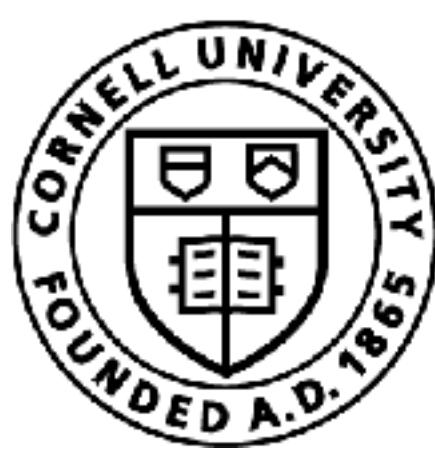


Markov Decision Process II

Sanjiban Choudhury



Cornell Bowers CIS
Computer Science

Learning

Robot
Decision
Making

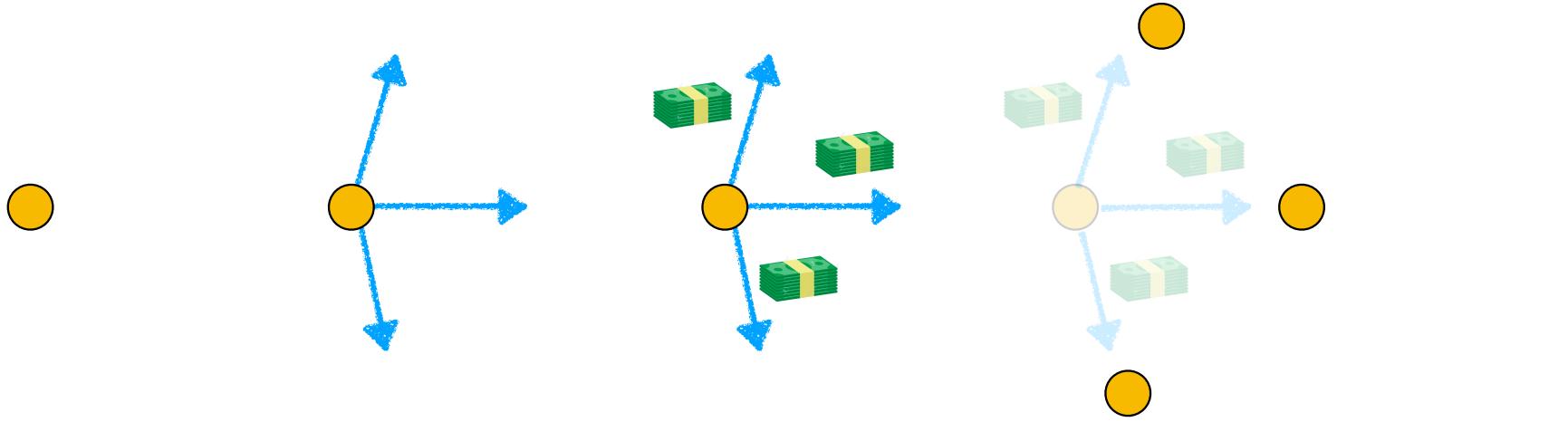
Today!

Recap

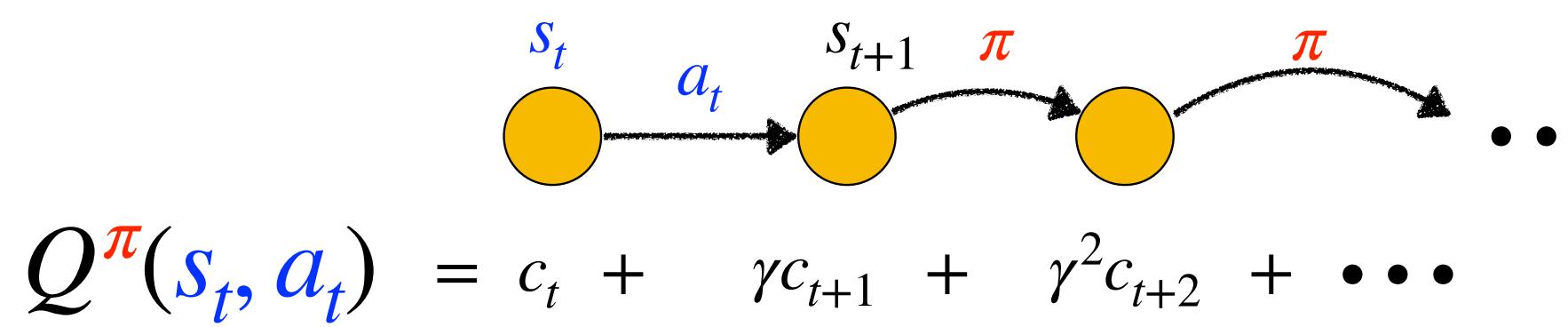
Markov Decision Process

A mathematical framework for modeling sequential decision making

$$\langle S, A, C, \mathcal{T} \rangle$$



Value of a state-action



Expected discounted sum of cost from starting at a state, executing action and following a policy from then on

$$Q^\pi(s_t, a_t) = c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{T}(s_t, a_t)} V^\pi(s_{t+1})$$

Dynamic Programming all the way!

