Dynamic Programming

→ policy evaluation

→ policy improvement — Value iteration

$\pi(a|s) \to$ policy

R1 π_1 R2 π_2 R3 π_3

3, 3, 5   3, 5, 7   5, 10, 15

Avg(3,3,5)		

Avg(3,5,7)		

Avg(5,10,15)		

$v_{\pi_1}(s) =$

4	5	3
0	2	2
1	1	0

$v_{\pi_2}(s) =$

5	10	12
14	13	3
2	5	0

$v_{\pi_3}(s) =$

10	12	13
15	14	4
3	6	0

9

$v_{\pi_3}(s) > v_{\pi_2}(s) > v_{\pi_1}(s)$

Policy Evaluation

$$\underbrace{v_\pi(s)}_{?} = \sum_{a \in A} \pi(a|s) \left[\sum_{s' \in S} p(s'|s,a)\left(r(s,a,s') + \gamma \underbrace{v_\pi(s')}_{?\ 0}\right) \right]$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{Bellmans eqn.}}$$

init,

$v_\pi(s) = 0$

Iterate:

$$v_{New}(s) = \sum_{a \in A} \pi(a|s) \left[\sum_{s' \in S} p(s'|s,a)\left(r(s,a,s') + \gamma v_{old}(s')\right) \right]$$

-1

$v_{OLD}(s) = v_{New}(s)$

$\{(0,0), (0,1) (1,0) \ldots\}$

until $\|v_{OLD} - v_{New}\| \le tol$

0.0001

0	0	0
0	0	0
0	0	0

diff1 = 5

-1	-1	-1
-1	-1	-1
-1	-1	0

v_{OLD}

diff2 = 3

v_{New}

Policy improvement

value iteration

$\pi(a|s) \rightarrow$ Not known

we need to come up with optimal policy

$\pi^*(a|s)$

$$v_*(s) = \max_{a \in A} \Big(\underbrace{\sum_{s' \in S} p(s'|s,a) \big(r(s,a,s') + \gamma v_*(s') \big)}_{q(s,a)} \Big)$$

Bellmans optimality equation

$*$ $\quad V_{OLD}(s) = 0$

$$V_{NEW}(s) = \max_{a \in A} \Big(\sum_{s' \in S} p(s'|s,a) \big(r(s,a,s') + \gamma V_{OLD}(s') \big) \Big)$$

repeat until $\| V_{OLD}(s) - V_{NEW}(s) \| \leq \Delta$

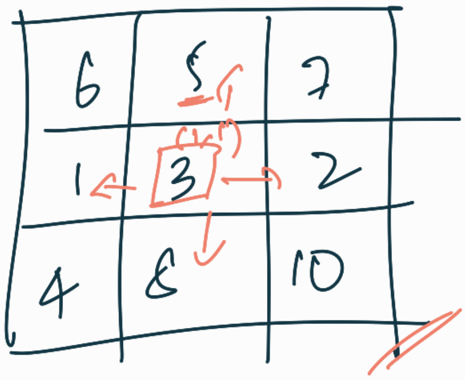
Ex, $\Delta = 0.001$

Finally after convergence

$V_{NEW}(s) \approx v^*(s)$

Optimal state Value function

$v^*(s) \xrightarrow{\text{gives}} \pi^*(a|s)$ optimal policy



one-step reward : -1

$5 - 1 = 4$