Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 23, 2021

Bellman optimality operator

The Bellman optimality operator that we worked with so far is of the form:

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

Or in component form:

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

Smoothed Bellman operator

The presence of the *max* in L can be problematic because of nthe ondifferentiability.

It is therefore convenient to use the following smoothed operator introduced by Rust (1988):

$$(L_\tau v)(s) \triangleq \tau \log \sum_{a \in \mathcal{A}(s)} \exp\left((1/\tau)\left(r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v(j)\right)\right) \quad .$$

The temperature parameter τ allows us to control the approximation error of L. We can show in fact that $\lim_{\tau \to 0} L_\tau v = Lv$. Furthermore, we will see that $v_\gamma^\star$ can also be recovered as a fixed point, in the limit of $\tau \to 0$.

Approximate operator

Let $L_\theta : \mathcal{V} \to \mathcal{V}$ be an approximate Bellman operator indexed by θ, and $v^\star_{\gamma,\theta}$ the corresponding fixed point.

Theorem (lemma 2.1 of Rust 1994) Let $\{L_\theta\}$ be a family of *contraction mappings* converging pointwise, that is $\lim_{\theta\to\infty} L_\theta v = Lv, \quad \forall v \in \mathcal{V}$, then:

1. $\|v^\star_{\gamma,\theta} - v^\star_\gamma\| \leq \frac{\|L_\theta v^\star_\gamma - L v^\star_\gamma\|}{(1-\gamma)}$
2. $\lim_{\theta\to\infty} \|v^\star_{\gamma,\theta} - v^\star_\gamma\| = 0$

Proof

(Usual trick: add and subtract + triangle inequality)

$$\begin{aligned}\|v^\star_{\gamma,\theta} - v^\star_\gamma\| &= \|L_\theta v^\star_{\gamma,\theta} - L v^\star_\gamma\| \\ &\leq \|L_\theta v^\star_{\gamma,\theta} - L_\theta v^\star_\gamma\| + \|L_\theta v^\star_\gamma - L v^\star_\gamma\| \\ &\leq \gamma\|v^\star_{\gamma,\theta} - v^\star_\gamma\| + \|L_\theta v^\star_\gamma - L v^\star_\gamma\| \quad .\end{aligned}$$

Therefore:

$$\|v^\star_{\gamma,\theta} - v^\star_\gamma\| \leq \frac{\|L_\theta v^\star_\gamma - L v^\star_\gamma\|}{(1-\gamma)} \quad .$$

Application to smoothed Bellman operator

Because $\{L_\tau\}$ is a family of contraction mappings converging pointwise as $\tau \to 0$, then we also have that:

1. $\|v^\star_{\gamma,\tau} - v^\star_\gamma\| \leq \frac{\|L_\tau v^\star_\gamma - L v^\star_\gamma\|}{(1-\gamma)}$
2. $\lim_{\tau \to 0} \|v^\star_{\gamma,\tau} - v^\star_\gamma\| = 0$

Successive Approximation

Because L_τ is a $\gamma - \textit{contraction}$, we can apply the method of successive approximation aka. value iteration as-is.

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

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

- Compute for each $s \in \mathcal{S}$:

$$v^{(k+1)}(s) = \tau \log \sum_{a \in \mathcal{A}(s)} \exp\left((1/\tau)\left(r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^{(k)}(j)\right)\right)$$

Return:

- $$d_\tau(a|s) = \frac{\exp\left((1/\tau)\left(r(s,a)+\gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^{(k)}(j)\right)\right)}{\sum_{a' \in \mathcal{A}(s)} \exp\left((1/\tau)\left(r(s,a')+\gamma \sum_{j \in \mathcal{S}} p(j|s,a') v^{(k)}(j)\right)\right)}$$

Newton-Kantorovich

Rather than solving $L_\tau v = v$, we will instead work with the operator $B_\tau v \triangleq L_\tau v - v$ and solve the nonlinear system of equations $B_\tau v = 0$.

- Given $v^{(0)} \in \mathcal{V}, \epsilon > 0$
- Repeat:
  - Find $\Delta^{(k)}$ by solving for Δ in $B'_\tau(v^{(k)})\Delta = B_\tau(v^{(k)})$
  - Set $v^{(k+1)} = v^{(k)} - \Delta^{(k)}$
  - Terminate if $\|v^{(k+1)} - v^{(k)}\| \leq \epsilon$
- Return $v^{(k)}$

This is essentially "policy iteration" for the smooth Bellman optimality equations.

