DSE 2023 Summer School Lausanne

Lecture 2: Advances in DP Theory

John Stachurski

2023

Loosely based on

- Chapters 8 and 9 of Dynamic Programming: Foundations by Thomas Sargent and John Stachurski
- Completely Abstract Dynamic Programming by Thomas Sargent and John Stachurski

Inspired by

- Abstract Dynamic Programming by Dimitri Bertsekas

Topics

Handling a large range of dynamic programs

- recursive preferences
- quantile preferences
- adversarial agents
- continuous time, etc.

Generalization $\implies$ abstraction $\implies$ clearer proofs

- clarifies optimality conditions
- clarifies relationships between DPs

Motivation

Consider a **Markov decision process** (MDP) with

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)

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

Flow **reward** $r(x, a)$ is received at (x, a)

Given **discount factor** $\beta \in (0, 1)$, **lifetime rewards** are

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

The **Bellman equation** is

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

A **feasible policy** is a map $\sigma \colon \mathsf{X} \to \mathsf{A}$ with

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

- $\Sigma :=$ all feasible policies

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

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

The **Bellman operator** is

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

For each $\sigma \in \Sigma$, we introduce the **policy operator**

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

Note:

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

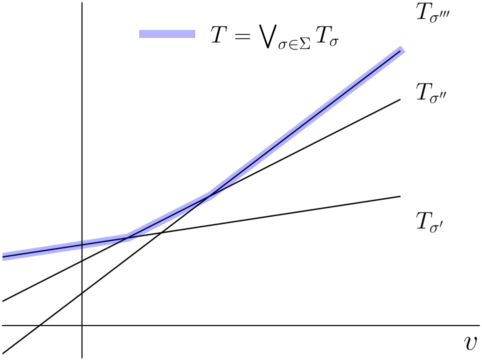


Figure: T is the pointwise max. of $\{T_\sigma\}_{\sigma \in \Sigma}$

Let

- $r_\sigma(x) := r(x, \sigma(x)) =$ rewards under σ
- $P_\sigma(x, x') := P(x, \sigma(x), x') =$ transitions under σ

Note that P_σ is Markov dynamics for the state under σ

When it exists, the **lifetime value** v_σ of σ obeys

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

where

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

Passing the expectation through the sum yields

$$v_\sigma(x) = \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}[r_\sigma(X_t) \,|\, X_0 = x]$$

$$= \sum_{t=0}^{\infty} \beta^t \sum_{x'} r_\sigma(x') P_\sigma^t(x, x')$$

Using operator / matrix notation, this is

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

$$= (I - \beta P_\sigma)^{-1} r_\sigma \quad \text{(Neumann series lemma)}$$

Recall that the policy operator corresponding to σ is

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

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

Clearly

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

Fact. : $T_\sigma^k v \to v_\sigma$ as $k \to \infty$ for all $v \in \mathbb{R}^{\mathsf{X}}$ ($\because$ Banach)

Defining optimality

We define the **value function** via

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

Equivalently,

$$v^* := \bigvee_\sigma v_\sigma$$

A policy σ is called **optimal** if $v_\sigma = v^*$

MDP Optimality

Theorem. For an MDP with Bellman operator T and value function v^*,

1. v^* is the unique fixed point of T in $\mathbb{R}^{\mathsf{X}}$
2. T is a contraction mapping on $\mathbb{R}^{\mathsf{X}}$
3. A feasible policy is optimal if and only it is v^*-greedy
4. At least one optimal policy exists

