CS 285 HW 4

Yanlai Yang

November 1, 2020

1 Problem 1

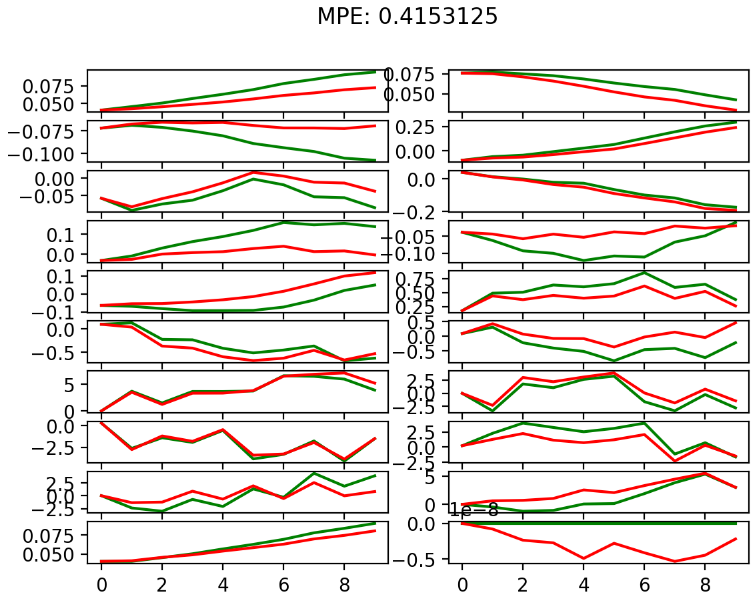

Figure 1: The qualitative model predictions for MBRL with neural network architecture 1x32 and 500 training steps on cheetah-cs285-v0.

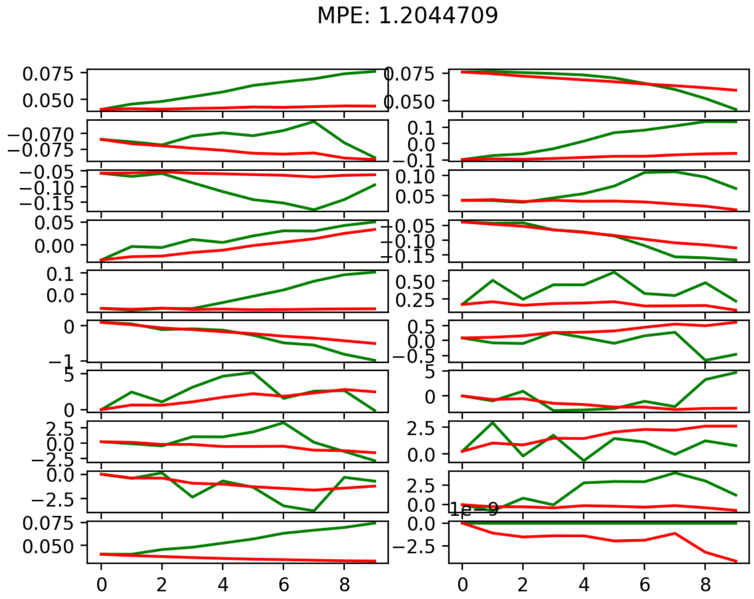

Figure 2: The qualitative model predictions for MBRL with neural network architecture 2x250 and 5 training steps on cheetah-cs285-v0.

