Optimality in Markov Decision Processes

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- Recap
  - In the bandit setting, we were learning.
  - Now we are starting with computation (of the optimal policy).

- Today:
  - Is there a simple way to characterize the optimal policy?
    - The Bellman Optimality Equations
  - The state-action visitation distribution

Recap

The Objective

- A "stationary" policy $\pi : S \mapsto A$
  - "stationary" means not history dependent
  - we could also consider π to be random and a function of the history
- Sampling a trajectory: from a given policy π starting at state s_0:
  - 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 $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
- Objective: given state starting state s,

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

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

Infinite horizon Discounted Setting

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

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

Policy $\pi : S \mapsto A$

Quantities that allow us to reason 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 $Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| (s_0, a_0) = (s, a), \pi\right]$

Bellman Consistency Equations:

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

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

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

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

Notation

- For a distribution D over a finite set $\mathscr{X}$,

$$E_{x\sim D}[f(x)] = \sum_{x\in\mathscr{X}} D(x)f(x)$$

- $P(\,\cdot\,|\,s,a)$ is a distribution, where $P(s'\,|\,s,a)$ specifies the probability of the transition $(s,a) \to s'$
- We will use notation:

$$E_{s'\sim P(\cdot|s,a)}[f(s')] = \sum_{s'\in S} P(s'\,|\,s,a)f(s')$$

And, if we are short on space and when it is clear, sometimes:

$$E_{s'\sim P(s,a)}[f(s')] = \sum_{s'\in S} P(s'\,|\,s,a)f(s')$$

Today:

Optimality in Markov Decision Processes

Property 1 of an Optimal Policy $\pi^\star$

Even if we consider policies which are randomized and history dependent, the policy which optimizes the the value (starting from any state s) is deterministic and memoryless.

- Defs:
  - "NonStat+Rand": the set of all non-stationary (history dependent), randomized policies.
  - "Stat+Det": the set of all deterministic, stationary (memoryless), policies.
- For any s, we have that:

$$max_{\pi \in \text{NonStat+Rand}} V^\pi(s) = max_{\pi \in \text{Stat+Det}} V^\pi(s)$$

[see theorem 1.7 in AJKS—no need to understand the proof]

- Part of the reason why: the transition function $P(s_{t+1} \mid s_t, a_t)$ is no a function of t.

  So knowledge of s_t implies that using the history doesn't alter the next state distribution.
- (Until we say otherwise) we limit ourselves to only consider det. stationary policies.

Property 2 of an Optimal Policy $\pi^\star$

- The optimal value at state s is defined as:

$$V^\star(s) = \max_\pi V^\pi(s)$$

Note the above permits the optimizing policy to be a function of the starting state s.

- There always exists a deterministic policy $\pi^\star$ such that, for all states s,

$$V^{\pi^\star}(s) = V^\star(s)$$

[see theorem 1.7 in AJKS—no need to understand the proof]