Standard algorithms

Algorithm 1: VFI

input $v_0 \in \mathbb{R}^{\mathsf{X}}$

input τ, a tolerance level for error

$\varepsilon \leftarrow +\infty$

$k \leftarrow 0$

while $\varepsilon > \tau$ **do**

$\quad v_{k+1} \leftarrow T v_k$

$\quad \varepsilon \leftarrow \|v_k - v_{k+1}\|_\infty$

$\quad k \leftarrow k + 1$

end

Compute a v_k-greedy policy σ

return σ

Algorithm 2: HPI

input $\sigma_0 \in \Sigma$, set $k \leftarrow 0$ and $\varepsilon \leftarrow 1$

while $\varepsilon > 0$ **do**

$v_k \leftarrow$ the lifetime value of σ_k

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

$\varepsilon \leftarrow \mathbb{1}\{\sigma_k \neq \sigma_{k+1}\}$

$k \leftarrow k + 1$

end

return σ_k

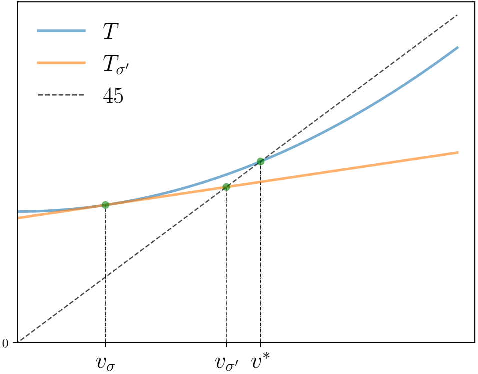


Figure: HPI as a version of Newton's method

Algorithm 3: OPI

input v_0, an initial guess of v^*

input τ, a tolerance level for error

input $m \in \mathbb{N}$, a step size

$k \leftarrow 0$

$\varepsilon \leftarrow +\infty$

while $\varepsilon > \tau$ **do**

  $\sigma_k \leftarrow$ a v_k-greedy policy

  $v_{k+1} \leftarrow T_{\sigma_k}^m v_k$

  $\varepsilon \leftarrow \|v_k - v_{k+1}\|_\infty$

  $k \leftarrow k + 1$

end

return σ_k

Proposition. Under the stated condition, VFI, HPI and OPI all converge

Moreover, HPI converges to an exact optimal policy in finitely many steps

For details and proofs see Ch. 5 of https://dp.quantecon.org/

Modifications and extensions

Let's now look at some extensions to the basic model

We can switch to the **expected value function**

$$g(x,a) := \sum_{x'} v(x')P(x,a,x')$$

with "Bellman operator"

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

- Does R have the same properties as T?
- What are the equivalent algorithms and do they converge?

We can introduce **Epstein–Zin preferences**, as in

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

- Is T still a contraction?
- Are the previous optimality results still valid?
- Do VFI, OPI, HPI converge?

We can introduce **risk-sensitive preferences**, as in

$$(Tv)(x) =$$

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

- Is T still a contraction?
- Are the previous optimality results still valid?
- Do VFI, OPI, HPI converge?

We can introduce risk-sensitive preferences with state-dependent discounting, as in

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

- Is T still a contraction?
- Are the previous optimality results still valid?
- Do VFI, OPI, HPI converge?

Many, many extensions and combinations we can consider

- ambiguity
- expected values in an Epstein–Zin framework
- expected values + ambiguity + state-dependent discounting
- integrated value functions in a risk-sensitive framework in continuous time
- Q-learning, etc., etc.

Is there any unifying theory?

Or are all these problems too diverse?

Abstraction Level 1: RDPs

1. Construct a DP framework based on an abstraction of the Bellman equation
2. State optimality results in this framework
3. Connect with applications

Builds on work by

- Eric Denardo
- Dimitri Bertsekas
- Takashi Kamihigashi

Recursive Decision Problems

We begin with a generic version of the Bellman 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**)
- $B(x, a, v) =$ total lifetime rewards
  - contingent on current state-action pair (x, a)
  - using v to evaluate future states

Formally, a **recursive decision process** (RDP) is a triple

$$\mathscr{R} = (\Gamma, V, B), \qquad \text{where...}$$

1. Γ is a nonempty correspondence from X to A

called the **feasible correspondence**

which generates:

- the **feasible state-action pairs**

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

- the set of **feasible policies**

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

2. V is a subset of $\mathbb{R}^{\mathsf{X}}$ called the **value space**

$\rightarrow$ candidates for the value function

3. B maps $\mathsf{G} \times V$ to $\mathbb{R}$, called the **value aggregator**, satisfies

(a) **monotonicity**:

$$v \leqslant w \implies B(x, a, v) \leqslant B(x, a, w)$$

(b) **consistency**:

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

Example. Every MDP is an RDP

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

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

- monotonicity and consistency conditions are trivial to check
- from $v(x) = \max_{a \in \Gamma(x)} B(x, a, v)$ we recover the MDP Bellman equation

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

Example. Consider an **optimal stopping** problem with

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

Let $V = \mathbb{R}^{\mathsf{X}}$

If $\Gamma(x) = \{0, 1\}$ and

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

then (Γ, V, B) is an RDP with the same Bellman equation

Example. Consider an MDP with **state-dependent discounting**, so that

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

Let $V = \mathbb{R}^{\mathsf{X}}$ and

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

Now (Γ, V, B) is an RDP with the same Bellman equation

Example. Consider a modified MDP with **risk-sensitive preferences**, so that

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

for nonzero θ

With $V = \mathbb{R}^{\mathsf{X}}$ and

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

we obtain an RDP with the same Bellman equation

Example. Consider a modified MDP with **Epstein–Zin preferences**, so that

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

for nonzero α, γ

With $V =$ the strictly positive functions in $\mathbb{R}^{\mathsf{X}}$ and

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

we obtain an RDP with the same Bellman equation

Example. Consider a modified MDP with **quantile preferences**, so that

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

where

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

With $V = \mathbb{R}^{\mathsf{X}}$ and

$$B(x, a, v) = r(x, a) + \beta (R^a_\tau v)(x)$$

we obtain an RDP with the same Bellman equation

Example. Consider a **shortest path problem** on graph $\mathscr{G} = (\mathsf{X}, E)$

- $c(x, x') =$ 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 the minimum cost path from x to a specified vertex d

No discounting (so cannot use MDP theory)

The Bellman equation is

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

Let $V = \mathbb{R}^{\mathsf{X}}$

Let $\Gamma(x) = \mathscr{O}(x)$ and

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

This is an RDP with the same Bellman equation

Policies

Consider an arbitrary RDP (Γ, 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_t with action $\sigma(X_t)$ 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 Epstein–Zin policy operator is

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

Optimality

To define optimality for RDPs, we use the natural generalizations...

Lifetime value

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

Suppose T_σ has a unique fixed point in V

We denote this function by v_σ and call it the **σ-value function**

We interpret this function as the lifetime value of following σ

We call $\mathscr{R}$ **well-posed** if T_σ has a unique fixed point in V for all $\sigma \in \Sigma$

Example. Let $\mathscr{R}$ be the RDP generated by an MDP

Recall that

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

This operator has the unique fixed point

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

- Hence $\mathscr{R}$ is well-posed
- $v_\sigma(x) = \mathbb{E}_x \sum_{t \geqslant 0} \beta^t r(X_t, \sigma(X_t))$ = lifetime value

Example. For the Epstein–Zin RDP,

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

and

$$V := \text{the strictly positive functions in } \mathbb{R}^{\mathsf{X}}$$

- Is this RDP well-posed?

Greedy Policies

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

A policy σ is called v-**greedy** if

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

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

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

Key idea:

$$Tv = v \quad \iff \quad v \text{ satisfies the Bellman equation}$$

Optimality

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

The **value function** is defined by $v^* = \bigvee v_\sigma$

More explicitly,

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

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

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

$$v_\sigma = v^*$$

Howard policy iteration for RDPs

input $\sigma_0 \in \Sigma$, an initial guess of σ^*
$k \leftarrow 0$
$\varepsilon \leftarrow 1$
while $\varepsilon > 0$ **do**
    $v_k \leftarrow$ the unique fixed point of T_{σ_k}
    $\sigma_{k+1} \leftarrow$ a v_k greedy policy
    $\varepsilon \leftarrow \mathbb{1}\{\sigma_k \neq \sigma_{k+1}\}$
    $k \leftarrow k + 1$
end
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 T_σ has a unique fixed point in V

This is the minimum requirement

What else?

