Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 23, 2021

Policy iteration

- Given: $d^{(0)} \in \mathcal{D}^{MD}$
  - Repeat:
  - **Policy evaluation**: find $v^{(k)}$ by solving for v in $(I - \gamma P_{d^{(k)}})v = r_{d^{(k)}}$
  - **Policy improvement**: choose $d^{(k+1)} \in \arg\max_{d \in \mathcal{D}} \left\{ r_d + \gamma P_d v^{(k)} \right\}$, breaking ties with $d^{(k+1)} = d^{(k)}$ if possible.
  - Terminate if $d^{(k+1)} = d^{(k)}$
- Return the policy $(d^\star)^\infty \triangleq (d^{(k)})^\infty$

In practice: component-wise maximization

Don't forget that when maximizing over the set of deterministic decision rules, this means that in practice we should simply take the maximum over actions in a component-wise fashion, ie: $d^{(k+1)}(s) \in \arg\max_{a\in\mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j\in\mathcal{S}} p(j|s,a) v^{(k)}(j) \right\}$.

Improvement step

Rather than choosing $d^{(k+1)} \in \arg\max_{d \in \mathcal{D}} \left\{ r_d + \gamma P_d v^{(k)} \right\}$, we could also pick any $d^{(k+1)} \in \mathcal{D}^{MD}$ such that $r_{d^{(k+1)}} + \gamma P_{d^{(k+1)}} v^{(k)} \geq r_{d^{(k)}} + \gamma P_{d^{(k)}} v^{(k)}$ with strict inequality in at least one state. This is what Sutton & Barto call *generalized policy iteration*.

While this is true in finite state and action MDPs, this procedure may terminate with suboptimal policies in the general case over compact sets.

Monotonicity

Theorem Let $v^{(k)}$ and $v^{(k+1)}$ be two successive iterates of policy iteration, then $v^{(k+1)} \geq v^{(k)}$.

Proof

In the policy improvement step of policy iteration, we choose the next decision rule as $d^{(k+1)} \in \arg\max_{d \in \mathcal{D}^{MD}} \left\{ r_d + \gamma P_d v^{(k)} \right\}$. Therefore:

$$r_{d^{(k+1)}} + \gamma P_{d^{(k+1)}} v^{(k)} \geq r_{d^{(k)}} + \gamma P_{d^{(k)}} v^{(k)} = v^{(k)} \quad .$$

where the right-hand side follows from the fact that we found $v^{(k)}$ by solving for v in $(I + \gamma P_{d^{(k)}})v = r_{d^{(k)}}$.

Proof

Rearranging the terms in the inequality gives us:

$$r_{d^{(k+1)}} \geq \left(I - \gamma P_{d^{(k+1)}} v^{(k)}\right) v^{(k)} \quad .$$

Multiplying both sides by $\left(I - \gamma P_{d^{(k+1)}} v^{(k)}\right)^{-1}$ gives us:

$$\left(I - \gamma P_{d^{(k+1)}} v^{(k)}\right)^{-1} r_{d^{(k+1)}} = v^{(k+1)} \geq v^{(k)} \quad .$$

Proof

In order to make sure that the order of inequality remains the same in the above proof, we need to show that $\left(I - \gamma P_{d^{(k+1)}} v^{(k)}\right)^{-1}$ is a *positive* operator. That is, $(I - \gamma P_d)^{-1} u \geq 0$ for $u \geq 0, u \in \mathcal{V}, d \in \mathcal{D}^{MR}$, which we write as $(I - \gamma P_d)^{-1} \geq 0$.

Positive operator

Theorem Let $\gamma \in [0, 1), u, v \in \mathcal{V}$, then for any $d \in \mathcal{D}^{MR}$:

1. if $u \geq 0$, then $(I - \gamma P_d)^{-1} u \geq 0$ and $(I - \gamma P_d)^{-1} u \geq u$
2. if $u \geq v$, then $(I - \gamma P_d)^{-1} u \geq (I - \gamma P_d)^{-1} v$
3. if $u \geq 0$, then $u^\top (I - \gamma P_d)^{-1} \geq 0$ and $u^\top (I - \gamma P_d)^{-1} \geq u^\top$

Proof

Because P_d is a stochastic matrix and $\sigma(\gamma P_d) < 1$, $(I - \gamma P_d)^{-1}$ has a Neumann series expansion where each term is positive:

$$(I - \gamma P_d)^{-1} u = u + \gamma P_d u + \gamma^2 P_d u + \ldots \geq u \geq 0 \quad .$$

2 is a subcase of 1 with u set to $u - v$, 3 is obtained from 1 by taking the transpose.

Convergence in the finite state and action case

Theorem Let $\mathcal{S}$ be finite and for each $s \in \mathcal{S}$, $\mathcal{A}(s)$ is finite. Policy iteration terminates in a finite number of iterations and returns a discount optimal policy $(d^\star)^\infty$.

