Advances in Dynamic Programming

Theory and Applications

Thomas J. Sargent and John Stachurski

2023

Plan

1. Review traditional dynamic programs (MDPs)
2. Generalize to recursive decision problems (RDPs)
3. Discuss RDP optimality
4. Identify types of RDPs
5. Consider some applications

RDPs are a generalization of MDPs that can handle

- recursive preferences
- robust control
- ambiguity
- state-dependent discounting
- adversarial agents
- negative discount rates
- jump processes (continuous time)
- etc., etc.

Aims

Provide conditions under which

- the value function satisfies Bellman's equation
- Bellman's principle of optimality holds
- at least one optimal policy exists
- standard algorithms converge
    - value function iteration (VFI)
    - optimistic policy iteration (OPI)
    - Howard policy iteration (HPI)

The RDP framework builds on work by

- Eric Denardo
- Dimitri Bertsekas
- Takashi Kamihigashi
- etc.

Further references

- Dynamic Programming (Vol 1)
- Completely Abstract Dynamic Programming

Terminology

Let T be a self-map on metric space $U = (U, \rho)$

We call T **globally stable** on U when

1. T has a unique fixed point $\bar{v}$ in V and
2. $T^k v \to \bar{v}$ as $k \to \infty$ for all $v \in V$

Let T be a self-map on partially ordered space $U = (U, \leqslant)$

We call T **order preserving** on U when

1. $u, v \in U$ and $u \leqslant v$ implies $Tu \leqslant Tv$

Fixed point theory: quick reminders

Our old friend **Banach**

Theorem. If

1. (U, d) is a complete metric space
2. T is a contraction of modulus λ on U

then T has a unique fixed point u^* in U and

$$d(T^k u, u^*) \leqslant \lambda^k d(u, u^*) \quad \text{for all } k \in \mathbb{N} \text{ and } u \in U$$

In particular, T is globally stable on U

An "eventual contraction" result:

- X is finite and T is a self-map on $V \subset \mathbb{R}^{\mathsf{X}}$
- V is closed

Thm. T is globally stable on $V \subset \mathbb{R}^{\mathsf{X}}$ whenever there exists a positive linear operator L on $\mathbb{R}^{\mathsf{X}}$ such that

1. $\rho(L) < 1$ and
2. for all $u, v \in V$, we have

$$|Tu - Tv| \leqslant L|u - v|$$

(For a proof see Ch. 6)

Du's Theorem

Let

- X be a finite set
- $I := [v_1, v_2]$ be a nonempty order interval in $(\mathbb{R}^{\mathsf{X}}, \leqslant)$
- T be an order-preserving self-map on I

Thm. T is globally stable on I if either

1. T is concave and $Tv_1 \gg v_1$ or
2. T is convex and $Tv_2 \ll v_2$

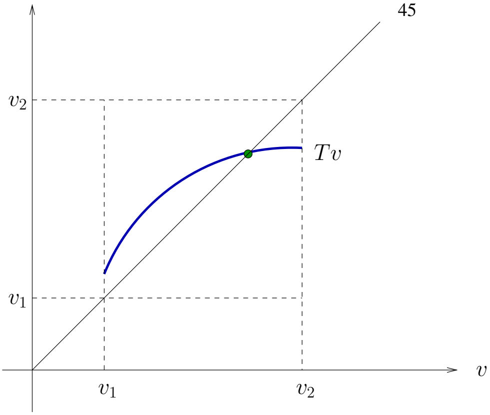


Figure: Concave case (one-dimensional)

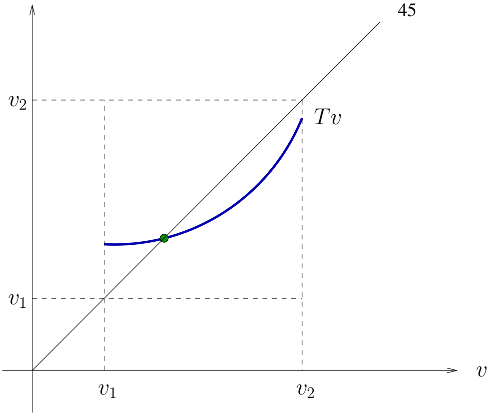


Figure: Convex case (one-dimensional)

Review

Let's quickly review Markov decision processes (MDPs)

- the most standard framework for dynamic programming
- includes many standard economic applications
- useful for intuition
- but limited in some ways

Aims:

- Recall basic concepts
- Lay groundwork for discussing RDPs

States and Actions

We take as given

1. a finite set X called the **state space** and
2. a finite set A called the **action space**

Actions are restricted by a **feasible correspondence** Γ from X to A

- $\Gamma(x) =$ actions available in state x (nonempty)

