# Dynamic Programming

- Planning by dynamic programming assumes full knowledge of the MDP

For Prediction / Evaluation
* Input : MDP $\langle S, A, P, R, \gamma \rangle$ and policy $\pi$
* Output : Value function $v_\pi$

For Control
* Input : MDP $\langle S, A, P, R, \gamma \rangle$
* Output : Optimal value function $v_*$ and optimal policy $\pi_*$

| Problem | Bellman Equation | Algorithm |
|---|---|---|
| Prediction | Bellman Equation Expectation | Iterative Policy Evaluation |
| Control | Bellman Expectation + greedy Policy Improvement | Policy Iteration |
| Control | Bellman Optimality Equation | Value Iteration |

Example : Grid World

| | | | |
|---|---|---|---|
| 1 G | 2 | 3 | 4 |
| 5 | 6 R | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 G |

Action ↔ ↕

Reward is -1 for all transition

A bot is required to traversed a grid of 4×4 dimensions to reach its goal (1 or 16) Each step is associated with a reward of -1. There are 2 terminal states here : 1 and 16 and 14 non-terminal states given by [2, 3 ... 16] Consider a random policy for which at every state, the probability of every action {up, down, left, right} is equal to 0.25.

Initializing $V_0$ for the Random Policy to all 0s

| | | | |
|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 |

Lets calculate $V_2$ for all the states of 6:

$$V_1(6) = \sum_{a \in \{u,d,l,r\}} \pi(a|6) \sum_{s',r'} p(s',r|6,a)\left[r + \gamma V_0(s')\right]$$

$$= \sum_{a \in \{u,d,l,r\}} \underbrace{\pi(a|6)}_{0.25} \sum_{s'} p(s'|6,a)\left[\underbrace{r}_{=-1} + \gamma \underbrace{V_0(s')}_{=0 \,\forall s'}\right]$$

$$= 0.25 \times \{-p(2|6,u) - p(10|6,d) - p(5|6,l) - p(7|6,r)\}$$

$$= 0.25 * \{-1-1-1-1\}$$

$$= -1$$

$$\Rightarrow V_1(6) = -1$$

Similarly for all non-terminal states

$$\boxed{V_1(s) = -1}$$

For terminal states $p(s'|s,a) = 0$

Hence $V_k(1) = V_k(16) = 0$ for all $k$.

So $V_1$ for the Random policy is given by

| | | | |
|---|---|---|---|
| 0.0 | -1.0 | -1.0 | -1.0 |
| -1.0 | -1.0 | -1.0 | -1.0 |
| -1.0 | -1.0 | -1.0 | -1.0 |
| -1.0 | -1.0 | -1.0 | 0.0 |

For $V_2(s)$, Assume $\gamma = 1$

$$V_2(6) = \sum_{a \in \{u,d,l,r\}} \pi(a|6) \sum_{s'} p(s'|6,a)\left[r + \gamma V_1(s')\right]$$

$\pi(a|6) = 0.25 \ \forall a$ ; $r = -1$ ; $V_1(s') = -1$ for all non-terminal $s'$

$$= 0.25 * \{ p(2|6,u)[-1-\gamma] + p(10|6,d)[-1-\gamma] + p(5|6,l)[-1-\gamma] + p(7|6,r)[-1-\gamma] \}$$

$= 0.25 * \{-2-2-2-2\}$

$= -2$

| | | | |
|---|---|---|---|
| 1 (G) | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 (G) |

