Infinite Horizon MDPs: Value and Policy Iteration

Lucas Janson

CS/Stat 184(0): Introduction to Reinforcement Learning

Fall 2024

Today

- Recap
- Value Iteration
- Policy Iteration

Infinite Horizon MDPs:

- An MDP: $\mathcal{M} = \{\mu, S, A, P, r, \gamma\}$
  - 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:

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

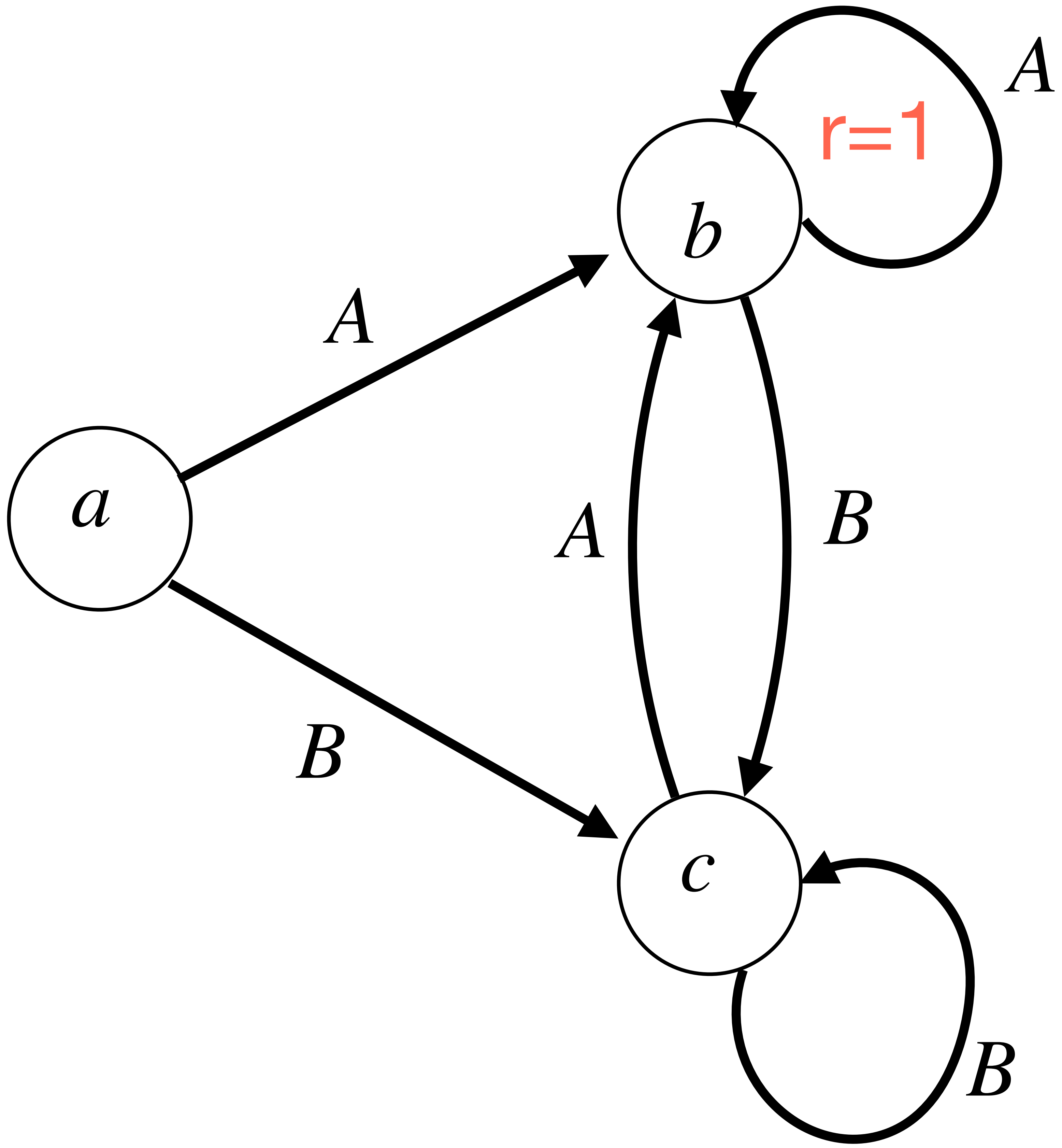
Value function and Q functions:

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

$$0 \le V^\pi(s), Q^\pi(s, a) \le 1/(1 - \gamma)$$

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?
  $V^\pi(a) = \gamma^2/(1-\gamma)$

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

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

Bellman Consistency (theorem)