Given Γ, we define the **feasible state-action pairs**

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

Dynamics

A **stochastic kernel** from G to X is a map

$$P \colon \mathsf{G} \times \mathsf{X} \to [0,1]$$

satisfying

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

- next period state x' is drawn from $P(x, a, \cdot)$

Rewards

Flow reward $r(x, a)$ is received at $(x, a) \in \mathsf{G}$

Lifetime rewards are

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

where $\beta \in (0, 1)$

- β is called the **discount factor**
- r is called the **reward function**

The **Bellman equation** is

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

Formal definition

Let's summarize the primitives:

Given X and A, an **MDP** is a tuple (Γ, β, r, P) where

1. Γ is a nonempty correspondence from X $\to$ A
2. β is a constant in $(0, 1)$
3. r is a function from G to $\mathbb{R}$
4. P is a stochastic kernel from G to X

Policies

A **feasible policy** is a $\sigma \in \mathsf{A}^{\mathsf{X}}$ such that

$$\sigma(x) \in \Gamma(x) \text{ for all } x \in \mathsf{X}$$

Choose $\sigma \iff$

$$\text{respond to state } X_t \text{ with action } \sigma(X_t) \text{ at all } t \geqslant 0$$

Notation:

$$\Sigma := \text{the set of all feasible policies}$$

Closed loop dynamics

Choosing $\sigma \in \Sigma \implies$

1. flow rewards at x are

$$r_\sigma(x) := r(x, \sigma(x))$$

2. the state process $(X_t)_{t \geqslant 0}$ follows

$$P_\sigma(x, x') := P(x, \sigma(x), x')$$

(We say that $(X_t)_{t \geqslant 0}$ is P_σ-**Markov**)

If

$$(X_t)_{t \geqslant 0} \text{ is } P_\sigma\text{-Markov with } X_0 = x$$

then the **lifetime value of σ** starting from x is

$$v_\sigma(x) := \mathbb{E}_x \sum_{t \geqslant 0} \beta^t r(X_t, \sigma(X_t))$$

$$= \mathbb{E}_x \sum_{t \geqslant 0} \beta^t r_\sigma(X_t)$$

$$= \sum_{t \geqslant 0} \beta^t \mathbb{E}_x r_\sigma(X_t)$$

$$= \sum_{t \geqslant 0} \beta^t (P_\sigma^t r_\sigma)(x)$$

In vector notation,

$$v_\sigma = \sum_{t \geqslant 0} (\beta P_\sigma)^t \, r_\sigma$$

Since $\beta < 1$,

$$v_\sigma = (I - \beta P_\sigma)^{-1} \, r_\sigma$$

Policy Operators

Given $\sigma \in \Sigma$, the **policy operator** T_σ is defined by

$$(T_\sigma\, v)(x) = r(x, \sigma(x)) + \beta \sum_{x'} v(x') P(x, \sigma(x), x')$$

In vector notation,

$$T_\sigma\, v = r_\sigma + \beta P_\sigma v$$

Lemma. T_σ is order-preserving on $\mathbb{R}^{\mathsf{X}}$

Proof: $v \leqslant w \implies P_\sigma v \leqslant P_\sigma w \implies T_\sigma\, v \leqslant T_\sigma\, w$

Lemma. T_σ is a contraction of modulus β on $\mathbb{R}^{\mathsf{X}}$

Proof: Let $|v| := v \vee (-v)$ and fix v, w in $\mathbb{R}^{\mathsf{X}}$

We have

$$
\begin{aligned}
|T_\sigma v - T_\sigma w| &= \beta \; |P_\sigma v - P_\sigma w| \\
&\leqslant \beta \, P_\sigma \, |v - w| \\
&\leqslant \beta \, P_\sigma \, \|v - w\|_\infty \mathbb{1} \\
&= \beta \, \|v - w\|_\infty \mathbb{1}
\end{aligned}
$$

Finally, $|a| \leqslant |b|$ implies $\|a\|_\infty \leqslant \|b\|_\infty$

Lemma. T_σ is a contraction of modulus β on $\mathbb{R}^{\mathsf{X}}$

Proof: Let $|v| := v \vee (-v)$ and fix v, w in $\mathbb{R}^{\mathsf{X}}$

We have

$$
\begin{aligned}
|T_\sigma v - T_\sigma w| &= \beta \; |P_\sigma v - P_\sigma w| \\
&\leqslant \beta \, P_\sigma \, |v - w| \\
&\leqslant \beta \, P_\sigma \, \|v - w\|_\infty \mathbb{1} \\
&= \beta \, \|v - w\|_\infty \mathbb{1}
\end{aligned}
$$

