CS 285 Homework 4

Sagnik Bhattacharya SID: 3033583960

November 6, 2021

Question 1

Figures 1, 2, and 3 show the qualitative model predictions for different architectures and number of agent training steps per iteration. As we can see in Figure 2, the two-layer network with 250 units in each layer, where the agent was trained for 500 steps per iteration performed best. In Figure 3 we can see that the small number (5) of agent train steps per iteration really hindered the performance of the model. As seen in Figure 1, the shallowness and narrowness of the one-layer, 32-hidden unit neural network was also not as good as Figure 2, but it was better than the third case as the agent was trained for 500 steps per iteration, which allowed it to surpass the performance of the undertrained deeper and wider model.

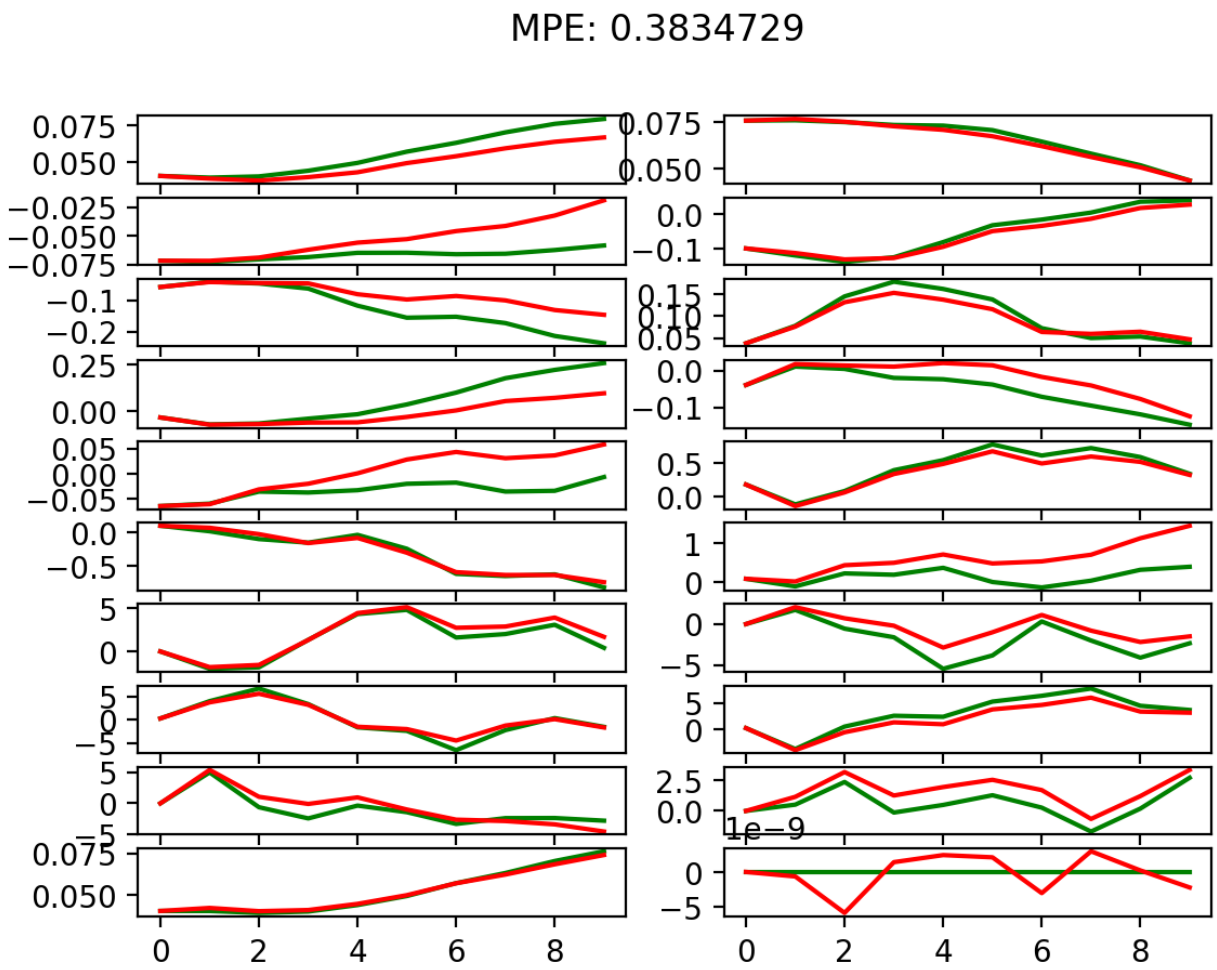

Figure 1: Qualitative model predictions for a one-layer, 32-wide neural network. 500 agent train steps per iteration.

