Report of Deep RL Assignment 1: Imitation Learning

Ke Wang, the First year Ph.D. student at EECS, Berkeley
Email: kewang@berkeley.edu

2.2. Behavior Cloning Results

Task name	Network size	Data	Epochs	Mean reward	Std reward
Ant	Dense(64,32,32)	1000	2000	4251.9	197.6
Hopper	Dense(64,32,32)	1000	2000	1236.1	453.5
Ant expert				4834.0	92.4
Hopper expert				3778.2	4.0

Table.1. Behavior Cloning results under different tasks. Comparison between Ant and Hopper tasks with the same network size, data amount, epoch number. The results show that Behavior Cloning method could get a relative comparable performance to the expert under Ant task while Hopper task does not.

2.3. How BC agent's performance varies with the epoch number

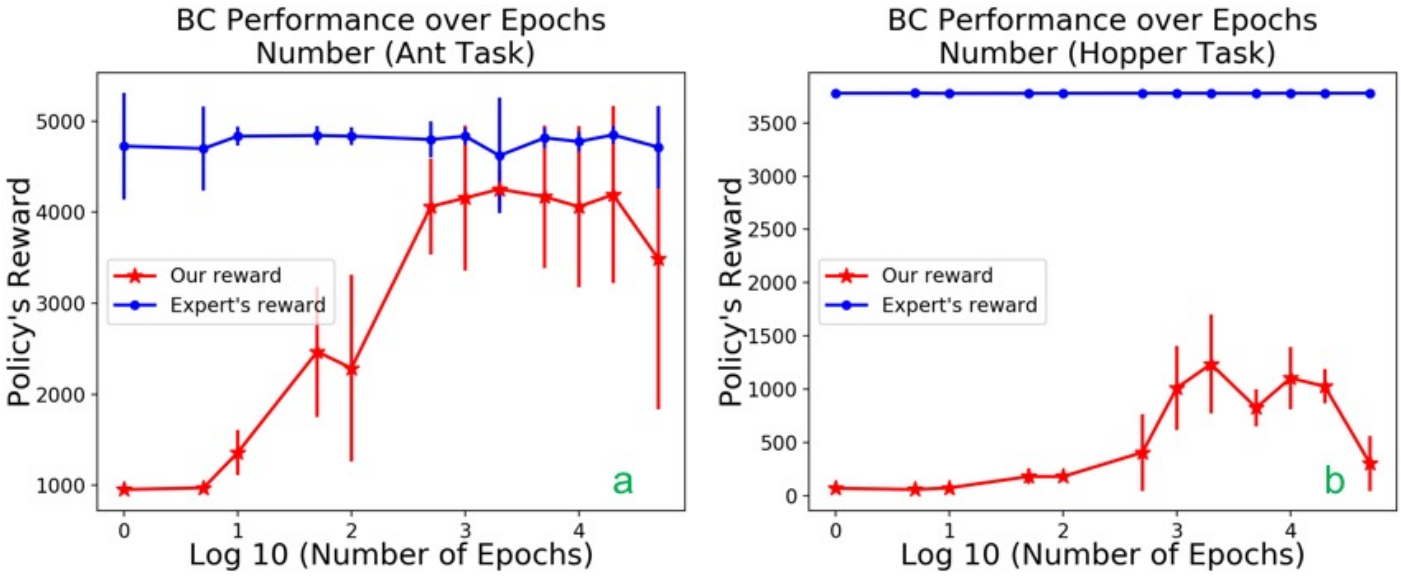

Fig.1. The effect of epoch numbers on the Behavior Cloning performance, including epochs number = {1,5,10,20,50,100,500,1000,2000,5000,10000,20000,50000}. Behavior Cloning performances under: a) Ant task, b) Hopper task are shown on the figures with standard deviation. The results show that with the increase of epoch number, the performance would firstly increase correspondingly, and decrease after a certain epoch number. What we use here are Adam optimizer with learning rate equals to 1e-3, dense neural network and relu activation function. Epoch number is one of the most important hyper parameters when doing training process. A lower epoch number is not able to provide enough information while a larger one would result in over fitting. In this regard, investigating how BC agent's performance varies with epoch number is meaningful and helpful for our parameter settings.

