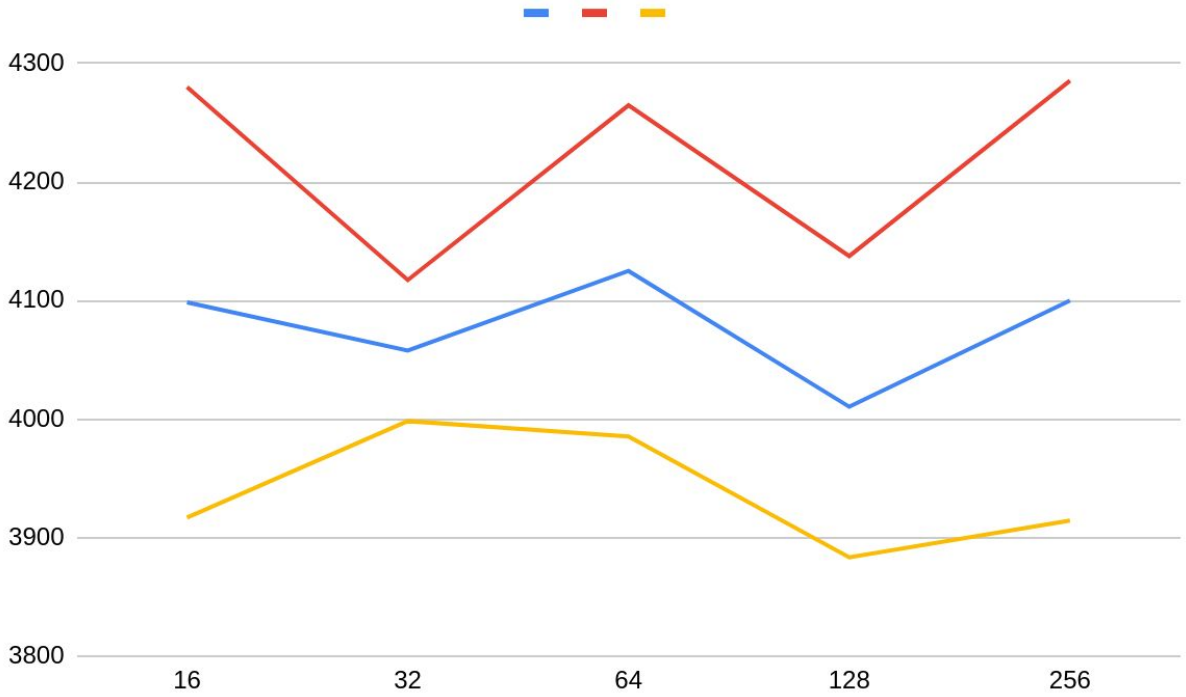
Deep RL hw1 report

1.2 All the parameters of the script are at the default state, except for the random seed. The following table contains the values of the total reward on evaluation sample for two environments: Half-cheetah and Hopper. Lower summarized their average and standard deviation. The last line presents the total return in the initial data collection.

seed	Hopper	Cheetah
1	630.13	2888.59
2	758.9	2572.65
3	317.35	3056.59
4	945.22	2819.28
5	237.55	2657.13
6	847.78	2594.42
7	624.37	2827.08
avg	623.0428571	2773.677143
std	262.9533903	175.2638046
initial	3772.67	4205.78

It's clear that for Half-cheetah the reward on evaluation sample is greater than 30% of the initial reward. For Hopper, reward is lower than 30% of the initial one.

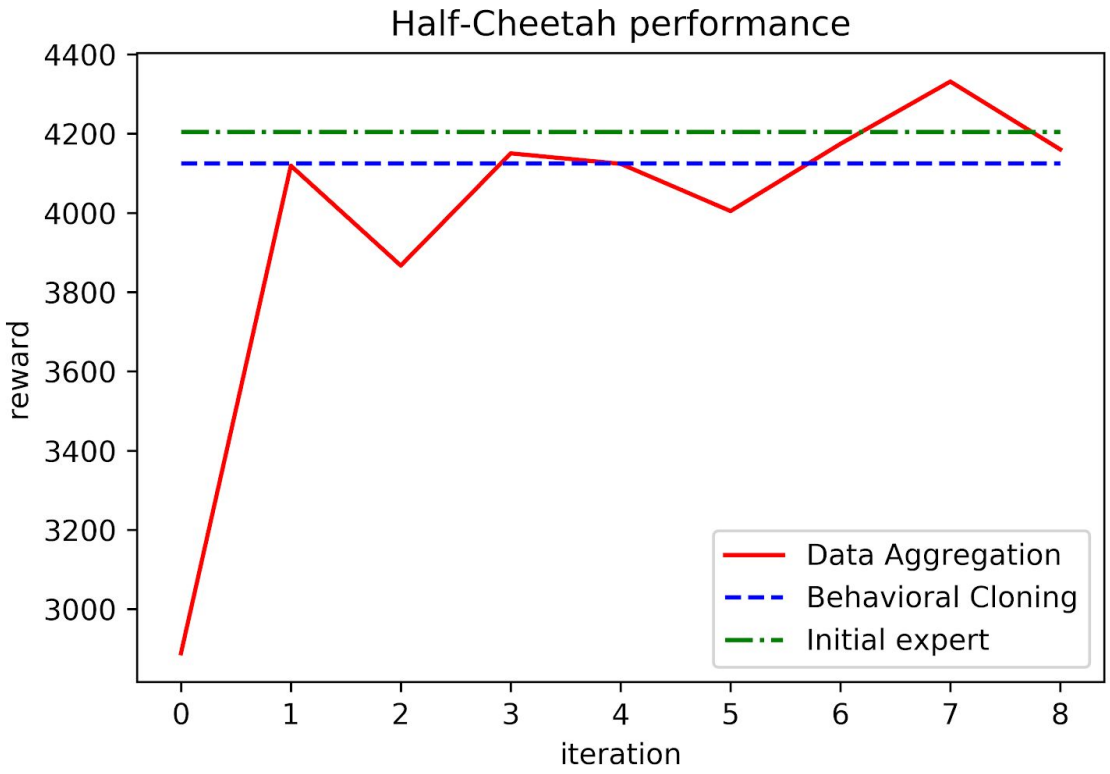
1.3 For Half-cheetah environment, the following plot represents the dependence of the total reward on evaluation sample on the width of the MLP representing the policy. 8000 steps seemed enough for convergence of the Adam procedure for every width value that is considered, so `--num_agent_train_steps_per_iter` was set to `8000`. For `--seed` in {1,..,5} and `--size` in {16, 32, 64, 128, 256} we collect the total reward on evaluation sample. All the other parameters are left as default.
Here we report the average and standard deviation over the trials.



Blue line characterizes the average value of the reward. Red is avg+std, while yellow is avg-std.

We built the chart to study generalization behaviour of the learning model. From the results, we conclude that the default value 64 for the width of the nework is the optimal one.

2.1 For the environment Half-cheetah we present the “reward on evaluation sample” curve.
To provide a fair comparison with the results from 1.3, we changed the number of iterations to 9.
All the other parameters are left at default.


Half-Cheetah performance
4400
4200
4000
3800
3600
3400
3200
3000
reward
0
1
2
3
4
5
6
7
8
iteration
Data Aggregation
Behavioral Cloning
Initial expert