RL Coursework 2

Alona Rimon
CID 06007066
Department of Computing
MSc Advanced Computing

Question 1: Tuning the DQN

Question 1.1: Hyperparameters

Label	Hyperparameter	Value
A	Number of neurons in each hidden layer	128
B	Number of hidden layers	2
C	Learning rate	0.00025
D	Size of the replay buffer	10000
E	Number of episodes	300
F	Constant epsilon (for ϵ-greedy behaviour)	I used decay epsilon instead (see details below)
G	Reward scale factor	1
H	Batch size	64
I	Relaxation period - target network update frequency	100

Table 1: Hyperparameters

Changes in design I have applied:

1. **Optimizer** - I used AdamW optimizer instead of SGD.
2. **Weight decay** - I used weight decay coefficient $1e^{-2}$.
3. **Decay epsilon (exploration schedule)** - In order to stabilise and enhance convergence, Instead of using a constant epsilon, I applied decaying epsilon, calculated with:

   $$\text{epsilon} = \max(0.005,\ 1/\text{episode number})$$

   I used a decaying epsilon, proportional to the inverse of the episode number, to encourage more exploration at the beginning of the training and convergence as the training progressed. I limited this approach by setting a minimum value of 0.005 for epsilon to maintain exploration in later stages of training.

Question 1.2: Learning curve

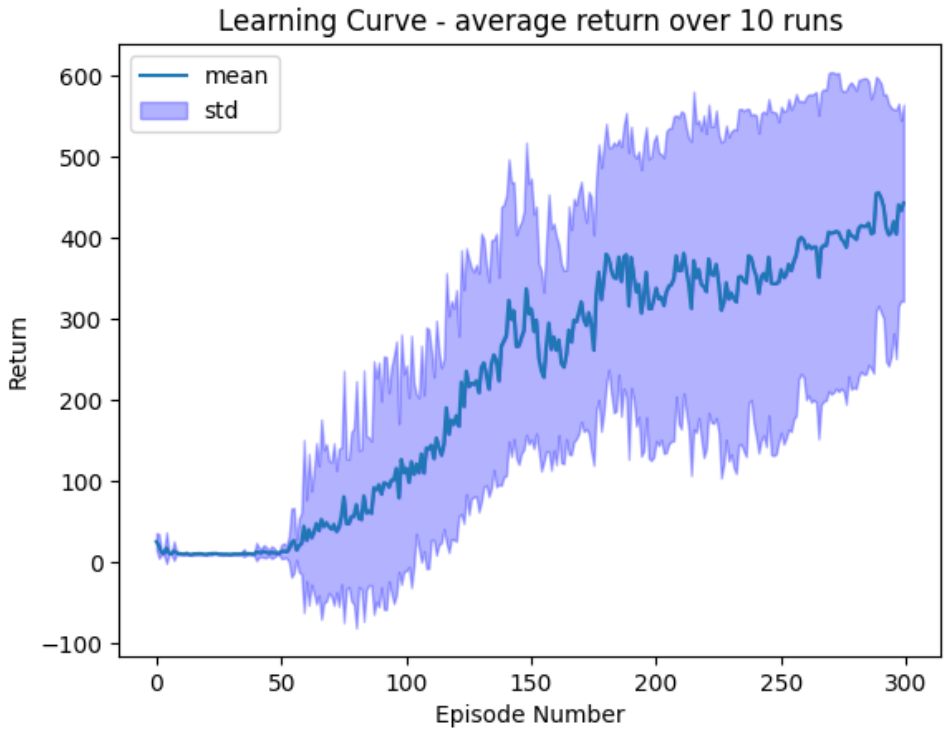


Figure 1: Learning curve showing average and standard deviation of return achieved over 10 training runs.

Since we receive a reward of +1 for each step the agent takes, the return value for each episode equals the episode length. We can see that in the first 50 episodes, the episode length is low, but the agent gradually reaches higher lengths. The learning curve is not smooth and shows high variance and low stability, which may be due to the sampling and exploration methods we are using.