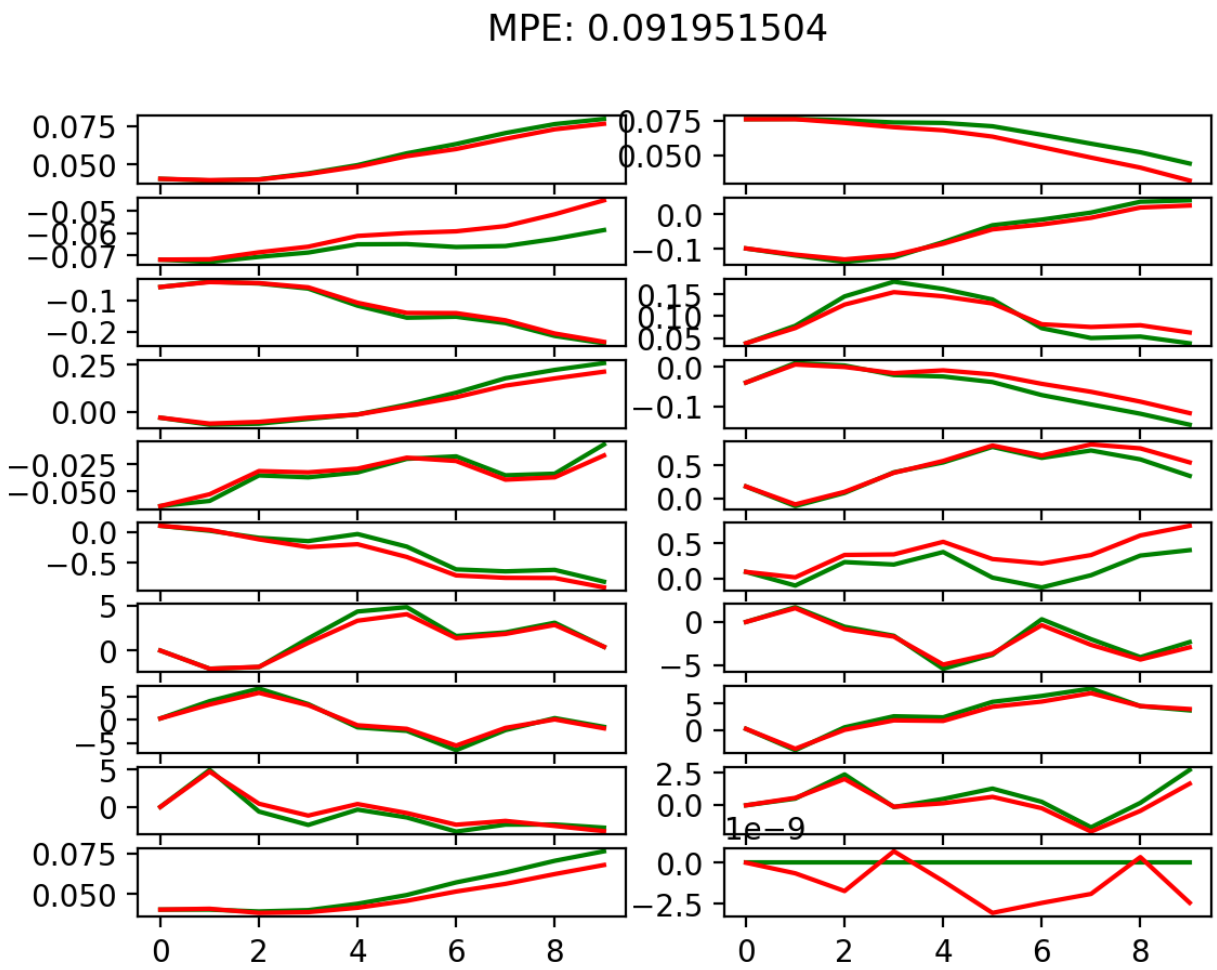

Figure 2: Qualitative model predictions for a two-layer, 250-wide neural network. 500 agent train steps per iteration.

