OLCAR - Exercise 3 – Reinforcement Learning

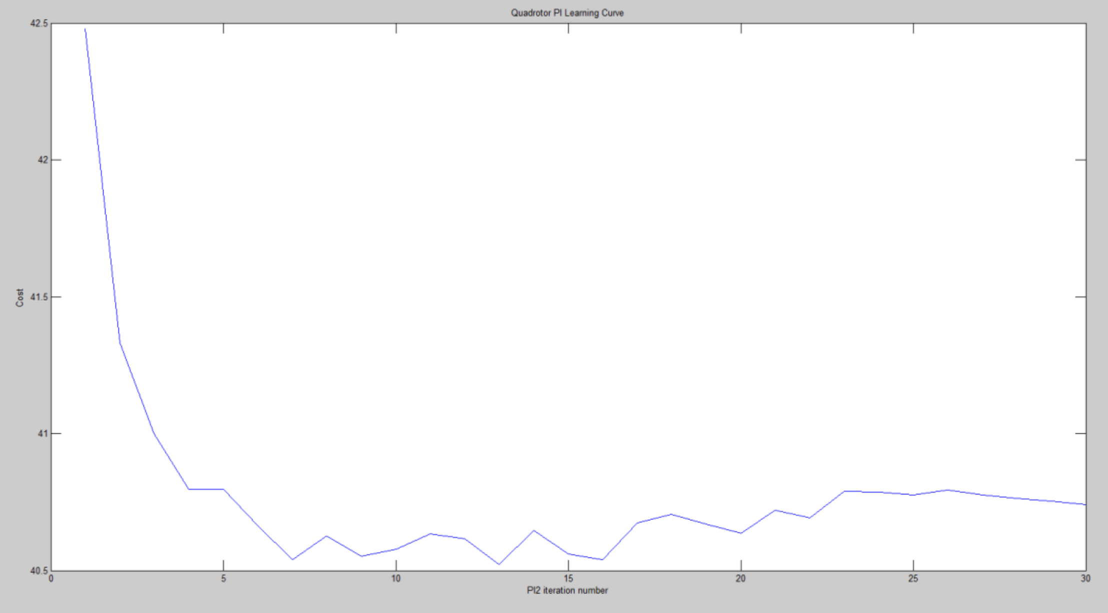
Answers to question related to programming exercise

Handout: 12.05.2015
Due: 26.05.2015

Yi Hao Ng, Raphael Stadler
(Team Number 10)

1. How much cost improvement did you obtain using PI2 learning?

By using PI2 the cost could be reduced by approximately 5 % (from 42.5 to 40.5).



2. How does the exploration noise (Task.std_noise) affect the learning curve? What happens if you decrease/increase it?

If the noise is chosen to be too big (e.g. to 0.015), in the part of the rollout calculations, the integration using ode45 fails. Decreasing the exploration noise (e.g. to 0.00015) makes the algorithm converge slower or doesn't allow the algorithm to converge at all in the allowed number of iterations.

3. The tuning parameter Task.num_reuse specifies how many (of the best) rollouts are saved, carried over and reused in the next learning iteration. Why does it make sense to keep some of the best rollouts for the next update?

In the algorithm, we performed importance sampling (sampling in areas with lower accumulated costs). This importance sampling is done K times (#rollouts) in each learning iteration. To further improve the quality of the optimal control estimation, we keep the important rollouts (with good sample weightings alpha) for the next iteration. This makes it more probable to improve our optimal control estimation in the next update by using the estimated parameters. The fact that we are exploring more often in more important areas makes the algorithm converge faster.

4. How does the quality of your initial guess affect the PI2 learning? For example, what happens if you limit your ILQC iterations to only 1?

The performance of the PI2 learning is directly affected by the quality of the initial guess. If the initial guess is of bad quality, the cost of the initial guess is really high (as it is e.g. the case when only 1 ILQC iteration is performed, which even makes the quadrotor "crash"). Having such a high-cost initial guess causes the algorithm not to be able to converge (in the allowed maximum number of iterations) and the cost may oscillate a lot throughout the iterations.

5. While executing your program, you might have noticed that the cost is not always strictly decreasing during learning. What is your explanation for this behaviour?

During the PI2 algorithm, when choosing the delta theta (the difference of the basis function parameters) the probabilistic exploration noise is included in the calculations. Because of this stochasticity, exploration gets possible. For such an exploration, it is possible that the found solution is not as good as the one from the previous iteration.