Finally, $|a| \leqslant |b|$ implies $\|a\|_\infty \leqslant \|b\|_\infty$

Key idea: Lifetime value of $\sigma \iff$ fixed point of T_σ

Indeed,

$$
\begin{aligned}
v = T_\sigma v \iff & \; v = r_\sigma + \beta P_\sigma v \\
\iff & \; v = (I - \beta P_\sigma)^{-1} r_\sigma \\
\iff & \; v = v_\sigma
\end{aligned}
$$

Greedy Policies

Fix $v \in \mathbb{R}^{\mathsf{X}}$

A policy σ is called ***v*-greedy** if

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

for all $x \in \mathsf{X}$

Note: at least one v-greedy policy exists in Σ

The Bellman Operator

The **Bellman operator** is the self-map on $\mathbb{R}^{\mathsf{X}}$ defined by

$$(Tv)(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x'} v(x') P(x, a, x') \right\}$$

- $Tv = v \iff v$ satisfies the Bellman equation

Note

$$(Tv)(x) = \max_{\sigma \in \Sigma} \left\{ r(x, \sigma(x)) + \beta \sum_{x'} v(x') P(x, \sigma(x), x') \right\}$$

Equivalently, $Tv = \bigvee_{\sigma} T_{\sigma} v$

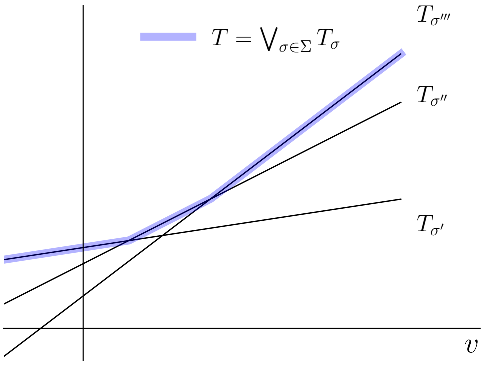


Figure: Visualization in one dimension

Theorem. T is globally stable on $\mathbb{R}^{\mathsf{X}}$

Proof: Easy to check that T is a contraction of modulus β

Optimality

The **value function** $v^* \in \mathbb{R}^{\mathsf{X}}$ is defined by

$$v^*(x) := \max_{\sigma \in \Sigma} v_\sigma(x) \qquad (x \in \mathsf{X})$$

$$= \text{max lifetime value from state } x$$

A policy $\sigma \in \Sigma$ is called **optimal** if

$$v_\sigma = v^*$$

Howard policy iteration (HPI)

input $\sigma_0 \in \Sigma$

$k \leftarrow 0$

repeat

$v_k \leftarrow (I - \beta P_{\sigma_k})^{-1} r_{\sigma_k}$

$\sigma_{k+1} \leftarrow$ a v_k greedy policy

$k \leftarrow k + 1$

until $\sigma_k = \sigma_{k-1}$

return σ_k

Theorem. For any MDP with value function v^*,

1. v^* is the unique solution to the Bellman equation in $\mathbb{R}^\mathsf{X}$
2. A feasible policy is optimal if and only it is v^*-greedy
3. At least one optimal policy exists
4. HPI returns an exact optimal policy in finitely many steps

Remark: Point (2) is called **Bellman's principle of optimality**

Where to now?

Researchers are pushing past the boundaries of MDPs

Problems that do not fit the MDP framework include

1. models with nonlinear recursive preferences
2. models with stochastic discounting
3. recursive equilibria in economic geography, production, etc.
4. problems with ambiguity, adversarial agents
5. various combinations of the above, etc.

Where to now?

Researchers are pushing past the boundaries of MDPs

Problems that do not fit the MDP framework include

1. models with nonlinear recursive preferences
2. models with stochastic discounting
3. recursive equilibria in economic geography, production, etc.
4. problems with ambiguity, adversarial agents
5. various combinations of the above, etc.

Our plan

1. Construct an abstract DP framework that includes MDPs as a special case

2. State optimality results in this framework

3. Provide sufficient conditions for several important cases

4. Connect with applications

Recursive Decision Problems

We begin with a generic version of Bellman's equation:

$$v(x) = \max_{a \in \Gamma(x)} B(x, a, v)$$

- $x \in$ a finite set X (the **state space**)
- $a \in$ a finite set A (the **action space**)
- v is a candidate value function
- $B(x, a, v) =$ total lifetime rewards given x, a, v

More formally...

