Structural Properties of Markov Decision Processes

(INF8250AE: Introduction to Reinforcement Learning)

Amir-massoud Farahmand

Polytechnique Montréal & Mila

Table of Contents

Goal

We study some important properties of value functions and MDPs.

- Bellman equation
- Bellman operator
  - Monotonicity
  - Contraction
- Focus on discounted tasks
- We show important consequences such as
  - The uniqueness of the solution to the Bellman equations
  - Error bounds on value error
  - Fixed point of T^* is the optimal value function

We refer to these frequently in studying and analyzing RL/Planning algorithms.

Learning Objectives

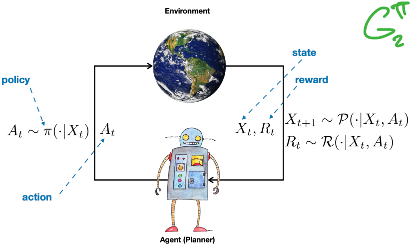
You need to

- Remember: Bellman Equation, Bellman Operator, Greedy Policy, Banach Fixed Point theorem
- Understand: What do Bellman equation encode and why do we use them? What do contraction and monotonicity mean?
- Apply: Contraction property

Bellman Equations

Return

$$\overbrace{R_1, \underbrace{R_2, R_3, R_4, \dots}_{G_2^\pi}}^{G_1^\pi} \quad \dots$$



Consider the sequence of rewards $(R_1, R_2, \dots)$ generated after the agent starts at state $X_1 = x$ and follows policy π. Given the discount factor $0 \leq \gamma < 1$, the return is

$$G_t^\pi \triangleq \sum_{k \geq t} \gamma^{k-t} R_k.$$

Recursive Property of Return

Comparing G_t^{π} and G_{t+1}^{π}, we observe that

$$G_t^{\pi} = R_t + \gamma G_{t+1}^{\pi}. \tag{1}$$

Interpretation: The return at the current time is equal to the immediate reward plus the *discounted* return at the next time step.

- Return is a random variable (r.v.).
- If we repeat the experiment from the same state x, the return would be different.
- Its distribution, however, is the same.
- Q: When would repeated runs lead to the same return?

From Return to the Bellman Equation

$V^\pi(X_{t+1}) = \mathbb{E}[G^\pi_{t+1} | ..]$

We take (conditional) expectation of G_t^π (conditioned on state x), and expand the return as in (1):

$$
\begin{aligned}
V^\pi(x) &= \mathbb{E}\left[G_t^\pi \mid X_t = x\right] \\
&= \mathbb{E}\left[R_t + \gamma G_{t+1}^\pi \mid X_t = x\right] \\
&= \mathbb{E}\left[R(X_t, A_t) \mid X_t = x\right] + \gamma \mathbb{E}\left[G_{t+1}^\pi \mid X_t = x\right] \\
&= r^\pi(x) + \gamma \mathbb{E}\left[V^\pi(X_{t+1}) \mid X_t = x\right]. \qquad (2)
\end{aligned}
$$

Neither side is random anymore!

Expanding $\mathbb{E}\left[V^\pi(X_{t+1}) \mid X_t = x\right]$

What does $\mathbb{E}\left[V^\pi(X_{t+1}) \mid X_t = x\right]$ mean?
It is the expected value of $V^\pi(X_{t+1})$ when

- the agent is at state x at time t
- chooses action $A \sim \pi(\cdot|x)$
- goes to a state $X_{t+1} \sim \mathcal{P}(\cdot|x, \boxed{A})$

That is:

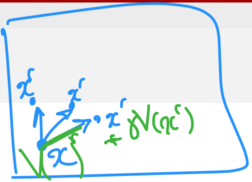
$$\mathbb{E}\left[V^\pi(X_{t+1}) \mid X_t = x\right] = \int \mathcal{P}(\mathrm{d}x'|x, a)\pi(\mathrm{d}a|x)V^\pi(x'). \qquad (3)$$

Deterministic $\pi \to \sum_{x'} P(x'|x, \pi(x)) V^\pi(x')$

For countable state-action spaces, we have

$$\mathbb{E}\left[V^\pi(X_{t+1}) \mid X_t = x\right] = \sum_{x',a} \mathcal{P}(x'|x, a)\pi(a|x)V^\pi(x').$$

Bellman Equation for a Policy π

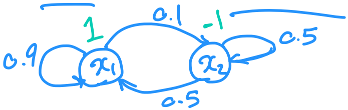


By (2) and (3), we get that for any $x \in \mathcal{X}$, we have

$$V^{\pi}(x) = r^{\pi}(x) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a|x)V^{\pi}(x'). \qquad (4)$$

This is the Bellman equation for a policy π.

Interpretation: The value of following a policy π starting from the state x is the reward that the π-following agent receives at that state plus the discounted average (expected) value that the agent receives at the next-state.



$$\begin{cases} V^{\pi}(x_1) = +1 + \gamma\left[0.9 V^{\pi}(x_1) + 0.1 V^{\pi}(x_2)\right] \\ V^{\pi}(x_2) = -1 + \gamma\left[0.5 V^{\pi}(x_1) + 0.5 V^{\pi}(x_2)\right] \end{cases}$$

Bellman Equation for a Policy π

Using the $\mathcal{P}^\pi$ notation:

$$V^\pi(x) = r^\pi(x) + \gamma \int \mathcal{P}^\pi(\mathrm{d}x'|x) V^\pi(x').$$

Or even more compactly,

$$V^\pi = r^\pi + \gamma \mathcal{P}^\pi V^\pi.$$

Remark

Recall that $(\mathcal{P}^\pi)(A|x) \triangleq \int_{\mathcal{X}} \mathcal{P}(\mathrm{d}y|x,a)\pi(\mathrm{d}a|x)\mathbb{I}_{\{y \in A\}}$

Bellman Equation for a Policy π (Q^π)

The Bellman equation for the action-value function Q^π:

$$
\begin{aligned}
Q^\pi(x,a) &= r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^\pi(x') \\
&= r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) \pi(\mathrm{d}a'|x') Q^\pi(x',a'). \qquad (5)
\end{aligned}
$$

$= Q^\pi(x', \pi(x'))$

More compactly:

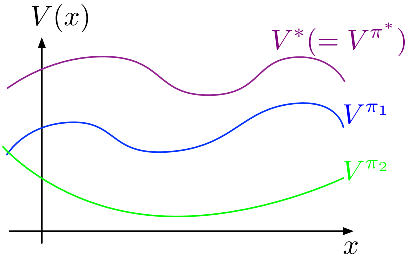
$$Q^\pi = r + \gamma \mathcal{P} V^\pi,$$

with the understanding that V^π and Q^π are related as $V^\pi(x) = \int \pi(\mathrm{d}a|x) Q^\pi(x,a)$.

