Reinforcement Learning in Video Games

Valentí Torrents Vila

June 16, 2024

Resum– Resum del projecte, màxim 10 línies. ....

Paraules clau– Paraules clau del treball, màxim 2 línies . ...

Abstract– Versió en anglès del resum . ...

Keywords– Versió en anglès de les paraules clau. ....

1 INTRODUCTION

1.1 Preliminary Information

REINFORCEMENT learning, or RL, is a type of machine learning that revolves around how an intelligent agent should behave in an environment in order to achieve an specific goal. Unlike other machine learning paradigms, like supervised learning or unsupervised learning, RL's goal is to generate an intelligent agent that learns on its own while interacting with a dynamic environment. Through trial and error, the agent must chose which of the possible actions yield the most reward (both immediate and long-term), until achieving the necessary "knowledge" to go through the environment without any problems and solving it.

1.2 Objectives and expected results

The goal of this project is to dive into the world of Reinforcement Learning while using the Gym library effectively. Going from the fundamentals of RL, the project aims to understand and implement different RL algorithms. Starting with a quick introduction to this world, and going through

- E-mail: valentovi55@gmail.com
- Specialisation: Computation
- Work tutored by: Jordi Casas Roma
- Year 2023/2024

the basics with different tabular methods, our final objective is to go deep into the most complex methods applicable to video games. The following objectives mark the trajectory of this project:

1. Develop a foundational understanding of Reinforcement Learning and proper use of the Gym library.
2. Study and implement tabular methods such as Dynaim Programming, Monte Carlo, or Q-learning to create an agent capable of solving simple environments.
3. Study and comprehend some non-tabular methods, like policy-based methods and Deep Q Networks (DQN).
4. Apply the acquired knowledge to solve more complex and sophisticated environments, enhancing the agents' capabilities.
5. Generate a series of agents, each more complex, capable of achieving desired results in various games, while documenting findings and utilizing visualization tools to ensure a thorough exploration of Reinforcement Learning in video games.

1.3 Metodology

I have chosen to work using the Agile methodology for my project, working in iterative sprints, each lasting two weeks, which will be shown in the Gantt Diagram in figure 16. Using this methodology allows continual reassessment and adjustment if necessary since reinforcement learning is a dynamic field and it will need some adaptability. The tasks shown in the Gantt Diagram will be accomplished in sprints of up to two weeks, depending on the workload. Through the regular review sessions with my tutor, we will check the gradual progression of the project, and provide checkpoints for evaluation, adaptation, and resolution of new obstacles as they appear.

1.4 Planning

The project planning is visualized through a Gantt Diagram in figure 16. This will be a dynamic tool to help the development of the project. At first it shows the main objectives and milestones necessary and, over time, using the Agile methodology, the Gantt diagram will show more, depending on the obstacles faced and the necessary changes that appear in the process.

2 State of the art

Having talked about the basics of this project, we can now get into a more detailed explanation of what is the State of the art like, and how our first implementations of reinforcement learning methods has been like.

2.1 Introduction to RL

As we have just explained in the preliminary information, RL stands at an in-between point of artificial intelligence and decision-making. Thanks to that, it offers a powerful structure for creating and teaching agents to navigate through sequential decisions in dynamic environments. Simply put, RL revolves around the interaction between an agent and a specific environment. Through observations, this agent can perceive its environment and, based on its current state, selects one action or another. Furthermore, and depending on the specifics of the environment, the agent receives feedback in the form of rewards. Depending on the actions taken by the agent in each situation, these rewards can vary, indicating the desirability of said actions. Through these rewards, the agent is guided towards learning optimal strategies, which let it maximize these rewards over time. To further understand how Reinforcement Learning works, we have to talk about its components:

- The environment is, basically, the problem space with which the agent interacts. All the states, possible actions, rewards and rules that the agent has to abide come from the environment. Its dynamics dictate how the agent's actions modify future states and rewards, and even though we usually see dynamic environments in RL, there can also be static ones.
- The agent is the principal entity in RL, and the one that takes the decisions, earns the rewards, and goes through the environment. By navigating through the environment, and recieving different inputs from it, the agent learns to take better decisions. Its goal is to follow a behaviour (policy) which maximizes the rewards over time, solving then the task at hand.
- Actions are the choices available to the agent at each step within the environment. Depending on which action the agent takes, its surroundings will be influenced one way or another. Actions can be discrete(finite options) or continuous.
- The states repressent the current position or configuration of the environment. Depending on the moment and the position of the agent the state changes, and with it all the relevant information used by the agent to make decisions.
- Rewards, as we said before, are provided by the environment to the agent, and are signals that evaluate its actions. They can present inmediate or delayed feedback, and are the resposibles for the changes in the agent's behaviour over time, since its objective is to maximize these rewards in order to solve the environment.
- Policies define the agent's behavior by relating all states to actions. They mark the strategy followed by the agent to achieve its objectives. These policies can be deterministic, where each state has a specific action, or stochastic, where all actions have a probability of happening depending on the state and policy parameters. The objective of all Reinforcement Learning algorithms is to achieve the optimal policy v_π that maximizes long-term rewards, and leads the agent to solving the environment.