Question 2: Visualise your DQN policy

Question 2.1: Slices of the greedy policy action

In the following figures, we present plots of the greedy policy decisions based on the Q-value estimates learned as described in Section 1. These plots show the decisions for states where the cart's position is fixed to zero (centre of the track), with four different fixed velocity values (one for each plot) and a range of pole angles and angular velocities. For each velocity, we observe different behaviours in the policy:

- **The regions of the plot where the agent chooses to push left or right**
  - **Expectation from an optimal agent**
    With velocity 0 - Since the position and the velocity are fixed at zero, I would expect the optimal policy to only push in the direction that will immediately stabilize the pole and contradict its tilt trajectory:
    1. If both pole angle and angular velocity are positive (i.e., to the right), the optimal policy should push to the right.
    2. If both are negative (i.e., to the left) the policy should push to the left.
    3. * If the angle and the angular velocity are in opposite directions (the pole is already moving toward the vertical stable position), the optimal agent policy will pick a balanced decision that depends on the specific values and the proportion between them. For example, with a strong left angle, the optimal policy would likely push left even if the pole's angular velocity is slightly positive, and similarly for the opposite case. If they naturally balance each other (the coordinates lying exactly on the decision boundary), both actions will have similar probabilities. Of course, since we use a greedy policy, the decision will eventually be deterministic.
    4. The point with 0 angle and 0 angular velocity (0,0) is a critical point where we expect that the policy will learn similar Q values for both actions and by that will have a similar probability for action in both directions (so it will take the arg-max choice out of two very similar and high values)
  - **My agent's learned policy** My agent's learned policy approximately matches the optimal agent's policy described above.
- **The general shape of the action decision boundary**
  - **Expectation from an optimal agent**
    We expect the general shape of the decision boundary to be a linear diagonal with a certain negative slope. Essentially, the agent's goal is to stabilize the pole by considering both the angle and the angular velocity. When both are in the same direction, the action will align to counteract the pole's fall. If they are in opposite directions (opposite signs), the relationship between the magnitudes of these factors determines the optimal decision, aiming to balance the two.*** This relationship analysis is elaborated in the paragraph above.**
  - **My agent's learned policy** The decision boundary of my agent's learned policy is piecewise linear, that is, approximately linear concerning the angle and angular velocity with a negative slope, aligning closely with the optimal policy.
- **The symmetries of the action decision boundary when the velocity is 0**
  - **Expectation from an optimal agent** When the velocity is zero, I would expect the optimal agent's decision boundary to be symmetrical around (0,0) with respect to the pole angle and angular velocity. This is because, with both position and velocity at zero, the agent's sole focus should be on manipulating the pole to stabilize it. The optimal actions to the left or right should be perfectly opposite and depend only on the relationship between

the angle and angular velocity. This is because the environment is symmetrical when the position and velocity are zero—pushing left or right will have the exact same effect but in opposite directions.

  - **My agent's learned policy** My agent's learned policy when velocity is zero is indeed symmetrical around (0,0) and therefore matches the expectations from an optimal one.

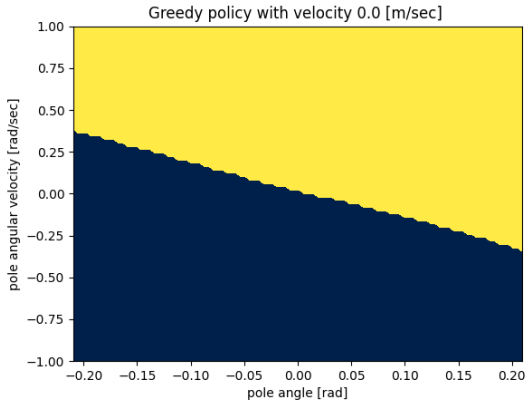
- **How the action decision boundary shifts as velocity increases**

  - **Expectation from an optimal agent** When the velocity is to the right, there is the risk of terminating at the right edge. So for an optimal agent- the long-term goal is to redirect the cart in order to stop steering right and avoid eventually reaching the right edge and terminating. However, due to the pole's and cart's inertia, if the agent pushes left without a sufficient leftward angle, the pole will tilt right and reach the termination angle. So to achieve both goals- steer left to avoid the wall and prevent the pole from falling to the right - the agent must induce a leftward tilt. Therefore, the immediate optimal action will be pushing to the right. As the velocity to the right increases, the agent will prefer pushing to the right even for high leftward angles and angular velocities. So, When the velocity to

