# Basics of Markov Decision Process

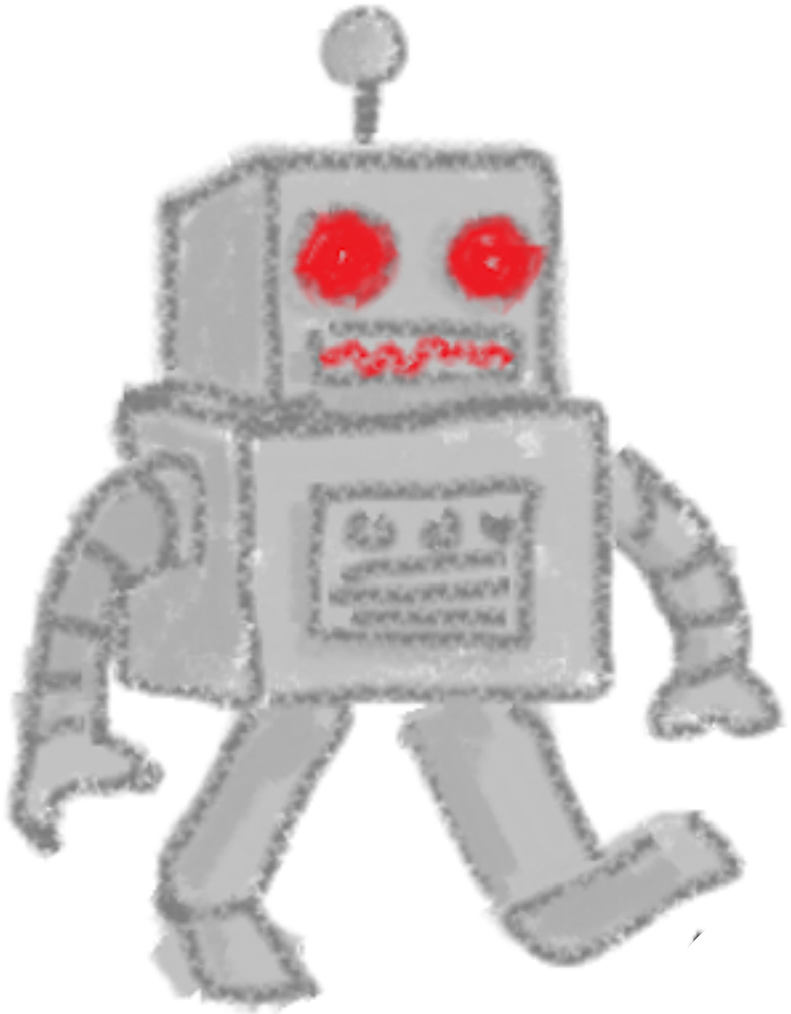
# Announcement:

1. For wait list, we will need to prioritize CS students and seniors (cs-course-enroll@cornell.edu)

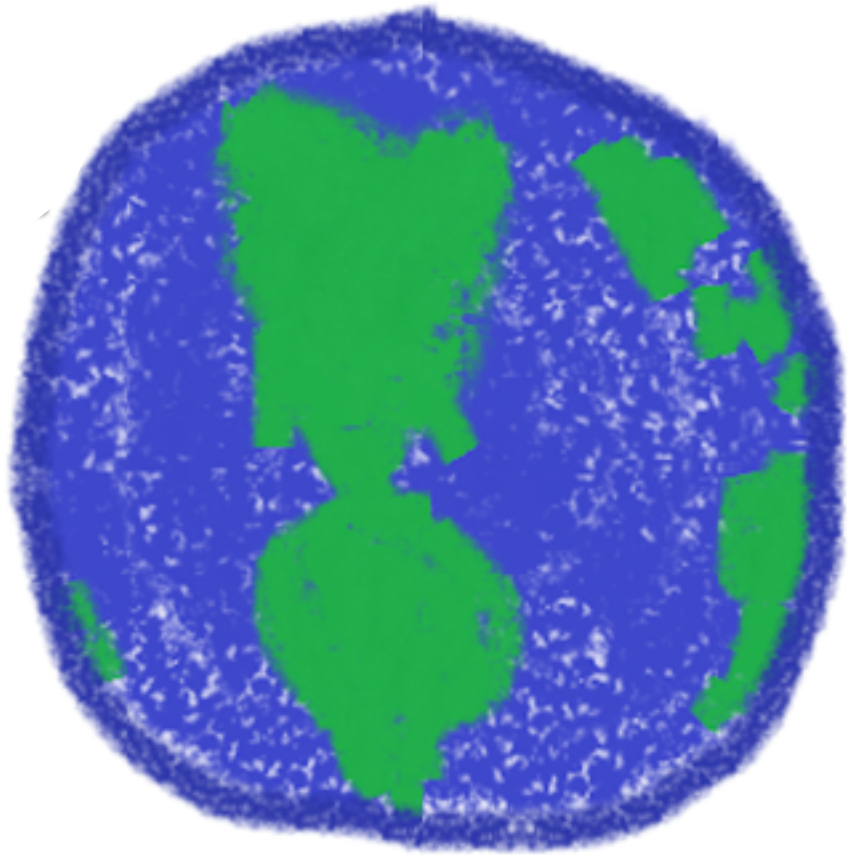
2. Clarification on the attendance bonus (5%)

# Recap:

Learning Agent



Environment



$$\pi(s) \rightarrow a$$

**Policy**: determine **action** based on **state**

**Multiple Steps**

Send **reward** and **next state** from a Markovian transition dynamics

$$r(s, a), s' \sim P(\cdot \mid s, a)$$

# Recap:

$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto (A)$

# Recap:

$$\mathscr{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto (A)$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \,|\, s_h, a_h)\right]$

# Recap:

$$\mathcal{M} = \{S, A, P, r, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto (A)$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Q function $Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| (s_0, a_0) = (s, a), a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

# Bellman Equation for V/Q-function:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

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

# Bellman Equation for V/Q-function:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, \pi(a))} V^\pi(s')$$

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

What is the BE for Q function??

# Bellman Equation for V/Q-function:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

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

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s,a), a_h = \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

What is the BE for Q function??

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

# Bellman Equation for Q-function:

$$\forall s, a: \quad Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')$$

# Today:

1. We have $A^S$ many policies, which one is the optimal policy $\pi^\star$?

2. Key property of the optimal policy $\pi^\star$

3. State-action distributions

# Definition of 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]

# Definition of 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 $\pi$, everywhere!**

# Definition of Optimal Policy $\pi^\star$

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

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

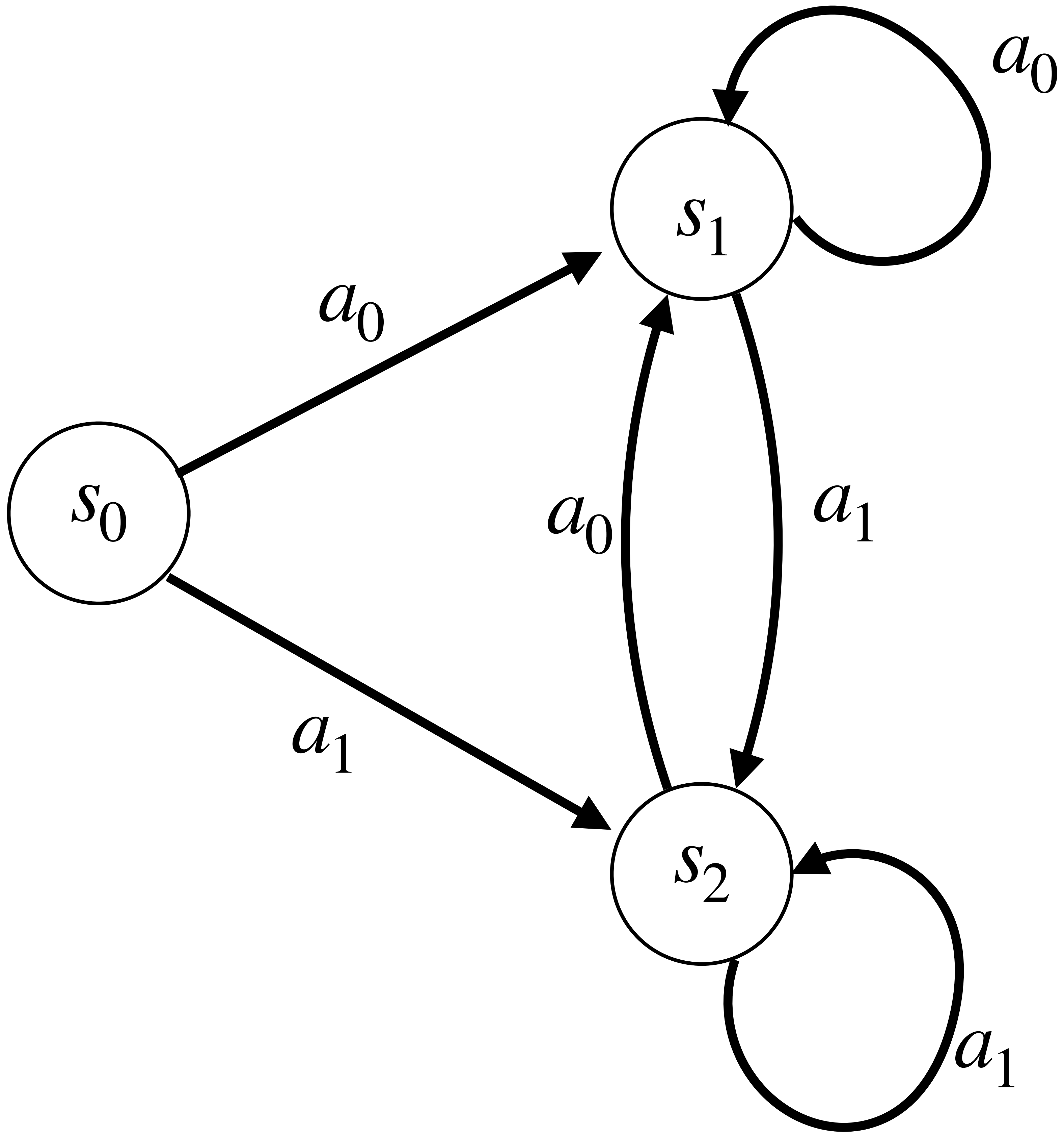
[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph—no need to understand the proof]

