Last Time

(S, A, T, R, γ)

Conditional Probability Dist.

Last Time

- How is a **Markov decision process** defined?

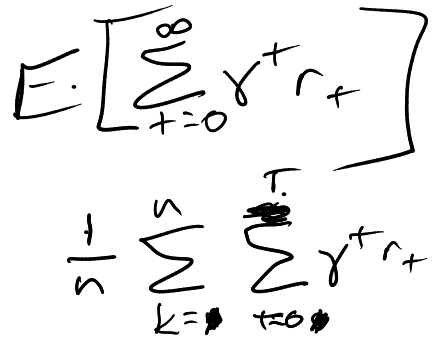
Last Time

- How is a **Markov decision process** defined?
- What is a **policy**?

$\pi : S \rightsquigarrow A$

Last Time

- How is a **Markov decision process** defined?
- What is a **policy**?
- How do we **evaluate** policies?



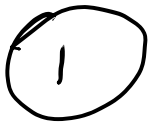
Guiding Questions

Guiding Questions

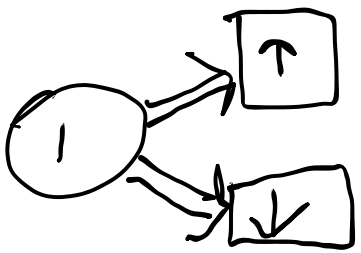
- How do we reason about the **future consequences** of actions in an MDP?
- What are the basic **algorithms for solving MDPs**?

