Rheinische Friedrich-Wilhelms-Universität Bonn
Institut für Informatik
Abteilung VI
Autonome Intelligente Systeme

Prof. Dr. Sven Behnke

Postanschrift: 53012 Bonn
Sitz: Friedrich-Ebert-Allee 144

Robot Learning

Assignment 8

Due Tuesday, July 11th, before class.

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 = 0.6 (half length of pole), m_c = 2 (mass of cart), m_p = 0.5 (mass of pole), g = 9.81m/s^2 (gravity), and
−50N ≤ F ≤ +50N (force applied to the cart).
The control interval shall be 0.01s.
A zero-mean noise vector ξ with diagonal covariance matrix Σ=diag(0.001, 0.002, 0.001, 0.002) shall be added after each control interval to the state vector (position, speed, angle, angular speed).

8.1) Implement a discrete-time simulator for this system.
Visualize, how the state evolves over 1s for the initial conditions (position = –1m, velocity = 0.2m/s, angle = 0.2rad, angular velocity = –0.4 rad/s) when no force is applied (F=0).

5 points

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

F=min(50, max(–50, 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 [−0.1rad, +0.1rad] and the position of the cart within [−0.1m, +0.1m].
The system fails, if the absolute pole angle is larger than 1.0 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

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

10 points