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

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

# Example of Optimal Policy $\pi^\star$

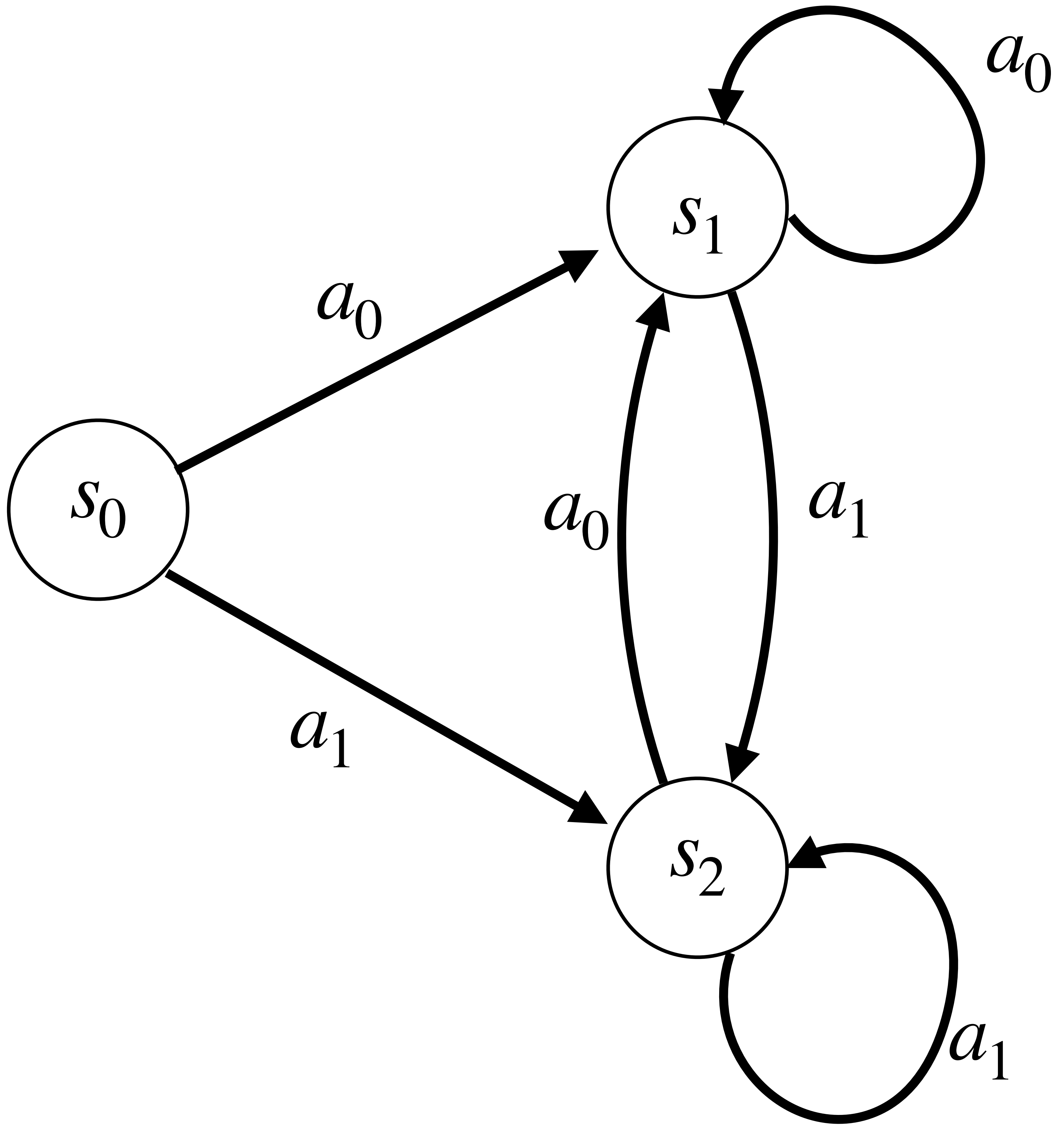
Consider the following **deterministic** MDP w/ 3 states & 2 actions




