1 Problem 1

1.1 Provide a plot of the dynamics model predictions when the predictions are mostly accurate.

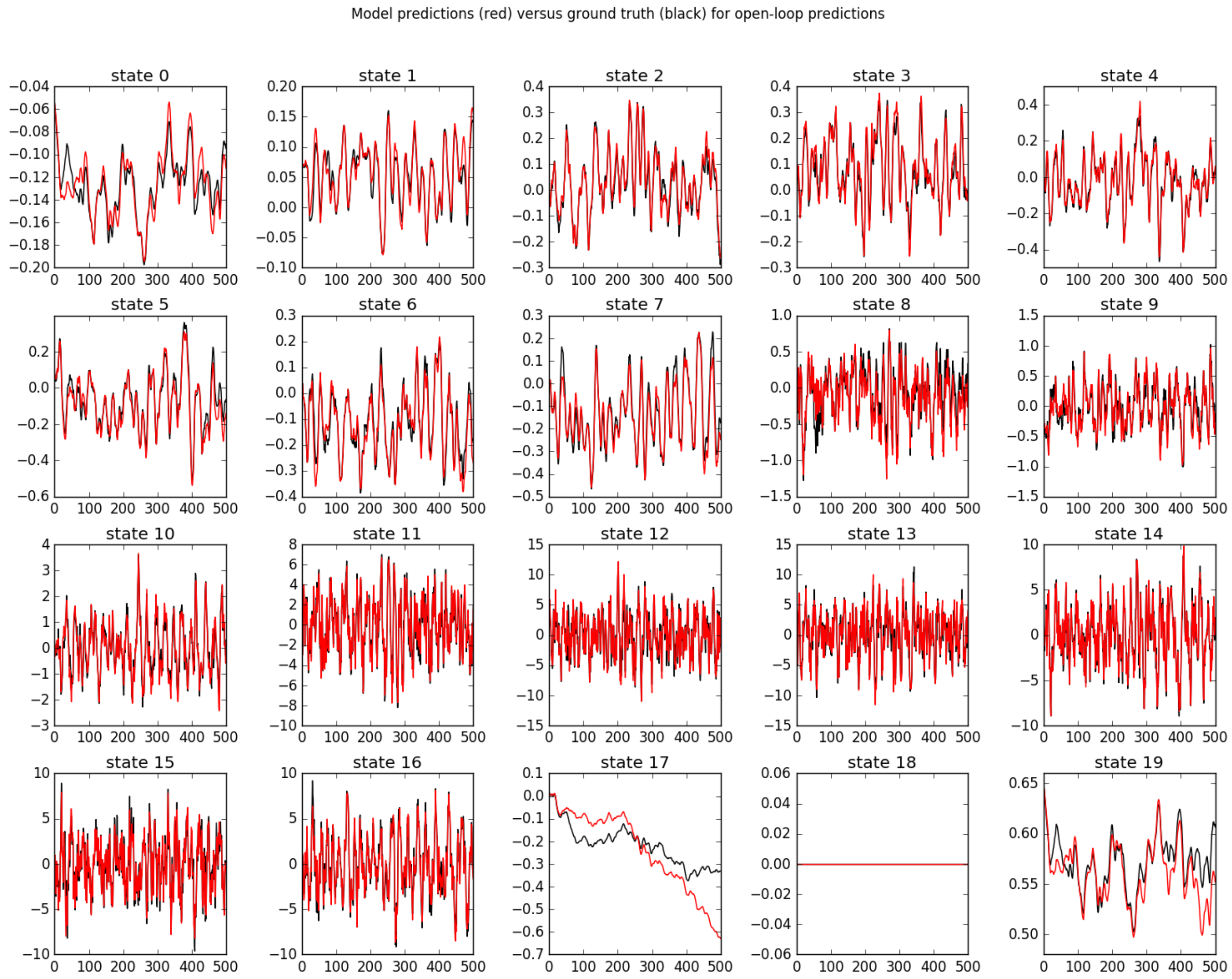

Figure 1: The plot of my most accurate dynamic model predictions.

