[Mustafa Suman] Assignment 4

Due November 03, 11:59 pm

2 Analysis

2.1 Alternative Simulation Lemma

Proof:

$$Q^{\pi} - \hat{Q}^{\pi} = (I - \gamma \cdot P^{\pi})^{-1} r - \hat{Q}^{\pi}$$

$$= (I - \gamma \cdot P^{\pi})^{-1} r - (I - \gamma \cdot P^{\pi})^{-1} (I - \gamma \cdot P^{\pi}) \hat{Q}^{\pi}$$

$$= (I - \gamma \cdot P^{\pi})^{-1} (I - \gamma \cdot \hat{P}^{\pi}) \hat{Q}^{\pi} - (I - \gamma \cdot P^{\pi})^{-1} (I - \gamma \cdot P^{\pi}) \hat{Q}^{\pi}$$

$$= \gamma (I - \gamma \cdot P^{\pi})^{-1} (P^{\pi} - \hat{P}^{\pi}) \cdot \hat{Q}^{\pi}$$

$$= \gamma (I - \gamma \cdot P^{\pi})^{-1} (P\pi - \hat{P}\pi) \cdot \hat{Q}^{\pi}$$

$$= \gamma (I - \gamma \cdot P^{\pi})^{-1} (P - \hat{P}) \cdot \pi \hat{Q}^{\pi}$$