Proof Because of the monotonicity property of the sequence $\{v^{(k)}\}$ and the fact that there is a finite number of deterministic decision rules, policy iteration must terminate in a finite number of steps under the given termination criterion. Because the last iterate satisfies:

$$v^{(k)} = r_{d^{(k+1)}} + \gamma P_{d^{(k+1)}} v^{(k)} = \max_{d \in \mathcal{D}^{MD}} \left\{ r_d + \gamma P_d v^{(k)} \right\} \quad ,$$

$d^{(k)}$ solves the optimality equation and $v_{d^{(k)}} = v^\star_\gamma$.

Newton’s method

Nonlinear system of equations

At a high level, solving nonlinear system of equations entails answering the problem:

$$\begin{aligned} &\text{find } \; x^\star \in \mathbb{R}^n \\ &\text{such that } \; f(x^\star) = 0 \\ &\text{given } \; f : \mathbb{R}^n \to \mathbb{R}^n \end{aligned}$$

Unlike the case of linear equations, nonlinear system of equations rarely admit closed-form solutions

Spivak notation: recap

Let $f : \mathbb{R}^n \to \mathbb{R}^m$

1. $Df(x)$: derivative of f at x (a linear map)
2. $D_i f(x_1, \ldots, x_n), i\{\in 1, \ldots, n\}$: the partial derivative of f with respect to the i-th argument.
   - Eg: $D_1 f(x, y)$: partial derivative of f with respect to x
3. $D_v f(x)$: the directional derivative of f at x in the direction of v (general concept: Gâteaux derivative)

The matrix of Df at x is called the *Jacobian* matrix, which we denote by $f'(x) \in \mathbb{R}^{m \times n}$.

Newton's method

Let $f : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto f(x)$ be a continuously differentiable function of $x \in \mathbb{R}^n$

- Given $x^{(0)} \in \mathbb{R}^n, \epsilon > 0$
- Repeat:
  - Find $\Delta^{(k)}$ by solving for Δ in $\left[Df(x^{(k)})\right] \Delta = f(x^{(k)})$
  - Set $x^{(k+1)} = x^{(k)} - \Delta^{(k)}$
  - Terminate if $\|x^{(k+1)} - x^{(k)}\| \leq \epsilon$
- Return $x^{(k)}$

Taylor approximation

If f is differentiable at $x^{(k)}$ then:

$$f(x^\star) = f(x^{(k)}) + Df(x^{(k)})(x^\star - x^{(k)}) + R(x^\star - x^{(k)}) \quad .$$

where $R(x^\star - x^{(k)})$ is a remainder term such that $\lim_{h \to 0} R(h)/\|h\| = 0$. As $x^{(k)}$ gets close to $x^\star$, the remainder term becomes negligeable and we have: Therefore, we can approximate $\Delta^{(k)} \triangleq x^\star - x^{(k)}$ by solving for Δ in:

$$Df(x^{(k)})\Delta = -f(x^{(k)})$$

and $x^{(k+1)} = x^{(k)} + \Delta^{(k)}$.

Newton Attraction Theorem

Theorem (simplified statement of 10.2.2 in O&R) Let $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ be differentiable in an open neighborhood $S_0 \subset D$ of a point $x^\star \in D$ and that $f(x^\star) = 0$. Furthermore, assume that Df is continuous at $x^\star$ and $Df(x^\star)$ is nonsingular. Then $x^\star$ is a point of attraction for the sequence of iterates:

$$x^{(k+1)} = x^{(k)} - Df(x^{(k)})f(x^{(k)}), \quad k = 0, 1, \ldots$$

An attractive feature of Newton's method is that it can exhibits quadratic convergence, that is we can show that there exists a λ such that: $\|x^{(k+1)} - x^\star\| \leq \lambda \|x^{(k)} - x^\star\|^2$.

Variants

> Newton's method may not be *norm-reducing*, ie it need not be the case that $\|f(x^{(k+1)}) \leq \|f(x^{(k)}\|, \quad k = 0, 1, \ldots$

1. To address this, it is customary to use a *damping* parameter ω_k:

$$x^{(k+1)} = x^{(k)} - \omega_k [Df(x^{(k)})]^{-1} f(x^{(k)}) \quad .$$

2. Furthermore, to ensure that $Df(x^{(k)})$ is nonsingular, we could also use:

$$x^{(k+1)} = x^{(k)} - [Df(x^{(k)}) + \lambda_k I]^{-1} f(x^{(k)}) \quad .$$

where λ_k is a scalar parameter chosen so that the inverse exists.

Variants

3. For computational reason, we could also allow ourselves to use a *stale* derivative information. That is:

$$x^{(k+1)} = x^{(k)} - [Df(x^{p(k)})]^{-1} f(x^{(k)}) \quad ,$$

   where $p(k)$ is an integer less than or equal to k. If $p(k) = k$, then we get back the original Newton's method whereas $p(k) = 0$ gives what Ortega and Rheinboldt call the *simplified Newton method*.

