**Results**

The results of the DQN, where all previous extensions have been combined, can be seen in table 1 below. In the table, these results are compared to previously obtained results for the previous extensions. For each of the extensions, the agent was trained 10 times. For each of the 10 training rounds, the trained agent was tested by running 10 episodes. The results of the tests were evaluated based on three metrics:

1. *No. Times converged:* Out of the 10 training rounds, this is the amount of rounds where the model converged. The model was considered to have converged once the mean return of the 100 latest rewards was > 195.
2. *Convergence time:* If, during a training round, the model converged, this is the number of episodes that passed before the model converged. This is counted from the start of the 100 episodes whose mean was > 195. For example, if episodes 43-143 are the first 100 consecutive episodes with a mean of > 195, the convergence time is 43 episodes.
3. *Return per episode:* For each training round, the return (cumulative reward) of each episode was averaged. The cumulative reward of an episode is equal to the length (number of steps) of that episode. The episode ends if the pole falls over, the cart moves too far to one side, or if the maximum episode length of 200 steps is reached. This means that the maximum return of an episode is 200, and anything lower means that the agent failed the task of balancing the pole on top of the cart.

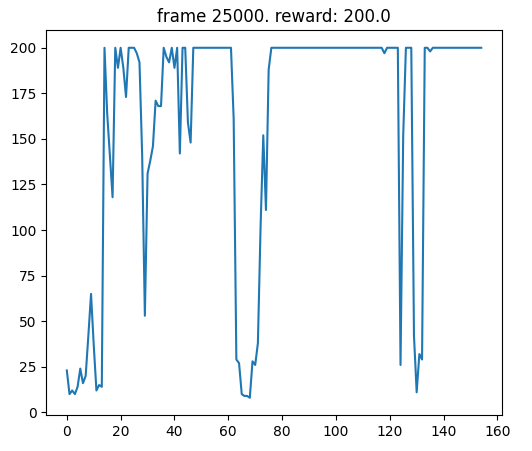
**Table 1.** The table shows the results from training the agent 10 times. The table includes the total amount of times the model converged, as well as the means and standard deviations of the convergence time and returns per episode, computed across the 10 training rounds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **No. Times converged** | **Convergence time – Mean**  **(No. episodes)** | **Convergence time – Standard deviation**  **(No. Episodes)** | **Return per episode - Mean** | **Return per episode – Standard deviation** |
| Base DQN | 1 | 43.0 | N/A | 193.1 | 13.0 |
| Double-Q DQN | 2 | 37.5 | 10.6 | 190.3 | 20.2 |
| n-step DQN | 2 | 20.0 | 5.7 | 199.6 | 1.4 |
| Dueling DQN | 0 | N/A | N/A | 162.7 | 66.3 |
| Base DQN with prioritized experience replay | 2 | 25.0 | 7.1 | 185.6 | 17.2 |
| DQN with all extensions combined | 0 | N/A | N/A | 196.3 | 11.9 |

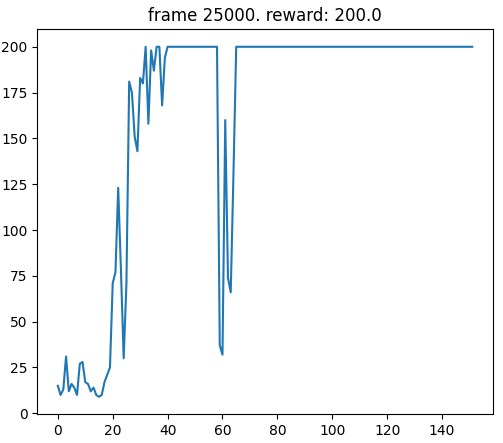
**Discussion and conclusions**

As mentioned in the previous reports, most of the DQNs appear stable and effective, yielding high mean returns per episode. The main exception is the Dueling DQN, which could be due to it being overly complex for the relatively simple Cart Pole environment. However, due to lack of data, these conclusions are uncertain. More, and possibly longer, training rounds would have to be performed before more certain conclusions can be drawn. This is especially true regarding the convergence time.

All of this is also true for the DQN with all extensions combined. As seen in table 1, it has a mean return per episode of 196.3, which is near the maximum 200 and second only to the n-step DQN. Thus, it appears to be an effective model, inheriting the strengths of the extensions that were included. However, figures 1 and 2 below, along with the fact that it is the only model along with the Dueling DQN to never meet the convergence criterium during the 10 training rounds, indicate that it is still somewhat unstable. This could be a sign of it being a bit too complex, possibly due to the inclusion of the Dueling DQN extension.



**Figure 1.** The graph shows an example of the rewards over time for a given training round. This particular example was generated when training the DQN with all extensions combined. It shows the average reward of the last 10 episodes, plotted over time, for the duration of that training round.



**Figure 2.** The graph shows the rewards over time, like in figure 1, but for a training round for the n-step DQN.