Tópicos especiais em Inteligência Computacional II:
Aprendizado por Reforço
Exercício 5: Indirect reinforcement learning

Wouter Caarls
wouter@ele.puc-rio.br

May 23, 2019

Indirect reinforcement learning uses a learned transition model to generalize experience in the space of the system dynamics. The goal of this exercise is to investigate its effectiveness for the pendulum swing-up problem.

1 Exercise

Exercise 1.1 Adapt the State-Value Actor Critic algorithm from Exercise 4 to keep track of all experienced transitions, and use these at the end of every episode to train a neural network on the system dynamics. If you used a different algorithm for Exercise 4, an example solution can be found in `ac.m`. If you prefer another reinforcement learning algorithm or supervised learning technique, you may apply those as well.

Exercise 1.2 Use random simulated transitions to create a testing dataset to characterize the network's accuracy as a function of the number of real episodes of training data.

Exercise 1.3 Use the trained neural network to generate simulated experience in order to update the value function/policy. Try both generating random experience and simulated trajectories of a certain length (say, max 20 steps).

Exercise 1.4 Create a testing harness to systematically investigate the algorithm's performance by running 10 experiments and calulating the mean and standard deviation of the rise time and end performance. Define the rise time as the first time the performance is higher than -1000 three episodes in a row. Use `errorbaralpha` to plot the mean and standard error, defined as $\frac{\sigma}{\sqrt{N}}$ ($N = 10$).

Exercise 1.5 Using your testing harness, investigate the sample complexity vs computational complexity trade-off by varying the amount of simulated experience per episode. Do this for both the random experience and simulated trajectories. Use `tic` and `toc` to measure computation time.

Exercise 1.6 Analyze the results. Are there diminishing returns? Is there also a trade-off between end performance and rise time? Is there a significant difference between random experience and trajectories? Etc.