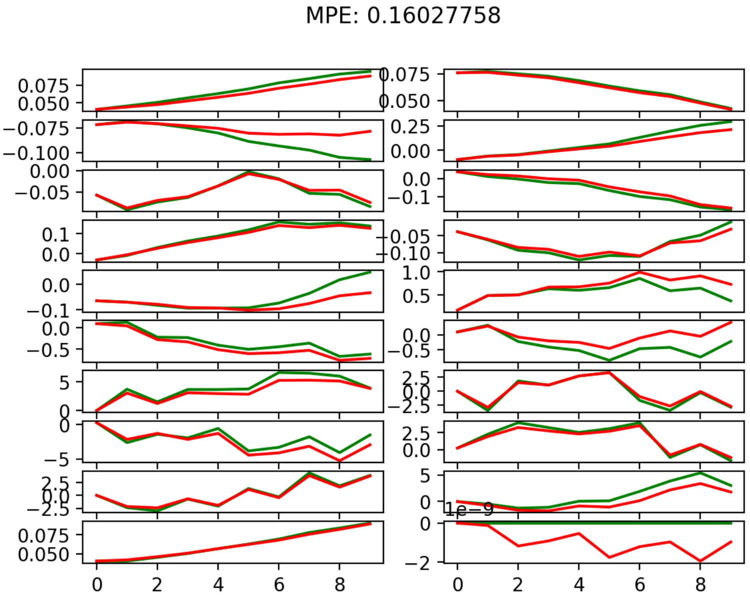

Figure 3: The qualitative model predictions for MBRL with neural network architecture 2x250 and 500 training steps on cheetah-cs285-v0.

Comment: The neural network architecture 2x250 with 500 training steps works best, because the neural network architecture is more expressive and a sufficient number of training steps is taken. The 1x32 neural network model is not expressive enough, and 5 training steps is not sufficient for the model to be reasonably trained.

2 Problem 2

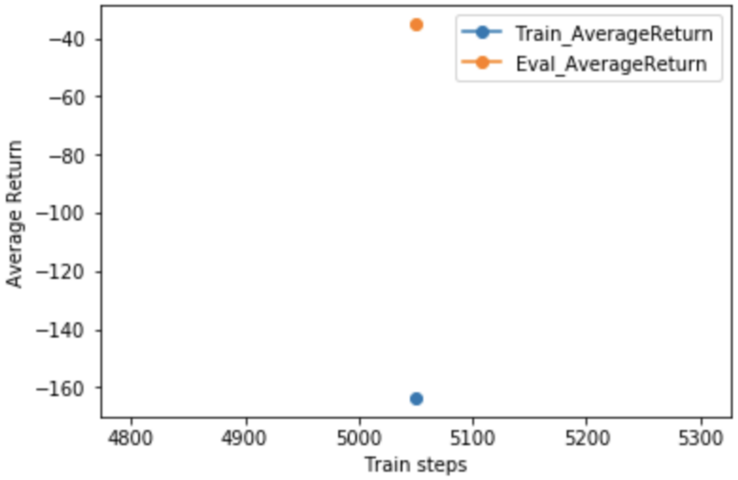

Figure 4: Comparison of Training Average Return (which was the execution of random actions) to Evaluation Average Return (which was the execution of MPC using a model that was trained on the randomly collected training data).

3 Problem 3

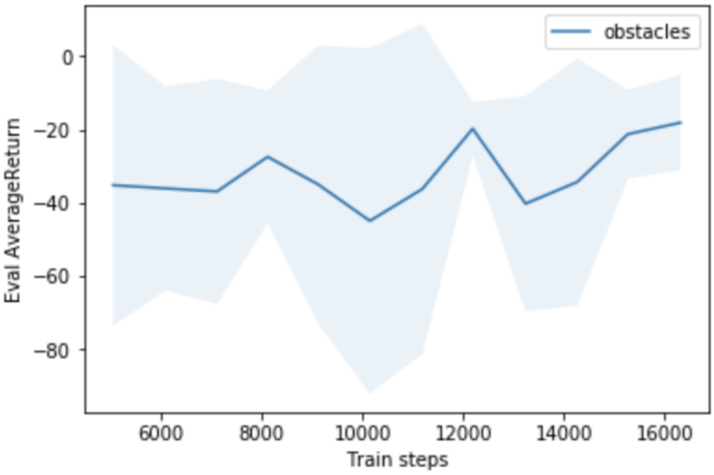

Figure 5: The performance plot of MBRL with on-policy data collection and iterative model training on obstacles-cs285-v0. The shaded area denotes the standard deviation.