Explicit update

The main update in the above algorithm reads as:

$$v^{(k+1)} = v^{(k)} - \left(B'_\tau v^{(k)}\right)^{-1} B_\tau v^{(k)} = v^{(k)} - \left(I - L'_\tau v^{(k)}\right)^{-1} (I - L_\tau v^{(k)}) \quad .$$

Here, L' denotes the Gâteaux derivative (G-derivative) of L at $v^{(k)}$.
That is, the linear operator $L'v : \mathcal{V} \to \mathcal{V}$:

$$(L'_\tau v)u \triangleq \lim_{t \to 0} \frac{L_\tau(v + tu) - L_\tau v}{t}$$

Remember that the above reads as $L'_\tau(v)(u)$: a mapping which takes v as input and returns a mapping which we then evaluate at u. $L'_\tau(v)$ is the returned mapping. $L'_\tau(v)(u)$ is the value of the returned mapping evaluated at u.

Derivative of the smoothed Bellman operator

The Gâteaux derivative of L_τ is given by:

$$((L'_\tau v)u)(s) = \gamma \sum_{a \in \mathcal{A}(s)} d_\tau(a|s) \sum_{j \in \mathcal{S}} p(j|s,a) u(s')$$

$$d_\tau(a|s) = \frac{\exp\left((1/\tau)\left(r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v(j)\right)\right)}{\sum_{a' \in \mathcal{A}(s)} \exp\left((1/\tau)\left(r(s,a') + \gamma \sum_{j \in \mathcal{S}} p(j|s,a') v(j)\right)\right)}$$

In physics, d_τ is called the Boltzmann distribution.

Newton-Kantorovich (NK) Theorem

If L_τ has continuous first and second derivatives such that $\|L''_\tau v\| \leq c$ for any $v \in \mathcal{V}$, then given any initial guess $v_0 \in \mathcal{V}$ such that:

$$\|I - L_\tau v_0\| = \eta \leq \frac{(1-\gamma)^2}{2c} \quad ,$$

then:

$$\|v^{(k)} - v^\star_{\gamma,\tau}\| \leq \frac{1}{2^k}\left(\frac{2c\eta}{(1-\gamma)^2}\right)^{2^k} \frac{(1-\gamma)^2}{c} \quad .$$

This implies that:

$$\|v^{(k+1)} - v^\star_{\gamma,\tau}\| \leq c'\|v^{(k)} - v^\star_{\gamma,\tau}\|^2 \quad ,$$

for some c'. Therefore, if v_0 is chosen appropriately, we can achieve a quadratic rate of convergence.

Global convergence

As opposed to PI for discrete state and action spaces and the usual Bellman optimality equations, here we can no longer leverage the argument that the space of decision rules is finite. In fact, our decision rules here are Markov randomized and that set is infinite.

Because of the convergence of NK is local, we can enlarge the region of attraction by first running successive approximation (VI) until we're close enough to the region of attraction, then switch to NK.

Polyalgorithm:

1. Run VI until close enough to region of attraction
2. Switch to NK

Neumann series expansion

Remember that $L'_\tau v$ is a linear operator that belongs in the space of all linear operators from $\mathcal{V}$ to $\mathcal{V}$. Therefore, we can establish existence of an inverse based on its spectral radius.

Theorem Let $0 \leq \gamma < 1$, then $\sigma([L'_\tau v)]) < 1$ and $(I - L'_\tau v)^{-1}$ exists for all $v \in \mathcal{V}$. Furthermore, $(I - L'_\tau v)^{-1} = \sum_{t=0}^{\infty} \left(L'_\tau v\right)^t$

“Modified” Newton-Kantorovich

We can develop a “modifed” counterpart to the above procedure that mimics “modifed policy iteration” where we only take a few terms in the Neumann series expansion.

