Reinforcement Learning: Project Report

06-04-2025

By Ella Kemperman, student number *s1119890*, Ioana-Anastasia Orasanu, student number *s1112418*, Rostyslav Redchyts, student number *s1113227*, Mila van Rijsbergen student number *s1128515*.

1. Introduction

Reinforcement Learning (RL) has become crucial in Machine Learning by contributing to the AI revolution through approaches such as RLHF (Reinforcement Learning with Human Feedback). In this report, some RL algorithms are discussed in order to get insight into how they differ.

The environment in which the algorithms were tested was specifically designed for this project. Thus, it involves an agent, also referred to as 'the cat', navigating a 2-dimensional grid while chasing a target, 'the mouse'. The environment (see Figure 1) is discrete and both the agent and the target have four possible movements. Two obstacles were added to the environment: a "tree" cell which the cat can traverse but the mouse cannot, and walls in between certain cells, impermeable for the cat, permeable for the mouse. The agent receives a penalty of -1 for each step that resulted in an empty cell, a reduced penalty of -0.1 for landing on a tree, and a reward of +20 for catching the mouse (regardless of the cell type).

The project consists of comparing two algorithms from Dynamic Programming (DP) (see Section 2); a Monte-Carlo (MC) method (see Section 3), and two Temporal Difference (TD) algorithms (see Section 4). Additionally, Deep Q-learning has been implemented (see Appendix A). Finally, the performance of the learned policies was compared as they were being refined during learning, in addition to the policies' final returns. This is done in Section 5.

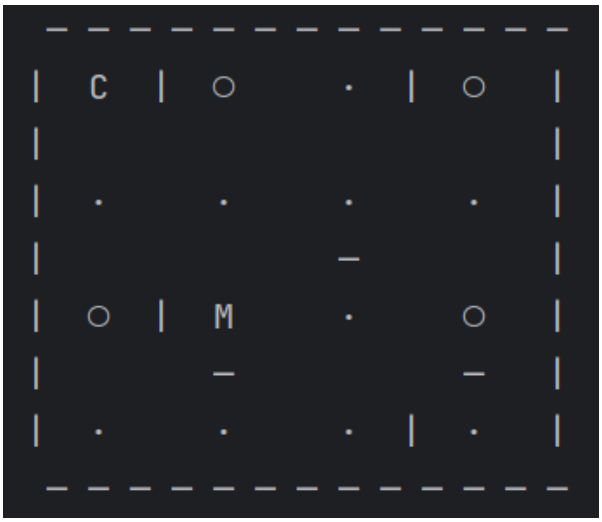


Figure 1: A visual representation of the grid environment. The '|'s indicate walls, the 'o's indicate trees, 'C' is the agent, and 'M' is the mouse.

2. Dynamic Programming

The idea behind the first family of algorithms discussed — Dynamic Programming (DP) — lies in the approach of dividing a large problem into smaller sub-problems, such as improving a current policy, based on the current estimate of the value function. Solving these smaller subproblems is guaranteed to lead to an optimal solution to the main problem.

For this project, two DP algorithms were implemented and tested: Policy Iteration (Section 2.1) and Value Iteration (Section 2.2).