Remark

The difference with the Bellman equation for V^π is that the choice of action at the first time step is pre-specified, instead of being selected by policy π.

Optimal Policy and Value Function



Recall that the optimal policy π^* is a policy that satisfies $\pi^* \geq \pi$ for any (stationary Markov) policy π. It satisfies

$$\pi^* \leftarrow \underset{\pi \in \Pi}{\operatorname{argmax}} V^\pi.$$

Given an optimal policy, the optimal value function would be V^{π^*}.

Bellman Equations for Optimal Value Functions

Does the optimal value function V^{π^*} satisfy a recursive relation similar to the Bellman equation for a policy π?

Short answer: Yes!

But we have to be a bit careful. Why?! We have to go through a few steps of argument.

Bellman Equations for Optimal Value Functions

The argument goes through three claims:

1. There exists a unique value function V^* that satisfies the following equation: For any $x \in \mathcal{X}$, we have

$$V^*(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}. \qquad (6)$$

   This equation is called the Bellman optimality equation for the value function.
2. V^* is indeed the same as V^{π^*}, the optimal value function when π is restricted to be within the space of stationary policies.
3. For discounted continuing MDPs, we can always find a stationary policy that is optimal within the space of all stationary and non-stationary policies.

In summary: V^* exists and is equal to V^{π^*}.

T: Finite

$\begin{cases} \pi_t^* \\ V_t^* = V_t^{\pi_t^*} \end{cases}$

Bellman Equations for Optimal Value Functions (Q^*)

Optimal action-value function:

$$Q^*(x,a) = r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) \max_{a' \in \mathcal{A}} Q^*(x',a'). \qquad (7)$$

Solutions of the Bellman Equations?

We have defined the Bellman equations for a fixed policy π and the Bellman optimality equation. Some reasonable questions:

- Is there only one solution V^π (or Q^π) satisfying (4) and (5)?
- Is there only one solution V^* (or Q^*) satisfying the Bellman optimality equations (6) and (7)?

We shall prove that their solutions are unique. We need some tools before doing so.

Greedy Policy

Optimal Policy from the Optimal Value Function

- If we know V^* or Q^*, we can find an optimal policy π^*.
- It is a deterministic policy.
- For any $x \in \mathcal{X}$, the optimal policy is

$T = 1 \rightarrow$

$\pi^*(x) = \operatorname{argmax}_a r(x, a)$

$$
\begin{aligned}
\pi^*(x) &= \operatorname*{argmax}_{a \in \mathcal{A}} Q^*(x, a) \\
&= \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}x' | x, a) V^*(x') \right\}.
\end{aligned}
$$

Optimal Policy from the Optimal Value Function

$$\pi^*(x) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}.$$

Interpretation: Suppose that the agent is at state x. To act optimally,

- It needs to act optimality both at the current time step (Now) and in the Future time steps.
- Suppose that we know that the agent is going to act optimally in the Future. This means that when it get to the next state $X' \sim \mathcal{P}(\cdot|x,a)$,
  - it follows the optimal policy π^*.
  - The value of following the optimal policy is going to be $V^*(X')$.

  $V^{\pi^*}(x') = V^*(x')$
- (continued ...)

Optimal Policy from the Optimal Value Function

$\neq \operatorname*{argmax}_{a \in \mathcal{A}} r(x,a)$

$$\pi^*(x) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}.$$

Interpretation: Suppose that the agent is state x. To act optimally,

$\pi_g(x; Q^*) = \pi^*(x)$

V^*

- ...
- Since we do not know where the agent will be at the next time step, the expected performance of acting optimally in the Future is $\int \mathcal{P}(\mathrm{d}x'|x,a)V^*(x')$.
- As we are dealing with discounted tasks, the performance of the agent at the current state x is going to be $r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)V^*(x')$.
- To act optimally Now, the agent should choose an action that maximizes this value.

Greedy Policy

The mapping that selects an action by choosing the maximizer of the (action-) value function is called the greedy policy.

- For $Q \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$, the greedy policy $\pi_g : \mathcal{X} \times \mathcal{B}(\mathcal{X} \times \mathcal{A}) \to \mathcal{A}$ is

$\pi_g(x; Q_1) \neq \pi_g(x; Q_2)$

$$\pi_g(x; Q) = \operatorname*{argmax}_{a \in \mathcal{A}} Q(x, a).$$

$Q_1 \neq Q_2$

- For $V \in \mathcal{B}(\mathcal{X})$, the greedy policy is

$$\pi_g(x; V) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x, a) V(x') \right\}.$$

- We use $\pi_g(V)$ and $\pi_g(Q)$ to denote functions from $\mathcal{X}$ to $\mathcal{A}$.
- $\pi_g(V^*) = \pi_g(Q^*) = \pi^*$. $V^{\pi_g(V^*)} = V^{\pi^*} = V^*$.
- Q: What is the difference between greedy policy and ε-greedy policy?

Greedy Policy

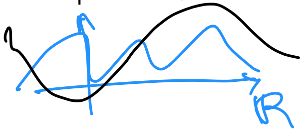
Intuition behind the greedy policy:

- Action selection based on the local information.
- Does not look at all of the future possibilities.
- Only one step ahead (for V) or even no-step ahead (for Q) in order to pick the action. This is myopic.
- Given V^* or Q^*, however, the selected action is going to be the optimal one.
- This is because the optimal value functions encodes the information about the future, so we do not need to explicitly consider all possible futures.

Bellman Operators

Bellman Operators

- The Bellman equations can be seen as the fixed point equation of certain operators known as the Bellman operators.
  - What this means become clear soon.
- Let us review what an operator is.
- An operator (or mapping) $L : \mathcal{Z} \to \mathcal{Z}$ takes a member of space $\mathcal{Z}$ and returns another member of $\mathcal{Z}$.
  - If $\mathcal{Z} = \mathbb{R}$ and $L : z \mapsto z^2$. So $L(5) = 25$ (this is the usual function).
  - If $\mathcal{Z}$ is the space of smooth functions defined on domain $\mathbb{R}$, $L : z \mapsto \frac{d}{dx} z$, is the differentiation operator. So $L(\sin(x)) = \cos(x)$.



