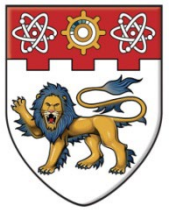


NANYANG TECHNOLOGICAL UNIVERSITY

Artificial Intelligence

MDP and RL

Lu Shijian

Assoc Prof, SCSE

Shijian.Lu@ntu.edu.sg
N4-02C-101

Problem01

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 external 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 cost of items in warehouse. 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.

Please write a MDP formulation for the above problem.

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.

Problem01

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:

- 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

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

In summary, the reward function becomes

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

Problem01

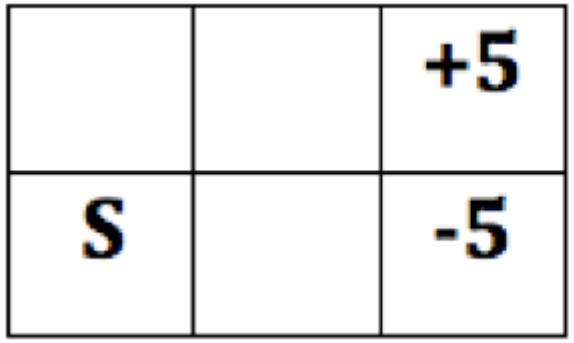
The transition function $T(s' = j|s = i, a)$ has three cases:

- If $j > i + a$, then $T(j|s = i\ , a) = 0$. That means after sale, the remaining in the warehouse cannot be more than 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

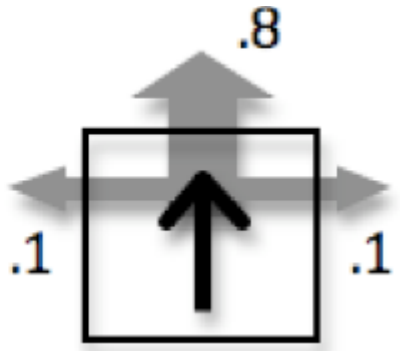
$$T(j|i, a) = p(D_t \geq i + a) = \sum_{d=i+a}^{\infty} p(D_t = d)$$

Problem02

This Gridworld MDP operates like to the one we saw in class. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability .8. With probability .1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.



Gridworld MDP



Transition function

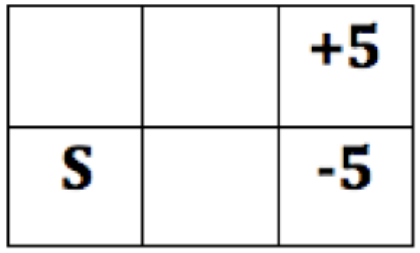
Problem02

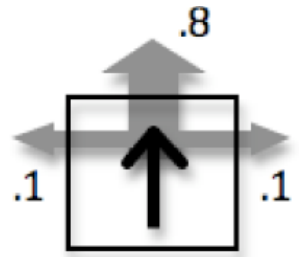
a) Suppose the agent knows transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V0 is 0 everywhere and compute V*i* for times *i* = 1, 2). (Assume values of termination states ((1, 3), (2, 3)) are always 0).

   This is a MDP problem,

   - We have 6 states: $S = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\}$
   - We have 4 actions at each state: $A = \{North, South, West, East\}$
   - The reward discount is $\gamma = 0.9$
   - V0 is 0 for all states: $V_0\{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3)\} = 0$
   - Values of termination states are always 0: $V_{0,1,2}\big((1,3),(2,3)\big) = 0$

Problem02





a) Suppose the agent knows transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V0 is 0 everywhere and compute V*i* for times *i* = 1, 2). (Assume values of termination states ((1, 3), (2, 3)) are always 0).