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

# Example of Optimal Policy $\pi^\star$

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



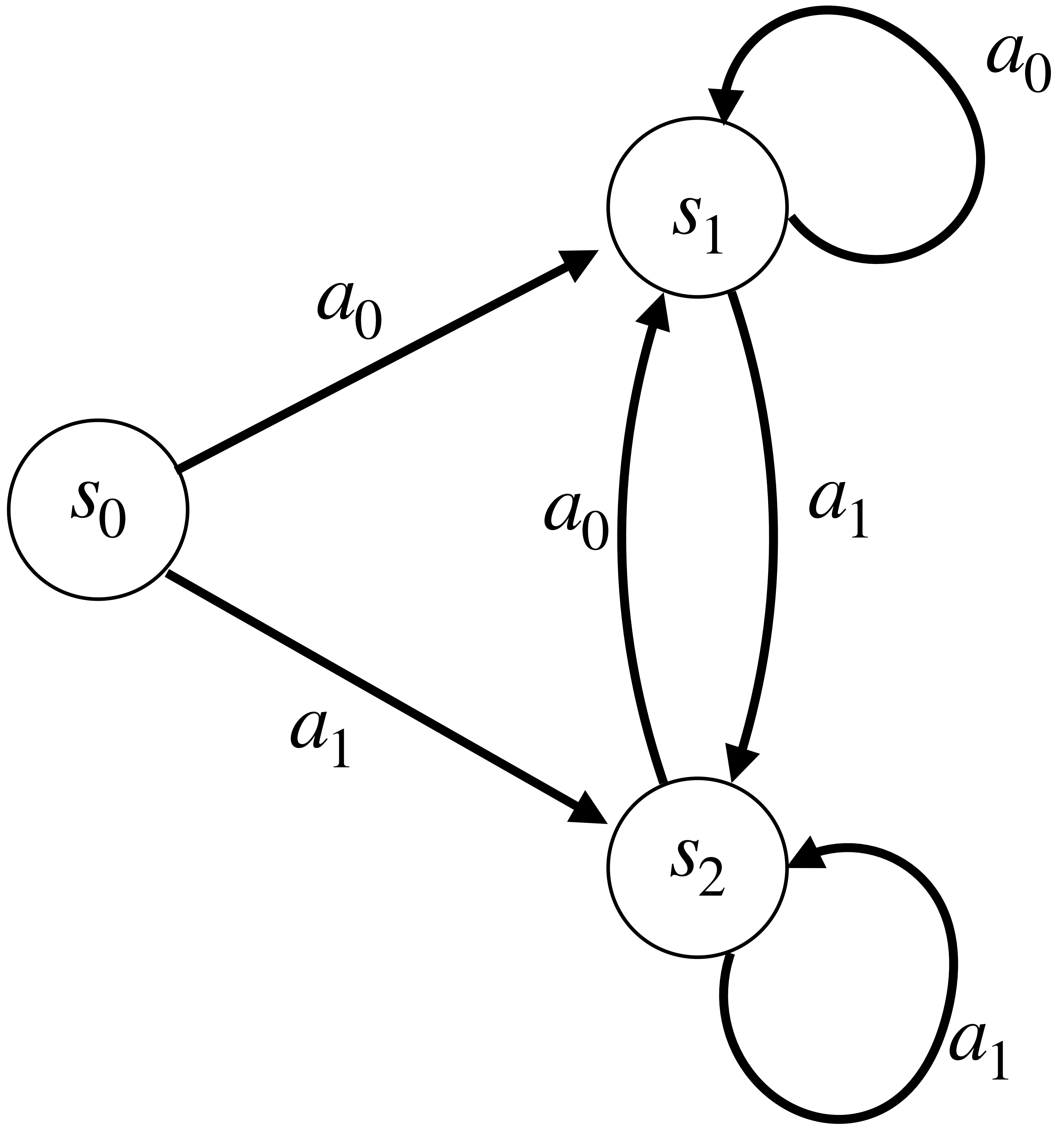

Reward: $r(s_1, a_0) = 1$, 0 everywhere else

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



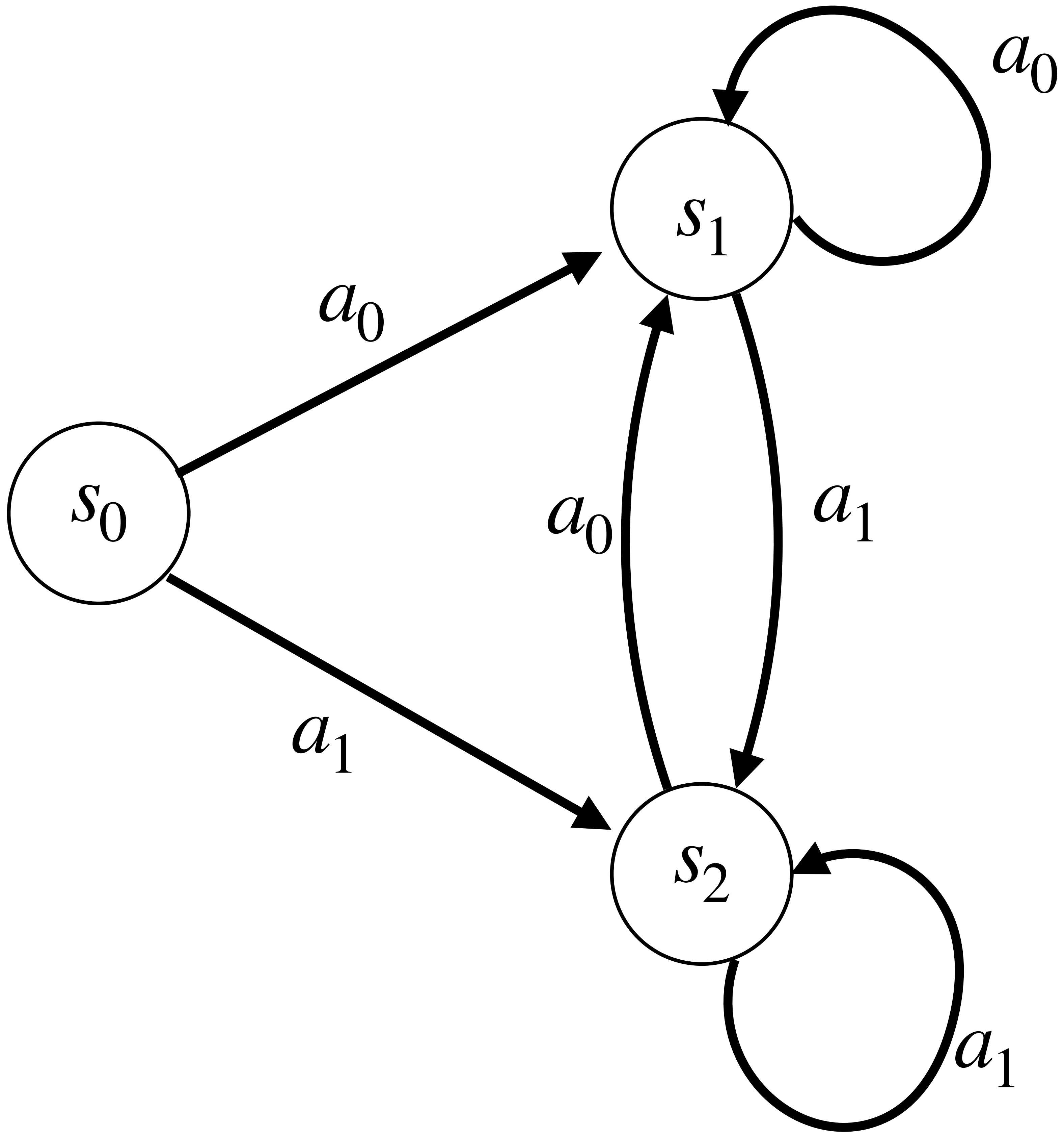

Reward: $r(s_1, a_0) = 1$, 0 everywhere else

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

$$\pi^\star(s) = a_0, \forall s$$

# 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

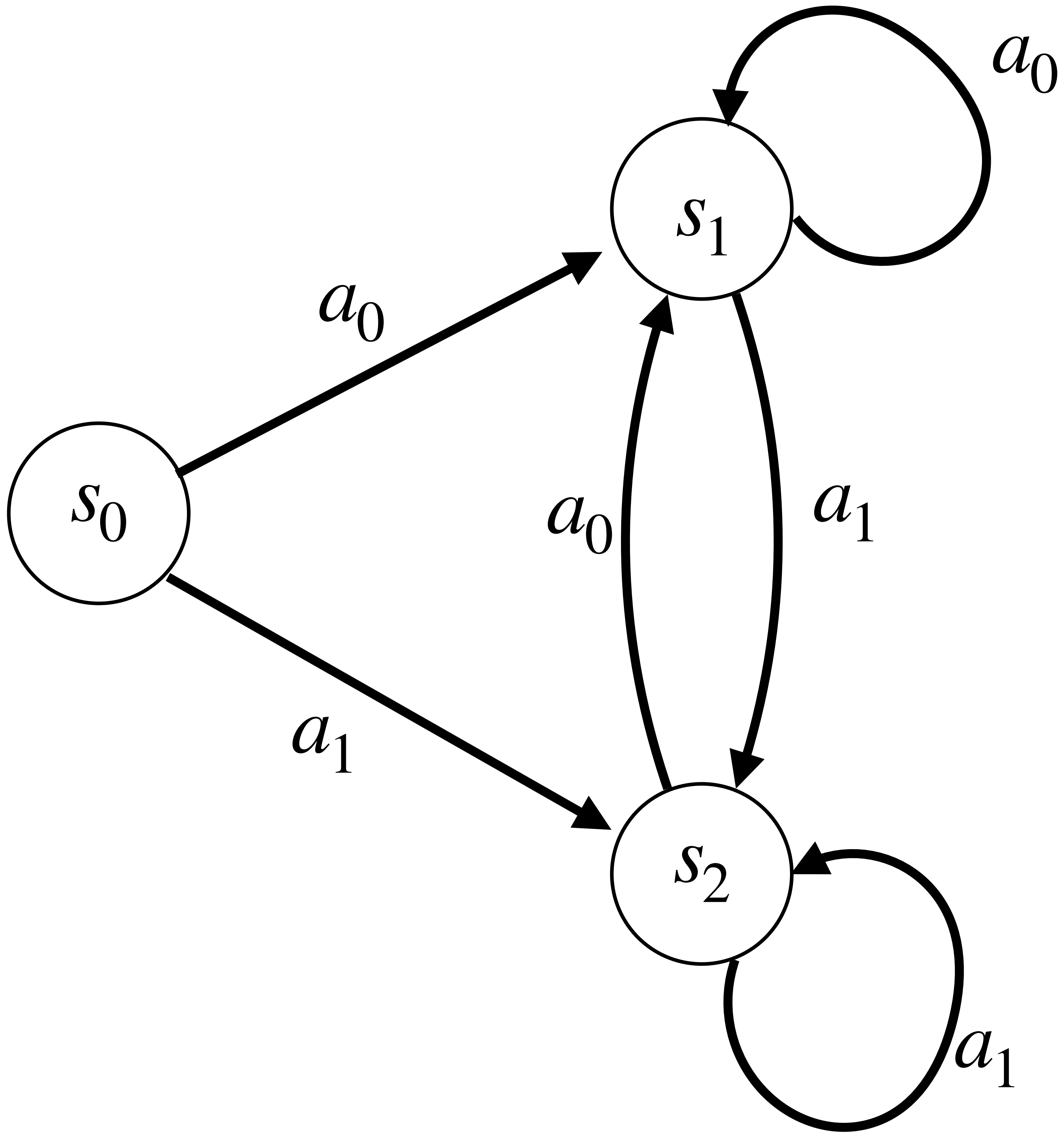
Let's say $\gamma \in (0,1)$
What's the optimal policy?

$$\pi^\star(s) = a_0, \forall s$$

$$V^\star(s_0) = \frac{\gamma}{1-\gamma}, V^\star(s_1) = \frac{1}{1-\gamma}, V^\star(s_2) = \frac{\gamma}{1-\gamma}$$