By interacting iteratively with its environment, and learning from its experience, RL algorithms generates agents

that learn autonomously how to behave, solving different tasks, from video games to controlling robots. By learning through trial and error, and without supervision, RL ends up being perfectly adequate to solving scenarios where the agent must adapt to dynamic and complex environments.

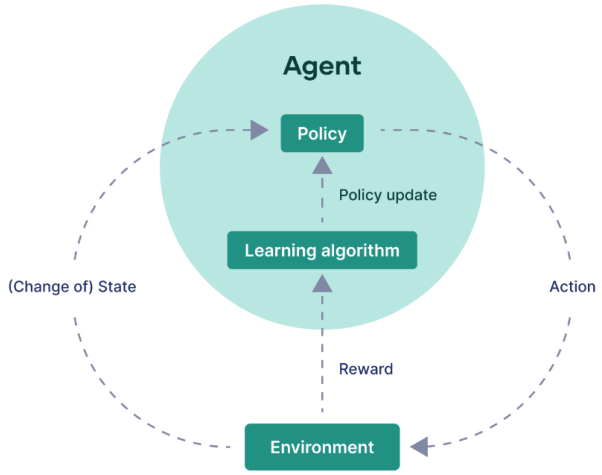

Fig. 1: RL General Framework

2.2 OpenAI Gymnasium

OpenAI Gym serves as a versatile platform for the development and evaluation of reinforcement learning algorithms. Offering a diverse array of environments spanning from classic control problems to complex simulated scenarios, Gym provides a standardized interface for interacting with different tasks. Each environment in Gym is encapsulated as a Markov decision process (MDP), allowing agents to interact with states, take actions, receive rewards, and transition between states based on probabilistic dynamics. With its user-friendly API and extensive documentation, OpenAI Gym[1] empowers researchers and practitioners to prototype, benchmark, and iterate on various reinforcement learning techniques with ease. Moreover, Gym's compatibility with popular RL libraries and frameworks fosters collaboration and accelerates the advancement of RL research and applications.

2.3 RL Methods

In the current state of reinforcement learning (RL), researchers employ tabular and non-tabular methods. Tabular methods, like dynamic programming and Monte Carlo, excel in small-scale tasks with discrete state and action spaces. They offer theoretical guarantees but struggle with larger, continuous spaces. Non-tabular methods, including deep reinforcement learning (DRL), leverage function approximation, particularly neural networks, to handle complex, high-dimensional environments. DRL algorithms like deep Q-networks (DQN) and policy gradients have demonstrated success in diverse applications, from robotics to gaming. Both approaches complement each other, with tabular methods providing theoretical foundations and non-tabular methods offering scalability and adaptability to real-world problems.

[1]OpenAI Gym `https://gymnasium.farama.org/`.

3 Tabular Methods

To get to understand the basics of RL we first have to talk about the tabular methods. These are the simplest ones in RL, since they store a specific action for each state in a table-like structure. The tabular methods we will talk about are Dynamic Programming, Monte Carlo methods, and Q-learning. To do so, we will work with finite Markov Decission Processes(MDP), trying to estimate value functions, which are functions of states that tell use how positive is for the agent to be in a specific state.

3.1 Markov Decission Processes

A Markov Decission Process(MDP) is a mathematical model used in RL to make sequential decisions in a stochastic environment. Basically, an MDP describes a system(or environment) where an agent makes decisions which results are subject to uncertainty. Just like the parts of a RL environment, a MDP contains states, actions and rewards. However, a MDP also contains a Transition Model, which describes the transition probabilities from one state to another after taking a certain action. In order to solve a finite MDP, we can use different tabular methods, which we are going to talk about now.

3.2 Dynamic Programming

Dynamic Programming(DP) is used to find the optimal policies to follow in order to solve a finite MDP. In a finite MDP, states, actions, and rewards are finite, and their dynamics are represented by transition probabilities. In order to achieve a good policy in RL in general, we use value functions. With DP we can compute these value functions by satisfying the Bellman optimality equations.
The Bellman Optimality Equations express the principle of optimality, that an optimal policy can be decomposed into smaller subproblems, each of which is solved optimally. In the field of MDPs, the Bellman Optimality Equations are:

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma v_*(s') \right]$$

Fig. 2: Equation for State-Value Function (V)

$$q_*(s, a) = \sum_{s',r} p(s', r \mid s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$

Fig. 3: Equation for Action-Value Function (Q)

In our project, we employed policy iteration to find this optimal value function. First, we initialized a policy, and then evaluated and improved it iteratively until it converged.

To keep evaluating the policy we had, we estimated the value function with the current policy using the Bellman Equations. Afterwards, we selected the actions that maximized the expected returns in order to update the policy. After repeating this 2 steps several times and observing no further improvements, we had the optimized value function and policy that let us guide the decision-making through the MDP.

