Artificial Intelligence and Machine Learning

Giuseppe Valente
Sapienza - University of Rome
Engineering in Computer Science

Contents

Chapter 1

Introduction

This work aims to study and solve Reinforcement Learning (RL) problem, using models like Q-learning and Deep Q-Network (DQN). We will introduce a basic problem called Frozen Lake, and we will study learning strategies that help the agent to solve the environment. The challenges in RL are understanding how an agent can explore and exploit an environment effectively to achieve its goal. Q-learning is a model-free RL algorithm that teaches an agent to assign values to each action it might take, conditioned on the agent being in a particular state. For any finite Markov decision process, Q-learning finds an optimal policy (maximizing the expected value of the total reward over any and all successive steps), starting from the current state. While effective for simple environments, its scalability is limited when dealing with larger or more complex state spaces. To address this limitation, DQN employs deep neural networks to approximate the Q-value function, enabling the application of RL to high-dimensional and complex environments. With this study, we will see the theoretical foundation and the practical implementation of these methods, evaluating their performance in solving the Frozen Lake problem.

Chapter 2

Frozen Lake

Frozen lake involves crossing a frozen lake from start to goal without falling into any holes by walking over the frozen lake. Frozen Lake is available into two versions: deterministic and non-deterministic setting the variable

Listing 2.1: slippery variable

```
is_slippery = True # Set the Non Deterministic environment
is_slippery = False # Set the Deterministic environment
```

In the non-determinisitc environment the agent may not always move in the intended direction while in the determinisitc environment it always moves in the intended location. In this project we use the simplest version of the Frozen Lake that is a 4x4 grid where the agent starts to position (0,0) with the goal located at the position (3,3). Holes in the ice are distributed on the cells of the grid. The agent makes moves until they reach the goal or fall in a hole. The figure 2.1 shown an example ov the Frozen Lake environment:

Figure 2.1: Frozen Lake 4x4 environment

2.1 Action Space

The agent has 4 possible actions to use:

- 0: Move left
- 1: Move down
- 2: Move right
- 3: Move up

2.2 Observation Space

The observation is an integer value representing the current position of the agent. The image in figure 2.2 shown the observation space

Figure 2.2: Frozen Lake 4x4 environment: Observation Space

2.3 Rewards

The rewards are the feedback obtain by the agent while perform an action when it is a specific state. The agent will try to maximize the rewards during the execution. In this environment the rewards are:

- Reach the goal $\rightarrow +1$
- Reach the hole $\rightarrow 0$
- Reach the frozen $\rightarrow 0$

2.4 Episode ends

The episodes ends in the following cases:

- Termination (The agent fall into hole or achieve the goal)
- Truncation (The time length of the episode, in this case is 100)

The condition of termination and truncation are given by the environment after agent takes an action

2.5 Non-Deterministic environment

We know in deterministic environement an action taken by the agent in a given state will go always in the expected destination state. This is not true for the non-deterministic environment where if the agent wants to take the action "move left" it will taken with a

$$\Pr = \frac{1}{3} \tag{2.1}$$

2.6 Brief of Q-Learning

Reinforcement Learning (RL) is one of three basic machine learning ofmachine learning concerned with how an agent should take actions in a dynamic environment to maximize a reward. Finding a balance between exploration (of uncharted environment) and exploitation (of current knowledge) with the goal of maximize the cumulative reward. The environment is typically stated in the form of a Markov decision process (MDP). The purpose of reinforcement learning is for the agent to learn an optimal policy that maximizes the reward function or other user-provided reinforcement signal that accumulates from immediate rewards.

2.6.1 Agent's learning task

A basic reinforcement learning agent interacts with its environment in discrete time steps. At each time step t, the agent receives the current state S_t and reward R_t. It then chooses an action A_t from the set of available actions, which is subsequently sent to the environment. The environment moves to a new state S_{t+1} and the reward R_{t+1} associated with the transition (S_t,A_t,S_{t+1}) is determined. The goal of a reinforcement learning agent is to learn a policy

2.6.2 Q-Function

The Q-Function (the state-action) function is a function that helps the agent to decide which action to take in a given state in order to maximize its expected cumulative reward. In general we can write the Q-Function using the Bellman equation:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right) \tag{2.2}$$

Q(s, a): Current estimate of the expected cumulative reward for taking action a in state

s and following the optimal policy thereafter.
Q(s, a): The current Q-value for the state-action pair (s, a).
α**:** The *learning rate*. It controls how much the new information should override the old information. A higher value means the agent will quickly adapt to new experiences.
r: The *immediate reward* received after taking action a in state s.
γ**:** The *discount factor*. It determines how much the agent values future rewards compared to immediate rewards. A value close to 1 means the agent cares about long-term rewards, while a value close to 0 means the agent prioritizes immediate rewards.
$\max_{a'} Q(s', a')$**:** The maximum Q-value over all possible actions a' in the next state s'. This represents the best expected reward the agent can get from the next state onward.
Q(s, a): The original Q-value before the update, which is updated based on the new experience.