1.2 For which state dimension are the predictions the most inaccurate? Give a possible reason why the predictions are inaccurate.

The predictions of 17th dimension (state 17) is the most inaccurate. It might be because state 17 rarely changes during random actions. Therefore, the fitted model has not seen changes in state 17 and overfits. It should also have something to do with the instability of open loop control.

2 Problem 2

	ReturnAvg	ReturnStd
Random Policy	-156.836	44.0595
model-based controller (MPC)	13.0374	29.8065

```
10-19 09:19:30 HalfCheetah_q2_19-10-2018_09-19-30 INFO Gathering random dataset
10-19 09:19:31 HalfCheetah_q2_19-10-2018_09-19-30 INFO Creating policy
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO Random policy
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO --------- ---------
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnAvg -156.836
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnMax -92.4933
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnMin -251.131
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnStd 44.0595
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO --------- ---------
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : total  0.0 (0.0%)
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : other  0.0 (0.0%)
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG
10-19 09:19:38 HalfCheetah_q2_19-10-2018_09-19-30 INFO Training policy....
10-19 09:19:41 HalfCheetah_q2_19-10-2018_09-19-30 INFO Evaluating policy...
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO Trained policy
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ----------------- -----------
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnAvg       13.0374
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnMax       53.6981
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnMin      -32.5848
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ReturnStd       29.8065
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO TrainingLossFinal 0.0275348
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO TrainingLossStart 0.999107
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 INFO ----------------- -----------
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : total  419.0 (100.0%)
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : get action 414.5 (98.9%)
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : train policy 3.3 (0.8%)
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : env step 0.7 (0.2%)
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG : other  0.6 (0.1%)
10-19 09:26:37 HalfCheetah_q2_19-10-2018_09-19-30 DEBUG
```

3 Problem 3a

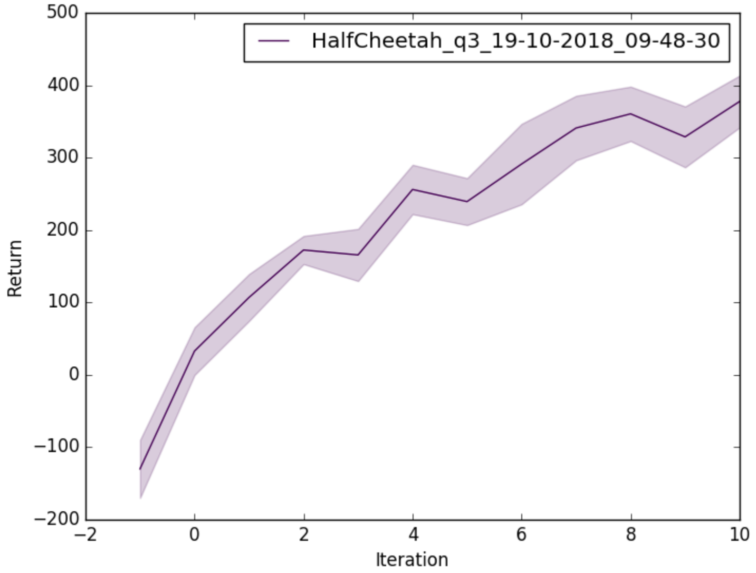

Figure 2: The returns versus iteration of model-based controller (MPC).