Bellman Operators

$\pi_1 \to T^{\pi_1}$

$\pi_2 \to T^{\pi_2}$

Definition (Bellman Operators for policy π)

Given a policy $\pi : \mathcal{X} \to \mathcal{M}(\mathcal{A})$, the Bellman operators $T^\pi : \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ and $T^\pi : \mathcal{B}(\mathcal{X} \times \mathcal{A}) \to \mathcal{B}(\mathcal{X} \times \mathcal{A})$ are defined as the mappings that take V (or Q) and return new functions defined for all $x \in \mathcal{X}$ (for V) or all $(x, a) \in \mathcal{X} \times \mathcal{A}$ (for Q):

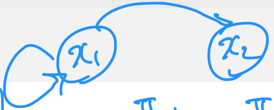
$V^\pi(x')$

$$(T^\pi V)(x) \triangleq r^\pi(x) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a|x)V^{\pi}(x'),$$

$$(T^\pi Q)(x,a) \triangleq r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a'|x')Q(x',a'),$$

$T^\pi V^\pi = V^\pi$

Bellman Operators



$$P^\pi = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, \quad r^\pi = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \rightarrow T^\pi V = r^\pi + \gamma P^\pi V$$

$$= \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} + \gamma \begin{bmatrix} P_{11} & P_{12} \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

If π is deterministic:

$$(T^\pi V)(x) \triangleq r^\pi(x) + \gamma \int \mathcal{P}(\mathrm{d}x'|x, \pi(x))V(x'),$$

$$(T^\pi Q)(x, a) \triangleq r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x, a)Q(x', \pi(x')).$$

Bellman Operators and Bellman Equation

Recall that

$$V^\pi(x) = r^\pi(x) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a|x)V^\pi(x'),$$

$$Q^\pi(x,a) = r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a'|x')Q^\pi(x',a').$$

Using the Bellman operator T^π, we can write them compactly as

$$\boxed{\begin{aligned} V^\pi &= T^\pi V^\pi, \\ Q^\pi &= T^\pi Q^\pi. \end{aligned}}$$

$V^\pi = r^\pi + \gamma P^\pi V^\pi$

$T^\pi : V \mapsto r^\pi + \gamma P^\pi V$

This is a compact form of the Bellman equations.

Bellman Optimality Operators

Definition (Bellman Optimality Operators)

The Bellman operators $T^*: \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ and $T^*: \mathcal{B}(\mathcal{X} \times \mathcal{A}) \to \mathcal{B}(\mathcal{X} \times \mathcal{A})$ are defined as the mapping

$$(T^*V)(x) \triangleq \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V(x') \right\},$$

$$(T^*Q)(x,a) \triangleq r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) \max_{a' \in \mathcal{A}} Q(x',a'),$$

defined for all $x \in \mathcal{X}$ (for V) or all $(x,a) \in \mathcal{X} \times \mathcal{A}$ (for Q).

Bellman Optimality Operators and Bellman Optimality Equation

Comparing the definition of the Bellman optimality operators with the Bellman equations

$$V^*(x) = \boxed{\max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}},$$

$$Q^*(x,a) = r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) \max_{a' \in \mathcal{A}} Q^*(x',a'),$$

we see that

$$V^* = T^* V^*,$$

$$Q^* = T^* Q^*.$$

$V^\pi = T^\pi V^\pi$

$Q^\pi = T^\pi Q^\pi$

Properties of the Bellman Operators

Properties of the Bellman Operators

The Bellman operators have some important properties. The properties that matters for us the most are

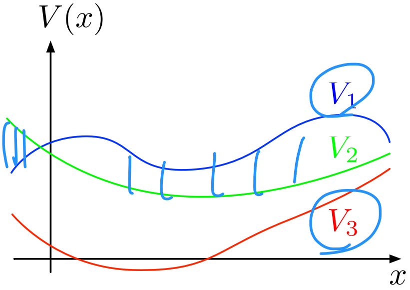
- Monotonicity
- Contraction

They are used in

- basic proofs such as the existence and uniqueness of the solution to the Bellman equations.
- (directly or indirectly) design of many RL/Planning algorithms.

Monotonicity

For two functions $V_1, V_2 \in \mathcal{B}(\mathcal{X})$, we use $V_1 \leq V_2$ if and only if $V_1(x) \leq V_2(x)$ for all $x \in \mathcal{X}$.



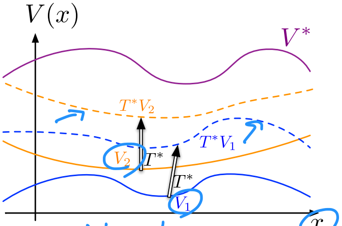
- $V_3 \leq V_1$ and $V_3 \leq V_2$,
- Neither $V_2 \leq V_1$, nor $V_1 \leq V_2$.

Monotonicity

Lemma (Monotonicity)

Fix a policy π. *If* $V_1, V_2 \in \mathcal{B}(\mathcal{X})$, *and* $V_1 \leq V_2$, *then we have*

$$\mathcal{T}^\pi V_1 \leq \mathcal{T}^\pi V_2,$$
$$\mathcal{T}^* V_1 \leq \mathcal{T}^* V_2.$$



$V_1 \leq V_2 \quad \mathcal{T}^* V_1 \leq \mathcal{T}^* V_2$

Monotonicity (Proof)

We only prove the first claim. Let us expand $T^\pi V_1$. As $V_1(x') \le V_2(x')$ for any $x' \in \mathcal{X}$, we get that for any $x \in \mathcal{X}$,

$$
\begin{aligned}
(T^\pi V_1)(x) &= r^\pi(x) + \gamma \int \mathcal{P}^\pi(\mathrm{d}x'|x) \underbrace{V_1(x')}_{\le V_2(x')} \\
&\le r^\pi(x) + \gamma \int \mathcal{P}^\pi(\mathrm{d}x'|x) V_2(x') = (T^\pi V_2)(x).
\end{aligned}
$$

Therefore, $T^\pi V_1 \le T^\pi V_2$.

Contraction Mapping and Banach Fixed Point Theorem

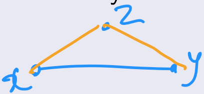
Another important property of the Bellman operators is their contraction property.
What does that mean?
Let us review some mathematical background before proving that the Bellman operators are contraction. We quote several results from Hunter and Nachtergaele [2001].

Metric

Definition (Metric)

A metric or a distance function on $\mathcal{Z}$ is a function $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with the following properties:

- $d(x, y) \geq 0$ for all $x, y \in \mathcal{Z}$; and $d(x, y) = 0$ if and only if $x = y$.
- $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{Z}$ (symmetry).
- $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in \mathcal{Z}$ (triangle inequality).



A metric space $(\mathcal{Z}, d)$ is a set $\mathcal{Z}$ equipped with a metric d.

Example

Let $\mathcal{Z} = \mathbb{R}$ and $d(x, y) = |x - y|$. These together define a metric space $(\mathbb{R}, d)$.