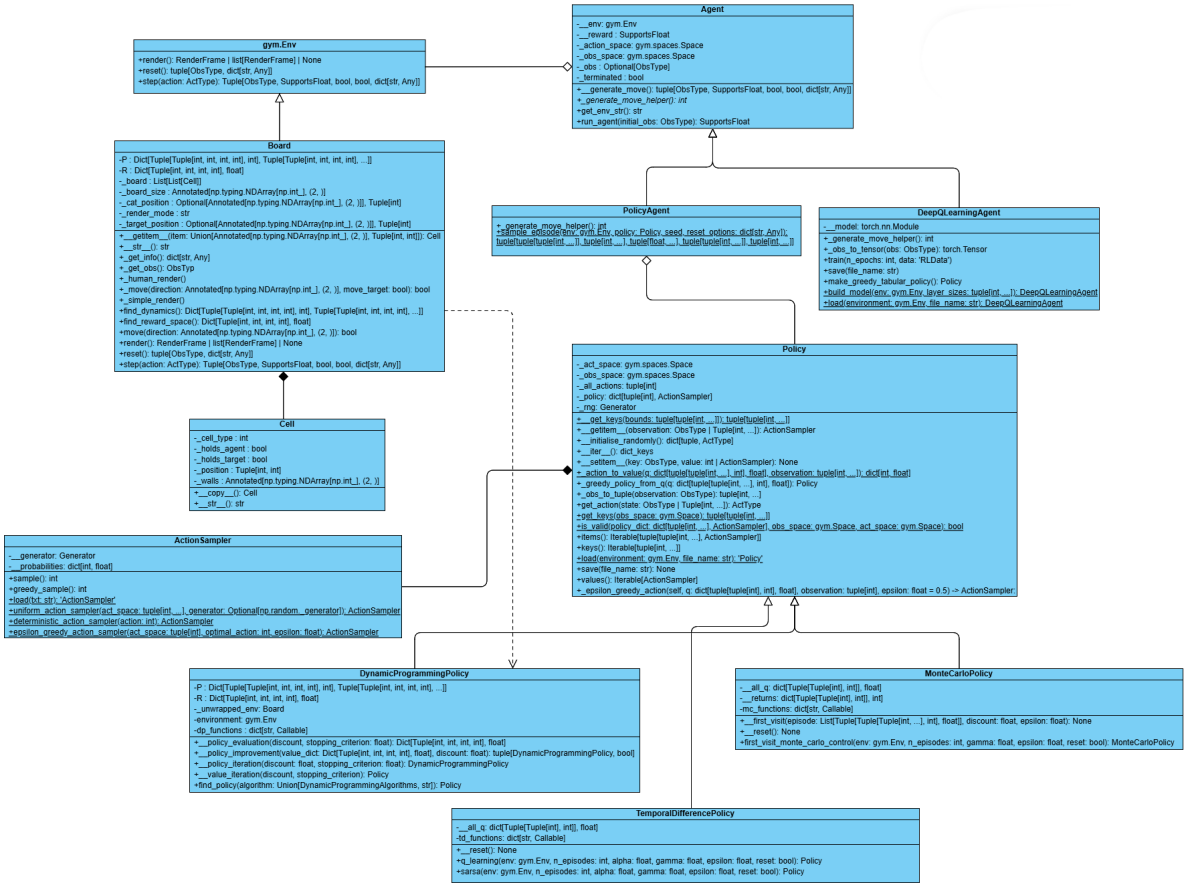


Figure 2: Class diagram of the implementation, made with Visual Paradigm. Note that the alias `gym` was used for the gymnasium package and the alias `np` was used for the numpy package.

2.1. Policy Iteration

The main idea of Policy Iteration is to find an optimal solution through planning. Specifically, the process consists of iteratively evaluating the current policy (Policy Evaluation) and improving this policy, based on the updated estimate of the value function (Policy Improvement), until convergence. Refer to Figure 3 for a full pseudocode.

The state values are updated using the Bellman expectation equation, estimating future rewards by considering the rewards of the successor states weighted by the probability of reaching that state. Convergence, although possibly expensive, is guaranteed. To avoid increased computation costs, a delta value is defined as the maximal value difference through all states after the current iteration. Once it is low enough, the algorithm is terminated.

The policy is refined by selecting the action that maximises expected rewards for each state, where the update is based on the value function.

Implementation

Implemented in the `DynamicProgrammingPolicy` class, the algorithm stores policies and state values as a dictionary. The policy iteration and policy improvement parts were separated into two methods, the former calling policy evaluation and then policy improvement, until no more changes to the policy are made.