All the states marked in Red in above are identical ~~for~~ to 6 for the purpose of calculating the value function.
Hence for all these states

$$\boxed{V_2(s) = -2}$$

For all the remaining states i.e 2, 5, 12 and 15 $V_2$ can be calculate as follows:

$$V_2(2) = \sum_{a \in \{u,d,l,r\}} \underbrace{\pi(a|2)}_{=0.25} \sum_{s'} p(s'|2,a) \left[ \underbrace{r}_{=-1} + \gamma \underbrace{V_1(s')}_{=0} \right]$$

$$= 0.25 * \{ p(2|2,u)[-1-\gamma] + p(6|2,d)[-1-\gamma] + p(1|2,l)[-1-\gamma*0] - p(3|2,r)[-1-\gamma] \}$$

$$= 0.25 * \{-2-2-1-2\}$$

$$= -1.75$$

$$\Rightarrow \boxed{V_2(2) = -1.75}$$

$V_2$ for the random policy

| | | | |
|---|---|---|---|
| 0.0 | -1.7 | -2.0 | -2.0 |
| -1.7 | -2.0 | -2.0 | -2.0 |
| -2.0 | -2.0 | -2.0 | -1.7 |
| -2.0 | -2.0 | -1.7 | 0.0 |

If we repeat this step several times we get $V_\infty$

∴ NOTE : Check PPT for the Solution & diagrams.

1. Problem

Frozen Lake Solution

1. Using Value Iteration

given state Representation

| | | | |
|---|---|---|---|
| S | F | F | F |
| F | H | F | H |
| F | F | F | H |
| H | F | F | G |

S = State
F = Frozen
H = Hole
G = Gold
$\gamma = 0.9$

Bellman Equation for Value Iteration

Value function

$$V(s) = \max_a \sum_{s'} p(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]$$

given
G = 1.0 (terminal State)
H = 0 (game Over)
other state = 0

Initial Value table

| | | | |
|---|---|---|---|
| S | 0 | 0 | 0 |
| 0 | H | 0 | H |
| 0 | 0 | 0 | H |
| H | 0 | 0 | G |

Formula
Step 1: for Iteration Value Update

$$V(s) = \max_a \sum_{s'} p(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]$$

Initial $V(s) = 0$ $V(G) = 1$

Iteration 1:

$G = 0$

State adjacent to the goal update

1. (3,2) Best move Right to (3,3)

$$V_{(3,2)} = 0.9 \times 1.0 = 0.9$$

2. (2,3) Best move Down to (3,3)

$$V_{(2,3)} = 0.9 \times 1.0 = 0.9$$

3. (3,1) Best move Right to (3,2)

$$V_{(3,1)} = 0.9 \times 0.9 = 0.98$$

4. (2,2) Best move Right (2,3)

$$V_{(2,2)} = 0.9 \times 0.9 = 0.98$$

5. (2,1) Best move Right (2,2)

$$V_{(2,1)} = 0.9 \times 0.98 = 0.97$$

6. (1,2) Best move Down to (2,2)

$$V_{(1,2)} = 0.9 \times 0.98 = 0.97$$

Iteration 2:

Repeating the same steps, value propagate backward

1. (1,1) (avoiding hole):

$$V_{(1,0)} = 0.9 \times 0.97 = 0.96$$

2. (0,2)

$$V_{(0,2)} = 0.9 \times 0.97 = 0.96$$

3. (0,1)

$$V_{(0,1)} = 0.9 \times 0.96 = 0.95$$

4. (0,0) (Start state)

$$V_{(0,0)} = 0.9 \times 0.95 = 0.94$$

| | | | | |
|---|---|---|---|---|
| S | 0.94 | 0.95 | 0.96 | ~~0.97~~ |
| 0.96 | H | 0.97 | H | 0. |
| 0.97 | 0.98 | 0.99 | H | |
| H | 0.98 | 0.99 | | |

Extracting the Optimal Policy

(0,0) → Right (Towards higher value)