A **recursive decision process** (RDP) is a triple (Γ, V, B), where

1. Γ is a nonempty correspondence from X to A

- called the **feasible correspondence**
- generates the **feasible state-action pairs**

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

2. V is a nonempty subset of $\mathbb{R}^{\mathsf{X}}$

- called the **value space**
- A set of candidates for the value function

3. B is a map from $\mathsf{G} \times V$ to $\mathbb{R}$ satisfying

(a) **monotonicity**:

$$v, w \in V \text{ and } v \leqslant w \implies B(x, a, v) \leqslant B(x, a, w)$$

for all $(x, a) \in \mathsf{G}$

(b) **consistency**:

$$w(x) := B(x, \sigma(x), v) \text{ is in } V \text{ whenever } \sigma \in \Sigma \text{ and } v \in V$$

- B is called the **value aggregator**

Example. Consider an MDP with Bellman equation

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x'} v(x') P(x,a,x') \right\} \tag{1}$$

Proof: Take Γ as is, $V := \mathbb{R}^{\mathsf{X}}$ and

$$B(x,a,v) := r(x,a) + \beta \sum_{x'} v(x') P(x,a,x')$$

Now (Γ, V, B) is an RDP

- monotonicity and consistency conditions are trivial to check
- setting $v(x) = \max_{a \in \Gamma(x)} B(x,a,v)$ recovers (1)

Example. **Optimal stopping**

$$v(x) = \max \left\{ e(x), c(x) + \beta \sum_{x'} v(x') P(x, x') \right\} \tag{2}$$

Set $V := \mathbb{R}^{\mathsf{X}}$, $\Gamma(x) := \{0, 1\}$ and

$$B(x, a, v) := a e(x) + (1 - a) \left[c(x) + \beta \sum_{x'} v(x') P(x, x') \right]$$

Then (Γ, V, B) is an RDP — checking conditions is trivial

- Setting $v(x) = \max_{a \in \Gamma(x)} B(x, a, v)$ recovers (2)

Example. **State-dependent discounting**

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \sum_{x'} v(x') \beta(x,a,x') P(x,a,x') \right\} \qquad (3)$$

Take Γ as is, $V := \mathbb{R}^{\mathsf{X}}$ and

$$B(x,a,v) := r(x,a) + \sum_{x'} v(x') \beta(x,a,x') P(x,a,x')$$

- (Γ, V, B) is an RDP
- Setting $v(x) = \max_{a \in \Gamma(x)} B(x,a,v)$ recovers (2)

Example. **Risk-sensitive preferences**

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \frac{1}{\theta} \ln \left(\sum_{x'} \exp(\theta v(x')) P(x,a,x') \right) \right\}$$

for nonzero θ

Take Γ as is, $V := \mathbb{R}^{\mathsf{X}}$ and

$$B(x,a,v) := r(x,a) + \beta \frac{1}{\theta} \ln \left(\sum_{x'} \exp(\theta v(x')) P(x,a,x') \right)$$

- an RDP with Bellman equation $v(x) = \max_{a \in \Gamma(x)} B(x,a,v)$

Example. **Quantile preferences**

$$v(x) = \max_{a \in \Gamma(x)} \{ r(x,a) + \beta (R_\tau^a v)(x) \}$$

where

$$(R_\tau^a v)(x) := \tau\text{-th quantile of } v(X') \text{ when } X' \sim P(x, a, \cdot)$$

Take Γ as is, $V := \mathbb{R}^{\mathsf{X}}$ and

$$B(x,a,v) := r(x,a) + \beta (R_\tau^a v)(x)$$

- an RDP with Bellman equation $v(x) = \max_{a \in \Gamma(x)} B(x,a,v)$

Example. **Epstein–Zin preferences**

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a)^\alpha + \beta \left(\sum_{x'} v(x')^\gamma P(x,a,x') \right)^{\alpha/\gamma} \right\}^{1/\alpha}$$

for nonzero α, γ and $r \geqslant 0$

Take Γ as is, $V := (0,\infty)^{\mathsf{X}}$ and

$$B(x,a,v) := \left\{ r(x,a)^\alpha + \beta \left(\sum_{x'} v(x')^\gamma P(x,a,x') \right)^{\alpha/\gamma} \right\}^{1/\alpha}$$

- an RDP with Bellman equation $v(x) = \max_{a \in \Gamma(x)} B(x,a,v)$

Example. **Shortest path problem** on digraph $\mathscr{G} = (\mathsf{X}, E)$

- $c(x, x') =$ cost of traversing edge $(x, x') \in E$
- the direct successors of x denoted by

$$\mathscr{O}(x) := \{x' \in \mathsf{X} : (x, x') \in E\}$$

Aim:

find minimum cost path from x to destination $d \in \mathsf{X}$

No discounting — not an MDP

The Bellman equation is

$$v(x) = \min_{x' \in \mathscr{O}(x)} \{c(x, x') + v(x')\} \tag{4}$$

Set $V := \mathbb{R}^{\mathsf{X}}$, $\Gamma(x) := \mathscr{O}(x)$ and

$$B(x, x', v) := c(x, x') + v(x')$$

Then (Γ, V, B) is an RDP and

$$v(x) = \min_{a \in \Gamma(x)} B(x, a, v)$$

recovers (4) — this is minimization, which we treat in Ch. 9

RDP setting can also handle

- ambiguity (smooth ambiguity model)
- adversarial agents
- negative discount rates
- recursive equilibria in economic geography, production, etc.
- various combinations of the above, etc.
- jump processes (continuous time)