3.3 Monte Carlo Methods

Even though Monte Carlo and DP are both tabular methods, Monte Carlo methods do not require a model of the environment to operate. Instead, they rely on trial and error, sampling trajectories by interacting with the environment.
In our Monte Carlo policy evaluation process, we began by executing episodes according to the current policy, collecting state-action-reward trajectories. After each episode, we calculated returns for each visited state, providing estimates of their values. These returns were then averaged over multiple episodes to update the value function incrementally. Optionally, we improved the policy based on the estimated values, possibly selecting actions greedily to maximize returns. This iterative refinement continued over multiple episodes until convergence, where the estimated value function approached the optimal one as the number of episodes increased.

3.4 Q-learning

Now that we have talked about both Monte Carlo and DP methods, we can get to understand temporal-difference (TD) learning, one of the most important methods of RL. TD is a mix of DP and Monte Carlo ideas, since it can learn directly from experience like Monte Carlo, and also update estimates from past estimates without having to wait for the final outcome, like DP.
From all the different TD methods, we just have worked with Off-Policy Q-learning, because it presents a different logic from the rest of the tabular methods. In Q-learning, the agent maintains an estimate of the value of each state-action pair represented by the action-value function Q(s, a). This function quantifies the expected cumulative reward that the agent will receive by taking action a in state s, and then following the optimal policy. With Q-learning, we aim to iteratively improve this estimate until it converges to the optimal action-value function. However, in off-policy Q-learning, we find that the agent learns from experiences generated by following a different policy than the one being learned.
Off-policy Q-learning allows agents to learn from past experiences generated under different policies, leading to faster learning. It effectively balances exploration (trying new actions) and exploitation (leveraging known information) in complex environments. This approach is versatile, enabling both policy evaluation and improvement, and benefits from experience replay to stabilize learning. Additionally, it's suitable for batch learning settings, where agents learn from fixed datasets, making it practical in scenarios where online exploration is challenging or costly.

3.5 Tabular Methods testing and results

To further understand the algorithms behind all three tabular methods we have talked about, we applied them to the Frozen Lake environment from the Gymnasium library.

Fig. 4: Gymnasium Frozen Lake 4x4

With DP, we followed the pseudocode in algorithm 1 to achieve the optimal policy:

Algorithm 1: Policy Evaluation

Data: π, the policy to be evaluated,
A threshold $\theta > 0$ accuracy of estimation,
$V(s)$ initialized arbitrarily.
while $\Delta > \theta$ **do**
  $\Delta \leftarrow 0$
  foreach $s \in S$ **do**
    $v \leftarrow V(s)$
    $V(s) \leftarrow$
    $\sum_a \pi(a|s) \sum'_s, rp(s', r|s, a)[r + \gamma V(s')]$
    $\Delta \leftarrow max(\Delta, |v - V(s)|$

The policy generated was
$\pi = [0, 3, 3, 3, 0, 0, 0, 0, 3, 1, 0, 0, 0, 2, 1, 0]$, with which the agent got to the end 82% of the time. If we used the 8x8 Frozen Lake environment, this percentage went to a 100%, since there was a possible route with no chances of failing.

AQUI PARLEM DE MONTECARLO

For Q-learning, the pseudocode was the algorithm 2.

The policy generated was the same as in DP, making them equally efficient at solving the 4x4 Frozen Lake environment.

Algorithm 2: Q-learning Off-policy

Data: Step size $\alpha \in (0, 1]$,
$\epsilon > 0$
$Q(s, a)$ for all $s \in S^+, a \in A(s)$ initialized arbitrarily.
foreach *episode* **do**
  Initialize S
  foreach *step of episode* **do**
    Choose A from S using policy derived from Q(ϵ-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow$
     $Q(S, A)+\alpha[R+\gamma max_a Q(S', a)-Q(S, A)]$
    $S \leftarrow S'$
  until S is terminal

4 APPROXIMATE SOLUTION METHODS

Now that we have understood the basics of RL, we can dig into a more complex field. Unlike when we were working with tabular methods, most of the state spaces we will be working with RL will be extremely complex and large. That is why we cannot aim to find an optimal policy v_π, but instead to find an approximate solution. To do so, the main tool we will be using is Generalization. Through generalizing the problems we face, we treat them like similar problems we have already faced, thus achieving proper results without the need of dedicating enormous quantities of time and data. In this second section of the project, we will be talking about different Approximate Solution Methods and how to implement them in different environments.

4.1 On-Policy Prediction

We begin this chapter focusing on the utility of function approximation to estimate the state-value function from on-policy data. This means that, through using a known non-optimal policy π, we can estimate the state-value function v_π. The main difference we will work with from now on, is that instead of representing v_π as a table, we will represent it as a parametrized functional form with a weight vector $w \in R^d$. Now, instead of using any of the previous formulas to calculate the value of a state s, we will approximate it with the weight vector as $\hat{v}(s, w) \approx v^\pi(s)$. One of the cases in which we can use $\hat{v}$ may be in a neural network, where $\hat{v}$ is the function computed, using w as the vector of connection weights in through all the layers of the network. Also, we can use w to define the split points of a decission tree which computes the function $\hat{v}$.
One thing we have to keep in mind for future applications of these methods, is that the dimensionality of w may be way less than the number of states. What this means is that unlike in tabular methods, where we changed only the value of the state we updated, every time we update a state, the values of other states will change too, due to generalization between them. This generalization provides a more powerful learning potential, but also a more difficult management and understanding.
This chapter sets the foundation for understanding on-policy approximations in reinforcement learning, highlighting their theoretical underpinnings and practical implications in both fully and partially observable environments.

