MARKOV DYNAMIC PROGRAMMING

Quantitative Economics 2023

Piotr Żoch

December 4, 2023

A TYPICAL PROBLEM

- The planner chooses a path of actions $(A_t)_{t \geq 0}$ to maximize

$$\mathbb{E} \sum_{t=0}^{\infty} \beta^t r(X_t, A_t)$$

  where $(X_t)_{t \geq 0}$ is state process.

- X is a finite set: state space.
- A is a finite set: action space.
- Γ is a correspondence from X to A. Intuitively: the set of actions feasible given the state.

MDP

- Given A and X a finite Markov decision process (MDP) is a tuple $\mathscr{M} = (\Gamma, P, r, \beta)$ where
    1. $\Gamma : X \to A$ is a nonempty correspondence from X to A defining feasible state-action pairs

$$G := \{(x, a) \in X \times A : a \in \Gamma(x)\}$$

    2. a stochastic kernel P from G to X:

$$\sum_{x' \in X} P\left(x, a, x'\right) = 1 \text{ for all } (x, a) \in G.$$

    3. a function r from G to $\mathbb{R}$ is a reward function
    4. $\beta \in (0, 1)$ is a discount factor.

MDP

- The Bellman equation associated with $\mathscr{M}$ is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x' \in \mathsf{X}} P\left(x, a, x'\right) v\left(x'\right) \right\} \text{ for all } x \in \mathsf{X}.$$

- This is an equation in the unknown function $v \in \mathbb{R}^{\mathsf{X}}$.
- We will show that the solution to the Bellman equation equals to the largest possible value of the objective function in the sequence problem:

$$\max \mathbb{E} \sum_{t=0}^{\infty} \beta^t r(X_t, A_t), \quad \text{subject to } A_t \in \Gamma(X_t) \text{ for all } t \geq 0.$$

POLICIES

- Let Σ be the set of all feasible policies given $\mathscr{M}$:

$$\Sigma := \left\{ \sigma \in \mathsf{A}^{\mathsf{X}} : \sigma(x) \in \Gamma(x) \text{ for all } x \in \mathsf{X} \right\}.$$

- For any $\sigma \in \Sigma$ we have P_σ is a stochastic kernel from X to X:

$$P_\sigma\left(x, x'\right) := P\left(x, \sigma(x), x'\right) \text{ for all } \left(x, x'\right) \in \mathsf{X} \quad \text{so } P_\sigma \in \mathcal{M}\left(\mathbb{R}^{\mathsf{X}}\right).$$

- Similarly, for any $\sigma \in \Sigma$ we have r_σ, a function from X to $\mathbb{R}$:

$$r_\sigma(x) := r(x, \sigma(x)) \text{ for all } x \in \mathsf{X} \quad \text{so } r_\sigma \in \mathbb{R}^{\mathsf{X}}.$$

POLICIES

- Define $\mathbb{E}_{x_0}[\cdot] := \mathbb{E}[\cdot \mid X_0 = x_0]$. The lifetime value of following $\sigma \in \Sigma$ from x is

$$v_\sigma(x) := \mathbb{E}_x \left[\sum_{t=0}^{\infty} \beta^t r_\sigma(X_t) \right]$$

where X_t is P_σ-Markov with $X_0 = x$.

- Since $\beta \in (0, 1)$, we can calculate

$$v_\sigma(x) = \sum_{t=0}^{\infty} \beta^t P_\sigma^t r_\sigma = (I - \beta P_\sigma)^{-1} r_\sigma.$$

POLICY OPERATOR

- Define the policy operator T_σ:

$$(T_\sigma v)(x) := r(x, \sigma(x)) + \beta \sum_{x' \in \mathsf{X}} v(x) P\big(x, \sigma(x), x'\big) \text{ for all } x \in \mathsf{X}.$$

- We denote a fixed point of T_σ by v_σ.
- We will now prove T_σ is a contraction of modulus β on $\mathbb{R}^{\mathsf{X}}$ under norm $\|\cdot\|_\infty$.
- We will also show that T_σ is order-preserving: if $v \leq w$ then $T_\sigma v \leq T_\sigma w$.

POLICY OPERATOR

- Take any $v, w \in \mathbb{R}^{\mathsf{X}}$ and $\sigma \in \Sigma$.
- Fix $x \in \mathsf{X}$. We have

$$
\begin{aligned}
\left|(T_\sigma v)(x) - (T_\sigma w)(x)\right| &= \beta \left| \sum_{x' \in \mathsf{X}} \left(v(x') - w(x')\right) P(x, \sigma(x), x') \right| \\
&\leq \beta \sum_{x' \in \mathsf{X}} \left|v(x') - w(x')\right| P(x, \sigma(x), x') \\
&\leq \beta \left\|v - w\right\|_\infty
\end{aligned}
$$

- Since it is true regardless of x, we have

$$
\left\|T_\sigma v - T_\sigma w\right\|_\infty \leq \beta \left\|v - w\right\|_\infty .
$$

POLICY OPERATOR

- To show that it is order preserving take any $v, w \in \mathbb{R}^X$ and $\sigma \in \Sigma$.
- $v \leq w$ implies $P_\sigma v \leq P_\sigma w$. We can write

$$Tv = r_\sigma + \beta P_\sigma v \text{ and } Tw = r_\sigma + \beta P_\sigma w.$$

so $Tv \leq Tw$.