Norm

$x \in \mathbb{R}^2 \to \|x\|_2$
$x \in \mathbb{R} \quad |x|$

Definition (Norm)

A norm on a linear space $\mathcal{Z}$ is a function $\|\cdot\| : \mathcal{Z} \to \mathbb{R}$ with the following properties:

- (non-negative) For all $x \in \mathcal{Z}$, $\|x\| \geq 0$.
- (homogenous) For all $x \in \mathcal{Z}$ and $\lambda \in \mathbb{R}$, $\|\lambda x\| = |\lambda| \, \|x\|$.
- (triangle inequality) For all $x, y \in \mathcal{Z}$, $\|x + y\| \leq \|x\| + \|y\|$.
- (strictly positive) If for a $x \in \mathcal{Z}$, we have that $\|x\| = 0$, it implies that $x = 0$.

Remark

We can use a norm to define a distance between two points in a linear space $\mathcal{Z}$ by defining $d(x, y) = \|x - y\|$. This gives us a metric space $(\mathcal{Z}, d)$.

Norm

Example

Let $\mathcal{Z} = \mathbb{R}^d$ $(d \geq 1)$. The following norms are often used:

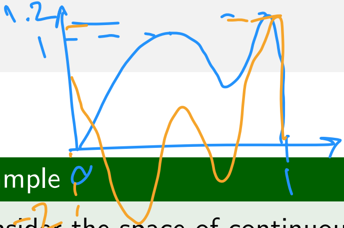
ℓ_p $\qquad \|x\|_2$

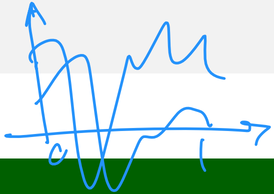
$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{d} |x_i|^p}, \qquad 1 \leq p < \infty,$$

$$\|x\|_\infty = \max_{i=1,\ldots,d} |x_i|.$$

$x = \begin{bmatrix} 1 \\ -5 \\ 3 \end{bmatrix} \rightarrow \|x\|_\infty = ?$ 5

Norm





Example

Consider the space of continuous functions with domain $[0, 1]$. It is denoted by $\mathcal{C}([0, 1])$. This plays the rule of $\mathcal{Z}$. We define the following norm for a function $f \in \mathcal{C}([0, 1])$:

$$\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|.$$

This is called the supremum or uniform norm. Given this norm, $(\mathcal{C}([0, 1]), \|\cdot\|_\infty)$ would be a normed linear space.

Norm

$f(x) = x$ $\quad x \in [0,1]$ $\quad (0,1)$

$\max\limits_{x \in [0,1]} f(x) = 1$ $\quad x = 1 \rightarrow \sup\limits_{x \in (0,1)} f(x) = 1$

For $V \in \mathcal{B}(\mathcal{X})$ and $Q \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$, their supremum norms are

$$\|V\|_\infty = \sup_{\boxed{x \in \mathcal{X}}} |V(x)|,$$

$$\|Q\|_\infty = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |Q(x, a)|,$$

max $\cong$ sup

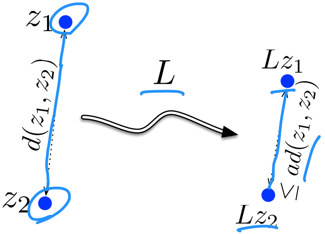
This is simply the maximum value of the value function V or action-value function Q over the state or state-action spaces.

Contraction Mapping

Definition (Contraction Mapping)

Let $(\mathcal{Z}, d)$ be a metric space. A mapping $L : \mathcal{Z} \to \mathcal{Z}$ is a contraction mapping (or contraction) if there exists a constant $0 \leq a < 1$ such that for all $z_1, z_2 \in \mathcal{Z}$, we have

$$\boxed{d(L(z_1), L(z_2)) \leq a d(z_1, z_2).}$$



Contraction Mapping

Example

Let $\mathcal{Z} = \mathbb{R}$ and $d(z_1, z_2) = |z_1 - z_2|$. Consider the mapping $L : z \mapsto az$ for $a \in \mathbb{R}$.
For any $z_1, z_2 \in \mathbb{R}$, we have

$$\begin{aligned} d(L(z_1), L(z_2)) &= |L(z_1) - L(z_2)| = |az_1 - az_2| \\ &= |a||z_1 - z_2| = |a| d(z_1, z_2). \end{aligned}$$

So if $|a| < 1$, this is a contraction mapping.

Why Do We Care About Contraction Mapping?

Some reasons:

- It describes the stability behaviour of a dynamical system.[1]
    - Stability is related to the uniqueness of where the dynamical system converges.
- The contraction property can be used to show the uniqueness of solution of certain equations.
- The contraction property can sometimes be used when solving equations.

[1] There are several notions of stability used in control theory.

Why Do We Care About Contraction Mapping?

As an example of its relation to stability, let $z_0 \in \mathcal{Z}$ and consider a mapping $L : z \mapsto az$ for some $a \in \mathbb{R}$. Define the dynamical system

$$z_{k+1} = Lz_k, \qquad k = 0, 1, \ldots .$$

The dynamical system described by this mapping generates

$$\begin{aligned} & z_0 \\ & z_1 = az_0 \\ & z_2 = az_1 = a^2 z_0 \\ & \vdots \\ & z_k = az_{k-1} = a^k z_0 . \end{aligned}$$

Why Do We Care About Contraction Mapping?

$$z_k = a^k z_0.$$

- If $|a| < 1$, z_k converges to zero, no matter what z_0 is.
- If $a = 1$, we have $z_k = z_0$. So depending on z_0, it converges to different points.
- For $a = -1$, the sequence would oscillate between $+z_0$ and $-z_0$.
- If $|a| > 1$, the sequence diverges (unless $z_0 = 0$).

Remark

The case of converge is the same as the case of L being a contraction map.

Why Do We Care About Contraction Mapping?

$L = 1 \to Lz = z$ $\quad$ $a = 2$

$Lz = az \to Lz = z \to az = z \to 2z = z$

Definition (Fixed Point)

If $L : \mathcal{Z} \to \mathcal{Z}$, then a point $z \in \mathcal{Z}$ such that $\quad z = 0$

$$\boxed{Lz = z}$$

is called a fixed point of L.

Why Do We Care About Contraction Mapping?

$a = -\frac{1}{2} \to z^* = \frac{-1}{-\frac{1}{2}} = 2$

$b = 1 \qquad z_0 = 10 \to z_1 = (a+1)z_0 + b = +\frac{1}{2} \times 10 + 1 = 6$

$z_2 = \frac{1}{2} \times 6 + 1 = 4$

The concept of fixed point and the solution of an equation is closely related.