4 Problem 3b

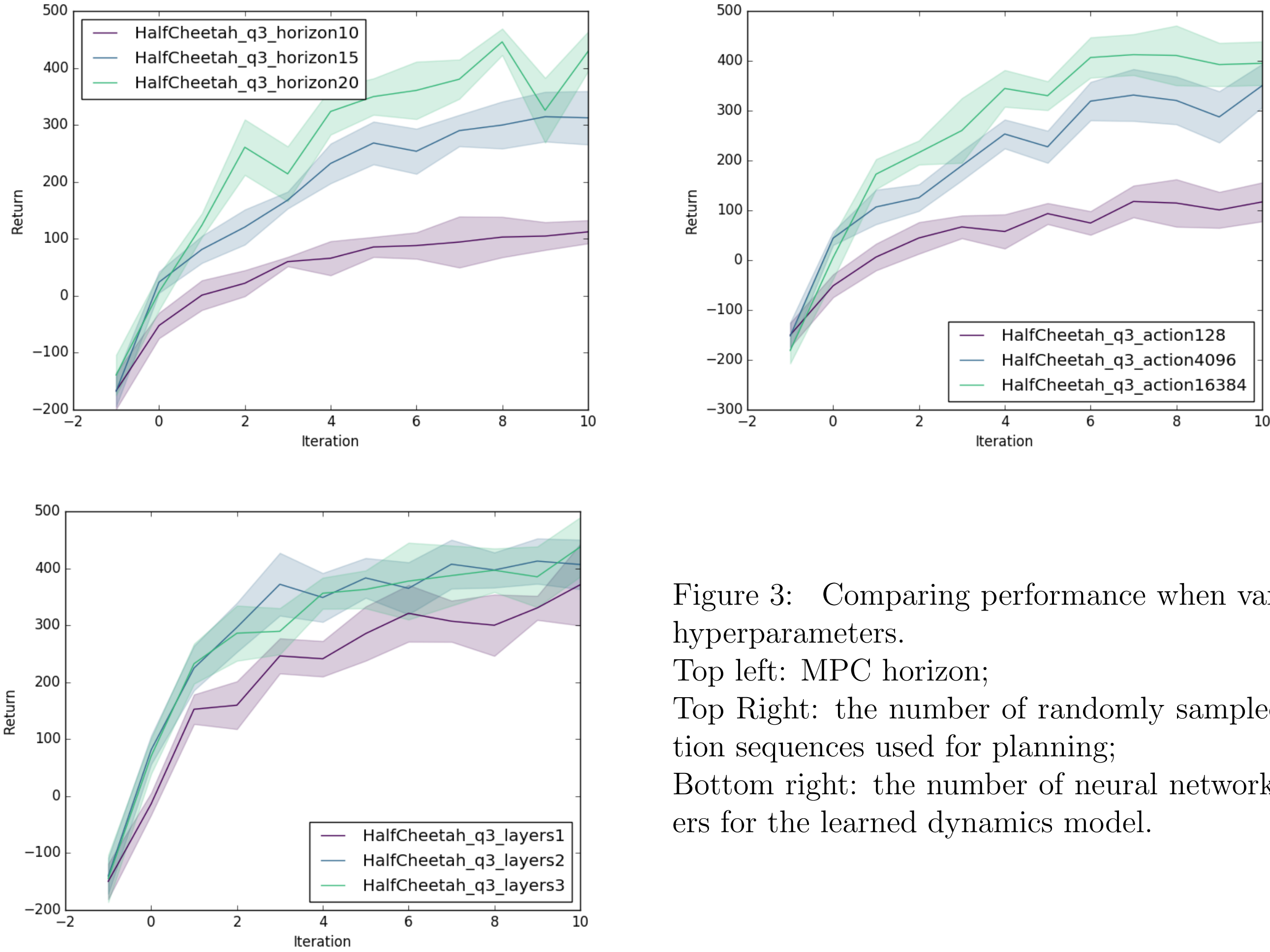

Figure 3: Comparing performance when varying hyperparameters.
Top left: MPC horizon;
Top Right: the number of randomly sampled action sequences used for planning;
Bottom right: the number of neural network layers for the learned dynamics model.

5 bonus

Instead of performing action selection using random action sequences, use the Cross Entropy Method (CEM). (See pseudo-code on Wikipedia.) How much does CEM improve performance?