# 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?

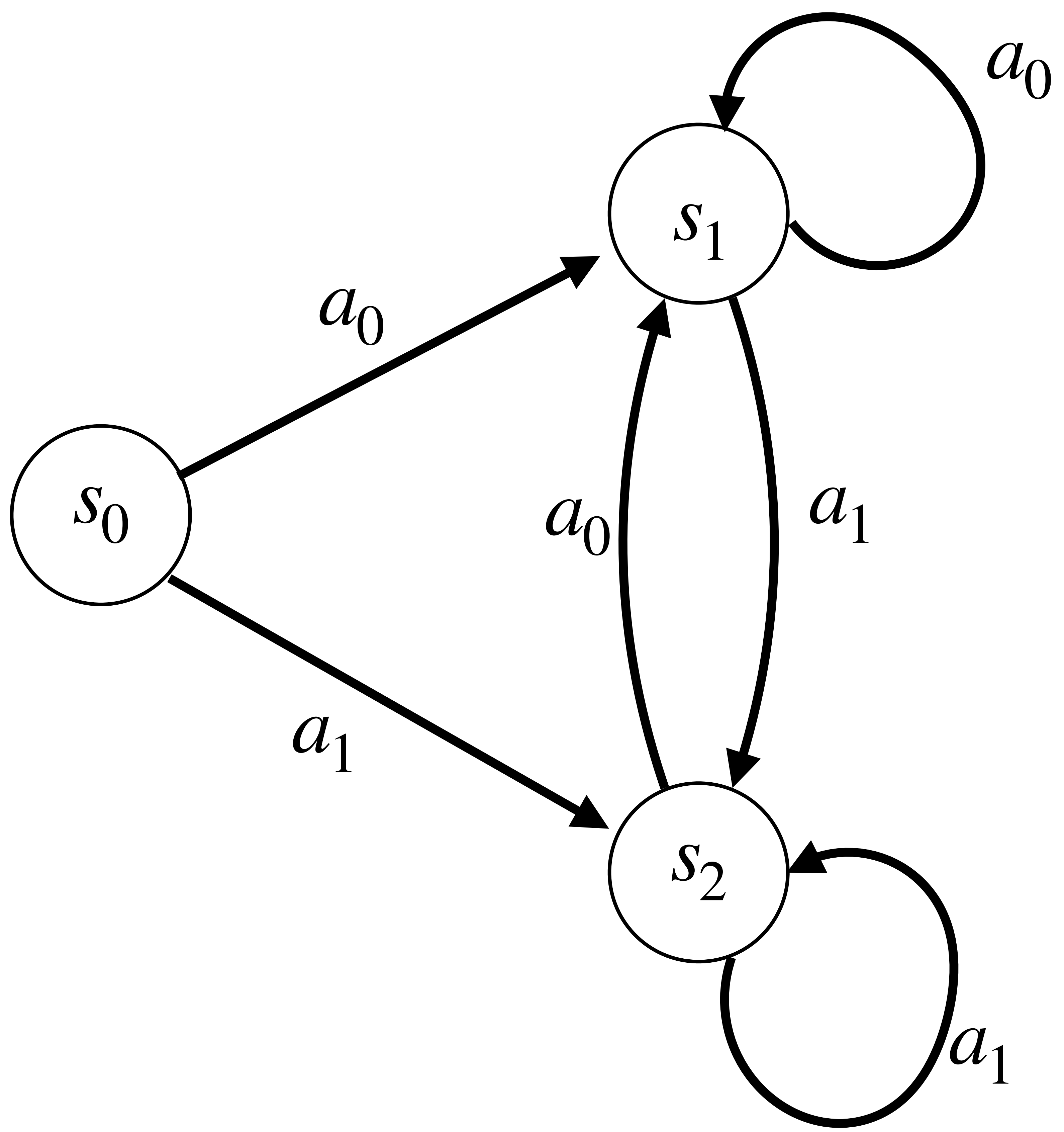
$$\pi^\star(s) = a_0, \forall s$$

$$V^\star(s_0) = \frac{\gamma}{1-\gamma}, V^\star(s_1) = \frac{1}{1-\gamma}, V^\star(s_2) = \frac{\gamma}{1-\gamma}$$

What about policy $\pi(s) = a_1, \forall s$

# 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?

$$\pi^\star(s) = a_0, \forall s$$

$$V^\star(s_0) = \frac{\gamma}{1-\gamma}, V^\star(s_1) = \frac{1}{1-\gamma}, V^\star(s_2) = \frac{\gamma}{1-\gamma}$$

What about policy $\pi(s) = a_1, \forall s$

$$V^\pi(s_0) = 0, V^\pi(s_1) = 0, V^\pi(s_2) = 0$$

# Summary so far:

Every discounted MDP has a deterministic optimal policy, that **dominates other policies everywhere** (proof is out of the scope)

$$V^\star(s) \geq V^\pi(s), \forall \pi, \forall s$$

# Outline

1. We have $A^S$ many policies, which one is the optimal policy $\pi^\star$?

**2. Key property of the optimal policy $\pi^\star$**

3. State-action distributions

# Bellman Optimality

**Theorem 1**: Bellman Optimality

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

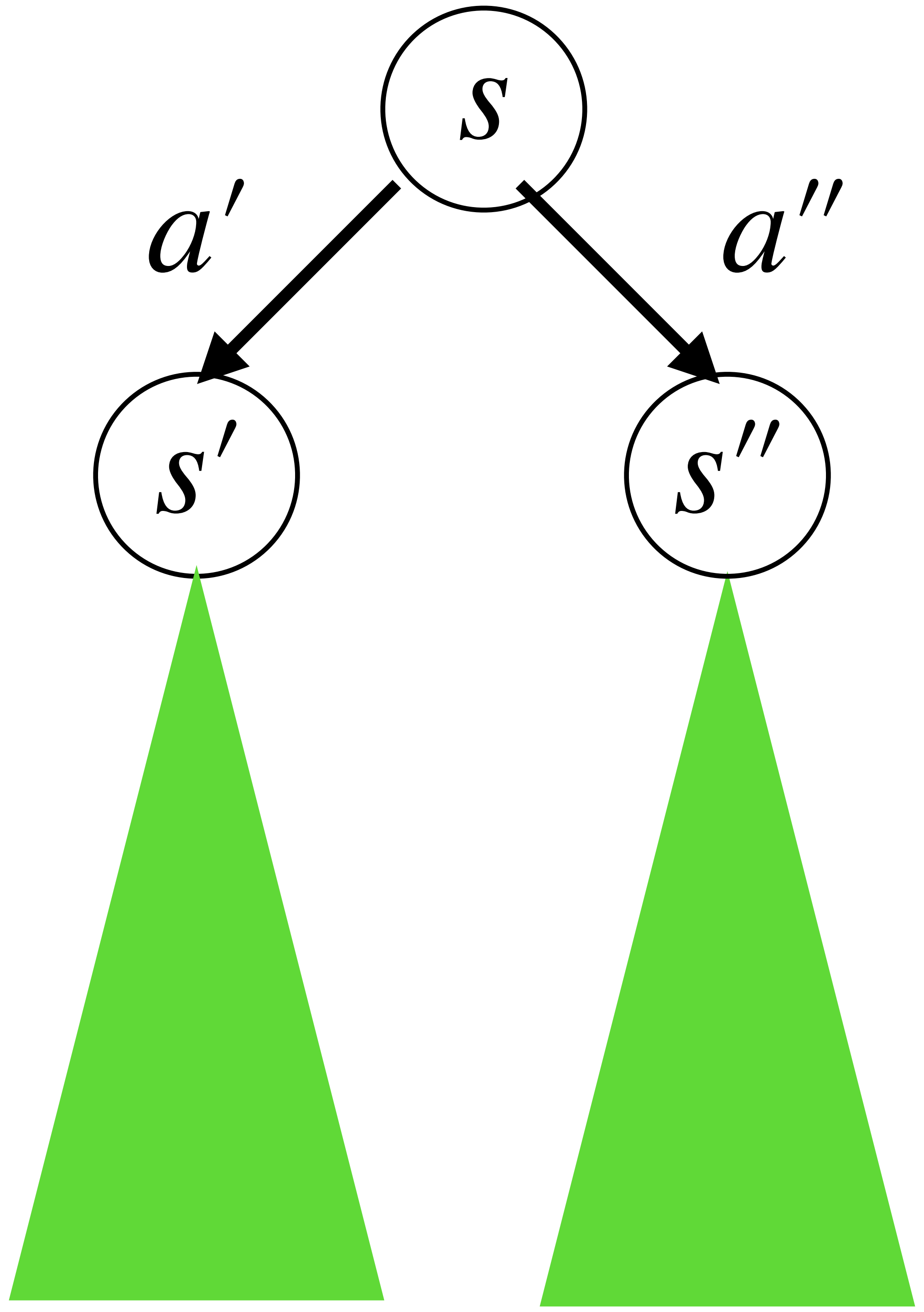
## Understanding Bellman Optimality

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

## Understanding Bellman Optimality

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

Q: If we know the optimal value at $s', s''$, i.e.,
$V^\star(s'), V^\star(s'')$, what we do at $s$?