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$; $V(\textit{terminal}) \doteq 0$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))[r + \gamma V(s')]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Figure 3: Policy Iteration pseudocode (Sutton & Barto, 2018, p. 80)

2.2. Value Iteration

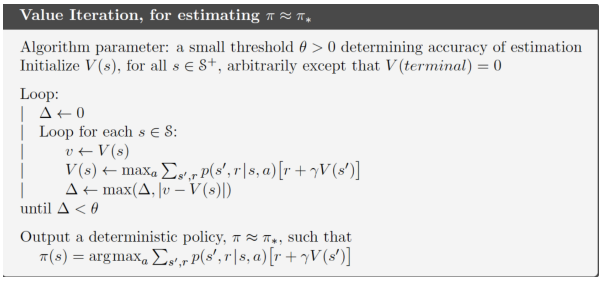
Value iteration is based on the principle of optimality, which states that a policy π achieves the optimal value from state s, $v_\pi(s) = v^*(s)$ if and only if from any state s' reachable from s, π achieves the optimal value $v_\pi(s') = v^*(s')$. The algorithm outputs a new approximately optimal policy.

The process starts by selecting a random value function, over which is iterated using the Bellman equations as an update rule, updating the values for all states. Once the optimal value is found, it takes the actions such that the maximum in the Bellman optimality equation is satisfied. For the pseudocode, see Figure 4.

Implementation

Implemented in the `DynamicProgrammingPolicy` class, as seen in Figure 2, this class inherits from `Policy`. A dictionary was used, with each state mapping to the corresponding value. The dictionary gets updated by selecting the value where the action is maximal based on the Bellman equation.

Convergence is determined by actively checking the value of Δ at all states. Once it becomes small enough, the process ends. Since the action taken in each state was already updated during the running of the algorithm, the policy is now approximately optimal.



Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

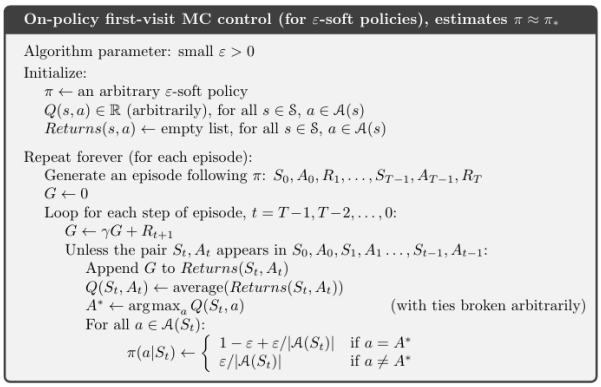
Loop:
  $\Delta \leftarrow 0$
  Loop for each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$

Figure 4: Policy Iteration pseudocode (Sutton & Barto, 2018, p. 80)

3. Monte-Carlo

In RL, Monte-Carlo (MC) methods update the model using the data obtained from running simulations of the environment, thus avoiding the need for the full knowledge of the environment. MC updates its values differently, depending on the algorithm being used. In this report, the MC First-Visit on-policy algorithm was used. The pseudocode is provided in Figure 5.



On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$
Initialize:
  $\pi \leftarrow$ an arbitrary ε-soft policy
  $Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
  $Returns(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):
  Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      Append G to $Returns(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow$ average$(Returns(S_t, A_t))$
      $A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)
      For all $a \in \mathcal{A}(S_t)$:
$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

Figure 5: MC on-policy first-visit pseudocode (Sutton & Barto, 2018, p. 101)

3.1. First-Visit

With First-Visit, after each episode, the program takes the learned information and calculates the returns at each time step. For the first time a state-action pair is discovered in the episode, the return from that time point is added to a list of returns, which is averaged to obtain the new state-action value function for that state-action pair.

3.2. On-Policy

After each episode, the program looks at all the acquired state-action values, determining the best action for each state based on the highest value, for each state-action pair. The policy is updated for a certain state by acting ϵ-greedy on the value function for that state. The probability for each action given ϵ is then given by Equation 1.

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \varepsilon/|A(s)| & \text{if} \quad a = \arg\max_a Q(s, a) \\ \varepsilon/|A(s)| & \text{if} \quad a \neq \arg\max_a Q(s, a) \end{cases} \tag{1}$$

Thus a better policy is guaranteed, but not always an optimal one.