See Ch.s 8–10 and discussion below

RDP theory

So far we have just defined RDPs — little structure

Next steps

1. define lifetime values
2. define optimality
3. provide conditions for optimality
4. discuss algorithms

Policies

Fix arbitrary RDP $\mathscr{R} = (\Gamma, V, B)$

A **feasible policy** is a

$$\sigma \in \mathsf{A}^{\mathsf{X}} \text{ such that } \sigma(x) \in \Gamma(x) \text{ for all } x \in \mathsf{X}$$

- respond to state x with action $a := \sigma(x)$ at all $t \geqslant 0$
- $\Sigma :=$ the set of all feasible policies

Policy Operators

Fix $\sigma \in \Sigma$

The corresponding **policy operator** T_σ is defined at $v \in V$ by

$$(T_\sigma\, v)(x) = B(x, \sigma(x), v) \qquad (x \in \mathsf{X})$$

Lemma. T_σ is an order-preserving self-map on V

Proof: Immediate from monotonicity and consistency

Example. The EZ policy operator is

$$(T_\sigma \, v)(x) = \left\{ r(x, \sigma(x)) + \beta \left(\sum_{x'} v(x')^\gamma P(x, \sigma(x), x') \right)^{\alpha/\gamma} \right\}^{1/\alpha}$$

Example. The state-dependent discounting policy operator is

$$(T_\sigma \, v)(x) = r(x, \sigma(x)) + \sum_{x'} v(x') \beta(x, \sigma(x), x') P(x, \sigma(x), x')$$

Lifetime value

Let $\mathscr{R} := (\Gamma, V, B)$ be an RDP and let σ be any policy

If T_σ has a unique fixed point in V we denote it by v_σ

- Interpretation: v_σ is the lifetime value of following σ

We call $\mathscr{R}$ **well-posed** if

$$T_\sigma \text{ has a unique fixed point in } V \text{ for all } \sigma \in \Sigma$$

- A minimal condition for discussing optimality

Why is the "lifetime value" interpretation valid?

Example. Let $\mathscr{R}$ be the RDP generated by MDP (Γ, β, r, P)

In this case

$$(T_\sigma \, v)(x) = r(x, \sigma(x)) + \beta \sum_{x'} v(x') P(x, \sigma(x), x')$$

Note $\mathscr{R}$ is well-posed because T_σ has unique fixed point

$$v_\sigma = \sum_{t \geqslant 0} \beta^t P^t r_\sigma = (I - \beta P_\sigma)^{-1} \, r_\sigma$$

Moreover,

$$v_\sigma(x) = \mathbb{E}_x \sum_{t \geqslant 0} \beta^t r(X_t, \sigma(X_t)) = \text{ lifetime value under } \sigma$$

Example. Consider the Epstein–Zin RDP

A fixed point of T_σ obeys

$$v(x) = \left\{ r(x, \sigma(x))^\alpha + \beta \left[\sum_{x'} v(x')^\gamma P(x, \sigma(x), x') \right]^{\alpha/\gamma} \right\}^{1/\alpha}$$

A fixed point of this equation is how we define lifetime value from each state under EZ preferences

- Is this RDP well-posed?

Greedy Policies

Given $v \in \mathbb{R}^{\mathsf{X}}$, a policy σ is called **v-greedy** if

$$\sigma(x) \in \operatorname*{argmax}_{a \in \Gamma(x)} B(x, a, v) \quad \text{for all } x \in \mathsf{X}$$

- Note: at least one v-greedy policy exists in Σ

The **Bellman operator** is the self-map on $\mathbb{R}^{\mathsf{X}}$ defined by

$$(Tv)(x) = \max_{a \in \Gamma(x)} B(x, a, v)$$

- Note: $Tv = v \iff v$ satisfies the Bellman equation

Note that

$$\sigma \text{ is } v\text{-greedy} \quad \iff \quad T_\sigma v = Tv$$

Proof: Fix $v \in V$

By definition, σ is v-greedy if and only if

$$\sigma(x) \in \operatorname*{argmax}_{a \in \Gamma(x)} B(x, a, v) \qquad \forall \, x \in \mathsf{X}$$

This is equivalent to

$$B(x, \sigma(x), v) = \max_{a \in \Gamma(x)} B(x, a, v) = (Tv)(x) \qquad \forall \, x \in \mathsf{X}$$

Note that

$$\sigma \text{ is } v\text{-greedy} \quad \iff \quad T_\sigma \, v = Tv$$

Proof: Fix $v \in V$

By definition, σ is v-greedy if and only if

$$\sigma(x) \in \operatorname*{argmax}_{a \in \Gamma(x)} B(x, a, v) \qquad \forall \, x \in \mathsf{X}$$

