Value Iteration

Recap: Optimal Policy $\pi^\star$

For infinite horizon discounted MDP, there always exists a deterministic policy

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^\pi(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

i.e., $\pi^\star$ dominates any other policy π, everywhere!

We often denote $V^\star$, $Q^\star$ in short for $V^{\pi^\star}$, $Q^{\pi^\star}$

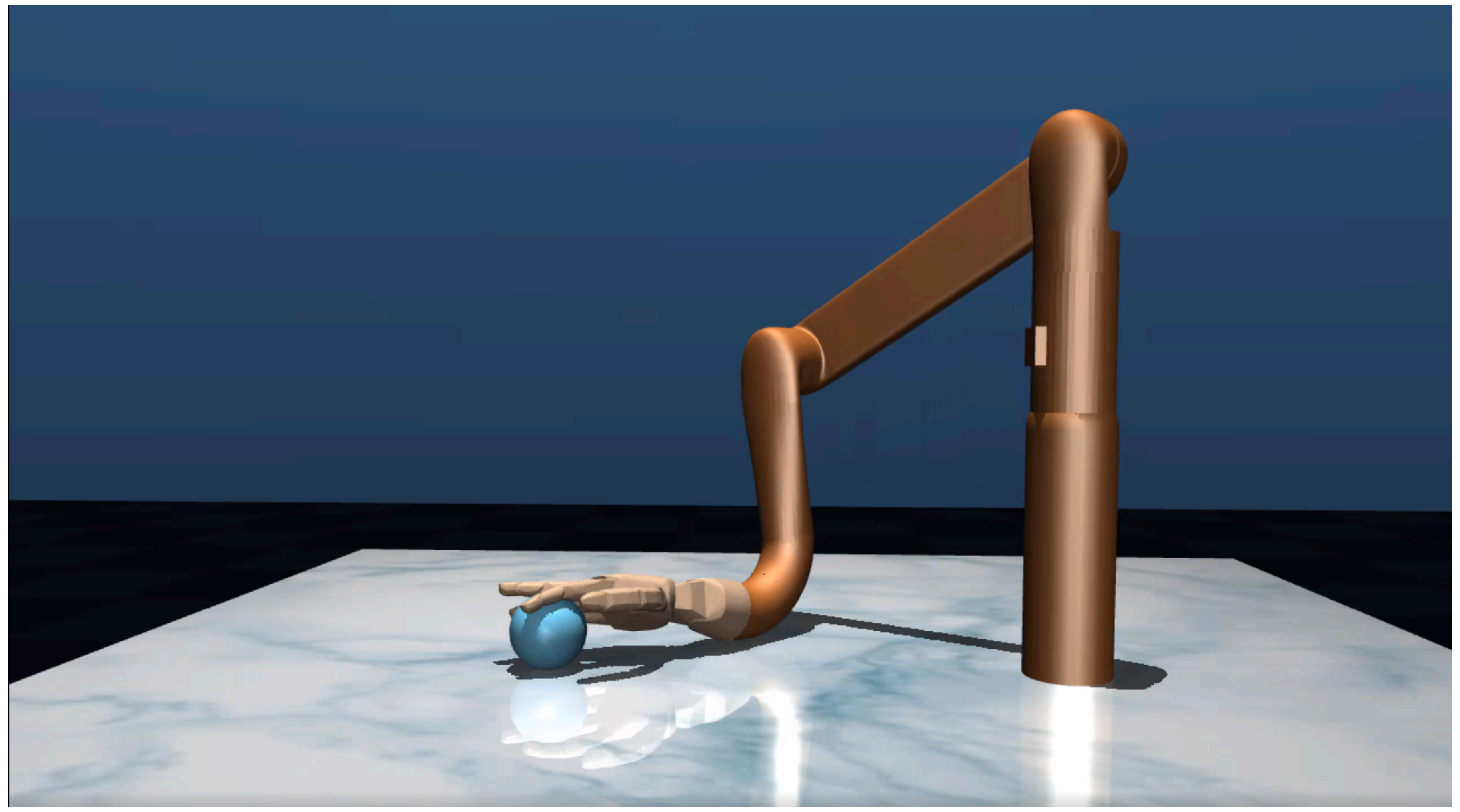
$V^\star(s)$: the maximum possible value we could possibly achieve