$$= \gamma (I - \gamma \cdot P^{\pi})^{-1} (P - \hat{P}) \cdot \hat{V}^{\pi}$$

2.2 Statements

1. True.
2. False.
3. True.
4. False.

Problem 1

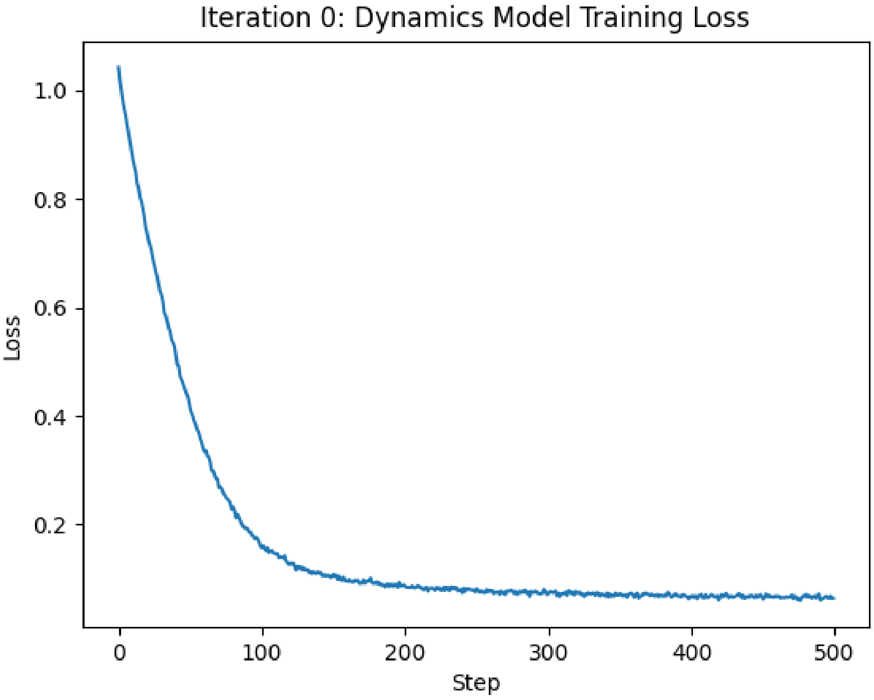


Figure 1: halfcheetah_0_iter; default settings

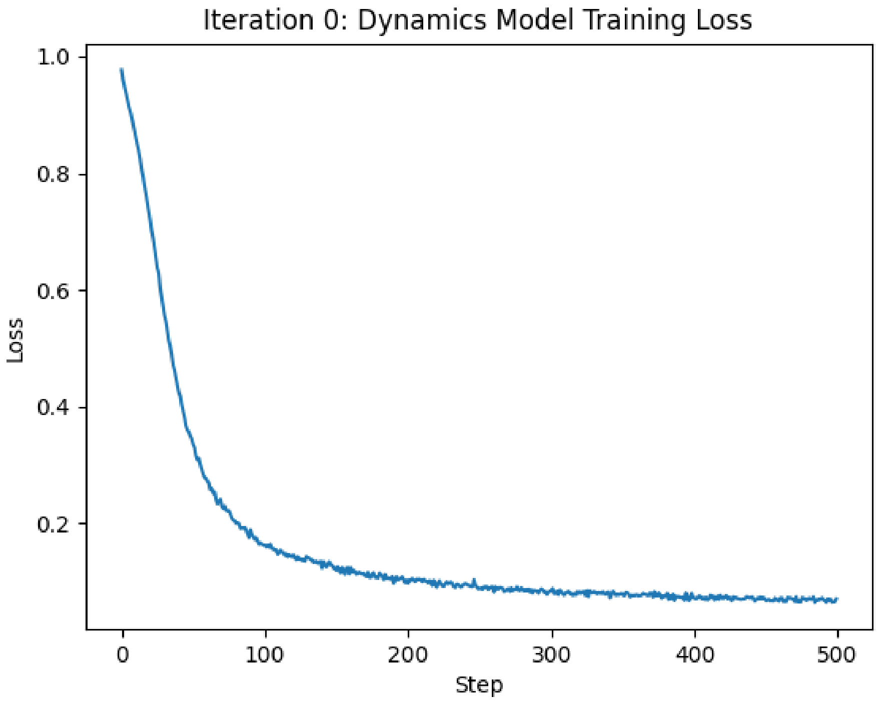


Figure 2: halfcheetah_0_iter; 3 layers