Stability

Let $\mathscr{R}$ be an RDP

We call $\mathscr{R}$ **globally stable** if, for all $\sigma \in \Sigma$, the operator T_σ is globally stable on V

That is, for all $\sigma \in \Sigma$,

1. T_σ has a unique fixed point v_σ in V and
2. $\lim_{k \to \infty} T_\sigma^k v = v_\sigma$ for all $v \in V$

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

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

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 optimal policy in finitely many steps
5. VFI and OPI converge

Proof: See Ch 8

Types of RDPs

The optimality properties require global stability of all T_σ

We can check this directly

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 classification approach

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

Contracting RDPs

We call $\mathscr{R}$ **contracting** if $\exists\ \beta < 1$ such that

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

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

Proof: Easy to show that each T_σ is a contraction on V

(Main idea dates back to Denardo 1967)

Eventually Contracting RDPs

We call $\mathscr{R}$ **eventually contracting** if there is an $L \geqslant 0$ such that $\rho(L) < 1$ and

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

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

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

Proof: See the book

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

Proof: See the book

Application: 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 embed into the RDP framework by taking

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

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

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

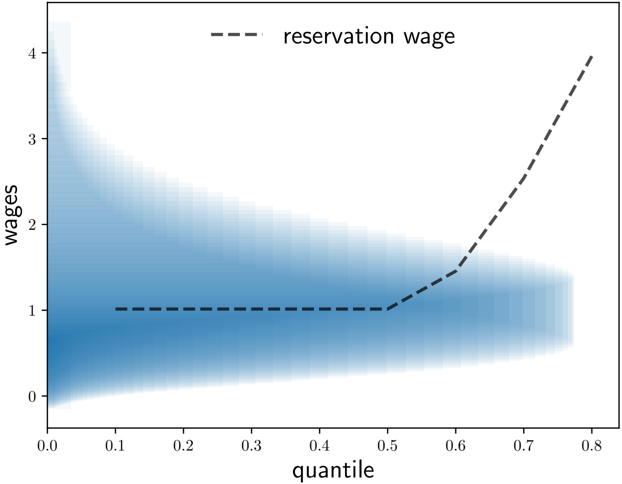
$$v(w) = \max\left\{ \frac{w}{1-\beta},\, c + \beta (R_\tau v)(w) \right\}$$

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

- Proof: See the text

Since V is closed, $\mathscr{R}$ is 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

Abstraction Level 2: ADPs

We define an **abstract dynamic program** (**ADP**) to be a pair

$$\mathscr{A} = (V, \{T_\sigma\}_{\sigma \in \Sigma}), \qquad \text{where}$$

1. $V = (V, \preceq)$ is a partially ordered set and
2. $\{T_\sigma\}_{\sigma \in \Sigma}$ is a family of self-maps on V

Below,

- elements of Σ will be referred to as **policies**
- elements of $\{T_\sigma\}$ are called **policy operators**

If T_σ has a unique fixed point, then we

- denote it v_σ
- call it the **σ-value function**

Interpretation:

- V is a set of candidate value functions
- Σ is a set of feasible policies
- the lifetime value of $\sigma \in \Sigma$ is v_σ
- we seek a greatest element in $\{v_\sigma\}_{\sigma \in \Sigma}$

Example. Consider an RDP (Γ, V, B)

Let Σ be the set of feasible policies

Recall that, for each $\sigma \in \Sigma$ and $v \in V$,

$$(T_\sigma \, v)(x) = B(x, \sigma(x), v)$$

The pair $(V, \{T_\sigma\})$ is an ADP

Recall the expected value Bellman equation

$$g(y,a) =$$

$$\sum_{y'} \int \max_{a' \in \Gamma(y')} \left\{ r(y', \varepsilon', a') + \beta g(y', a') \right\} P(y, a, y') \varphi(\varepsilon') \, \mathrm{d}\varepsilon'$$

If $V = \mathbb{R}^{\mathsf{G}}$ and R_σ is defined by

$$(R_\sigma g)(y,a) =$$

$$\sum_{y'} \int \left\{ r(y', \varepsilon', \sigma(y)) + \beta g(y', \sigma(y)) \right\} P(y, a, y') \varphi(\varepsilon') \, \mathrm{d}\varepsilon'$$

then $(V, \{R_\sigma\})$ is an ADP

Benefits of ADP theory

- More abstraction means easier proofs
- Removing structure makes it easier to see connections
- Can handle a more diverse range of problems