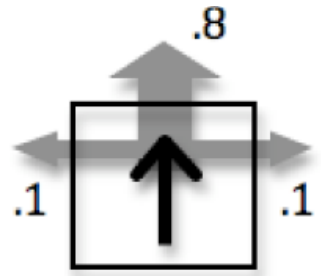
By applying the Bellman equation in two iterations

$$V(s) = max_{a \in A} \sum_{s'} p(s'|s,a)[r(s,a,s') + \gamma V(s')]$$

$S =$	(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
$V_2(S) =$	0	$0.8 \times 0.9 \times 4 + 0.1 \times -5 = 2.38$	0	$0.8 \times 0.9 \times 4 = 2.88$	$0.8 \times 5 + 0.1(0 + 0.9 * 4) = 4.36$	0

Problem02

		+5
S		-5



$$Q((2,2),N) \leftarrow \underbrace{P((2,1)|(2,2),N)}_{0.1} \cdot \left(\underbrace{R((2,2),N,(2,1))}_{0} + 0.9 \cdot \underbrace{V((2,1))}_{0} \right)$$

$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,2)|(2,2),N)}_{0.8} \cdot \left(\underbrace{R((2,2),N,(2,2))}_{0} + 0.9 \cdot \underbrace{V((2,2))}_{0} \right)$$

$$0.8 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,3)|(2,2),N)}_{0.1} \cdot \left(\underbrace{R((2,2),N,(2,3))}_{5} + 0.9 \cdot \underbrace{V((2,3))}_{0} \right)$$

$$0.1 \times (5 + 0.9 \times 0) = 0.5$$

$$Q((2,2),S) \leftarrow \underbrace{P((1,2)|(2,2),S)}_{0.8} \cdot \left(\underbrace{R((2,2),S,(1,2))}_{0} + 0.9 \cdot \underbrace{V((1,2))}_{0} \right)$$

$$0.8 \times (0 + 0.9 \times 0) = 0$$

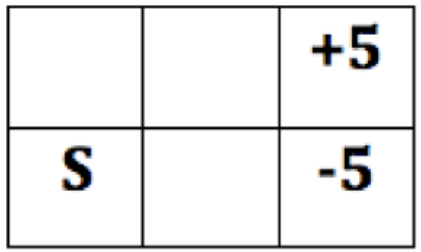
$$+ \underbrace{P((2,1)|(2,2),S)}_{0.1} \cdot \left(\underbrace{R((2,2),S,(2,1))}_{0} + 0.9 \cdot \underbrace{V((2,1))}_{0} \right)$$

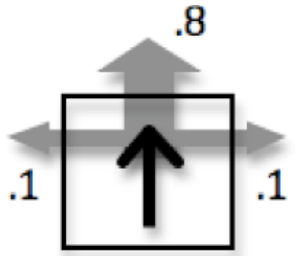
$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,3)|(2,2),S)}_{0.1} \cdot \left(\underbrace{R((2,2),S,(2,3))}_{5} + 0.9 \cdot \underbrace{V((2,3))}_{0} \right)$$

$$0.1 \times (5 + 0.9 \times 0) = 0.5$$

Problem02





$$Q((2,2),W) \leftarrow \underbrace{P((1,2)|(2,2),W)}_{0.1} \cdot \left(\underbrace{R((2,2),W,(1,2))}_{0} + 0.9 \cdot \underbrace{V((1,2))}_{0} \right)$$

$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,1)|(2,2),W)}_{0.8} \cdot \left(\underbrace{R((2,2),W,(2,1))}_{0} + 0.9 \cdot \underbrace{V((2,1))}_{0} \right)$$

$$0.8 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,2)|(2,2),W)}_{0.1} \cdot \left(\underbrace{R((2,2),W,(2,2))}_{0} + 0.9 \cdot \underbrace{V((2,2))}_{0} \right)$$

$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$Q((2,2),E) \leftarrow \underbrace{P((1,2)|(2,2),E)}_{0.1} \cdot \left(\underbrace{R((2,2),E,(1,2))}_{0} + 0.9 \cdot \underbrace{V((1,2))}_{0} \right)$$

$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,2)|(2,2),E)}_{0.1} \cdot \left(\underbrace{R((2,2),E,(2,2))}_{0} + 0.9 \cdot \underbrace{V((2,2))}_{0} \right)$$

$$0.1 \times (0 + 0.9 \times 0) = 0$$

$$+ \underbrace{P((2,3)|(2,2),E)}_{0.8} \cdot \left(\underbrace{R((2,2),E,(2,3))}_{5} + 0.9 \cdot \underbrace{V((2,3))}_{0} \right)$$

