Value Iteration : Convergence

Easwar Subramanian

TCS Innovation Labs, Hyderabad

Email : easwar.subramanian@tcs.com / cs5500.2020@iith.ac.in

August 30, 2022

Overview

Review

Policy Evaluation : Prediction

Given a Markov decision process is a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ and a policy π, we have,

- State value function of policy π:

$$V^{\pi}(s) \stackrel{\text{def}}{=} \mathbb{E}_{\pi}(G_t | s_t = s) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right)$$

- Bellman evaluation equation:

$$V^{\pi}(s) \quad = \quad \mathbb{E}_{\pi}\left[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s\right] = \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}^{a}_{ss'}\left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s')\right]$$

- Iterative policy evaluation:

$$V_{k+1}(s) \quad \leftarrow \quad \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}^{a}_{ss'}\left[\mathcal{R}^{a}_{ss'} + \gamma V_k(s')\right]$$

$$V_k(s) \rightarrow V^{\pi}(s)$$

Finding Optimal Policy : Control

- Bellman Optimality Equation

$$V_*(s) \quad = \quad \max_{\pi} \left[\mathbb{E}_{\pi}\left[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s\right]\right] = \max_{a} \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} \left(\mathcal{R}^{a}_{ss'} + \gamma V_*(s')\right)\right]$$

- Value Iteration

$$V_{k+1}(s) \leftarrow \max_{a} \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'} \left(\mathcal{R}^{a}_{ss'} + \gamma V_k(s')\right)\right]$$

$$V_k(s) \rightarrow V_*(s)$$

- Policy Iteration
  - ★ Policy evaluation for policy π_k (k-th iteration)
  - ★ Policy Improvement $\pi_{k+1} \leftarrow \text{greedy}(\pi_k)$

$$\pi_k(s) \rightarrow \pi_*(s)$$

Policy Iteration : Algorithm

Algorithm Policy Iteration

Start with an initial policy π_1
for $i = 1, 2, \cdots, N$ **do**
    Evaluate $V^{\pi_i}(s) \quad \forall s \in \mathcal{S}$. That is,
    for $k = 1, 2, \cdots, K$ **do**
        For all $s \in \mathcal{S}$ calculate

$$V^{\pi_i}_{k+1}(s) \leftarrow \sum_a \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^{\pi_i}_k(s') \right]$$

    end for
    Perform policy Improvement

$$\pi_{i+1} = \text{greedy}(V^{\pi_i})$$

end for

Value Iteration : Algorithm

Algorithm Value Iteration

1: Start with an initial value function $V_1(\cdot)$;
2: **for** $k = 1, 2, \cdots, K$ **do**
3: **for** $s \in \mathcal{S}$ **do**
4: Calculate

$$V_{k+1}(s) \leftarrow \max_a \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V_k(s') \right) \right]$$

5: **end for**
6: **end for**

Value Iteration : Policy Evaluation

Iterative application of Bellman Evaluation Equation

Iterative update rule :

$$V_{k+1}^{\pi}(s) \leftarrow \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}_{ss'}^{a} \left[\mathcal{R}_{ss'}^{a} + \gamma V_{k}^{\pi}(s')\right]$$

The sequence of value functions $\{V_1^{\pi}, V_2^{\pi}, \cdots, \}$ converge to V^{π}

Technical Questions

- How do we know that value iteration converges to V_* ?
- Or that iterative policy evaluation converges to V_π ?
- And therefore that policy iteration converges to π_* ?
- Is the solution unique ?
- How fast do these algorithms converge ? (Depends on discount factor γ)
- These questions were resolved by **Banach Fixed Point Theorem / Contraction Mapping Theorem**

Slide Credit: David Silver's UCL course

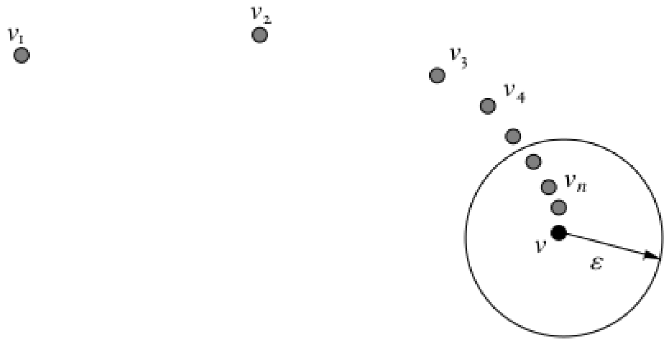
Proof of Value Iteration Convergence

Notion of Convergence

Convergence

Let $\mathcal{V}$ be a vector space. A sequence of vectors $\{v_n\} \in \mathcal{V}$ (with $n \in \mathbb{N}$) is said to converge to v if and only if