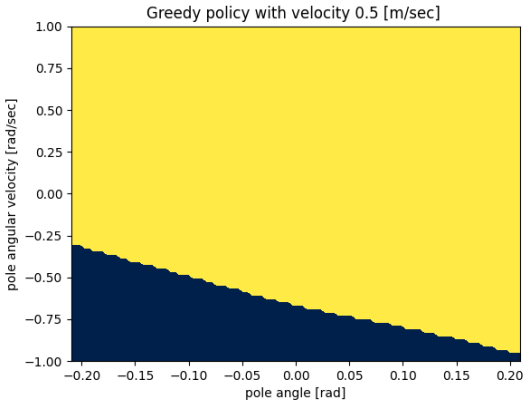
    the right increases, we expect the decision boundary to shift downward and leftward. This indicates that the policy increasingly favors pushing to the right—even in cases of a left tilt and pole movement. In other words, stronger leftward angles and angular velocities would be required to "convince" the agent's policy to choose to push left. This occurs because the situation is now asymmetrical around zero — there is a higher risk of termination at the right edge. This shift in behavior becomes more pronounced as the velocity to the right increases, due to both the system's growing inertia to the right and the heightened risk of termination at the right edge in the future.

    * I still expect that for near-terminating leftward angles and left angular velocities (in the leftmost corner of the grid), the optimal policy would favour pushing left rather than right, as pushing right would likely result in the immediate pole falling further to the left.

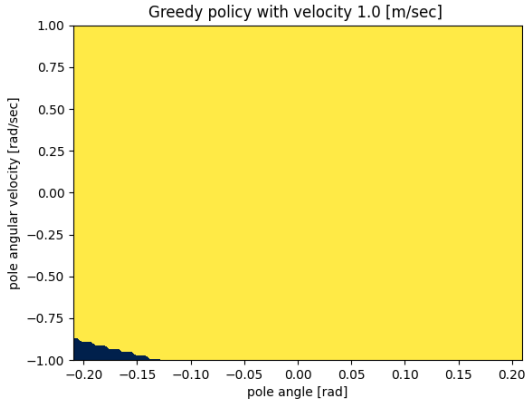
  - **My agent's learned policy** The decision boundary of my agent's policy is indeed shifting as described above, aligning with the optimal policy's decision boundary shift. However, there might be some difference in the extent of the shift - an overly enthusiastic adjustment. For example, at a velocity of 2 m/s, the policy might exhibit no leftward choices at all, even in cases of near-terminating left angles combined with strong leftward angular velocities. As noted above, for the leftmost corner of the grid (with an almost terminating left angle and strong left angular velocity) My agent's policy (pushing right) may be suboptimal and result in the pole reaching the left terminating angle.



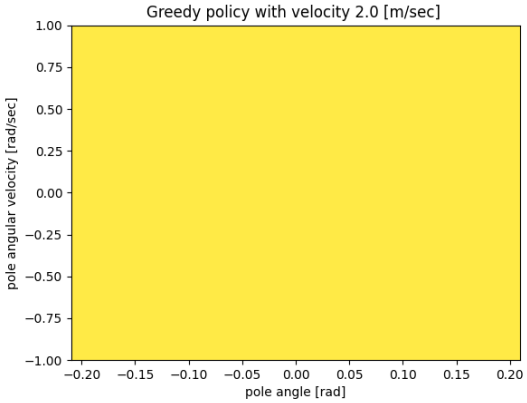
(a) Greedy policy with velocity 0.0 [m/s]



(b) Greedy policy with velocity 0.5 [m/s]



(c) Greedy policy with velocity 1.0 [m/s]



(d) Greedy policy with velocity 2.0 [m/s]

Figure 2: Greedy policy actions with respect to angle and angular velocity. In each figure, velocity is fixed to the specified value and position is fixed to center (zero).