Question for Today and Wed:

Given an MDP $\mathcal{M} = (S, A, P, r, \gamma)$, How to find $Q^\star$ and $\pi^\star$ (approximately)

Motivation for Finding the Optimal Policy

Find the strategy w/ the highest prob of winning
(i.e., a policy that maps the board position to the next move)



Find the strategy (i.e., a mapping from robot & ball configuration to torques) that picks the ball and moves it to a goal position ASAP

Outline:

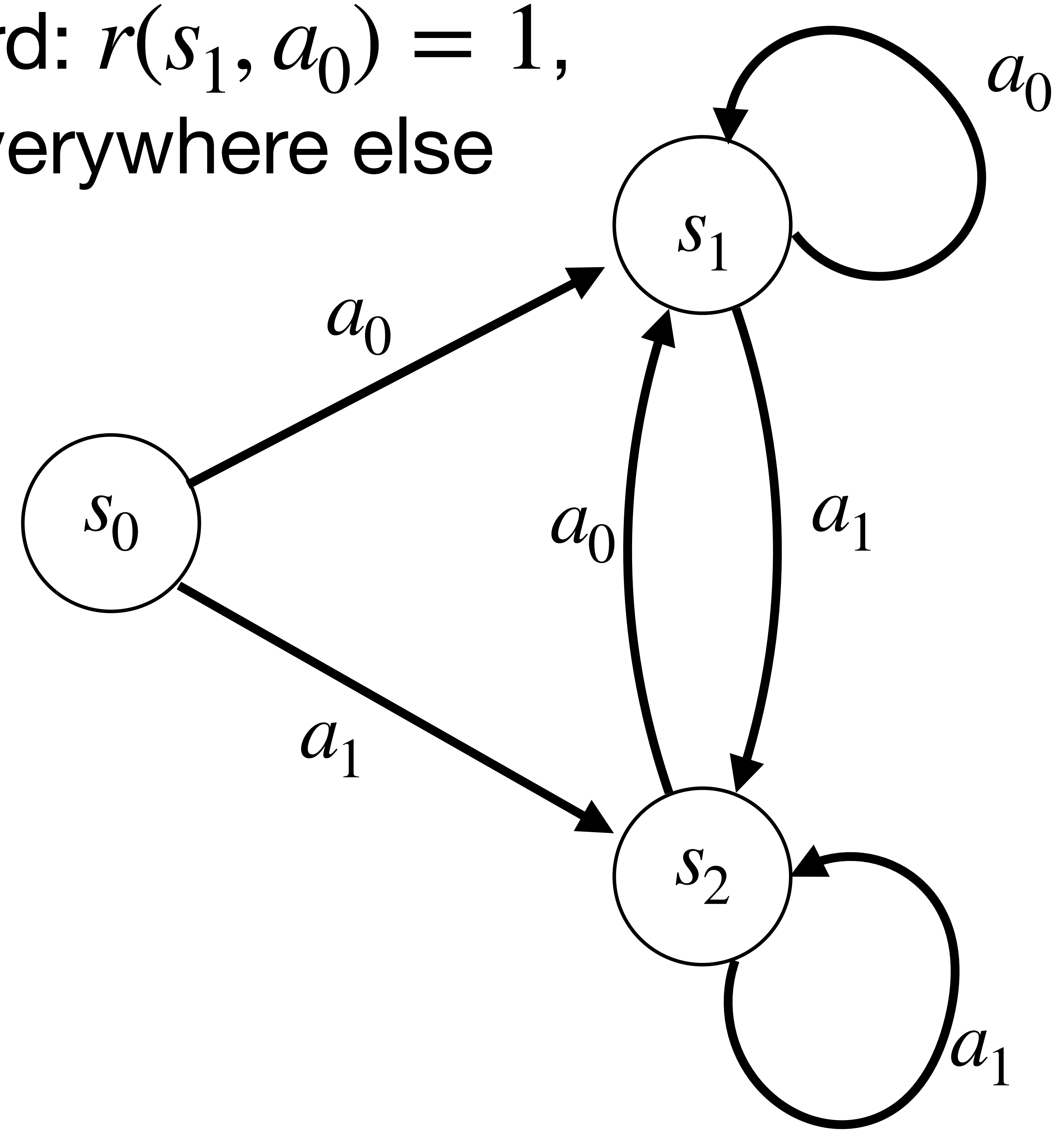
1: Bellman optimality

2: An Iterative Algorithm: Value Iteration

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions

Reward: $r(s_1, a_0) = 1$,
0 everywhere else

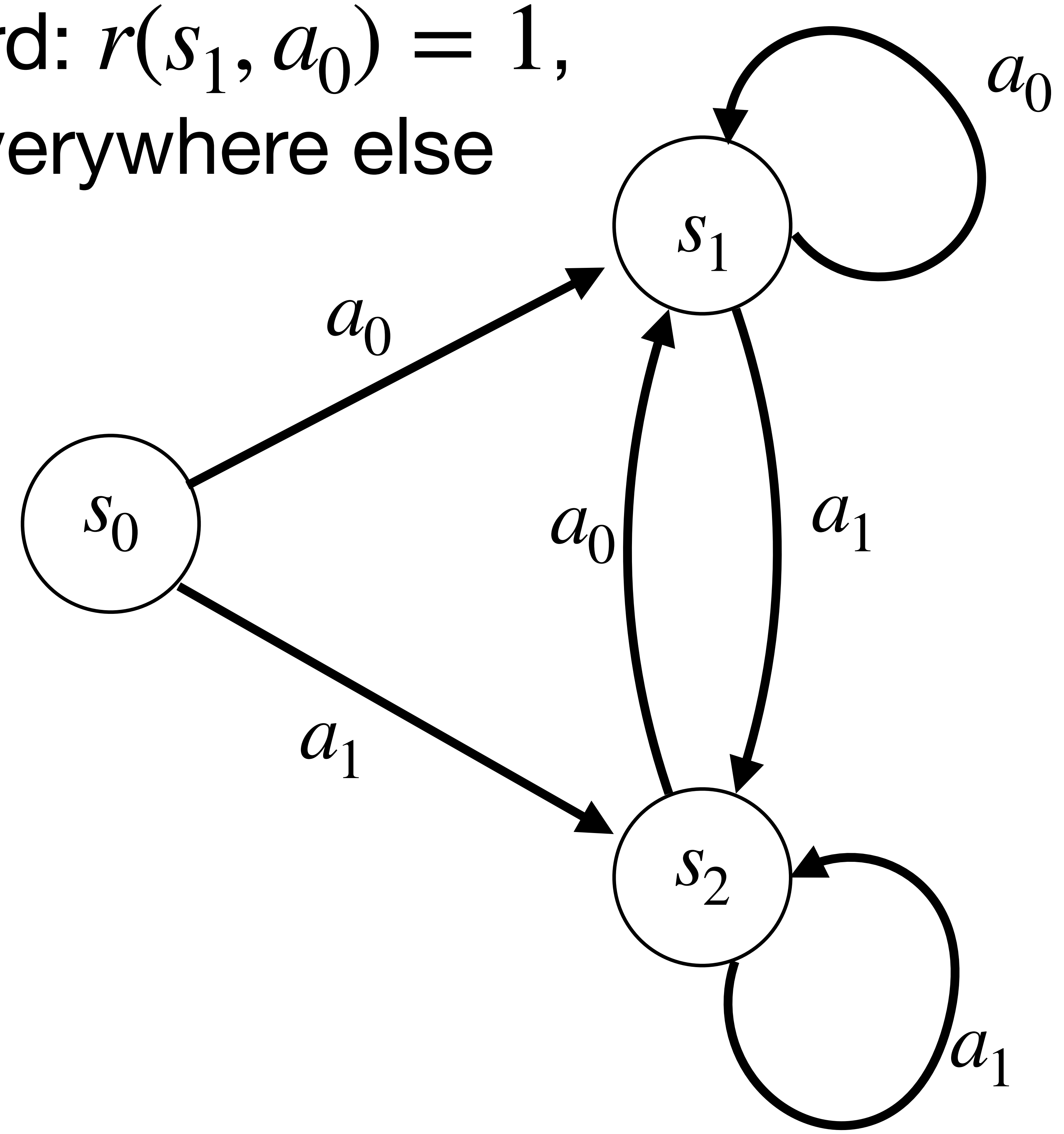


Let's say $\gamma \in (0,1)$
What's the optimal policy?

