



# AIL 7022: Reinforcement Learning

## Lecture 9: Towards Policy Iteration & Value Iteration

Instructor: Raunak Bhattacharyya



**ScAI**

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE  
INDIAN INSTITUTE OF TECHNOLOGY DELHI

# Recap

- IPE
- Model-free PE
- IS and Incremental
- Return to DP

# Plan

- Value function recursive definition & policy improvement
- Using the optimal value function
- Greedyfication & policy iteration

# Value Function Recursion

$$\begin{aligned}v_{\pi}(s) &\doteq \mathbb{E}_{\pi}[G_t \mid S_t = s] \\&= \mathbb{E}_{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a) \left[ r + \gamma \mathbb{E}_{\pi}[G_{t+1} | S_{t+1} = s'] \right] \\&= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[ r + \gamma v_{\pi}(s') \right], \quad \text{for all } s \in \mathcal{S}\end{aligned}$$

$$\begin{aligned}v_{\pi}(s) &\doteq \mathbb{E}_{\pi}[G_t \mid S_t = s] \\&= \mathbb{E}_{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\&= \mathbb{E}_{\pi}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s] \\&= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \left[ r + \gamma v_{\pi}(s') \right]\end{aligned}$$

# Policy Improvement

Let  $\pi$  and  $\pi'$  be any pair of deterministic policies such that, for all  $s \in \mathcal{S}$ ,

$$q_{\pi}(s, \pi'(s)) \geq v_{\pi}(s)$$

Then the policy  $\pi'$  must be as good as, or better than,  $\pi$ .

$$v_{\pi'}(s) \geq v_{\pi}(s)$$

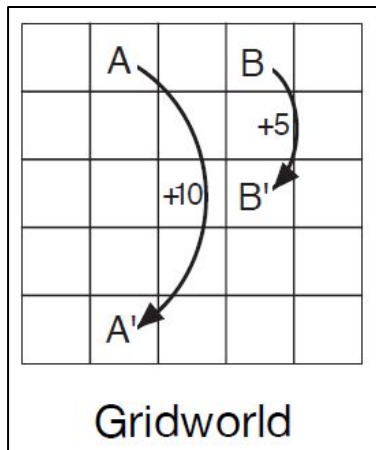
# Policy Improvement

$$\begin{aligned} v_{\pi}(s) &\leq q_{\pi}(s, \pi'(s)) \\ &= \mathbb{E}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s, A_t = \pi'(s)] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_{\pi}(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}[R_{t+2} + \gamma v_{\pi}(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})] \mid S_t = s] \\ &= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_{\pi}(S_{t+2}) \mid S_t = s] \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_{\pi}(S_{t+3}) \mid S_t = s] \\ &\vdots \\ &\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s] \\ &= v_{\pi'}(s). \end{aligned}$$

# Optimal Value Functions

$$v_*(s) \doteq \max_{\pi} v_{\pi}(s), \text{ for all } s \in \mathcal{S}.$$

$$q_*(s, a) \doteq \max_{\pi} q_{\pi}(s, a), \text{ for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}(s)$$