4.2 Value-function Approximation

Up until now, when we updated a specific state, we only shifted a tiny bit the overall behaviour of our agent towards the desirable policy, by modifying only the estimated value of the state we updated. Each of these updates is represented by $s \mapsto u$, where s is the state updated and u is the update target where s is shifting towards. Now, with generalization, the updates at s also change the estimated values of many other states. Machine learning methods that do this are called $supervisedlearningmethods$, and when the outputs are numbers, we call them $functionapproximation$. To generate an estimated value function using these methods, we take the $s \mapsto u$ of each update as training examples. By doing this, we enable these methods to be treated like any other learning method, and can now apply some of their algorithms. However, due to the necessity of having interactions and live updates, we need methods able to handle non stationary target functions.

4.3 PREDICTION OBJECTIVE

Through all this chapter, we have talked about the weight vector w but, how can we choose a proper w?
In tabular methods, the learned value function could end up being the true value function. However, here we cannot. Changing one state affects others, so we will never reach the correct value for all of them at the same time. That is why we need to give more importance to some of the states, specifying a distribution $\mu(s) \geq 0, \sum_s \mu(s) = 1$, which represents how much we care about the error in each state. The error here is calculated with the difference between the approximate value $\hat{v}(s, w)$ and the true value $v_\pi(s)$. Knowing all of this, we can now create an objective function, $\overline{VE}(w)$, which we will want to minimize in order to obtain the best w.

$$\overline{VE}(\mathbf{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[v_\pi(s) - \hat{v}(s, \mathbf{w})\right]^2$$

Fig. 5: Mean Squared Value Error

At first, this formula seems too complicated to apply, since we are summing through an exponentially big space (μ), and we are working with the true value v_π which we don't know. However, there are some algorithms with which we can approximately optimize it.

4.4 Stochastic Gradient Descent

In order to approximately optimize what we just talked about, we can use Stochastic Gradient Descent, also known as SGD. To work with SGD, we need to assume a few things:
Since we will be updating w as we bounce through the state space according to the on-policy distribution, all states must be visited in proportion to μ, which is what we have chosen from the on-policy distribution. Furthermore, our value

function $\hat{v}(s, w)$ has to be differentiable with respect to w, necessary to calculate the gradient. And, as we said before, there needs to be a surrogate for $v_\pi(S_t)$, which is U_t.
With all this, SGD provides us with an update rule to apply to w for each visit to a state.

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t)$$

Fig. 6: SGD update rule

Where α is the step-size parameter, and $\nabla \hat{v}(S_t, w)$ is the gradient. Following this rule, SGD methods adjust w by a small ammount in the direction that would reduce most the error in that specific visit to that state.
Even though it may not seem obvious at first, it is necessary to carefully select which is the value of α. If the step on a certain direction was too big, we could end up with a value function with zero error in a certain state, but an abysmal error in others. Taking small steps, and adjusting w little by little, will let SGD methods converge to a local optimum.
We now focus on the target U_t. If the expected value of U_t in a state S_t equals the true value $v_\pi(s)$, we call U_t unbiased, and w will undeniably converge into a local optimum of $\overline{VE}$.

4.4.1 Examples of Gradient Descend methods

Monte Carlo algorithm requires us to generate the states through interaction with the environment, following a policy π. If U_t is created by following π, the target of Monte Carlo ($U_t \rightarrow G_t$) has to be an unbiased approximation of $v_\pi(S_t)$. That is why, if we apply a gradient descent on Monte Carlo, it is going to find a locally optimal solution.

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[G_t - \hat{v}(S_t, \mathbf{w}) \right] \nabla \hat{v}(S_t, \mathbf{w})$$

Fig. 7: Gradient MC

However, having to wait until the end of each episode can make the learning slower. Therefore, we can contemplate other alternatives, like bootstrapping.
When bootstrapping, rather than waiting until the end of an episode to make updates based on the true return, we update at each time step using the current estimates of future returns. This means that bootstrapping targets, such as n-step returns or DP target, all are biased. Up until now, the target U_t has been independent of w_t. However, with bootstrapping it isn't the case, which means that the update rule is not a true gradient step anymore but a Semi-gradient, which makes guaranteing convergence impossible. In spite of its complications, we can still work with semi-gradient methods, such as semi-gradient TD(0), which uses $U_t = R_{t+1} + \gamma \hat{v}(S_{t+1}, w)$ as its target. Using this, the update rule, looks like this:
And as we saw in the previous chapter, where TD methods like Q-learning were faster than Monte Carlo, here we can say the same. (POTSER POSAR REFERENCIA A QUAN FACI LES PROVES).

