Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 15, 2021

Overview

- The Bellman optimality operator is a contraction
- Bounds on value iteration

Optimality equations

Today, we will see that the nonlinear equations:

$$v(s) = \max_{a \in \mathcal{A}(s)} r(s,a) + \sum_{j \in \mathcal{S}} p(j|s,a)v(j) \quad ,$$

called the **optimality equations** have a unique solution, and that solution coincides with $v_\gamma^\star$ – the value of the MDP. In vector notation:

$$v = \max_{d \in \mathcal{D}^{MD}} r_d + \gamma P_d v = Lv \quad ,$$

where L is the Bellman optimality operator. We will show that:

1. L is a contraction
2. $v_\gamma^\star$ is the unique fixed point of L.

The Bellman optimality operator is a contraction

Theorem If $\gamma \in [0, 1)$, then L is a contraction mapping.

Proof

As usual, we assume that $\mathcal{S}$ is discrete and $L : \mathcal{V} \to \mathcal{V}$. We want to show that there exists a $\lambda \in [0, 1)$ such that $\|Lv - Lu\| \leq \lambda \|v - u\|, \forall u, v \in \mathcal{V}$.

Assume that $Lv(s) \geq Lu(s)$:

$$0 \leq Lv(s) - Lu(s)$$

$$= \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v(j) \right\} - \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) - \gamma \sum_{s \in \mathcal{S}} p(j|s,a) u(j) \right\}$$

Proof

Now let $a_s^\star \in \arg\max_{a\in\mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j\in\mathcal{S}} p(j|s,a)v(j) \right\}$.

$$\leq r(s,a_s^\star) + \gamma \sum_{j\in\mathcal{S}} p(j|s,a_s^\star)v(j) - r(s,a_s^\star) - \gamma \sum_{s\in\mathcal{S}} p(j|s,a)u(j)$$

$$= \gamma \sum_{j\in\mathcal{S}} p(j|s,a_s^\star)\left(v(j) - u(j)\right) \leq \gamma \sum_{j\in\mathcal{A}(s)} p(j|s,a) \underbrace{\|v-u\|}_{\max_{i\in\mathcal{S}}|v_i - u_i|} = \gamma\|v-u\| \ .$$

Therefore:

$$|Lv(s) - Lu(s)| \leq \gamma\|v-u\|$$

$$\Rightarrow \max_{s\in\mathcal{S}} |Lv(s) - Lu(s)| \triangleq \|Lv - Lu\| \leq \gamma\|v-u\| \ .$$

We can repeat the same argument for $Lu(s) \geq Lv(s)$.

Bellman equations: existence of unique solution

Theorem Let $\gamma \in [0, 1)$ with $\mathcal{S}$ finite or countable and bounded reward function:

1. There exists a unique $v^\star \in \mathcal{V}$ such that $Lv^\star = v^\star$.
2. This $v^\star$ is equal to $v^\star_\gamma$, ie: $v^\star = v^\star_\gamma = \max_{\pi \in \Pi^{MR}} v_{\pi,\gamma}$

Proof

- Part 1. follows directly from the fact that L is a γ-contraction under the sup norm.
- Part 2. Doesn't come for free! In fact, we need to first (thm. 6.2.2 in Puterman) that if there exists a $v \in \mathcal{V}$ such that:
  - when $v \geq Lv$ then $v \geq v^\star_\gamma$
  - when $v \leq Lv$ then $v \leq v^\star_\gamma$
  - then if $v = Lv$, this v must be the only element of $\mathcal{V}$ with this property and that $v = v^\star_\gamma$.

Recap

Consequence:

- We now know that the Bellman equations have a unique solution.
- The solution to the Bellman equations gives us the value of the MDP, ie: $v^\star_\gamma$.

What's next:

- How to find optimal policies
- How to find the solution to the Bellman equations numerically.

Optimal policies

So far, we have shown that $v_\gamma^\star$ exists and can be found as the solution to the Bellman equations. What we obtain out of this nonlinear system of equations is $v_\gamma^\star$: not an optimal policy just yet.

