Infinite Horizon MDPs

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2023

Today

- ✓ Recap
- Infinite Horizon MDPs
  - Policy Evaluation
  - Optimality & the Bellman Equations
  - Value Iteration
  - Policy Iteration

Recap

Bellman Consistency (theorem)

- For a fixed policy, $\pi := \{\pi_0, \pi_1, \ldots, \pi_{H-1}\}, \pi_h : S \mapsto A, \forall h,$
- By definition, $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$
- At $H-1$, $Q_{H-1}^\pi(s,a) = r(s,a), \ V_{H-1}^\pi(s) = r(s, \pi_{H-1}(s))$
- Bellman consistency conditions: for a given policy π,
    - $V_h^\pi(s) = r(s, \pi_h(s)) + \mathbb{E}_{s' \sim P(\cdot|s,\pi_h(s))}\left[V_{h+1}^\pi(s')\right]$
    - $Q_h^\pi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right]$

Computation of V^π via Backward Induction

- For a fixed policy, $\pi := \left\{\pi_0, \pi_1, \ldots, \pi_{H-1}\right\}, \pi_h : S \mapsto A, \forall h,$

  Bellman consistency $\Longrightarrow$ we can compute V_h^π, assuming we know the MDP.

- Init: $V_H^\pi(s) = 0, \ \forall s \in S$
- For $h = H - 1, \ldots 0$, set:

  $V_h^\pi(s) = r(s, \pi_h(s)) + \mathbb{E}_{s' \sim P(\cdot|s, \pi_h(s))} \left[V_{h+1}^\pi(s') \right], \ \forall s \in S$

- What is the per iteration computational complexity of DP?

  (assume scalar $+, -, \times, \div$ are $O(1)$ operations)
- What is the total computational complexity of DP?

Properties of an Optimal Policy $\pi^\star$

- Let Π be the set of all time dependent, history dependent, stochastic policies.

- **Theorem:** Every finite horizon MDP has a deterministic, history-independent optimal policy, that dominates all other policies, everywhere.
  - i.e. there exists a policy $\pi^\star := \left\{\pi_0^\star, \pi_1^\star, \ldots, \pi_{H-1}^\star\right\}, \pi_h^\star : S \mapsto A$ such that

$$V_h^{\pi^\star}(s) \geq V_h^{\pi}(s) \quad \forall s, h, \forall \pi \in \Pi$$

The Bellman Equations

- A function $V = \{V_0, \ldots V_{H-1}\},\ V_h : S \to R$ satisfies the Bellman equations if
$$V_h(s) = \max_a \left\{ r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{h+1}(s') \right] \right\}, \forall s$$
(assume $V_H = 0$).

- **Theorem:**
  - V satisfies the Bellman equations if and only if $V = V^\star$.
  - The optimal policy is: $\pi_h^\star(s) = \arg\max_a \left\{ r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{h+1}^\star(s') \right] \right\}$.