3.3. Implementation

Implemented in the `MonteCarloPolicy` class (see Figure 2), it first samples an episode using itself as a policy, afterwards updating its internal Q function, a dictionary mapping state-action to a reward, based on the episode returns. The policy is updated by creating an ϵ-greedy `ActionSampler` on the optimal action. Note that ϵ does not decay, so the policy will always stay ϵ-soft.

4. Temporal Difference

Temporal Difference (TD) Learning algorithms are another set of algorithms that learn the environment from experience rather than by computing the optimal policy based on a model. TD Learning achieves this by learning the optimal state-action value function $q^*(s, a)$ by sampling many episodes. Although similar to MC methods, TD learning does not use the full return computed at the end of the episode; rather it uses bootstrapping to allow it to update the $Q(s, a)$ after each step taken in the episode. This enables it to be used in environments without terminal states. Compared to Monte-Carlo, TD algorithms generally have higher bias but lower variance.

Two main variants of TD learning are discussed in this report: SARSA, an on-policy control algorithm (Section 4.1), and Q-learning, an off-policy control algorithm (Section 4.2).

4.1. SARSA

SARSA is an on-policy TD learning control algorithm. SARSA is short for 'State, Action, Reward, State, Action', which refers to how the state-action value function is updated at each time step. The goal of the algorithm is to approximate $Q(s, a) \approx q^*(s, a)$. This is done by sampling many episodes, and for each step of the episode, updating the value function using the following update rule:

$$Q(S_t, A_t) := Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)) \tag{2}$$

Note that α is the learning rate, γ is the discount factor, S_t the state at time t, and A_t the action at time t. The next action (A_{t+1}) is generated by acting ϵ-greedy on Q for the given state-action pair.

The algorithm first initialises all state-action pairs, $Q(s, a) = 0$. It will then repeat the following over a specified number of episodes. It first picks an action taken from the starting state of the episode, after which it will first take the determined action. A new state S_{t+1}, uses this new state together by acting ϵ-greedy on the value function to get a new action A_{t+1}. Afterwards, Equation 2 is used to update the state-action value function. This is repeated until the episode is finished.

Compared to Q-Learning, SARSA generally chooses a safer path during training, which makes it more likely to converge to an optimal policy while requiring more steps than Q-Learning.

Implementation

Implemented as a method of the `TemporalDifferencePolicy` class (see Figure 2) it learns over n episodes. Each episode decides its actions ϵ-greedy, updating its state-action value function Q along the way, which is represented by a dictionary mapping state-action pairs to their value. When finished, it computes a new policy by acting greedily on the value function. This policy is then returned. The policy derived from Q is ϵ-greedy. However, this implementation acts greedily in the final policy greedily, since it is assumed to have arrived at an approximately optimal policy.

4.2. Q-Learning

Q-learning is an off-policy algorithm, meaning it learns the value of the optimal policy independently of the action values (**poole-2017**). Although similar to value iteration, q-learning aims to estimate the optimal state-action value function (Q). Convergence is also guaranteed for sufficient experience. The goal of Q-learning is to determine the optimal action based on the current state. The update rule of Q-learning is as follows:

$$Q(S, A) := Q(S, A) + \alpha(R + \gamma \max_a Q(S', A') - Q(S, A)) \tag{3}$$

The algorithm works by initializing a Q-table, which stores the action-value sets. The process continues by selecting and performing an action, after which the reward is measured and the Q-table gets updated. The process ends when a terminal state s is reached.