MDP Example: Up-Down Problem

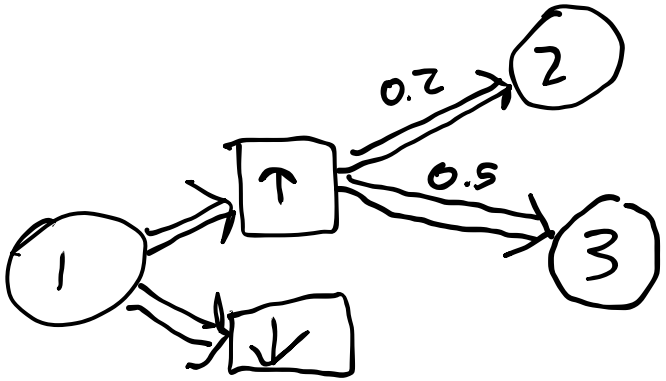
MDP Example: Up-Down Problem



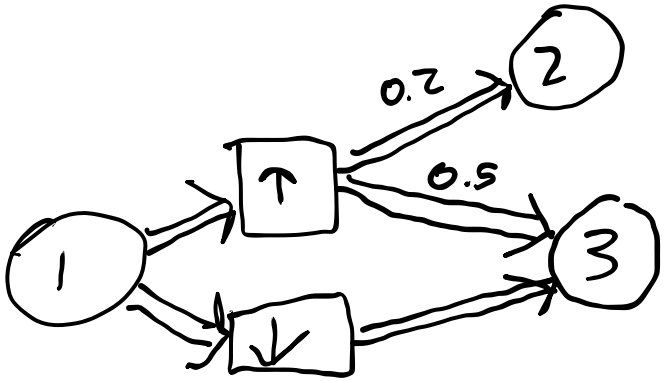
MDP Example: Up-Down Problem



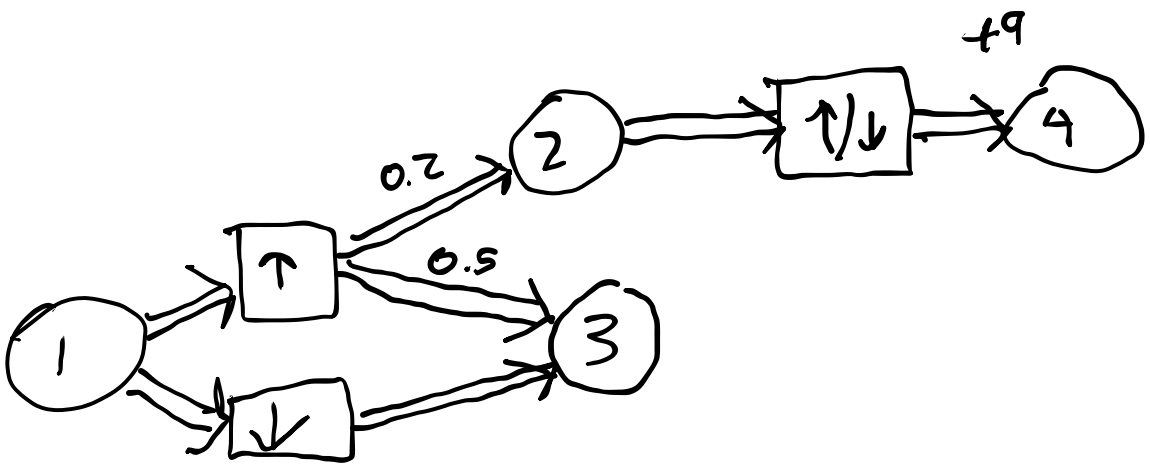
MDP Example: Up-Down Problem



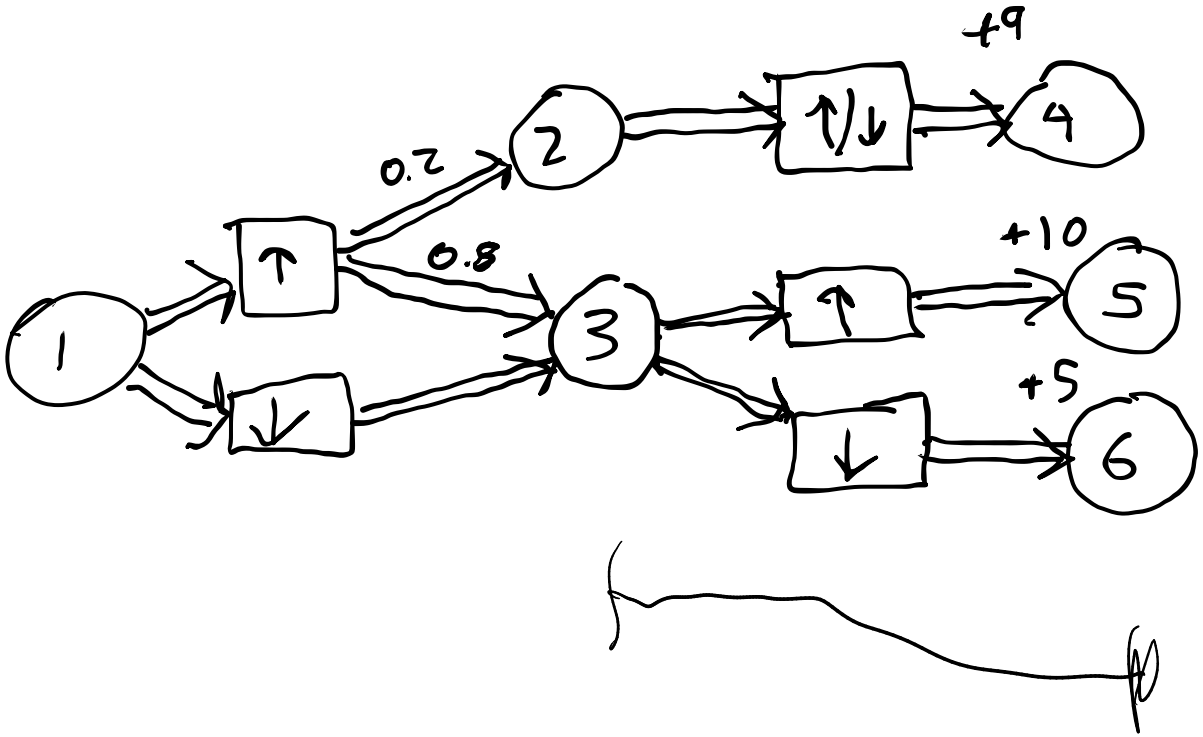
MDP Example: Up-Down Problem



MDP Example: Up-Down Problem



MDP Example: Up-Down Problem



s	$\pi(s)$
1	↓
3	↑

Value Functions

$$E[r_0 \mid s_0 = s] = R(s, \pi(s))$$

maximize $V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right]$

$= \gamma^0 R(s, \pi(s)) + E\left[\sum_{t=1}^{\infty} \gamma^t r_t \mid s_0 = s\right]$

$= R(s, \pi(s)) + E\left[\sum_{t=1}^{\infty} \gamma^t r_t \mid s_1 \sim T(s, \pi(s))\right]$

$+ \gamma E\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 \sim T(s, \pi(s))\right]$