Computation of $V^\star$ with Dynamic Programming

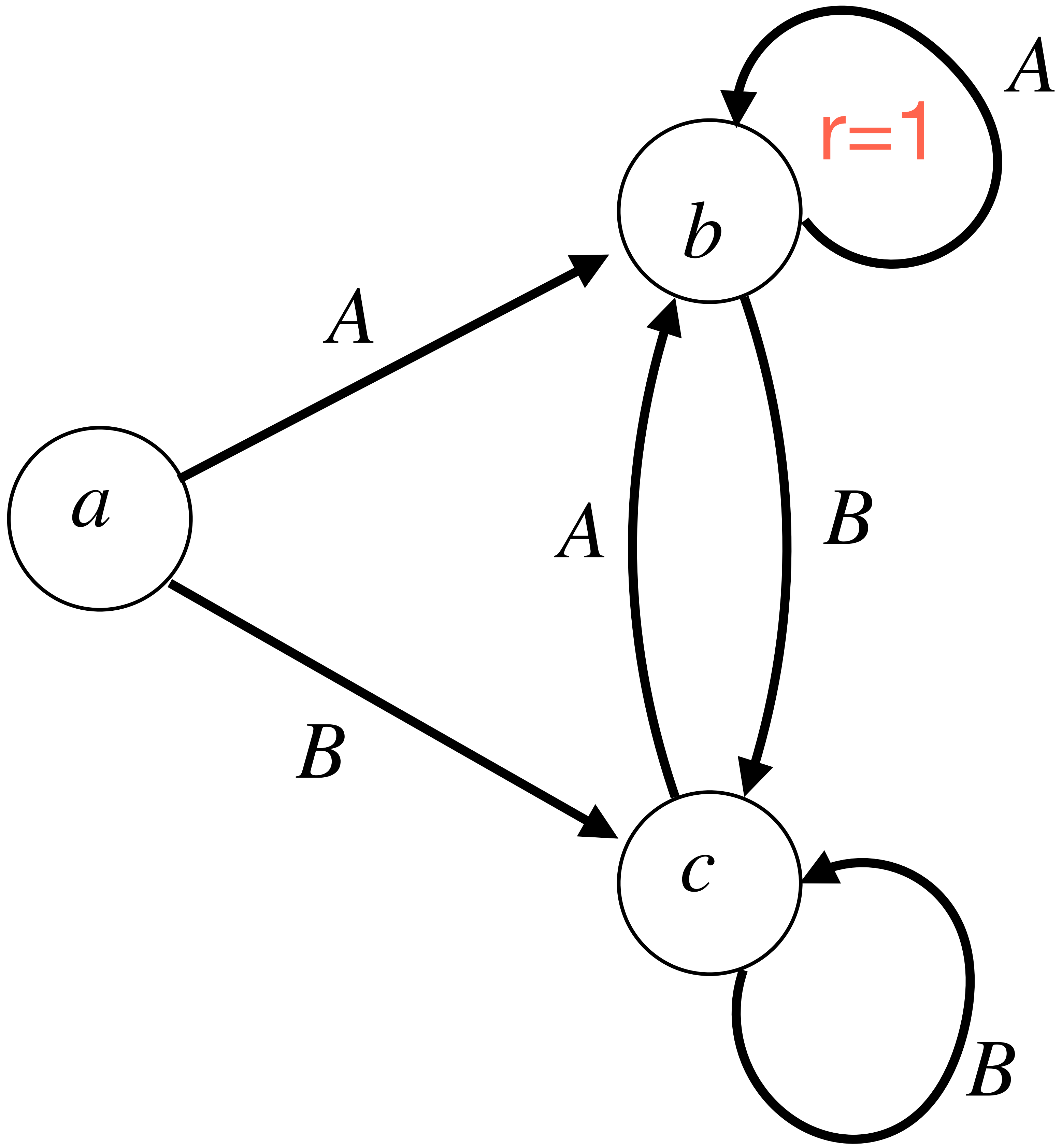
- Theorem: the following Dynamic Programming algorithm correctly computes $\pi^\star$ and $V^\star$
  Prf: the Bellman equations directly lead to this backwards induction.

- Initialize: $V_H^\pi(s) = 0 \ \forall s \in S$
  For t= $H-1, \ldots 0$, set:
  - $V_h^\star(s) = \max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{h+1}^\star(s') \right] \right], \ \forall s \in S$
  - $\pi_h^\star(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_{h+1}^\star(s') \right] \right], \ \forall s \in S$

- What is the per iteration computational complexity of DP?
  (assume scalar $+$, $-$, $\times$, $\div$ are $O(1)$ operations)
- What is the total computational complexity of DP?

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions, with $H = 3$



Reward: $r(b, A) = 1$, & 0 everywhere else

- What's the optimal policy?
  $\pi_h^\star(s) = A, \forall s, h$

- What is optimal value function, $V^{\pi^\star} = V^\star$?

$V_2^\star(a) = 0, V_2^\star(b) = 1, V_2^\star(c) = 0$

$V_1^\star(a) = 1, V_1^\star(b) = 2, V_1^\star(c) = 1$

$V_0^\star(a) = 2, V_0^\star(b) = 3, V_0^\star(c) = 2$

Today:

Today

- Recap
- Infinite Horizon MDPs
  - Policy Evaluation
  - Optimality & the Bellman Equations
  - Value Iteration
  - Policy Iteration

Infinite Horizon MDPs:

- An MDP: $\mathscr{M} = \{\mu, S, A, P, r, \gamma\}$
  - μ, S, A, $P : S \times A \mapsto \Delta(S)$, $r : S \times A \to [0,1]$ same as before
  - instead of finite horizon H, we have a discount factor $\gamma \in [0,1)$

