CS 285 HW 1

Yanlai Yang

September 4, 2020

0 Note

As I mentioned in Piazza private post @44, I made a change to the MLP_policy.py file, where I changed the default MSE loss to the log probability loss. All the results shown in this report is based on this modification of the loss function.

1 Question 1.2

Environment Name	Mean	Std	Expert Mean	Percentage
Ant-v2	4664.2	80.1	4713.7	98.9%
Hopper-v2	1038.6	100.9	3772.7	27.5%

Table 1: Performance of Behavioral Cloning on the Ant environment and the Hopper environment. In the table "mean" and "std" refer to the mean and standard deviation of the trained policy's return over multiple rollouts, and "percentage" refers to mean / expert_mean * 100%. Both policies are trained with a neural network with 2 hidden layers, 64 neurons in each hidden layer, a max replay buffer size of 1 million, and 1,000 training iterations. The evaluation batch size is 5,000 and the max number of steps per episode is 1,000.

2 Question 1.3

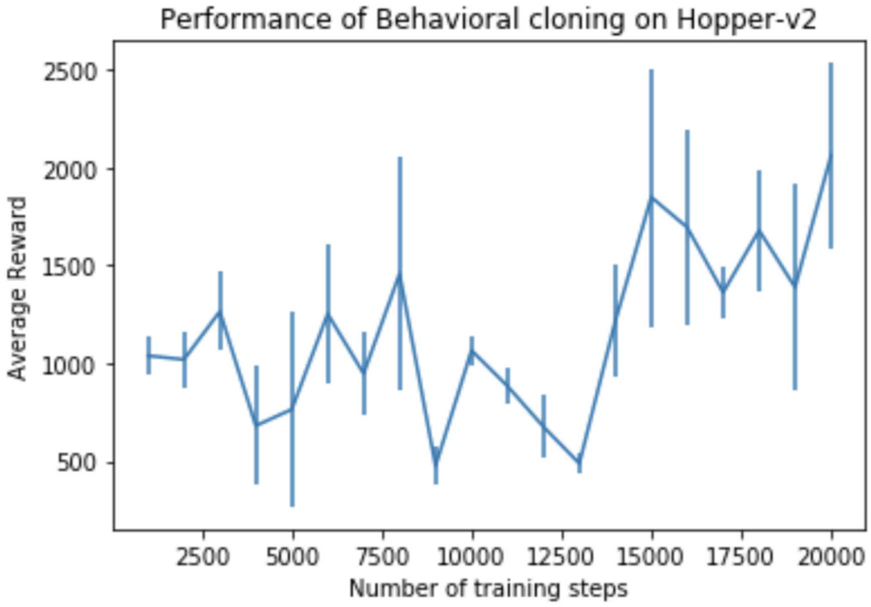

Figure 1: The Performance of behavioral cloning on the Hopper-v2 environment with respect to the number of training steps. The error bars in the figure indicate the standard deviation of the trained policy's return over multiple rollouts, with evaluation batch size be 5,000 and max number of steps per episode be 1,000. Except the number of training steps, the training setup is exactly the same as described in the caption of Table 1. I chose this hyperparameter because I want to examine whether the poor performance of behavioral cloning on the Hopper-v2 environment (as shown in Table 1) is due to under-training.

3 Question 2.2

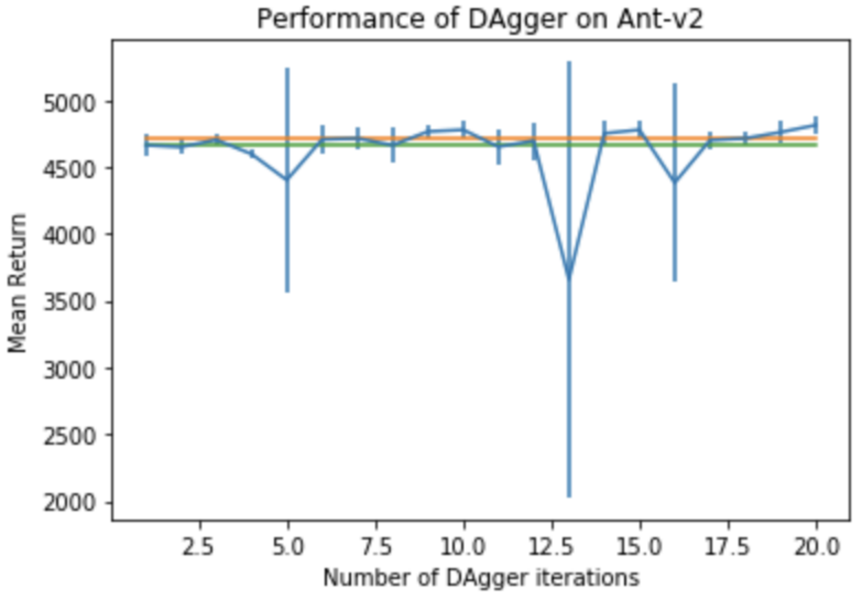

Figure 2: The performance of DAgger on the Hopper-v2 environment with respect to the number of DAgger iterations. The error bars in the figure indicate the standard deviation of the trained policy's return over multiple rollouts, with evaluation batch size be 5,000 and max number of steps per episode be 1,000. The orange line shows the performance of the expert. The green line shows the performance of the behavioral cloning agent. The training setup is exactly the same as described in the caption of Table 1.

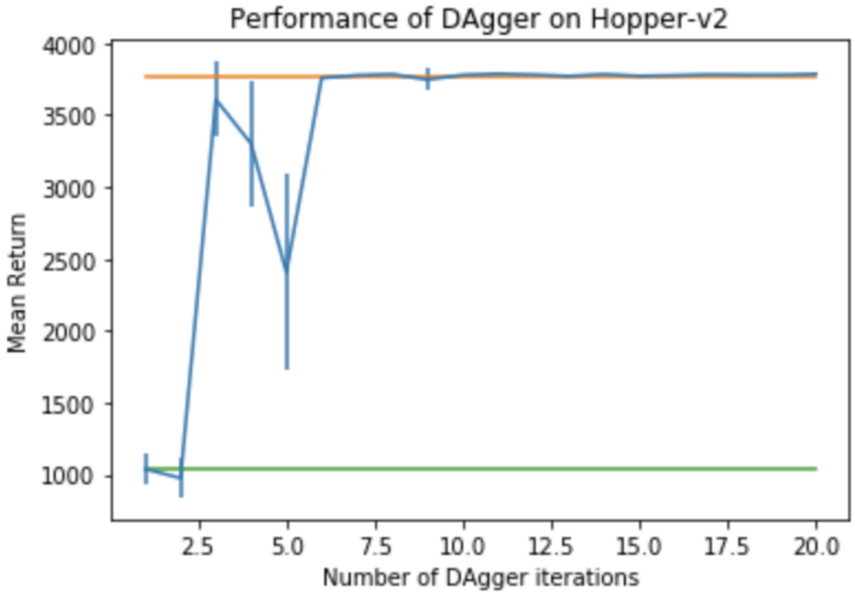

Figure 3: The performance of DAgger on the Ant-v2 environment with respect to the number of DAgger iterations. The error bars in the figure indicate the standard deviation of the trained policy's return over multiple rollouts, with evaluation batch size be 5,000 and max number of steps per episode be 1,000. The orange line shows the performance of the expert.The green line shows the performance of the behavioral cloning agent. The training setup is exactly the same as described in the caption of Table 1.