

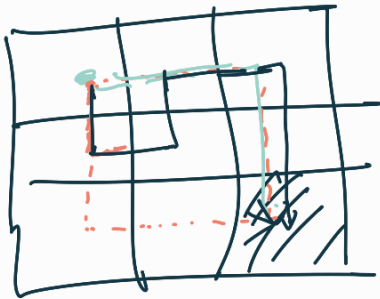
# Dynamic Programming

→ policy evaluation

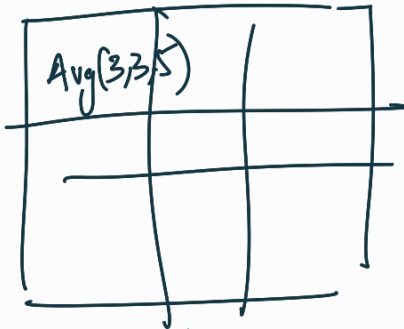
→ policy improvement - Value iteration

$\pi(a/s)$  → policy

$R_1 \quad \pi_1$   
=



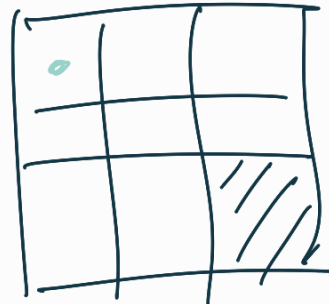
3, 3, 5  
↓



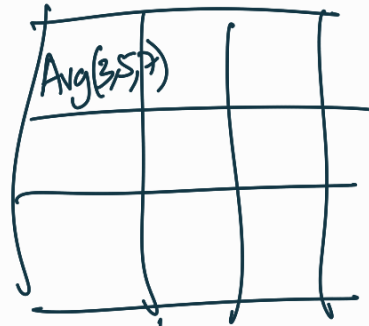
$V_{\pi_1}(s) =$

4	5	3
0	2	2
1	1	0

$R_2 \quad \pi_2$



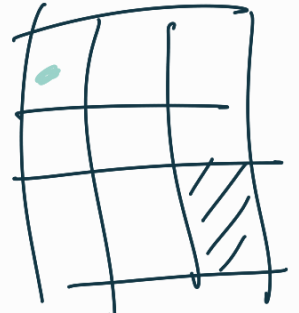
3, 5, 7  
↓



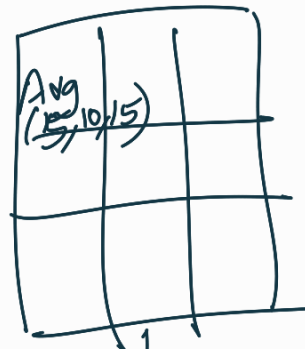
$V_{\pi_2}(s) =$

5	10	12
14	13	3
2	5	0

$R_3 \quad \pi_3$



5, 10, 15  
↓



$V_{\pi_3}(s) =$

10	12	13
15	14	4
3	6	0

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

# Policy Evaluation

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

Bellmans eqn.

*(Note: In the original image, there are red annotations: a red arrow points from the first  $V_{\pi}(s)$  to a question mark below it; another red arrow points from the  $V_{\pi}(s')$  term to a question mark below it, which is also labeled with a 0.)*

init,  $V_{\pi}(s) = 0 \leftarrow$

Iterate:

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

$V_{\text{old}}(s) = V_{\text{New}}(s)$

$\{(0,0), (0,1), (1,0), \dots\}$  until  $\|V_{\text{old}} - V_{\text{New}}\| \leq \text{tol}$

0	0	0
0	0	0
0	0	0

0.0001

diff 1 = 5

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

diff 2 = 3


$V_{\text{new}}$

1  
2

# Policy improvement

## value iteration

$\pi(a|s) \rightarrow$  Not known  
we need to come up  
with optimal policy  
 $\pi^*(a|s)$

$$\underline{V_*}(s) = \max_{a \in A} \left( \sum_{s' \in S} \underbrace{p(s'|s, a)}_{q(s, a)} \left( \underbrace{r(s, a, s') + \gamma V_*(s')}_{\text{Bellmans optimality equation}} \right) \right)$$

\*  $V_{old}(s) = 0$

$$\underline{V_{new}}(s) = \max_{a \in A} \left( \sum_{s' \in S} p(s'|s, a) (r(s, a, s') + \gamma V_{old}(s')) \right)$$

repeat until  $\|V_{old}^{(s)} - V_{new}^{(s)}\| \leq \Delta$

Ex,  $\Delta = 0.001$

Finally after convergence  
 $\underline{V_{new}}(s) \approx \underline{V^*(s)}$   
Optimal state  
Value function

$V^*(s)$   $\xrightarrow{\text{gives}}$

$\pi^*(a|s)$  optimal policy

6	5	7
1	3	2
4	8	10

one-step reward: -1

$5 - 1 = 4$   
 $0 \leftarrow \begin{matrix} \uparrow \\ \rightarrow 1 \\ \downarrow \end{matrix}$