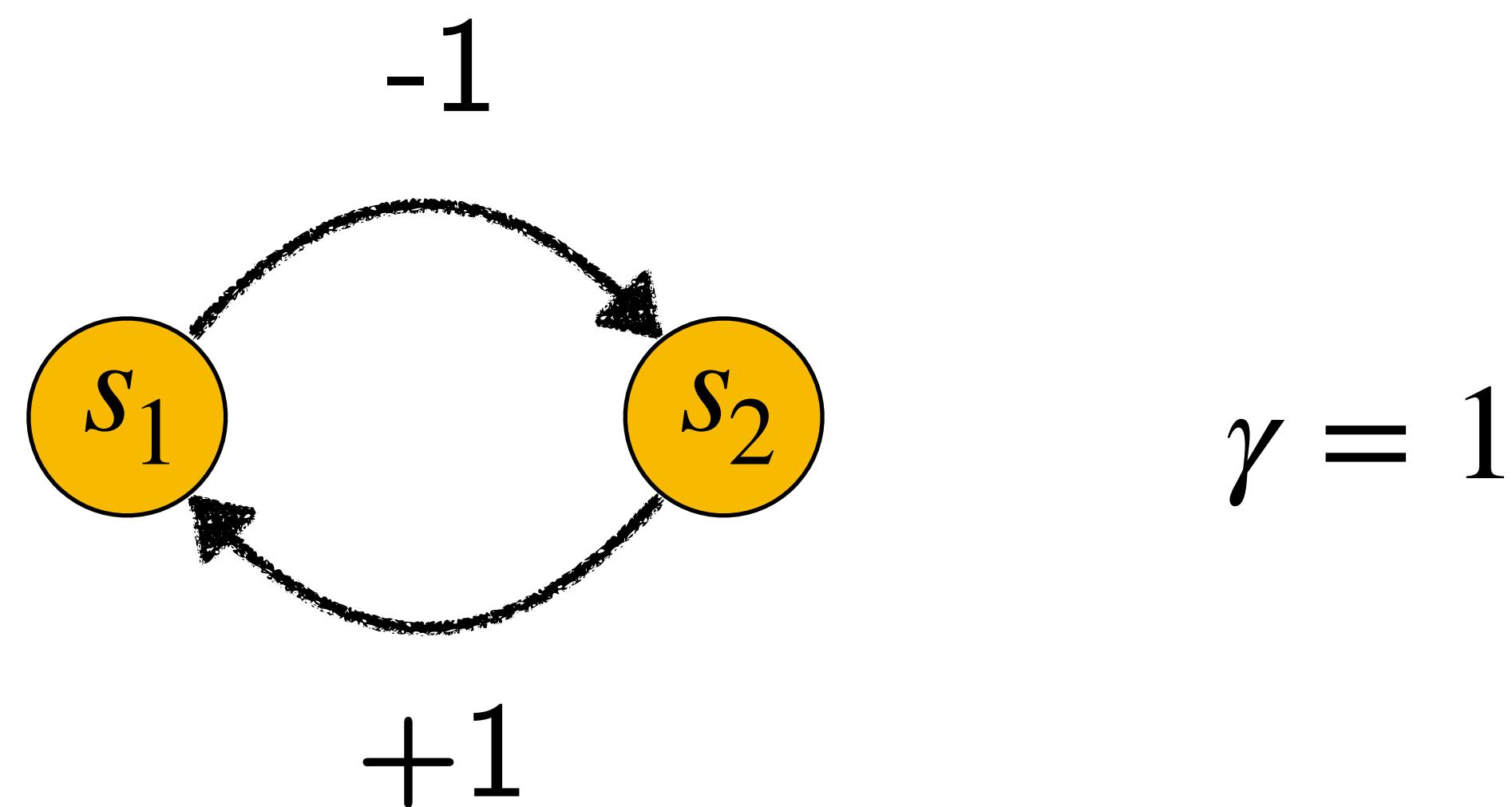
Time: 16										
0	1	2	3	4	5	6	7	8	9	
0	14	14	13	14	14	14	14	2	1	0
1	14	13	12	14	14	14	14	3	2	1
2	13	12	11	14	14	14	14	4	3	2
3	12	11	10	9	8	7	6	5	4	3
4	13	12	11	14	14	14	14	6	5	4
5	14	13	12	14	14	14	14	7	6	5
6	14	14	13	14	14	14	14	8	7	6
7	14	14	14	13	12	11	10	9	8	7
8	14	14	14	14	13	12	11	10	9	8
9	14	14	14	14	14	13	12	11	10	9

$$V^*(s_t) = \min_a [c(s_t, a) + V^*(s_{t+1})]$$

$$\pi^*(s_t) = \arg \min_a [c(s_t, a) + V^*(s_{t+1})]$$

0	1	2	3	4	5	6	7	8	9	
0	x	x	↓	x	x	x	x	→	→	x
1	x	→	↓	x	x	x	x	→	→	↑
2	→	→	↓	x	x	x	x	→	→	↑
3	→	→	→	→	→	→	→	→	→	↑
4	→	→	↑	x	x	x	x	→	→	↑
5	x	→	↑	x	x	x	x	→	→	↑
6	x	x	↑	x	x	x	x	→	→	↑
7	x	x	x	→	→	→	→	→	→	↑
8	x	x	x	x	→	→	→	→	→	↑
9	x	x	x	x	x	→	→	→	→	↑

Does value iteration converge?