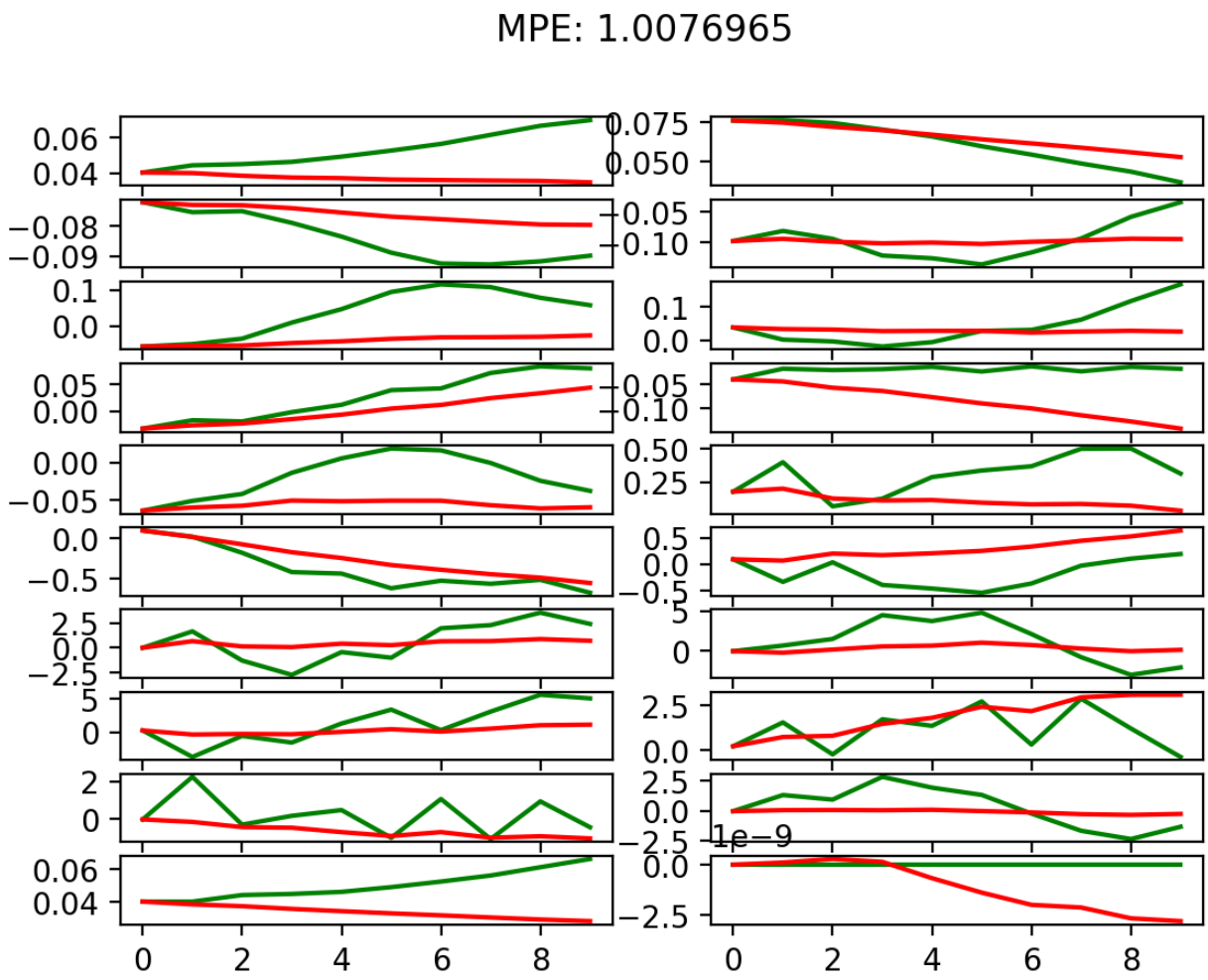

Figure 3: Qualitative model predictions for a two-layer, 250-wide neural network. 5 agent train steps per iteration.

Question 2

Figure 4 shows the evaluation average return for the MPC agent trained for a single iteration on randomly collected data. Figure 5 shows the training average returns for the same agent in the same scenario.

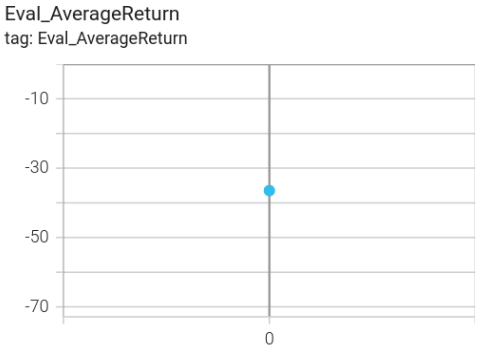

Figure 4: Average return during evaluation of the MPC agent trained with randomly collected data.