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

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

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

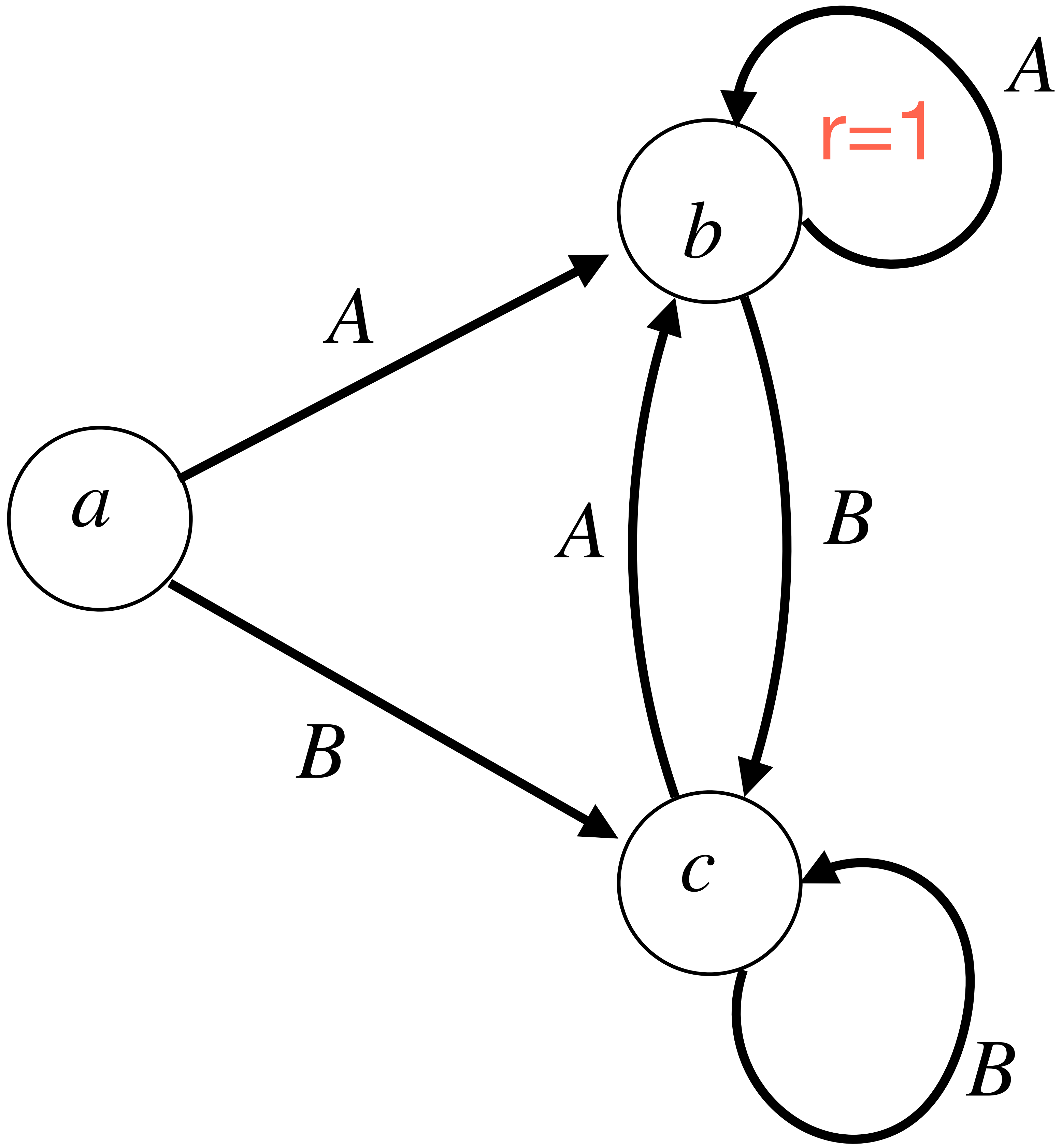
- How do we use this to find a solution?

- What is the time complexity?

- Do you see how to write this with matrix algebra?