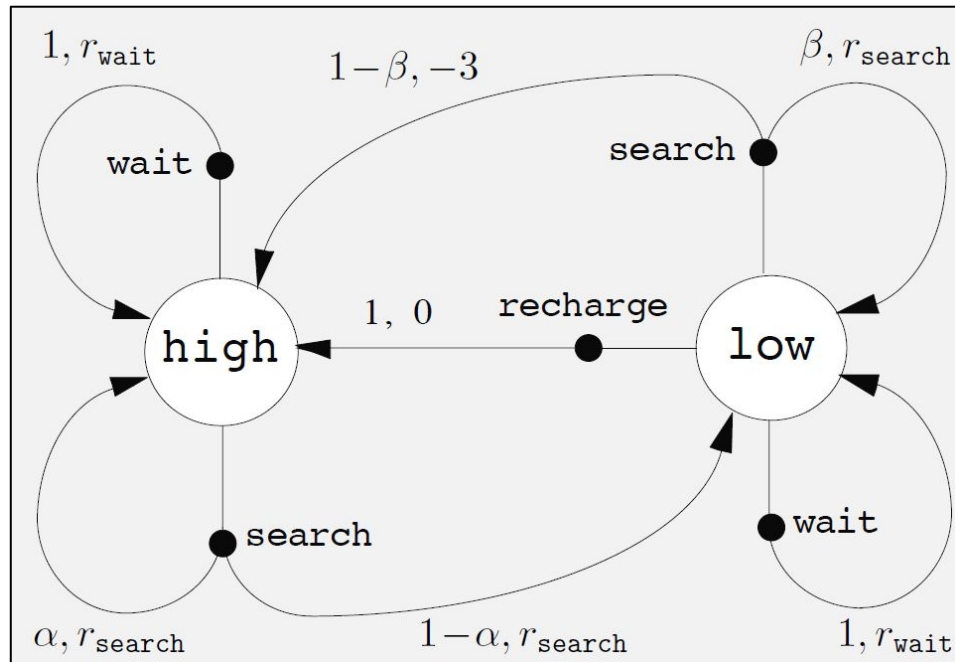
|      |      |      |      |      |
|------|------|------|------|------|
| 3.3  | 8.8  | 4.4  | 5.3  | 1.5  |
| 1.5  | 3.0  | 2.3  | 1.9  | 0.5  |
| 0.1  | 0.7  | 0.7  | 0.4  | -0.4 |
| -1.0 | -0.4 | -0.4 | -0.6 | -1.2 |
| -1.9 | -1.3 | -1.2 | -1.4 | -2.0 |

|      |      |      |      |      |
|------|------|------|------|------|
| 22.0 | 24.4 | 22.0 | 19.4 | 17.5 |
| 19.8 | 22.0 | 19.8 | 17.8 | 16.0 |
| 17.8 | 19.8 | 17.8 | 16.0 | 14.4 |
| 16.0 | 17.8 | 16.0 | 14.4 | 13.0 |
| 14.4 | 16.0 | 14.4 | 13.0 | 11.7 |

$v_*$

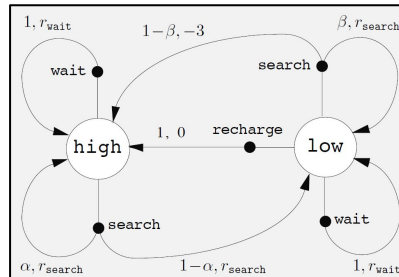
Which action to take?

# Example: Can Collecting Robot



| $s$  | $a$      | $s'$ | $p(s'   s, a)$ | $r(s, a, s')$       |
|------|----------|------|----------------|---------------------|
| high | search   | high | $\alpha$       | $r_{\text{search}}$ |
| high | search   | low  | $1 - \alpha$   | $r_{\text{search}}$ |
| low  | search   | high | $1 - \beta$    | $-3$                |
| low  | search   | low  | $\beta$        | $r_{\text{search}}$ |
| high | wait     | high | 1              | $r_{\text{wait}}$   |
| high | wait     | low  | 0              | -                   |
| low  | wait     | high | 0              | -                   |
| low  | wait     | low  | 1              | $r_{\text{wait}}$   |
| low  | recharge | high | 1              | 0                   |
| low  | recharge | low  | 0              | -                   |

