AE598 HW2 Deep Q Network Report

Mikihisa Yuasa
Department of Aerospace Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801, USA
Email: myuasa2@illinois.edu

1. Introduction

In this homework, deep Q network (DQN) [1] algorithm is applied to Discrete Pendulum environment. In the following sections, the resulted plots of the algorithm and evaluations are discussed.

2. Methodology

For the implemented DQN algorithm, following hyperparameters are chosen: the bath size is 128, memory buffer size is 10,000, starting value of ϵ is 0.9, final value of ϵ is 0.05, decaying rate of ϵ is 1,000, discount factor γ is 0.99, and update rate of the target network is 0.005.

There are four different agents to be evaluated: one with both the replay memory and target network, one without the replay memory and with the target network, one with the replay memory and without the target network, and the other without the replay memory and target network. Each of these agents is trained for 10,000 episodes, and AdamW algorithm [2] is used as the optimizer for each, whose learning rate is 1×10^{-4}.

3. Results

For all the four agents, the learning curves are plotted to track how they learn the environment and for the ablation study as shown in Fig. 1. For the readability, the plots are smoothed for 1,000 episodes.

In addition, the policy of the agent with the replay memory and target net is plotted as in Fig. 2.

Also, Fig. 3 is the value function of the agent with the replay memory and target net.

Fig. 4 is a sample trajectory of the agent with the replay memory and target net

In Table 1, the averaged episodic rewards are shown for each agent after running 1,000 episodes for the ablation study.

4. Discussion

As shown in Fig. 1, the learning curve of the agent with replay and target and one without replay and with target show lower variances and higher averages during the training when compared to the others.

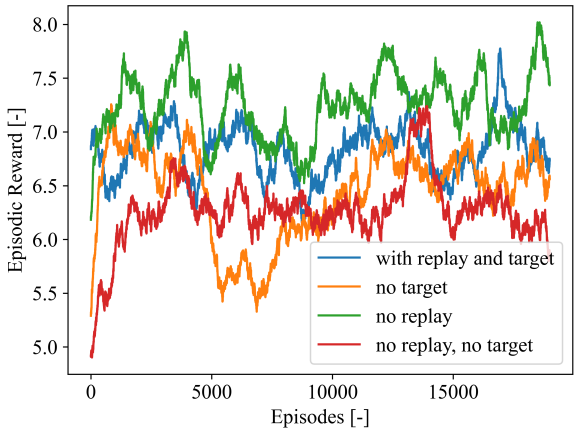

Figure 1: Learning curves of the four agents for the ablation study.

TABLE 1: The effects of replay and separating the target Q-network. Each agent runs 1,000 episodes, and the episodic rewards are averaged

With reply, with target Q	With reply, without target Q	Without reply, with target Q	Without reply, without target Q
87.335	-37.219	-18.058	87.754

In Fig. 2, the torque value is in a gradation at $(0, 0)$ position. This implies that, when the pendulum is closed to the upright position, there is more nuanced torque maneuver to make it upright. Fig. 3 marks the highest value at the $(0, 0)$ position, which is consistent with the environment setup which gives reward of 1 when the pendulum is upright. Fig. 4 shows that the pendulum vibrates around the upright position, and this suggests the agent learning the policy from the environment.

In Table 1, the agent with replay and target and one without replay and target have the highest averaged episodic rewards, and it suggests these two agents learned the policy better that the other two. This result, however, contradicts with that suggested by Fig. 1: the agent with replay and

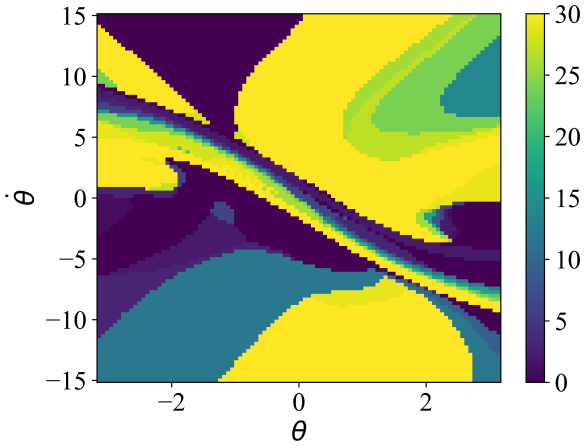

Figure 2: Policy of the agent with the replay memory and target net. The plotted value is τ

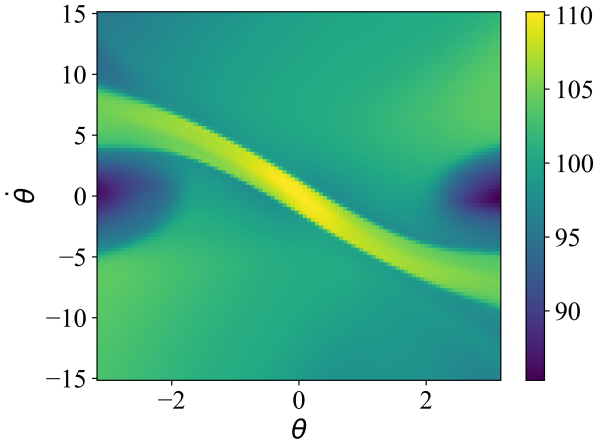

Figure 3: Value function plot of the agent with the replay memory and target net.

target shows consistent performance both in Fig. 1 and Table 1, but the agent without replay and with target and without replay and target show their performances only in Fig. 1 and Table 1, respectively. This contradiction can be solved by training the agents with different hyperparameters or averaging the rewards over more trained agents. For this homework, due to the limitation of the available computational resources, only one set hyperparameters is used, and the agents are trained only once.

5. Conclusion

Based on Section 4, the agent trained with the replay memory and target Q network consistently shows that it learned the policy to make the pendulum upright. The results also suggest that more numbers of training and trials

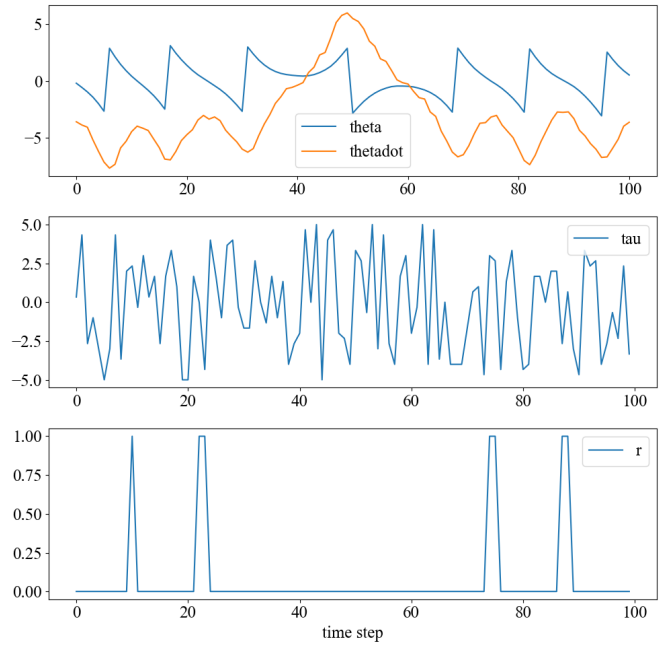

Figure 4: Sample trajectory of the agent with the replay memory and target net.

with different hyperparameter set will be helpful to draw stronger conclusion for the performances over the agents with different replay and target configurations.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540.

[2] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Jan. 2019. arXiv:1711.05101 [cs, math].