What is $V^*(s_1)$? What is $V^*(s_2)$?

What is the effect of discount factor?

Gamma: 0.0

0	1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1

0	x	x	x	x	x	x	x	x	x	x
1	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x	x	x	x
3	x	x	x	x	x	x	x	x	x	x
4	x	x	x	x	x	x	x	x	x	x
5	x	x	x	x	x	x	x	x	x	x
6	x	x	x	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x
8	x	x	x	x	x	x	x	x	x	x
9	x	x	x	x	x	x	x	x	x	x

Activity!



Think-Pair-Share

Think (30 sec): What are some attributes of a **hard** MDP?

Pair: Find a partner

Share (45 sec): Partners exchange
ideas

Policy Iteration

How frequently does the best action change?

$\sigma = 0$	10	10	10	10	10	10	10	10	10	
$\sigma = 1$	10	10	10	10	10	10	10	10	10	
$\sigma = 2$	10	10	10	10	10	10	10	10	10	
$\sigma = 3$	10	10	10	10	10	10	10	10	10	
$\sigma = 4$	10	10	10	10	10	10	10	10	10	
$\sigma = 5$	10	10	10	10	10	10	10	10	10	
$\sigma = 6$	10	10	10	10	10	10	10	10	10	
$\sigma = 7$	10	10	10	10	10	10	10	10	10	
$\sigma = 8$	10	10	10	10	10	10	10	10	10	
$\sigma = 9$	10	10	10	10	10	10	10	10	10	
	0	1	2	3	4	5	6	7	8	9

Values

$\sigma = 0$	x	x	x	x	x	x	x	x	x	x
$\sigma = 1$	x	x	x	x	x	x	x	x	x	x
$\sigma = 2$	x	x	x	x	x	x	x	x	x	x
$\sigma = 3$	x	x	x	x	x	x	x	x	x	x
$\sigma = 4$	x	x	x	x	x	x	x	x	x	x
$\sigma = 5$	x	x	x	x	x	x	x	x	x	x
$\sigma = 6$	x	x	x	x	x	x	x	x	x	x
$\sigma = 7$	x	x	x	x	x	x	x	x	x	x
$\sigma = 8$	x	x	x	x	x	x	x	x	x	x
$\sigma = 9$	x	x	x	x	x	x	x	x	x	x
	0	1	2	3	4	5	6	7	8	9

Policy



Policy converges **faster**
than the value

Can we iterate over **policies**?

Policy Iteration

Init with some policy π

Repeat forever

Evaluate policy

$$V^\pi(s) = c(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s, a)} V^\pi(s')$$

Improve policy

$$\pi^+(s) = \arg \min_a c(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s, a)} V^\pi(s')$$

Init with some policy π

Iter: 0

0	-	→	→	→	→	→	→	→	→	→	↑
1	-	→	→	→	→	→	→	→	→	→	↑
2	-	→	→	→	→	→	→	→	→	→	↑
3	-	→	→	→	→	→	→	→	→	→	↑
4	-	→	→	→	→	→	→	→	→	→	↑
5	-	→	→	→	→	→	→	→	→	→	↑
6	-	→	→	→	→	→	→	→	→	→	↑
7	-	→	→	→	→	→	→	→	→	→	↑
8	-	→	→	→	→	→	→	→	→	→	↑
9	-	→	→	→	→	→	→	→	→	→	↑
	↓	0	1	2	3	4	5	6	7	8	9

Iteration 1

Iter: 1

0	74	75	76	77	77	77	77	2	1	0
1	74	75	76	77	77	77	77	3	2	1
2	74	75	76	77	77	77	77	3.9	3	2
3	55	56	56	57	50	40	26	4.9	3.9	3
4	74	75	76	77	77	77	77	5.9	4.9	3.9
5	74	75	76	77	77	77	77	6.8	5.9	4.9
6	74	75	76	77	77	77	77	7.7	6.8	5.9
7	15	14	13	12	11	10	9.6	8.6	7.7	6.8
8	16	15	14	13	12	11	10	9.6	8.6	7.7
9	17	16	15	14	13	12	11	10	9.6	8.6
	0	1	2	3	4	5	6	7	8	9

0	x	←	←	x	x	x	x	→	→	x
1	x	←	←	x	x	x	x	→	→	↑
2	↓	↓	↓	x	x	x	x	→	→	↑
m	x	←	←	→	→	→	→	→	→	↑
4	↑	↑	↑	x	x	x	x	→	→	↑
5	x	←	←	x	x	x	x	→	→	↑
6	↓	↓	↓	x	x	x	x	→	→	↑
7	→	→	→	→	→	→	→	→	→	↑
8	→	→	→	→	→	→	→	→	→	↑
9	→	→	→	→	→	→	→	→	→	↑
	0	1	2	3	4	5	6	7	8	9

$$V^\pi(s) = c(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,a)} V^\pi(s')$$

$$\pi^+(s) = \arg \min_a c(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,a)} V^\pi(s')$$