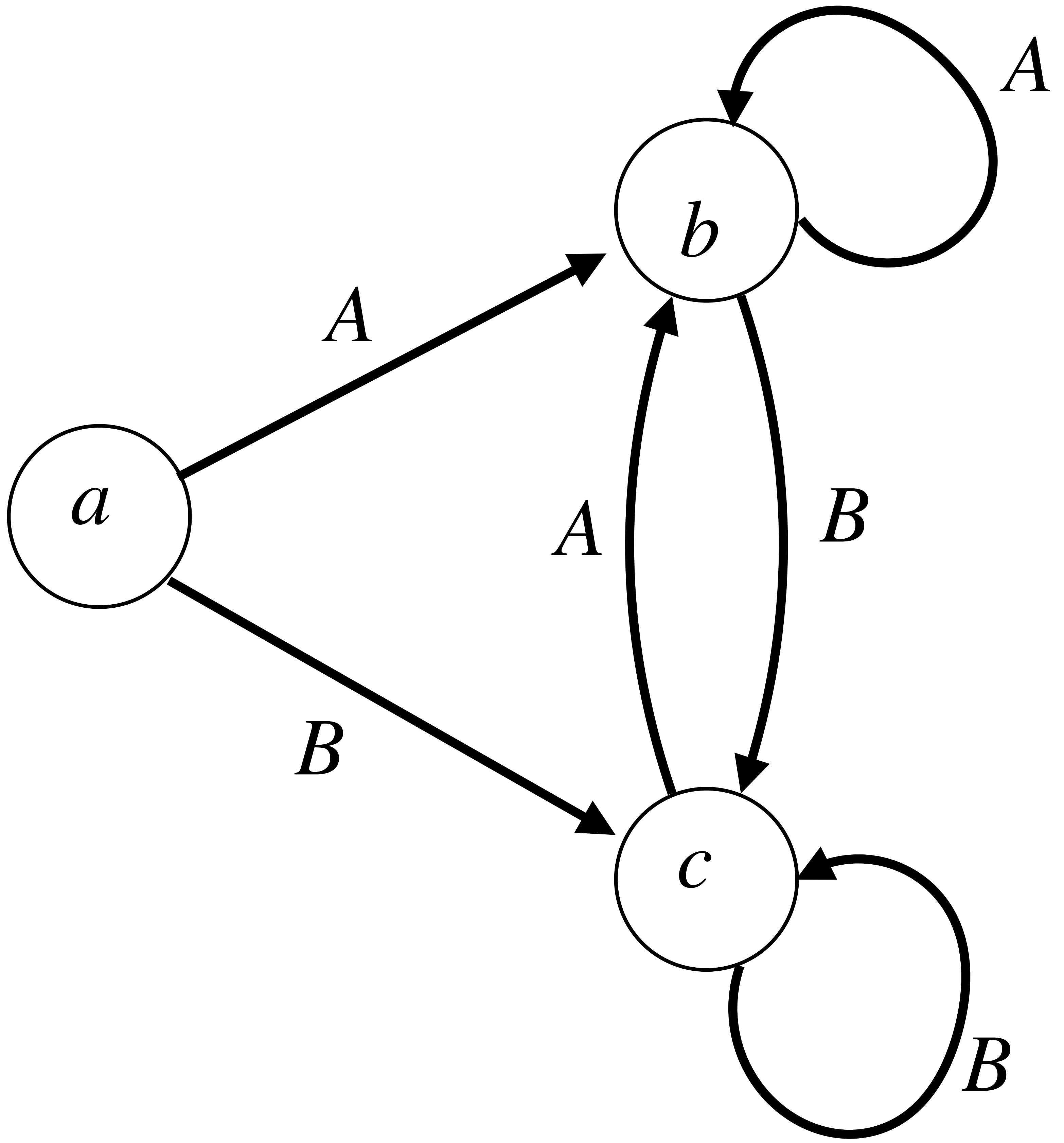
- There is an optimal policy that simultaneously dominates all π, from any starting state.
- Intuition:

$$V^\pi(s) = r(s, \pi(s)) + \gamma E_{s' \sim P(\cdot|s,\pi(s))}[V^\pi(s')]$$
$$\leq r(s, \pi(s)) + \gamma E_{s' \sim P(\cdot|s,\pi(s))}\left[\max_{\tilde{\pi}} V^{\tilde{\pi}}(s')\right]$$

($\Longrightarrow$ after reaching any state s', we can ignore how we got to s' and instead choose the next action at s' to optimize the long term future only as a function of s')

Example of Optimal Policy $\pi^\star$

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



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

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

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

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

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

$$V^\pi(a) = 0, V^\pi(b) = 0, V^\pi(c) = 0$$

Summary so far:

Every discounted MDP has some deterministic optimal policy, that dominates all other policies, everywhere

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

So we have, $V^\star = V^{\pi^\star}$ and $Q^\star = Q^{\pi^\star}$.