Given $v \in V$, a policy σ in Σ is called **v-greedy** if

$$T_\sigma \, v \succeq T_\tau \, v \quad \text{for all } \tau \in \Sigma \tag{1}$$

Example. For an RDP we have

$$(T_\sigma \, v)(x) = B(x, \sigma(x), v)$$

so (1) holds iff

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

- ADP definitions generalize RDP definitions

Fix an ADP $\mathscr{A} = (V, \{T_\sigma\})$

We define the **Bellman operator** via

$$Tv := \bigvee_\sigma T_\sigma\, v$$

(if it exists)

We say that $v \in V$ satisfies the **Bellman equation** if $Tv = v$

Properties

We say that $\mathscr{A} = (V, \{T_\sigma\})$ is

- **well-posed** if T_σ has one fixed point in V for each $\sigma \in \Sigma$
- **order stable** if (V, T_σ) is order stable for each $\sigma \in \Sigma$
- **max-stable** if $\mathscr{A}$ is order stable, each $v \in V$ has at least one greedy policy, and T has at least one fixed point in V

Note: order stability is a regularity property — see Ch 9

Let $\mathscr{A}$ be a well-posed ADP

A policy $\sigma \in \Sigma$ is called **optimal** for $\mathscr{A}$ if

$$v_\tau \preceq v_\sigma \text{ for all } \tau \in \Sigma$$

We set $v^* := \bigvee_\sigma v_\sigma$ and call v^* the **value function**

We define a self-map H on V via

$$H\, v = v_\sigma \quad \text{where} \quad \sigma \text{ is } v\text{-greedy}$$

Iterating with H is an abstract version of HPI

Max-Optimality

Theorem. If $\mathscr{A}$ is max-stable, then

1. v^* exists in V
2. v^* is the unique solution to the Bellman equation in V
3. a policy is optimal if and only if it is v^*-greedy
4. at least one optimal policy exists

If, in addition, Σ is finite, then HPI $\to v^*$ in finitely many steps

Proof: See Ch. 9

Subordinate ADPs

Let $\mathscr{A} := (V, \{T_\sigma\})$ and $\hat{\mathscr{A}} := (\hat{V}, \{\hat{T}_\sigma\})$ be ADPs

We say that $\hat{\mathscr{A}}$ is **subordinate** to $\mathscr{A}$ if $\exists$

1. an order-preserving map F from V onto $\hat{V}$ and
2. order-preserving maps $\{G_\sigma\}_{\sigma \in \Sigma}$ from $\hat{V}$ to V

such that

$$T_\sigma = G_\sigma \circ F \quad \text{and} \quad \hat{T}_\sigma = F \circ G_\sigma \qquad \text{for all } \sigma \in \Sigma$$

Let $G = \bigvee_\sigma G_\sigma$

Theorem. If

1. $\mathscr{A}$ is max-stable and
2. $\hat{\mathscr{A}}$ is subordinate to $\mathscr{A}$,

then $\hat{\mathscr{A}}$ is also max-stable and the Bellman operators are related by

$$T = G \circ F \quad \text{and} \quad \hat{T} = F \circ G$$

while the value functions are related by

$$v^* = G\,\hat{v}^* \quad \text{and} \quad \hat{v}^* = F\,v^*$$

Moreover,

1. if σ is optimal for $\mathscr{A}$, then σ is optimal for $\hat{\mathscr{A}}$, and
2. if $G_\sigma\,\hat{v}^* = G\,\hat{v}^*$, then σ is optimal for $\mathscr{A}$

Application

Consider an Epstein–Zin dynamic program with Bellman equation

$$v(w, e) = \max_{0 \leqslant s \leqslant w} \left\{ r(w, s, e)^{\alpha} + \beta \left(\sum_{e'} v(s, e')^{\gamma} \varphi(e') \right)^{\alpha/\gamma} \right\}^{1/\alpha}$$

Here

- w is current wealth (discretized)
- s is savings (discretized)
- e is an IID endowment shock with range E
- β is a constant in $(0, 1)$ and r is a reward function