Policy Iteration

Iter: 0									
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
0	1	2	3	4	5	6	7	8	9
1	0	1	2	3	4	5	6	7	8
2	0	1	2	3	4	5	6	7	8
3	0	1	2	3	4	5	6	7	8
4	0	1	2	3	4	5	6	7	8
5	0	1	2	3	4	5	6	7	8
6	0	1	2	3	4	5	6	7	8
7	0	1	2	3	4	5	6	7	8
8	0	1	2	3	4	5	6	7	8
9	0	1	2	3	4	5	6	7	8

$$V^\pi(s) = c(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,a)} V^\pi(s')$$

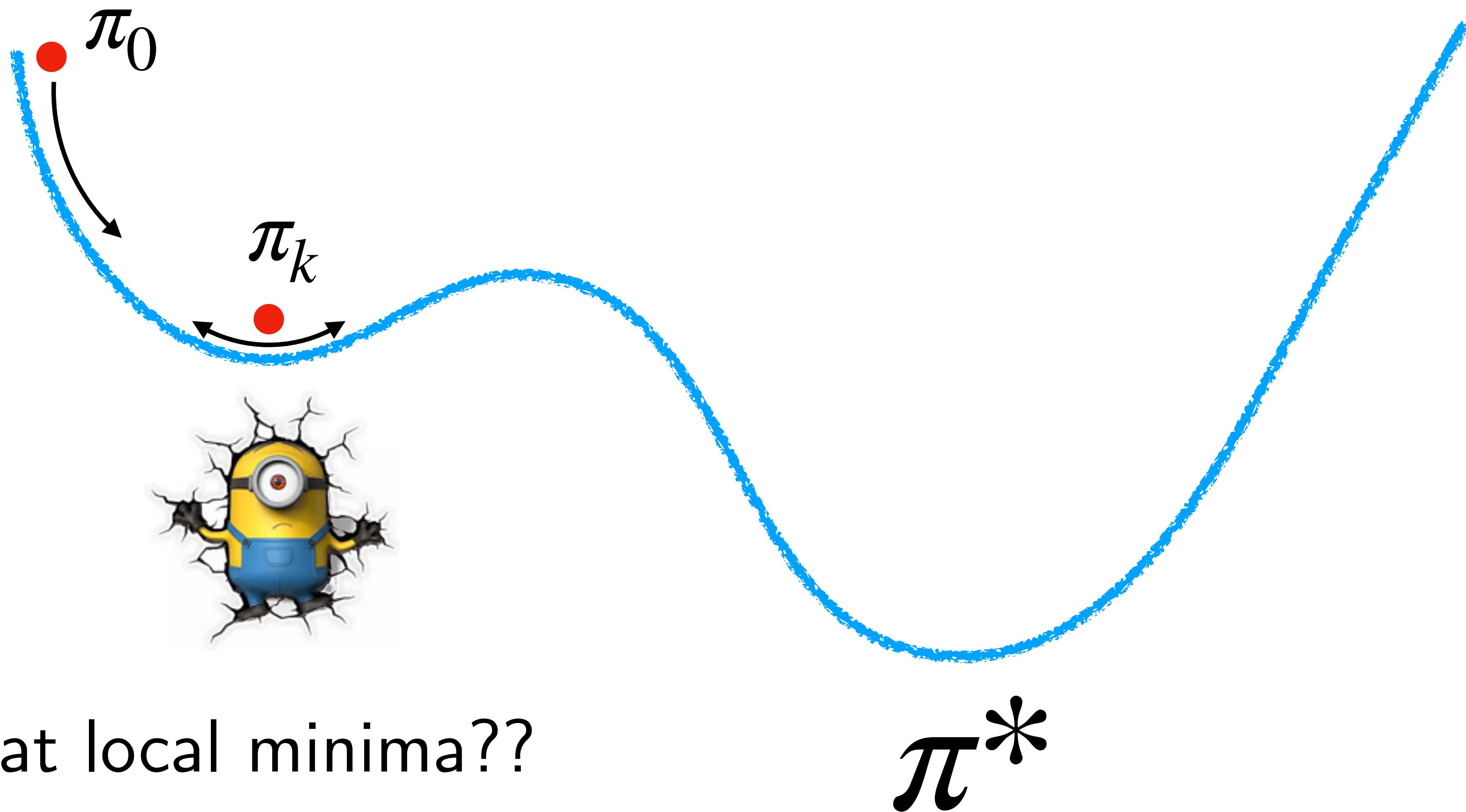
$$\pi^+(s) = \arg \min_a c(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(s,a)} V^\pi(s')$$



So many questions

- Q1. Does policy iteration converge to the optimal policy?
- Q2. Does it converge faster than value iteration?
- Q3. Can we bootstrap policy evaluation?

Q1. Does policy iteration converge to optimal policy?



Q1. Does policy iteration converge to optimal policy?

Proof has 2 step argument

Step 1: Policy iteration *monotonically* improves

Step 2: Once it reaches the optimal policy, it does not change

Performance Difference Lemma



Q1. Does policy iteration converge to optimal policy?

2 step argument

 Step 1: Policy iteration *monotonically* improves

All advantages ≤ 0 implies monotonic performance improvement

 Step 2: Once it reaches the optimal policy, it does not change

Advantage ≥ 0 for the optimal policy

PDL answers *all* ...