2.6.3 Hyperparameters

They are parameters set before the learning process begins, play a crucial role in determining the efficiency and effectiveness of the learning process.

- **Learning rate (α)**: Controls how much the Q-values are adjusted during each update. A larger learning rate leads to faster updates, but can also result in instability. A smaller learning rate may lead to slower learning.
- **Discount factor (γ)**: Determines how much the agent values future rewards compared to immediate rewards. A discount factor close to 1 means the agent values long-term rewards, while a value close to 0 means the agent focuses on immediate rewards.
- **Exploration rate (ϵ)**: In the ϵ-greedy strategy, this parameter controls the probability with which the agent chooses a random action (exploration) instead of the action with the highest Q-value (exploitation). It helps balance exploration of new actions with exploiting the known best actions.
- **Number of episodes (N)**: Specifies the number of training episodes or iterations over which the agent will interact with the environment and learn. More episodes typically lead to better learning but require more computation time.

Choosing the right set of hyperparameters is crucial for training an effective model. Poor choices of hyperparameters can lead to slow convergence, overfitting, or even complete failure of the algorithm to learn useful behavior.

2.7 Brief of DQN

2.8 Solving the Deterministic Environment

2.8.1 Deterministic Environment

In a *deterministic* environment, the outcome of any action taken by the agent is completely predictable. For a given state s and action a, the next state s' is always the same. The environment follows strict rules, meaning that there is no randomness involved in

the transition from one state to another. The reward associated with an action is also deterministic, meaning the agent always receives the same reward for performing the same action in the same state. The agent can rely on complete predictability and plan its actions with full certainty about the consequences. In deterministic environments, the Q-value can be updated directly based on the immediate reward and the known future reward, so α isn't needed. Then the update rule for the Q-function is:

$$Q(s, a) \leftarrow r + \gamma Q(s', a') \tag{2.3}$$

2.8.2 Learning function

The following algorithm shows the pseudocode of the learning function

Algorithm 1 Q-learning Algorithm (with ε-greedy)

Input: discount factor γ, exploration rate ε, number of episodes N
Initialize: Q-table $Q(s, a)$ for all states s and actions a, set episode count $i = 1$
Initialize: $\varepsilon_{min} \leftarrow 0.01$, $\varepsilon_{start} \leftarrow \varepsilon$, $k = 0.00001$
while $i \leq N$ **do**
    Initialize starting state s_1
    while not terminated or truncated s_t **do**
        Choose action a_t based on the ϵ-greedy policy:
        if random number $r \leq \epsilon$ **then**
            Select a random action a_t ▷ Exploration
        else
            Select action $a_t \leftarrow \arg\max_a Q(s_t, a)$ ▷ Exploitation
        end if
        Take action a_t, observe reward r_t and next state s_{t+1}
        $Q(s_t, a_t) \leftarrow r_t + \gamma Q(s_{t+1}, a')$ ▷ Update Q-Value
        $s_t \leftarrow s_{t+1}$
    end while
    $i \leftarrow i + 1$
    $\varepsilon \leftarrow \max\left(\varepsilon_{\min}, \varepsilon_{\text{start}} \cdot e^{-k \cdot (i+1)}\right)$
end while
Output: $\pi \leftarrow \{s : \arg\max Q[s] \mid s \in \text{range(for all states s)}\}$

Input parameters

The algorithm takes in input the discount factor (γ), the exploration rate (ϵ), the number of iterations (episodes) N.

- **Discount factor (γ)** is a value $0 \leq \gamma \leq 1$. A discount factor close to 1 means the agent values long-term rewards, while a value close to 0 means the agent focuses on immediate rewards.
- **Exploration rate (ϵ)** is a value $0 \leq \epsilon \leq 1$ It helps balance exploration of new actions with exploiting the known best actions.
- **Number of episodes (N)** is an integer value $N \geq 1$ It specifies the number of training with the agent that interacts with the environement

Q-table

In this environement instance the Q-table is represented as a matrix (2D array) composed by:

- $4x4 = 16$ rows representing the **states**
- 4 columns representing the **actions**

Then the final Q-Table is composed by $16x4 = 64$ elements In this implementation the Q-Table is zeroized.

