Understanding (Exact) Dynamic Programming through Bellman Operators

Ashwin Rao

ICME, Stanford University

January 14, 2019

Overview

Vector Space of Value Functions

- Assume State pace $\mathcal{S}$ consists of n states: $\{s_1, s_2, \ldots, s_n\}$
- Assume Action space $\mathcal{A}$ consists of m actions $\{a_1, a_2, \ldots, a_m\}$
- This exposition extends easily to continuous state/action spaces too
- We denote a stochastic policy as $\pi(a|s)$ (probability of "a given s")
- Abusing notation, deterministic policy denoted as $\pi(s) = a$
- Consider a n-dim vector space, each dim corresponding to a state in $\mathcal{S}$
- A vector in this space is a specific Value Function (VF) $\mathbf{v}$: $\mathcal{S} \rightarrow \mathbb{R}$
- With coordinates $[\mathbf{v}(s_1), \mathbf{v}(s_2), \ldots, \mathbf{v}(s_n)]$
- Value Function (VF) for a policy π is denoted as $\mathbf{v}_\pi : \mathcal{S} \rightarrow \mathbb{R}$
- Optimal VF denoted as $\mathbf{v}_* : \mathcal{S} \rightarrow \mathbb{R}$ such that for any $s \in \mathcal{S}$,

$$\mathbf{v}_*(s) = \max_\pi \mathbf{v}_\pi(s)$$

Some more notation

- Denote $\mathcal{R}_s^a$ as the Expected Reward upon action a in state s
- Denote $\mathcal{P}_{s,s'}^a$ as the probability of transition $s \rightarrow s'$ upon action a
- Define

$$\mathbf{R}_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \cdot \mathcal{R}_s^a$$

$$\mathbf{P}_\pi(s, s') = \sum_{a \in \mathcal{A}} \pi(a|s) \cdot \mathcal{P}_{s,s'}^a$$

- Denote $\mathbf{R}_\pi$ as the vector $[\mathbf{R}_\pi(s_1), \mathbf{R}_\pi(s_2), \ldots, \mathbf{R}_\pi(s_n)]$
- Denote $\mathbf{P}_\pi$ as the matrix $[\mathbf{P}_\pi(s_i, s_{i'})], 1 \leq i, i' \leq n$
- Denote γ as the MDP discount factor

Bellman Operators $\mathbf{B}_\pi$ and $\mathbf{B}_*$

- We define operators that transform a VF vector to another VF vector
- *Bellman Policy Operator* $\mathbf{B}_\pi$ (for policy π) operating on VF vector $\mathbf{v}$:

$$\mathbf{B}_\pi \mathbf{v} = \mathbf{R}_\pi + \gamma \mathbf{P}_\pi \cdot \mathbf{v}$$

- $\mathbf{B}_\pi$ is a linear operator with fixed point $\mathbf{v}_\pi$, meaning $\mathbf{B}_\pi \mathbf{v}_\pi = \mathbf{v}_\pi$
- *Bellman Optimality Operator* $\mathbf{B}_*$ operating on VF vector $\mathbf{v}$:

$$(\mathbf{B}_* \mathbf{v})(s) = \max_a \{\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a \cdot \mathbf{v}(s')\}$$

- $\mathbf{B}_*$ is a non-linear operator with fixed point $\mathbf{v}_*$, meaning $\mathbf{B}_* \mathbf{v}_* = \mathbf{v}_*$
- Define a function G mapping a VF $\mathbf{v}$ to a deterministic "greedy" policy $G(\mathbf{v})$ as follows:

$$G(\mathbf{v})(s) = \arg\max_a \{\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{s,s'}^a \cdot \mathbf{v}(s')\}$$

- $\mathbf{B}_{G(\mathbf{v})} \mathbf{v} = \mathbf{B}_* \mathbf{v}$ for any VF $\mathbf{v}$ (Policy $G(\mathbf{v})$ achieves the max in $\mathbf{B}_*$)

Contraction and Monotonicity of Operators

- Both $\mathbf{B}_\pi$ and $\mathbf{B}_*$ are γ-contraction operators in L^∞ norm, meaning:
- For any two VFs $\mathbf{v_1}$ and $\mathbf{v_2}$,

$$\|\mathbf{B}_\pi \mathbf{v_1} - \mathbf{B}_\pi \mathbf{v_2}\|_\infty \leq \gamma \|\mathbf{v_1} - \mathbf{v_2}\|_\infty$$

$$\|\mathbf{B}_* \mathbf{v_1} - \mathbf{B}_* \mathbf{v_2}\|_\infty \leq \gamma \|\mathbf{v_1} - \mathbf{v_2}\|_\infty$$

- So we can invoke Contraction Mapping Theorem to claim fixed point
- We use the notation $\mathbf{v_1} \leq \mathbf{v_2}$ for any two VFs $\mathbf{v_1}, \mathbf{v_2}$ to mean:

$$\mathbf{v_1}(s) \leq \mathbf{v_2}(s) \text{ for all } s \in \mathcal{S}$$

- Also, both $\mathbf{B}_\pi$ and $\mathbf{B}_*$ are monotonic, meaning:
- For any two VFs $\mathbf{v_1}$ and $\mathbf{v_2}$,

$$\mathbf{v_1} \leq \mathbf{v_2} \Rightarrow \mathbf{B}_\pi \mathbf{v_1} \leq \mathbf{B}_\pi \mathbf{v_2}$$

$$\mathbf{v_1} \leq \mathbf{v_2} \Rightarrow \mathbf{B}_* \mathbf{v_1} \leq \mathbf{B}_* \mathbf{v_2}$$