4. Combining the above:

$$x^{(k+1)} = x^{(k)} - \omega_k [Df(x^{p(k)}) + \lambda_k I]^{-1} f(x^{(k)}) \quad ,$$

   with Newton's method corresponding to $\omega_k = 1$, $p(k) = k$, $\lambda_k = 0$.

Solving the optimality equations as root-finding problem

We have seen the optimality equations can be viewed as a fixed point problem of the form $Lv = v$ where L is defined as:

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

Equivalently, the above can be viewed as a **root finding** problem:

$$Lv - v = 0 \quad .$$

Accordingly, we define the operator $Bv \triangleq Lv - v$, or more explicitely:

$$Bv \triangleq \max_{d \in \mathcal{D}^{MD}} \{r_d + (\gamma P_d - I)v\} \quad .$$

Beyond derivatives

The presence of the max operator in the Bellman optimality equation is problematic for a direct application of Newton's method using the usual notion of derivative. While Newton's method has been studied by Kantorovich for the case where $f : D \subset X \to Y$ where X and Y are Banach spaces, this is still not enough for us. The right notion to use is that of so-called *partially ordered topological vector space* (PTL) (Vandergraft, 1967)

The formal treatment of policy iteration as Newton's method under the PTL setting is due Puterman and Brumelle (1979), based on a generalization of Vandergraft (1967) to the nondifferentiable setting in Brumelle and Puterman (1976).

Convex functions

A set $\mathcal{X} \in \mathbb{R}^n$ is *convex* if any two points in $\mathcal{X}$ can be connected by a straight line segment lying entirely inside $\mathcal{X}$, that is:

- Given any $x \in \mathcal{X}$ and $y \in \mathcal{X}$, $\alpha x + (1-\alpha)y \in \mathcal{X}$ for all $\alpha \in [0,1]$.

A function is *convex* if its **domain** is a convex set and if for any two points $x \in \mathcal{X}$ and $y \in \mathcal{X}$:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \quad \forall \alpha \in [0,1]$$

First-order characterization

If a function f is convex and differentiable, then:

$$f(x) + Df(x)(y - x) \leq f(y) \quad ,$$

for all x and y in the domain of f.

This means that for **convex functions**, the first-order Taylor approximation of f is a *global underestimator* of f: ie. its graph is always above all of its tangents.

Support inequality

Let $\mathcal{D}_v^{MD}$ denote the set of v-improving decision rules, ie $d_v \in \mathcal{D}_v^{MD}$ means that:

$$d_v \in \arg\max_{d \in \mathcal{D}^{MD}} \{r_d + (I - \gamma P_d)\, v\}$$

Theorem For any $u, v \in \mathcal{V}$ and $d_v \in \mathcal{D}_v^{MD}$:

$$Bu \geq Bv + (\gamma P_{d_v} - I)(u - v) \quad .$$

Proof

By definition:

$$Bu = \max_{d \in \mathcal{D}^{MD}} \{r_d + (\gamma P_d - I)u\} \geq r_{d_v} + \left(\gamma P_{d_v} - I\right) u$$

Because d_v is v-improving:

$$Bv = r_{d_v} + \left(\gamma P_{d_v} - I\right) v \quad .$$

Therefore:

$$Bu = Bv + (Bu - Bv) \geq Bv + \left(\gamma P_{d_v} - I\right)(u - v)$$

Closed-form expression for policy iteration

Theorem Let $\{v^{(k)}\}$ be the sequence of value functions produced by policy iteration, and $d_{v^{(k)}} \in \mathcal{D}^{MD}_{d_{v^{(k)}}}$

$$v^{(k+1)} = v^{(k)} - (\gamma P_{d_v^{(k)}} - I)^{-1} B v^{(k)} \quad .$$

Proof

Using the closed-form expression for $v_{d_{v^{(k)}}}$:

$$v^{(k+1)} \triangleq v_{d_{v^{(k)}}} = \left(I - \gamma P_{d_{v^{(k)}}}\right)^{-1} r_{d_{v^{(k)}}} \quad .$$

Adding and subtracting:

$$\begin{aligned} v^{(k+1)} &= \left(I - \gamma P_{d_{v^{(k)}}}\right)^{-1} r_{d_{v^{(k)}}} - v^{(k)} + v^{(k)} \\ &= v^{(k)} - \left(\gamma P_{d_{v^{(k)}}} - I\right)^{-1} \left(r_{d_{v^{(k)}}} + \left(\gamma P_{d_{v^{(k)}}} - I\right) v^{(k)}\right) \\ &= v^{(k)} - \left(\gamma P_{d_{v^{(k)}}} - I\right)^{-1} B v^{(k)} \quad . \end{aligned}$$

Differentiable case

What if instead of using the nondifferentiable Bellman equations we would instead use a differentiable approximation?

This is the topic for next class on Wednesday: the smooth Bellman equations and the begining of the section on approximate dynamic programming.