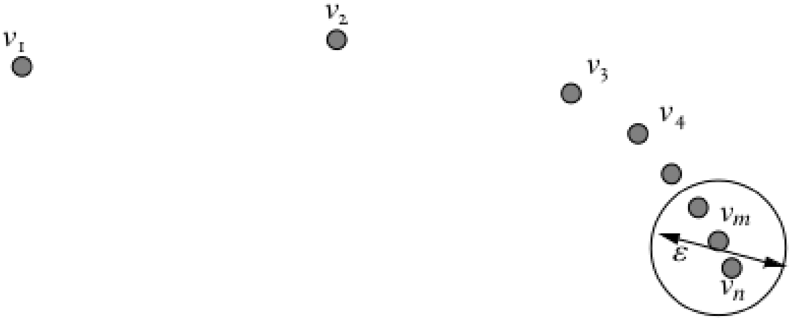
$$\lim_{n \to \infty} \|v_n - v\| = 0$$



Cauchy Sequence

Cauchy Sequence

A sequence of vectors $\{v_n\} \in \mathcal{V}$ (with $n \in \mathbb{N}$) is said to be a Cauchy sequence, if and only if, for each $\varepsilon > 0$, there exists an N_ε such that $\|v_n - v_m\| \leq \varepsilon$ for any $n, m > N_\varepsilon$



Notion of Completeness

Completeness

A **normed vector space** $(\mathcal{V}, \|\cdot\|)$ is complete, if and only if, every Cauchy sequence in $\mathcal{V}$ converges to a point in $\mathcal{V}$

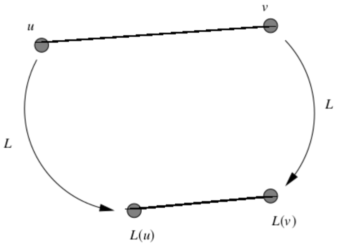
Contractions

Contractions

Let $(\mathcal{V}, \|\cdot\|)$ be a normed vector space and and let $L : \mathcal{V} \to \mathcal{V}$. We say that L is a contraction, or a contraction mapping, if there is a real number $\gamma \in [0, 1)$, such that

$$\big\|L(v) - L(u)\big\| \leq \gamma \|v - u\|$$

for all v and u in $\mathcal{V}$, where the term γ is called a Lipschitz coefficient for L.



Notion of Fixed Point

Fixed Point

A vector $v \in \mathcal{V}$ is a fixed point of the map $L : \mathcal{V} \rightarrow \mathcal{V}$ if $L(v) = v$

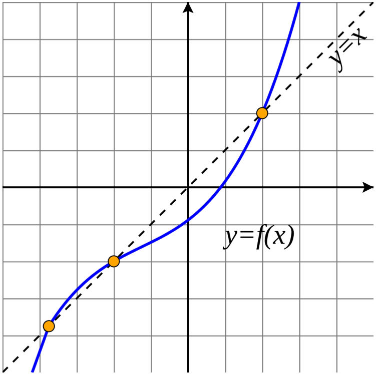
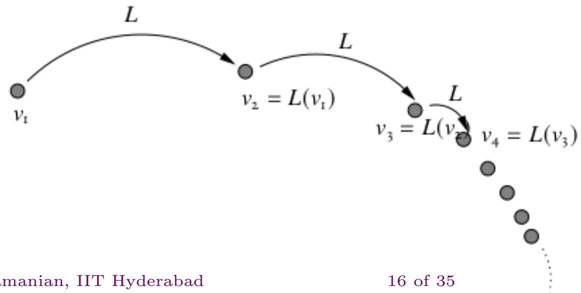


Figure: Fixed Point : Illustration

Banach Fixed Point Theorem

Theorem

Let $<\mathcal{V}, \|\cdot\|) >$ be a complete normed vector space and let $L : \mathcal{V} \to \mathcal{V}$ be a γ-contraction mapping. Then iterative application of L converges to a unique fixed point in $\mathcal{V}$ independent of the starting point



Value Function Space

- $\mathcal{S}$ is a discrete state space with $|\mathcal{S}| = d$
- $\mathcal{A}_s \subseteq \mathcal{A}$ be the non-empty subset of actions allowed from state s
- $\mathcal{V}$ be a vector space of set of all bounded real valued functions from $\mathcal{S}$ to $\mathbb{R}$
- Measure the distance between state value functions $u, v \in \mathcal{V}$ using the max-norm defined as follows

$$\|u - v\| = \|u - v\|_\infty = \max_{s \in \mathcal{S}} \big| u(s) - v(s) \big| \qquad s \in \mathcal{S}; u, v \in \mathcal{V}$$

  ★ Largest distance between state values

- The space $\mathcal{V}$ is complete

Bellman Evaluation Operator

$$V_{k+1}^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}_{ss'}^{a} \left[\mathcal{R}_{ss'}^{a} + \gamma V_{k}^{\pi}(s')\right]$$