- Objective: find policy π that maximizes our expected, discounted future reward:

$$\max_{\pi} \mathbb{E} \left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \ldots\ldots \middle| \pi \right]$$

The Setting and Our Objective

- Consider a deterministic, stationary policy $\pi : S \mapsto A$
  - stationary means not history or time dependent
- Sampling a trajectory τ on an episode: for a given policy π
  - Sample an initial state $s_0 \sim \mu$:
  - For $t = 0,1,2,\ldots\infty$
    - Take action $a_t = \pi(s_t)$
    - Observe reward $r_t = r(s_t, a_t)$
    - Transition to (and observe) s_{t+1} where $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
- The infinite trajectory: $\tau = \{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, \}$

Today

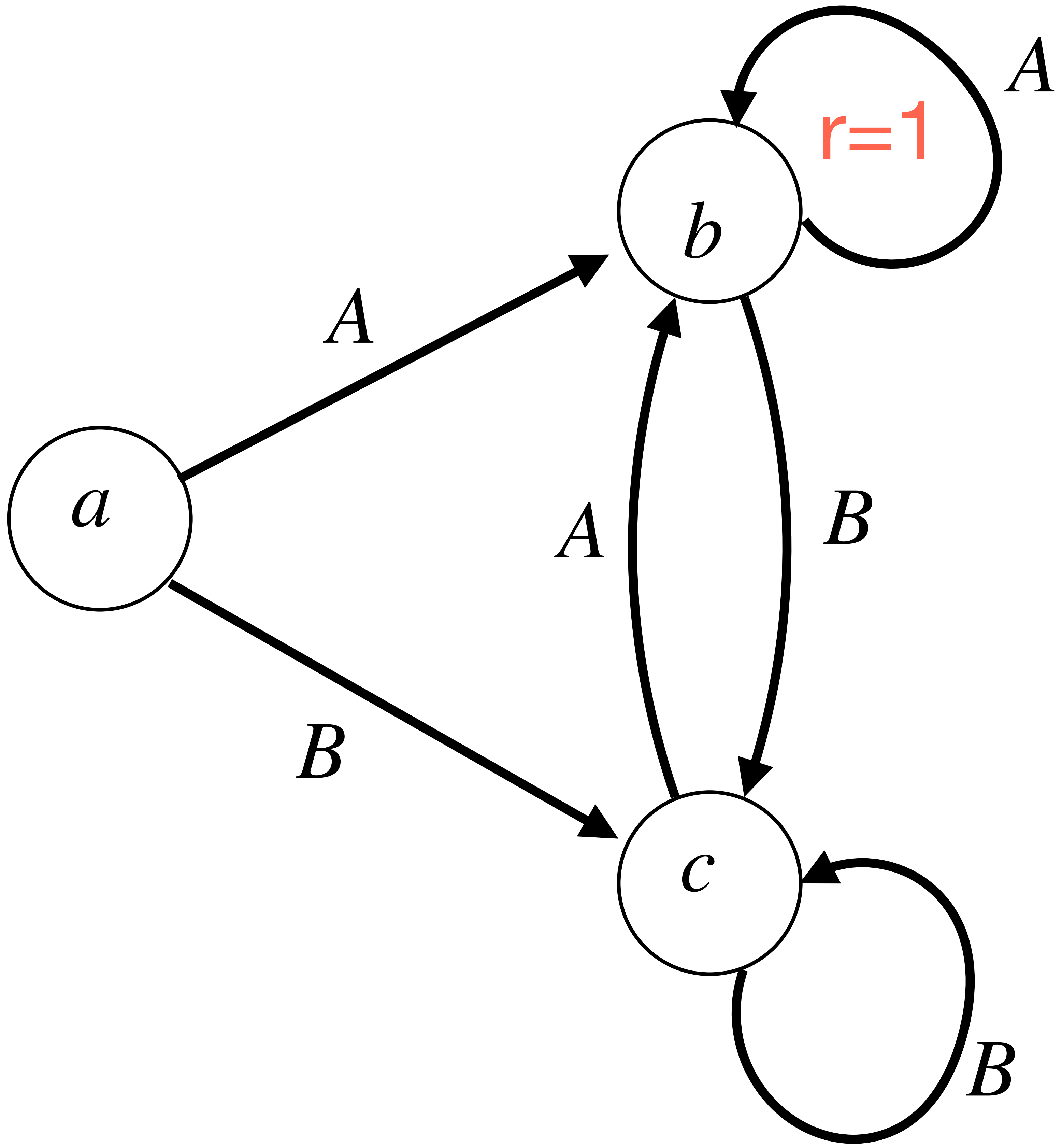
- Recap
- Infinite Horizon MDPs
  - ✓ Policy Evaluation
  - Optimality & the Bellman Equations
  - Value Iteration
  - Policy Iteration

