Value Iteration

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning
Fall 2022

Today

- Recap
- Today:
  - An Iterative Algorithm: Value Iteration
  - Visitation distributions

Recap

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:

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

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

Summary so far:

Every discounted MDP has some deterministic optimal policy $\pi^\star$,
that dominates all other policies, everywhere

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

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

Bellman (Optimality) Equations

- $V^\star$ satisfies Bellman Equations:

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

- If V satisfies the Bellman Equations,

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

then $V = V^\star$.

- The optimal policy is:

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

Today:

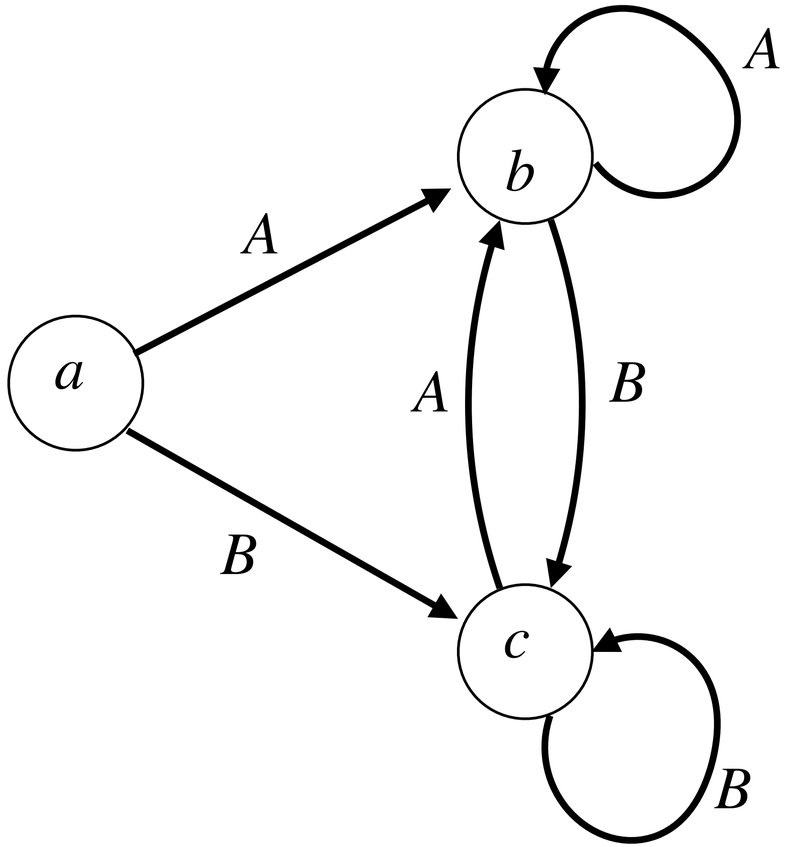
Value Iteration

Question for Today:

Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$,
how can we (approximately) find $\pi^\star$?