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions

Reward: $r(s_1, a_0) = 1$,
0 everywhere else



If we were told $V^\star(s_1)$ & $V^\star(s_2)$,
how to compute $V^\star(s_0)$

Bellman Optimality

Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right], \forall s$$

$$V^\star(s) = \left[r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi^\star(s))} V^\star(s') \right] \quad \text{(By BE of } \pi^\star\text{)}$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

If we took this $\arg\max_a$ at s, then follow $\pi^\star$, we **would have higher value**

Contradicts to the fact that $V^\star(s)$ is the maximum value at s one could possibly achieve

Bellman Optimality

Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Recall that $Q^\star(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s')$

This implies that $\arg\max_a Q^\star(s,a)$ is an optimal policy

An optimal policy should pick this action at s

Outline:

1: Bellman optimality

2: An Iterative Algorithm: Value Iteration

Define the Bellman optimality for $Q^\star$

We now know:

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

What's the Q version?

$$Q^\star(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^\star(s',a')$$

To estimate $Q^\star$, we will use the fix-point iterative approach again

Define the Bellman operator

Given a function $Q : S \times A \mapsto \mathbb{R}$,

$$\mathcal{T}Q : S \times A \mapsto \mathbb{R},$$

$$\big(\mathcal{T}Q\big)(s, a) := r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in A} Q(s', a'), \forall s, a \in S \times A$$

We can express $Q \in \mathbb{R}^{|S||A|}$, so $\mathcal{T}Q \in \mathbb{R}^{|S||A|}$

i.e., think about $\mathcal{T}$ as a (non-linear) mapping that maps from $\mathbb{R}^{|S||A|}$ to $\mathbb{R}^{|S||A|}$

High Level idea for Algorithm Design

Fix-point iteration again!

Recall Bellman Optimality for $Q^\star$:

$$Q^\star(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^\star(s',a')$$

We have $Q^\star = \mathcal{T} Q^\star$,

i.e., $Q^\star$ is a fix-point solution of $Q = \mathcal{T} Q$

Value Iteration Algorithm:

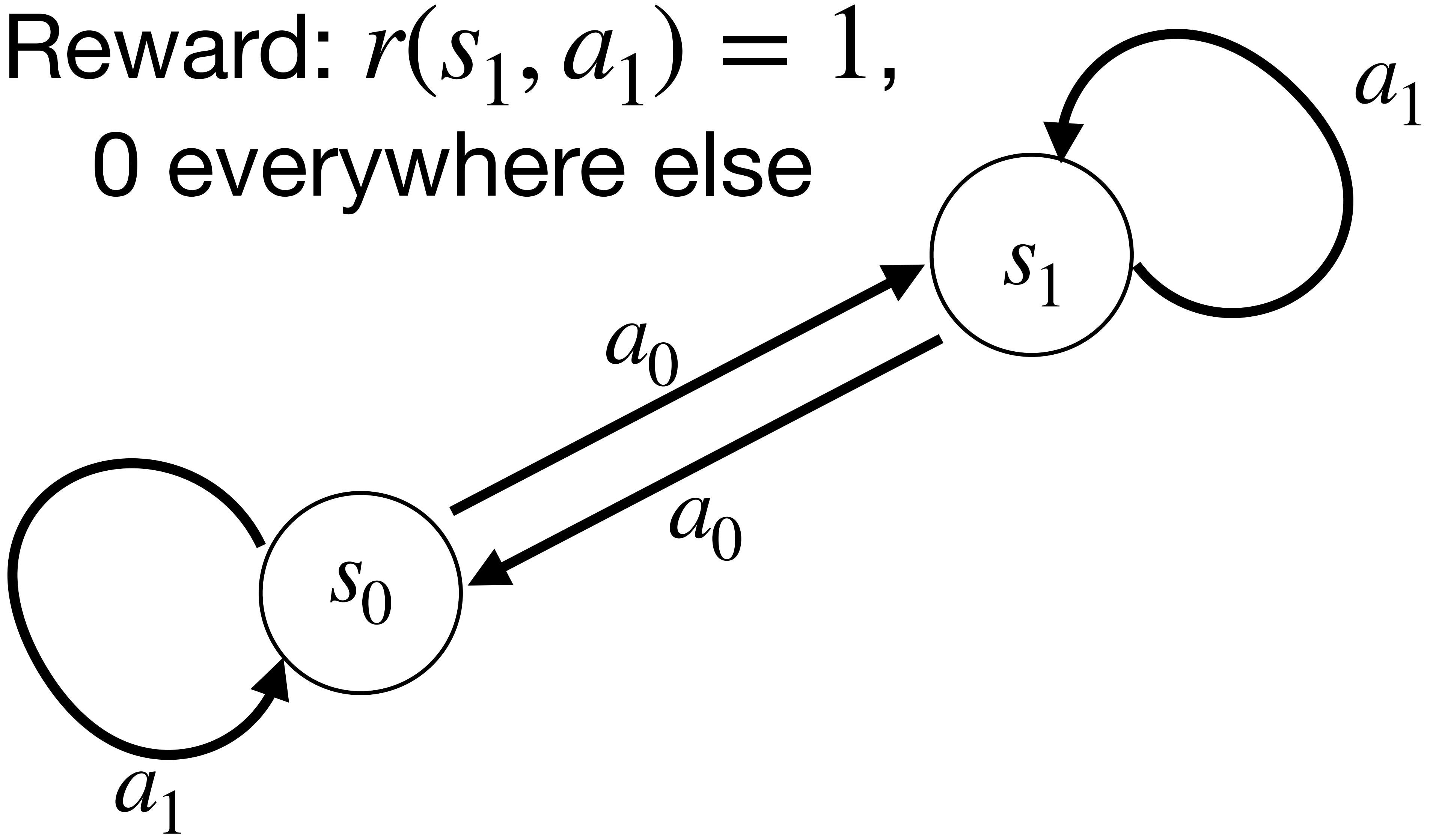
1. Initialization: Q^0

2. Iterate until convergence: $Q^{t+1} \Leftarrow \mathcal{T} Q^t$

We hope $Q^t \to Q^\star$, as $t \to \infty$

Exercise

Reward: $r(s_1, a_1) = 1$,
0 everywhere else



$$Q^{t+1}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^t(s', a')$$

$$Q^0(s,a) = 0, \forall s, a$$

1. Compute $Q^1(s,a), Q^2(s,a), \forall s, a$

2. Compute $\|Q^i - Q^\star\|_\infty$ for $i \in \{0,1,2\}$

3. Find a conlusion on how $\|Q^i - Q^\star\|_\infty$ behaves as i increases

Convergence of Value Iteration:

Lemma [contraction]: Given any Q, Q', we have:

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \gamma \|Q - Q'\|_\infty$$

Proof:

$$|(\mathcal{T}Q)(s,a) - (\mathcal{T}Q')(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q(s',a') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} \max_{a'} Q'(s',a') \right) \right.$$

$$\le \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left| \left(\max_{a'} Q(s',a') - \max_{a'} Q'(s',a') \right) \right| \quad \le \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right|$$

$$\le \gamma \max_{s'} \max_{a'} \left| \left(Q(s',a') - Q'(s',a') \right) \right| \quad = \gamma \|Q - Q'\|_\infty$$

Convergence of Value Iteration:

Lemma [Convergence]: Given Q^0, we have:

$$\|Q^t - Q^\star\|_\infty \leq \gamma^t \|Q^0 - Q^\star\|_\infty$$

$$\|Q^{t+1} - Q^\star\|_\infty = \|\mathcal{T} Q^t - \mathcal{T} Q^\star\|_\infty \leq \gamma \|Q^t - Q^\star\|_\infty$$

$$\ldots \leq \gamma^{t+1} \| \widehat{Q}^0 - Q^\star\|_\infty$$

Summary so far:

VI (a fix point iteration alg):

$$Q^{t+1} \Leftarrow \mathcal{T} Q^t$$

VI convergence (via contraction)

i.e., $\|Q^t - Q^\star\|_\infty \leq \gamma^t \|Q^0 - Q^\star\|_\infty$

What about the policy? Ultimately, we do want $\pi^\star$...