Bellman Optimality Equations

Theorem 1: $V^\star$ satisfies the following Bellman Equations:

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

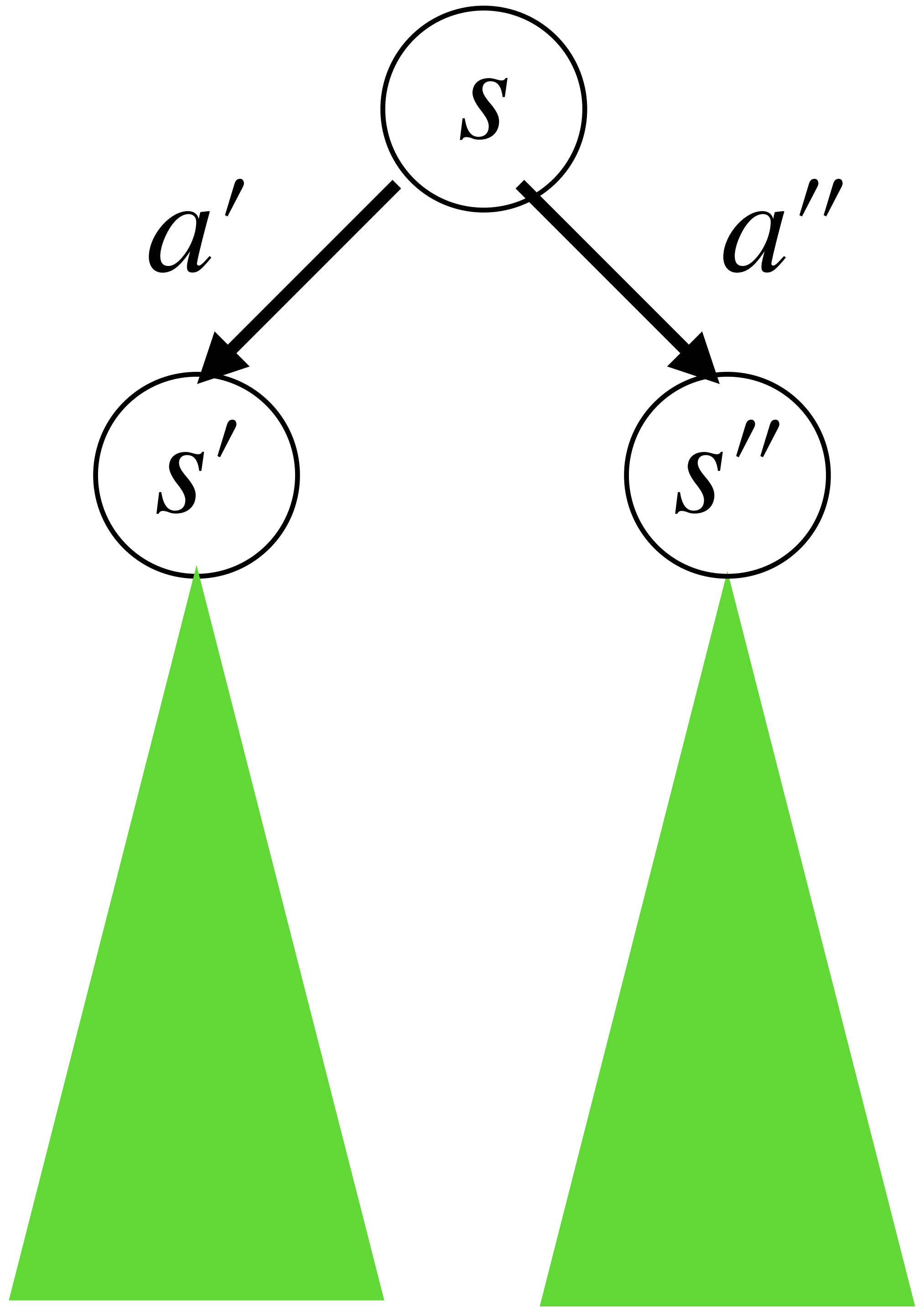
Also, if $\widehat{\pi}(s) = \arg\max_a Q^\star(s,a)$, then $\widehat{\pi}$ is an optimal policy.