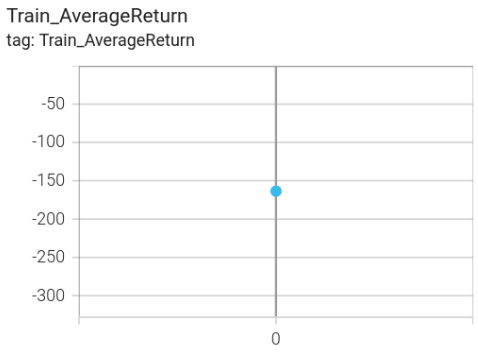

Figure 5: Average return during training of the MPC agent trained with randomly collected data.

Question 3

Please see Figures 6, 7, and 8 respectively for the evaluation average returns of on-policy data collection and iterative model training in the obstacles, reacher, and cheetah environments.

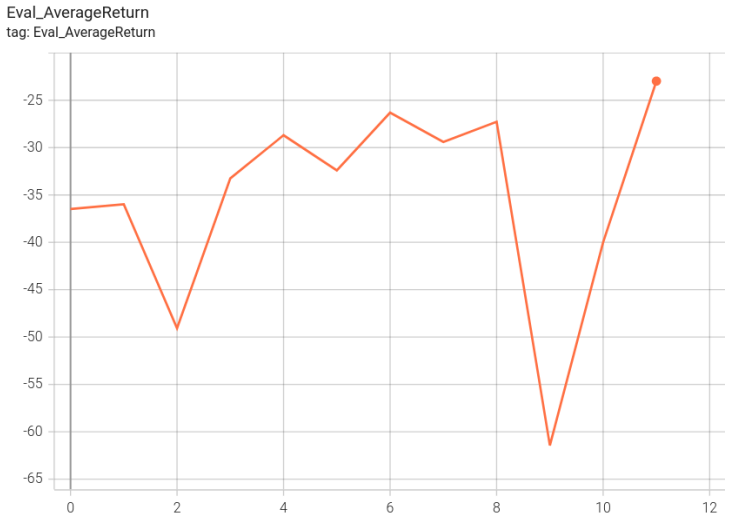

Figure 6: Evaluation average returns for on-policy data collection and iterative model training in the obstacles environment.