Value function and Q functions:

- Quantities that allow us to reason about the policy's long-term effect:
- Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, \pi\right]$
- Q function $Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), \pi\right]$
- What are upper and lower bounds on V^π and Q^π

Example of Policy Evaluation (e.g. computing V^π and Q^π)

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



Reward: $r(b, A) = 1$, & 0 everywhere else

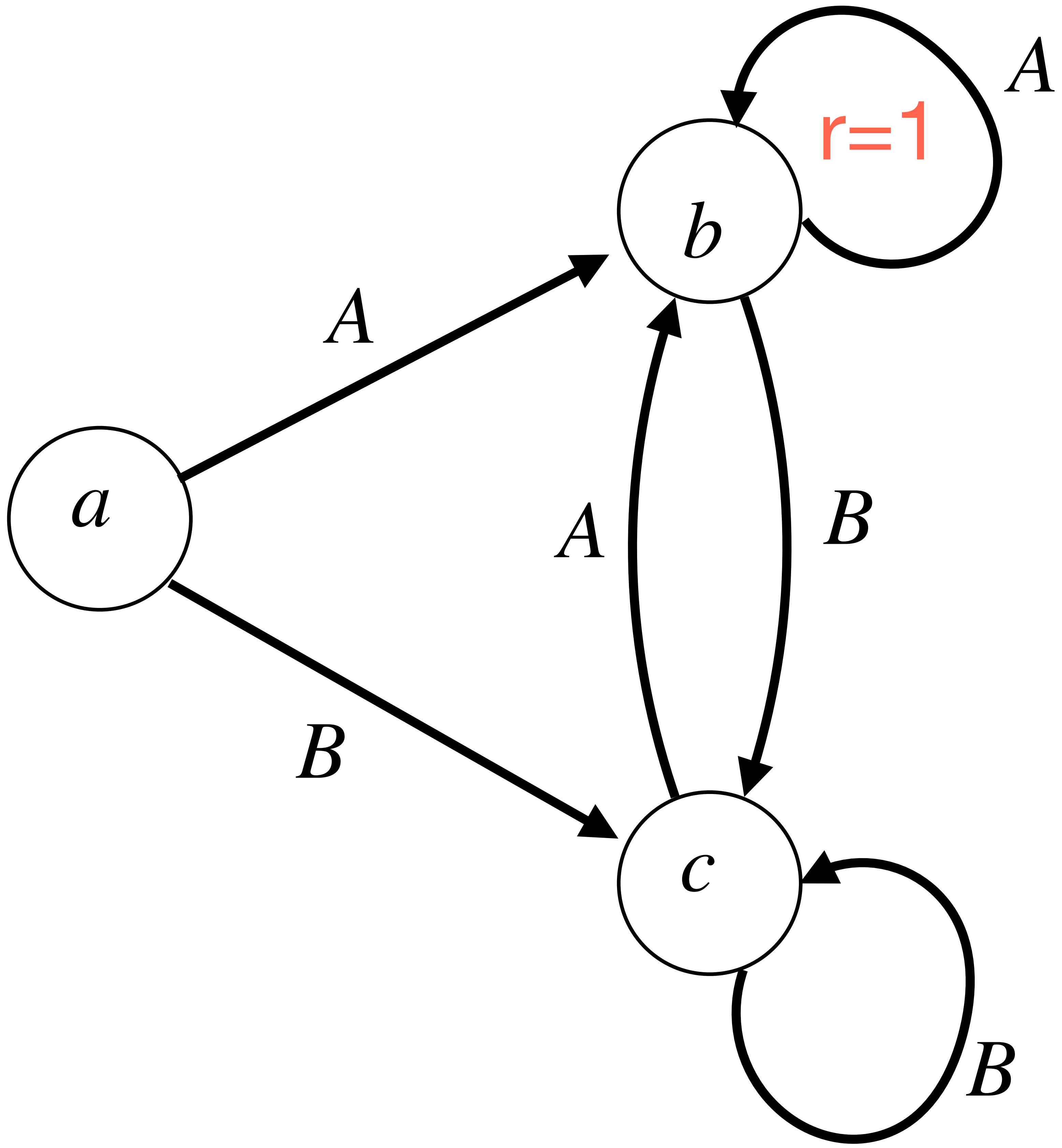
- Consider the policy
  $\pi(a) = B, \pi(b) = A, \pi(c) = A$
- What is V^π?
  $V^\pi(a) =$