$z_3 = \frac{1}{2} \times 4 + 1 = 3$

Given an equation $f(z) = 0$, we can convert it to a fixed point equation $Lz = z$ by defining

$z_4 = \frac{3}{2} + 1 = 2.5$

$z_5 = 1.25 + 1 = 2.25$

$$L : z \mapsto f(z) + z.$$

$z_6 = \frac{2.25}{2} + 1 = 2.125$

Then, if $Lz^* = z^*$ for a z^*, we get that $f(z^*) = 0$, i.e., the fixed point of L is the same as the solution of $f(z) = 0$.

$f(z) = az + b$: $L : z \mapsto f(z) + z = (a+1)z + b$

$f(z^*) = 0 \to z^* = -\frac{b}{a}$: $Lz^* = z^* \to |a+1| < 1 \to -2 < a < 0$

Banach Fixed Point Theorem

Theorem (Banach Fixed Point Theorem)

If $L : \mathcal{Z} \rightarrow \mathcal{Z}$ is a contraction mapping on a complete metric space (Z, d), then there exists a unique $z^* \in \mathcal{Z}$ such that $Lz^* = z^*$. Furthermore, the point z^* can be found by choosing an arbitrary $z_0 \in \mathcal{Z}$ and defining $z_{k+1} = Lz_k$. We have $z_k \rightarrow z^*$.

Simple Exercise

Exercise

Suppose that we want to solve $cz + b = 0$ *for* $z \in \mathbb{R}$ *and constants* $c, b \in \mathbb{R}$.

- *Choose a mapping* $L : \mathbb{R} \to \mathbb{R}$ *such that its fixed point is the same as the solution of this equation.*
- *For what range of* c *is this mapping a contraction?*
- *Let* $c = -0.5$ *and* $b = 1$. *If we start from* $z_0 = 0$, *what is the sequence of* z_0, z_1, z_2 *that we obtain by computing* $z_{k+1} = Lz_k$?

Bellman Operator is a Contraction

Lemma (Contraction)

For any π, the Bellman operator T^{π} is a γ-contraction mapping.
The Bellman optimality operator T^{*} is a γ-contraction mapping.

They are specifically γ-contraction w.r.t. the metric d defined based on the supremum norm: $d(V_1, V_2) = \|V_1 - V_2\|_{\infty}$ (and similar for Q). We have

$$\|TV_1 - TV_2\|_{\infty} \leq \gamma \|V_1 - V_2\|_{\infty},$$
$$\|TQ_1 - TQ_2\|_{\infty} \leq \gamma \|Q_1 - Q_2\|_{\infty}.$$

Bellman Operator is a Contraction (Proof)

We only show it for the Bellman operator $T^\pi : \mathcal{B}(\mathcal{X} \times \mathcal{A}) \to \mathcal{B}(\mathcal{X} \times \mathcal{A})$.

Consider two action-value functions $Q_1, Q_2 \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$.

Consider the metric $d(Q_1, Q_2) = \|Q_1 - Q_2\|_\infty$.

We show the contraction w.r.t. this metric.

For any $(x, a) \in \mathcal{X} \times \mathcal{A}$, we have

$\int P(x') Q_1(x') - \int P(x') Q_2(x')$

$= \int P(x') (Q_1(x') - Q_2(x'))$

$$|(T^\pi Q_1)(x, a) - (T^\pi Q_2)(x, a)| =$$

$$\left| \left[r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x, a) \pi(\mathrm{d}a'|x') Q_1(x', a') \right] - \right.$$

$$\left. \left[r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x, a) \pi(\mathrm{d}a'|x') Q_2(x', a') \right] \right|$$

$$= \gamma \left| \int \mathcal{P}(\mathrm{d}x'|x, a) \pi(\mathrm{d}a'|x') \left(Q_1(x', a') - Q_2(x', a') \right) \right| .$$

Bellman Operator is a Contraction (Proof)

Let us upper bound the right-hand side (RHS).
We have an integral of the form $\left|\int P(\mathrm{d}x) f(x)\right|$ (or a summation $\left|\sum_x P(x) f(x)\right|$ for a countable state space). This can be upper bounded as

$$
\begin{aligned}
\left|\int P(\mathrm{d}x) f(x)\right| \leq \int |P(\mathrm{d}x) f(x)| &= \int |P(\mathrm{d}x)| |f(x)| \\
&\leq \int P(\mathrm{d}x) \sup_{x \in \mathcal{X}} |f(x)| \\
&= \sup_{x \in \mathcal{X}} |f(x)| \underbrace{\int P(\mathrm{d}x)}_{=1} = \|f\|_\infty,
\end{aligned}
$$

where we used $\int P(\mathrm{d}x) = 1$.

Bellman Operator is a Contraction (Proof)

In our case, we get that

$$
\begin{aligned}
&|(T^\pi Q_1)(x,a) - (T^\pi Q_2)(x,a)| = \\
&\gamma \left| \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a'|x') \left(Q_1(x',a') - Q_2(x',a') \right) \right| \\
&\leq \gamma \int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a'|x') \left| Q_1(x',a') - Q_2(x',a') \right| \\
&\leq \gamma \left\| Q_1 - Q_2 \right\|_\infty \underbrace{\int \mathcal{P}(\mathrm{d}x'|x,a)\pi(\mathrm{d}a'|x')}_{=1} \\
&= \gamma \left\| Q_1 - Q_2 \right\|_\infty .
\end{aligned}
$$

This inequality holds for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, so it holds for its supremum over $\mathcal{X} \times \mathcal{A}$ too, i.e.,

$0 \leq \gamma < 1$

$$
\left\| (T^\pi Q_1) - (T^\pi Q_2) \right\|_\infty \leq \gamma \left\| Q_1 - Q_2 \right\|_\infty .
$$

This shows that T^π is a γ-contraction.

Consequences of Monotonicity and Contraction

Consequences of Monotonicity and Contraction

Bellman operators are

- Monotonic
- γ-contraction

Several consequences:

$\equiv$ solutions

- Bellman equations have unique fixed points.
- Error bounds on the difference between V and V^* when $V \approx T^* V$.
- V^* is the optimal value function V^{π^*}.
- An optimal policy belongs within the space of stationary policies.