2.8.3 Exploration and Exploitation strategy

The agent will balance the exploration and the exploitation, we need to use a very high ε parameter (closes to 1) to obtain an optimal policy.
During the iterations the algorithm will update ε value using the formula:

$$\varepsilon \leftarrow \max\left(\varepsilon_{\min}, \varepsilon_{\text{start}} \cdot e^{-k\cdot(i+1)}\right) \tag{2.4}$$

2.8.4 Output

In output the agent will return the policy learned, and the Q-Table updated.

2.8.5 Results

We learned three agents with different parameters, and collected cumulative rewards during the learning phase then the following are the results: From the 2.3 we can see:

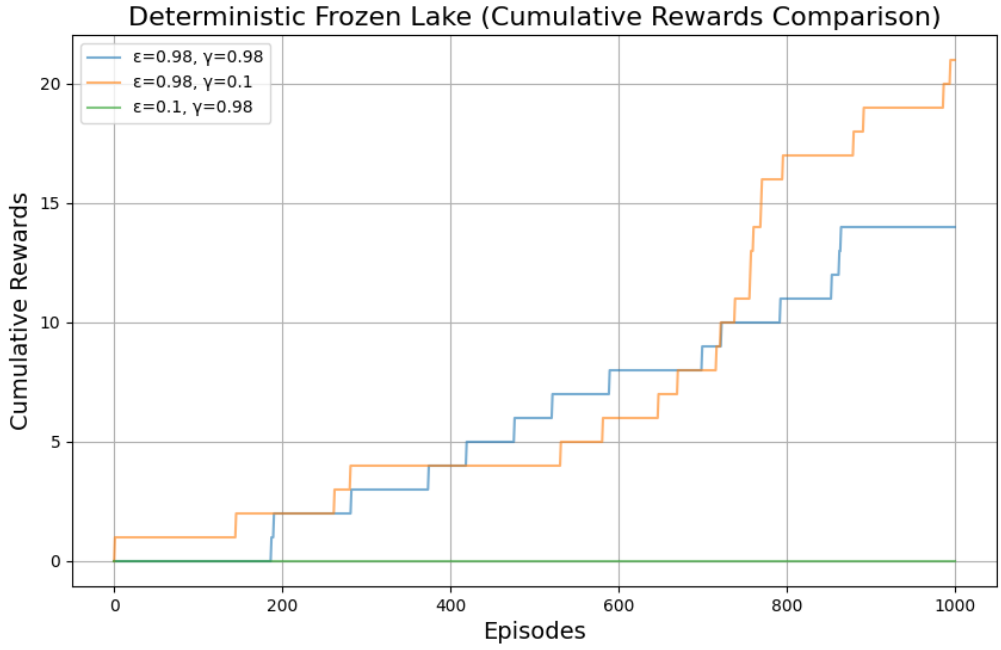

Figure 2.3: Agents comparison

- The green agent (with $\varepsilon = 0.1$) will not reach out the optimal policy, because the agent prefers exploitation instead of balance with exploration
- The orange agent has an high $\varepsilon = 0.98$ that allows the agent to balance exploitation and exploration. The second parameter $\gamma = 0.1$ indicates, that the agent prefers immadiate reward to future reward, and we can see the learning process start immediatly.
- The blu agent has $\varepsilon = 0.98$ that allows to balance exploitation and exploration. The second parameter $\gamma = 0.98$ indicates, that the agent prefers future rewards to immediate rewards, and we can see the learning process starts after some episodes.