0,1 → Right

0,2 → Right

0,3 → Down

1,3 → Down

2,3 → Right

3,2 → Right

3,3 → goal

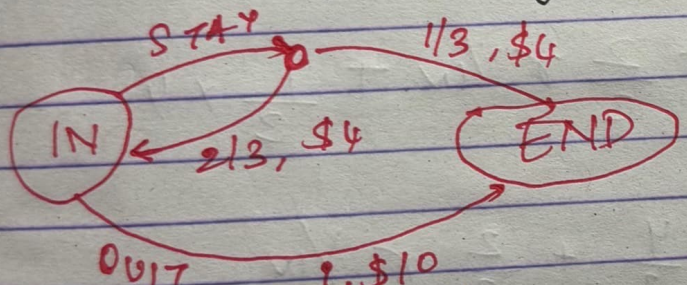
final O/P

| | | | |
|---|---|---|---|
| → | → | → | ↓ |
| ↓ | H | → | H |
| ↓ | → | → | H |
| H | ↓ | → | G |

2. Problem

At any time, the game is specified by two states: IN and END

When the game in the state IN, the agent can choose the action from the set $A_i = \{STAY, QUIT\}$
If QUIT is selected, then the game will go to state END and the agent recive a reward $10 with a probability of 1. If STAY is selected, the game will stay at IN state & the agent receive a reward $4 with a probability of 2/3 or transition to state END. and the agent recive a reward $4 with a probability of 1/3




STATE Transition Probabilities.

| Action | Next State | Prob | Reward |
|---|---|---|---|
| QUIT | END | 1.0 | 10 |
| STAY | IN | 2/3 | 4 |
| STAY | END | 1/3 | 4 |

$$V(s) = \max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]$$

$\gamma$ : b/w 0 & 1 ~~= 0.5~~

= 0.9

$V_{(END)} = 0$

State = IN given that the agent

For QUIT:

$$V_{quit}(IN) = 10 + 0.\gamma = 10$$

For STAY:

$$V_{STAY}(IN) = \left(\frac{2}{3} \times (4 + \gamma V_{(IN)})\right) + \left(\frac{1}{3} \times (4 + \gamma V_{(END)}\right)$$

$V_{END} = 0$

$$= \frac{2}{3} \times 4 + \frac{2}{3}\gamma V_{IN} + \frac{1}{3} \times 4$$

$$= \frac{8}{3} + \frac{4}{3} + \frac{2}{3}\gamma V_{IN}$$

$$V_{STAY}(IN) = 4 + \frac{2}{3}\gamma V_{IN}$$

∴ the agent selects STAY with probability 0.5 and QUIT with 0.5, we take weighted sum

$$V(IN) = 0.5 \times \left(4 + \frac{2}{3}\gamma V_{IN}\right) + 0.5 \times 10$$

$$V_{IN} = 0.5 \times 0.4 + 0.5 \times \frac{2}{3}\gamma V_{IN} + 5$$

$$= 2 + \frac{1}{3}\gamma V_{IN} + 5$$

$$= 7 + \frac{1}{3}\gamma V_{IN} + 5$$

$$V_{IN} - \frac{1}{3}\gamma V_{IN} = 7$$

$$\left(1 - \frac{1}{3}\gamma\right) V_{IN} = 7$$

$$V_{IN} = \frac{7}{1 - \frac{1}{3}\gamma}$$

$\boxed{\gamma = 0.9}$

$$V_{IN} = \frac{7}{1 - \frac{1}{3} \times 0.9} = \boxed{10}$$

---

$V_{(IN)} = 0$

Iteration 1 = $V_{IN} = 0.5 \times (4 + 0.6 \times 0) + 0.5 \times 10$

$\Rightarrow V_{IN} = 9$

Iteration 2 = $V_{IN} = 0.5 \times (4 + 0.6 \times 7) + 0.5 \times 10$

$= 9.1$

Iteration 3 = $V_{IN} = 0.5 \times (4 + 0.6 \times 9.1 + 0.5 \times 10$

$= 9.73$

4 $V_{IN} = 0.5 \times (4 + 0.6 \times 9.73) + 0.5 \times 10$

$= 9.919$

5 $V_{IN} = 9.9757$

6 $V_{IN} = 9.9927$

7 $V_{IN} = 9.9978$

8 $V_{IN} = 9.999$