Assignment 1 – From Q-learning to Deep Q-learning (DQN)

Submitted by :

Gur TBD

Gal Benmaman 205633589

## Section 1 – Tabular Q learning

1. Value iteration is an algorithm that calculate the optimate state values V(s) by iteratively updating them using bellman optimality equation:

It means we need to have information about the probability of each state and reward based on the back state and action. If this information is lack, so the algorithm don’t work properly

1. Model-free methods like SARSA and Q-Learning don’t require a predefined set of rules or transition probability. Instead, they interact directly with the environment during training. As the agent explores various state-action pairs, it observes the resulting next state and reward, which it uses to estimate the Q-values (expected cumulative rewards) for each action. The learning process allows the agent to learn the optimal policy.
2. Both algorithms, SARSA an Q-Learning, are model-free methods, SARSA is an On-policy model, it updates the Q-Value based on the actions the agent actually takes according to the policy. While Q-Learning is an off-policy model, it updates the Q-value by the optimal action which received the highest reward. Q- learing will choose the best reward action, even if the probability of the action is very low (like winning a lottery prize- high reward with low probability)
3. The balance of exploration and exploitation by epsilon-greedy allows the agent to converge to the optimal action-state, (exploitation), but in other hand it allow the agent to explore new actions, which vary the current action to explore even better actions with highest reward (explore). if we use greedy search the agent is in risk to converge in local-optima action
4. Tbd – report of the script

What is the best KPI for the hyper parameter tuning?

Check about the version of frozen lake

Add figure of the state ( 4\*4 table with the lakes)

## Section 2 – Deep Q-learning

1. Sampling experiences in random order from the replay memory improves the generalization of learning. By using non-consecutive states , the agent learn to deal with varies states, otherwise the agent will be at risk to overfit the action of each sets of consecutive set of states
2. Tbd

the loss increases during the training

Do we need to compute the average reward of at least 475 by the train agent or test agent?

Tbd – report of the script

## Section 3 – Improved DQN

Tbd – report of the script