Question 2.2: Slices of the Q function

In the following figures, we present plots of the greedy Q-value estimates learned as described in Section 1. These plots show the maximal Q value estimate for states where the cart's position is fixed to zero (centre of the track), with four different fixed velocity values (one for each plot) and a range of pole angles and angular velocities.

- **The regions of the plot where values are relatively higher or lower**
  - **Expectation from an optimal agent** Since the reward is +1 for each step, the return reflects the length of the episode. Therefore, the action-value function for each state represents the estimated duration the agent can maintain balance and position without termination, starting from the current state.

    For each velocity, we expect varying Q values corresponding to the stability of the system in each state: There are expected to be generally high values in places where the risk of the episode terminating soon is the lowest.

With velocity 0, The lowest values are expected in the up-right and bottom-left corners - where the angle and angular velocity are both strong in the same direction, (the angle is close to the termination angle ±12 or approaching it with high angular velocity). The values are expected to be highest when both the angle and angular velocities are small and remain high even as they increase, as long as it is in opposite directions and they naturally balance each other. These Q values indicate a lower chance of termination when the pole's angle is small or decreasing (with opposite-direction velocity).

- **My agent's learned values** My agent's relative Q values are approximately aligned with the optimal Q values I just described.

- **The range of values that the agent has learned**

  - **Expectation from an optimal agent** Since the Q-values represent the expected length of the episodes - e.g. estimated duration during which the agent avoids termination (as explained above), they must always be **non-negative.** I expect the value to vary according to the location on the grid (as described above) with low values (can be close to zero) for almost terminating angles and angular velocities and very high estimated return (unbounded, aspire to infinity) for stable areas (those change as the velocity increases as described below). As the velocity to the right increases and reaches 2 m/s, we expect the maximum Q value to decrease, reflecting the growing challenge of avoiding the right edge while balancing the pole.
  - **My agent's learned values** The range of the learned Q-values observed is between 300-860. The relatively high lower bound of the range, even for nearly terminating angles, suggests potential overestimation (maximization bias) during learning, where the approximator predicts higher returns than are realistic.

    * Notice - Episodes always terminate within 500 steps during training. However, the learned Q values represent the estimated return, so it makes sense that they are not bounded by the training episode length, indicating generalization to cases beyond the empirical return encountered during training.

- **The symmetries of the learned values when the velocity is 0**

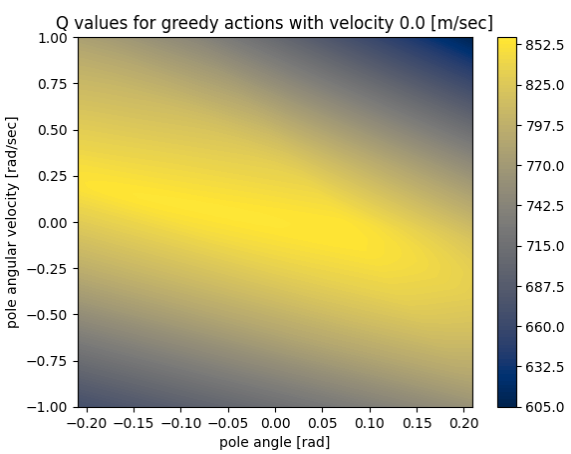
  - **Expectation from an optimal agent** When the velocity is zero and the cart's position is centred, the environment is symmetrical with respect to angle and angular velocity. In this case, we expect the Q-values learned by the agent to also exhibit symmetry around (0,0) with respect to the angle and angular velocity, reflecting the environment's inherent symmetry. Symmetric Q values around (0,0) represent balanced decision-making. High Q-values should correspond to stable states, representing a higher estimated episode length. These values should depend solely on the relationship between the angle and angular velocity, as described earlier in this section.
  - **My agent's learned values** The learned Q-values of my agent are indeed symmetrical around (0,0) and align closely with the values expected from an optimal agent.