Example of Optimal Policy $\pi^\star$

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

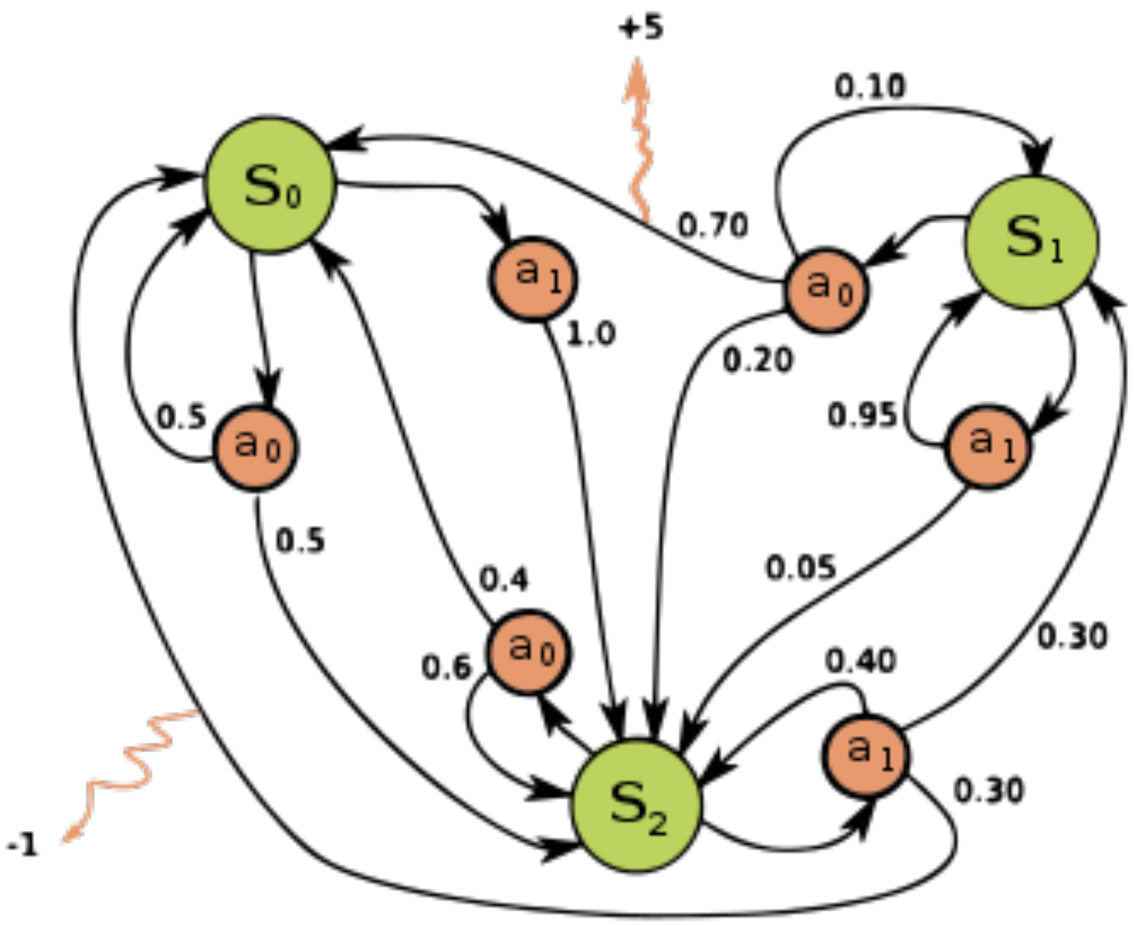


$$\pi^\star(s) = A, \forall s$$

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

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

What about this one…



Let's design an algorithm that computes $V^\star / Q^\star$ for any given MDP

Can we efficiently compute in a optimal policy?

(polynomial in $|S|, |A|$, and other relevant quantities)

- Suppose we can efficiently compute $V^{\pi}(s)$ for any given $\pi : S \mapsto A$.
  - Brute force search would to find $\pi^{\star}$ would still take $|A|^{|S|}$ time.

- Can we construct an iterative algorithm based on the BEs?
  - Will it converge?
  - What is the computation time to get an approximate solution?

Detour: fix-point solution

Consider $x^\star = f(x^\star), \quad f : [a, b] \mapsto [a, b]$

A naive approach to find $x^\star$:

Initialize $x^0 \in [a, b]$, repeat: $x^{t+1} = f(x^t)$

If f is a contraction mapping,
i.e., $\forall x, x', \ |f(x) - f(x')| \leq \gamma |x - x'|$, for some $\gamma \in [0,1)$, then:
$x^t \to x^\star$, as $t \to \infty$

Define Bellman Operator $\mathcal{T}$:

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

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

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

- Bellman equations in terms of $\mathcal{T}$:

$$\mathcal{T} V = V$$

Value Iteration Algorithm:

1. Initialization: $V^0 : \|V^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$

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

Guarantee of VI:

We will see this fix-point iteration converges, i.e., $V^t \to V^\star$, as $t \to \infty$

Alternative Version: Bellman Operator $\mathcal{T}$ on Q

(HW2 Q2 is the Q-version of the Bellman Equations)

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

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

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

What is the Per-Iteration Computational Complexity?