This is equivalent to

$$B(x, \sigma(x), v) = \max_{a \in \Gamma(x)} B(x, a, v) = (Tv)(x) \qquad \forall \, x \in \mathsf{X}$$

Optimality

Let $\mathscr{R}$ be a well-posed RDP

The **value function** v^* is defined by

$$v^*(x) := \max_{\sigma \in \Sigma} v_\sigma(x) \qquad (x \in \mathsf{X})$$

$$= \text{max lifetime value from state } x$$

A policy $\sigma \in \Sigma$ is called **optimal** if

$$v_\sigma = v^*$$

HPI for well-posed RDPs

input $\sigma_0 \in \Sigma$

$k \leftarrow 0$

repeat

$v_k \leftarrow$ the unique fixed point of T_{σ_k}

$\sigma_{k+1} \leftarrow$ a v_k greedy policy

$k \leftarrow k + 1$

until $\sigma_k = \sigma_{k-1}$

return σ_k

Let $\mathscr{R}$ be an RDP

Key question:

What assumptions to we need for optimality?

Obviously $\mathscr{R}$ must be well-posed

- each σ must have a uniquely defined lifetime value

This is the minimum requirement

What else do we need?

We call $\mathscr{R}$ **globally stable** if

$$T_\sigma \text{ is globally stable on } V \text{ for all } \sigma \in \Sigma$$

Let $\mathscr{R}$ be an RDP

Key question:

What assumptions to we need for optimality?

Obviously $\mathscr{R}$ must be well-posed

- each σ must have a uniquely defined lifetime value

This is the <u>minimum</u> requirement

What else do we need?

We call $\mathscr{R}$ **globally stable** if

$$T_\sigma \text{ is globally stable on } V \text{ for all } \sigma \in \Sigma$$

Let $\mathscr{R}$ be a well-posed RDP with value function v^*

Theorem. If $\mathscr{R}$ is globally stable, then

1. v^* is in V
2. v^* is the unique solution to the Bellman equation in V
3. A feasible policy is optimal if and only it is v^*-greedy
4. At least one optimal policy exists
5. HPI returns an exact optimal policy in finitely many steps

Proof: See Ch. 9

Types of RDPs

The key condition above is global stability

We can check this directly — show each T_σ is globally stable

We can also

1. identify classes of RDPs that are globally stable
2. show that a given application belongs to one of these classes

Let's discuss the second approach

Below $\mathscr{R} = (\Gamma, V, B)$ is a fixed RDP

Contracting RDPs

We call $\mathscr{R}$ **contracting** if

$$\exists\, \beta < 1 \text{ such that } \quad |B(x, a, v) - B(x, a, w)| \leqslant \beta \|v - w\|_\infty$$

for all $(x, a) \in \mathsf{G}$ and $v, w \in V$

Thm. If $\mathscr{R}$ is contracting and V is closed, then $\mathscr{R}$ is globally stable

- Hence all optimality results on slide 58 hold

Proof: Easy to show each T_σ is a mod-β contraction on $(V, \|\cdot\|_\infty)$

(Main idea dates back to Denardo 1967)

Eventually Contracting RDPs

We call $\mathscr{R}$ **eventually contracting** if $\exists$ an $L \geqslant 0$ s.t.

1. $\rho(L) < 1$ and
2. for all $(x, a) \in \mathsf{G}$ and $v, w \in V$,

$$|B(x,a,v) - B(x,a,w)| \leqslant \sum_{x'} |v(x') - w(x')| L(x,x')$$

Thm. If $\mathscr{R}$ is eventually contracting and V is closed, then $\mathscr{R}$ is globally stable

- Hence all optimality results on slide 58 hold

Proof: Let $\mathscr{R}$ be eventually contracting

Fixing $\sigma \in \Sigma$ and $v, w \in V$,

$$|(T_\sigma \, v)(x) - (T_\sigma \, w)(x)| = |B(x, \sigma(x), v) - B(x, \sigma(x), w)|$$

$$\leqslant \sum_{x'} |v(x') - w(x')| L(x, x')$$

In other words,

$$|T_\sigma \, v - T_\sigma \, w| \leqslant L|v - w|$$

The claim now follows from the result on slide 8

Concave RDPs

We call $\mathscr{R}$ **concave** if

1. $V = [v_1, v_2]$
2. $B(x, a, v_1) > v_1(x)$ for all $(x, a) \in \mathsf{G}$ and
3. $v \mapsto B(x, a, v)$ is concave for all $(x, a) \in \mathsf{G}$

Thm. If $\mathscr{R}$ is concave, then $\mathscr{R}$ is globally stable

- Hence all optimality results on slide 58 hold

Proof: Let $\mathscr{R}$ be concave and fix $\sigma \in \Sigma$

Recall that T_σ is an order-preserving self-map on V

Also, $v \mapsto B(x, \sigma(x), v) = (T_\sigma \, v)(x)$ is concave for all $x \in \mathsf{X}$