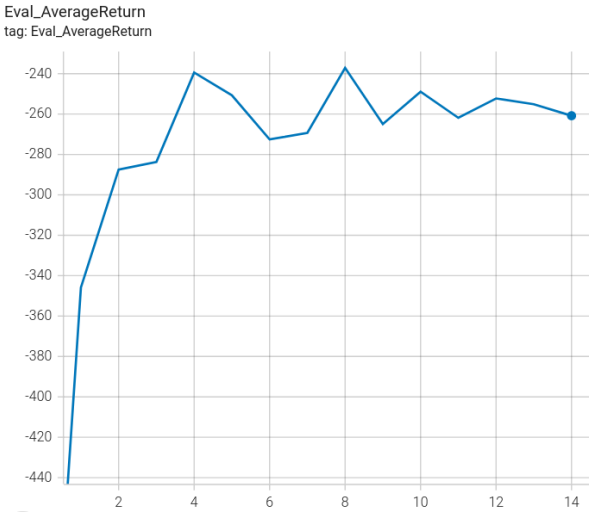

Figure 7: Evaluation average returns for on-policy data collection and iterative model training in the reacher environment.

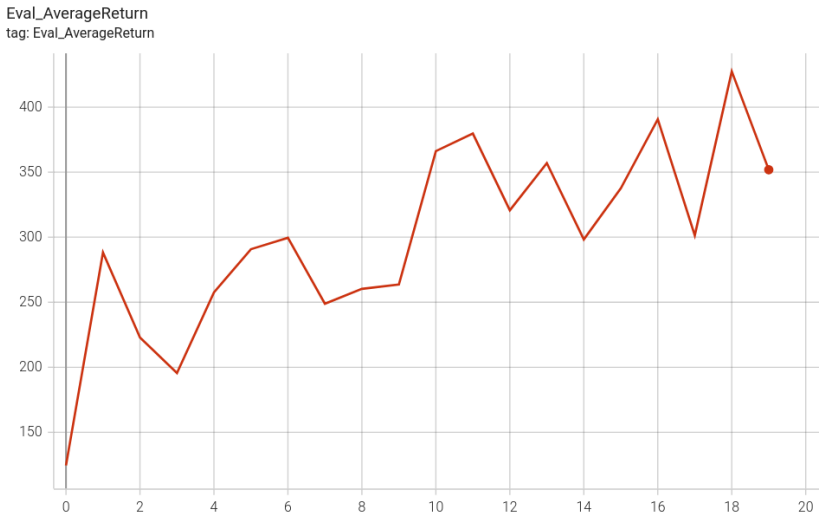

Figure 8: Evaluation average returns for on-policy data collection and iterative model training in the cheetah environment.

Question 4

Please see Figures 9 for the effect of ensemble size, **??** for the effect of the number of candidate sequences, and 11 for the effect of the planning horizon on performance.

