CS 285 Set 1

Erich Liang

Due: 9/13/21

1 Section 1, Question 2

Task	Policy Return Mean	Policy Return Standard Deviation
Ant	4239.77	1171.55
Hopper	737.08	318.49

Table 1: A comparison of how the behavior cloning policy performs on the ant task and the hopper task. For this comparison, the MLP used to learn actions from observations was a fully connected neural net with two hidden layers each of width 64, and ReLU regularization was placed between each layer. During training, 1000 gradient steps per iteration were used, and each gradient step was based on 100 sampled expert data points. Each task was only ran for one iteration, and the learning rate used was 5e-3. For evaluation, the batch size used was 10000 data points with maximum episode length of 1000.

2 Section 1, Question 3

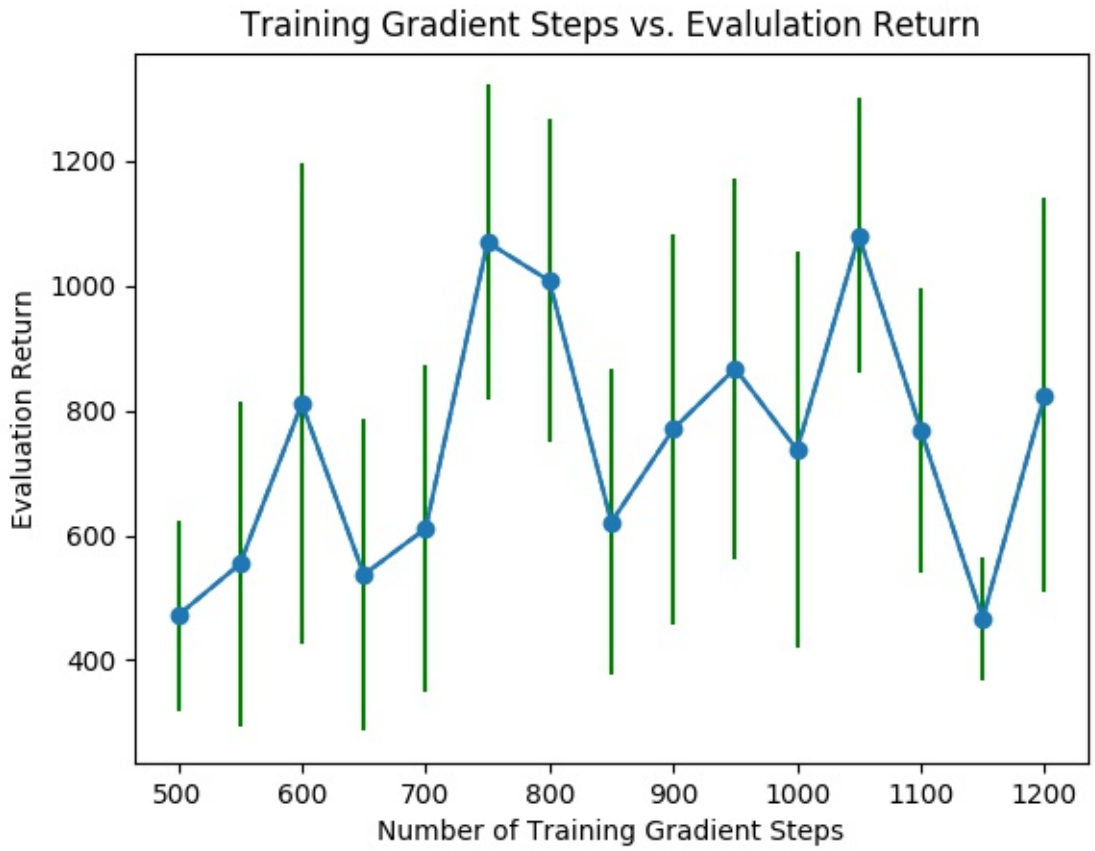

Figure 1: A behavior cloning experiment ran on the Hopper task. The hyperparameter chosen for this experiment was the number of training gradient steps used, ranging from 500 to 1200. All other parameters were kept the same as those in Question 1 Section 2. When the Hopper task was ran previously with 1000 training gradient steps, the evaluation return was not very high, so this experiment was ran to help determine whether the Hopper task was very complex and required more gradient steps (and thus more exposure to data) during behavior cloning algorithm, or if the Hopper task was very simple and the behavior cloning algorithm was overfitting at 1000 training gradient steps. Standard deviation of evaluation returns are plotted as error bars in green.

3 Section 2, Question 2

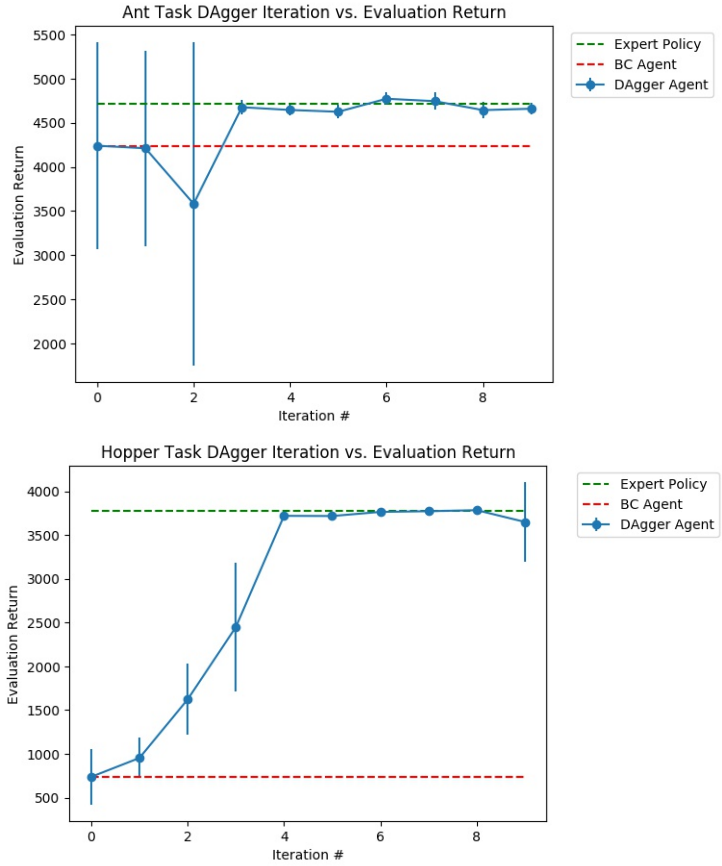

Figure 2: A comparison of behavior cloning agent, DAgger agent, and expert policy for the Ant task and the Hopper task. For these comparisons, the MLP used to learn actions from observations was a fully connected neural net with two hidden layers each of width 64, and ReLU regularization was placed between each layer. During training, 1000 gradient steps per iteration were used, and each gradient step was based on 100 sampled expert data points. Each task was only ran for one iteration and the learning rate used was 5e-3. For evaluation, the batch size used was 10000 data points with maximum episode length of 1000.