RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
COMPUTER SCIENCE VI
AUTONOMOUS INTELLIGENT SYSTEMS
Prof. Dr. Sven Behnke
Endenicher Allee 19a

Robot Learning

Assignment 6

Due Tuesday, May 21th, before class.

6.1) The dynamics of a cart-pole system is given by the following equations:

$$\ddot{x} = \frac{F - m_p\, l\, (\ddot{\theta}\, \cos\theta - \dot{\theta}^2\, \sin\theta)}{m_c + m_p}$$

$$\ddot{\theta} = \frac{g\, \sin\theta\, (m_c + m_p) - (F + m_p l\, \dot{\theta}^2\, \sin\theta) \cos\theta}{\frac{4}{3} l\, (m_c + m_p) - m_p\, l\, \cos^2\theta}$$

The parameters are l = 0.8 (half length of pole), m_c = 6 (mass of cart), m_p = 3 (mass of pole), g = 9.81m/s^2 (gravity), and −100N ≤ F ≤ +100N (force applied to the cart).
The control interval shall be 0.01s.
A zero-mean Gaussian noise vector ξ with diagonal covariance matrix Σ=diag(0.004, 0.04, 0.001, 0.01) shall be added after each control interval to the state vector (position, speed, angle, angular speed).

Implement a discrete-time simulator for this system.
Visualize, how the state evolves over 1s for the initial conditions
(position = -1m, velocity = 0.25m/s, angle = 0.3rad, angular velocity = -0.7 rad/s)
when no force is applied (F=0).

5 points

6.2) Find a linear (saturated) state-feedback policy

F = min(100, max(–100, k_1*position + k_2*velocity + k_3*angle + k_4*angular_velocity))

that moves the cart from the initial state to the target state region, described by the angle within [−0.05rad, +0.05rad] and the position of the cart within [−0.1m, +0.1m].
The system fails, if the absolute pole angle is larger than 1.0 rad or the absolute cart position is larger than 5 m. In case of a failure, the episode is stopped.
The final reward is computed by −(N − t), where N=1000 gives the maximum episode length and t is the time step, where the failure occurred. This means, that a later failure is better than an earlier failure. In case of the state being within the target region, the reward is 0 and the episode is continued (since the system might leave the target region again). In every other situation, the reward is −1.
Visualize the state trajectory of your policy and compute the return.

5 points

6.3) Improve your initial policy using a policy gradient method.
Document the learning process and the final system behavior.

10 points