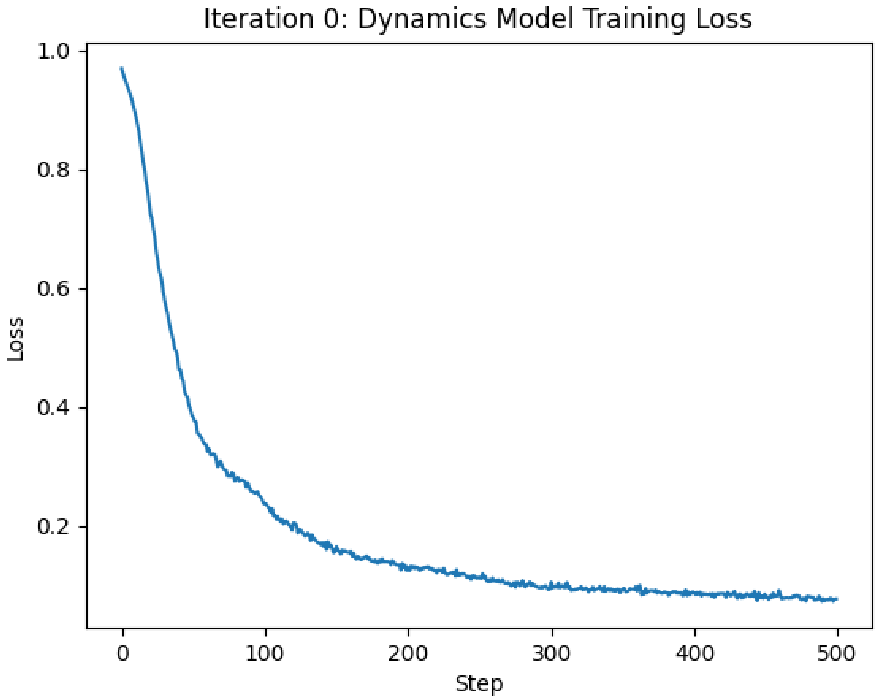


Figure 3: halfcheetah_0_iter; 5 layers

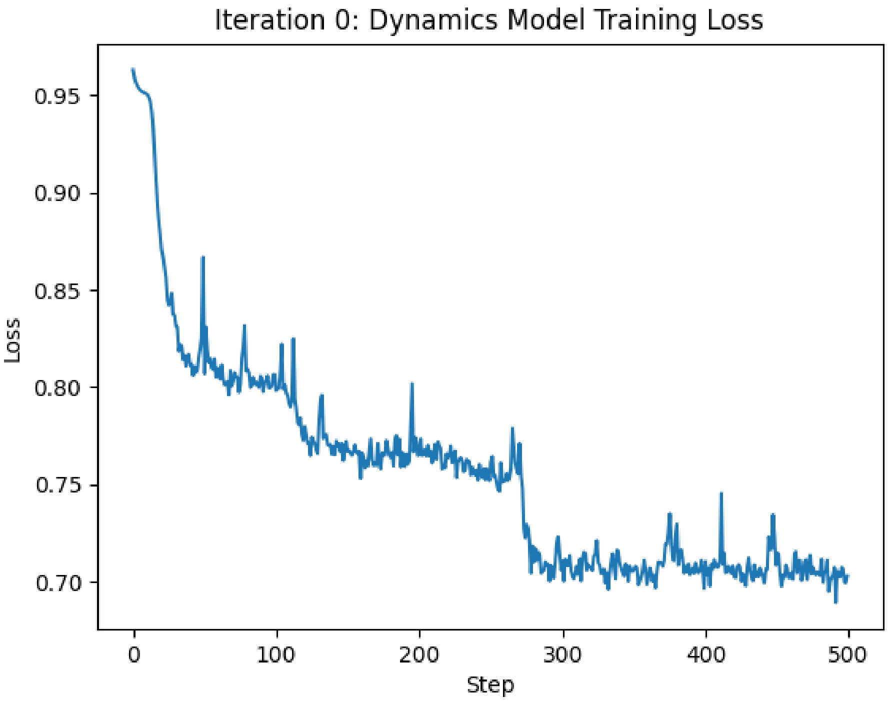
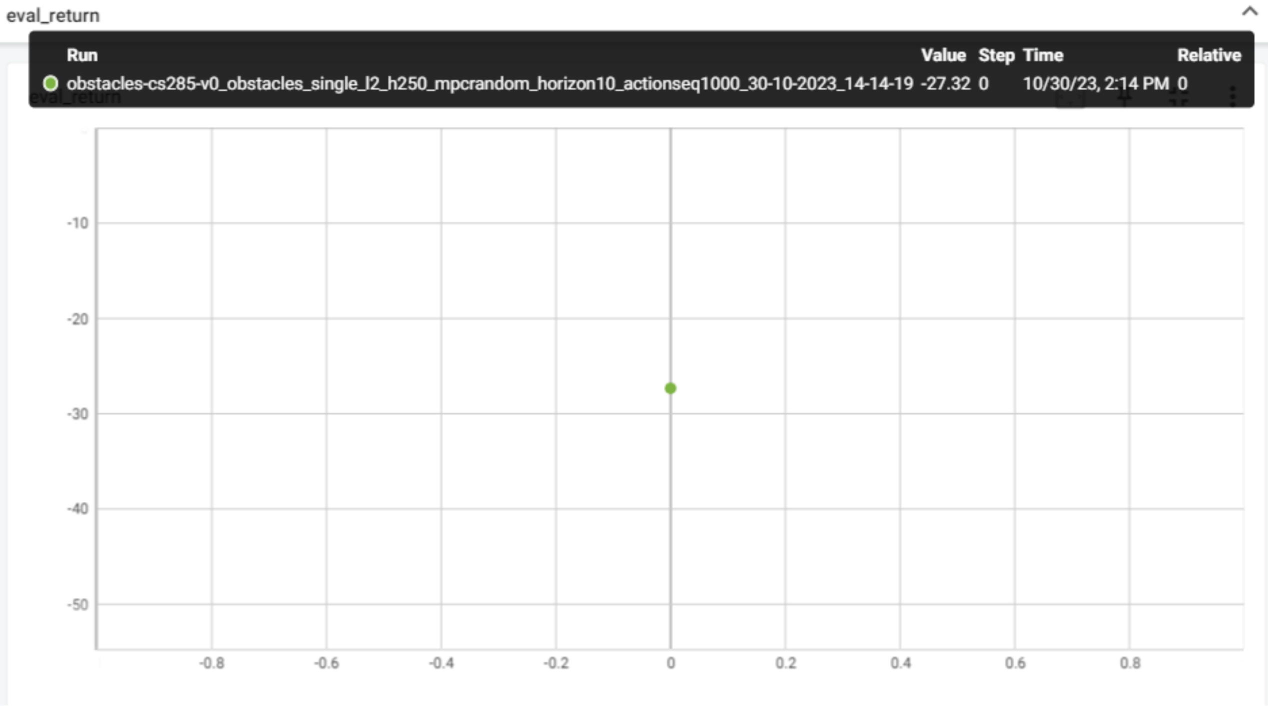