- Making the update $V^{t+1} \leftarrow \mathcal{T} V^t$ explicit:
  - Define Q^{t+1}:

$$\forall s, a \;\; Q^{t+1}(s,a) = r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s, a) V^t(s')$$

  - Set V^{t+1}:

$$\forall s \;\; V^{t+1}(s) = \max_a Q^{t+1}(s,a)$$

- What is the order of the number of basic arithmetic operations?

$$O(|S|^2 |A|)$$

With matrix multiplication?

- Making the update $V^{t+1} \leftarrow \mathcal{T} V^t$ explicit:
  - Define Q^{t+1}:

$$\forall s, a \;\; Q^{t+1}(s,a) = r(s,a) + \gamma \sum_{s' \in S} P(s' \mid s, a) V^t(s')$$

  - Set V^{t+1}:

$$\forall s \;\; V^{t+1}(s) = \max_a Q^{t+1}(s,a)$$

- In terms of matrix multiplication,

let us view r as a vector, $r \in \mathbb{R}^{|S| \cdot |A|}$ and P as a matrix $P \in \mathbb{R}^{|S| \cdot |A| \times |S|}$

$$Q^{t+1} = r + \gamma P V^t$$

Outline:

1: An Iterative Algorithm: Value Iteration
(a fix-point iteration algorithm again)

2: Convergence? How fast?
(Via the **contraction** argument again!)

Convergence of Value Iteration:

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

$$\|\mathscr{T}V - \mathscr{T}V'\|_\infty \leq \gamma \|V - V'\|_\infty$$

Convergence of Value Iteration:

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

$$\|\mathcal{T}V - \mathcal{T}V'\|_\infty \leq \gamma \|V - V'\|_\infty$$

Proof:

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

Convergence of Value Iteration:

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

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

Proof:

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

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

Computational Complexity of VI

VI will return a V^t s.t. $\|V^t - V^\star\|_\infty \leq \epsilon$ in no more than,

$$\frac{\ln(\| V^0 - V^\star \|_\infty / \epsilon)}{\ln(1/\gamma)} \leq \frac{\ln(\| V^0 - V^\star \|_\infty / \epsilon)}{(1-\gamma)} \text{ iterations.}$$

So the computational complexity for an ϵ-accurate solution is

$$O\left(\frac{|S|^2|A|}{1-\gamma} \ln\left(\frac{1}{\epsilon(1-\gamma)}\right)\right)$$

But what about the policy we find with VI?

Theorem: For any V, let $\pi(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then

$$V^{\pi}(s) \geq V^{\star}(s) - \frac{2\gamma}{1-\gamma} \| V - V^{\star} \|_{\infty}$$

But what about the policy we find with VI?

Theorem: For any V, let $\pi(s) = \arg\max_a \left[r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V(s') \right]$, then

$$V^{\pi}(s) \geq V^{\star}(s) - \frac{2\gamma}{1-\gamma} \| V - V^{\star} \|_{\infty}$$

Runtime: After $\frac{\ln\left(2/\left((1-\gamma)^2 \epsilon\right)\right)}{1-\gamma}$ iterations of PI, we have: $V^{\pi^t}(s) \geq V^{\star}(s) - \epsilon$,

and, the total runtime of VI is:

$$O\left(\frac{|S|^2 |A|}{1-\gamma} \ln\left(1/\left((1-\gamma)^2 \epsilon\right)\right) \right)$$

We replace $\epsilon \leftarrow (1-\gamma)\epsilon/2$, then VI will return V^t s.t. $\|V^t - V^{\star}\|_{\infty} \leq (1-\gamma)\epsilon/2$.

Thus, $V^{\pi^t}(s) \geq V^{\star}(s) - \frac{2\gamma}{1-\gamma} \|V^t - V^{\star}\|_{\infty} \geq V^{\star}(s) - \epsilon$

Proof of $\|V^\star - V^\pi\| \leq \frac{2\gamma}{1-\gamma}\|V - V^\star\|_\infty$

Proof:

- To prove the theorem, it suffice to show that:

$$\|V^\pi - V^\star\| \leq \frac{2\gamma}{1-\gamma}\|V - V^\star\|_\infty$$

- We have:

$$\begin{aligned} V^\star(s) - V^\pi(s) &= Q^\star(s, \pi^\star(s)) - Q^\pi(s, \pi(s)) \\ &= \underbrace{Q^\star(s, \pi^\star(s)) - Q^\star(s, \pi(s))}_{\text{Term 1}} + \underbrace{Q^\star(s, \pi(s)) - Q^\pi(s, \pi(s))}_{\text{Term 2}} \end{aligned}$$

- For term 2,

$$\begin{aligned} Q^\star(s, \pi(s)) - Q^\pi(s, \pi(s)) &= \gamma \mathbb{E}_{s' \sim P(s, \pi(s))}\left(V^\star(s') - V^\pi(s')\right) \\ &\leq \gamma \|V^\star - V^\pi\|_\infty \end{aligned}$$

Continuing….