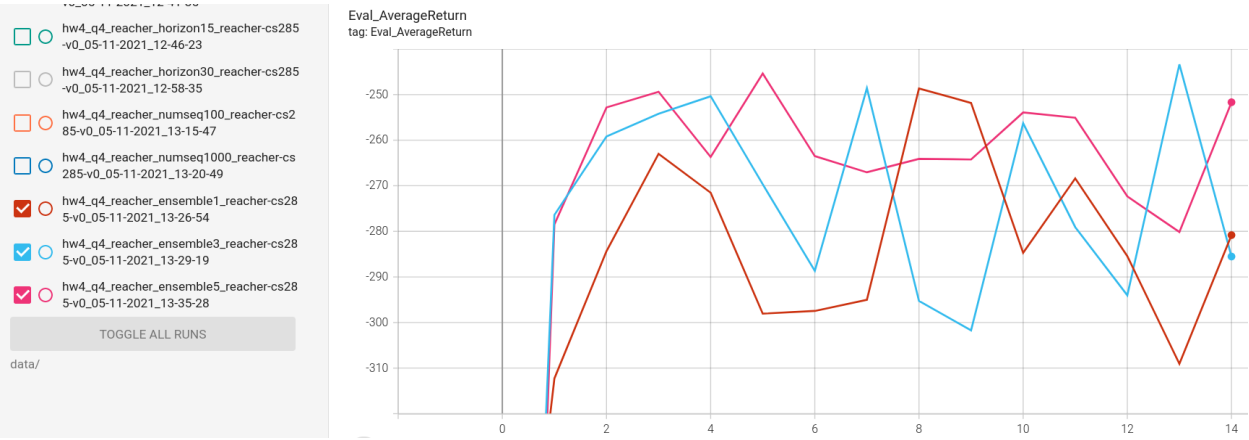

Figure 9: As we can see in this figure, increasing ensemble size leads to better evaluation returns, due to a reduction in variance, while keeping bias the same. Thus using a larger ensemble allows us to get better performance at the cost of extra compute.

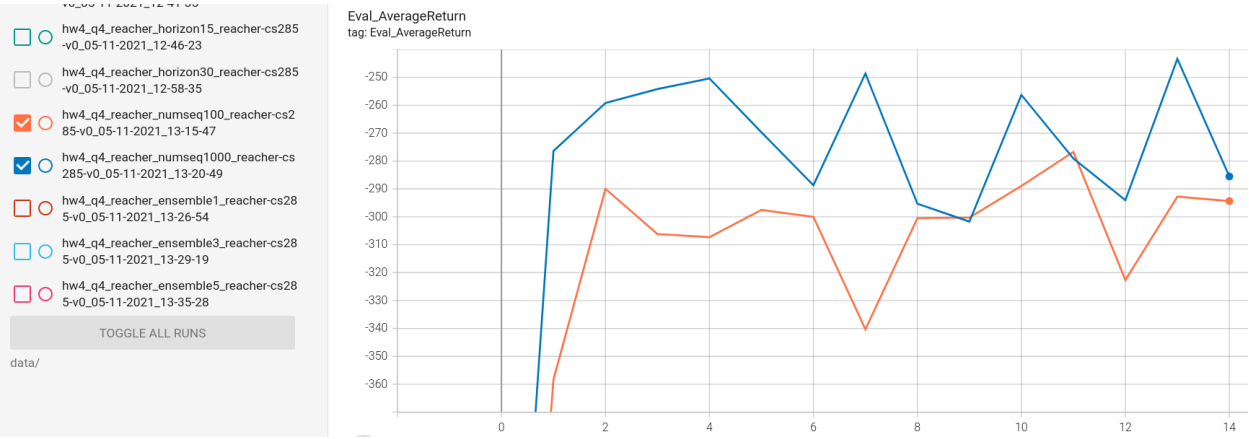

Figure 10: As we can see in this figure, increasing the number of candidate action sequences also increases performance, as having more candidate sequences makes it more likely that we will find a better-performing sequence. This also comes at the cost of extra compute.