In the following, we will show that there exists a **stationary deterministic** optimal policy.

Optimal policies

Theorem Let $\mathcal{S}$ be discrete, and assume that the sup in $\mathcal{L}v = \sup_{d \in \mathcal{D}^{MD}} \{r_d + \gamma P_d v\}$ is attained for all $v \in \mathcal{V}$, then:

1. There exists a conserving decision rule $d^\star \in \mathcal{D}^{MD}$, ie:

$$L_{d^\star} v_\gamma^\star = r_{d^\star} + \gamma P_{d^\star} v_\gamma^\star = v_\gamma^\star \quad .$$

   and the stationary policy $(d^\star)^\infty$ is optimal.
2. $v_\gamma^\star = \sup_{\pi \in \Pi^{MR}} v_{\pi,\gamma} = \sup_{d \in \mathcal{D}^{MD}} v_{d^\infty,\gamma}$

Proof Since $v_\gamma^\star$ is the unique solution of $Lv = v$, then:

$$L_{d^\star} v_\gamma^\star = r_{d^\star} + \gamma P_{d^\star} v_\gamma^\star = v_\gamma^\star = L v_\gamma^\star$$

Proof

By the application of Neumann's lemma for policy evaluation(last lecture), we have that for any $d \in \mathcal{D}^{MD}$, $v_{d^\infty,\gamma}$ is the solution to:

$$v_{d^\infty,\gamma} = L_d v_{d^\infty,\gamma} = r_d + \gamma P_{d^\infty,\gamma} = (I - \gamma P_d)^{-1} r_d \quad .$$

Going back to our theorem:

$$v_\gamma^\star = L v_\gamma^\star = \underbrace{r_{d^\star} + \gamma P_{d^\star} v_\gamma^\star = L_{d^\star} v_\gamma^\star}_{\text{because conserving}} \quad .$$

Therefore:

$$v_\gamma^\star = r_{d^\star} + \gamma P_{d^\star} v_\gamma^\star = v_{(d^\star)^\infty,\gamma} \quad .$$

Important thing to remember

The important consequence of the above is that we can now say that: $v^\star_\gamma = \sup_{\pi \in \Pi^{MR}} v_{\pi,\gamma} = \sup_{d \in \mathcal{D}^{MD}} v_{d^\infty,\gamma}$.
This is a big deal because we went from searching over the space of nonstationary history-dependent randomized policies to only searching over the space of Markov deterministic decision rules, resulting in stationary deterministic Markovian policies.

Practical consequence

If we identified $v^\star_\gamma$, then we can derive an optimal stationary policy $(d^\star)^\infty$ by taking:

$$d^\star(s) \in \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^\star_\gamma(j) \right\}$$

Numerical methods

Value iteration

This algorithm corresponds to the **method of successive approximation**, which comes directly from the constructive proof in Banach fixed point theorem.

Given: $v^{(0)}$, and some tolerance $\epsilon > 0$.

While $\|v^{(k+1)} - v^{(k)}\| \leq \epsilon(1-\gamma)/2\gamma$:

- Compute for each $s \in \mathcal{S}$:
  $v^{(k+1)}(s) = \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^{(k)}(j) \right\}$

Return:

- $d_\epsilon(s) \in \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^{(k+1)}(j) \right\}$

Termination criterion

Theorem Upon termination of value iteration with the above criterion, the last iterate is within $\epsilon/2$ of the optimal value function, ie: $\|v^{(k+1)}) - v_\gamma^\star\| < \epsilon/2$.