- **How the values change as velocity increases**

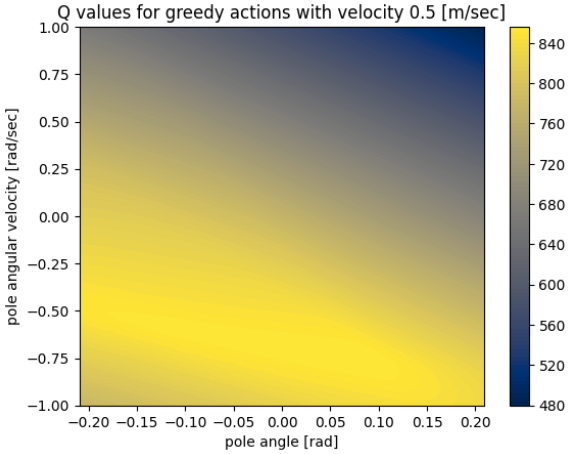
  - **Expectation from an optimal agent** The distribution of the Q-values should no longer be symmetrical around (0,0) as the environment loses its symmetry when the velocity to the right increases. In this scenario, the agent's Q-value distribution is expected to shift leftward and downward. These values now reflect the recovery potential from the increased risk of termination at the right edge while still balancing the pole (the agent's ability to steer the cart in the opposite direction without causing the pole to fall to the right). As a result, tilting left provides a better chance of recovery (explained in detail in 2.1), causing the highest Q-values to shift leftward and downward instead of remaining balanced around (0,0). So, we expect the Q values distribution to shift, with higher values estimated for states where the pole tilts and moves left, while lower values for the opposite. When the

velocity is very high to the right, we expect that the Q values will be generally low in all of the grid, with a slight increase in the far down-left corner (and a decrease in the top right). As the velocity to the right increases, the chance for recovery decreases. So, we expect to get higher Q values when the pole has a leftward angle and angular velocity, providing a relatively better chance for recovery, but still lower than what we got in the velocity 0 situation - reflecting the challenging scenario.

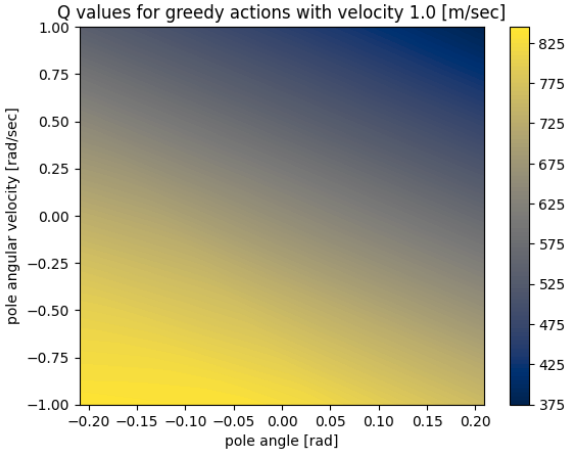
– **My agent's learned values** My agent's learned values trend generally aligns with the optimal value estimation described above, However, it might be overestimating (presenting maximization bias) - we see that my agent estimates at least 335 return value even with velocity 2 m/s to the right, and strong right angle and angular velocity.



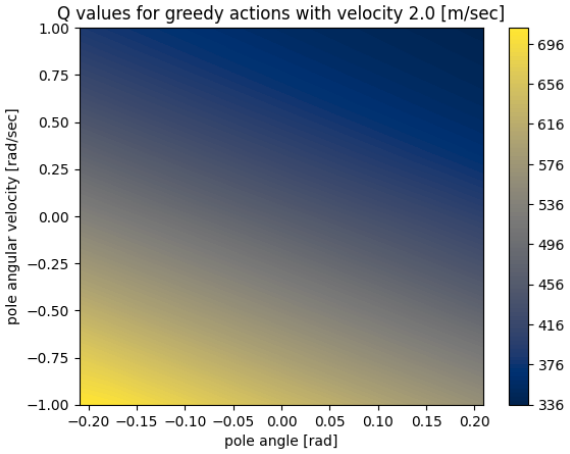
(a) Q values with velocity 0.0 [m/s]



(b) Q values with velocity 0.5 [m/s]



(c) Q values with velocity 1.0 [m/s]



(d) Q values with velocity 2.0 [m/s]

Figure 3: Greedy (maximal) Q values with respect to angle and angular velocity. In each figure, velocity is fixed to the specified value and position is fixed to the centre (zero).