CS 285 HW 1

Tasks	Eval. Average Return	Eval. Std. Return	Expert Policy Return	Number of Training Itr.
Ant-v2	2765.54	1019.01	4713.65	1000
HalfCheetah	145.10	75.05	4205.78	100

Table 1: BC on two tasks with the same number of replay buffer (1000 timesteps), the same training batch size (1000 timesteps, note this is not the default value), the same network size (2 layers each 64 units), the same learning rate and the same number of rollouts (5) to take the average to obtain evaluation return. The Ant task is successful with an average evaluation return of 4793.90, whose expert policy return on the same rollout is 4713.65. The HalfCheetah task is a failure with an average evaluation return of only 145.10, whose expert policy return is 4205.78. The only difference between the two tasks is the number of training iterations.

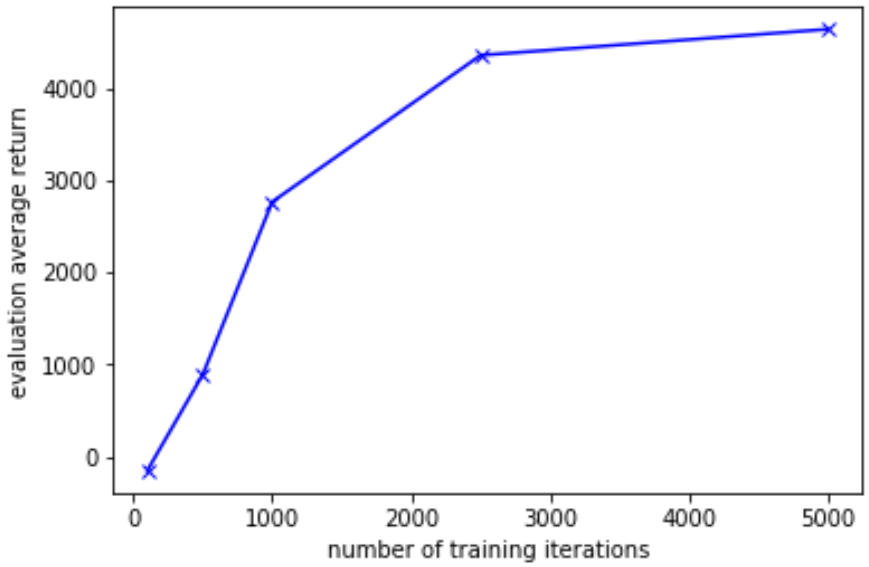

Figure 1: The Ant task with the same hyperparameters as those in Table 1, except the one, the number of training iterations, that varies. The plot is the evaluation average return versus the number of training iterations. The reason for this choice of this varying hyperparameter is that just like in supervised learning that learning with many epochs often makes the loss approach the minimum, here more iterations of training means more gradient descent steps and make the difference between the expert action distribution and the trained policy action distribution smaller.

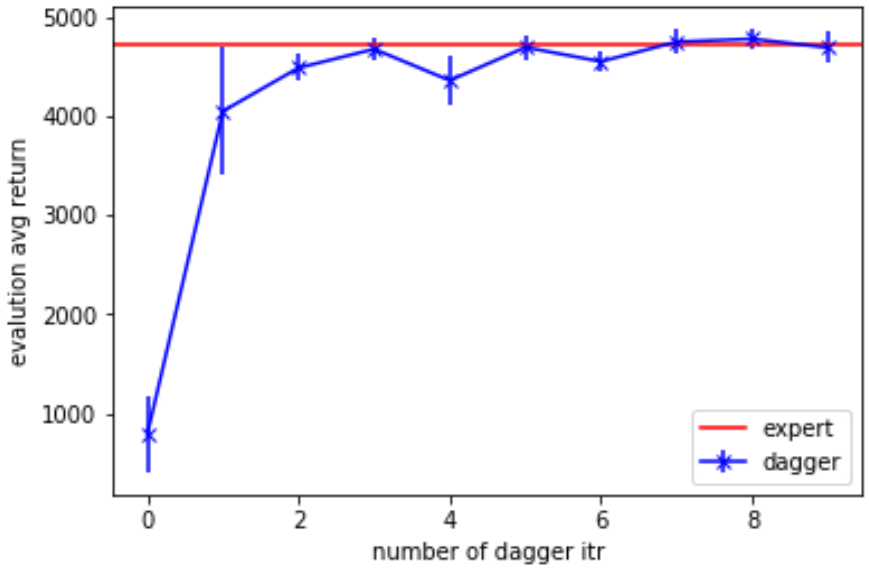

Figure 2a: The DAgger learning curve on the Ant task. All hyperparameters are by default (1000 buffer size, 100 training batch size, 5000 evaluation timesteps, 1000 rollout length, 2 layer 64 unit MLP). The red line is the expert evaluation reward return.

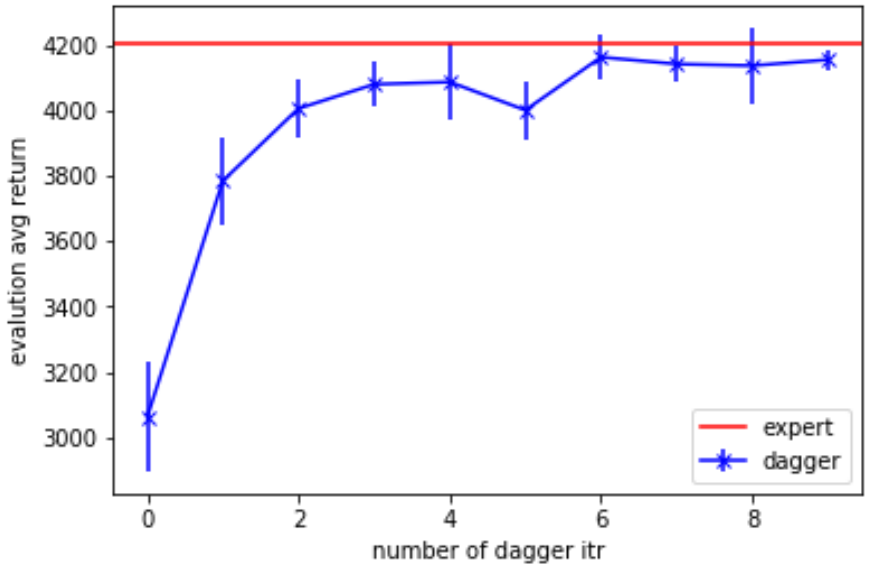

Figure 2b: The DAgger learning curve on the HalfCheetah task. All hyperparameters are by default (1000 buffer size, 100 training batch size, 5000 evaluation timesteps, 1000 rollout length, 2 layer 64 unit MLP). The red line is the expert evaluation reward return.