Homework 2

Pan Li
CSCI-GA 3033-090: Deep Reinforcement Learning

October 2, 2020

Solution 1. *Here are some descriptions of my solution for Part 1: I have made significant changes on the starter codes and make them work for behavioral cloning. To generate the training data, I have played the game for a very long time (1̃2 hours) and set different step thresholds to obtain different amounts of car racing samples. I have attached a screenshot of the training samples. After that. I take the collected data to train the behavioral cloning agent. I have also attached a screenshot of the generated actions, as well as the sample-performance plot. To improve the performance of the agent, I have designed a delicate deep learning model based on CNNs. The model takes the states as input and feed the into a convolutional layer with a subsequent max-pooling layer. I then flatten the output and feed them to another fully connected layer with 40 units and the dropout mechanism to avoid overfitting. In the end, I feed the output to a fully connected layer to predict the final predicted action. I have attached a screenshot of the training loss-epoch plot. I have also implemented the DAGGER algorithm using a teacher action network, and have obtained even better performance as shown in the Figure 3. The codes can be found in the zip file.*

Figure 1: A Screenshot of the Training Data

Figure 2: A Screenshot of the Generated Action

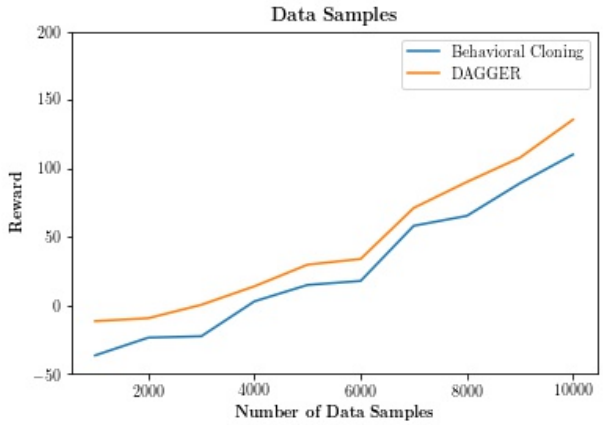

Figure 3: Plot of Relations between Number of Samples and the Performance Reward (Behavorial Cloning and DAGGER)

Solution 2. *Section 2.1: My guess is that the best policy will bet all when you have a capital of 50 and the possible dividends of it, like 25, 12/13, etc.. Note that you can choose whether to include only 50 as the dividend or to include all the dividends, thus creating an entire family of possible solutions.*

Section 2.2: The problem has an optimal policy solution, because at the number 50 when you bet everything, you can win the game with probability p_h, thus it will be optimal to bet everything at this number. However, at the case of 51, we need to figure out the way to obtain additional reward from the extra 1 dollar. If the return is positive, we can bet the excessive amounts over and over again till we reach 75; if the return is negative, we need to reach 25 if we lose the bet, and it is a much worse condition.

Section 2.3: The plots have been attached. I have tried different θ values ranging from 1e-3 to 1e-8, and the results are stable.

Section 2.4:

$$q_{k+1}(s,a) = E[R_{t+1} + max_{a'}\gamma q_k(s',a')] = \sum_{s',r} p(s',r|s,a)[r + max_{a'}\gamma q_k(s',a')] \qquad (1)$$

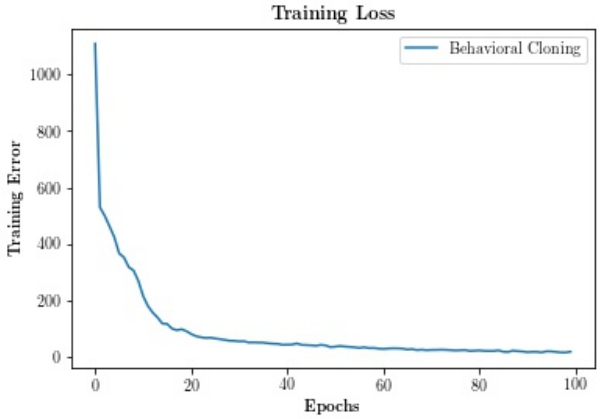

Figure 4: Plot of Relations between Training Epochs and Training Loss

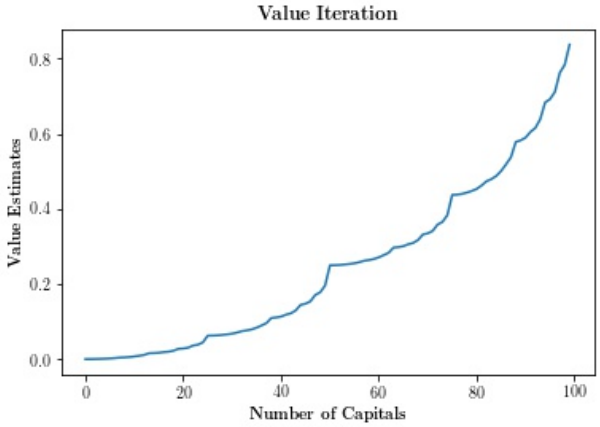

Figure 5: Plot of Value Iteration when ph=0.25

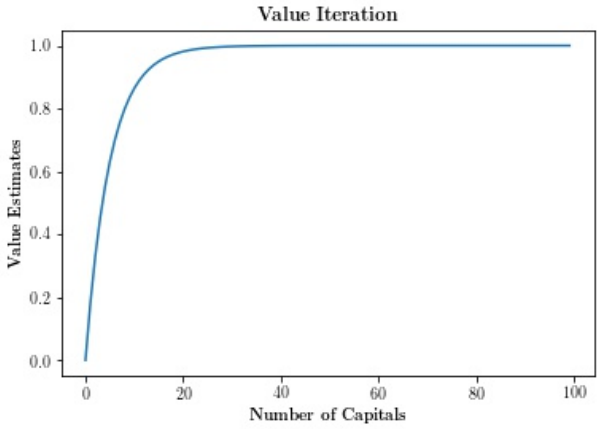

Figure 6: Plot of Value Iteration when ph=0.55