Denote,

$$\begin{aligned} \mathcal{P}^{\pi}(s'|s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}_{ss'}^{a} \\ \mathcal{R}^{\pi}(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s'} \mathcal{P}_{ss'}^{a} \mathcal{R}_{ss'}^{a} = \mathbb{E}(r_{t+1}|s_t = s) \end{aligned}$$

Then, we can write,

$$V^{\pi} = \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} V^{\pi} \quad \text{(or)} \ V_{k+1} = \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} V_{k}$$

Define **Bellman Evaluation Operator** $(\mathcal{L}^{\pi} : \mathcal{V} \to \mathcal{V})$ as,

$$L^{\pi}(v) = \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} v$$

Bellman Optimality Operator

$$V_{k+1}(s) = \max_{a} \left[\sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \left(\mathcal{R}_{ss'}^{a} + \gamma V_k(s') \right) \right]$$

Denote,

$$\begin{aligned} \mathcal{P}^a(s'|s) &= \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \\ \mathcal{R}^a(s) &= \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \mathcal{R}_{ss'}^{a} \end{aligned}$$

Then, we can write,

$$V_{k+1} = \max_{a \in \mathcal{A}} \left[\mathcal{R}^a + \gamma \mathcal{P}^a V_k \right]$$

Definte **Bellman Optimality Operator** : $(\mathcal{L} : \mathcal{V} \rightarrow \mathcal{V})$ as

$$L(v) = \max_{a \in \mathcal{A}} \left[\mathcal{R}^a + \gamma \mathcal{P}^a v \right]$$

Remark : Note that since value functions are a mapping from state space to real numbers one can also think of $\mathcal{L}^{\pi}$ and $\mathcal{L}$ as mappings from $\mathbb{R}^d \rightarrow \mathbb{R}^d$

Fixed Points of Maps $\mathcal{L}^\pi$ and $\mathcal{L}$

We can see that V^π is a fixed point of function $\mathcal{L}^\pi$

$$\mathcal{L}^\pi V^\pi = V^\pi$$

and V_* is a fixed point of operator $\mathcal{L}$

$$\mathcal{L} V_* = V_*$$

Bellman Evaluation Operator is a Contraction

Recall that Bellman evaluation operator is given by $L^\pi : \mathcal{V} \to \mathcal{V}$

$$L^\pi(v) = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi v$$

- This operator is γ contraction. i.e., it makes value fuctions closer by at least γ.

Proof.

For any two value functions u and v in the space $\mathcal{V}$, we have,

$$\begin{aligned} \left\|L^\pi(u) - L^\pi(v)\right\|_\infty &= \left\|(\mathcal{R}^\pi + \gamma\mathcal{P}^\pi u) - (\mathcal{R}^\pi + \gamma\mathcal{P}^\pi v)\right\|_\infty \\ &= \left\|\gamma\mathcal{P}^\pi(u - v)\right\|_\infty \left(\leq \gamma\|P^\pi\|_\infty\left\|(u-v)\right\|_\infty = \gamma\left\|(u-v)\right\|_\infty\right) \\ &\leq \left\|\gamma\mathcal{P}^\pi\|u - v\|_\infty\right\|_\infty \\ &\leq \gamma\|u - v\|_\infty \end{aligned}$$

(We used for every $x \in \mathbb{R}^n$, and A is a $m \times n$ matrix, $\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$) $\square$

Convergence of Bellman Updates

- Banach fixed-point theorem guarantees that iteratively applying evaluation operator $\mathcal{L}^\pi$ to any function $V \in \mathcal{V}$ will converge to a unique function $V^\pi \in V$
- Iterative policy evaluation converges to V^π
- Policy iteration converges on V^*
- Similarly, the Bellman optimality operator $(\mathcal{L} : \mathcal{V} \to \mathcal{V})$

  $$L(v) = \max_{a \in \mathcal{A}} [\mathcal{R}^a + \gamma \mathcal{P}^a v] \quad (\text{ A similar argument as } L^\pi)$$

  is also a γ contraction and hence iteratively applying optimality operator $\mathcal{L}$ to any funciton $V \in \mathcal{V}$ will converge to a unique function $V_* \in V$
- Does $V_* = \max_\pi V^\pi(\cdot)$? (Yes, it does)

DP Algorithms : A Closer Look

DP Algorithms : Terminology

Let us consider the policy evaluation formula

$$V_{k+1}(s) \leftarrow \sum_{a} \pi(s,a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V_k(s')\right]$$

DP Algorithms : Terminology

Let us consider the policy evaluation formula

One-step lookahead

$$V_{k+1}(s) \leftarrow \sum_{a} \pi(s,a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V_k(s')\right]$$