We can see an highlight of the Q-Table learned by agents (blue and green agents) that help us to understand the final Q-Table learned by the two agents

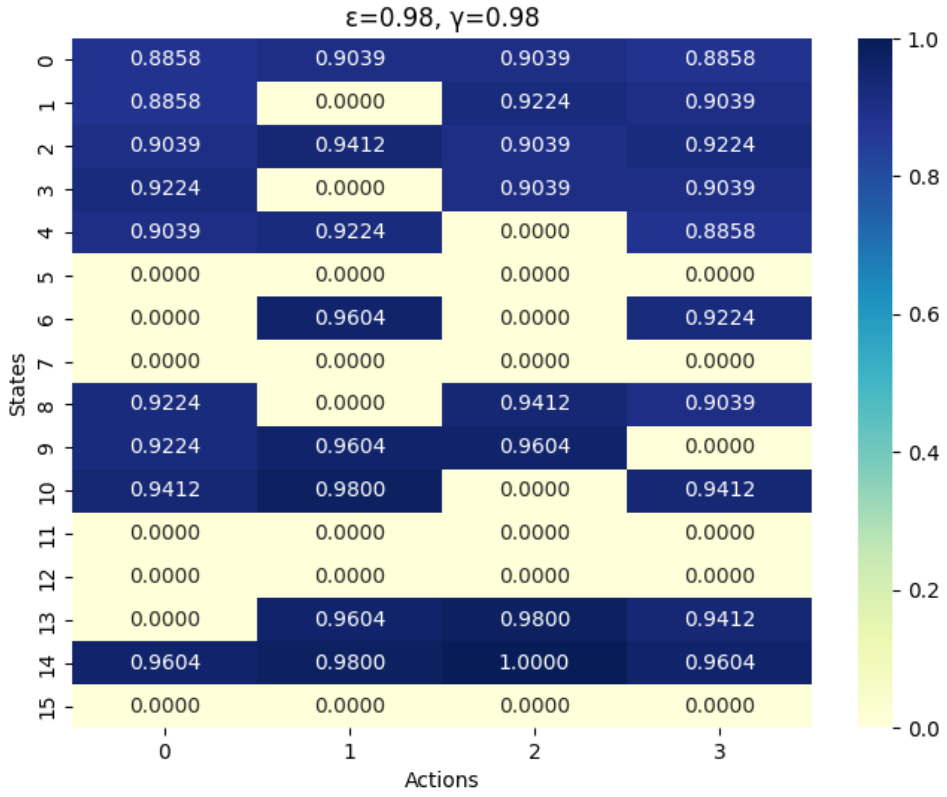

Figure 2.4: Q-Table of blue agent

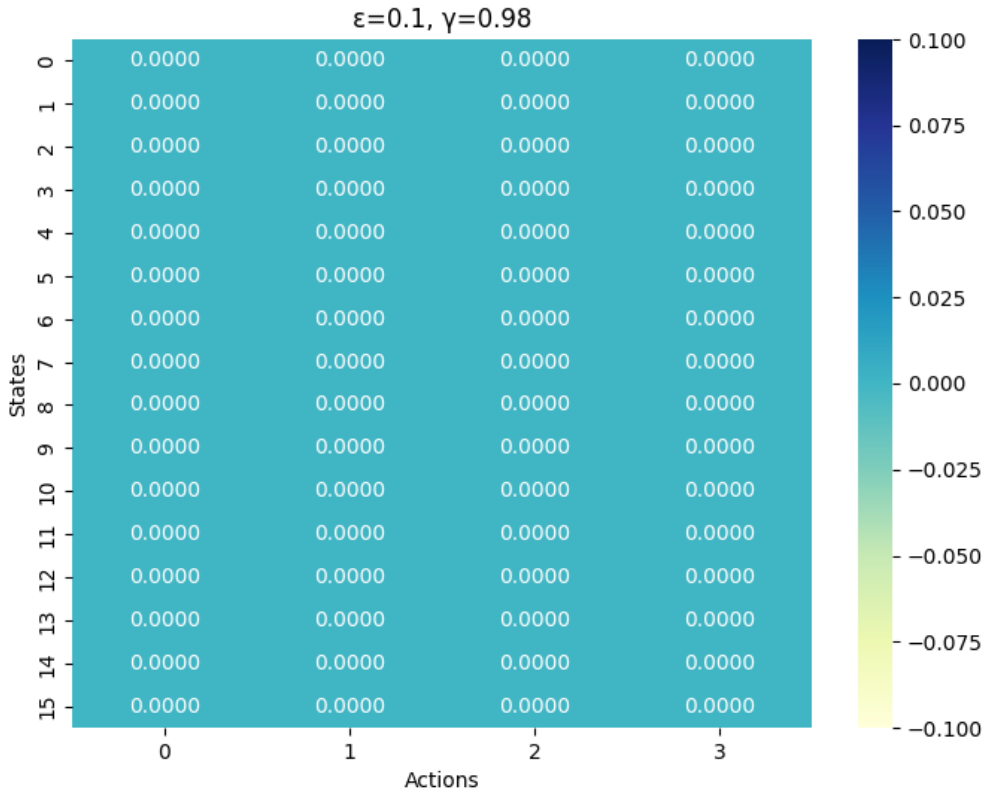

Figure 2.5: Q-Table of green agent

Using the policies learned by the agents is possible to see success rate that in a deterministic environement must be 100% In the figure 2.6 are shown the $\log(\frac{wins}{episodes})$ of the three agents:

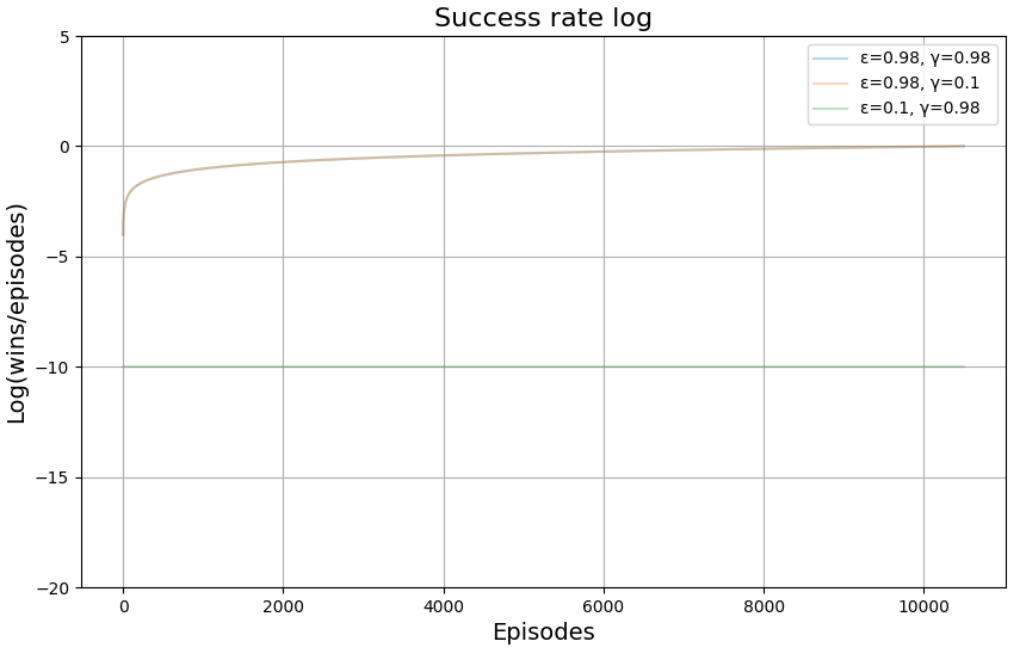

Figure 2.6: Success rate log

- The green agent has no rewards and for convenience is designed as a constant negative line, because it never converges to the optimal policy
- The other two learned agents are overlapping (you can see the brown curve that is the union of the blue and orange), and they converge to the optimal policy

2.9 Solving the Non-Deterministic Environment

2.9.1 Learning Function

The following algorithm shows how the agent is learned in a non-deterministic environment

Algorithm 2 Q-learning Algorithm (with ε-greedy)

Input: discount factor γ, exploration rate ε, learning rate α, number of episodes N
Initialize: Q-table $Q(s, a)$ for all states s and actions a, set episode count $i = 1$
Initialize: $\varepsilon_{min} \leftarrow 0.01$, $\varepsilon_{start} \leftarrow \varepsilon$, $k = 0.00005$
while $i \leq N$ **do**
    Initialize starting state s_1
    while not terminated or truncated s_t **do**
        Choose action a_t based on the ϵ-greedy policy:
        if random number $r \leq \epsilon$ **then**
            Select a random action a_t ▷ Exploration
        else
            Select action $a_t \leftarrow \arg\max_a Q(s_t, a)$ ▷ Exploitation
        end if
        Take action a_t, observe reward r_t and next state s_{t+1}
        $Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right)$ ▷ Update Q-Value
        $s_t \leftarrow s_{t+1}$
    end while
    $i \leftarrow i + 1$
    $\varepsilon \leftarrow \max\left(\varepsilon_{\min}, \varepsilon_{\text{start}} \cdot e^{-k \cdot (i+1)}\right)$
end while
Output: $\pi \leftarrow \{s : \arg\max Q[s] \mid s \in \text{range(for all states s)}\}$

Difference

In non-deterministic environment we have another input parameter the learning rate. It controls how much the Q-values are adjusted during each up-date. A larger learning rate leads to faster updates, but can also result in instability.