Let's use Bellman Consistency for computing V^π

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



$$\pi(a) = B, \quad \pi(b) = \pi(c) = A$$

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

The Bellman Equations

- A function $V : S \to R$ satisfies the Bellman equations if

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

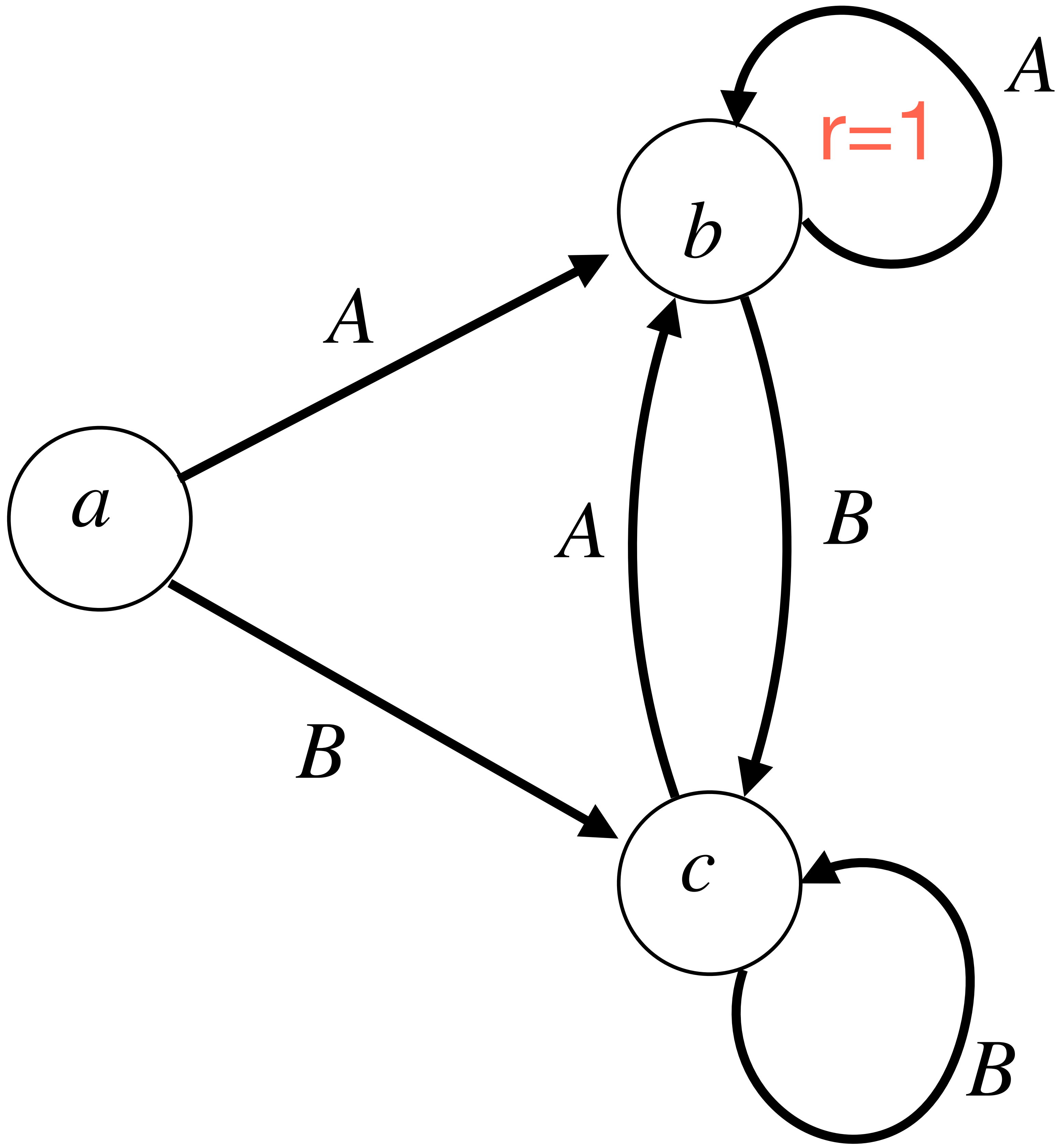
- **Theorem:**
  - V satisfies the Bellman equations if and only if $V = V^\star$.
  - The optimal policy is: $\pi^\star(s) = \arg\max_a \left\{ r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V^\star(s') \right] \right\}$.

Today

- Recap
- Value Iteration
- Policy Iteration

Exercise: use the BE to the purported $\pi^\star$ is optimal

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



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

- What's the optimal policy?
  $\pi^\star(s) = A, \forall s$
- What is optimal value function, $V^{\pi^\star} = V^\star$?

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

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

Detour: fix-point solution

- Suppose we want to find an $x^\star$ s.t. $x^\star = f(x^\star), \quad f: [a, b] \mapsto [a, b]$
- A naive approach to find $x^\star$:
  - Initialize $x^0 \in [a, b]$, repeat: $x^{t+1} = f(x^t)$
- Suppose f is a contraction mapping: $\forall x, x'$, $|f(x) - f(x')| \leq \gamma |x - x'|$, for $\gamma \in [0,1)$. Then it converges, i.e. $x^t \to x^\star$, as $t \to \infty$.
- Observe $|x^t - x^\star| = |f(x^{t-1}) - f(x^\star)| \leq \gamma |x^{t-1} - x^\star|$
- If we want $|x^t - x^\star| \leq \epsilon$, then how should we set t?
  - Want t such that $\gamma^t(b - a) \leq \epsilon$
  - $\implies t \geq -\ln\big(\epsilon/(b-a)\big)/\ln(\gamma)$
  - $\implies t \geq \ln\big((b-a)/\epsilon\big)/(1-\gamma)$ $\qquad$ $[\ln(1+x) \leq x,$ set $x = \gamma - 1]$