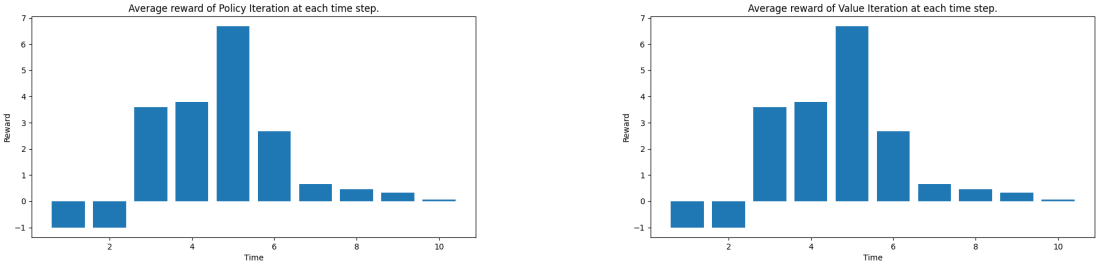
Of great importance is that Q-learning has maximisation bias. A solution to this is Double Q-learning, which uses two Q-functions but does not get updated at the same time.

Implementation

Implemented also as a method of the `TemporalDifferencePolicy` class (see Figure 2) Q-Learning is extremely similar to that of SARSA. The main difference in the implementation is that the action chosen to update the Q dictionary is a greedy action on Q, instead of the actual action taken. Other than that, there are no differences from SARSA.

5. Comparison and Discussion

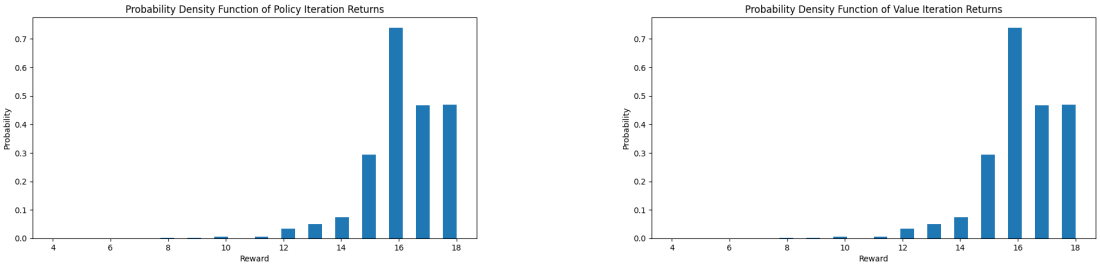
From the plots below, the results derived from Policy Iteration and Value Iteration are strikingly similar. This indicates that both algorithms likely converged to an optimal policy. Both policies seem to mostly reach the terminal state after taking 3-6 steps with the highest average reward being at the fifth time stamp. This indicates that most episodes end quickly, suggesting a great performance of the determined policy.



(a) Policy Iteration **(b)** Value Iteration

Figure 6: Average reward at each time step for both Policy Iteration and Value Iteration.

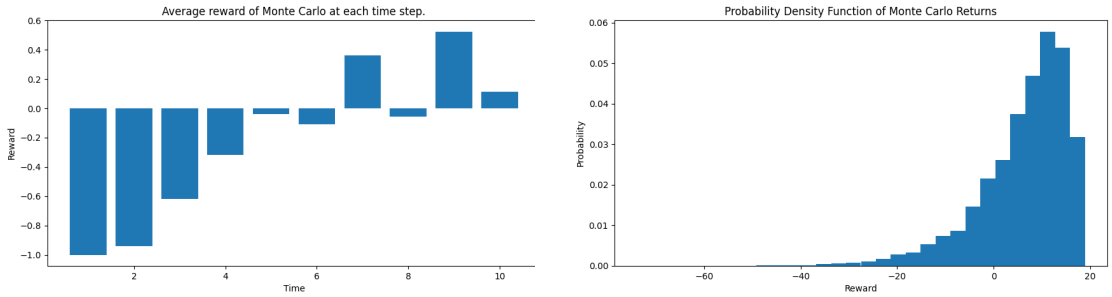
This interpretation is further supported by the plots of the probability density functions of the returns derived from both policies. The plots are once again (almost) identical and indicate that the highest reward likelihood for these policies is around 16, aligning with the 3-6 steps with the rewards of -1 from before.



(a) Policy Iteration **(b)** Value Iteration

Figure 7: Return probability distributions for Policy Iteration and Value Iteration.