Uniqueness of Fixed Points

$V_1 = T^\pi V_1$ $\quad$ ~~$V_1 \neq V_2$~~

$V_2 = T^\pi V_2$ $\quad$ $V_1 = V_2$

Proposition (Uniqueness of Fixed Points)

The operators T^π and T^* have unique fixed points, denoted by V^π and V^*, i.e.,

$$V^\pi = T^\pi V^\pi,$$
$$V^* = T^* V^*.$$

They can be computed from any $V_0 \in \mathcal{B}(\mathcal{X})$ by iteratively computing $V_{k+1} \leftarrow T^* V_k$ (and similar for V^π using T^π instead) for $k = 0, 1, \ldots$. We have that $V_k \to V^*$ (and similarly, $V_k \to V^\pi$).

The same result is true for Q^π and Q^*.

Uniqueness of Fixed Points (Proof)

- Consider the space of bounded functions $\mathcal{B}(\mathcal{X})$ with the metric d based on the uniform norm, i.e., $d(V_1, V_2) = \|V_1 - V_2\|_\infty$. The space $(\mathcal{B}(\mathcal{X}), d)$ is a complete metric space.
- For any π, the operator T^π is a γ-contraction. Likewise, T^* has the same property too (Lemma 13).
- By the Banach fixed point theorem (Theorem 12), they have a unique fixed point. Moreover, any sequence (V_k) with $V_0 \in \mathcal{B}(\mathcal{X})$ and $V_{k+1} \leftarrow T^\pi V_k$ $(k = 0, 1, \dots)$ is convergent, which means that $\lim_{k\to\infty} \|V_k - V^\pi\|_\infty = 0$.

Value of the Greedy Policy of V^* is V^*

Proposition

We have $T^{\pi}V^* = T^*V^*$ if and only if $V^{\pi} = V^*$.

Value of the Greedy Policy of V^* is V^* (Proof)

Proof of $T^\pi V^* = T^* V^* \implies V^\pi = V^*$:
Assume that $T^\pi V^* = T^* V^*$.
As V^* is the solution of the Bellman optimality equation, we have $T^* V^* = V^*$. Therefore,

$$T^\pi V^* = T^* V^* = V^*.$$

This shows that V^* is a fixed point of T^π.
The fixed point of T^π, however, is unique (Proposition 14) and is equal to V^π.
So V^π and V^* should be the same, i.e., $V^\pi = V^*$.

Value of the Greedy Policy of V^* is V^* (Proof)

Proof of $V^\pi = V^* \implies T^\pi V^* = T^* V^*$:
We apply T^π to both sides of $V^* = V^\pi$ to get

$$T^\pi V^* = T^\pi V^\pi.$$

As V^π is the solution of the Bellman equation for policy π, we have $T^\pi V^\pi = V^\pi$. Therefore,

$$T^\pi V^* = T^\pi V^\pi = V^\pi.$$

By assumption, $V^\pi = V^*$. So we have $T^\pi V^* = V^\pi = V^*$.
On the other hand, we have $V^* = T^* V^*$, so

$$T^\pi V^* = V^* = T^* V^*,$$

which is the desired result.

Value of the Greedy Policy of V^* is V^*

Discussion:

- If $T^\pi V^* = T^* V^*$ for some policy π, the value function V^π of that policy is the same as the fixed point of T^*, which is V^*. ✓
- We have not yet shown that the fixed point of T^* is an optimal value function, in the sense that it is

$$\pi^* \leftarrow \operatorname*{argmax}_{\pi \in \Pi} V^\pi(x) \qquad \text{(for all } x \in \mathcal{X}\text{)}$$

$V^* \stackrel{?}{=} V^{\pi^*}$

  over the space of all stationary policies Π (or even more generally, over the set of all non-stationary policies)
- But it is indeed true!

Value of the Greedy Policy of V^* is V^*

To see the connection to the greedy policy:

- Given V^*, the greedy policy selects
  $\pi_g(x; V^*) = \mathrm{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}$.
- So $T^{\pi_g(V^*)} V^* = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}$
- Compare with $T^* V^*$, i.e.,
  $(T^* V^*)(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\}$.
- So $\boxed{T^{\pi_g(V^*)} V^* = T^* V^*}$.
- This proposition states that the value of following $\pi_g(V^*)$, that is $V^{\pi_g(V^*)}$, is the same as V^*.
- The practical consequence is that if we find V^* and its greedy policy $\pi_g(V^*)$, the value of following the greedy is V^*.
- Practical Consequence: To find an optimal policy, we can find V^* first and then follow its greedy policy $\pi_g(V^*)$.

What if $V \approx T^*V$?

- If we find a V such that $\boxed{V = T^*V}$, we know that $V = V^*$ (similar for T^π and Q).

uniqueness

- What if $\boxed{V \approx T^*V?}$ What can be said about the closeness of V to V^*?
- Practically important, because we often can only solve the Bellman equations approximately, because of various sources of errors
  - Computational
  - Approximation
  - Statistical

An Error Bound based on the Bellman Error

Proposition

For any $V \in \mathcal{B}(\mathcal{X})$ or $Q \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$, we have

$$\|V - V^*\|_\infty \le \frac{\|V - T^*V\|_\infty}{1-\gamma}, \qquad \|Q - Q^*\|_\infty \le \frac{\|Q - T^*Q\|_\infty}{1-\gamma}.$$

$\| V - V^\pi \|$

$\|Q - Q^\pi\|_\infty \leq \frac{\|Q - T^\pi Q\|_\infty}{1-\gamma}$

The quantity $\mathrm{BR}(V) \triangleq T^\pi V - V$ and $\mathrm{BR}^*(V) \triangleq T^* V - V$ are called Bellman Residuals.

Their norms are called Bellman Errors.

$0 = \|V^\pi - V^\pi\|_\infty \leq \frac{\|0\|_\infty}{1-\gamma} = 0$

$V = V^\pi \longrightarrow \mathrm{BR}(V) = T^\pi V^\pi - V^\pi = 0 \rightarrow$

An Error Bound based on the Bellman Error (Proof)

We want to upper bound $\|V - V^*\|_\infty$.
We start from $V - V^*$, and add and subtract T^*V to $V - V^*$.
We then take the supremum norm, and use the triangle inequality to get

$\|① + ②\| \leq \|①\| + \|②\|$

$$
\begin{aligned}
V - V^* &= V \underbrace{\boxed{- T^*V + T^*V}}_{} - V^* \qquad (= 0) \\
\Rightarrow \|V - V^*\|_\infty &= \|\underbrace{V - T^*V}_{①} + \underbrace{T^*V - V^*}_{②}\|_\infty \\
&\leq \|V - T^*V\|_\infty + \|T^*V - V^*\|_\infty .
\end{aligned}
$$

$V^* = T^*V^*$

$\|T^*V - T^*V^*\|_\infty \leq \gamma \|V - V^*\|_\infty .$

An Error Bound based on the Bellman Error (Proof)

Let us focus on the term $\|T^*V - V^*\|_\infty$. Two observations:

- $V^* = T^*V^*$.
- The Bellman optimality operator is a γ-contraction w.r.t. the supremum norm.