Value Iteration Algorithm:

1. Initialization: $V^0(s) = 0, \ \forall s$
2. For $t = 0, \ldots T - 1$

   $V^{t+1}(s) = \max_a \left\{ r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) V^t(s') \right\}, \ \forall s$
3. Return: $V^T(s)$

   $\pi(s) = \arg\max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s,a)} V^T(s') \right\}$

- What is the per iteration computational complexity of VI?
  (assume scalar $+, -, \times, \div$ are $O(1)$ operations)
- Guarantee: VI is fix-point iteration, which contracts, so $V^t \to V^\star$, as $t \to \infty$

Define Bellman Operator $\mathscr{T}$:

- Any function $V : S \mapsto \mathbb{R}$ can also be viewed as a vector in $V \in \mathbb{R}^{|S|}$.
- Define $\mathscr{T} : \mathbb{R}^{|S|} \mapsto \mathbb{R}^{|S|}$, where

$$\left(\mathscr{T}V\right)(s) := \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$$

- Bellman equations: $V = \mathscr{T}V$
- Value iteration: $V^{t+1} \leftarrow \mathscr{T}V^t$

Convergence of Value Iteration:

- The "infinity norm": For any vector $x \in R^d$, define $|x|_\infty = \max_i |x_i|$
- Theorem: Given any V, V', we have: $\|\mathcal{T}V - \mathcal{T}V'\|_\infty \leq \gamma \|V - V'\|_\infty$

$$
\begin{aligned}
|(\mathcal{T}V)(s) - (\mathcal{T}V')(s)| &= \left| \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right\} - \max_a \left\{ r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V'(s') \right\} \right| \\
&\leq \max_a \left| r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V'(s') \right) \right| \\
&= \gamma \max_a \left| \mathbb{E}_{s' \sim P(s,a)} [V(s') - V'(s')] \right| \\
&\leq \gamma \max_a \mathbb{E}_{s' \sim P(s,a)} [|V(s') - V'(s')|] \\
&\leq \gamma \max_s |V(s') - V'(s')| = \gamma \|V - V'\|_\infty
\end{aligned}
$$

- Corollary: If $T = \frac{1}{1-\gamma} \ln\left(\frac{1}{\epsilon(1-\gamma)}\right)$ iterations, VI will return V^T s.t. $\|V^T - V^\star\|_\infty \leq \epsilon$.

  VI then has computational complexity $O\left(|S|^2 |A| T\right)$.

Today

- Recap
- Value Iteration
- Policy Iteration

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots T-1$
  1. **Policy Evaluation**: given π^t, compute $Q^{\pi^t}(s,a)$:
  2. **Policy Improvement**: set $\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$

- What's the computational complexity per iteration?
  Let's do this in parts:
  - Computing V^{π^t}:
  - Computing Q^{π^t} with V^{π^t}:
  - Computing π^{t+1} with Q^{π^t}:

  Per iteration complexity:
- What about convergence?

Convergence of Policy Iteration:

- Theorem: PI has two properties:
  - montone improvement: $V^{\pi^{t+1}}(s) \geq V^{\pi^t}(s)$
  - "contraction": $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \gamma \|V^{\pi^t} - V^\star\|_\infty$

- Corollary: If we set $T = \dfrac{1}{1-\gamma} \ln\left(\dfrac{1}{\epsilon(1-\gamma)}\right)$ iterations,

  PI will return a policy π^{t+1} s.t. $\|V^{\pi^{t+1}} - V^\star\|_\infty \leq \epsilon$

  - with total computational complexity $O\Big(\big(|S|^3 + |S|^2|A|\big)T\Big)$.

Monotonic Improvement Proof

- First, let us show that $\mathcal{T} V^{\pi^t} \geq V^{\pi^t}$.
$$\begin{aligned} \mathcal{T} V^{\pi^t}(s) &= \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\pi^t}(s') \right] \\ &\geq r(s, \pi^t(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^t(s))} V^{\pi^t}(s') \\ &= V^{\pi^t} \end{aligned}$$

- By construction of π^{t+1}:
$$\mathcal{T} V^{\pi^t}(s) = r(s, \pi^{t+1}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\pi^{t+1}(s))} V^{\pi^t}(s')$$

- Using last two claims:
$$\begin{aligned} V^{\pi^{t+1}}(s) - V^{\pi^t}(s) &\geq V^{\pi^{t+1}}(s) - \mathcal{T} V^{\pi^t}(s) \\ &= \gamma \mathbb{E}_{s' \sim P(s,\pi^{t+1}(s))} \left[V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \right] \end{aligned}$$

Today

- Recap
- Value Iteration
- Policy Iteration

Summary:

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

Attendance:
bit.ly/3RcTC9T

Feedback:
bit.ly/3RHtlxy