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w}) \right] \nabla \hat{v}(S, \mathbf{w})$$

Fig. 8: Semi-gradient TD

4.5 Linear Methods

Linear methods in function approximation are crucial for reinforcement learning due to their simplicity and effectiveness. These methods approximate the state-value function as a linear combination of feature vectors and weights. Specifically, the approximate value function $\hat{v}(s, \mathbf{w})$ is represented as:

$$\hat{v}(s, \mathbf{w}) \doteq \mathbf{w}^\top \mathbf{x}(s) = \sum_{i=1}^{d} w_i x_i(s)$$

Here, $\mathbf{x}(s)$ is a feature vector representing the state s, and $\mathbf{w}$ is the weight vector.
When using linear methods in Monte Carlo methods, the return G_t is used to update the weights in a way that the observed returns fit the linear approximation of the value function. This results in a simple and efficient update rule:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left(G_t - \hat{v}(S_t, \mathbf{w}) \right) \mathbf{x}(S_t)$$

Here, the linear approximation ensures that the weight update is proportional to the difference between the return and the estimated value, scaled by the feature vector $\mathbf{x}(S_t)$. This means the weight updates directly reflect how well the linear combination of features is predicting the actual returns, allowing for straightforward and interpretable adjustments.
In the context of gradient TD methods, linear function approximation leads to the semi-gradient TD(0) update rule:

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left[R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w}) \right] \mathbf{x}(S)$$

Here, the linear method simplifies the computation of the gradient $\nabla \hat{v}(S, \mathbf{w})$ to just the feature vector $\mathbf{x}(S)$. This simplification allows the algorithm to efficiently adjust weights based on the temporal difference error, making the updates more computationally feasible and easier to analyze. The linear relationship ensures that the changes in the weights are directly proportional to the features, maintaining a consistent and interpretable mapping between states and value estimates.
Feature construction is crucial for the effectiveness of linear methods, and in this project we focused on using polynomial features, which allow linear models to represent complex interactions among state variables.

Given a state with k variables $(s_1, s_2, \ldots, s_k)$, a polynomial feature of order n is:

$$x_i(s) = \prod_{j=1}^{k} s_j^{c_{ij}}$$

where c_{ij} are integers from $\{0, 1, \ldots, n\}$. These features capture higher-order interactions.
As an example, suppose we have two variables s_1 and s_2. The second-order polynomial features would be:

$$x(s) = (1, s_1, s_2, s_1^2, s_1 s_2, s_2^2)^\top$$

This vector includes linear, squared, and interaction terms, enabling the approximation of complex value functions.

4.6 Approximate solution methods testing and results

In this section, we present our experiments using the CartPole environment from the Gymnasium library to evaluate the performance of Gradient Monte Carlo (GMC) and Gradient Temporal Difference (GTD) methods. The goal was to analyze the efficacy of these methods in reinforcement learning tasks, specifically in terms of their ability to balance the pole and maintain the cart within bounds.

4.6.1 Introduction to the CartPole Environment

The CartPole environment is a classic reinforcement learning problem where a cart, which can move horizontally along a track, has a pole attached to it. The state space consists of four continuous variables: cart position, cart velocity, pole angle θ, and pole angular velocity $\hat{\theta}$. The action space is discrete with two possible actions: pushing the cart left or right. We can see a representation of it in figure 9

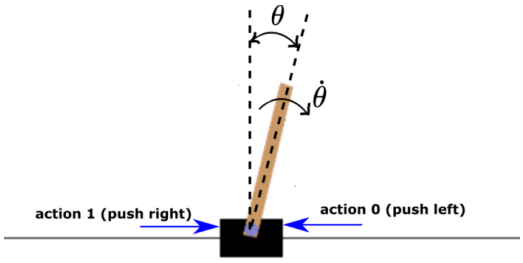

Fig. 9: Illustration of the CartPole environment

The primary goal is to keep the pole balanced for as long as possible. The environment terminates when the pole falls past a certain angle or the cart moves too far from the center. The agent receives a reward for each time step the pole remains upright, encouraging strategies that maintain balance. Balancing the pole requires continuous and precise adjustments to the cart's position and velocity, making the CartPole environment an excellent testbed for evaluating reinforcement learning algorithms.

4.6.2 Experimental Setup

Both Gradient Monte Carlo and Gradient Temporal Difference methods were implemented and tested under the same conditions. The algorithms were trained over multiple episodes, with performance metrics such as the average reward per episode recorded.