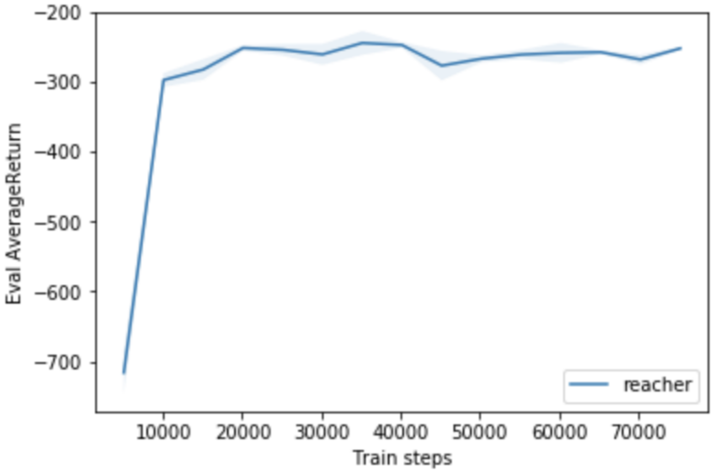

Figure 6: The performance plot of MBRL with on-policy data collection and iterative model training on reacher-cs285-v0. The shaded area denotes the standard deviation.

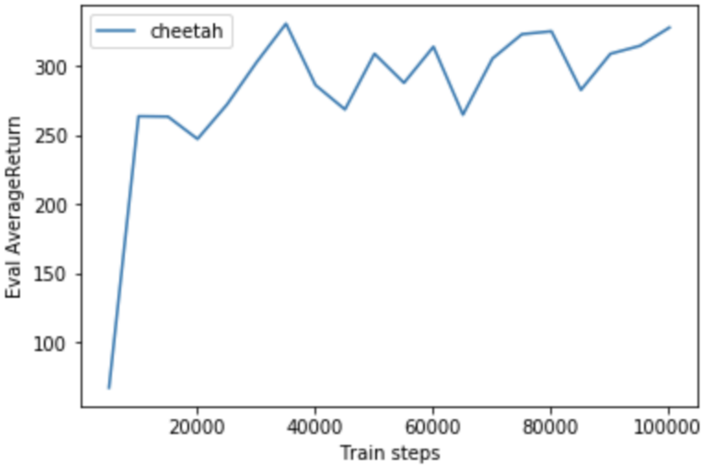

Figure 7: The performance plot of MBRL with on-policy data collection and iterative model training on cheetah-cs285-v0. The shaded area denotes the standard deviation.

4 Problem 4

4.1 Experiments with the MPC planning horizon

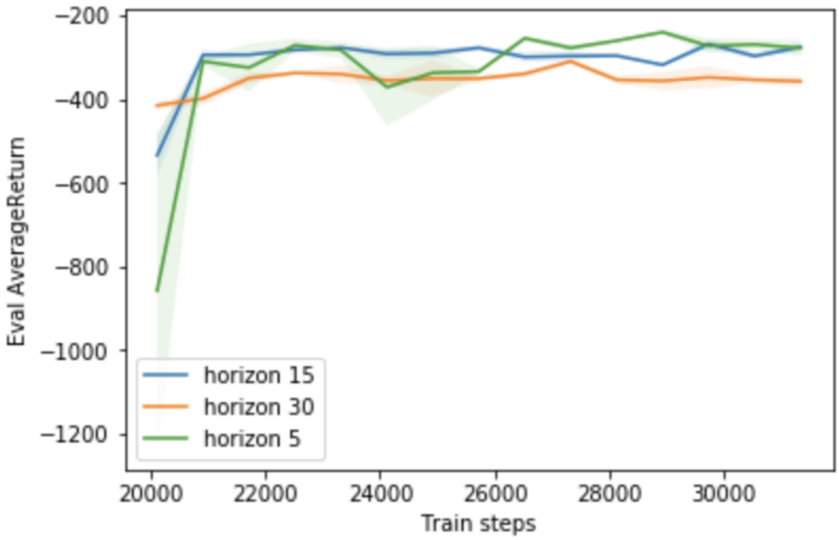

Figure 8: The performance plot of MBRL with different MPC planning horizons. It turns out that a long MPC planning horizon (30 steps in this experiment) starts off with a higher average return, but end up have a lower average return after several training iterations, compared to that of a shorter MPC planning horizon. The algorithms with horizon length 5 and length 15 reach similar asymptotic performance in this experiment, though the one with horizon 15 start off higher. Explanation of this pattern: in this specific environment, having a shorter horizon is more desirable since completing the task does not require very long horizon planning, and the model error will not accumulate as much when the horizon is short. The shaded area denotes the standard deviation.