DP Algorithms : Terminology

Let us consider the policy evaluation formula

One-step lookahead

$$V_{k+1}(s) \leftarrow \sum_{a} \pi(s,a) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V_k(s') \right]$$

Backup

DP Algorithms : Terminology

Let us consider the policy evaluation formula

One-step lookahead

$$V_{k+1}(s) \leftarrow \sum_a \pi(s,a) \boxed{\sum_{s'}} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V_k(s') \right]$$

Full backup

Backup

DP Algorithms : Terminology

Let us consider the policy evaluation formula

Bootstrap

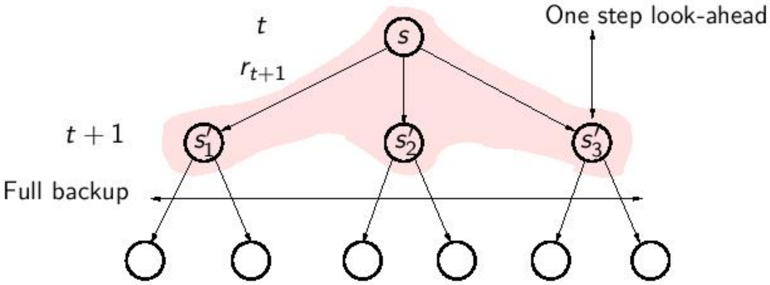
One-step lookahead

$$V_{k+1}(s) \leftarrow \sum_{a} \pi(s,a) \boxed{\sum_{s'}} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V_k(s')\right]$$

Full backup

Backup

DP Algorithms : Schematic View



$$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}(s') \right]$$

$$V_{k+1}(s) \leftarrow \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V_{k}(s') \right]$$

Drawbacks of DP Algorithms

- Requires full prior knowledge of the dynamics of the environment
- DP uses full width back-ups
  - Every successor state and action is considered
- Can be implemented only on small or medium sized discrete state spaces
  - For large problems, DP suffers from Bellman's curse of dimensionality

Model Free Prediction : Key Idea

$$
\begin{aligned}
V^{\pi}(s) &\stackrel{\text{def}}{=} \mathbb{E}_{\pi}(G_t | s_t = s) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right) \\
&= \mathbb{E}_{\pi}\left[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s\right]
\end{aligned}
$$

How can we estimate the expectations?
Use samples!

Appendix

Vector Space

A vector space over a field $\mathcal{F}$ is a set $\mathcal{V}$ together with two operations that satisfy the certain axioms (eight in number)

- **Vector addition** $+ : \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$, takes any two vectors v and w and assigns to them a third vector which is commonly written as $v + w$, and called the sum of these two vectors. (The resultant vector is also an element of the set $\mathcal{V}$ i.e. $v + w \in \mathcal{V}$)
- **Scalar multiplication** $\cdot : \mathcal{F} \times \mathcal{V} \rightarrow \mathcal{V}$ takes any scalar a and any vector v and gives another vector av. (Similarly, the vector av is an element of the set $\mathcal{V}$, i.e. $av \in \mathcal{V}$)

Elements of $\mathcal{V}$ are commonly called vectors; Elements of $\mathcal{F}$ are commonly called scalars.

Norms

Norm assigns a (non-negative) size (or length) to each element of the vector space $\mathcal{V}$

Norm

Given a vector space $\mathcal{V}$, a function $f : \mathcal{V} \to \mathbb{R}^+ \cup \{0\}$ is a norm on the vector space $\mathcal{V}$ if and only if

- **Zero norm** : If $f(v) = 0$ for some $v \in \mathcal{V}$ then, $v = 0$
- **Scalar Multiplication** : For any $\lambda \in \mathbb{R}$ $f(\lambda v) = |\lambda| f(v), \quad \forall v \in \mathcal{V}$
- **Triangle inequality** : For any $v, u \in \mathcal{V}$, we have

$$f(v + u) \leq f(v) + f(u)$$

A **normed vector space** is a pair $(\mathcal{V}, \|\cdot\|)$ where $\mathcal{V}$ is a vector space and $\|\cdot\|$ is a norm on $\mathcal{V}$

Norms : Examples

Let $\boldsymbol{v} = (v_1, v_2, \cdots, v_d)$ be a vector in $\mathcal{V}$

- L_1 or Absolute Value Norm

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{d} |v_i|$$

- L_2 or Euclidean Norm

$$\|\boldsymbol{v}\|_2 = \sqrt{v_1^2 + v^2 + \cdots + v_d^2}$$

- L_p norm

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{d} |v_d|^p \right)^{\frac{1}{p}}$$

- L_∞ or Max Norm

$$\|\boldsymbol{v}\|_\infty = \max_{i \in \{1, \cdots, d\}} |v_i|$$