Deep RL Homework 1: Imitation Learning

Louis TILLOY

For the whole homework, the network used is a 4 layers feedforward neural network :
— layer 1 : 50 hidden units, bias, Relu activation

— layer 2 : 50 hidden units, bias, Relu activation

— layer 3 : 50 hidden units, bias, Relu activation

— layer 4 : n_{labels} units, no bias, no activation function

question 2.2

500,000 training steps 1,000 time-steps 100 rollouts (Clones trained on 100 rollouts of 1,000 time-steps data)	Expert Humanoid	Cloned Humanoid	Expert HalfCheetah	Cloned HalfCheetah
mean	10414	9460	4140	4154
standard deviation	39	2785	86	72

Table 1 : We can see that the behavioral imitation on the humanoid has a very large standard deviation, it is mainly due to the fact that for some rollouts, he falls almost instantly at the beginning. The HalfCheetah has a better score since it is really stable and therefore does not suffer from such problems.

question 2.3

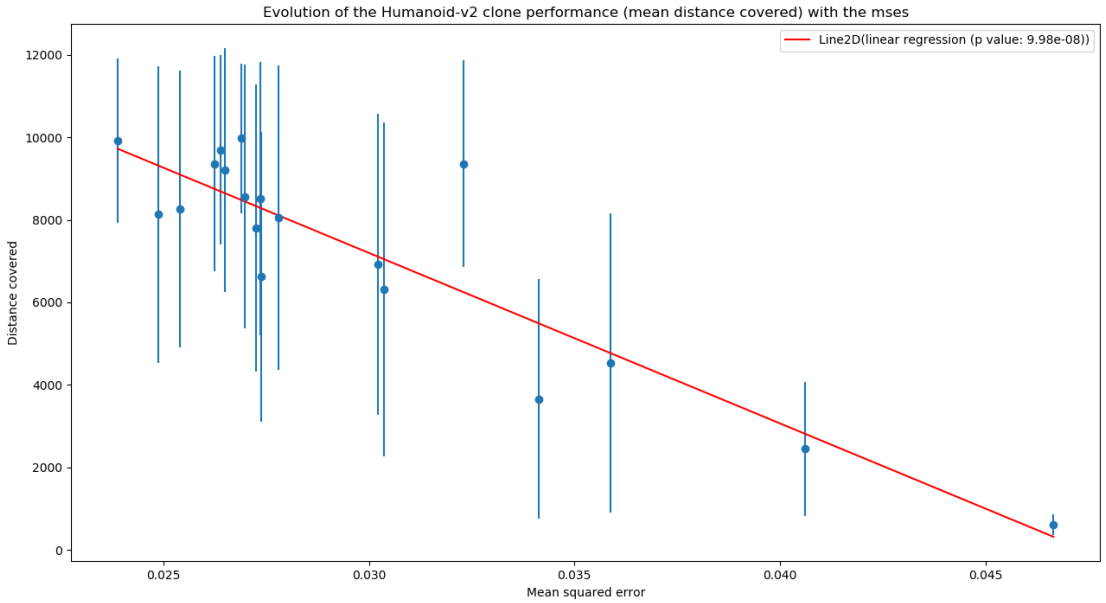

FIGURE 1 – Evolution of the **Humanoid-v2** performance in terms of mean distance covered in 1000 time-steps as a function of mean squared error (the mean squared error has each time been computed on the whole validation set, representing 10% of the whole data). I chose this hyper parameter for the Humanoid model because it is similar to the Tightrope walker example seen in class (indeed the Humanoid cannot get up once on the ground, the simulation actually stops, just like the tightrope walker cannot get back and the rope) and I wanted to see if the relation between the distance and the mean squared error was also linear like with the 0-1 loss seen in class. Because the standard deviation is really high, it is hard to conclude, but the results show a well fitted linear relation, as expected.

