CS285 Homework 1

Matheus Medeiros Centa

May 2020

1 Results

The three figures summarize the results obtained. If the value of an hyperparameter is not mentioned, the experiment uses its default value. The default hyperparameters are: a network architecture of 2 fully-connected layers of 64 units, a learning rate of 0.005 and a batch size of 1000. All evaluations are made over 10000 steps, which guarantees at least 10 different episodes given the limit of at most 1000 steps per episode.

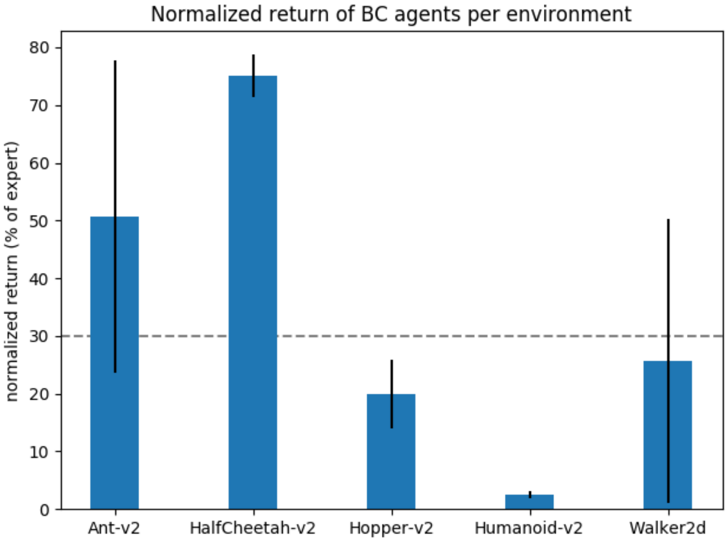

Figure 1: Normalized return of behavior cloning agents on five different environments. The returns are normalized so that the average of the provided expert returns of the respective environment is 100% and the error bars represent the standard deviation of the data.

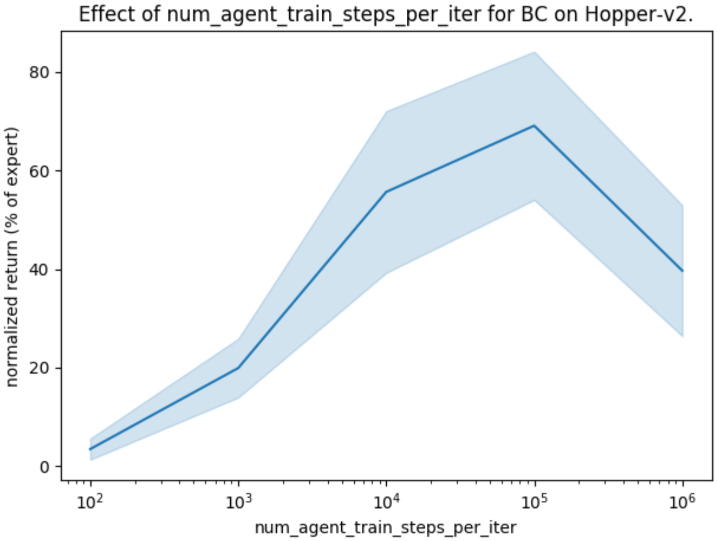

Figure 2: Normalized return of behaviour cloning versus the value of the hyperparameter `num_agent_steps_per_iter` on the Hopper-v2 environment. The returns are normalized so that the average of expert returns is 100% and the band represents the standard deviation of the data. In behavior cloning mode, the chosen parameter controls how many optimizer steps are taken during training. The intuition behind the choice came from the observation that behavior cloning performed worse on more complex tasks on Figure 1, which led me to believe that the worse performing environments would benefit from more training.

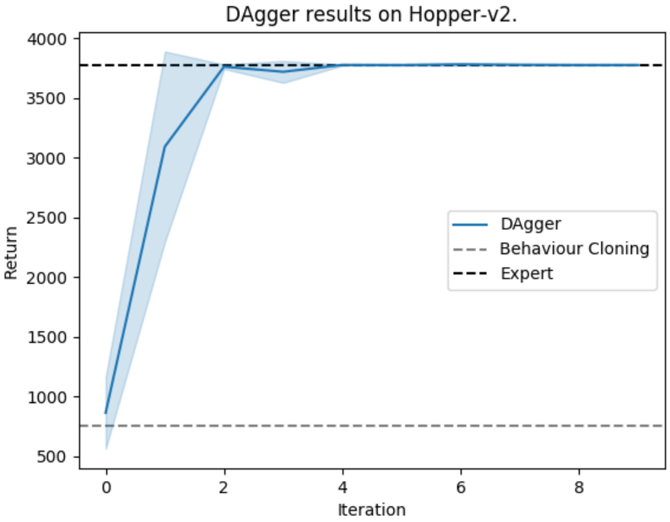

Figure 3: Returns of a DAgger agent on Hopper-v2. The horizontal lines represent the performance obtained with the behavior cloning method and the performance obtained by the expert agent. The band represents the standard deviation of the data. The data represented is an aggregation of results obtained using three different seeds. We can observe that the DAgger iterations helped improve performance significantly, surpassing the results from Figure 2 using fewer training steps.