Figure 4: halfcheetah_0_iter; 20 layers

Problem 2


eval_return

	Run	Value	Step	Time	Relative
●	obstacles-cs285-v0_obstacles_single_l2_h250_mpcrandom_horizon10_actionseq1000_30-10-2023_14-14-19	-27.32	0	10/30/23, 2:14 PM	0

Figure 5: obstacles_1_iter.yaml eval is above minus seventy

Problem 3

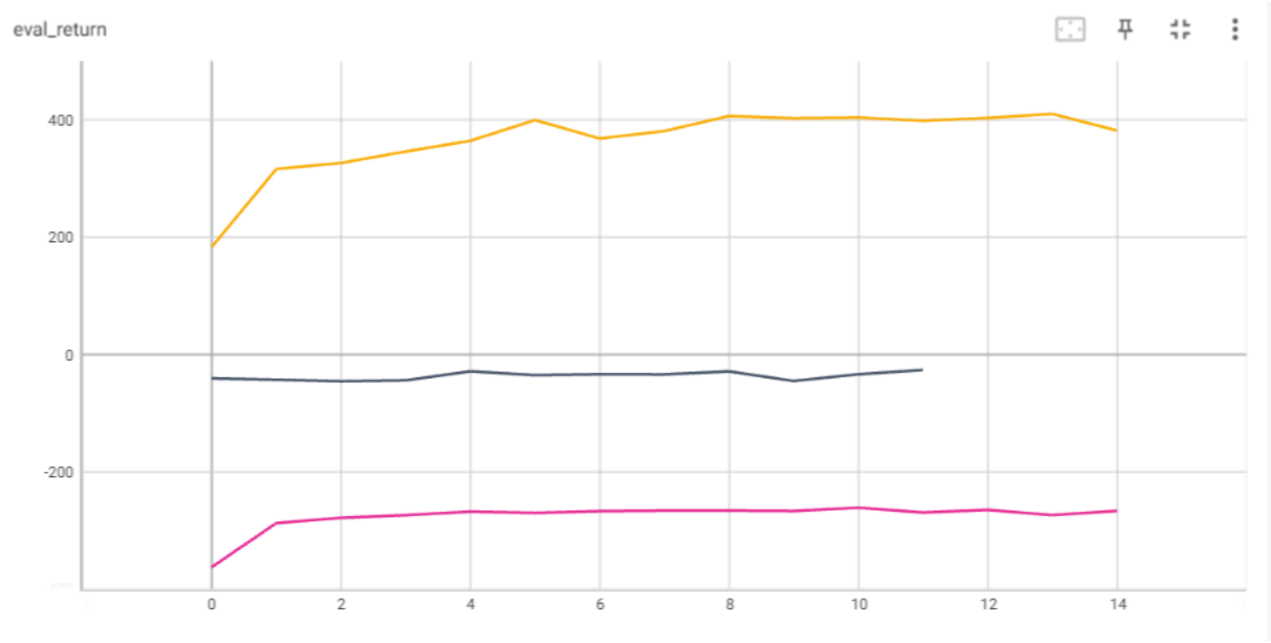


Figure 6: yellow = cheetah, black = obstacles, orange = reacher

Problem 4

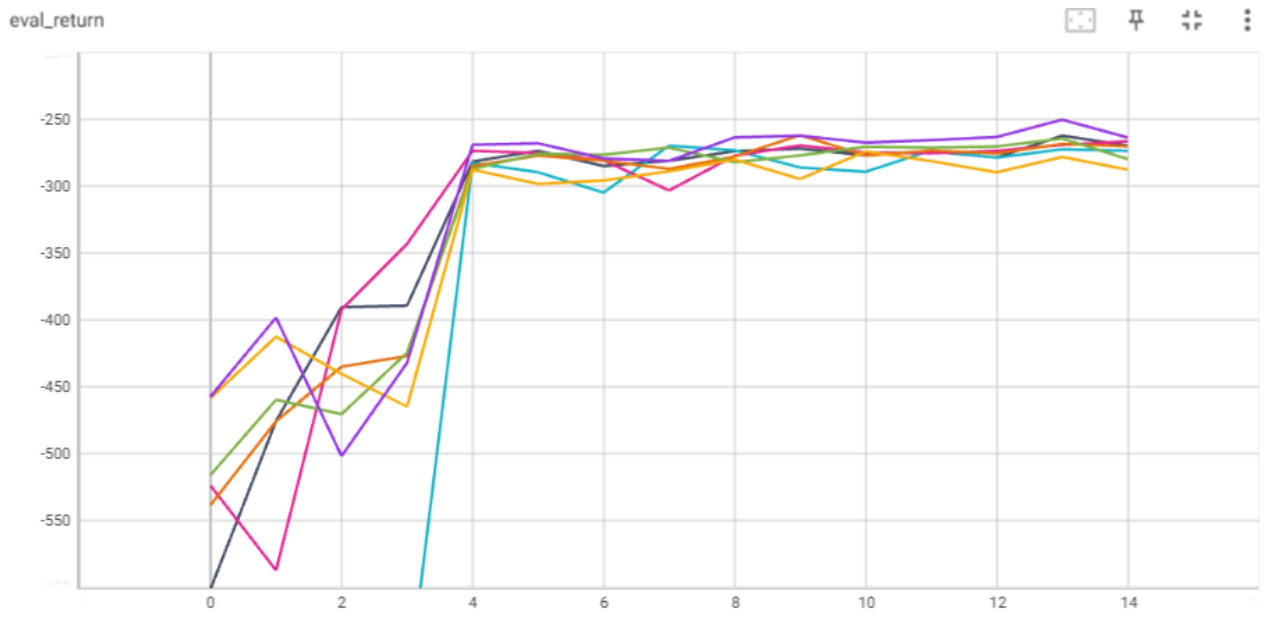


Figure 7: dark blue = default, blue = ensemble size d(ecreased), pink = ensemble size i(increased), orange = horizon i, purple = Horizon d, green = # ac_sequences d, orange (light) = # ac_sequences i

First of all, we can observe that after iteration 4, all of the approaches seem to perform pretty similar. Before that, there is some discrepancy.

- Ensemble size: The default performs very similar compared to the increased one. The decreased one is performing far worse at the beginning. Overall, it has the lowest starting and steepest learning phase.
- Horizon: No big impact. The decreased as well as the increased version perform worse than the default at the beginning. Might be due to randomness though.
- #Action sequences: No big impact/variation.