Policy Evaluation

- $\mathbf{B}_\pi$ satisfies the conditions of Contraction Mapping Theorem
- $\mathbf{B}_\pi$ has a unique fixed point $\mathbf{v}_\pi$, meaning $\mathbf{B}_\pi \mathbf{v}_\pi = \mathbf{v}_\pi$
- This is a succinct representation of Bellman Expectation Equation
- Starting with any VF $\mathbf{v}$ and repeatedly applying $\mathbf{B}_\pi$, we will reach $\mathbf{v}_\pi$

$$\lim_{N \to \infty} \mathbf{B}_\pi^N \mathbf{v} = \mathbf{v}_\pi \text{ for any VF } \mathbf{v}$$

- This is a succinct representation of the Policy Evaluation Algorithm

Policy Improvement

- Let π_k and $\mathbf{v}_{\pi_\mathbf{k}}$ denote the Policy and the VF for the Policy in iteration k of Policy Iteration
- Policy Improvement Step is: $\pi_{k+1} = G(\mathbf{v}_{\pi_\mathbf{k}})$, i.e. deterministic greedy
- Earlier we argued that $\mathbf{B}_* \mathbf{v} = \mathbf{B}_{G(\mathbf{v})} \mathbf{v}$ for any VF $\mathbf{v}$. Therefore,

$$\mathbf{B}_* \mathbf{v}_{\pi_\mathbf{k}} = \mathbf{B}_{G(\mathbf{v}_{\pi_\mathbf{k}})} \mathbf{v}_{\pi_\mathbf{k}} = \mathbf{B}_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \tag{1}$$

- We also know from operator definitions that $\mathbf{B}_* \mathbf{v} \geq \mathbf{B}_\pi \mathbf{v}$ for all $\pi, \mathbf{v}$

$$\mathbf{B}_* \mathbf{v}_{\pi_\mathbf{k}} \geq \mathbf{B}_{\pi_k} \mathbf{v}_{\pi_\mathbf{k}} = \mathbf{v}_{\pi_\mathbf{k}} \tag{2}$$

- Combining (1) and (2), we get:

$$\mathbf{B}_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \geq \mathbf{v}_{\pi_\mathbf{k}}$$

- Monotonicity of $\mathbf{B}_{\pi_{k+1}}$ implies

$$\mathbf{B}^N_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \geq \ldots \mathbf{B}^2_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \geq \mathbf{B}_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \geq \mathbf{v}_{\pi_\mathbf{k}}$$

$$\mathbf{v}_{\pi_{\mathbf{k+1}}} = \lim_{N \to \infty} \mathbf{B}^N_{\pi_{k+1}} \mathbf{v}_{\pi_\mathbf{k}} \geq \mathbf{v}_{\pi_\mathbf{k}}$$

Policy Iteration

- We have shown that in iteration $k+1$ of Policy Iteration, $\mathbf{v}_{\pi_{\mathbf{k+1}}} \geq \mathbf{v}_{\pi_{\mathbf{k}}}$
- If $\mathbf{v}_{\pi_{\mathbf{k+1}}} = \mathbf{v}_{\pi_{\mathbf{k}}}$, the above inequalities would hold as equalities
- So this would mean $\mathbf{B}_*\mathbf{v}_{\pi_{\mathbf{k}}} = \mathbf{v}_{\pi_{\mathbf{k}}}$
- But $\mathbf{B}_*$ has a unique fixed point $\mathbf{v}_*$
- So this would mean $\mathbf{v}_{\pi_{\mathbf{k}}} = \mathbf{v}_*$
- Thus, at each iteration, Policy Iteration either strictly improves the VF or achieves the optimal VF $\mathbf{v}_*$

Value Iteration

- $\mathbf{B}_*$ satisfies the conditions of Contraction Mapping Theorem
- $\mathbf{B}_*$ has a unique fixed point $\mathbf{v}_*$, meaning $\mathbf{B}_* \mathbf{v}_* = \mathbf{v}_*$
- This is a succinct representation of Bellman Optimality Equation
- Starting with any VF $\mathbf{v}$ and repeatedly applying $\mathbf{B}_*$, we will reach $\mathbf{v}_*$

$$\lim_{N \to \infty} \mathbf{B}_*^N \mathbf{v} = \mathbf{v}_* \text{ for any VF } \mathbf{v}$$

- This is a succinct representation of the Value Iteration Algorithm

Greedy Policy from Optimal VF is an Optimal Policy

- Earlier we argued that $\mathbf{B}_{G(\mathbf{v})}\mathbf{v} = \mathbf{B}_*\mathbf{v}$ for any VF $\mathbf{v}$. Therefore,

$$\mathbf{B}_{G(\mathbf{v}_*)}\mathbf{v}_* = \mathbf{B}_*\mathbf{v}_*$$

- But $\mathbf{v}_*$ is the fixed point of $\mathbf{B}_*$, meaning $\mathbf{B}_*\mathbf{v}_* = \mathbf{v}_*$. Therefore,

$$\mathbf{B}_{G(\mathbf{v}_*)}\mathbf{v}_* = \mathbf{v}_*$$

- But we know that $\mathbf{B}_{G(\mathbf{v}_*)}$ has a unique fixed point $\mathbf{v}_{G(\mathbf{v}_*)}$. Therefore,

$$\mathbf{v}_* = \mathbf{v}_{G(\mathbf{v}_*)}$$

- This says that simply following the deterministic greedy policy $G(\mathbf{v}_*)$ (created from the Optimal VF $\mathbf{v}_*$) in fact achieves the Optimal VF $\mathbf{v}_*$
- In other words, $G(\mathbf{v}_*)$ is an Optimal (Deterministic) Policy