Thus,

$$\|T^*V - V^*\|_\infty = \|T^*V - T^*V^*\|_\infty \leq \gamma \|V - V^*\|_\infty .$$

Therefore,

$$\|V - V^*\|_\infty \leq \|V - T^*V\|_\infty + \gamma \|V - V^*\|_\infty .$$

Re-arranging this, we get

$$(1-\gamma) \|V - V^*\|_\infty \leq \|V - T^*V\|_\infty .$$

An Error Bound based on the Bellman Error (for policy π)

Proposition

For any $V \in \mathcal{B}(\mathcal{X})$ or $Q \in \mathcal{B}(\mathcal{X} \times \mathcal{A})$, and any $\pi \in \Pi$, we have

$$\|V - V^\pi\|_\infty \le \frac{\|V - T^\pi V\|_\infty}{1-\gamma}, \qquad \|Q - Q^\pi\|_\infty \le \frac{\|Q - T^\pi Q\|_\infty}{1-\gamma}.$$

V^* is the same as V^{π^*}

The fixed point of T^* is indeed the optimal value function within the space of stationary policies Π.
We use the monotonicity of T^*, in addition to contraction, to prove it.

V^* is the same as V^{π^*}

Proposition

Let V^* be the fixed point of T^*, i.e., $V^* = T^* V^*$. We have

$$V^*(x) = \sup_{\pi \in \Pi} V^\pi(x), \qquad \forall x \in \mathcal{X}.$$

Recall that Π is the space of stationary Markov policies.
We skip the proof! You can read it in the Foundations of Reinforcement Learning.

V^* is the same as V^{π^*} (Proof)

Overview:

- We show that $V^*(x) \leq \sup_{\pi \in \Pi} V^\pi(x)$.
- We show that $\sup_{\pi \in \Pi} V^\pi(x) \leq V^*(x)$.
- Combined, they show that $V^*(x) = \sup_{\pi \in \Pi} V^\pi(x)$.

V^* is the same as V^{π^*} (Proof)

Proof of $V^*(x) \leq \sup_{\pi \in \Pi} V^\pi(x)$:
From the error bound result (Proposition 17) with the choice of $V = V^*$, we get that for any $\pi \in \Pi$,

$$\|V^* - V^\pi\|_\infty \leq \frac{\|V^* - T^\pi V^*\|_\infty}{1-\gamma}. \tag{8}$$

Let $\varepsilon > 0$. Choose a policy π_ε such that

$$\|V^* - T^{\pi_\varepsilon} V^*\|_\infty \leq (1-\gamma)\varepsilon.$$

This is possible because we have

$$(T^* V^*)(x) = \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V^*(x') \right\},$$

so it is sufficient to pick a π_ε that solves the optimization problem up to $(1-\gamma)\varepsilon$-accuracy of the supremum at each state x (if we find the maximizer, then $\varepsilon = 0$).

V^* is the same as V^{π^*} (Proof)

Proof of $V^*(x) \leq \sup_{\pi \in \Pi} V^\pi(x)$ (Continued):
For policy π_ε, (8) shows that

$$\|V^* - V^{\pi_\varepsilon}\|_\infty \leq \varepsilon.$$

This means that

$$V^*(x) \leq V^{\pi_\varepsilon}(x) + \varepsilon, \qquad \forall x \in \mathcal{X}.$$

Notice that $V^{\pi_\varepsilon}(x) \leq \sup_{\pi \in \Pi} V^\pi(x)$ (as $\pi_\varepsilon \in \Pi$). We take $\varepsilon \to 0$ to get that for all $x \in \mathcal{X}$,

$$V^*(x) \leq \lim_{\varepsilon \to 0} \{V^{\pi_\varepsilon}(x) + \varepsilon\} \leq \lim_{\varepsilon \to 0} \left\{\sup_{\pi \in \Pi} V^\pi(x) + \varepsilon\right\} = \sup_{\pi \in \Pi} V^\pi(x). \tag{9}$$

This shows that V^*, the fixed point of T^*, is smaller or equal to the optimal value function within the space of stationary policies.

V^* is the same as V^{π^*} (Proof)

Proof of $\sup_{\pi \in \Pi} V^\pi(x) \leq V^*(x)$:
Consider any $\pi \in \Pi$. By the definition of T^π and T^*, for any $V \in \mathcal{B}(\mathcal{X})$, we have that for any $x \in \mathcal{X}$,

$$
\begin{aligned}
(T^\pi V)(x) &= \int \pi(\mathrm{d}a|x) \left[r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V(x') \right] \\
&\leq \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \int \mathcal{P}(\mathrm{d}x'|x,a) V(x') \right\} \\
&= (T^* V)(x).
\end{aligned}
$$

In particular, with the choice of $V = V^*$, we have

$$
T^\pi V^* \leq T^* V^*.
$$

V^* is the same as V^{π^*} (Proof)

Proof of $\sup_{\pi \in \Pi} V^\pi(x) \le V^*(x)$ **(Continued):**

$$T^\pi V^* \le T^* V^*.$$

As $T^* V^* = V^*$, we have

$$T^\pi V^* \le V^*. \tag{10}$$

We use the monotonicity of T^π (Lemma 3) to conclude that

$$T^\pi (T^\pi V^*) \le T^\pi V^*,$$

which by (10) shows that

$$(T^\pi)^2 V^* \le V^*.$$

We repeat this argument for k times to get that

$$(T^\pi)^k V^* \le V^*.$$

V^* is the same as V^{π^*} (Proof)

Proof of $\sup_{\pi \in \Pi} V^{\pi}(x) \leq V^*(x)$ **(Continued):**

$$(T^{\pi})^k V^* \leq V^*.$$

As $k \to \infty$, Proposition 14 shows that $(T^{\pi})^k V^*$ converges to V^{π} (the choice of V^* is irrelevant). Therefore,

$$V^{\pi} = \lim_{k \to \infty} (T^{\pi})^k V^* \leq V^*.$$

As this holds for any $\pi \in \Pi$, we take the supremum over $\pi \in \Pi$ to get

$$\sup_{\pi \in \Pi} V^{\pi} \leq V^*. \tag{11}$$

Inequalities (9) and (11) together show the desired result.

