Dynamic Programming

Chapter 9: Abstract Dynamic Programming

Thomas J. Sargent and John Stachurski

2024

Topics

1. Order stability
2. ADPs — definition
3. ADPs — optimality theory
4. Further applications

Abstract Dynamic Programs: Prelude

We saw in Ch. 8 that a well-posed RDP produces

- a set of feasible policies Σ and
- a set of policy operators $\{T_\sigma\}_{\sigma\in\Sigma}$ on the value space $V \subset \mathbb{R}^{\mathsf{X}}$
- a set of lifetime values $\{v_\sigma\}_{\sigma\in\Sigma} \subset V$

Optimality:

- $\{v_\sigma\}_{\sigma\in\Sigma}$ defines the value function v^* via $v^* = \vee_\sigma v_\sigma$
- an optimal policy is a $\sigma \in \Sigma$ obeying $v_\sigma = v^*$

To shed unnecessary structure before the optimality proofs, a natural idea is to

1. start directly with an abstract set of "policy operators" $\{T_\sigma\}$ acting on some abstract "value space" V (partially ordered)
2. define lifetime values and optimality as in the previous slide
3. investigate conditions on the family of operators $\{T_\sigma\}$ that lead to characterizations of optimality that we hope to obtain

This chapter pursues the plan of action listed above

But before discussing these "abstract dynamic programs," let's cover some order-theoretic results that will be useful for our plan

Order Stability

Let

- V be a partially ordered set
- T be a self-map on V with exactly one fixed point $\bar{v}$ in V

In this setting, we call T

- **upward stable** on V if $v \in V$ and $v \preceq T\,v$ implies $v \preceq \bar{v}$,
- **downward stable** on V if $v \in V$ and $T\,v \preceq v$ implies $\bar{v} \preceq v$
- **order stable** on V if T is both upward and downward stable on V

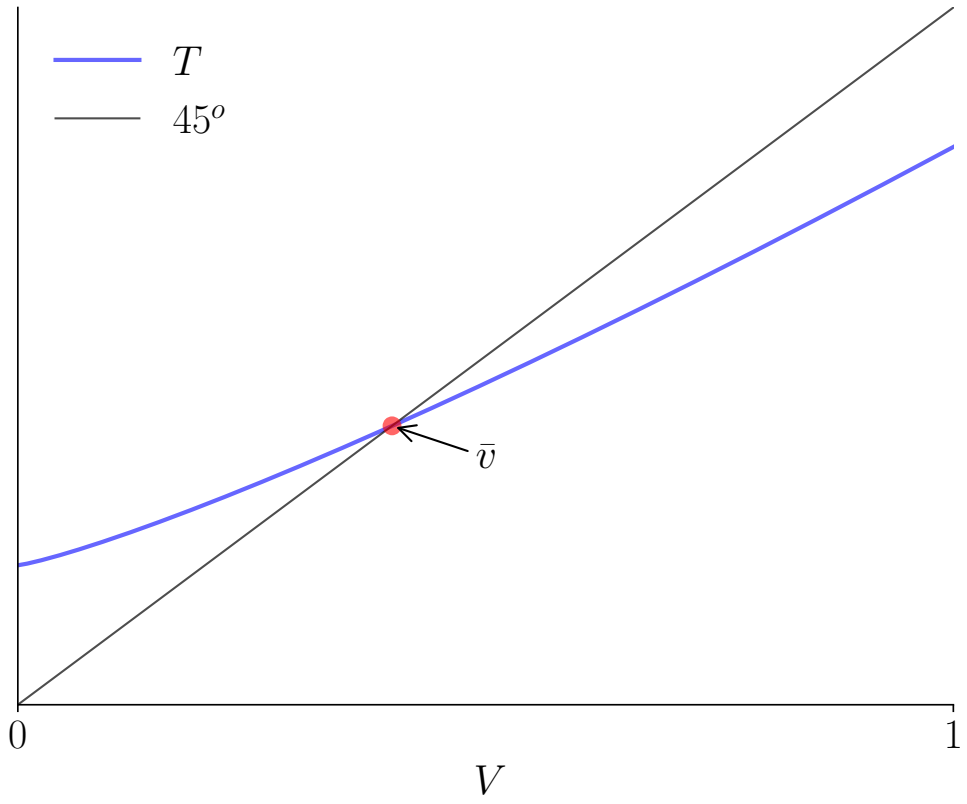

Figure: Order-stability

Example. Consider $Tv = r + Av$ on $\mathbb{R}^{\mathsf{X}}$ with $|\mathsf{X}| < \infty$