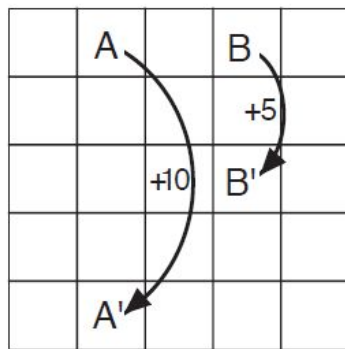
# Example: Can Collecting Robot



$$\begin{aligned}
 v_*(\mathbf{h}) &= \max \left\{ \begin{array}{l} p(\mathbf{h} | \mathbf{h}, \mathbf{s})[r(\mathbf{h}, \mathbf{s}, \mathbf{h}) + \gamma v_*(\mathbf{h})] + p(\mathbf{l} | \mathbf{h}, \mathbf{s})[r(\mathbf{h}, \mathbf{s}, \mathbf{l}) + \gamma v_*(\mathbf{l})], \\ p(\mathbf{h} | \mathbf{h}, \mathbf{w})[r(\mathbf{h}, \mathbf{w}, \mathbf{h}) + \gamma v_*(\mathbf{h})] + p(\mathbf{l} | \mathbf{h}, \mathbf{w})[r(\mathbf{h}, \mathbf{w}, \mathbf{l}) + \gamma v_*(\mathbf{l})] \end{array} \right\} \\
 &= \max \left\{ \begin{array}{l} \alpha[r_{\mathbf{s}} + \gamma v_*(\mathbf{h})] + (1 - \alpha)[r_{\mathbf{s}} + \gamma v_*(\mathbf{l})], \\ 1[r_{\mathbf{w}} + \gamma v_*(\mathbf{h})] + 0[r_{\mathbf{w}} + \gamma v_*(\mathbf{l})] \end{array} \right\} \\
 &= \max \left\{ \begin{array}{l} r_{\mathbf{s}} + \gamma[\alpha v_*(\mathbf{h}) + (1 - \alpha)v_*(\mathbf{l})], \\ r_{\mathbf{w}} + \gamma v_*(\mathbf{h}) \end{array} \right\}.
 \end{aligned}$$

$$v_*(\mathbf{l}) = \max \left\{ \begin{array}{l} \beta r_{\mathbf{s}} - 3(1 - \beta) + \gamma[(1 - \beta)v_*(\mathbf{h}) + \beta v_*(\mathbf{l})], \\ r_{\mathbf{w}} + \gamma v_*(\mathbf{l}), \\ \gamma v_*(\mathbf{h}) \end{array} \right\}$$

# From Optimal Value Functions to Policies

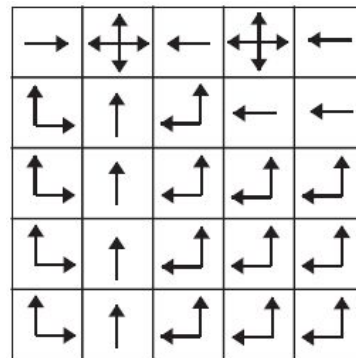


Gridworld

|      |      |      |      |      |
|------|------|------|------|------|
| 3.3  | 8.8  | 4.4  | 5.3  | 1.5  |
| 1.5  | 3.0  | 2.3  | 1.9  | 0.5  |
| 0.1  | 0.7  | 0.7  | 0.4  | -0.4 |
| -1.0 | -0.4 | -0.4 | -0.6 | -1.2 |
| -1.9 | -1.3 | -1.2 | -1.4 | -2.0 |

|      |      |      |      |      |
|------|------|------|------|------|
| 22.0 | 24.4 | 22.0 | 19.4 | 17.5 |
| 19.8 | 22.0 | 19.8 | 17.8 | 16.0 |
| 17.8 | 19.8 | 17.8 | 16.0 | 14.4 |
| 16.0 | 17.8 | 16.0 | 14.4 | 13.0 |
| 14.4 | 16.0 | 14.4 | 13.0 | 11.7 |

$v_*$



$\pi_*$

# Greedy Policy: Policy Improvement

$$\begin{aligned}\pi'(s) &\doteq \arg \max_a q_\pi(s, a) \\ &= \arg \max_a \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \arg \max_a \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_\pi(s')],\end{aligned}$$

$$\begin{aligned}v_{\pi'}(s) &= \max_a \mathbb{E}[R_{t+1} + \gamma v_{\pi'}(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_{\pi'}(s')].\end{aligned}$$

