Last Time

Last Time

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

Last Time

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

Last Time

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

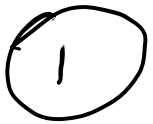
Guiding Questions

Guiding Questions

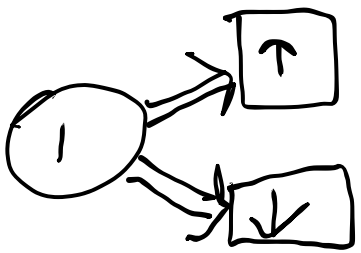
- How do we reason about the **future consequences** of actions in an MDP?
- What are the basic **algorithms for solving MDPs**?

MDP Example: Up-Down Problem

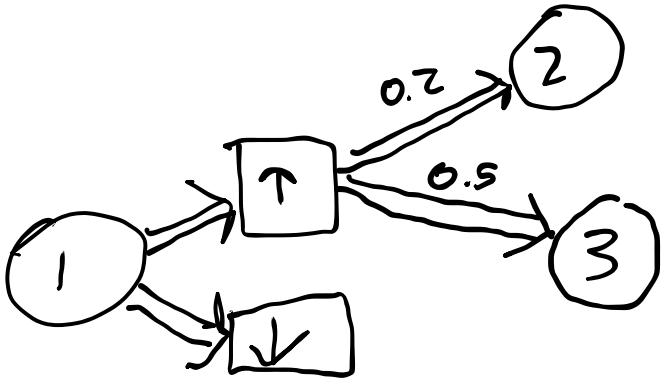
MDP Example: Up-Down Problem



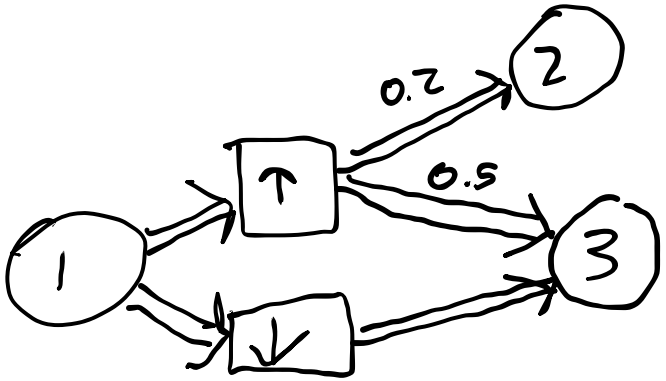
MDP Example: Up-Down Problem