- Let $Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')$, which implies that $Q(s, \pi^\star(s)) \leq Q(s, \pi^t(s))$

- Using this for term 1,

$$
\begin{aligned}
Q^\star(s, \pi^\star(s)) - Q^\star(s, \pi^t(s)) &\leq Q^\star(s, \pi^\star(s)) - Q(s, \pi^\star(s)) + Q(s, \pi(s)) - Q^\star(s, \pi(s)) \\
&\leq \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))}[V^\star(s') - V(s')] - \gamma \mathbb{E}_{s' \sim P(s, \pi(s))}[V(s') - V^\star(s')] \\
&\leq 2\gamma \| V - V^\star \|_\infty
\end{aligned}
$$

- Therefore, we have:
$V^\star(s) - V^{\pi^t}(s) \leq 2\gamma \| V - V^\star \|_\infty + \gamma \| V^\star - V^{\pi^t} \|_\infty$

- which complete the proof because we have shown:
$\| V^\star - V^{\pi^t} \|_\infty \leq 2\gamma \| V - V^\star \|_\infty + \gamma \| V^\star - V^{\pi^t} \|_\infty$

Understanding the sampling

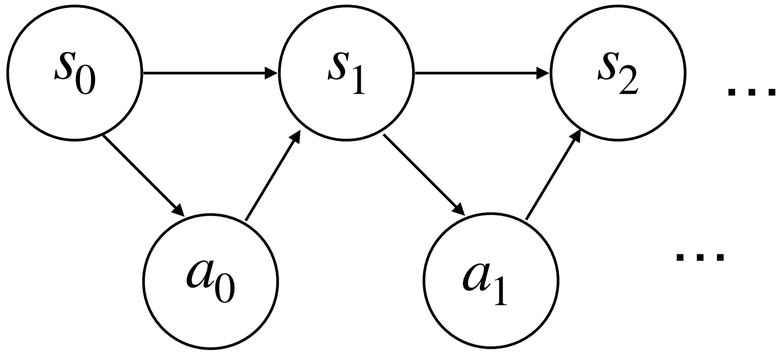
“Occupancy measures” are a helpful concept

Discounted State (action) Occupancy Measures

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

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

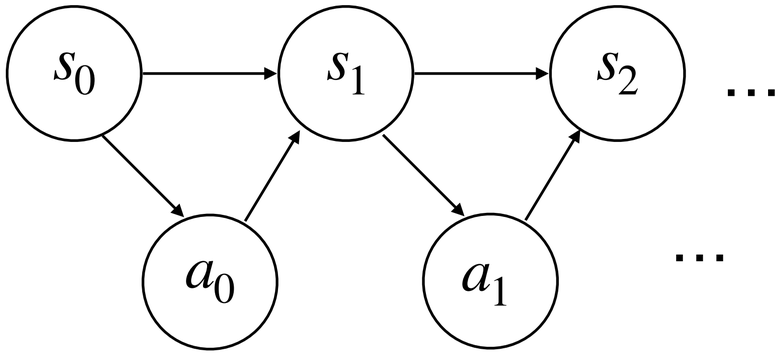
Let's write π as a delta distribution, i.e., $\pi(a \mid s) = \begin{cases} 1, & a = \pi(s), \\ 0, & \text{else} \end{cases}$



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

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

State-action distribution at time step h



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

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

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

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

Discounted Average State-action distribution

Probability of π visiting (s, a) at h, starting from s_0

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

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

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 Today

- Value iteration: an iterative algorithm with a “linear” convergence rate.
- The concept of an “occupancy measure”.

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