Reinforcement Learning Lecture: Homework 01 Solution

April 14, 2016

1 Exercise 1

This exercise was described in the book of Putterman (IMarkov Decision Processes: Discrete Stochastic Dynamic Programming).

The MDP formulation is $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, R, P_0\}$

As the capacity of the warehouse is W so it is clear that: the state space is $\mathcal{S} = \{0, 1, \ldots, W\}$, and the action space $\mathcal{A} = \{0, 1, 2 \ldots, W\}$.

The reward term $R(s_t, a_t)$ consists of three components:

- the cost of buying a_t items are $Buy(a_t)$
- cost for storing $(s_t + a_t)$. This cost is fixed and presumbly it is equal to $Store(s_t + a_t)$.
- the lost profit when the demand D_t is larger than $(s_t + a_t)$. Assume that the penalty is c.
- the selling price of D_t items, $f(D_t)$

The last two components are not clear yet since we do not know the exact value of D_t. More specifically, the third term can be computed as $Lost(s_t + a_t) = p(D_t \geq s_t + a_t) \times c$ where $p(D_t \geq s_t + a_t) = \sum_{d=s_t+a_t}^{\infty} p(D_t = d)$.

The fourth term is

$$Sell(s_t + a_t) = \sum_{d=0}^{s_a+a_t} p(D_t = d) f(d)$$

In summary, the reward function is

$$R(s_t, a_t) = Sell(s_t + a_t) - Buy(a_t) - Store(s_t + a_t) - Lost(s_t + a_t)$$

The transition function $T(s' = j|s = i, a)$ has three cases:

- if $j > i + a$, then $T(s' = j|s = i, a) = 0$. That means after even after sale the remaining in the warehouse can not exceed the current capacity.
- if $j \leq i + a$ and $j > 0$, that means the demands at time t does not exceed the capacity. Hence $T(j|i, a) = p(D_t = i + a - j)$.
- if $j = 0$, that means the demand is equal to or exceeds the capacity. Hence

$$T(j|i, a) = p(D_t \geq i + a) = \sum_{d=i+a}^{\infty} p(D_t = d)$$

Note: One could compute R in another form $R(s, a, s')$. In this way the demand D_t can be inferred as either $D_t = s + a - s'$ or $D_t \geq s + a$, hence you can compute all the cost and reward terms more easily.