$+ \gamma \underset{s' \sim T(s,a)}{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s'\right]$

myopic

$\pi(s) = \underset{a}{\text{argmax}}\, R(s,a)$

$\pi(s) = \text{argmax}\, Q^\pi(s,a)$

$$V^\pi(s) = \underbrace{R(s, \pi(s))}_{\text{Immediate reward}} + \gamma \underbrace{\underset{s' \sim T(s,a)}{E}\left[V^\pi(s')\right]}_{\text{Rewards in future}}$$

$$Q^\pi(s,a) \equiv R(s, \pi(s)) + \gamma E\left[V^\pi(s')\right]$$

$V = U$ in book

Exact

policy

Matrix Evaluation

$V^\pi = [0 \dots 0]$

$$V^\pi(s) = R(s, \pi(s)) + \gamma \underset{s' \sim T(s,a)}{E}\left[V^\pi(s')\right]$$

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s'|s,a) V^\pi(s')$$

iterative policy evaluation

$$V_i^\pi = R_i^\pi + \gamma \sum_{j \in \{1, \dots |S|\}} T_{ij}^\pi V_j^\pi$$

$$T_{ij} = T\left(s'=j \mid s=i, a=\pi(i)\right)$$

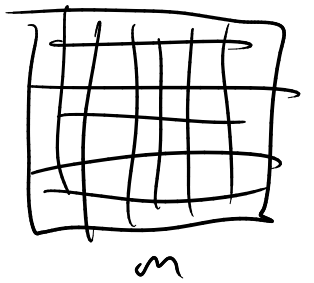
$$V^\pi = R^\pi + \gamma T^\pi V^\pi$$

$$V^\pi - \gamma T^\pi V^\pi = R^\pi$$

$$(I - \gamma T^\pi) V^\pi = R^\pi$$

$$\boxed{V^\pi = (I - \gamma T^\pi)^{-1} R^\pi}$$

$O(n^3)$



$|S| \times |S|$

$O(S^6)$ $\quad |S| = m^d$

Policy Iteration

Policy Iteration

Algorithm: Policy Iteration

Given: MDP (S, A, R, T, γ, b)

initialize π, π' (differently)

while $\pi \neq \pi'$

$\pi \leftarrow \pi'$

policy evaluation

$V^\pi \leftarrow (I - \gamma T^\pi)^{-1} R^\pi$

policy update

$\pi'(s) \leftarrow \underset{a \in A}{\operatorname{argmax}} \left(R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V^\pi(s') \right)$

$Q^\pi(s,a)$

return π ← optimal

Bellman's Equation

$\pi = \text{argmax}\ Q^{\pi}(s,a)$

$\pi^* = \text{argmax}\ Q^{*}(s,a)$

V^*

$$V^*(s) = R(s, \pi^*(s)) + \gamma \mathop{E}_{s' \sim T(s,\pi^*(s))} [V^*(s')]$$

$$V^*(s) = \max_a \{ R(s,a) + \gamma \mathop{E}_{s' \sim T(s,a)} [V^*(s')] \}$$

Value Function Policies

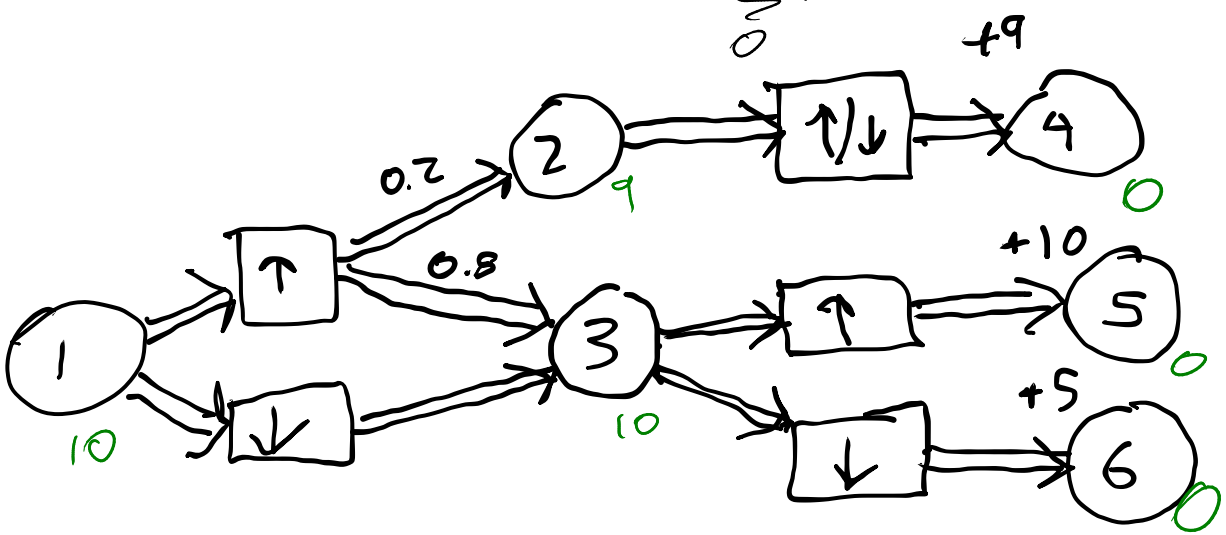
$$\pi^*(s) = \underset{a}{\operatorname{argmax}} \left\{ R(s,a) + \gamma \underset{s' \sim T(s,a)}{E} \left[V^*(s') \right] \right\}$$