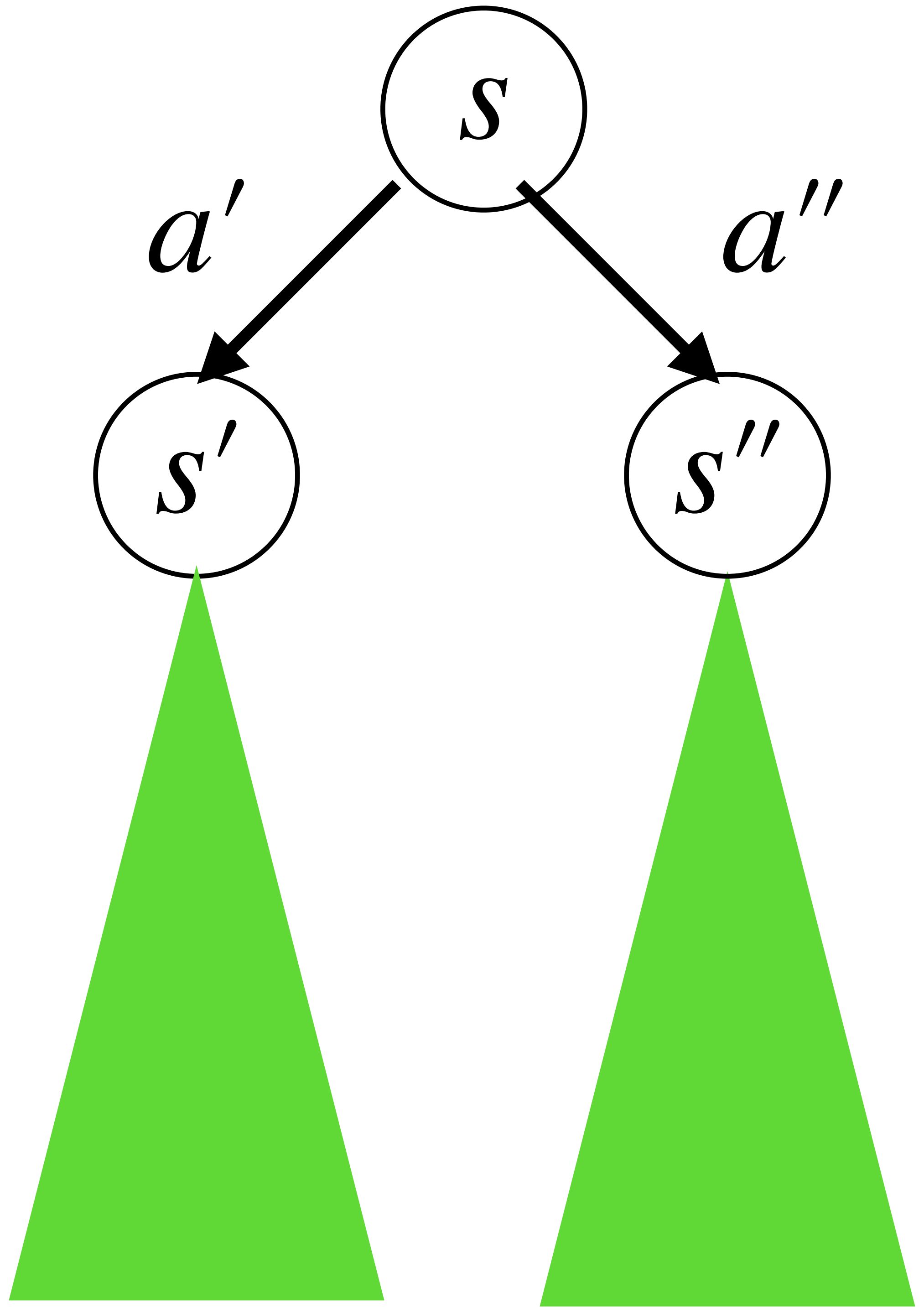


## Understanding Bellman Optimality

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

Q: If we know the optimal value at $s', s''$, i.e., $V^\star(s'), V^\star(s'')$, what we do at $s$?

1. Try $a'$, we get $Q^\star(s,a') := r(s,a') + \gamma V^\star(s')$

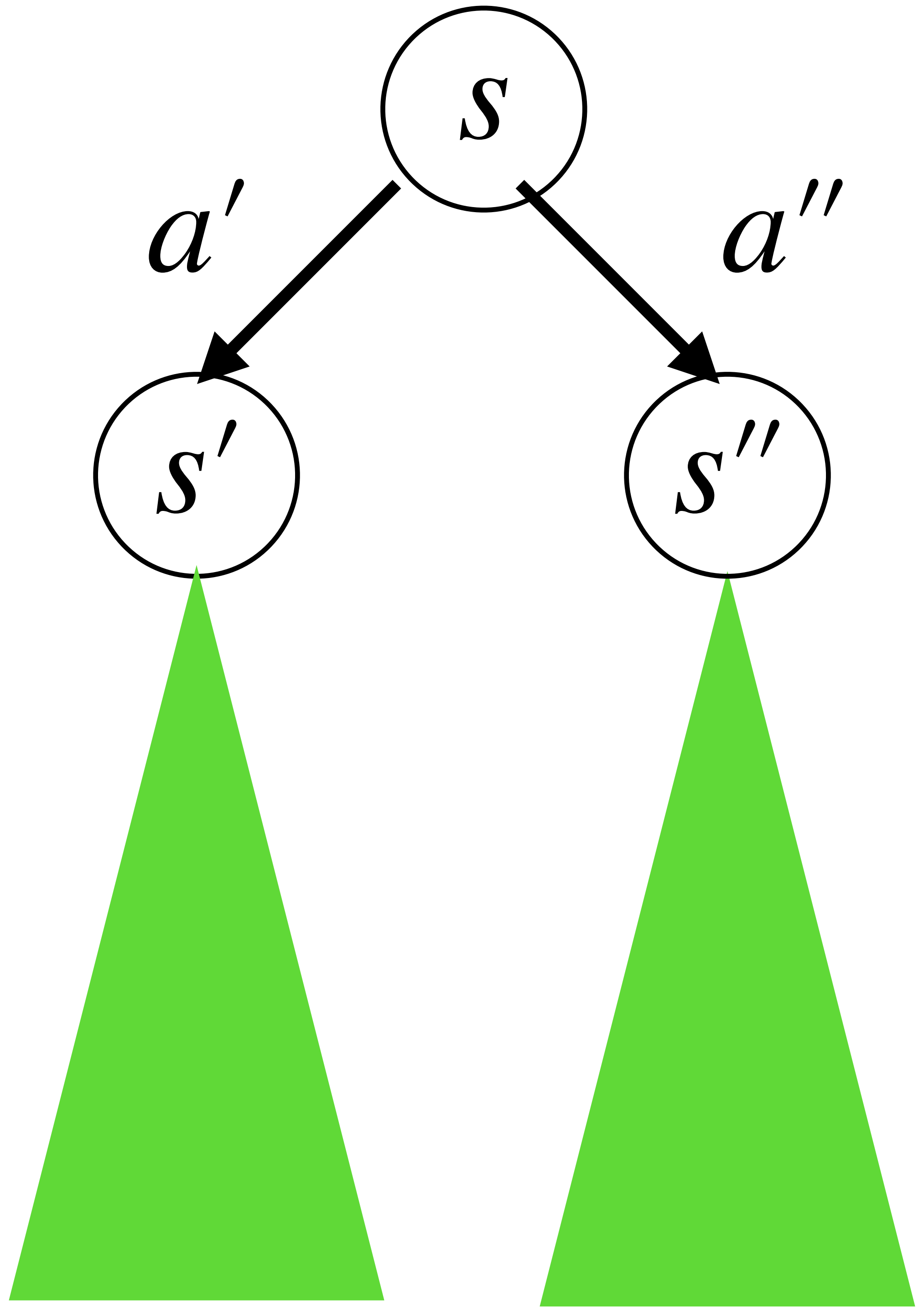


# Understanding Bellman Optimality

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

Q: If we know the optimal value at $s', s''$, i.e.,
$V^\star(s'), V^\star(s'')$, what we do at $s$?