In Imitation Learning

In Model Free Reinforcement Learning

In Model Based Reinforcement Learning

How it partners with online learning

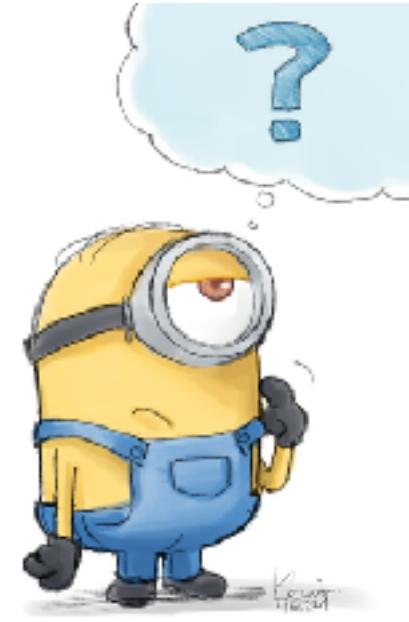


So many questions

✓ Q1. Does policy iteration converge to the optimal policy?

Q2. Does it converge faster than value iteration?

Q3. Can we bootstrap policy evaluation?



So many questions

✓ Q1. Does policy iteration converge to the optimal policy?

(`)/ Q2. Does it converge faster than value iteration?

Empirically yes, but no rigorous theory about when ..

Q3. Can we bootstrap policy evaluation?



So many questions

✓ Q1. Does policy iteration converge to the optimal policy?

(`)/ Q2. Does it converge faster than value iteration?

Empirically yes, but no rigorous theory about when ..

✓ Q3. Can we bootstrap policy evaluation?

Yes => Modified policy iteration

Messing
with
MDPs



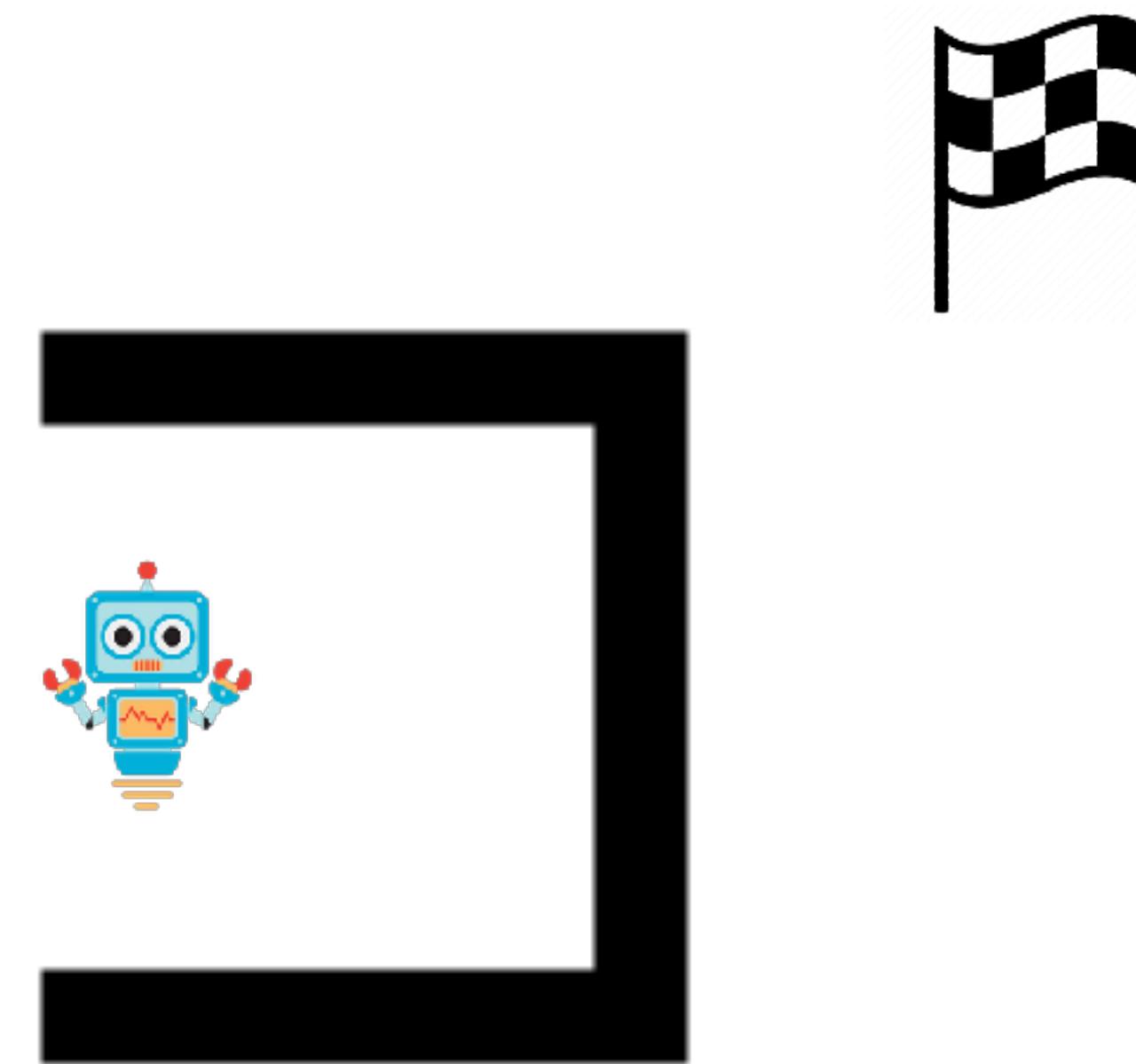
What happens as you increase the slipperiness of bridge?