The MC algorithm, on the other hand, failed to arrive at an optimal policy over 100 steps (see Figure 10). The plots from Figure 8 indicate a lower average reward for all time steps compared to the DP algorithms from earlier. Furthermore, the probability density function suggests that the policy, derived from First Visit Epsilon Greedy MC, results in a much higher likelihood of returning a negative reward than any of the DP policies. This can be explained by the fact that DP algorithms have a full context of the environment and use planning internally, while MC needs to rely on the limited information obtained through interaction with the environment.

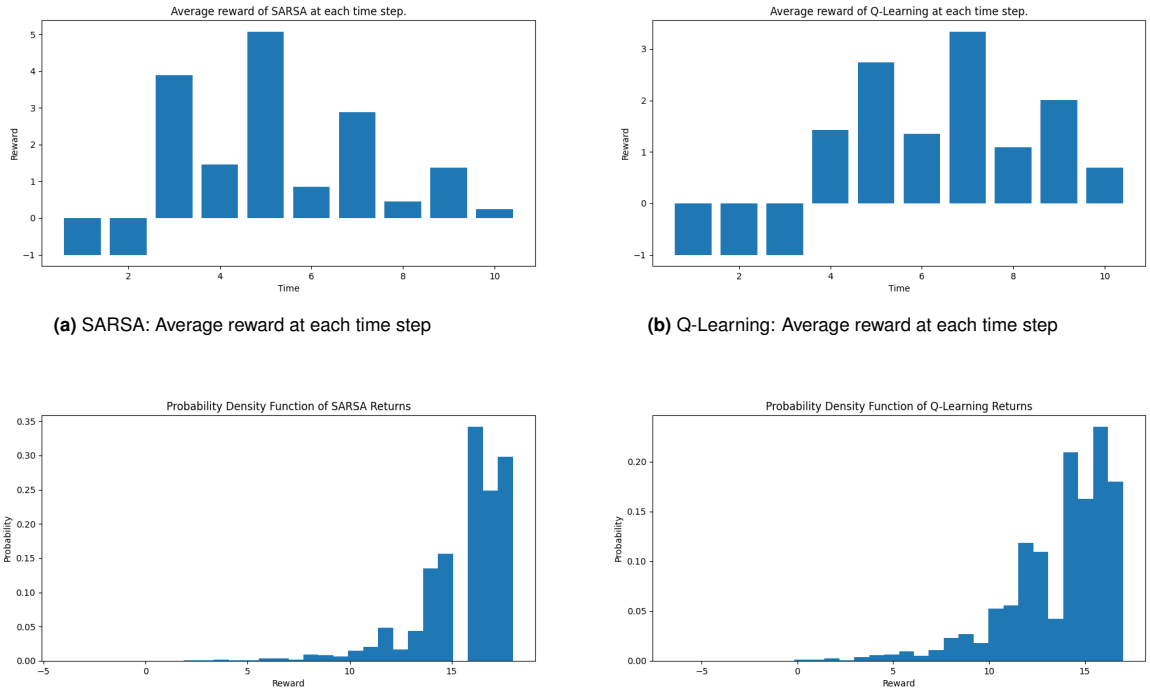


(a) Average reward at each time step

(b) Return probability distributions

Figure 8: Average reward over time and the Return probability distribution for Monte Carlo.

As for the TD algorithms (see Figure 9), both SARSA and Q-Learning seem to perform quite similarly, with SARSA displaying a slightly better performance. Similar to results from the DP algorithms, it can be seen that the first two steps yield a negative reward, though unlike Q-Learning, SARSA learned to obtain significant positive reward on the third step. The return probability distributions indicate a low chance of receiving a negative total reward, displaying a significantly higher performance compared to MC.



(a) SARSA: Average reward at each time step

(b) Q-Learning: Average reward at each time step

(c) SARSA: Return probability distribution

(d) Q-Learning: Return probability distribution

Figure 9: Comparison of SARSA and Q-Learning. Top row (a, b): Average reward over time. Bottom row (c, d): Return probability distributions.