1. Try $a'$, we get
$Q^\star(s,a') := r(s,a') + \gamma V^\star(s')$



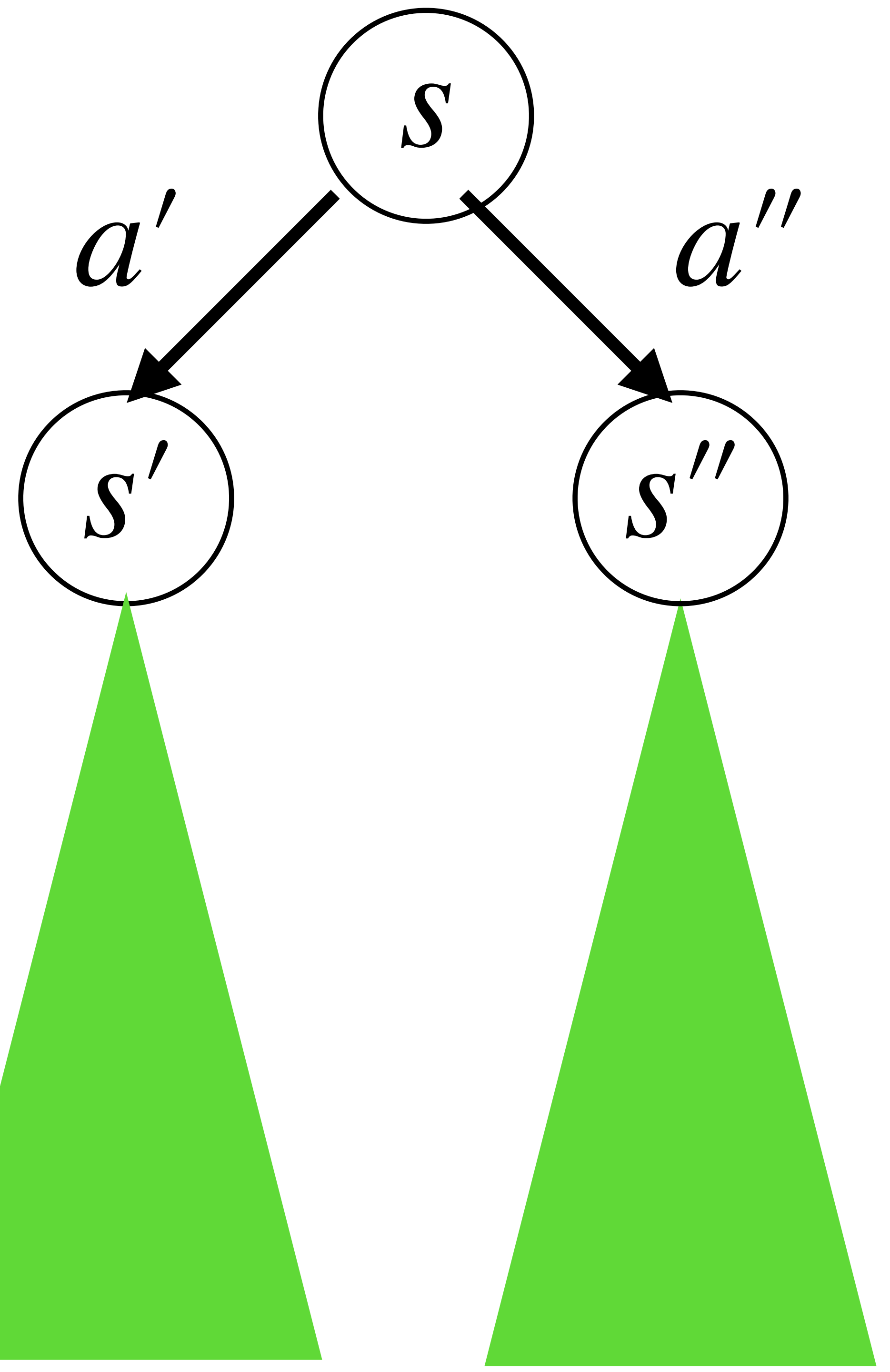

2. Try $a''$, we get
$Q^\star(s,a'') := r(s,a'') + \gamma V^\star(s'')$

## Understanding Bellman Optimality

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

Q: If we know the optimal value at $s', s''$, i.e.,
$V^\star(s'), V^\star(s'')$, what we do at $s$?

1. Try $a'$, we get
$Q^\star(s,a') := r(s,a') + \gamma V^\star(s')$




2. Try $a''$, we get
$Q^\star(s,a'') := r(s,a'') + \gamma V^\star(s'')$

$$V^\star(s) = \max_{a',a''} \left\{ Q^\star(s,a'), Q^\star(s,a'') \right\}$$

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\hat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\hat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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

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

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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

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

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

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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

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

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

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

$$\leq \mathbb{E} \left[ r(s, \widehat{\pi}(s)) + \gamma r(s', \widehat{\pi}(s')) + \ldots \right] = V^{\widehat{\pi}}(s)$$

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

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

# Proof of Bellman Optimality

**Theorem 1**: Bellman Optimality

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

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

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

Q: why?

# Summary so far:

Bellman Optimality and DP

**Theorem 1**: Bellman Optimality

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

# Summary so far:

Bellman Optimality and DP

**Theorem 1**: Bellman Optimality

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

Next:

Any function $V(s)$ that satisfies Bellman Optimality, MUST be equal to $V^\star$

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

Bellman Opt allows us to focus on just one step,

i.e., to check if $V = V^\star$,

we only need to check if $\left| V(s) - \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right] \right| = 0, \forall s,$

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

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

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

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

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

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

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

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

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

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

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

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

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

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left( \max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)$$

# Bellman Optimality

**Theorem 2:**

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all $s$,

then $V(s) = V^\star(s), \forall s$

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

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

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

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left( \max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)$$

$$\leq \max_{a_1, a_2, \ldots a_{k-1}} \gamma^k \mathbb{E}_{s_k} |V(s_k) - V^\star(s_k)|$$

# Summary so far:

1. $V^\star$ satisfies Bellman Optimality:

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

2. Any V that satisfies Bellman Optimality, i.e., $V(s) = \max_a \left[ r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right], \forall s,$

MUST be that $V(s) = V^\star(s)$, for all $s$

# Outline

1. We have $A^S$ many policies, which one is the optimal policy $\pi^\star$? ✓

2. Key property of the optimal policy $\pi^\star$: **Bellman Optimality** ✓

3. State-action distributions

What's the probability of $\pi$ visiting a particular state $s$?

# Discounted State (action) Occupancy Measures

Assume we start at $s_0$, following $\pi$ to step $h$, what's probability of seeing a trajectory:

$(s_0, a_0, s_1, a_1, \ldots, s_h, a_h)$?

# Discounted State (action) Occupancy Measures

Assume we start at $s_0$, following $\pi$ to step $h$, what's probability of seeing a trajectory:

$(s_0, a_0, s_1, a_1, \ldots, s_h, a_h)$?

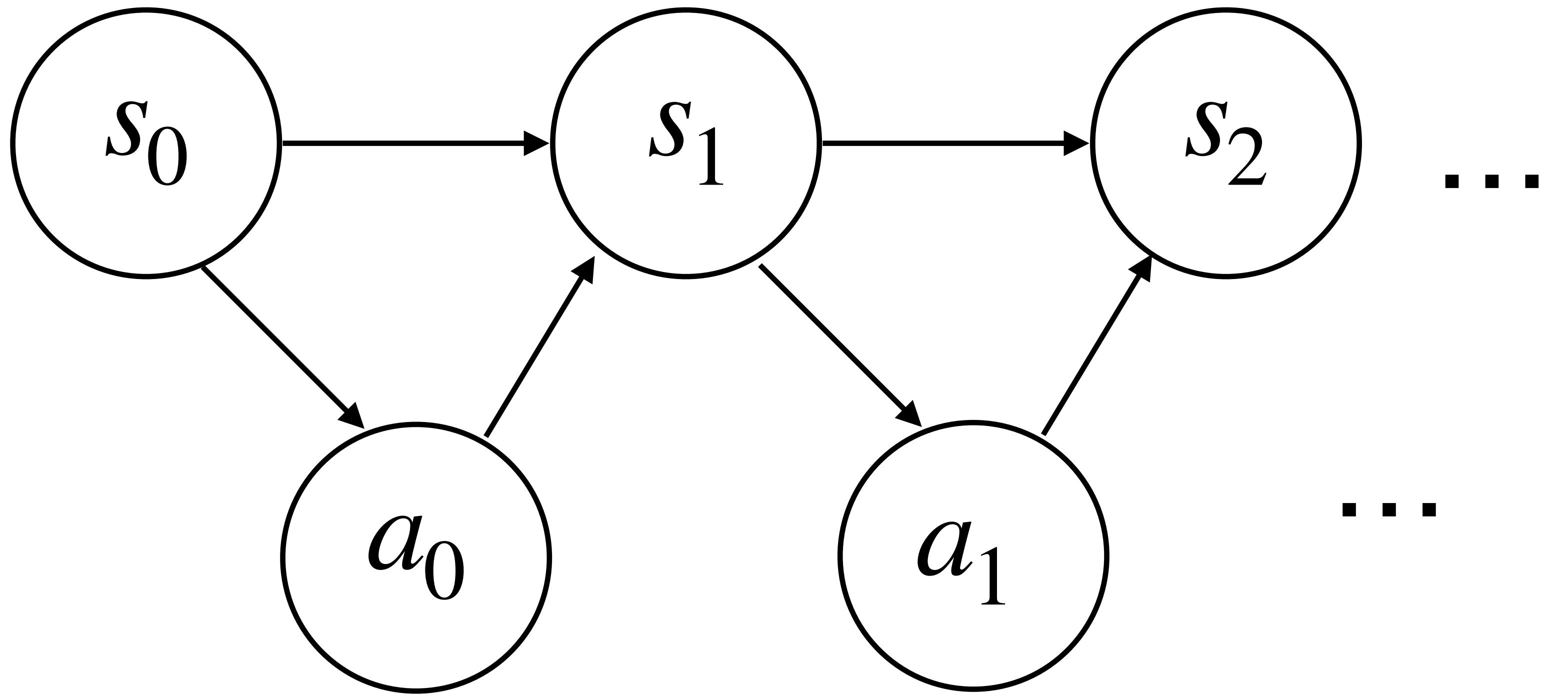
Let's write $\pi$ as a delta distribution, i.e., $\pi(a \mid s) = \begin{cases} 1, & a = \pi(s), \\ 0, & \text{else} \end{cases}$