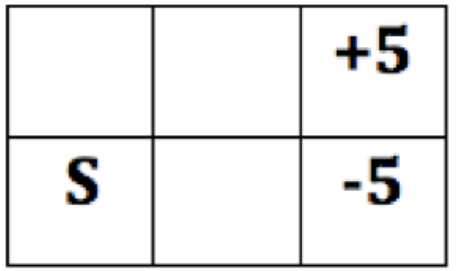
$$0.8 \times (5 + 0.9 \times 0) = 4$$

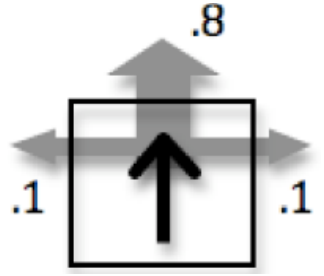
Problem02

b) Suppose the agent does not know the transition probabilities. What does it need to be able do (or have available) in order to learn the optimal policy?

This is a Monte-Carlo problem or Q-learning problem. Under the specified scenario the agent must be able to explore the world by taking actions and observing the effects.

Problem02





c) The agent starts with the policy that always chooses to go right, and executes the following three trials: 1) (1,1) – (1,2) – (1,3), 2) (1,1) – (1,2) – (2,2) – (2,3), and 3) (1,1) – (2,1) – (2,2) – (2,3). What are the Monte Carlo estimates for states (1,1) and (2,2), given these traces (assuming that the discount factor is 1)?

It is a Monte-Carlo problem. MC uses experience to learn an empirical state value function

$$V_\pi(s) = \frac{1}{N}\sum_{i=1}^{N} G_{i,s}$$

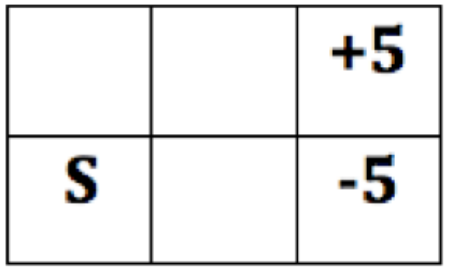
To compute estimates of two states, average the discounted rewards received in the three trajectories that went through the indicates states.

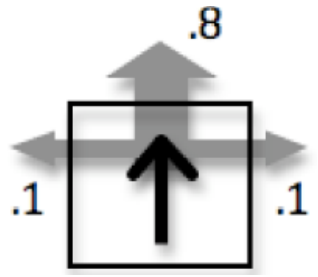
$V((1,1)) = (-5 + 5 + 5)/3 = 5/3 = 1.666$

$V((2,2)) = (5 + 5)/2 = 5$

Problem02

	(1,1)	(1,2)	(1,2)	(2,1)	(2,2)	2,3)
N						
S						
W						
E						





d) Using a learning rate of .1 and assuming initial values of 0, what updates does the Q-learning agent make after trials 1 and 2 above?

The general Q-learning update is:

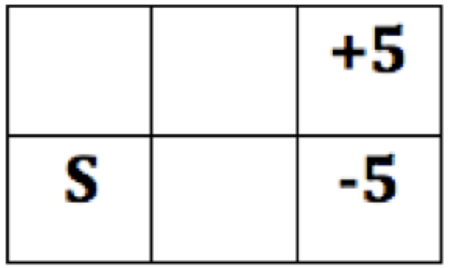
$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha[r + \gamma max_{a'} Q_{old}(s',a') - Q_{old}(s,a)]$$

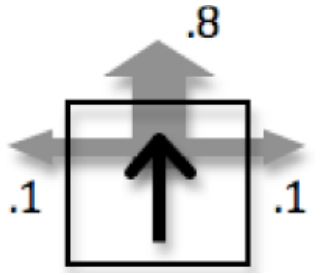
After trial 1 ((1,1) – (1,2) – (1,3)), all of the updates will be zero, except for:

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

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

Problem02





d) Using a learning rate of .1 and assuming initial values of 0, what updates does the Q-learning agent make after trials 1 and 2 above?

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 2 ((1,1) – (1,2) – (2,2) – (2,3)), the non-zero updates will be:

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

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

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