

# MARKOV DYNAMIC PROGRAMMING

---

QUANTITATIVE ECONOMICS 2024

Piotr Żoch

January 10, 2025

## A TYPICAL PROBLEM

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

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t r(X_t, A_t)$$

where  $(X_t)_{t \geq 0}$  is a state process ( $X_0$  is given).

- $X$  is a finite set: **state space**.
- $A$  is a finite set: **action space**.
- $\Gamma$  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  $\mathcal{M} = (\Gamma, P, r, \beta)$  where
  - $\Gamma : X \rightarrow A$  is a nonempty correspondence from  $X$  to  $A$  defining feasible state-action pairs

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

- a **stochastic kernel**  $P$  from  $G$  to  $X$ :

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

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

## MDP

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

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

- This is an equation in the unknown function  $v \in \mathbb{R}^X$  ( $\mathbb{R}^X$  is a set of all functions from  $X$  to  $\mathbb{R}$ ).
- 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  $\Sigma$  be the set of all **feasible policies** given  $\mathcal{M}$ :

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

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

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

Note: notational issue -  $\mathcal{M}$  here is not MDP, it is a set of Markov operators.

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

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

## 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  $P_\sigma$ -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**  $T_\sigma$ :

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

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

## POLICY OPERATOR

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

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

- Since it is true regardless of  $x$ , we have

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



## 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  $\mathcal{M}$  the **value function** is

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

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

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

## BELLMAN

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

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

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

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

# OPTIMALITY

## Theorem

*Let  $\mathcal{M}$  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.  $\lim_{k \rightarrow \infty} T^k v = v^*$  for all  $v \in \mathbb{R}^X$ ,*
- 3. Bellman's principle of optimality holds for  $\mathcal{M}$ ,*
- 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  $\Sigma$ .
- The required conditions are pretty weak. Important and somewhat hidden: sets are finite and  $r : G \rightarrow \mathbb{R}$ .

## 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

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

- We have  $\|Tv - Tw\|_\infty \leq \|T_\sigma v - T_\sigma w\|_\infty$  for all  $\sigma \in \Sigma$ .
- We showed earlier that  $T_\sigma$  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  $T_\sigma$ . 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  $T_\sigma$  is order preserving, so it must be that  $v_\sigma \leq \bar{v}$ . This is true for any  $\sigma$ , so  $v^* \leq \bar{v}$ .



## OPTIMALITY

- We can use  $T_\sigma$  to look for the value function (instead of value function iteration).
- Start with a guess  $v_0$ , find a greedy policy  $\sigma_0$  and calculate the fixed point of  $T_{\sigma_0}$ :

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

- Repeat the process with  $v_{\sigma_0}$  – 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

---

```
1: procedure HPI
2:    $k \leftarrow 1, \epsilon \leftarrow \tau + 1, v_k \leftarrow v_{\text{init}}$ 
3:   while  $\epsilon > \tau$  do
4:      $\sigma_k \leftarrow v_k$ -greedy policy
5:      $v_{k+1} = (I - \beta P_{\sigma_k})^{-1} r_{\sigma_k}$ 
6:      $\epsilon \leftarrow \|v_{k+1} - v_k\|_{\infty}, k \leftarrow k + 1$ 
7:   end while
8: 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  $v_{\sigma}$  can be expensive.

## OPTIMISTIC POLICY ITERATION

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

# OPI

---

**Algorithm** Optimistic Policy Iteration

---

```
1: procedure OPI
2:    $k \leftarrow 1, \epsilon \leftarrow \tau + 1, v_k \leftarrow v_{\text{init}}$ 
3:   while  $\epsilon > \tau$  do
4:      $\sigma_k \leftarrow v_k$ -greedy policy
5:      $v_{k+1} = T_{\sigma_k}^m v_k$ 
6:      $\epsilon \leftarrow \|v_{k+1} - v_k\|_{\infty}, k \leftarrow k + 1$ 
7:   end while
8: end procedure
```

---