Reinforcement Learning – exercise 01

Vien Ngo
Machine Learning & Robotics lab, University of Stuttgart
Universittsstrae 38, 70569 Stuttgart, Germany

April 7, 2016

1 Problem Formulation

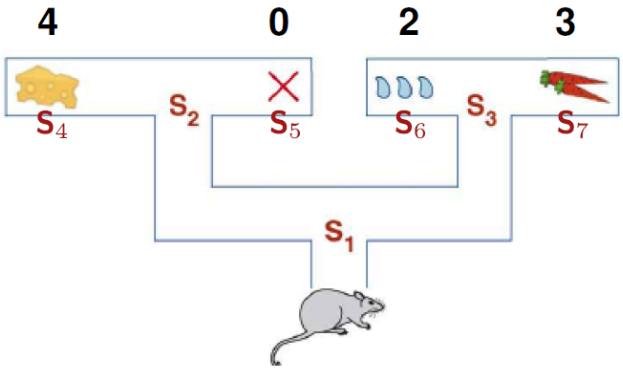
In this exercise you are asked to write an MDP formulation for the following problem.

Assume that you are a manager of a warehouse (with a maximum capacity of W items). Each month t, you know the current inventory (how many items left) in your warehouse. You might have a guess of the exernal demand in the next month $(t+1)$ with a distribution p (the probability that the external demand are j items is $p(D_t = j), j = 0, 1, 2, \ldots$). Based on this information, you decide to order additional items from a supplier. The cost might come from the storing cose of items in warehouse, and the penalties when you can not satisfy the external demand (this clearly comes from the fact that you lost the potential sales). Your objective is to maximize the profit. Use your own parameters for fixed costs to buy and store for each item and a fixed selling price.

Hint: Decision epochs are made at the beginning of each month, hence all events (more items arrive, fill external orders) would make states change. Actions are the amount of an order.

2 Optimal Policy

Consider the following T-maze:



We distinguish 7 states $s_1, .., s_7$ in the maze. The first 3 states are the T-junctions; only the transition to the last 4 states S_4, S_5, S_6, S_7 receive rewards $(4, 0, 2, 3)$, respectively. These 4 states are terminal states. At each T-junction we have two possible actions: left, right. Assume that the transition function is deterministic.

1. Write the above problem's MDP formulation
2. The expected discounted reward in slide 15 is with respect to the stochasticity of the policy $\pi(a|s)$ and the transition $T(s, a, s')$. Given that at state S_1 the mouse selects left with a probability of 0.5 and right with a probability of 0.5; at S_2, S_3 it selects left with a probability of 1.0. You are asked to compute η_γ.