# Discounted State (action) Occupancy Measures

Assume we start at $s_0$, following $\pi$ to step $h$, what's probability of seeing a trajectory:

$$(s_0, a_0, s_1, a_1, \ldots, s_h, a_h)?$$

Let's write $\pi$ as a delta distribution, i.e., $\pi(a \mid s) = \begin{cases} 1, & a = \pi(s), \\ 0, & \text{else} \end{cases}$

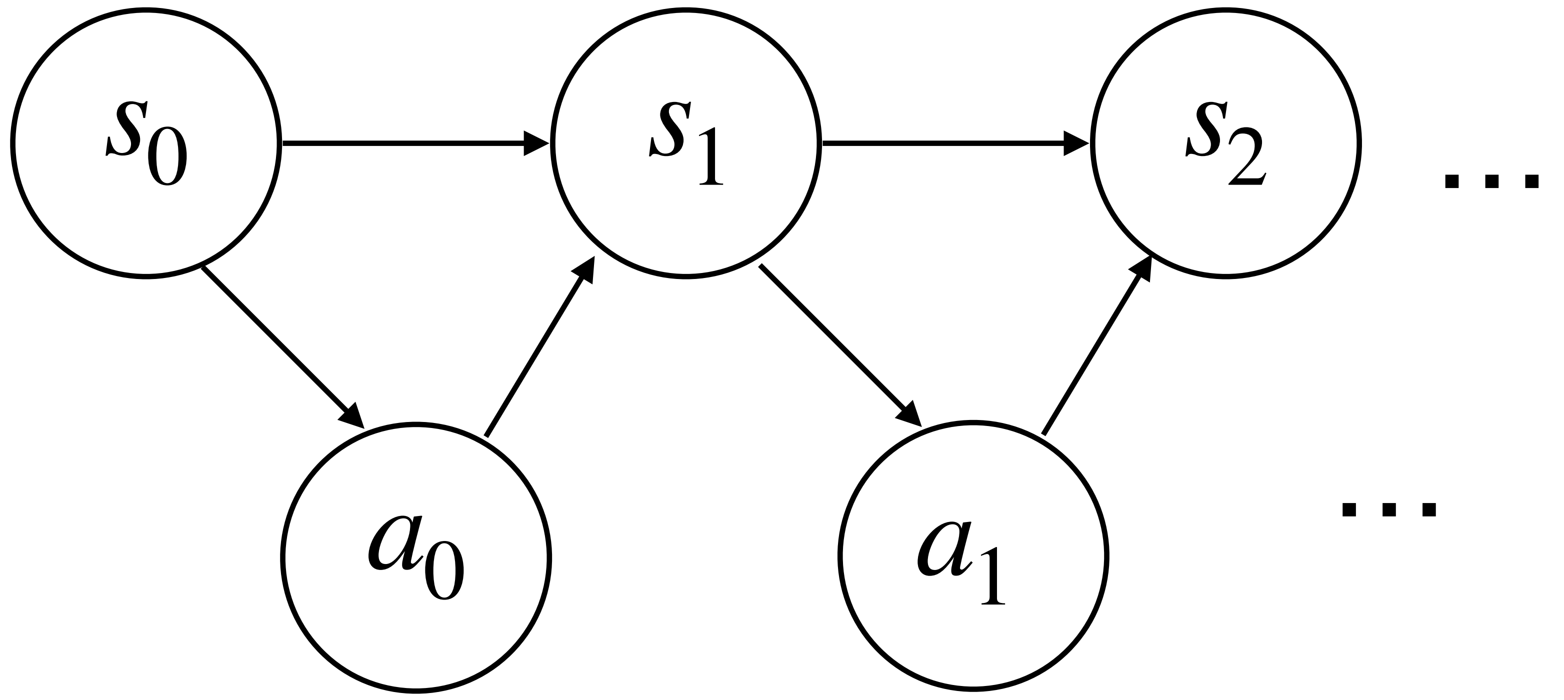


# Discounted State (action) Occupancy Measures

Assume we start at $s_0$, following $\pi$ to step $h$, what's probability of seeing a trajectory:

$$(s_0, a_0, s_1, a_1, \ldots, s_h, a_h)?$$

Let's write $\pi$ as a delta distribution, i.e., $\pi(a \mid s) = \begin{cases} 1, & a = \pi(s), \\ 0, & \text{else} \end{cases}$



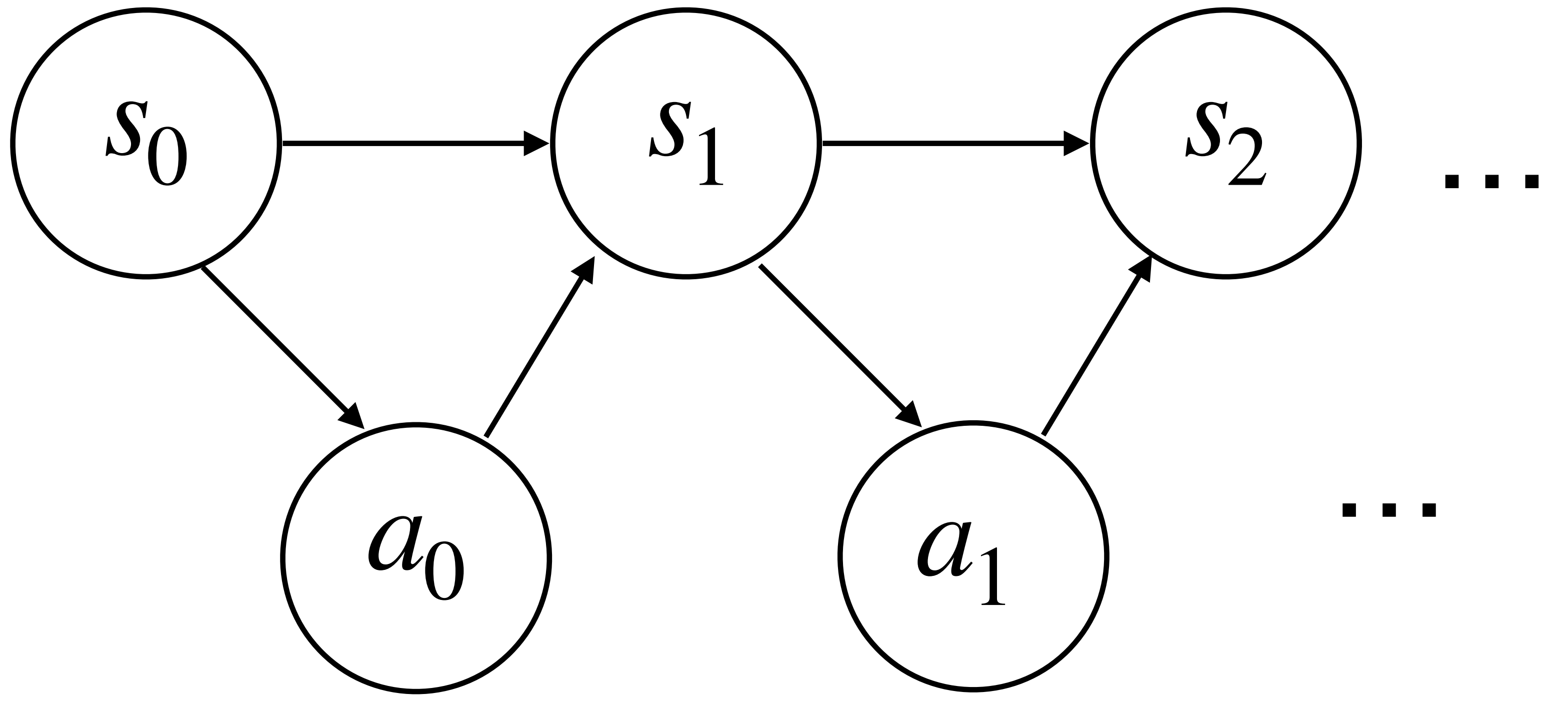

$\mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$

# Discounted State (action) Occupancy Measures

Assume we start at $s_0$, following $\pi$ to step $h$, what's probability of seeing a trajectory:

$$(s_0, a_0, s_1, a_1, \ldots, s_h, a_h)?$$

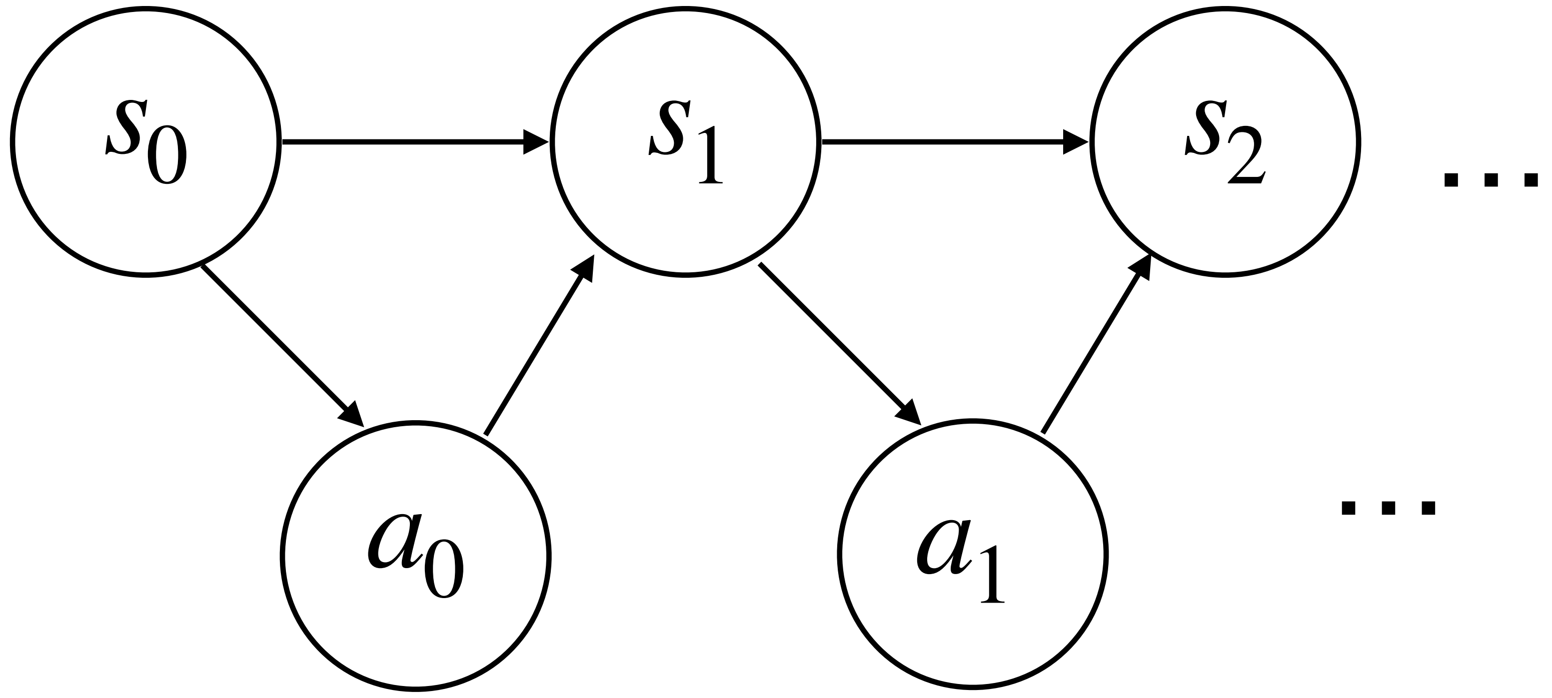
Let's write $\pi$ as a delta distribution, i.e., $\pi(a \mid s) = \begin{cases} 1, & a = \pi(s), \\ 0, & \text{else} \end{cases}$




$$\mathbb{P}^{\pi}(s_0, a_0, \ldots, s_h, a_h)$$

$$= \pi(a_0 \mid s_0) P(s_1 \mid s_0, a_0) \pi(a_1 \mid s_1) P(s_2 \mid s_1, a_1) \ldots P(s_h \mid s_{h-1}, a_{h-1}) \pi(a_h \mid s_h)$$

# State-action distribution at time step $h$

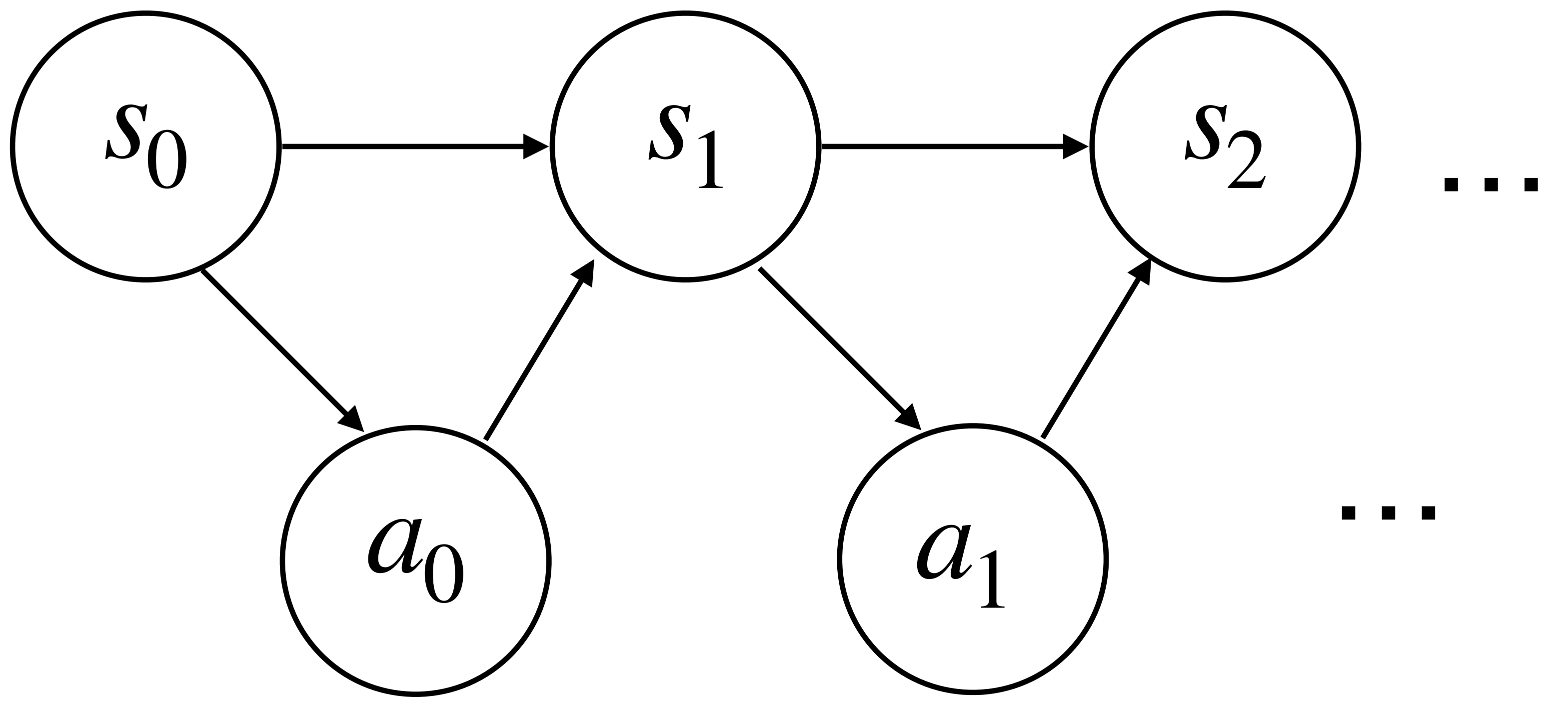



$$\mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$$

$$= \pi(a_0 \mid s_0) P(s_1 \mid s_0, a_0) \pi(a_1 \mid s_1) P(s_2 \mid s_1, a_1) \ldots P(s_h \mid s_{h-1}, a_{h-1}) \pi(a_h \mid s_h)$$

Q: what's the probability of $\pi$ visiting state ($s$,a) at time step h?

# State-action distribution at time step $h$




$$\mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$$

$$= \pi(a_0 \mid s_0) P(s_1 \mid s_0, a_0) \pi(a_1 \mid s_1) P(s_2 \mid s_1, a_1) \ldots P(s_h \mid s_{h-1}, a_{h-1}) \pi(a_h \mid s_h)$$

Q: what's the probability of $\pi$ visiting state ($s$,a) at time step h?

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$

Can you show that this is a valid distribution?

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$

Can you show that this is a valid distribution?

$$V^\pi(s_0) = \frac{1}{1 - \gamma} \sum_{s,a} d_{s_0}^\pi(s, a) r(s, a)$$

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$

Can you show that this is a valid distribution?

$$V^\pi(s_0) = \frac{1}{1 - \gamma} \sum_{s,a} d_{s_0}^\pi(s, a) r(s, a)$$

Can you show the above is true?

# Discounted Average State-action distribution

Probability of $\pi$ visiting $(s, a)$ at $h$, starting from $s_0$

$$\mathbb{P}_h^\pi(s, a; s_0) = \sum_{a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1} s_h = s, a_h = a)$$

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a; s_0)$$

Can you show that this is a valid distribution?

$$V^\pi(s_0) = \frac{1}{1 - \gamma} \sum_{s,a} d_{s_0}^\pi(s, a) r(s, a)$$

**HW0 questions!**

Can you show the above is true?

# Summary for today:

1. $\pi^\star$ **dominates** other policies, i.e., $V^\star(s) \geq V^\pi(s), \forall s, \pi$

2. Key property of the optimal policy $\pi^\star$: **Bellman Optimality**

(BE and B-Opt allow us to focus on one step)

3. State-action distribution: $d^\pi_{s_0}(s, a)$

# Summary for today:

1. $\pi^\star$ **dominates** other policies, i.e., $V^\star(s) \geq V^\pi(s), \forall s, \pi$

2. Key property of the optimal policy $\pi^\star$: **Bellman Optimality**

(BE and B-Opt allow us to focus on one step)

3. State-action distribution: $d^\pi_{s_0}(s, a)$

RL is notation heavy! But we will see these over and over again during the semester.