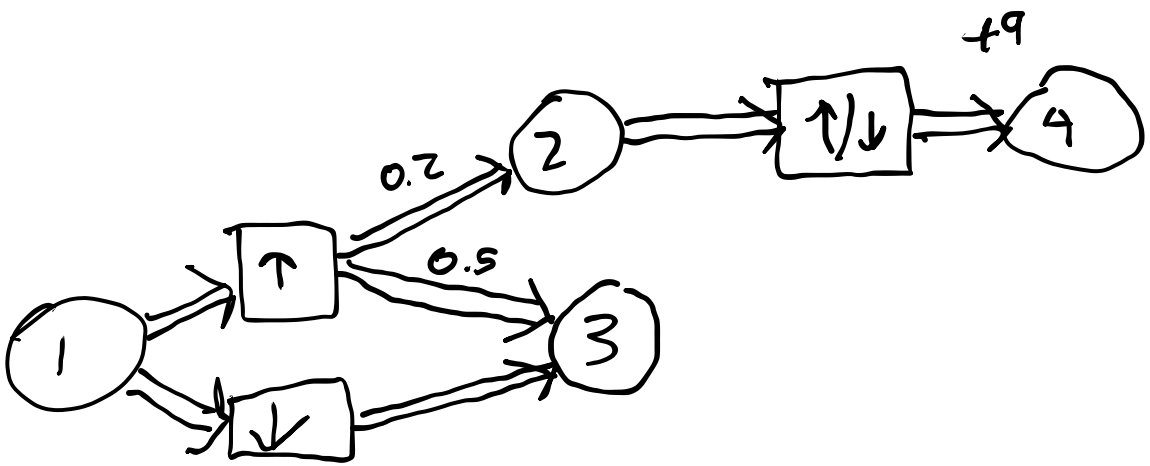
MDP Example: Up-Down Problem



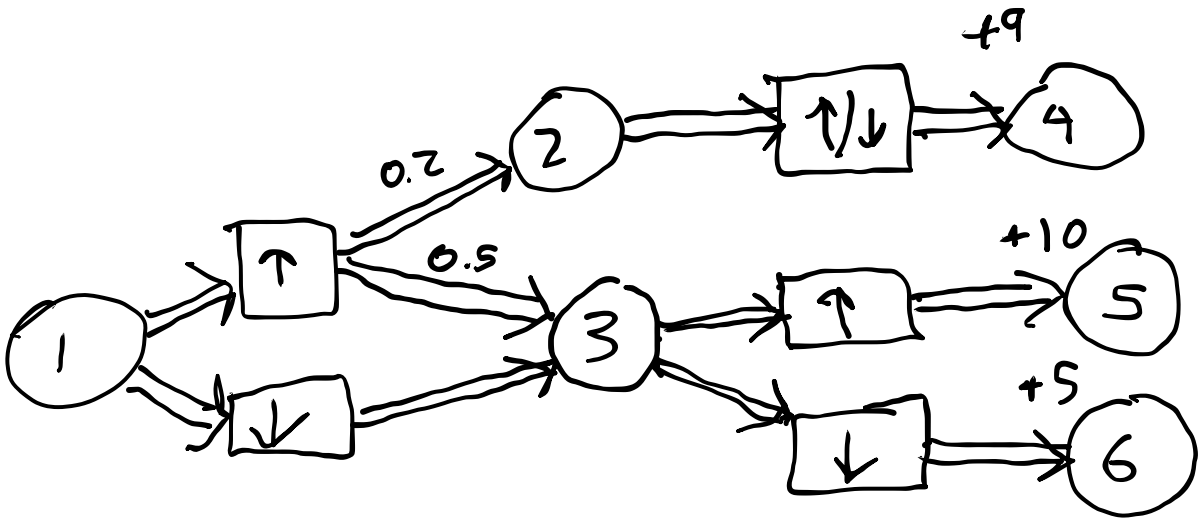
MDP Example: Up-Down Problem



MDP Example: Up-Down Problem



MDP Example: Up-Down Problem



Value Functions

Matrix Evaluation

Policy Iteration

Policy Iteration

Algorithm: Policy Iteration

Given: MDP (S, A, R, T, γ, b)

initialize π, π' (differently)

while $\pi \neq \pi'$

$\pi \leftarrow \pi'$

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

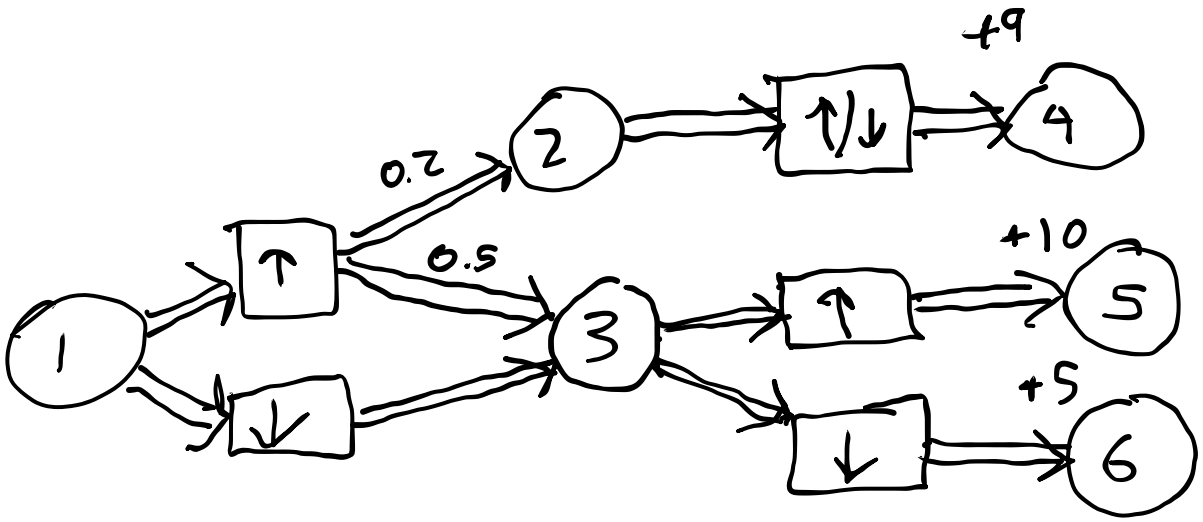
$\pi'(s) \leftarrow \underset{a \in A}{\operatorname{argmax}} \left(R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V^{\pi}(s') \right)$

