**Assignment ATAI Part I –**

**N-Step and Double-Q DQN**

Fredrik, Leon, Siebe, Dimi

**Deep Q-Network**

In this assignment a n-step Deep Q-Network is implemented. To be able to understand this network it is important to first understand how a normal Deep Q-network works. The Deep Q-network learns an optimal policy by combining neural networks and Q-learning for reinforcement learning. This neural network learns the Q values associated to each state/action pair by calculating the expected reward based on a mini-batch of the dataset. This mini-batch is selected randomly for every iteration and consists of state-action values (the predicted variables) and the ‘true’ Q-value (the target). As we don’t know the true Q-value we use an approximation taken from Q-learning which takes the first n immediate rewards and the estimated Q value from step n onwards. In a basic Deep Q-network only the first immediate reward is used whereafter the Q value of the action that achieves the highest long-term reward is added:

So for every iteration, a combination of the predicted variables and the target out is taken from the training dataset which is used to update the weights of the neural network Q(S,A) with the following formula:

\bold{w} = \bold{w}  + \; \alpha \;
\sum_{t \in B}  \frac{\partial \hat{q}_\bold{w}}{\partial \bold{w}}(S_t,A_t) 
[r_{t} + \gamma max_{a'}\hat{q}(S_{t+1},a',\color{green}{\bold{w'}}) -  \hat{q}(S_t,A_t,\bold{w})]\; 

In here a separate neural network is used for the target to make sure that the weights we are trying to learn are not part of the training formula and new policies are added to the experience using the epsilon-greedy policy to make sure that there is some exploration.

**Double Q-Learning in DQN**

The Double Q-learning is an extension of normal Q-learning, which solves the problem of the maximization bias, which comes from the overestimating of values in normal Q-learning. The problem is solved using two different estimates, one for the maximum action, and one providing the estimate of its value.

Another problem of vanilla Q-learning is the **moving target problem**. It appears when the policy being is constantly changing as the agent updates its estimates of the Q-values. This leads to problems in the convergence process and makes learning unstable.

All of those problems are fixed in our Double Q-learning in the DQN approach. The difference from normal Double Q-learning is that we use two different networks to determine the next action and the value.

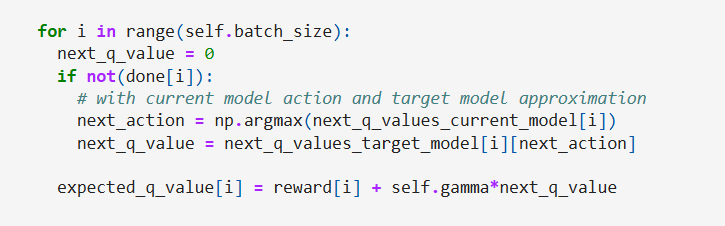
The formula:



Like shown in the formula we determine the next action by using the older network with weights w´to get the next action and the current network with weights w to get the value. Like this we can fix the moving target problem and the maximization bias problem.

**Changes in Code for Double Q-Learning:**

To implement this method we used the already existing solutions of Deep Q-Network. We figured out that the solutions for the Double-Q approach were already implemented as a comment. We commented out the calculation of the next\_q\_value for DQN and uncomment the part to calculate the next action and estimates for Double Q-Learning in DQN. The code is in the compute\_td\_loss\_our(self, target\_model) function:



**N-Step Deep Q-Network**

The n-step Q learning method uses almost the same steps as described above, only changing the target. As mentioned above it takes the first n discounted immediate rewards and the (also discounted) estimated Q value for the best action in state S(t+n) which can be seen in the following formula:



It therefore must save in the dataset not only the first reward and the next state, but the sum of discounted n rewards and the next state at step n.

**Changes in Code for N-Step Deep Q-Network**

For our implementation of n-Step DQN, we used the same base as provided in the lecture and modified almost exclusively the ReplayBuffer class:

In this first part, we allow for the n\_step and gamma parameters to be given when initializing a ReplayBufferNStep object.

A screen shot of a computer

Description automatically generated

Then we modified the “push” method to allow for the “truncated” attribute to be passed. When push is executed, we first add the experience to the n\_step\_buffer deque. After that we check if the buffer is full or if we have encountered a truncated or done state. While technically in reverse order, we will first explain what happens when the buffer is full:

If the buffer is full, we will calculate the n-Step version of the first experience in the n-step buffer. We keep the state and action, but we replace the reward with the discounted sum of rewards – calculated by the \_calculate\_n\_stop\_return( ) method. This method also gives us the updated next state and done attribute.