The policy operator corresponding to $\sigma \in \Sigma$ is

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

Proposition. If

- $\mathsf{X} := \mathsf{W} \times \mathsf{E}$ and
- $V := (0, \infty)^{\mathsf{X}}$,

then $\mathscr{A} = (V, \{T_\sigma\})$ is a max-stable ADP

(Details in Ch 9)

Next consider the operator

$$(B_\sigma\, h)(w) = \left\{ \sum_e \{ r(w, \sigma(w), e)^\alpha + \beta h(\sigma(w))^\alpha \}^{\gamma/\alpha} \varphi(e) \right\}^{1/\gamma},$$

where h is an element of $(0, \infty)^{\mathsf{W}}$

Define F at $v \in V$ by

$$(Fv)(w) = \left\{ \sum_e v(w, e)^\gamma \varphi(e) \right\}^{1/\gamma} \qquad (w \in \mathsf{W})$$

Then $\mathscr{B} = (F(V), \{B_\sigma\})$ is also an ADP

Moreover, $\mathscr{B}$ is subordinate to $\mathscr{A}$

To see, this, define G_σ by

$$(G_\sigma \, h)(w, e) = \{r(w, \sigma(w), e)^\alpha + \beta h(\sigma(w))^\alpha\}^{1/\alpha}$$

Then

- F and G_σ are order-preserving
- T_σ is equal to $G_\sigma \circ F$ and
- B_σ is equal to $F \circ G_\sigma$

Algorithm 4: Solving $\mathscr{A}$ via $\mathscr{B}$

input $\sigma_0 \in \Sigma$, set $k \leftarrow 0$ and $\varepsilon \leftarrow 1$

while $\varepsilon > 0$ **do**

$h_k \leftarrow$ the fixed point of B_{σ_k}

$\sigma_{k+1} \leftarrow$ an h_k-greedy policy, satisfying

$$\sigma_{k+1}(w) \in \operatorname*{argmax}_{0 \leqslant s \leqslant w} \left\{ \sum_e \{r(w,s,e)^\alpha + \beta h(s)^\alpha\}^{\gamma/\alpha} \varphi(e) \right\}^{1/\gamma}$$

$\varepsilon \leftarrow \mathbb{1}\{\sigma_k \neq \sigma_{k+1}\}$ and $k \leftarrow k+1$

end

Compute σ to satisfy

$$\sigma(w,e) \in \operatorname*{argmax}_{0 \leqslant s \leqslant w} \{r(w,s,e)^\alpha + \beta h_k(s)^\alpha\}^{1/\alpha}$$

return σ

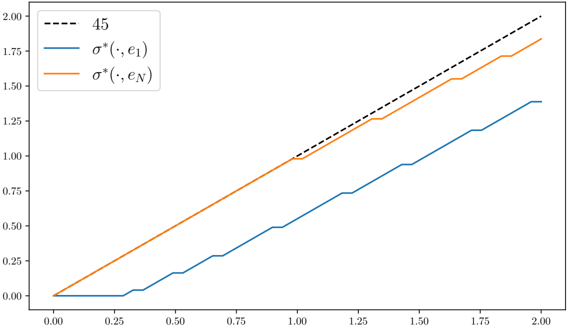


Figure: Optimal savings policy with Epstein–Zin preference

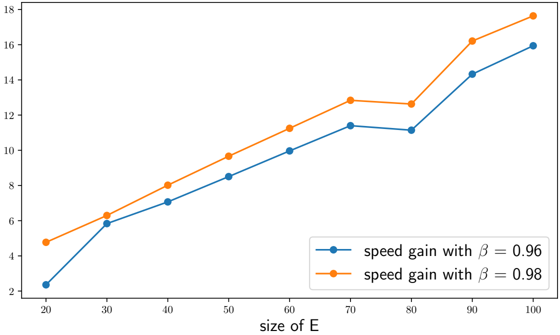


Figure: Speed gain from replacing $\mathscr{A}$ with subordinate model $\mathscr{B}$

For details of computations see

https://github.com/jstac/adps_public