, so no further specification will be given, except that

- the work reported on this question should be related to one of the previous tasks.
- you should aim at submitting solutions that are qualitatively different from the solutions of other students.
- you should focus on an extension of one of the above questions, and if you add here more than one piece of work, the marker will consider only one of them.
- up to 5 bonus marks will be given for any reasonable effort; up to 5 additional bonus marks for a creative approach; up to 5 additional bonus marks for an impressive result; and up to 5 additional bonus marks for a result of near publishable quality.
- bonus marks are not subject to negotiation and feedback on the bonus part may not be comprehensive.

For this quiestion, you have the option to enter a short description into the `answer_sheet.py` and/or to add a `pdf` file to your zip file that you are submitting.

Note: It is not necessary in any way to work on this bonus question, as you can get full marks already from Questions 1 - 4.

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:

Correctness Marking: As mentioned for most questions, we partly mark your submissions based on the correctness of the implemented functions. For predefined 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 1–4 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, 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, make sure to provide saved model parameters for DDPG trained on Racetrack as instructed in the respective Questions. In particular, make sure to save your model for Question 4 as `racetrack_latest.pt` in the exercise4 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 `rl2025` 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 this blog post.

You may also refer to the link below for instructions specific to the CodeGrade submission platform.

Late Submissions: All submissions are timestamped automatically and **we will mark the latest regular 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 time frame as for on-time submissions.

For additional information or any queries regarding late penalties and extension requests, follow the instructions stated on this page.

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". *Proceedings of Machine Learning Research* In: (2014), pp. 387-395.

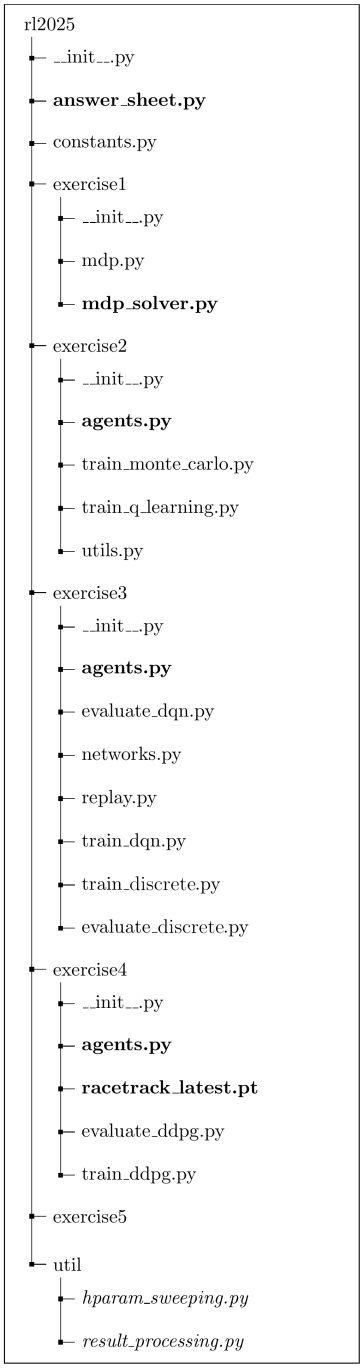

Figure 6: Required folder structure for submission. Files which need to be modified or created for this coursework are marked in bold. Files which may optionally be modified to facilitate completion of the coursework are italicised.