return π

Bellman's Equation

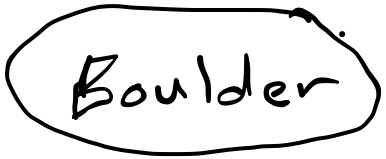
Value Function Policies

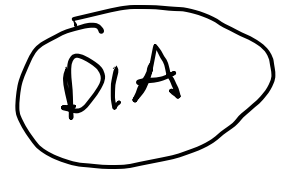
Backup by hand example



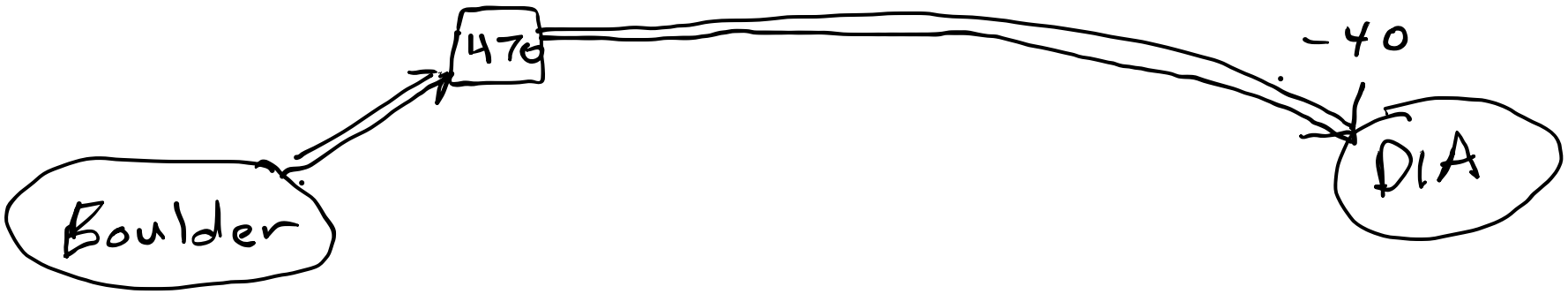
Breakout Rooms: DIA Run

Breakout Rooms: DIA Run

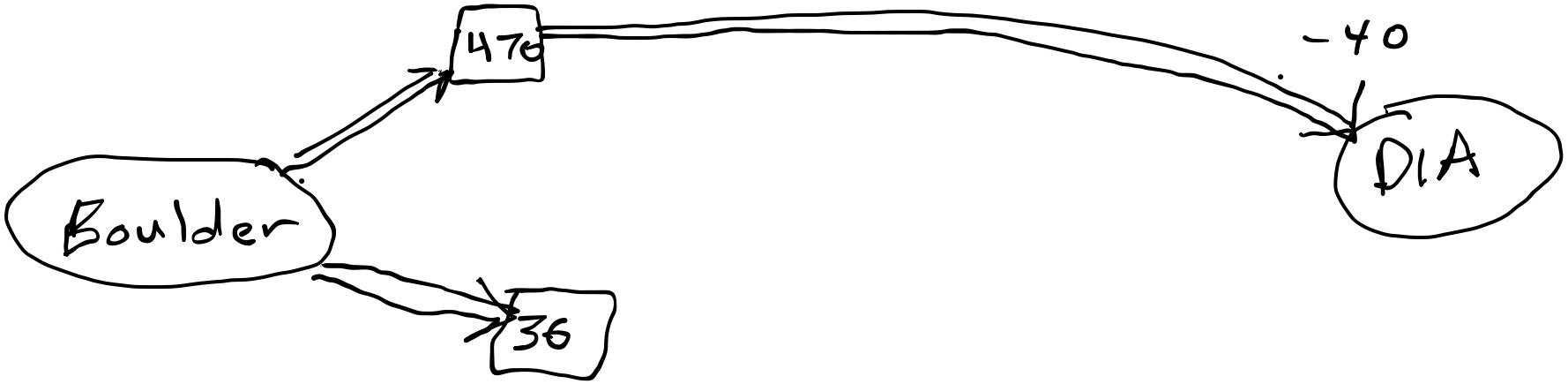




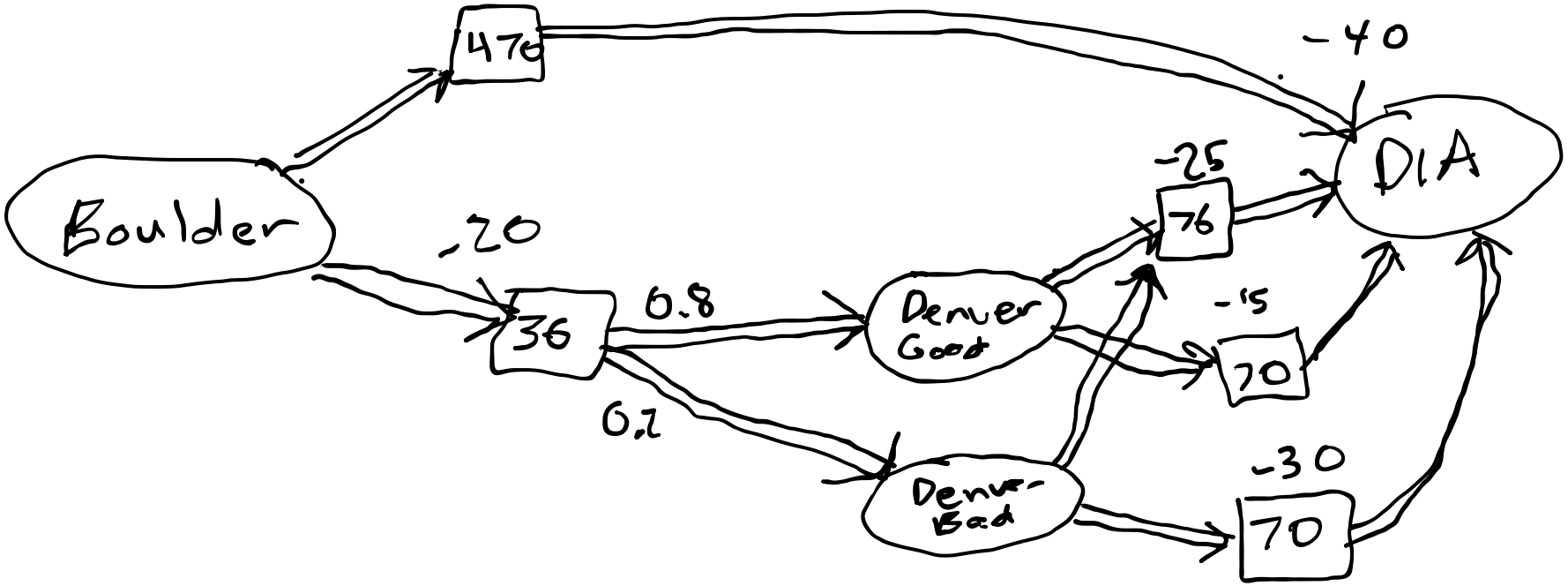
Breakout Rooms: DIA Run



Breakout Rooms: DIA Run



Breakout Rooms: DIA Run



Value Iteration

Value Iteration

Algorithm: Value Iteration

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

initialize V, \(\V'\) (differently)

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

$V \leftarrow V'$

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

return V'

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**?