Hence T_σ is a concave operator

Also, $T_\sigma \, v_2 \leqslant v_2$ because T_σ is a self-map on $V = [v_1, v_2]$

Finally, for any $x \in \mathsf{X}$,

$$(T_\sigma \, v_1)(x) = B(x, \sigma(x), v_1) > v_1(x)$$

Now apply Du's theorem on slide 9

Applications

We have just listed three classes of globally stable RDPs

1. contracting RDPS
2. eventually contracting RDPS
3. concave RDPS

Now let's look at some applications and how they fit in

Application 1: job search with quantile preferences

Set up:

- wage offer process $(W_t)_{t \geqslant 0}$ is P-Markov on finite set W
- discount factor $\beta \in (0, 1)$

The Bellman equation is

$$v(w) = \max \left\{ \frac{w}{1-\beta}, \, c + \beta (R_\tau v)(w) \right\}$$

Here

$$(R_\tau v)(w) := \tau\text{-th quantile of } v(W') \text{ when } W' \sim P(w, \cdot)$$

This problem studied in

- de Castro and Galvao (2019)
- de Castro, Galvao and Nunes (2022)
- de Castro and Galvao (2022)

We can embed the into the RDP framework by taking

- $\Gamma(w) := \{0, 1\}$
- $V := \mathbb{R}^{\mathsf{W}}$
- B given by

$$B(w, a, v) := a \frac{w}{1-\beta} + (1-a)[c + \beta (R_\tau v)(w)]$$

Now $\mathscr{R} := (\Gamma, V, B)$ is an RDP with Bellman equation

$$v(w) = \max_{a \in \Gamma(x)} B(x, a, v) = \max \left\{ \frac{w}{1-\beta}, \, c + \beta (R_\tau v)(w) \right\}$$

Proposition. $\mathscr{R}$ is a contracting RDP

<u>Proof</u>: The quantile map R_τ obeys (see Ch. 7)

$$R_\tau(v + \lambda) = R_\tau \, v + \lambda \text{ for all } v \in \mathbb{R}^{\mathsf{X}} \text{ and } \lambda \in \mathbb{R}$$

Hence

$$\begin{aligned} B(x, a, v + \lambda) &= a \frac{w}{1-\beta} + (1-a)[c + \beta(R_\tau(v + \lambda))(w)] \\ &= a \frac{w}{1-\beta} + (1-a)[c + \beta(R_\tau v)(w)] + (1-a)\beta\lambda \\ &\leqslant B(x, a, v) + \beta\lambda \end{aligned}$$

Hence

$$
\begin{aligned}
B(x,a,v) &= B(x,a,v'+v-v') \\
&\leqslant B(x,a,v'+\|v-v'\|_\infty) \leqslant B(x,a,v') + \beta\|v-v'\|_\infty
\end{aligned}
$$

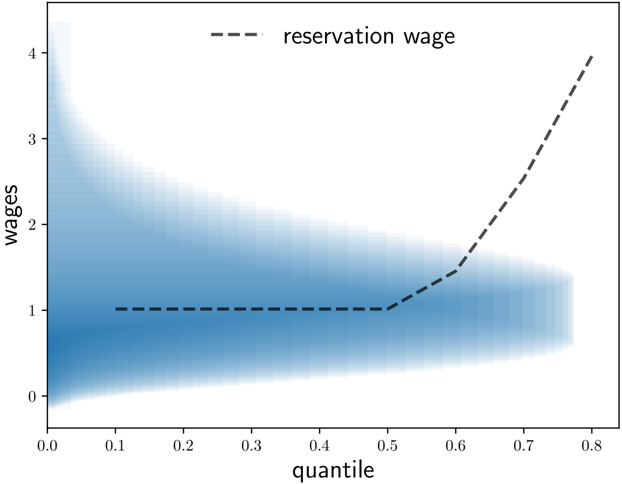
$$\therefore \quad B(x,a,v) - B(x,a,v') \leqslant \beta\|v-v'\|_\infty$$

Reversing the roles of v and v' gives

$$|B(x,a,v) - B(x,a,v')| \leqslant \beta\|v-v'\|_\infty$$

Hence $\mathscr{R}$ is contracting and, since V is closed, globally stable

Hence all optimality properties apply



reservation wage
wages
quantile
0
1
2
3
4
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Application 2: adversarial agents

Consider

$$v(x) = \max_{a \in \Gamma(x)} \inf_{d \in D} \left\{ r(x, a, d) + \beta \sum_{x'} v(x') P(x, a, d, x') \right\}$$

