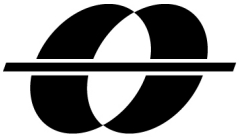

MÄLARDALEN UNIVERSITY
SWEDEN

DVA427 — Assignment 5

Reinforcement Learning Inverted Pendulum

Authors:
Petr Fejfar
Martin Váňa

Datum:
March 17, 2016

Contents

1 Task

In this assignment you will apply the reinforcement learning technique to learn optimal action strategies for automatic control. The plant to be considered is an inverted pendulum system. The constitution of the pendulum system is depicted in the following figure, where a pole is hinged to a cart which moves on a rail to its right and left. The pole has only one degree of freedom to rotate around the hinge point. The controller can apply a "left" or "right" force of fixed magnitude to the cart at discrete time intervals.

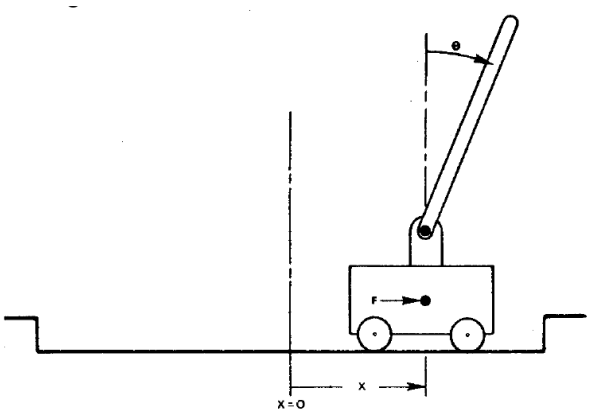

Figure 1: The constitution of the pendulum system

The states of the pendulum system are:

1. position of the cart(x)
2. angle of the pole (θ)
3. cart velocity on the rail ($\dot{x}$)
4. speed of pole rotation ($\dot{\theta}$)

We can use such states to describe the behaviour of the system as a set of non-linear differential equations as

$$\ddot{\theta}(t) = \frac{g\,sin(\theta(t)) + cos(\theta(t))\left[\frac{-F(t)-ml\dot{\theta}(t)^2 sin(\theta(t))}{m_c+m}\right]}{l\left[\frac{4}{3} - \frac{m\,cos(\theta(t))^2}{m_c+m}\right]}$$

$$\ddot{x}(t) = \frac{F(t) + ml\lfloor\dot{\theta}(t)^2 sin(\theta(t)) - \ddot{\theta}(t)cos(\theta(t))\rfloor}{m_c + m}$$

Where

$g = -9.8\frac{m}{s^2}$, acceleration due to gravity

$m_c = 10kg$, mass of the cart

$m = 0.1kg$, mass of the pole

$l = 0.5m$, half of the pole length

$F(t) = +10N$ or $-10N$, force applied to cart's mass center at time t.

The system survives if the cart position is within $[-2.4m, 2.4m]$ and pole angle is not more than 12 degrees. Suppose the initiate state of the system is ($x = 0$, $\dot{x} = 0$, $\theta = 0$, $\dot{\theta} = 0$) and the discrete time interval is 0.02 second. Now your task is to use the reinforcement learning method to learn the best strategy of selecting forces such that the survival time of the system is maximized.

In order to get response of the system by applying a force under a state, you have to do simulation of the real plant in terms of the system equations given above. A simulation function in MATLAB code is available from the "Labs" folder. But you have to program the simulation function if you use other languages in this assignment.

Your report has to clarify your approach and results by covering the following issues:

1. How do you characterize (discretize) states of the world in this learning task?
2. What is the reward function designed?
3. What is the function to be learned?
4. What learning algorithm is used?
5. How do you choose exploratory actions?
6. How many intervals can your system survive by using the learned strategy?

2 Analysis

2.1 State discretization

We have constricted continuous variables with reasonable values and then divided uniformly.

1. position of the cart(x) — constraints $\langle -2.4, 2.4 \rangle$, 8 intervals
2. angle of the pole (θ) — constraints $\langle -12°, 12° \rangle$, 28 intervals
3. cart velocity on the rail ($\dot{x}$) — constraints $\langle -1, 1 \rangle$, 10 intervals
4. speed of pole rotation ($\dot{\theta}$) — constraints $\langle -1, 1 \rangle$, 28 intervals

2.2 Reward function

Our reward function is simple. It returns 1 if the state is safe, 0 otherwise.

2.3 Q function

$Q^*(s, a)$ is the expected pay-off when taking the action a at the state s and following the optimal policy π^*.

$$Q(s, a) = Q(s, a) + \alpha[r^a_{ss'} + \gamma \max_a Q^*(s', a') - Q(s, a)]$$

,where s, s' are states, a, a' are actions, α is learning rate and γ is a discount factor and $r^a_{ss'}$ is a reward.

2.4 Learning algorithm

TODO Petr Fejfar

2.5 Random action selection

TODO Petr Fejfar

2.6 Results

TODO Petr Fejfar

3 Conclusion

TODO Petr Fejfar