Value Iteration : Convergence

Easwar Subramanian

TCS Innovation Labs, Hyderabad

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

September 9, 2021

Overview

Review

Optimal Value Functions

- Optimal state-value function $V_*(s)$, for a state s, is the maximum value function over all policies

$$V_*(s) = \max_{\pi} V^{\pi}(s)$$

- Similarly, the optimum action value function $Q_*(s, a)$ is given by

$$Q_*(s, a) = \max_{\pi} Q^{\pi}(s, a)$$

- An Optimal policy $\pi_*(\cdot)$ for an MDP is a policy that is *better than or equal to* all the other policies
  - '*better than* ' – defined using policy evaluation

Prediction and Control using Dynamic Programming

- Dynamic Programming assumes full knowledge of MDP
- Used for both **prediction** and **control** in an MDP
- Prediction
  - ★ Input MDP ($< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$) and policy π
  - ★ Output : $V^{\pi}(\cdot)$
- Control
  - ★ Input MDP ($< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$)
  - ★ Output : Optimal value function $V_*(\cdot)$ or optimal policy π_*

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**

Policy Iteration : Algorithm

Algorithm Policy Iteration

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

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

6: **end for**
7: Perform policy Improvement

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

8: **end for**

Possible Extensions

Synchronous Dynamic Programming

Problem	Bellman Equation	Algorithm
Prediction	Bellman Evaluation Equation	Policy Evaluation
Control	Bellman Evaluation Equation + Greedy Policy Improvement	Policy Iteration
Control	Bellman Optimality Equation	Value Iteration

- All the methods described above have synchronous backups
- All states are backed up in every iteration

Slide Credit: David Silver's UCL course

Asynchronous Dynamic Programming

- Updates to states are done individually, in any order
- For each selected state, apply the appropriate backup
- Can significantly reduce computation
- Convergence guarantees exist, if all states are selected sufficient number of times

Real Time Dynamic Programming

- Idea : update only states that are relevant to agent
- After each time step, we get s_t, a_t, r_{t+1}
- Perform the following update

$$V(s_t) \leftarrow \max_a \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^a_{s_t s'} \left(\mathcal{R}^a_{s_t s'} + \gamma V(s') \right) \right]$$

Concluding Remarks

- Recall that a (stochastic) policy is a distribution over actions given states
- Markov policy means that the policy depends only on the current state and not on the history
- Policies could be stationary or non-stationary
- In general, the optimal policy for an MDP need not be unique
- For finite horizon MDP, the optimal policy need not be even stationary
- For infinite horizon, an MDP admits an optimal policy that is deterministic and stationary. But there could other optimal policies that are stochastic and non-stationary.

Concluding Remarks

- The grid world problem is an example **stochastic shortest path** problem where we consider only policies that are 'proper'
  - ★ A policy that has a non-zero chance to finally reach the terminal state

  Under this assumption the theory on convergence will work out for even $\gamma = 1$.
- The total discounted return G_t could have infinite terms or $\gamma = 1$ but not both

MDP and RL setting

- **MDP Setting** : The agent has knowledge of the state transition matrices $\mathcal{P}^a_{ss'}$ and the reward function $\mathcal{R}$
- **RL Setting** : The agent does not have knowledge of the state transition matrices $\mathcal{P}^a_{ss'}$ and the reward function $\mathcal{R}$
  - ★ The goal in both cases are same; Determine optimal sequence of actions such that the total discounted future reward is maximum.
  - ★ Although, this course would assume Markovian structure to state transitions, in many (sequential) decision making problems we may have to consider the history as well.

Proof of Value Iteration Convergence

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**

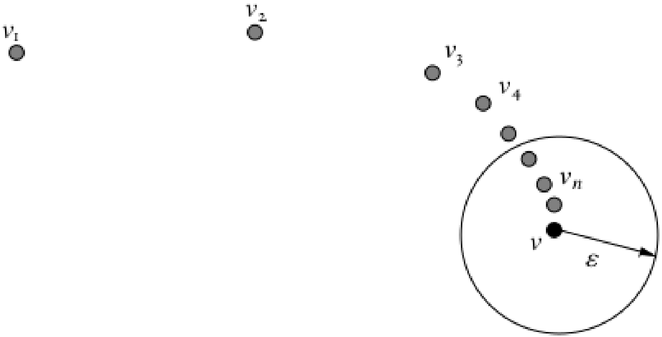
Slide Credit: David Silver's UCL course

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

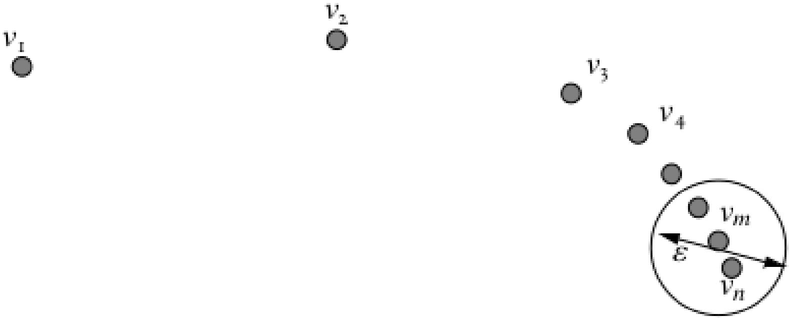
$$\lim_{n\to\infty} \|v_n - v\| = 0$$



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 $\mathcal{V}$

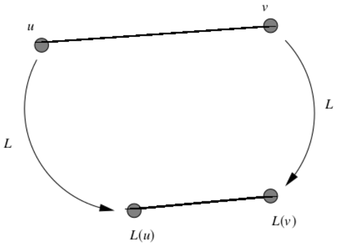
Contractions

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 $L : \mathcal{V} \to \mathcal{V}$ if $L(v) = v$

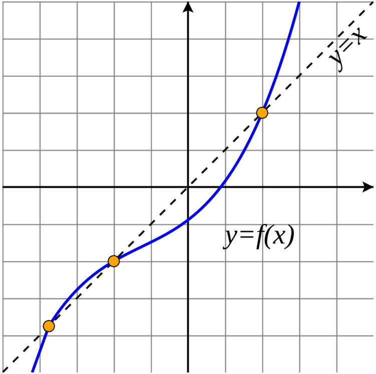
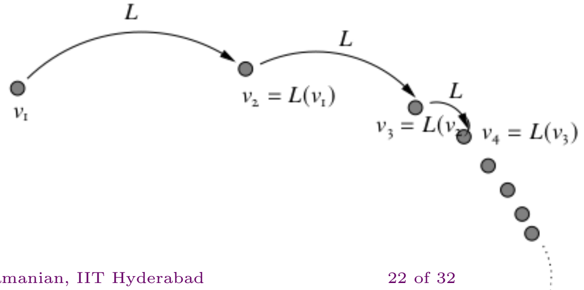


Figure: Fixed Point : Illustration

Banach Fixed Point Theorem

Theorem

Let $< \mathcal{V}, \|\cdot\|) >$ be a complete normed vector space and let $L : \mathcal{V} \rightarrow \mathcal{V}$ 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

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

  ★ 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

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

Denote,

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

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** : $(\mathcal{L} : \mathcal{V} \to \mathcal{V})$ 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 $\mathbb{R}^d \to \mathbb{R}^d$

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)

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** $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$, 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** $\cdot : \mathcal{F} \times \mathcal{V} \to \mathcal{V}$ 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

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