# Greedy Policy



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

$R_t = -1$   
on all transitions

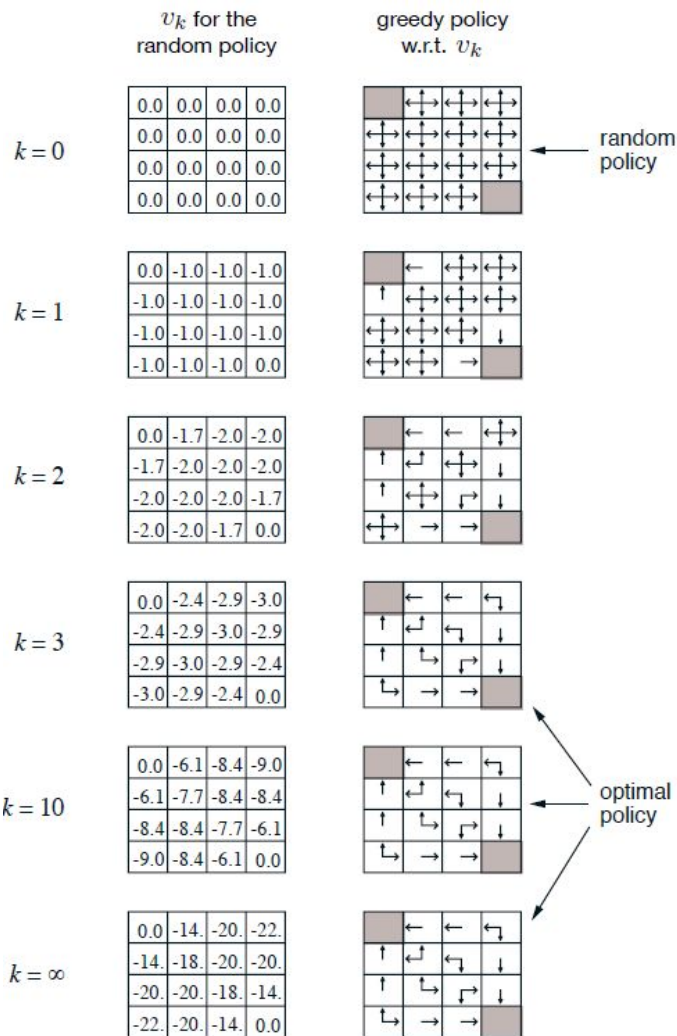


Fig 4.1, [Sutton & Barto](#)

# Policy Iteration

$$\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \dots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_*$$

## Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

### 1. Initialization

$V(s) \in \mathbb{R}$  and  $\pi(s) \in \mathcal{A}(s)$  arbitrarily for all  $s \in \mathcal{S}$ ;  $V(\text{terminal}) \doteq 0$

### 2. Policy Evaluation

Loop:

$\Delta \leftarrow 0$

Loop for each  $s \in \mathcal{S}$ :

$v \leftarrow V(s)$

$V(s) \leftarrow \sum_{s',r} p(s', r | s, \pi(s)) [r + \gamma V(s')]$

$\Delta \leftarrow \max(\Delta, |v - V(s)|)$

until  $\Delta < \theta$  (a small positive number determining the accuracy of estimation)

### 3. Policy Improvement

$\text{policy-stable} \leftarrow \text{true}$

For each  $s \in \mathcal{S}$ :

$\text{old-action} \leftarrow \pi(s)$

$\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$

If  $\text{old-action} \neq \pi(s)$ , then  $\text{policy-stable} \leftarrow \text{false}$

If  $\text{policy-stable}$ , then stop and return  $V \approx v_*$  and  $\pi \approx \pi_*$ ; else go to 2

# Demo: Policy Iteration

## GridWorld: Dynamic Programming Demo

Policy Evaluation (one sweep)