4.2 Experiments with the number of random action sequences considered

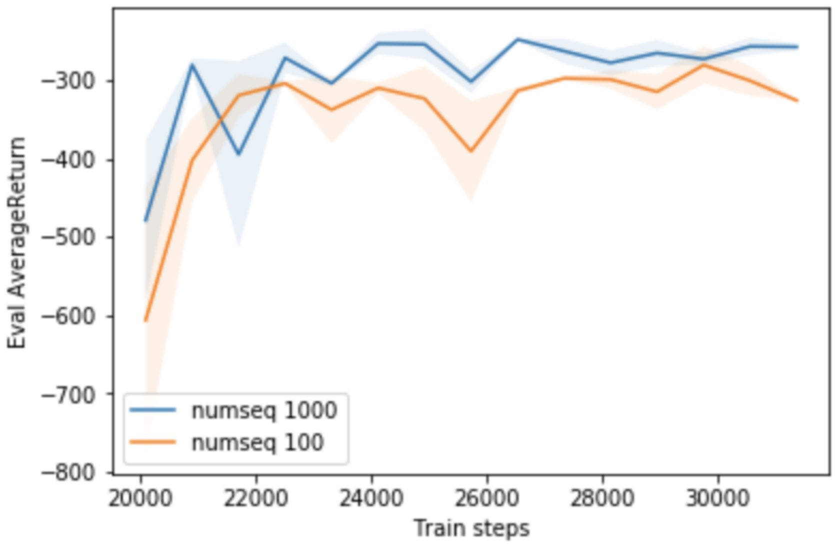

Figure 9: The performance plot of MBRL with different number of random action sequences considered. It turns out that considering 1000 random action sequences results in much better performance than considering only 100 random action sequences. This is intuitive because the more random action sequences we consider, the more likely the best first action we pick is an action that leads to a high-reward trajectory. The shaded area denotes the standard deviation.

4.3 Experiments with the number of models in the ensemble

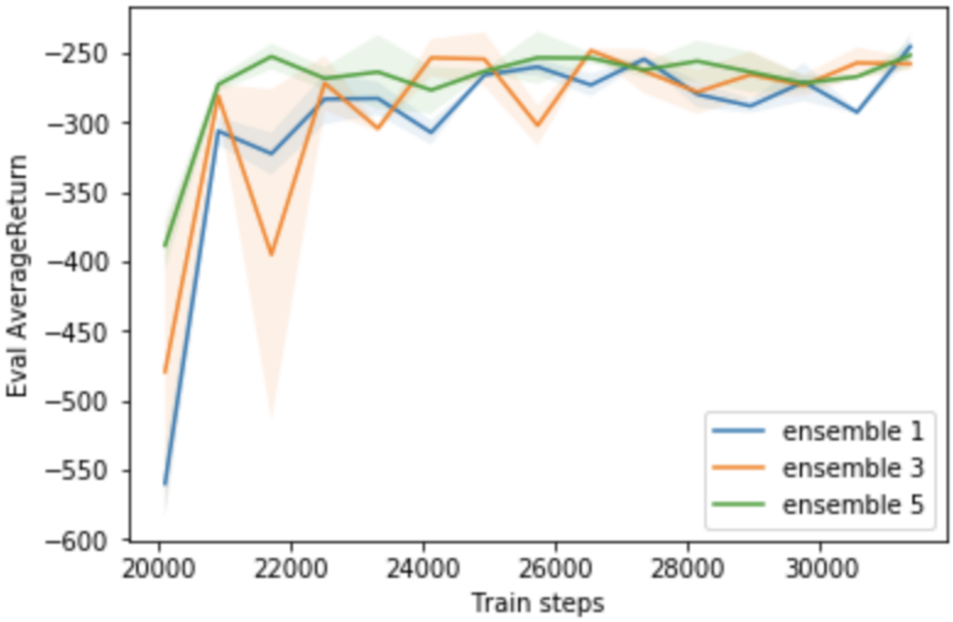

Figure 10: The performance plot of MBRL with different number of models in the ensemble. It turns out that using more models ensemble results in better performance, but only in the beginning phase of training. It is intuitive because using more models in the ensemble can correct the errors made in a minority of the models in the ensemble. As the training goes on and the models become more and more accurate, using more models in the ensemble does not benefit as much since 1 model can also make accurate predictions. The shaded area denotes the standard deviation.