4.6.3 Results and Discussion

At first, we measured both methods in terms of their ability to maximize the total reward over a series of episodes. The results were surprisingly poor, leaving us questioning what was wrong. We tried implementing the algorithms using regular gradient descend, as well as applying lineal methods. By using polynomial features, we went from having the 4 main features to having 15, including the bias. After applying these changes, we got to see some improvement, but it was not enough.
In the GMC implementation, with a learning rate of 0.001, a discount factor of 0.95, a exploration rate ϵ of 0.01 and training for 100000 episodes, the final weights of the features looked like this (see figure 10): As we can see, most of

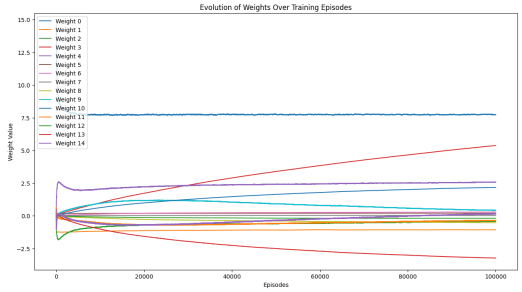

Fig. 10: Evolution of feature weights over time

the feature weights converged to reasonable values. However, the bias weight always went to a number close to 8. This made the whole model shift to a higher value when deciding which action to take, rendering the model as a failiure. However, by testing manually different values, we found that if we changed the bias' weight to a number close to 0.02 the model worked much better, going from an average reward per episode of 10 to an average of 150 steps until termination.
The Gradient Temporal Difference method, specifically using TD(0), updates the value function incrementally based on the temporal difference error. In the implementation of this method we tried different values for the parameters, and after seeing no kind of improvement, we decided to stick with the same we used in GMC. Although GTD was supposed to be faster at converging than GMC, the bias weight, as well as the others, remained a persistent issue, leading to less stable policy updates (see figure 11.

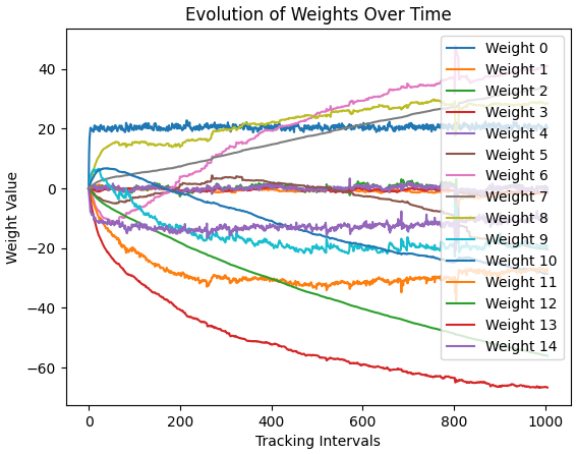

Fig. 11: Evolution of feature weights over time

After having trained the agent for 100000 episodes, and seeing some kind of nonsensical convergence, the average reward per episode was 15. Having reviewed the whole code, and finding no way of improving from the actual state, we decided to focus on the next section of the project in hopes of obtaining better results.

4.6.4 Conclusion

The experiments with the CartPole environment highlighted the challenges associated with the bias weight in both Gra-

dient Monte Carlo and Gradient Temporal Difference methods. While both methods showed potential in terms of learning and performance, addressing the bias weight issue is crucial for achieving stable and reliable results.

5 Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) combines reinforcement learning (RL) with deep learning. In RL, an agent learns by interacting with an environment to maximize rewards. DRL uses deep neural networks to handle complex, high-dimensional environments where traditional RL methods fall short.

5.1 About DRL

DRL excels in solving problems in large state spaces and complex environments, making it essential for advanced applications. It allows for learning from raw data, enabling more sophisticated decision-making.
Even though we focused on the RL implementation in the video games world, we have to know that DRL is also used in robotics and natural language processing primarily.

- **Video Games:** Achieving superhuman performance in games like Go, Chess, and Atari games.
- **Robotics:** Tasks like manipulation, locomotion, and navigation using sensory inputs.
- **Natural Language Processing:** Dialogue generation, machine translation, and text-based games.

5.2 Deep Q-Networks

Deep Q-Networks, or DQNs, combine Q-learning with deep neural networks, making it suitable for high-dimensional state spaces. DQN uses a deep neural network to approximate the Q-function, which predicts the maximum expected future reward for an action in a given state. Traditional Q-learning aims to learn the optimal Q-function, $Q^*(s, a)$.
In order to achieve proper stable and efficient learning, DQNs present several key components in its arquitecture:

- **Function Approximation with Neural Networks:** Instead of a lookup table, DQN employs a neural network to approximate the Q-values. This network takes the state s as input and outputs a Q-value for each possible action. The use of deep networks allows DQN to handle high-dimensional inputs, making them useful to process environments with image inputs.
- **Experience Replay:** To take into account the correlation between consecutive experiences, DQN uses experience replay, a mechanism that stores the agent's experiences in a replay buffer. During training, mini-batches of experiences are randomly sampled from this buffer to break the correlation and provide more stable and diverse training data.
- **Fixed Target Network:** Another technique to stabilize training in DQN is the use of a target network. This network, denoted as $Q(s, a; \theta^-)$, is a copy of the Q-network $Q(s, a; \theta)$ but with parameters θ^- that are updated less frequently. By holding the target network constant for a fixed number of iterations, DQN reduces the risk of divergence and oscillations in the learning process. The target Q-value for the loss function is computed as:

$$y = r + \gamma \max_{a'} Q(s', a'; \theta^-)$$

  where r is the reward, γ is the discount factor, and s' is the next state.
- **Loss Function and Optimization:** The objective of DQN is to minimize the difference between the predicted Q-values and the target Q-values. This is achieved by minimizing the mean squared error loss:

$$L(\theta) = E_{(s,a,r,s')\sim D}\left[(y - Q(s, a; \theta))^2\right]$$

  where $y = r + \gamma \max_{a'} Q(s', a'; \theta^-)$ is the target Q-value. The network parameters θ are updated using stochastic gradient descent or its variants.
- **Exploration-Exploitation Trade-off:** DQN balances exploration and exploitation using an ϵ-greedy policy. Initially, the agent explores the environment by selecting random actions with a high probability ϵ. Over time, ϵ is decayed, allowing the agent to exploit its learned policy more frequently. This trade-off is crucial for ensuring that the agent gathers diverse experiences while gradually improving its performance.
- **Handling High-Dimensional Inputs:** In many applications, the state space can be high-dimensional, such as images in video games. DQN processes these inputs using convolutional neural networks (CNNs) to extract spatial features.

5.2.1 Types of DQNs

Having talked about DQNs, we now need to know how different deep neural networks can help create our DQN:

- **Feedforward Neural Network (FNN) DQNs:** They use a fully connected neural network to approximate the Q-value function. The network consists of an input layer, multiple hidden layers with activation functions like ReLU, and an output layer. The input is the state representation, and the output layer provides Q-values for each action. The network is trained using the Bellman equation to update Q-values based on rewards and future estimates. This type is simple to implement but struggles with high-dimensional inputs.
- **Convolutional Neural Network (CNN) DQNs:** These are ideal for processing high-dimensional state spaces like images. The architecture includes convolutional layers to detect spatial features, followed by fully connected layers. The input is typically an image or a stack of images, and the output layer provides Q-values for each action. Training involves updating Q-values based on rewards and future estimates using the Bellman equation. CNNs excel at handling visual data but are computationally intensive.

To better the efficiency, and allow non-linearity to the networks we can use Rectified Linear Units (ReLU). ReLU is an activation function used in both Feedforward Neural Networks (FNNs) and Convolutional Neural Networks (CNNs) to introduce non-linearity and help the networks learn complex patterns. It outputs the input directly if it's positive, and zero otherwise. In FNNs, ReLU is applied after each hidden layer to improve training efficiency and model performance. In CNNs, it is used after each convolutional layer to activate features detected by filters, aiding in the learning of intricate patterns. ReLU is popular for its simplicity and effectiveness, preventing the vanishing gradient problem and simplifying computations.
Other deep neural networks to keep in mind are Recurrent Neural Networks (RNNs), including LSTMs and GRUs, for temporal dependencies; Actor-Critic Networks like A2C and A3C for combining policy and value functions; Deterministic Policy Gradient Networks such as DDPG and TD3 for continuous actions; Policy Gradient Networks like PPO and TRPO for stable policy optimization; and Generative Networks like VAEs and GANs for synthetic data and model-based reinforcement learning.

6 DQN Testing and results

In this project, we implemented two different DQN to test. The first one is a DQN using feedforward neural networks to solve the cartpole environment (same one used in section 4.6). It is a more simple approach that allows us to understand the basics of a DQN, and lets us compare its results with the stochastic gradient implementations. Then, we AQUI VA LA SEGONA IMPLEMENTACIO, UNA CNN AMB EL PONG MES COMPLEXA.

6.1 Cartpole Environment using FNN DQN

The development process involves creating a Deep Q-Learning Network (DQN) to train an agent to balance a pole on a cart. The key steps include initializing the environment, designing the neural network, implementing the training loop, and managing the exploration-exploitation trade-off. The process is iterative, involving frequent evaluation and adjustment of key parameters.

6.1.1 Initialization

First, we initialized the CartPole environment, providing state and reward information for the agent. Key parameters are defined for the DQN: the discount factor (gamma) which determines the importance of future rewards, typically set to 0.99; the exploration rate (epsilon) that controls the exploration-exploitation trade-off, initially set to 1 and decaying over time; and the total number of episodes for training the agent.

6.1.2 Neural Network Design

The DQN uses a feedforward neural network architecture. The input layer receives the state representation from the environment, which for CartPole is a vector representing position, velocity, angle, and angular velocity. The network includes two hidden layers with ReLU activation functions. The first hidden layer consists of 128 neurons, while the second hidden layer has 56 neurons. The output layer provides the Q-values for each possible action (move left or move right) and has two neurons with a linear activation function. The design also includes maintaining two networks: the main network and the target network. The main network is used to predict the Q-values for the current state, whereas the target network predicts the Q-values for the next state. The target network is updated periodically to stabilize the training process.