- α closes to 0 the learning rate is very slow, but the Q-Table will have stable values.
- α closes to 1 the learning rate is fast, but the Q-Table will have unstable values.

Another important difference is the Q-Function used, that in this case is 2.2

2.9.2 Results

Now we can see the results of three learned agents.

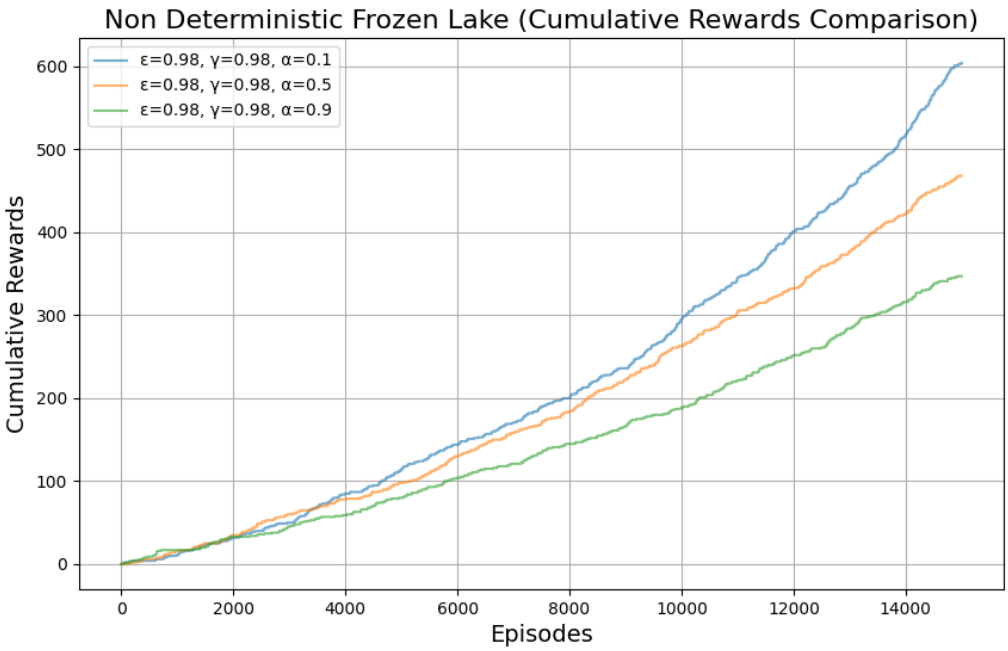

Figure 2.7: Agents comparison

In this analysis we are interesting to understand the learning rate parameter and how it changes the learning curve of the agents.

- The blu agent has $\alpha = 0.1$, this means that the agent will update with caution the new Q-Values, this causes a slow initial grows, but after a specific point, it will converge most speedly then each other to the convergence.
- The orange agent has $\alpha = 0.5$, and it represent the balanced situation between the two curves, it will increase most speedly at begin of the blue agent, but it will converge slowly the the blue agent, but speedly then the green agent
- The green agent has $\alpha = 0.9$, it will learn speedly to the begin, but this lead instability then it will reach the optimal policy most slowley then the other curves.

In the following picture are shown the Q-Table learned by the three agents:

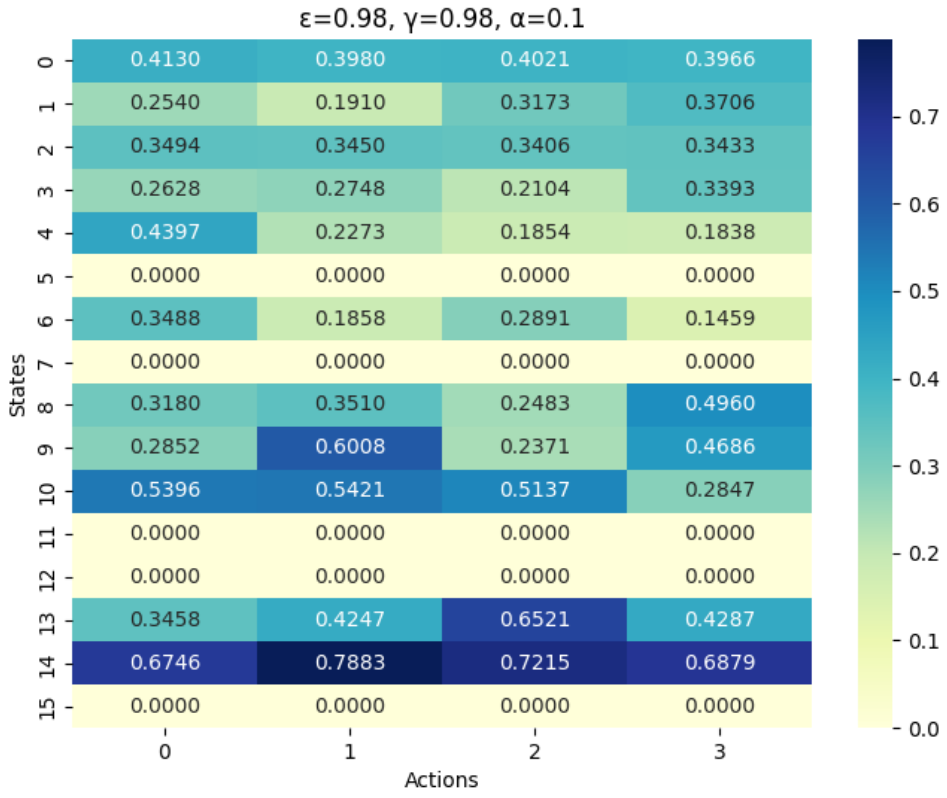

Figure 2.8: Q-Table of blue agent

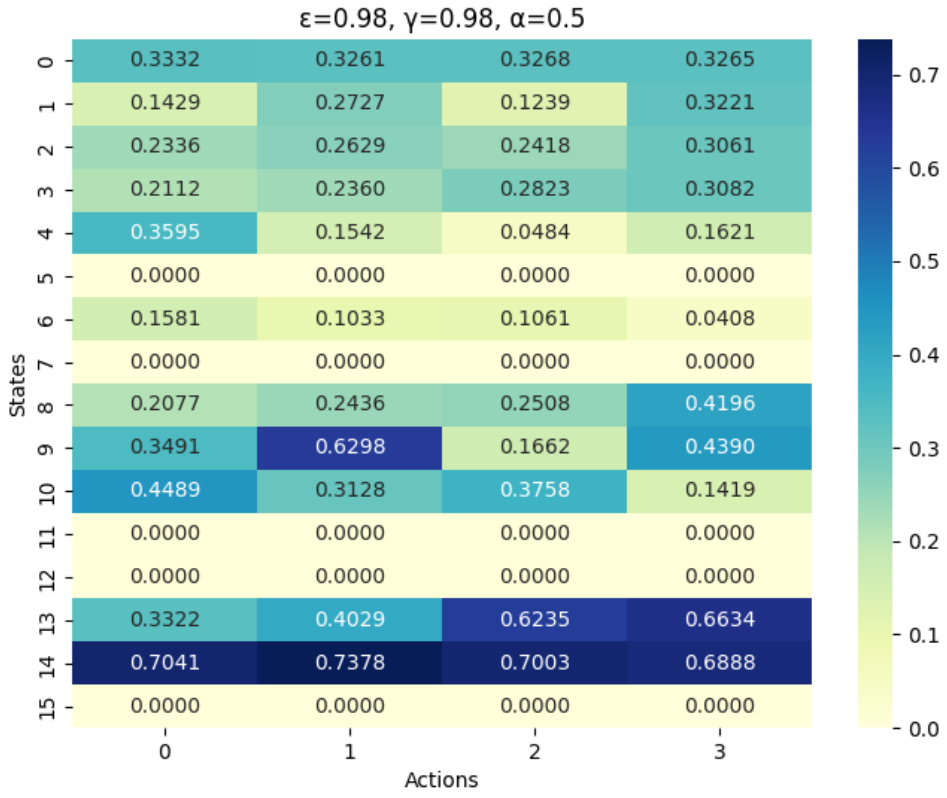

Figure 2.9: Q-Table of orange agent

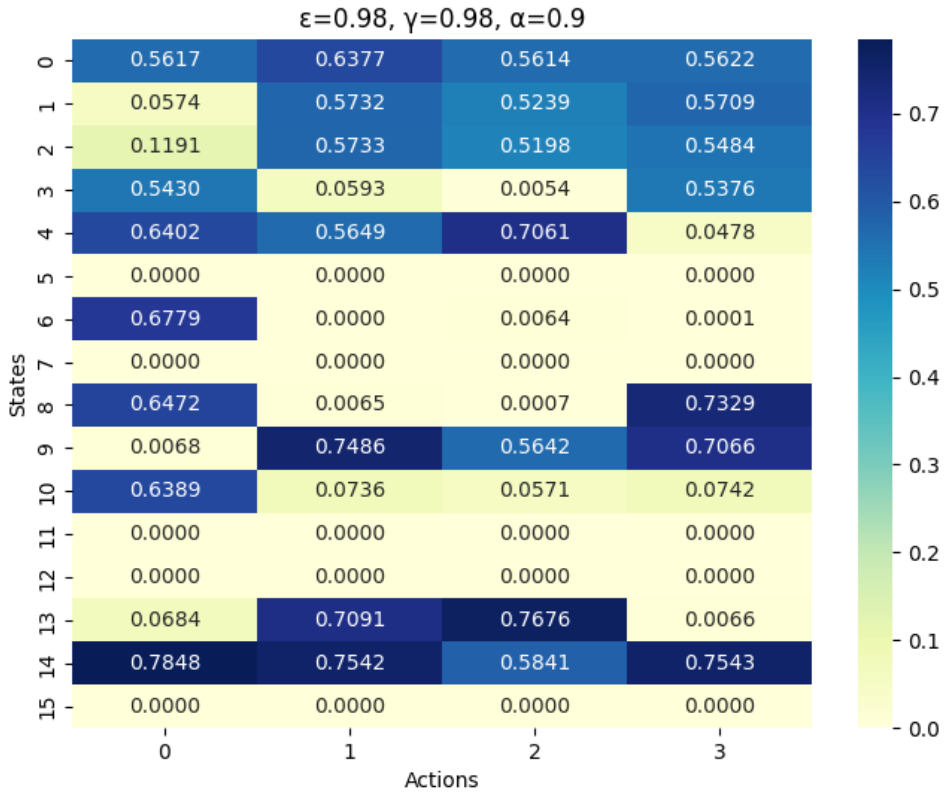

Figure 2.10: Q-Table of green agent