p_slip: 0.0

	0	1	2	3	4	5	6	7	8	9
0	14	13	12	25	25	25	25	2	1	0
1	13	12	11	25	25	25	25	3	2	1
2	12	11	10	25	25	25	25	3.9	3	2
3	11	10	9.6	8.6	7.7	6.8	5.9	4.9	3.9	3
4	12	11	10	25	25	25	25	5.9	4.9	3.9
5	13	12	11	25	25	25	25	6.8	5.9	4.9
6	14	13	12	25	25	25	25	7.7	6.8	5.9
7	15	14	13	12	11	10	9.6	8.6	7.7	6.8
8	16	15	14	13	12	11	10	9.6	8.6	7.7
9	17	16	15	14	13	12	11	10	9.6	8.6

	0	1	2	3	4	5	6	7	8	9
0	→	→	↓	x	x	x	x	→	→	x
1	→	→	↓	x	x	x	x	→	→	↑
2	→	→	↓	x	x	x	x	→	→	↑
3	→	→	→	→	→	→	→	→	→	↑
4	→	→	↑	x	x	x	x	→	→	↑
5	→	→	↑	x	x	x	x	→	→	↑
6	→	→	↑	x	x	x	x	→	→	↑
7	→	→	→	→	→	→	→	→	→	↑
8	→	→	→	→	→	→	→	→	→	↑
9	→	→	→	→	→	→	→	→	→	↑

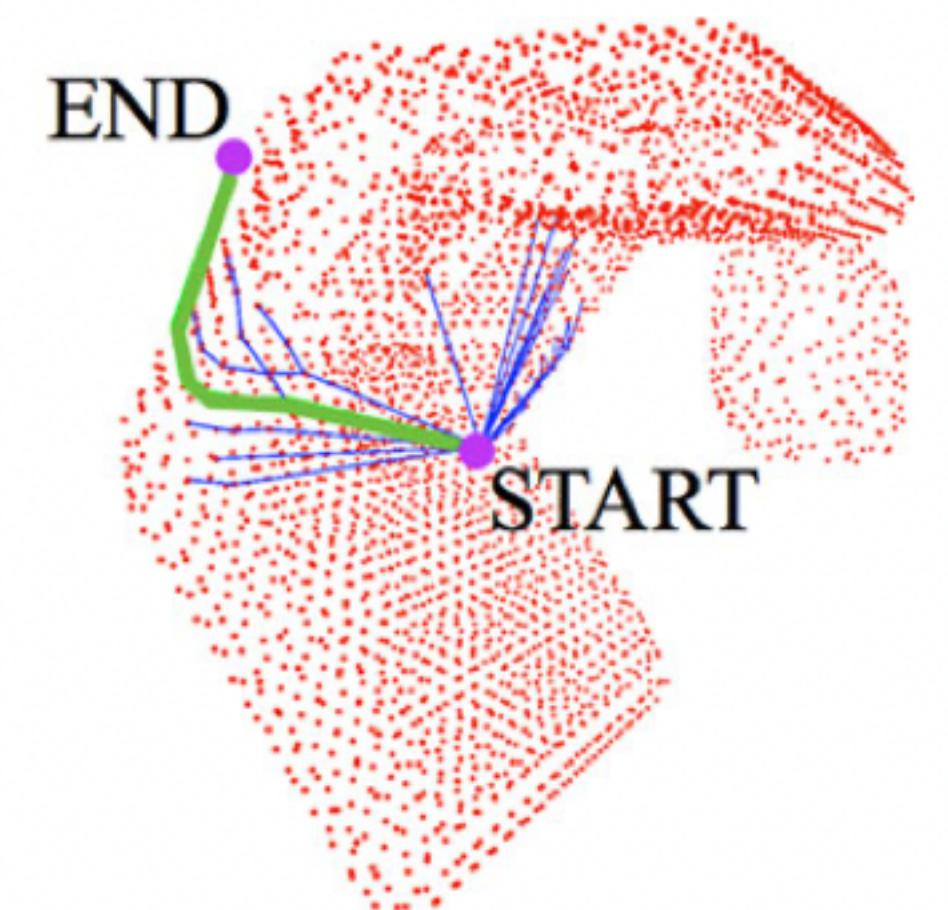
What happens if you change the cost function?