6.1.3 Replay Buffer

The replay buffer is initialized at the beginning of the code, and stores the experiences of the agent throughout training to allow it to learn from past interactions. Then, we randomly sample a batch of experiences from the replay buffer to train the neural network. By implementing a replay buffer and using batch sampling, the DQN is able to learn more efficiently and effectively, making better decisions over time as it is exposed to a wide variety of experiences.

6.1.4 Training Loop

For each episode, we reset the environment and initialized the variables to track rewards.
We then ran the episode until it terminated, or it got to 400 steps (to stop it from being too slow). In order to take a step, we first used the epsilon-gredy policy to select either a random action (exploration) or the one with the highest predicted Q-value (exploitation). After performing the action, we observed the next state, reward, and if it terminated. Finally, after storing the experience in the replay buffer, if it had enough experiences we sampled a batch and trained the main network, adjusting the target network periodically.

6.1.5 Results analysis

In figure 12 we can observe that, as training progresses, the agent explores less and exploits more, relying on learned knowledge to make decisions. This balance is crucial for effective learning.
The steady increase in the Q-value shown in figure 13indicates that the agent is learning to predict higher rewards, reflecting improved policy performance. Fluctuations may indicate exploration phases or instability in learning.
We can observe in figure 14 that the agent kept getting better, but not until the very end it accomplished perfect results.
A decreasing trend in figure 15 indicates that the network is improving its predictions. Spikes may occur due to significant updates in the target network or new exploration phases.
Overall, the results were really good. We tested different values for the discount factor, the exploration rate, and the total number of episodes to train for, and decided to work with the ones in the code REFERENCIA AL CODI(COM ES FA?). The parameter that was more troublesome was the epoch in keras.models.Model.fit, which decided how many complete passes through the entire training dataset the agent should do. At first it was set to 100, which made the convergence much faster, but the episodes and the overall running

time much larger. After trying lower values, and seeing that the computing times were way faster, we decided to stick with an epoch of 1 which, by giving up some efficiency, made the whole script work faster, reducing the overall computing time from 15 hours to 4. VULL MENCIONAR ELS VIDEOS AMB ELS RESULTATS DEL CARTPOLE, SE SUPOSA QUE HAN D'ESTAR A UN GITHUB I HE DE FER REFERENCIA AL GITHUB?

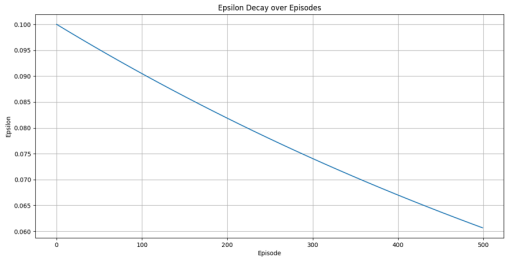

Fig. 12: Epsilon Decay over Episodes

Fig. 13: Average Q-value over Time

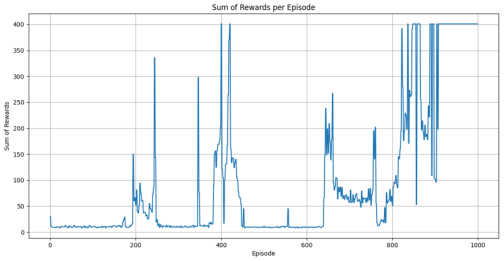

Fig. 14: Sum of Rewards per Episode

6.2 Pong Environment using CNN DQN

…. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. .

7 Conclusions

…. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. .

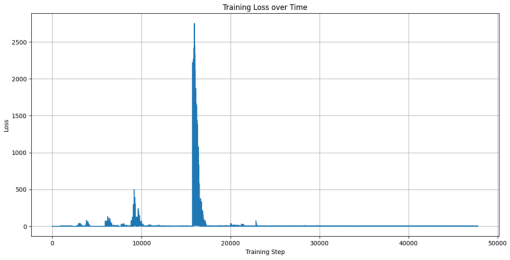

Fig. 15: Training Loss over Time

…. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. .

Special Thanks

… .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. . …. .. …. .. …. … ….. … ….. … … ….. …. .

References

https://gymnasium.farama.org/

Appendix

Tasks	Start Date	End Date	05/02/24	19/02/24	04/03/24	18/03/24	01/04/24	15/04/24	29/04/24	06/05/24	20/05/24	03/06/24	17/06/24
Basic understanding of RL	12/02/24	19/02/24											
Familiarize with Gym library	12/02/24	26/02/24											
Introduction to MDP	26/02/24	04/03/24											
Dynamic Programming	04/03/24	18/03/24											
Monte Carlo Methods	18/03/24	01/04/24											
Q-learning	01/04/24	15/04/24											
Documentation and Testing of Tabular Methods	26/02/24	29/04/24											
Introduction to Approximate Solution Methods	29/04/24	06/05/24											
On-Policy Methods	06/05/24	20/05/24											
Off-Policy Methods	20/05/24	03/06/24											
Policy Gradient Methods	03/06/24	17/06/24											
Documentation and Testing of non-Tabular Methods	17/06/24	30/06/24											

Fig. 16: Gantt Diagram