MDP and RL Tutorial – Solutions

3.1 One form of MDP formulation $\{S, A, T, R\}$ is as follows.
State space is $S = \{0,1,2, \dots, W\}$
Action space $A = \{0,1,2, \dots, W\}$
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 presumably it is equal to $Store(s_t + a_t)$.
- Assume the selling price of D_t items is $f(D_t)$. The total sale price 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

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

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

- If $j > i + a$, then $T(j|s = i, a) = 0$. That means even after sale the remaining in the warehouse cannot 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
$$T(j|i, a) = p(D_t \geq i + a) = \sum_{d=i+a}^{\infty} p(D_t = d)$$

3.2 (a) Apply the Bellman backups $V_{i+1}(s) = \max_a(\sum_{s'} T(s,a,s')(R(s,a,s') + \gamma V_i(s')))$ twice. We just show the computation for the max actions. Most of the terms will be zero, which are omitted here for compactness.

$S =$	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
$V_0(S) =$	0	0	0	0	0	0
$V_1(S) =$	0	0	0	0	$0.8 \times 5 = 4.0$	0
$V_2(S) =$	0	$0.9 \times 0.8 \times 4 + 0.1 \times -5 = 2.38$	0	$0.8 \times 0.9 \times 4.0 = 2.88$	$0.8 \times 5 = 4.0$	0

(b) The agent must be able to explore the world by taking actions and observing the effects.

(c) To compute the estimates, average the rewards received in the trajectories that went through the indicates states.

$$V((1,1)) = ((-5+5+5))/3 = 5/3 = 1.666$$
$$V((2,2)) = ((5+5))/2 = 5$$

(d) The general Q-learning update is:

$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha[r + \gamma \max_{a'} Q_{old}(s',a') - Q_{old}(s,a)]$$

After trial 1, all of the updates will be zero, expect for:

$$Q\big((1,2), right\big) = 0 + .1\,(-5 + 0.9 \times 0 - 0) = -0.5$$

After trial 2, the non-zero updates will be:

$$Q\big((1,2), right\big) = -0.5 + .1(0 + 0.9 \times 0 - (-0.5)) = -0.45$$
$$Q\big((2,2), right\big) = 0 + .1(5 + 0.9 \times 0 - 0) = 0.5$$