Choice $d \in D$ is made by the adversary

$P(x, a, d, \cdot)$ is a distribution over X for each feasible (x, a, d)

We assume that

- Γ is a nonempty correspondence from X to A
- D is nonempty

Set

$$B(x, a, v) := \min_{d \in D} \left\{ r(x, a, d) + \beta \sum_{x'} v(x') P(x, a, d, x') \right\}$$

Fix $\varepsilon > 0$, set

$$V = [v_1, v_2] \quad \text{where} \quad v_1 := \frac{\min r - \varepsilon}{1 - \beta} \quad \text{and} \quad v_2 := \frac{\max r}{1 - \beta}$$

Then (Γ, V, B) is an RDP and

$$v(x) = \max_{a \in \Gamma(x)} B(x, a, v)$$

recovers the adversarial agent Bellman equation

Prop. $\mathscr{R}$ is a concave RDP

<u>Proof</u> Fixing $(x, a) \in \mathsf{G}$ and $v, w \in V$, we have

$$B(x, a, \lambda v + (1 - \lambda) w) = \min_{d} \left\{ r + \beta \sum (\lambda v + (1 - \lambda) w) P \right\}$$

$$= \min_{d} \left\{ \lambda [r + \beta \sum v P] + (1 - \lambda)[r + \beta \sum w P] \right\}$$

Since $\min (f + g) \geqslant \min f + \min g$, we have

$$B(x, a, \lambda v + (1 - \lambda) w) \geqslant \lambda B(x, a, v) + (1 - \lambda) B(x, a, w)$$

For remaining minor details see Ch. 8

Application 3: inventory management

Consider an inventory problem with

$$v(y, z) = \max_{a \in \Gamma(x)} \left\{ r(y, a) + \beta(z) \sum_{z', y'} v(y', z') R(y, a, y') Q(z, z') \right\}$$

where

- y is inventory, a is current order
- $r(y, a)$ is current profits
- $\mathsf{X} := \{0, \ldots, K\}$
- $\beta(z) = 1/(1 + r(z))$ — time-varying interest rates

This is an RDP $\mathscr{R} = (\Gamma, B, V)$ with

- $V := \mathbb{R}^{\mathsf{X}}$ for $\mathsf{X} := \mathsf{Y} \times \mathsf{Z}$
- $\Gamma(y, z) := \{0, \ldots, K - y\}$ and

$$B((y, z), a, v) := r(y, a) + \beta(z) \sum_{z', y'} v(y', z') Q(z, z') R(y, a, y')$$

Proposition. If $L(z, z') := \beta(z) Q(z, z')$ obeys $\rho(L) < 1$, then $\mathscr{R}$ is eventually contracting

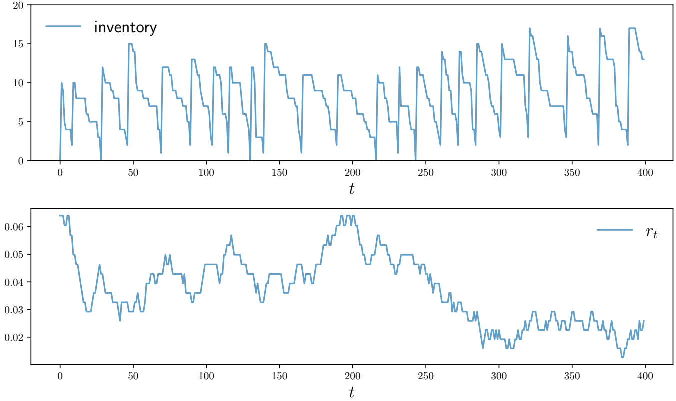
Proof: It suffices to show that, for all $(x, a) \in \mathsf{G}$ and $v, w \in V$,

$$|B(x, a, v) - B(x, a, w)| \leqslant \sum_{x'} |v(x') - w(x')| L(x, x')$$

This holds because

$$
\begin{aligned}
&|B((y,z),a,v) - B((y,z),a,w)| \\
&\qquad \leqslant \beta(z) \sum_{y',z'} \left| v(y',z') - w(y',z') \right| Q(z,z') R(y,a,y') \\
&\qquad \leqslant \beta(z) \sum_{y',z'} \left| v(y',z') - w(y',z') \right| Q(z,z') \\
&\qquad = \sum_{y',z'} \left| v(y',z') - w(y',z') \right| L(z,z')
\end{aligned}
$$

Since $\rho(L) < 1$, we are done


20
15
10
5
0
inventory
0
50
100
150
200
250
300
350
400
t
0.06
0.05
0.04
0.03
0.02
r_t
0
50
100
150
200
250
300
350
400
t

Further applications

Other applications covered in the book:

- Epstein–Zin — details
- Robust control
- Smooth ambiguity, etc.
- Equilibria in production models
- Jump processes (continuous time)
- etc.