Intuition for the Bellman Equations

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

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 the Bellman Equations

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

Proof:

- Denote:
$$\widehat{\pi}(s) := \arg\max_a [r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s')]$$
$$= \arg\max_a Q^\star(s,a)$$
- It suffices to show $V^\star(s) \le V^{\widehat{\pi}}(s)$, which would complete the proof.
- To see this completes the proof,
  - optimality of $V^\star$ implies $V^{\widehat{\pi}}(s) \le V^\star(s)$.
  - and so:
$$V^\star(s) \le V^{\widehat{\pi}}(s) \le V^\star(s).$$
  - Thus $V^{\widehat{\pi}}(s) = V^\star(s)$ and $\hat{\pi}$ is optimal.

Completing the proof: showing $V^\star(s) \leq V^{\widehat{\pi}}(s)$

- Recall: $\widehat{\pi}(s) := \arg\max_a [r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s')]$
- We have:

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

- Proceeding recursively,

$$\begin{aligned}
&\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} V^\star(s'') \right] \\
&\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s,\widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s',\widehat{\pi}(s'))} \left[r(s'', \widehat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'',\widehat{\pi}(s''))} V^\star(s''') \right] \right] \\
&\leq \mathbb{E} \left[r(s, \widehat{\pi}(s)) + \gamma r(s', \widehat{\pi}(s')) + \ldots \mid \widehat{\pi} \right] = V^{\widehat{\pi}}(s)
\end{aligned}$$

Summary so far:

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 that satisfies Bellman Optimality, MUST be equal to $V^\star$

Bellman Equations, Claim 2

Theorem 2: For any $V : S \to \mathbb{R}$, if
$V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right], \forall s,$ then $V = V^\star$.

Bellman Opt allows us to focus on just one step,
i.e., to check if $V = V^\star$, we only need to check if the above equation holds.

Proving Theorem 2

- Define the "maximal component distance" between V and $V^\star$:
$$\|V - V^\star\|_\infty = \max_s |V(s) - V^\star(s)|$$
- For V which satisfies the Bellman equations,
suppose we could show that $\|V - V^\star\|_\infty \le \gamma \|V - V^\star\|_\infty$.
$\Longrightarrow$ the proof is complete because
$$\|V - V^\star\|_\infty \le \gamma \|V - V^\star\|_\infty \le \gamma^2 \|V - V^\star\|_\infty \le \ldots \le \lim_{k\to\infty} \gamma^k \|V - V^\star\|_\infty = 0$$

Proof Continued…

- For V which satisfies the Bellman equations, we want to show $\|V - V^\star\|_\infty \le \gamma \|V - V^\star\|_\infty$.
- Technial observation: $|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$
- Using that V satisfies the Bellman equations, we have, for any s,

$$
\begin{aligned}
|V(s) - V^\star(s)| &= \left| \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right) - \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right) \right| \\
&\le \max_a \left| \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s') \right) - \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right) \right| \\
&= \gamma \max_a \left| \mathbb{E}_{s' \sim P(s,a)} [V(s') - V^\star(s')] \right| \\
&\le \gamma \max_a \mathbb{E}_{s' \sim P(s,a)} |V(s') - V^\star(s')| \\
&\le \gamma \max_a \max_{s'} |V(s') - V^\star(s')| \\
&= \gamma \max_{s'} |V(s') - V^\star(s')| \\
&= \gamma \|V - V^\star\|_\infty
\end{aligned}
$$

Summary Today

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

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

2. If V satisfies Bellman Optimality Equations, $V(s) = \max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then $V = V^\star$.

1-minute feedback form: https://bit.ly/3RHtlxy