GREEDY POLICIES

- Given MDP $\mathscr{M}$ the value function is

$$v^*(x) := \max_{\sigma \in \Sigma} v_\sigma(x) \text{ for all } x \in \mathsf{X}.$$

- We call a policy $\sigma \in \Sigma$ optimal if $v_\sigma = v^*$.
- We call a policy v-greedy if

$$\sigma(x) \in \operatorname*{argmax}_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x, a, x') \right\} \text{ for all } x \in \mathsf{X}.$$

BELLMAN

- We say that Bellman's principle of optimality holds for MDP $\mathscr{M}$ if

$$\sigma \in \Sigma \text{ is optimal for } \mathscr{M} \iff \sigma \text{ is } v^*\text{-greedy.}$$

- The Bellman operator corresponding to $\mathscr{M}$ is a self-map T on $\mathbb{R}^{\mathsf{X}}$ defined by

$$Tv(x) := \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x' \in \mathsf{X}} v(x') P(x, a, x') \right\} \text{ for all } x \in \mathsf{X}.$$

OPTIMALITY

Theorem

Let $\mathscr{M}$ be an MDP with Bellman operator T. Then

1. *v^* is the unique solution to the Bellman equation $v = Tv$ in $\mathbb{R}^X$,*
2. $\lim_{k\to\infty} T^k v = v^*$ *for all* $v \in \mathbb{R}^X$,
3. *Bellman's principle of optimality holds for $\mathscr{M}$,*
4. *at least one optimal policy exists.*

OPTIMALITY

- Instead of solving the (possibly hard) sequence problem we can solve the (possibly easier) functional equation $v = Tv$.
- Finding v-greedy policies is easier than looking at the entire set of feasible policies Σ.
- The required conditions are pretty weak. Important and somewhat hidden: sets are finite and $r : \mathsf{G} \to \mathbb{R}$.

OPTIMALITY

- We will prove (1) and (2).
- Two parts of the proof:
    1. Show there exists the unique fixed point of T.
    2. Show that the fixed point is v^*.

OPTIMALITY

- Fix v, w in $\mathbb{R}^X$. We have

$$
\begin{aligned}
\left|(Tv)(x) - (Tw)(x)\right| &= \left|\max_{\sigma \in \Sigma} (T_\sigma v)(x) - \max_{\sigma \in \Sigma} (T_\sigma w)(x)\right| \\
&\leq \max_{\sigma \in \Sigma} \left|(T_\sigma v)(x) - (T_\sigma w)(x)\right| \\
&= \|T_\sigma v - T_\sigma w\|_\infty
\end{aligned}
$$

- We have $\|Tv - Tw\|_\infty \leq \|T_\sigma v - T_\sigma w\|_\infty$ for all $\sigma \in \Sigma$.
- We showed earlier that T_σ is a contraction: $\|T_\sigma v - T_\sigma w\|_\infty \leq \beta \|v - w\|_\infty$.
- We thus have

$$\|Tv - Tw\|_\infty \leq \beta \|v - w\|_\infty \text{ for all } v, w \in \mathbb{R}^X.$$

OPTIMALITY

- By the Banach fixed point theorem T has a unique fixed point $\bar{v}$.
- We will now show that $\bar{v} = v^*$.
- Pick $\sigma \in \Sigma$ that is $\bar{v}$-greedy. By definition we have $T_\sigma \bar{v} = \bar{v} = T\bar{v}$. So $\bar{v}$ is a fixed point of T_σ. Because we defined v^* as $\max_{\sigma \in \Sigma} v_\sigma$ We have $\bar{v} \leq v^*$.
- Pick any $\sigma \in \Sigma$, We must have $T_\sigma v \leq Tv$ for any v. We know that T_σ is order preserving, so it must be that $v_\sigma \leq \bar{v}$. This is true for any σ, so $v^* \leq \bar{v}$.

OPTIMALITY

- We can use T_σ to look for the value function (instead of value function iteration).
- Start with a guess v_0, find a greedy policy σ_0 and calculate the fixed point of T_{σ_0}:

$$v_{\sigma_0} = \left(I - \beta P_{\sigma_0}\right)^{-1} r_{\sigma_0}.$$

- Repeat the process with v_{σ_0} – find a greedy policy and calculate the new fixed point.
- Do it until convergence.
- This algorithm is known as policy iteration or Howard's policy iteration

HPI

Algorithm Howard's Policy Iteration

procedure HPI
    $k \leftarrow 1, \epsilon \leftarrow \tau + 1, v_k \leftarrow v_{\text{init}}$
    while $\epsilon > \tau$ **do**
        $\sigma_k \leftarrow v_k$–greedy policy
        $v_{k+1} = \left(I - \beta P_{\sigma_k}\right)^{-1} r_{\sigma_k}$
        $\epsilon \leftarrow \|v_{k+1} - v_k\|_\infty, k \leftarrow k + 1$
    end while
end procedure

EXAMPLE

- HPI converges at a faster rate than VFI.
- In a finite state setting, the algorithm always converges to an exact optimal policy in a finite number of steps, regardless of the initial condition.
- Drawback: computing v_σ can be expensive.

OPTIMISTIC POLICY ITERATION

- This is a variant of HPI.
- Key difference: do not compute v_σ exactly.
- Instead, apply the policy operator T_σ to v_k for a fixed number of iterations, m.
- For $m \to \infty$ we have HPI; for $m = 1$ we have VFI.
- Often outperforms HPI and VFI, but this requires choosing m.

OPI

Algorithm Optimistic Policy Iteration

procedure OPI
  $k \leftarrow 1, \epsilon \leftarrow \tau + 1, v_k \leftarrow v_{\text{init}}$
  while $\epsilon > \tau$ **do**
    $\sigma_k \leftarrow v_k$–greedy policy
    $v_{k+1} = T_{\sigma_k}^m v_k$
    $\epsilon \leftarrow \|v_{k+1} - v_k\|_\infty, k \leftarrow k + 1$
  end while
end procedure