  $V^\pi(b) =$

  $V^\pi(c) =$

Example of Policy Evaluation (e.g. computing V^π and Q^π)

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



- Consider the policy
  $\pi(a) = B, \pi(b) = A, \pi(c) = A$
- What is V^π?
  $V^\pi(a) = \gamma^2/(1-\gamma)$

  $V^\pi(b) = 1/(1-\gamma)$

  $V^\pi(c) = \gamma/(1-\gamma)$

Reward: $r(b, A) = 1$, & 0 everywhere else

Bellman Consistency (theorem)

- Consider a fixed policy, $\pi : S \mapsto A$.
- By definition, $V^\pi(s) = Q^\pi(s, \pi(s))$
- Bellman consistency conditions:
  - $V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))}[V^\pi(s')]$
  - $Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^\pi(s')\right]$

(Optional) Proof: Bellman Consistency for V-function:

- By definition and by the "tower" property of conditional expectations:

$$V^\pi(s) = \mathbb{E}\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \ldots \middle| s_0 = s \right]$$

$$= \mathbb{E}\left[r(s_0, a_0) + \mathbb{E}\left[\gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \ldots \middle| s_0 = s, a_0, s_1 \right] \middle| s_0 = s \right]$$

- By the Markov property:

$$= \mathbb{E}\left[r(s_0, a_0) + \gamma \mathbb{E}\left[r(s_1, a_1) + \gamma r(s_2, a_2) + \ldots \middle| s_1 \right] \middle| s_0 = s \right]$$

$$= \mathbb{E}\left[r(s_0, a_0) + \gamma V^\pi(s_1) \middle| s_h = s \right]$$

$$= r(s, \pi(s)) + \gamma \sum_{s'} P(s' | s, \pi(s)) \; V^\pi(s')$$

Computation of V^{π}

- For a fixed policy, $\pi : S \mapsto A$, let's compute its V (and Q) value functions.
- We have the Bellman consistency conditions, for a given policy π

$$V^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V^{\pi}(s')$$

- How do we use this to find a solution?