$$V^* \Rightarrow \pi^*$$

$$\pi^* \Rightarrow V^* \qquad V^* = \left(I + \gamma T^{\pi^*} \right)^{-1} R^{\pi^*}$$

Backup by hand example

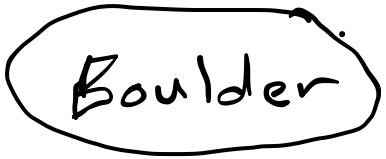
$$V^*(s) = \max_a \{ R(s,a) + \gamma \, E[V^*(s')] \}$$

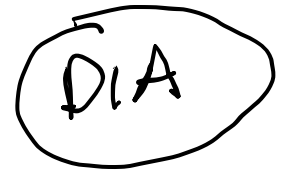


s	$V^*(s)$	a	R(s,a)	$E[V^*(s')]$
6	0			
5	0			
4	0			
3	10	↑	+10	0
		↓	+5	0
2	9	↑	+9	0
		↓	+9	0
1	10	↑	0	0.2·9 + 0.8·10 = 9.8
		↓	0	10

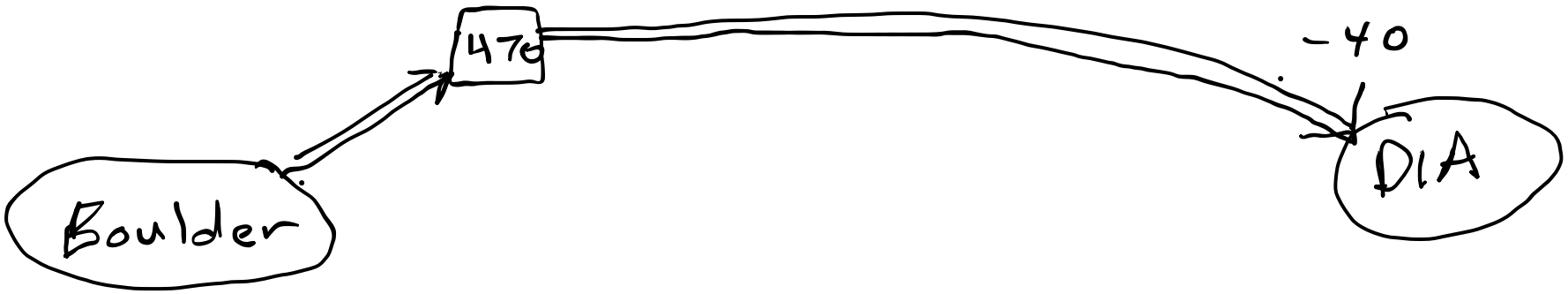
Breakout Rooms: DIA Run

Breakout Rooms: DIA Run

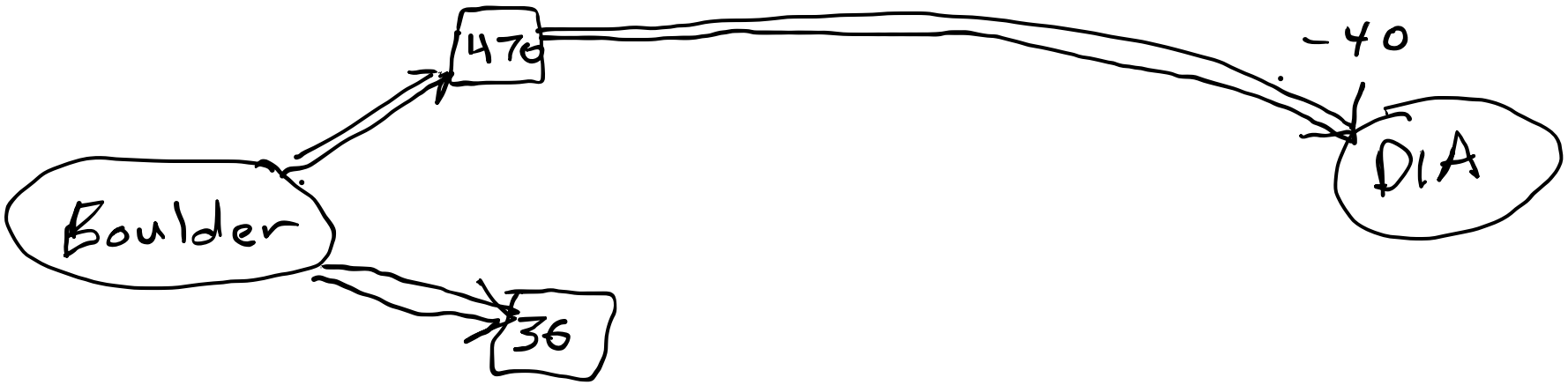




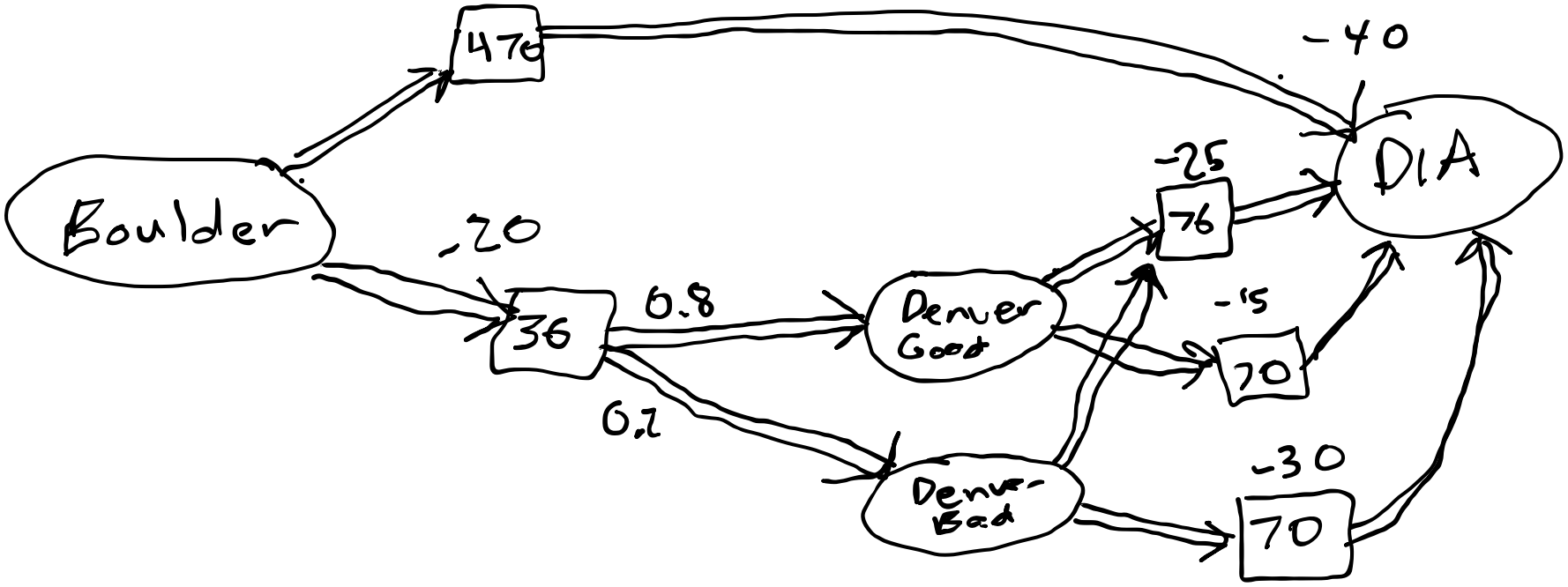
Breakout Rooms: DIA Run



Breakout Rooms: DIA Run



Breakout Rooms: DIA Run



Value Iteration

Value Iteration

HW2

Algorithm: Value Iteration

Given: MDP (S, A, R, T, γ, b), tolerance ϵ

initialize V, ~~V'~~ (differently)

while $\|V - V'\|_\infty < \epsilon$

$V \leftarrow V'$

$V'(s) \leftarrow \max_{a \in A} \left(R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V^\pi(s')\right) \quad \forall s \in S$

return $V' \approx V^*$

$\|V\|_\infty = \max_i V_i$

apply Bellman's

$\|V' - V^*\|_\infty < \varepsilon$

Guiding Questions

Guiding Questions

$V^*(s)$

$V(s)$

- How do we reason about the **future consequences** of actions in an MDP?
- What are the basic **algorithms for solving MDPs**?

Value Iter — apply Bellman's

Policy Iter < eval / update