Problem 5

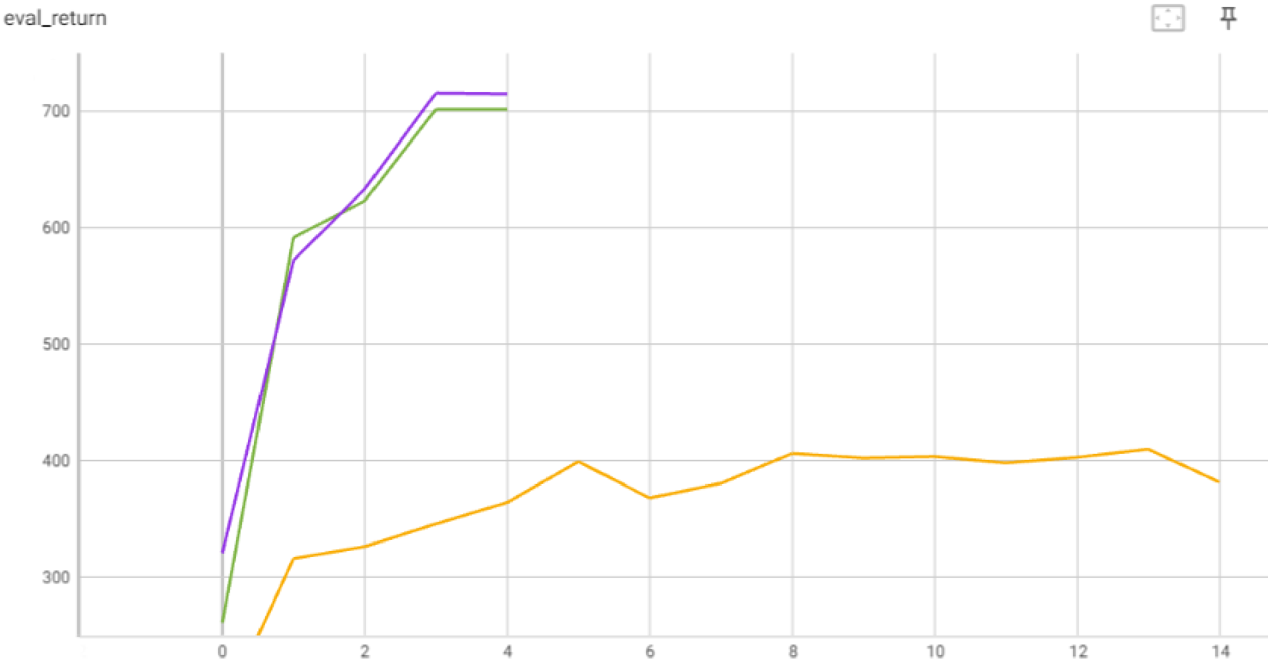


Figure 8: purple = n_iter 2, green = n_iter 4, orange = random shooting

The CEM strategy performs far better than the random shooting approach. The number of iterations does not seem to have a significant impact.

Problem 6

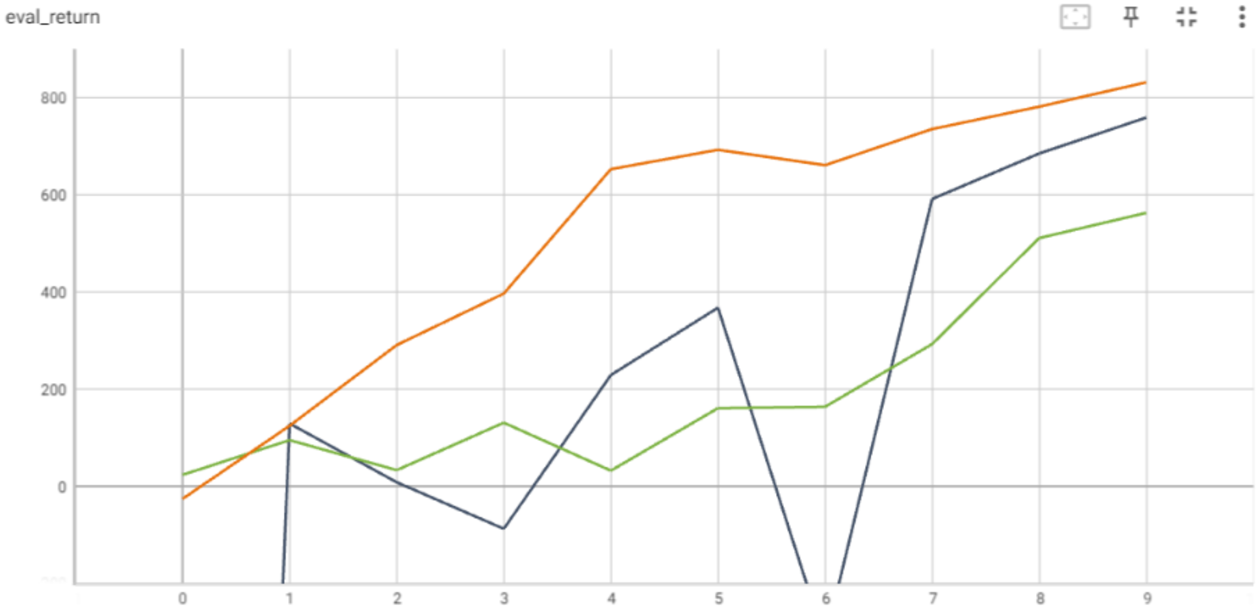


Figure 9: blue = model-free, green = Dyna-like, orange = full MBPO

We can see that the model-free and Dyna-like method perform quite similar, but the model-free method is way more volatile. The full MBPO approach clearly outperforms the other two, reaching higher rewards, doing so quicker and even in a more stable process!