In my result, the cross entropy method doesn't improve the result at all. I think is because the action space is small enough and our number of sample actions is large enough. This makes uniform sampling enough to cover most of the action space and gives good results.
At later iterations, CEM overfits to the original low cost region which makes it harder to find new low cost regions.

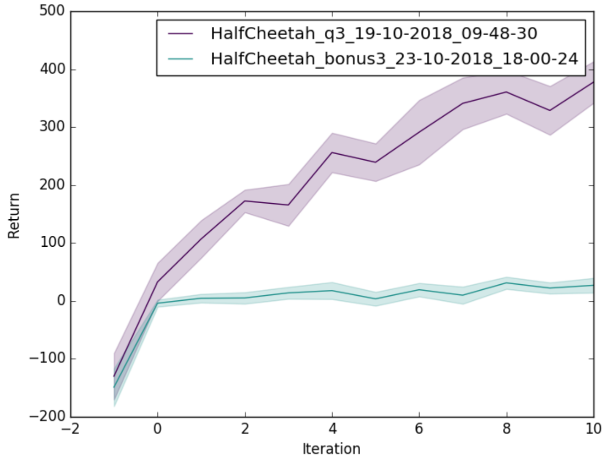

Figure 4: Comparing performance of random sampling and cross entropy method.
Violet: random sampling;
Cyan: cross entropy method (CEM)

```
10-20 17:35:50 HalfCheetah_q2_bonus INFO Gathering random dataset
10-20 17:35:51 HalfCheetah_q2_bonus INFO Creating policy
10-20 17:36:14 HalfCheetah_q2_bonus INFO Random policy
10-20 17:36:14 HalfCheetah_q2_bonus INFO --------- --------
10-20 17:36:14 HalfCheetah_q2_bonus INFO ReturnAvg -147.093
10-20 17:36:14 HalfCheetah_q2_bonus INFO ReturnMax -102.979
10-20 17:36:14 HalfCheetah_q2_bonus INFO ReturnMin -186.081
10-20 17:36:14 HalfCheetah_q2_bonus INFO ReturnStd  27.542
10-20 17:36:14 HalfCheetah_q2_bonus INFO --------- --------
10-20 17:36:14 HalfCheetah_q2_bonus DEBUG
10-20 17:36:14 HalfCheetah_q2_bonus DEBUG : total   0.0 (100.0%)
10-20 17:36:14 HalfCheetah_q2_bonus DEBUG : other   0.0 (100.0%)
10-20 17:36:14 HalfCheetah_q2_bonus DEBUG
10-20 17:36:14 HalfCheetah_q2_bonus INFO Training policy....
10-20 17:36:16 HalfCheetah_q2_bonus INFO Evaluating policy...
10-20 17:54:08 HalfCheetah_q2_bonus INFO Trained policy
10-20 17:54:08 HalfCheetah_q2_bonus INFO ----------------- -----------
10-20 17:54:08 HalfCheetah_q2_bonus INFO ReturnAvg        -11.5704
10-20 17:54:08 HalfCheetah_q2_bonus INFO ReturnMax         18.5725
10-20 17:54:08 HalfCheetah_q2_bonus INFO ReturnMin        -31.9431
10-20 17:54:08 HalfCheetah_q2_bonus INFO ReturnStd         13.4393
10-20 17:54:08 HalfCheetah_q2_bonus INFO TrainingLossFinal  0.0298326
10-20 17:54:08 HalfCheetah_q2_bonus INFO TrainingLossStart  1.04421
10-20 17:54:08 HalfCheetah_q2_bonus INFO ----------------- -----------
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG : total   1074.8 (100.0%)
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG : get action 1070.9 (99.6%)
```

```
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG : train policy 2.7 (0.3%)
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG : env step 0.8 (0.1%)
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG : other   0.3 (0.0%)
10-20 17:54:08 HalfCheetah_q2_bonus DEBUG
```