3.2. How BC agent's performance varies with the epoch number

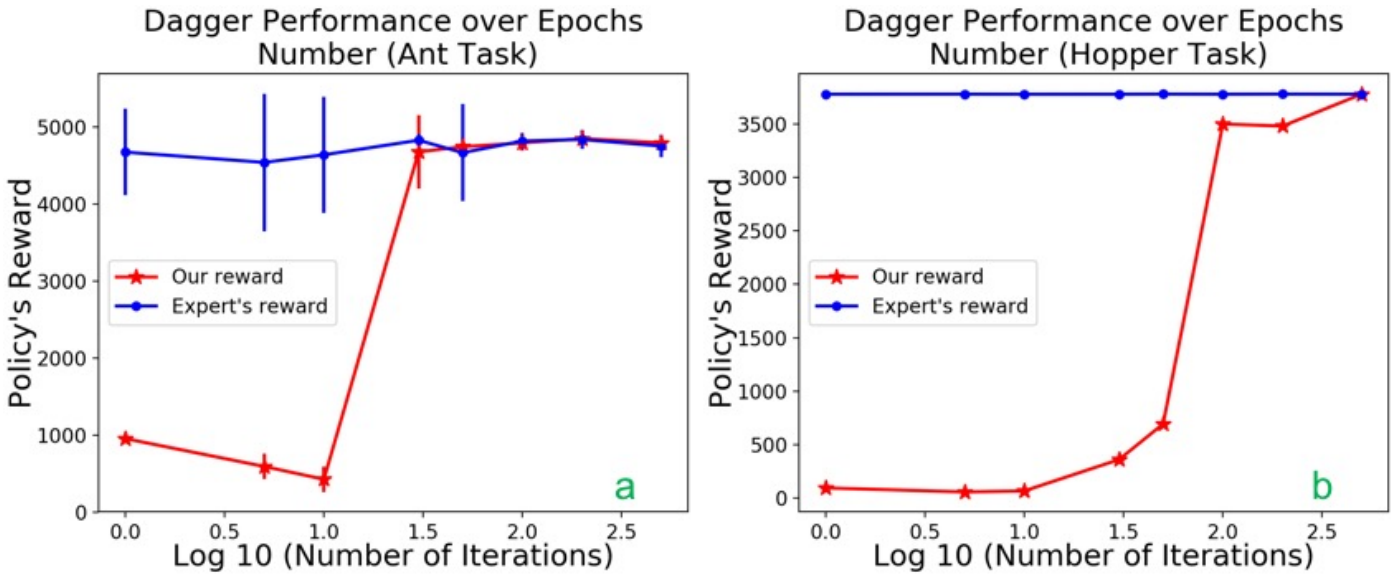

Fig.2. The effect of iteration numbers on the Dagger performance, including iteration number = {1,5,10,20,50,100,200,500}. Dagger performances under a) Ant task, b) Hopper task are shown on the figures with standard deviation. From the results, we could clearly see that by using dagger method, we could get a higher policy's reward and less standard deviation both in two tasks (Ant & Hopper). The architecture of out neural network is fully connected with three layers. The data amount increase 1000 in each iteration.

4.1. Bonus: Alternative policy architecture: Nonlinearities

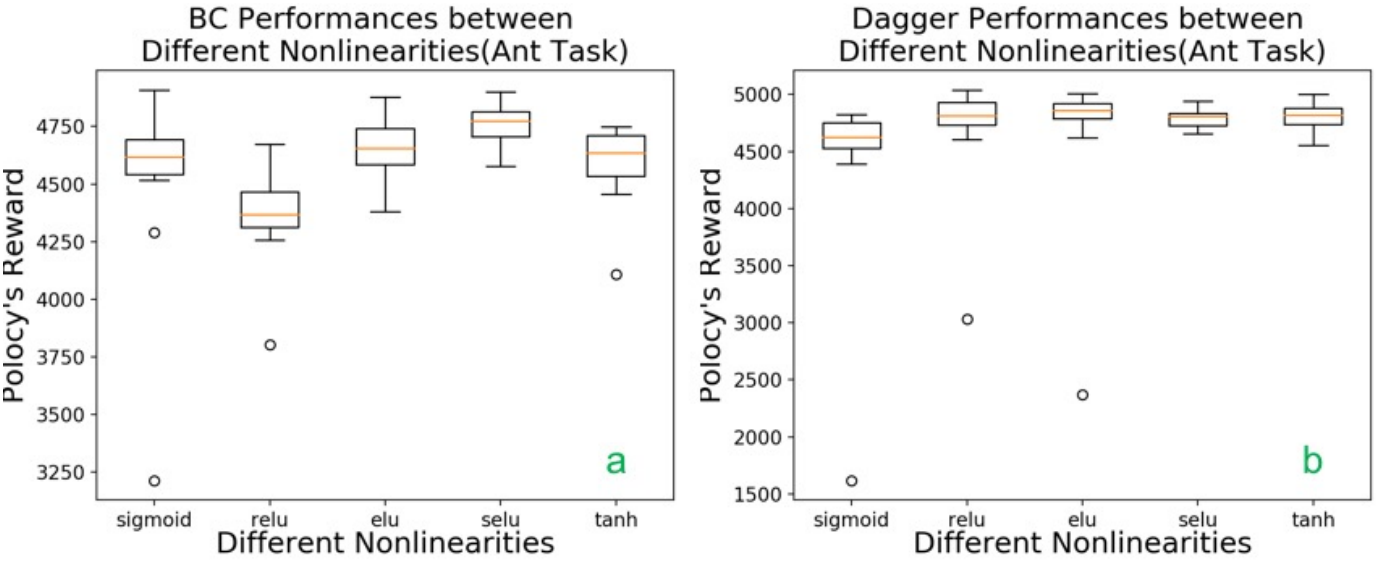

Fig.3. The effect of nonlinear activation functions on policy reward under Ant task using a) Behavior Cloning, b) Dagger. The nonlinearities we test here are {sigmoid, relu, elu, selu, tanh}, which are five common activation functions. The results shows that in Behavior Cloning, selu could achieve the highest reward while elu does in Dagger method.