question 3.2

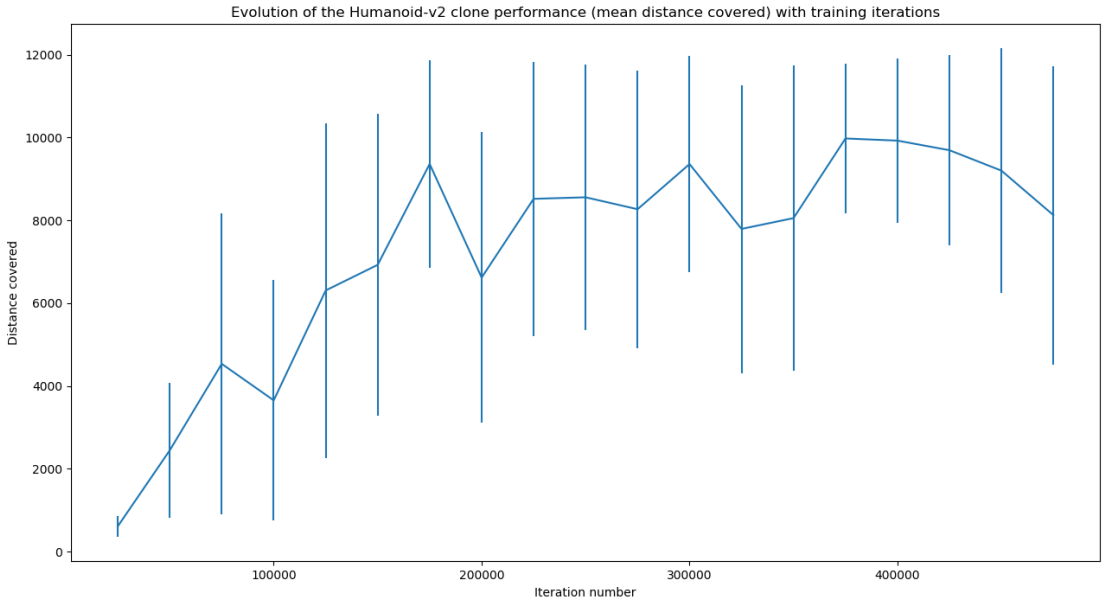

FIGURE 2 – Evolution of the **Humanoid-v2** performance in terms of mean distance covered in 1000 time-steps as a function of the number of training iterations (this figure is here to compare the classic Humanoid with the DAgger trained one). (All means and standard values were computed over 100 rollouts, the training data is composed of 100 rollouts of 1000 time-steps from the expert Humanoid-v2)

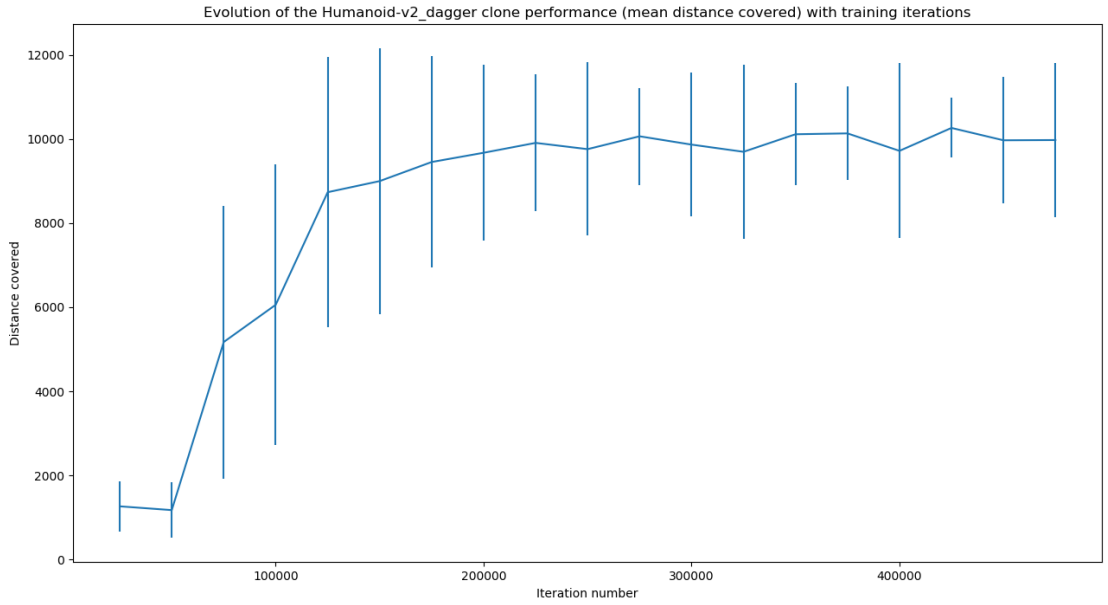

FIGURE 3 – Evolution of the **DAgger trained Humanoid-v2** performance in terms of mean distance covered in 1000 time-steps as a function of the number of training iterations. We can see that the DAgger trained Humanoid-v2 has a better performance than the classic Humanoid-v2. Indeed it scores about 800 times more points and has a standard deviation close to a third of the classic Humanoid-v2. In terms of training, the DAgger trained Humanoid can be trained with less iterations and manages to stabilize itself in good performance whereas the classic Humanoid oscillate more. Even though each iteration costs more on average, the training with DAgger is still faster. (All means and standard values were computed over 100 rollouts, the original training data is composed of 100 rollouts of 1000 time-steps from the expert Humanoid-v2. Every 25000 training steps, starting from training step 50000, the model is used to generate 10 rollouts of 1000 time-steps which are stored in the validation and training set)