When comparing the learning behaviour of First-Visit MC, Q-Learning, and SARSA (see Figure 10), it can be seen that both SARSA and Q-Learning stabilise quite quickly at the same average reward obtained, after about 50 learning steps. Monte-Carlo also stabilises at this point, although it is less stable and has a lower average value.

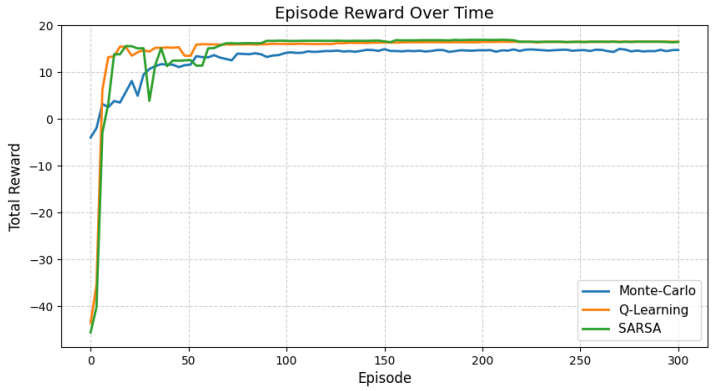


Figure 10: Comparison of the average reward evolution over the course of learning of First-Visit MC, Q-Learning, and SARSA

6. Conclusion

In this report, some common algorithms used in RL were discussed. These algorithms include Policy Iteration, Value Iteration, First-Visit Epsilon-Greedy MC, SARSA, and Q-learning. Having analysed data from all the algorithms (see Figure 11), it was found that Policy Iteration and Value Iteration arrived at the best-performing optimal policy. This result indicates that, for the cases where the environment is known, DP is generally a great choice.

SARSA and Q-Learning fall shortly behind, with SARSA slightly outperforming Q-Learning. Those algorithms did not have any direct knowledge of the environment and still arrived at a great policy after seeing 100 episodes of training in the environment used in this report. Q-Learning 'prefers' to act more riskily, resulting in a lower average reward.

First-Visit Epsilon-Greedy MC yielded the worst results of all the policies. The results clearly indicate that it performs significantly better than the random policy, indicating that it did learn from the environment. However, it still seems to converge approximately as fast as the TD algorithms. The main reason for the lower mean return of MC seems to be caused by a small amount of extremely low returns.

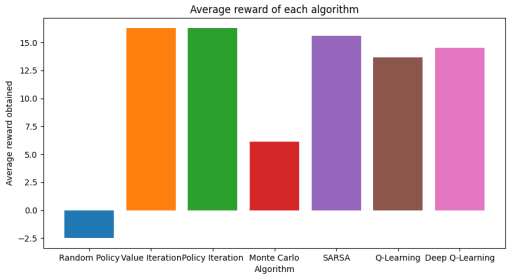


Figure 11: Comparison of the average (over 1000 episodes) final return achieved by each algorithm.

References

Qin, Y. (2025). Tabular Learning and Beyond. https://brightspace.ru.nl/d2l/le/content/502312/viewContent/3008418/View

Sutton, R. S., & Barto, A. G. (2018, November). *Reinforcement Learning, second edition*. MIT Press.

A. Deep Q-Learning

As an additional exercise, Deep Q-learning (DQL) was implemented. In contrast to the algorithms discussed in the main report, DQL is not tabular, meaning that it does not compute a lookup table for every single state. This is mainly useful when dealing with environments that have an extremely large observation space since it would be impossible to use tabular methods for these types of environments since training would take too long while likely also taking too much space to store all information.

A.1. Theoretical Deep Q-Learning

Deep Q-Learning uses a neural network to circumvent the need for lookup tables. At the output layer, this neural network needs to have one neuron for each possible action. The output value of a single neuron in the output layer represents the value of this action in the given state. From these output values, the optimal action can for instance be selected greedily by picking the action with the largest value (softmax could also be used, picking each action at the probability specified by softmax).