$$C(s) = ||s - s_{goal}||$$

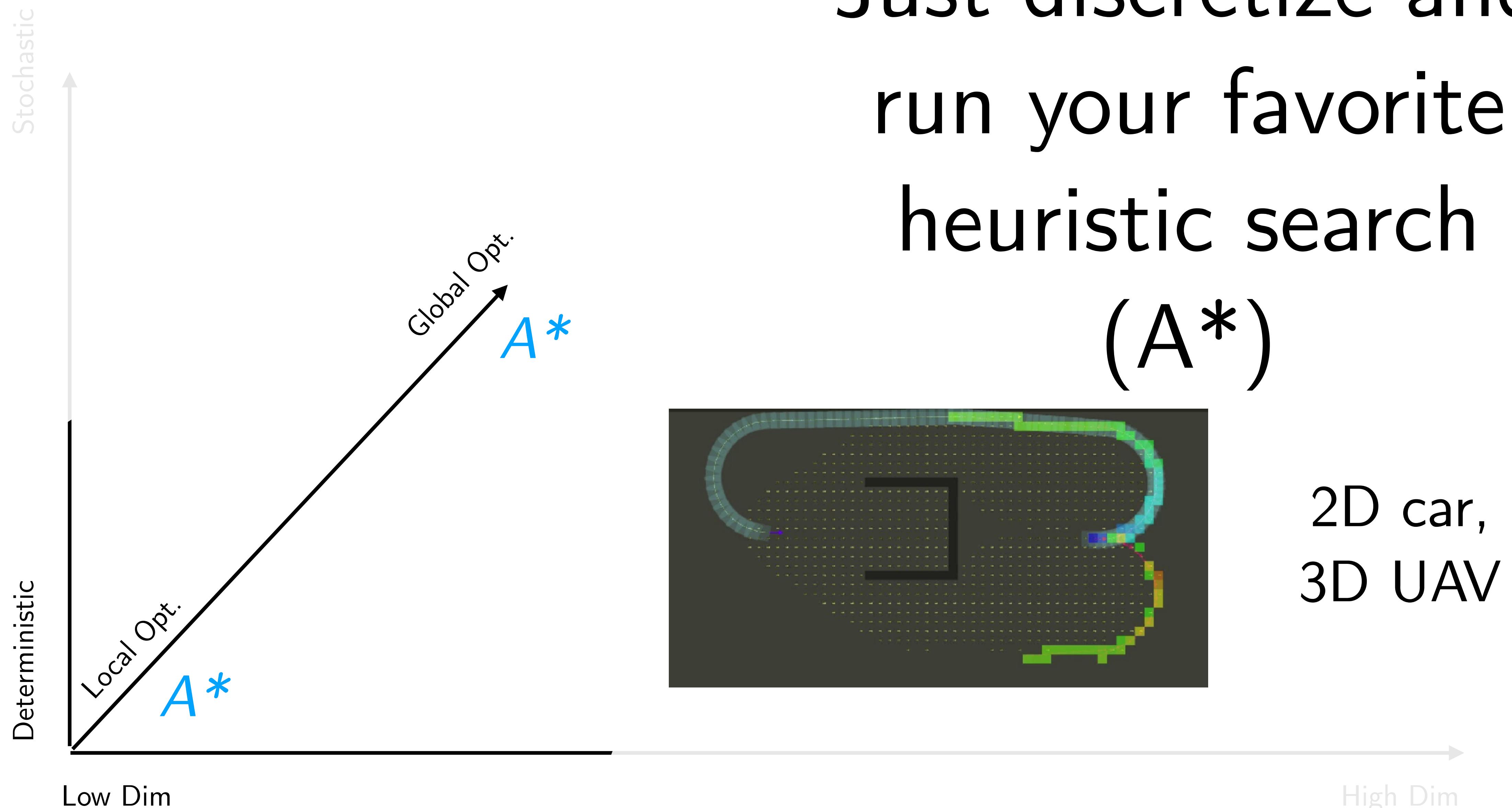
How do you convert a HARD MDP into an EASY one?