If we encounter a true state for the truncated or done attributes, we immediately process all remaining experiences in the n-step buffer (until it is empty), before the environment is reset in the DQN training method.

A screenshot of a computer code

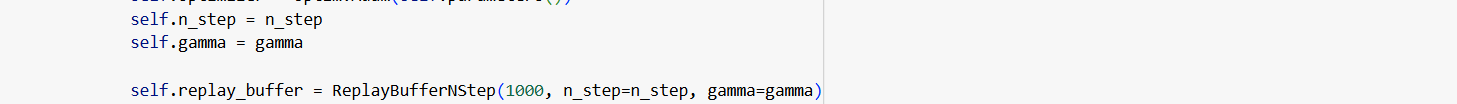
Description automatically generated

Finally, the calculation of the discounted return. We first set the reward, next\_state and done attributes to default values. Then we go through the experiences saved in the n-step buffer and for each add the rewards with corresponding discount (so first element will be ( 1 \* reward ), second will be (gamma \* reward), third will be ( (gamma \*\* 2) \* reward ) etc. We also constantly update the next\_state and done attributes, so we automatically have the correct values after we process all experiences. Then if we encounter a terminal state, we make sure to not continue adding rewards after it. At the end, we return the discounted sum of rewards and updated next\_state and done attributes.

A screen shot of a computer

Description automatically generated

The last changes are updating the ways in which the replay buffer in the DQN class is created (include n-Step and gamma) and the push method for it handled (include truncated when calling):





**Results**

For each of the methods, 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.

The results of the tests can be seen in table 1.

**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.

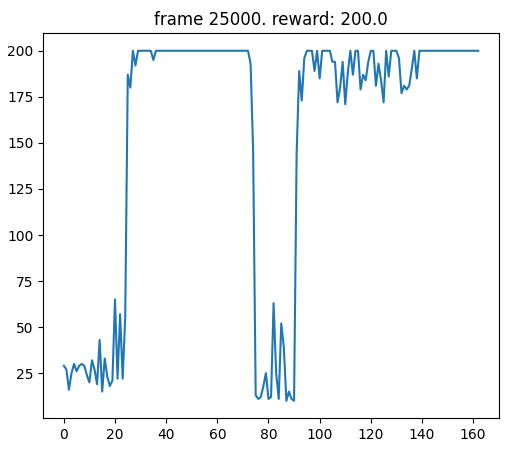
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **No. Times converged** | **Convergence time – Mean**  **(No. episodes)** | **Convergence time – Standard deviation**  **(No. Episodes)** | **Return per episode - Mean** | **Return per episode – Standard deviation** |
| Base DQN | 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 |

**Discussion**

Table 1 shows that for all methods, the model only converged 1, 2 and 2 times respectively, out of a possible maximum of 10. This makes it difficult to draw any conclusions from the convergence time of the different models, although the significantly lower mean and standard deviation of the n-step convergence time could indicate that the n-step DQN is trained more efficiently. Regardless, the models have mean returns per episode that are near the maximum of 200, indicating that they still performed well and comparatively equal. Since the n-step DQN had a higher mean return (almost exactly 200, i.e. the maximum), as well as a significantly lower standard deviation of the return, it could be considered more stable and effective than the other two models. Since only 10 rounds of training and testing were performed however, these conclusions are somewhat uncertain.

To be able to draw more certain conclusions, more rounds of training should be performed. Additionally, each round of training could be extended to cover more episodes. This would increase the likelihood of convergence and make it easier to draw conclusions from the convergence time of the different models. This would, however, require more time and computing power than what was available when evaluating the models.

Additionally, another reason to believe that the likelihood of convergence would be increased by extending the training rounds can be seen in figure 1 below. The model seemingly converges quite quickly, but after a while a sudden, sharp drop in the reward appears. This is likely due to the ε-greedy behaviour policy used during training, which causes the agent to sometimes take non-greedy actions that yield a significantly lower reward. Since ε decreases over time however, the likelihood of taking non-greedy actions also decreases over time. Thus, extending the training rounds would increase the likelihood of having 100 consecutive episodes in which the mean reward is not significantly affected by the agent taking non-greedy actions. Therefore, this would also increase the likelihood of the model converging according to the convergence metric mentioned above.



**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 Double-Q DQN. It shows the average reward of the last 10 episodes, plotted over time, for the duration of that training round.

**Conclusion**

Of the three DQN methods, the n-step DQN seems to converge faster, more effectively, and yield a more stable agent. 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.