The input layer of the neural network needs to be able to take an observation from the environment. In the case of the environment used in this report, there are 4 input neurons, the first two representing the agent position, the latter two representing the target position. However, for an agent that would for example observe the world with a camera, the flattened image could be used as input for the network.

In order to train a neural network, gradient descent (GD) is used. To allow GD to be used, there needs to be a loss function. According to Qin (2025), the loss for DQL can be computed using Equation 4:

$$L_t(\mathbf{w}_t) = \mathbb{E}\left[\left(R_{t+1} + \gamma \max_{a'}(S_{t+1}, a'; \mathbf{w}_{t-1}) - Q(S_t, A_t; \mathbf{w}_t)\right)^2\right] \tag{4}$$

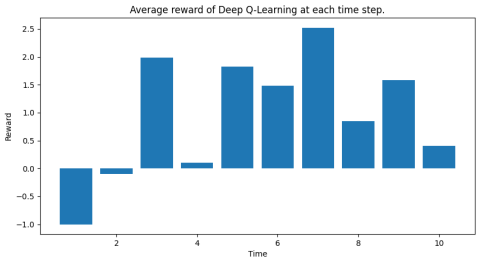
With t the time, R_t the reward at time t, S_t the state at time t, A_t the action taken at time t, $\mathbf{w}_t$ the weights of the model at time t, a' the set of actions that can be taken, and Q the model, that gives the value of a state and action computed using the weights of the model.

A.2. Practical Deep Q-Learning

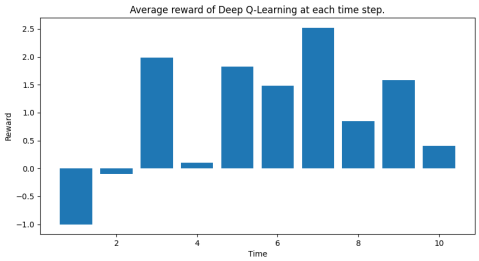
The basic implementation of DQL is fairly simple. Two models are maintained, the previous model and the current one. The model is trained by looping over many episodes, at each step in the episode, obtaining the reward and next state, then taking an action using some policy, and using this information to compute the loss using Equation 4.

There are, however, some issues with this implementation of DQL. Firstly, the target is determined by the previous model. Since this model is constantly updating, the algorithm is chasing a moving target, which can lead to algorithm instability (Qin, 2025). To solve this issue, instead of moving the target after every step, the target is moved after n steps. This somewhat improves model stability.

This improvement to the base algorithm was still not enough to make DQL work. This was caused by that the model trains on each episode only once, and that all data it trains on is sequential, making the model constantly need to adapt to the current episode. To circumvent this issue, experience replay was used. Many episodes were sampled, for each episode storing a tuple of (S, A, R, S', T) (with T indicating if S' is terminal). This data was then randomly shuffled and trained on using mini-batch GD over a few epochs. This allowed the model to converge on a good policy.



(a) The average reward received by Deep Q-Learning at each time step.



(b) Return probability distributions

Figure 12: Average reward over time and the Return probability distribution for Deep Q-Learning

The final model trained was trained on 100000 randomly sampled episodes, training on it in 10 epochs. A learning rate of 0.01 was used, a discount of 0.9 was used, and the batch size was set to 1024. The model was set to reset the target model every 10 batches. The final model receives the average rewards plotted in Figure 12a at every time step when the agent starts in the top left corner of the environment and the agent starts in the bottom right. Furthermore, the probability distribution of the return was plotted in Figure 12b. Note that there were a few extremely bad episodes in there, but in general, the model performs quite well, although not as well as SARSA and Q-Learning. Furthermore, the model does take quite some time to train, and sampling all the data required takes even longer.

In conclusion, the DQL algorithm discussed in this report seems to be able to somewhat approximate an optimal policy, and in cases where tabular algorithms are not possible, DQL could be a good alternative. However, when tabular solutions are feasible, it seems to be more efficient to just use a tabular algorithm such as Q-Learning instead.