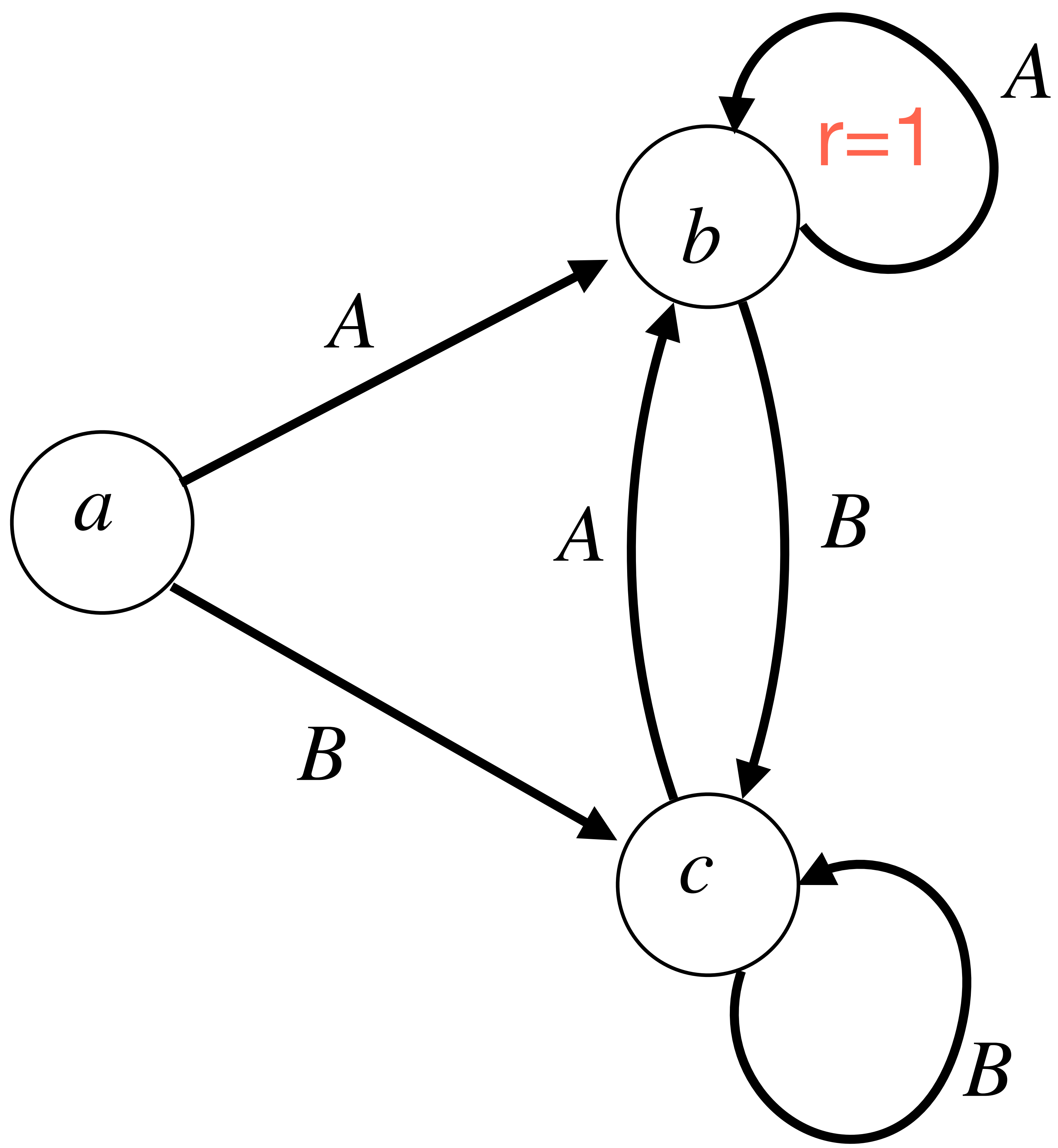
- What is the time complexity?

Today

- Recap
- Infinite Horizon MDPs
  - Policy Evaluation
  - ✓ Optimality & the Bellman Equations
  - Value Iteration
  - Policy Iteration

Example of Optimal Policy $\pi^\star$, discounted case

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



- What's the optimal policy?
  $\pi^\star(s) = A, \forall s$

- What is optimal value function, $V^{\pi^\star} = V^\star$?

$$V^\star(a) = \frac{\gamma}{1-\gamma}, \; V^\star(b) = \frac{1}{1-\gamma}, \; V^\star(c) = \frac{\gamma}{1-\gamma}$$

Reward: $r(b, A) = 1$, & 0 everywhere else

How do we compute $\pi^\star$ and $V^\star$?

- Naively, we could compute the value of all policies and take the best one.
- Suppose $|S|$ states, $|A|$ actions.
  How many different stationary polices are there?

Properties of an Optimal Policy $\pi^\star$

- **Theorem:** Every infinite horizon MDP has a stationary, deterministic optimal policy, that dominates all other policies, everywhere.
  - i.e. there exists a policy $\pi^\star : S \mapsto A$ such that

$$V^{\pi^\star}(s) \geq V^{\pi}(s) \quad \forall s, \; \forall \pi \in \Pi$$

(again Π is the set of all time dependent, history dependent, stochastic policies)

- $\implies$ we can write: $V^\star = V^{\pi^\star}$ and $Q^\star = Q^{\pi^\star}$.

Summary:

- **Discounted infinite horizon MDP:**
  - Concepts: Policy Eval; Bellman equations; Value Iteration

Attendance:
bit.ly/3RcTC9T

Feedback:
bit.ly/3RHtlxy