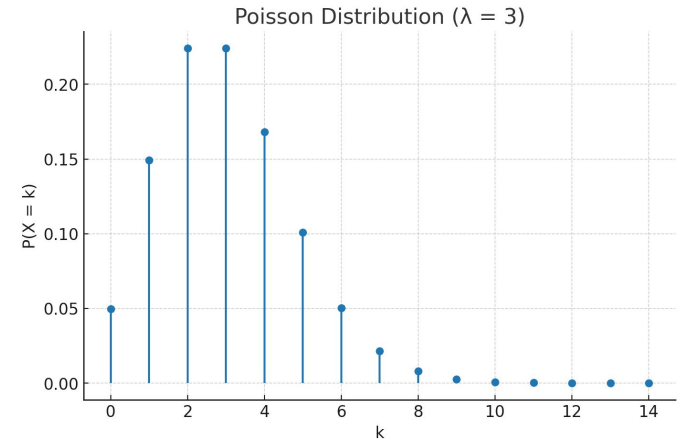
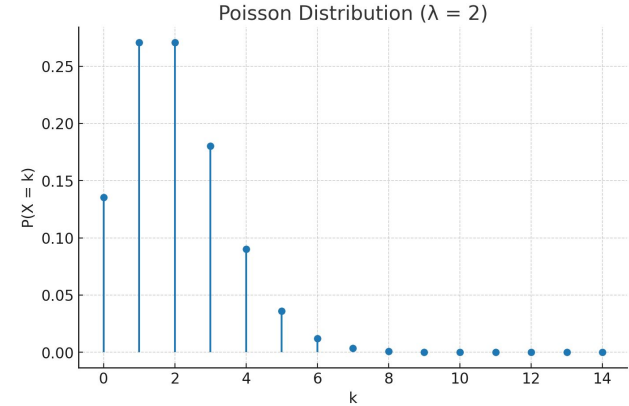
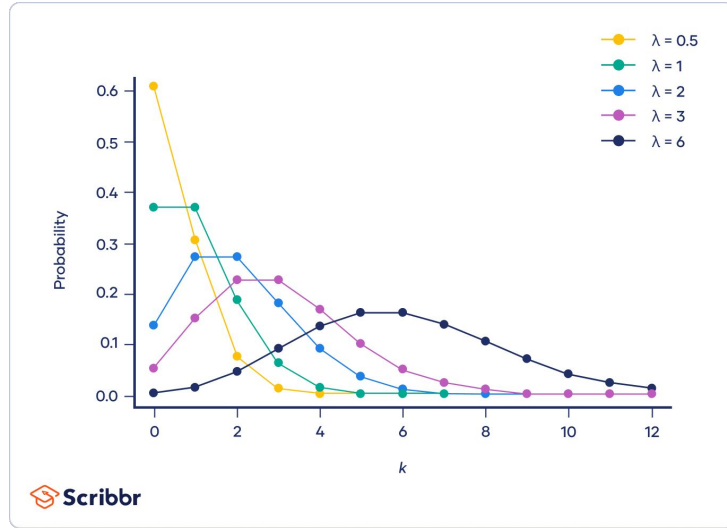
Policy Update