Consequences of Monotonicity and Contraction

Result	Monotonicity	Contraction
Uniqueness of Fixed Points	✗	✓
Error Upper Bounds	✗	✓
Fixed point of T^* is Optimal Value	✓	✓
Stationary Policies are All You Need	✓	✓

Table: The use of Monotonicity and Contraction Properties in the proof of various results

Sequence of the Bellman Operators

A New Notation: $\mathcal{P}f$

Definition

Given the transition probability kernel $\mathcal{P}$ and a function $f \in \mathcal{B}(\mathcal{X})$, we define $\mathcal{P}f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as the function

$$(\mathcal{P}f)(x, a) \triangleq \int_{\mathcal{X}} \mathcal{P}(\mathrm{d}y|x, a) f(y), \qquad \forall (x, a) \in \mathcal{X} \times \mathcal{A}.$$

Likewise, given the transition probability kernel induced by a policy π, we define $\mathcal{P}^\pi f : \mathcal{X} \to \mathbb{R}$ as $(\mathcal{P}^\pi f)(x) \triangleq \int_{\mathcal{X}} \mathcal{P}^\pi(\mathrm{d}y|x) f(y)$ for all $x \in \mathcal{X}$.

$\mathcal{P}^\pi f$ is the function whose value at a state x is the expected value of function f according to the distribution $\mathcal{P}^\pi(\cdot|x)$, that is,
$(\mathcal{P}^\pi f)(x) = \mathbb{E}_{X' \sim \mathcal{P}^\pi(\cdot|x)} [f(X')]$.

$V^\pi = r^\pi + \gamma \mathcal{P}^\pi V^\pi$

Following a Sequence of Policies

$(\mathcal{P}^{\pi})^m$

For a sequence of policies $\pi_{1:m} = (\pi_1, \dots, \pi_m)$, the transition probability kernel of following them in the order of π_1, then π_2, etc., is denoted by $\mathcal{P}^{\pi_1:\pi_m}$ or $\mathcal{P}^{\pi_{1:m}}$ and is

$$\mathcal{P}^{\pi_{1:m}}(A|x) \triangleq \int_{\mathcal{X}} \mathcal{P}^{\pi_1}(\mathrm{d}y|x)\mathcal{P}^{\pi_{2:m}}(A|y),$$

$$\mathcal{P}^{\pi_1:\pi_m}(B|x,a) \triangleq \int_{\mathcal{X}} \mathcal{P}(\mathrm{d}y|x,a)\mathcal{P}^{\pi_2:\pi_m}(A|y),$$

for deterministic policies, and similar for stochastic policies. This is the generalization of $(\mathcal{P}^{\pi})^m$ when $\pi_{1:m} = (\pi, \dots, \pi)$.

Sequence of the Bellman Operators

Suppose that we have two policies π_1 and π_2. What is the meaning of the operator $T^{\pi_1} T^{\pi_2}$?
To understand what it does, consider a value function V, and see what the effect of $T^{\pi_1} T^{\pi_2}$ on V is.
Denote $T^{\pi_2} V$ by U. We have

$$(T^{\pi_1} \underbrace{T^{\pi_2} V}_{U})(x) = (T^{\pi_1} U)(x) = r^{\pi_1}(x) + \gamma \int \mathcal{P}^{\pi_1}(\mathrm{d}z|x) U(z).$$

The function U at state $z \in \mathcal{X}$ is

$$U(z) = (T^{\pi_2} V)(z) = r^{\pi_2}(z) + \gamma \int \mathcal{P}^{\pi_2}(\mathrm{d}y|z) V(y).$$

Sequence of the Bellman Operators

Combining these two equations, we get

$$(T^{\pi_1} T^{\pi_2} V)(x) =$$

$$r^{\pi_1}(x) + \gamma \int \mathcal{P}^{\pi_1}(\mathrm{d}z|x) \left[r^{\pi_2}(z) + \gamma \int \mathcal{P}^{\pi_2}(\mathrm{d}y|z) V(y) \right] =$$

$$r^{\pi_1}(x) + \gamma \int \mathcal{P}^{\pi_1}(\mathrm{d}z|x) r^{\pi_2}(z) + \gamma^2 \int \mathcal{P}^{\pi_1}(\mathrm{d}z|x) \mathcal{P}^{\pi_2}(\mathrm{d}y|z) V(y) =$$

$$r^{\pi_1}(x) + \gamma (\mathcal{P}^{\pi_1} r^{\pi_2})(x) + \gamma^2 (\mathcal{P}^{\pi_1:\pi_2} V)(x).$$

Therefore, the function $T^{\pi_1} T^{\pi_2} V$ is

$$T^{\pi_1} T^{\pi_2} V = r^{\pi_1} + \gamma \mathcal{P}^{\pi_1} r^{\pi_2} + \gamma^2 \mathcal{P}^{\pi_1:\pi_2} V,$$

and the operator $T^{\pi_1} T^{\pi_2} : \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ is

$$T^{\pi_1} T^{\pi_2} : V \mapsto r^{\pi_1} + \gamma \mathcal{P}^{\pi_1} r^{\pi_2} + \gamma^2 \mathcal{P}^{\pi_1:\pi_2} V.$$

Sequence of the Bellman Operators

For a sequence of policies $\pi_1, \ldots, \pi_m$, we have

$$T^{\pi_1} T^{\pi_2} \ldots T^{\pi_m} V =$$
$$r^{\pi_1} + \gamma \mathcal{P}^{\pi_1} r^{\pi_2} + \gamma^2 \mathcal{P}^{\pi_1:\pi_2} r^{\pi_3} + \cdots + \gamma^{m-1} \mathcal{P}^{\pi_1:\pi_{m-1}} r^{\pi_m} + \gamma^m \mathcal{P}^{\pi_1:\pi_m} V$$
$$= \sum_{k=1}^{m} \gamma^{k-1} \mathcal{P}^{\pi_1:\pi_{k-1}} r^{\pi_k} + \gamma^m \mathcal{P}^{\pi_1:\pi_m} V.$$

Interpretation: the function $T^{\pi_1} T^{\pi_2} \ldots T^{\pi_m} V$ is the value function of following the non-stationary policy $\bar{\pi} = (\pi_1, \ldots, \pi_m)$ in a finite horizon MDP with the terminal reward of V.

Summary

HW
-Strike on Friday

- Bellman equations describe an important recursive properties of value functions.
- Bellman operators T^{π} and T^{*}.
- Greedy policy and the optimal policy.
- Monotonicity and contraction properties of the Bellman operators.
- Bellman equations have uniqued solutions.
- Bellman error $\|V - T^{*}V\|_{\infty}$ provides an upper bound on value error $\|V - V^{*}\|_{\infty}$.
- The solution of the Bellman optimality equation is the optimal value function.

References

Dimitri P. Bertsekas. *Abstract dynamic programming*. Athena Scientific Belmont, 2nd edition, 2018.

John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Company, 2001.