- Given $v^{(0)} \in \mathcal{V}, \epsilon > 0$, truncation n, initial guess $\tilde{\Delta}_0$
- Repeat:
  - Set $\tilde{\Delta}^{(1)} = \tilde{\Delta}_0$
  - Repeat from $i = 1, \ldots, n-1$
    - $\tilde{\Delta}^{(i+1)} = B_\tau v^{(k)} + (L'_\tau v^{(k)})\tilde{\Delta}^{(i)}$.
  - Let $\Delta^{(k)} = \tilde{\Delta}^{(n)}$
  - Set $v^{(k+1)} = v^{(k)} - \Delta^{(k)}$
  - Terminate if $\|v^{(k+1)} - v^{(k)}\| \leq \epsilon$
- Return $v^{(k)}$

Important pratical implication

Remember: $L'_\tau v^{(k)})$ is the Gâteaux derivative. Therefore, computing $(L'_\tau v^{(k)})\tilde{\Delta}^{(i)}$ in practice amounts to evaluating a Jacobian-vector product (jvp) JVPs and VJPs are the fundamental building blocks of automatic differentiation (AD) in machine learning. They allows to implement so-called "matrix-free" algorithms.

Using JAX for example, you could compute $(L'_\tau v^{(k)})\tilde{\Delta}^{(i)}$ with `jax.linearize` without forming any Jacobian matrix.

Linear programming formulation

Theorem 6.2.2a showed that if $v \geq Lv$ for some $v \in \mathcal{V}$ then it must be that $v \geq v^\star_\gamma$: ie v is an upper bound on $v^\star_\gamma$.

Definition $v \in \mathcal{V}$ is said to be γ-superharmonic if:

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

Theorem $v^\star_\gamma$ is the smallest γ-superharmonic vector

Proof $v^\star_\gamma$ satisfies the Bellman optimality equations:

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

Let $a^\star_s$ be a maximizer of the above. Therefore:

$$v^\star_\gamma(s) = r(s, a^\star_s) + \gamma \sum_{j \in \mathcal{S}} p(j|s, a^\star_s) v(j)$$

Proof

Let $v \in \mathcal{V}$ be γ-superharmonic. Then:

$$v \geq r_d + \gamma P_d v, \quad \forall d \in \mathcal{D}^{MD}$$

(just the definition in vector form) Re-arranging:

$$(I - \gamma P_d) v \geq r_d \quad .$$

Because $(I - \gamma P_d)^{-1}$ is a positive operator (see last lecture), then:

$$v \geq (I - \gamma P_d)^{-1} r_d \triangleq v_{d^\infty}, \quad \forall d \in \mathcal{D}^{MD} \quad .$$

Therefore, $v \geq v_\pi$ for any $\pi \in \Pi^{MD}$ (including optimal policies), and $v \geq v_\gamma^\star = \max_{\pi \in \Pi^{MD}} v_\pi$

LP formulation

Therefore, in the set of all superharmonic vectors, $v_\gamma^\star$ is the smallest.

This is the basis for our LP formulation: let's search the set of γ-superharmonic vectors and find the smallest.

Primal LP

Let $\alpha \in \mathbb{R}^{|\mathcal{S}|}$, $\sum_{i \in \mathcal{S}} \alpha(i) = 1$,

$$\begin{aligned} \text{minimize} \quad & \sum_{i \in \mathcal{S}} \alpha(i) v(i) \\ \text{subject to} \quad & v(s) - \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v(j) \geq r(s,a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}(s) \end{aligned}$$

Dual LP

$$\text{maximize} \quad \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}(s)} r(s,a) x(s,a)$$

$$\text{subject to} \quad \sum_{a \in \mathcal{A}(j)} x(s,a) - \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}(s)} p(j|s,a) x(s,a) = \alpha(j)$$

Policies

Theorem (6.9.1 in Puterman) Let $d \in \mathcal{D}^{MR}$ and for any $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$, and define:

$$x_d(s,a) \triangleq \sum_{i \in \mathcal{S}} \alpha(i) \sum_{t=1}^{\infty} \gamma^{(t-1)} P_{d^\infty}(S_t = s, A_t = a | S_1 = i)$$

1. x_d is a feasible solution to the dual problem
2. we can derive a stationary policy d_x^∞ from x with:

$$d_x^\infty(a|s) = \frac{x(s,a)}{\sum_{a' \in \mathcal{A}(s)} x(s,a')}$$

Interpretation

$x(s, a)$ is the total discounted probability joint probability that starting from the initial state α, the system is in state s taking action a.