Proof

$$\|v_{(d_\epsilon)^\infty,\gamma} - v_\gamma^\star\| \leq \|v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| + \|v^{(k+1)} - v_\gamma^\star\| \quad .$$

Looking at the first term:

$$\begin{aligned}
\|v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| &= \|L_{d_\epsilon} v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| \\
&\leq \|L_{d_\epsilon} v_{(d_\epsilon)^\infty,\gamma} - Lv^{(k+1)}\| + \|Lv^{(k+1)} - v^{(k+1)}\| \\
&= \|L_{d_\epsilon} v_{(d_\epsilon)^\infty,\gamma} - L_{d_\epsilon} v^{(k+1)}\| + \|Lv^{(k+1)} - Lv^{(k+1)}\| \\
&\leq \gamma\|v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| + \gamma\|v^{(k+1)} - v^{(k+1)}\|
\end{aligned}$$

Proof

Re-arranging the terms:

$$\|v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| \leq \frac{\gamma}{1-\gamma}\|v^{(k+1)} - v^{(k)}\|$$

We can also apply the same exact step to the second term and get:

$$\|v^{(k+1)} - v^\star_\gamma\| \leq \frac{\gamma}{1-\gamma}\|v^{(k+1)} - v^{(k)}\|$$

Therefore, since:

$$\|v_{(d_\epsilon)^\infty,\gamma} - v^\star_\gamma\| \leq \|v_{(d_\epsilon)^\infty,\gamma} - v^{(k+1)}\| + \|v^{(k+1)} - v^\star_\gamma\| \quad .$$

then:

$$\|v_{(d_\epsilon)^\infty,\gamma} - v^\star_\gamma\| \leq \frac{\gamma}{1-\gamma}\|v^{(k+1)} - v^{(k)}\| + \frac{\gamma}{1-\gamma}\|v^{(k+1)} - v^{(k)}\| \quad .$$

Proof

By our termination criterion, we have $\|v^{(k+1)} - v^{(k)}\| < \epsilon(1-\gamma)/2\gamma$, therefore:

$$\|v_{(d_\epsilon)^\infty,\gamma} - v^\star_\gamma\| \leq \epsilon \quad .$$

Using Newton’s method

Differentiation as Linearization

Definition (Differentiability). A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be differentiable at $x \in \mathbb{R}^n$ if there exists a linear map $\lambda : \mathbb{R}^n \to \mathbb{R}^m$ such that:

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - \lambda(h)\|}{\|h\|} = 0 \quad ,$$

where $h \in \mathbb{R}^n$.

We can show that if f is differentiable at x, then the linear map λ is unique.

Jacobian

Definition (Jacobian matrix). The Jacobian matrix of $f : \mathbb{R}^n \to \mathbb{R}^m$ at $x \in \mathbb{R}^n$ is the matrix of $Df(x)$ under the standard bases for $\mathbb{R}^n$ and $\mathbb{R}^m$. We denote this matrix by $f'(x) \in \mathbb{R}^{m\times n}$ which we obtain by concatenating the values of $Df(x)(e_i)$, $i = 1, \dots, n$ as columns: $f'(x) \triangleq [Df(x)(e_i), \dots, Df(x)(e_n)]$.

Chain rule

Lemma (Chain rule). If $g : \mathbb{R}^k \to \mathbb{R}^n$ is differentiable at $x \in \mathbb{R}^k$ and $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $g(x)$, then $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at x and

$$D(f \circ g)(x) = Df(g(x)) \circ Dg(x) \quad ,$$

and the matrix of $D(f \circ g)(x)$ is given by:

$$[D(f \circ g)(x)] = [Df(g(x))][Dg(x)] \quad .$$

Directional derivative

Definition (Directional derivative). Let $f : \mathbb{R}^n \to \mathbb{R}$, the directional derivative of f at $x \in \mathbb{R}^n$ in the direction of $v \in \mathbb{R}^n$ is the limit:

$$D_v f(x) \triangleq \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}$$

where $t \in \mathbb{R}$.

If the derivative of f at x exists, the directional directive is given by the value of the linear mapping obtained at this point and evaluated at v, ie: $D_v f(x) = Df(x)(v)$. Using the matrix of $Df(x)$ – the Jacobian – we also have that is given by the Jacobian-vector product $D_v f(x) = [Df(x)]v$ or $v^\top \nabla f(x)$ using the gradient notation.