Toggle Value Iteration

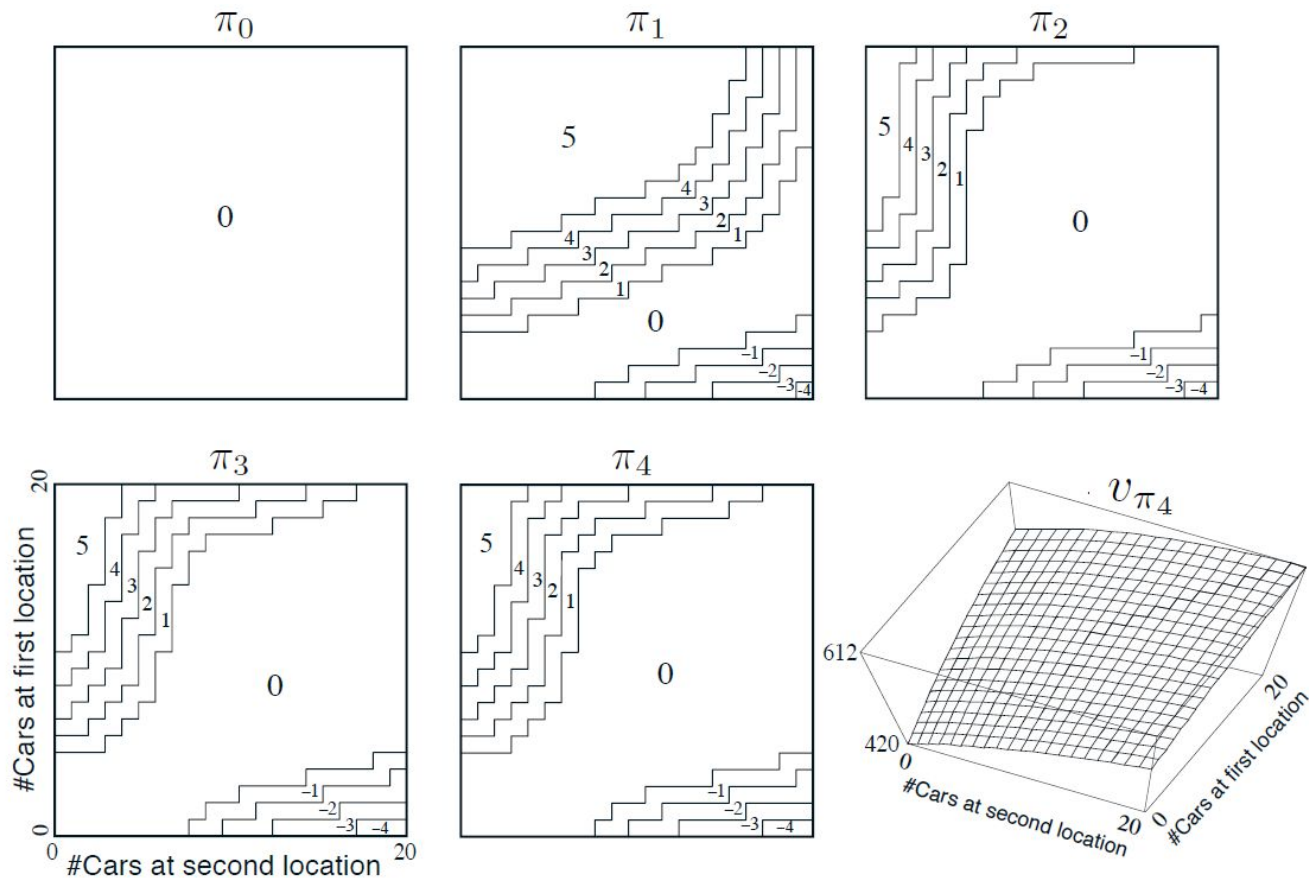
Reset

|           |           |           |                     |           |                     |                     |           |                     |           |
|-----------|-----------|-----------|---------------------|-----------|---------------------|---------------------|-----------|---------------------|-----------|
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |
| 0.00<br>↖ |           |           |                     |           | 0.00<br>↖           |                     |           |                     | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖<br>R -1.0 |           | 0.00<br>↖           | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           |           | 0.00<br>↖<br>R -1.0 | 0.00<br>↖<br>R -1.0 | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           |           | 0.00<br>↖<br>R 1.0  | 0.00<br>↖<br>R -1.0 | 0.00<br>↖ | 0.00<br>↖<br>R -1.0 | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           |           | 0.00<br>↖           | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖<br>R -1.0 | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖<br>R -1.0 |           | 0.00<br>↖<br>R -1.0 | 0.00<br>↖<br>R -1.0 | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |
| 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖           | 0.00<br>↖ | 0.00<br>↖           | 0.00<br>↖ |

# Example: Jack's Car Rental

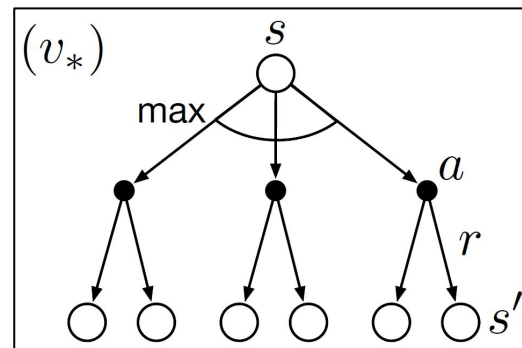


# Example: Jack's Car Rental



# Bellman Optimality Equations

$$\begin{aligned}v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s, a) \\&= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\&= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\&= \max_a \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_*(s')].\end{aligned}$$



$$\begin{aligned}v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s, A_t = a] \\&= \max_a \sum_{s', r} p(s', r \mid s, a) [r + \gamma v_k(s')],\end{aligned}$$

# Value Iteration

## Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold  $\theta > 0$  determining accuracy of estimation

Initialize  $V(s)$ , for all  $s \in \mathcal{S}^+$ , arbitrarily except that  $V(\text{terminal}) = 0$

Loop:

```
|  $\Delta \leftarrow 0$   
|   Loop for each  $s \in \mathcal{S}$ :  
|      $v \leftarrow V(s)$   
|      $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$   
|      $\Delta \leftarrow \max(\Delta, |v - V(s)|)$   
until  $\Delta < \theta$ 
```

Output a deterministic policy,  $\pi \approx \pi_*$ , such that

$$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$$