Solving continuous MDPs

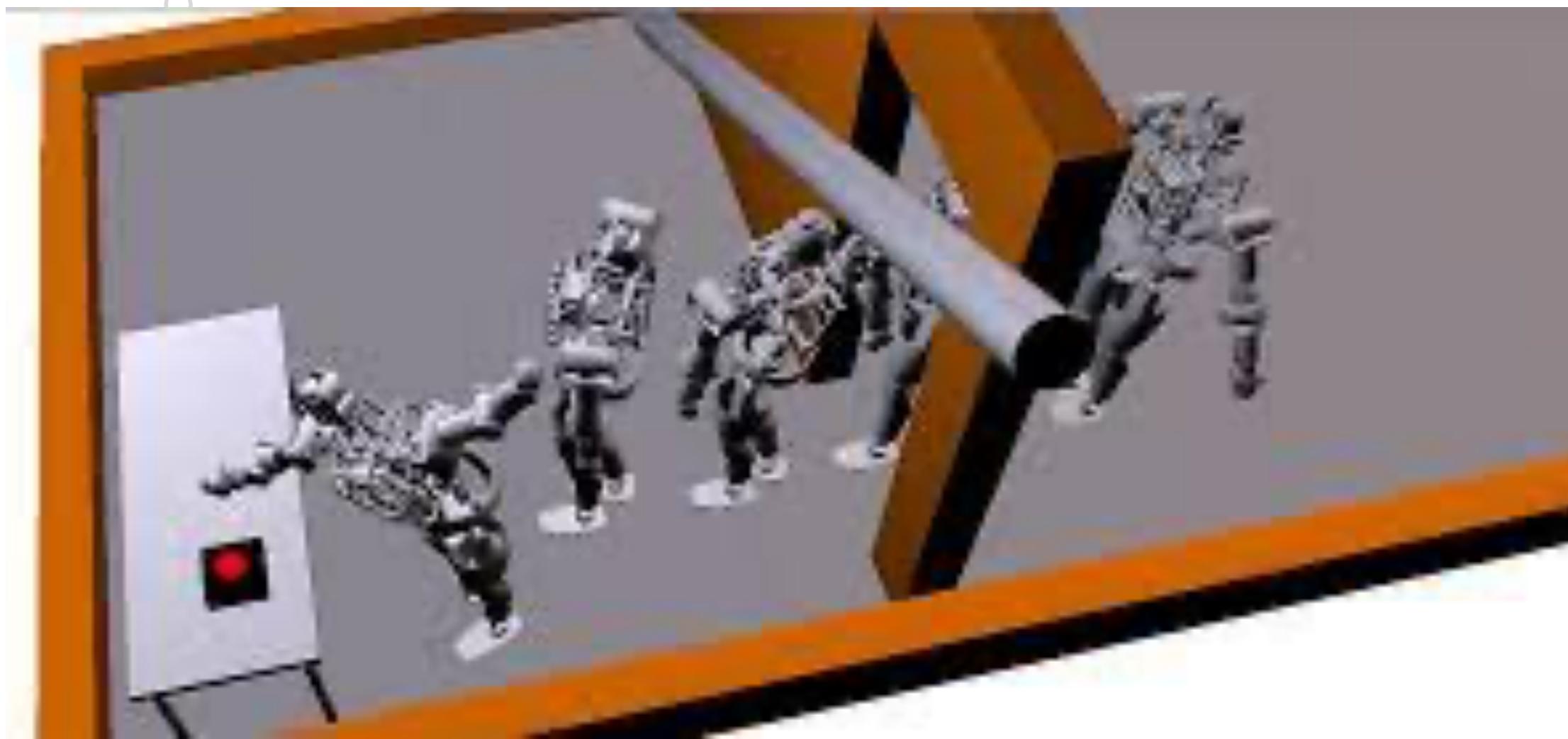




Just discretize and
run your favorite
heuristic search



Run some
sequential convex
trajectory optim.



Stochastic

Local Opt.

*A**

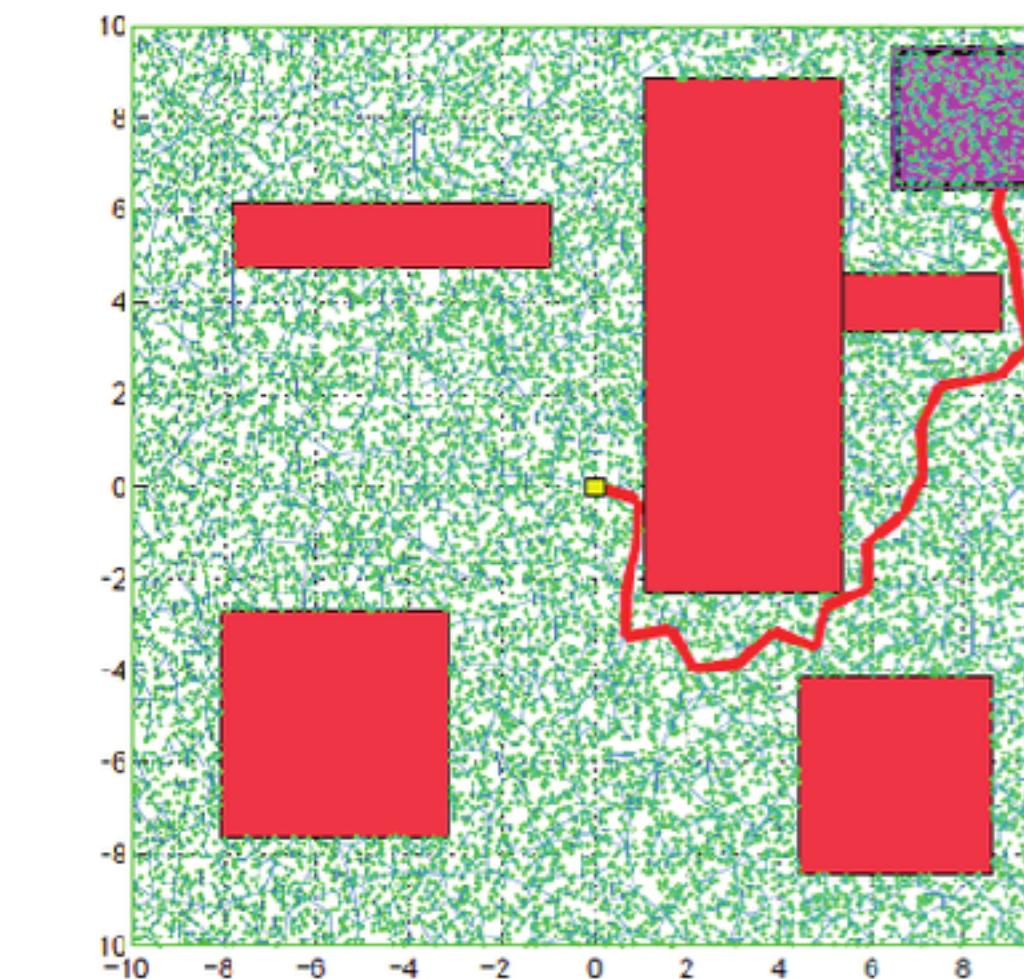
Low Dim

Global Opt.

*A**



Call a (slow) probabilistic
planner that “almost surely” gives
you the optimum
(Don’t hold your breath...)



*RRT**,
*BIT**

CHOMP,
TrajOPT

High Dim