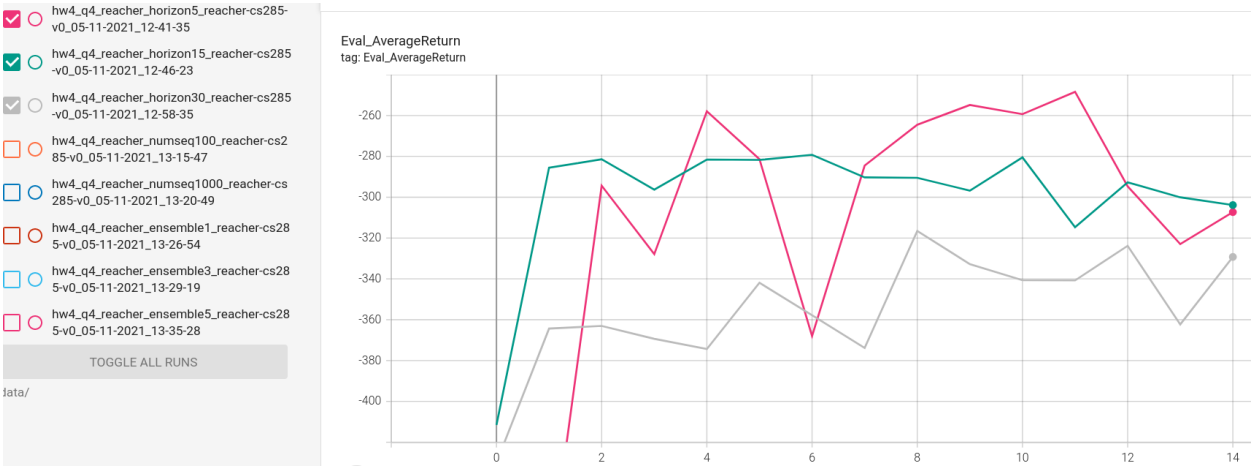

Figure 11: As we can see in this figure, increasing the planning horizon does not lead to better performance, probably because having a longer horizon can distract the model, since its predictions become worse at the later stages in the horizon, leading to suboptimal decisions overall. Thus we need to choose a moderately large horizon, one that is neither too short nor too long.

Question 5: CEM

Figure 12 shows a comparison of CEM with random shooting, as well as a comparison of different numbers of sampling iterations of CEM.

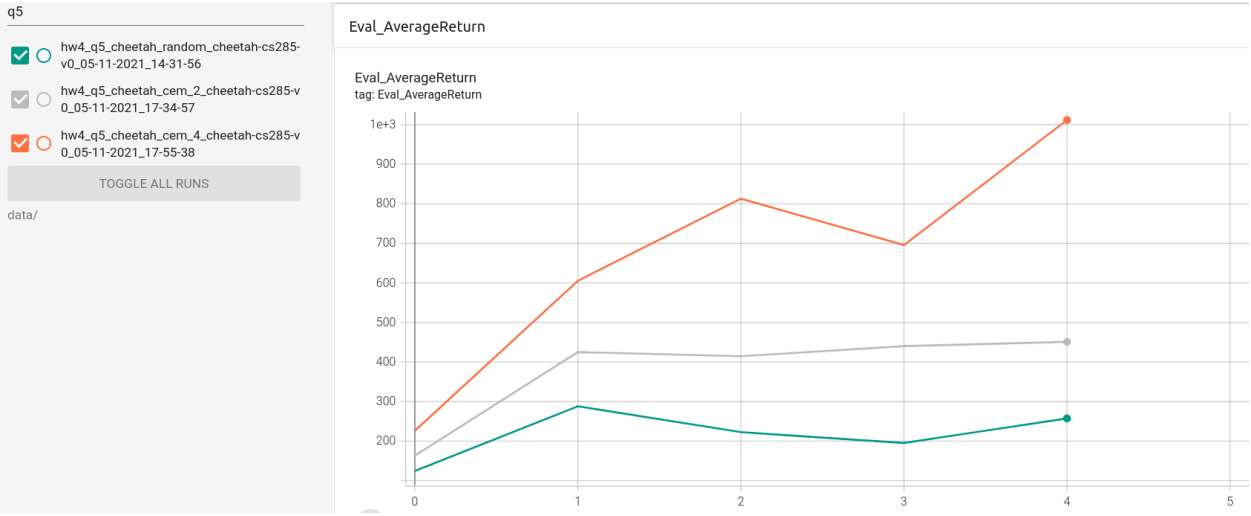

Figure 12: Green curve is random shooting, gray is CEM with 2 sampling iterations, and orange curve is CEM with 4 sampling iterations. As we can see, CEM performs better than random shooting, since it intelligently selects candidate action sequences based on performance. Running CEM for more iterations also leads to better performance, since it iteratively improves on its previous action sequences, by sampling closer to the elites, which are the sequences with the highest performance.