The obtained result end with the measure of the success rate. In the following figure is shown the overlap success rate of three agents. A very important fact is the agents learned an optimal policy, they do not have a success rate of 100%.

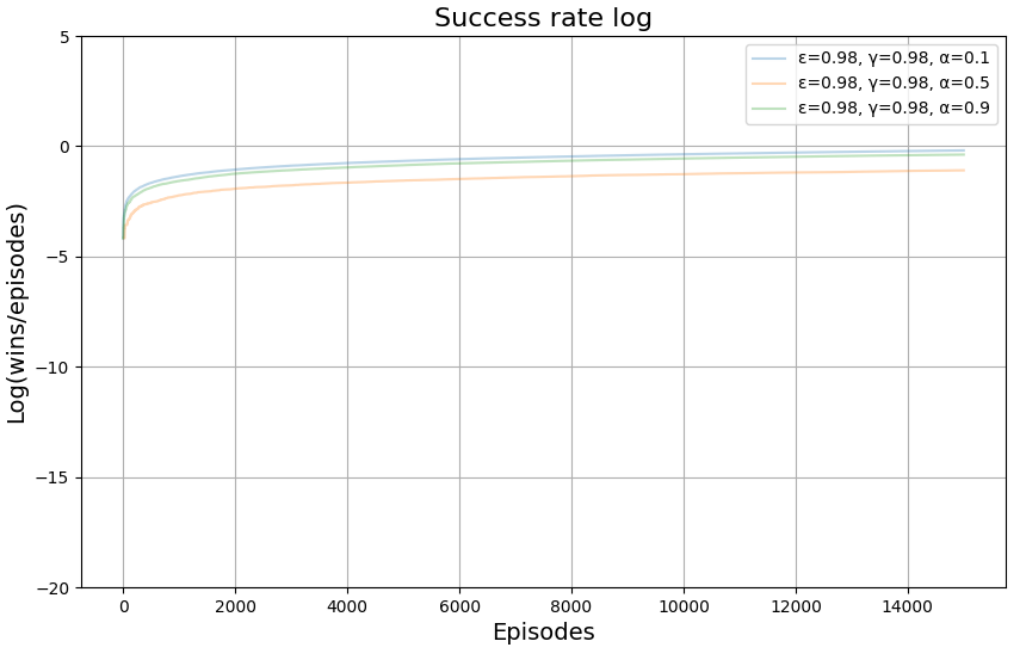

Figure 2.11: Success rate log