We assume that

- $r \in \mathbb{R}^{\mathsf{X}}$ and
- $A \in \mathcal{L}(\mathbb{R}^{\mathsf{X}})$ with $0 \leqslant A$ and $\rho(A) < 1$

Ex. Prove that T is order stable on $\mathbb{R}^{\mathsf{X}}$

Proof: By the NSL, T has a unique fixed point in $\mathbb{R}^{\mathsf{X}}$ given by

$$\bar{v} := (I - A)^{-1} r$$

Given $v \in \mathbb{R}^{\mathsf{X}}$ with $v \leqslant T\, v$, we have

$$v \leqslant r + Av \iff (I - A)v \leqslant r \iff v \leqslant (I - A)^{-1} r = \bar{v}$$

(We are using $(I - A)^{-1} = \sum_{t \geqslant 0} A^t \geqslant 0$)

This proves upward stability — downward stability is similar

Let X be finite, fix $V \subset \mathbb{R}^{\mathsf{X}}$, and let T be a self-map on V

Lemma. If T is order-preserving and globally stable on V, then T is order stable on V

Proof: By global stability, T has a unique fixed point $\bar{v}$ in V

Suppose $v \in V$ and $v \leqslant T\, v$

Since T is order-preserving, iterating yields $v \leqslant T^k\, v$ for all $k \in \mathbb{N}$

Taking the limit gives $v \leqslant \bar{v}$, so upward stability holds

The proof of downward stability is similar

ADPs

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

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

1. $V = (V, \preceq)$ is a partially ordered set
2. $\{T_\sigma\}_{\sigma \in \Sigma}$ is a family of self-maps on V
3. for all $v \in V$, the set $\{T_\sigma \, v\}_{\sigma \in \Sigma}$ has a least and greatest element

Below,

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

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

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

Note:

$\{T_\sigma\, v\}_{\sigma\in\Sigma}$ has a greatest element $\iff$

at least one v-greedy policy exists

Existence of a least element is needed only because we wish to consider minimization as well as maximization

For settings where only maximization is considered, this can be dropped from the list of assumptions

Example. Let $\mathcal{R} = (\Gamma, V, B)$ be an RDP with finite state space X

For each σ in Σ, let $(T_\sigma\, v)(x) = B(x, \sigma(x), v)$

The pair $\mathcal{A}_\mathcal{R} := (V, \{T_\sigma\})$ is an ADP

- V is partially ordered by $\leqslant$
- T_σ is a self-map on V for all $\sigma \in \Sigma$ and
- a v-greedy policy is obtained by choosing $\bar{\sigma} \in \Sigma$ such that

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

(Replacing argmax with argmin yields a least element of $\{T_\sigma\, v\}$)

In this setting, we call $\mathcal{A}_{\mathcal{R}}$ the ADP **generated by** $\mathcal{R}$

Example. Let $\mathcal{M} = (\Gamma, \beta, r, P)$ be an MDP

- state space is X
- policy operators $\{T_\sigma\}$ defined by $T_\sigma\, v = r_\sigma + \beta P_\sigma\, v$

Since every MDP is an RDP,

$$\mathcal{A}_{\mathcal{M}} := (\mathbb{R}^{\mathsf{X}}, \{T_\sigma\})$$

is an ADP

We call $\mathcal{A}_{\mathcal{M}}$ the ADP **generated by** $\mathcal{M}$

We have just shown that RDPs are ADPs

There are also ADPs that do not fit naturally into the RDP framework

The next example illustrates

Example. Recall from Ch. 5 the Q-factor MDP Bellman operator

$$(Sq)(x,a) = r(x,a) + \beta \sum_{x'} \max_{a' \in \Gamma(x')} q(x',a') P(x,a,x')$$

with $q \in \mathbb{R}^{\mathsf{G}}$ and $(x,a) \in \mathsf{G}$

The corresponding policy operators are given by

$$(S_\sigma\, q)(x,a) \coloneqq r(x,a) + \beta \sum_{x'} q(x',\sigma(x')) P(x,a,x')$$

Each S_σ is a self-map on $\mathbb{R}^{\mathsf{G}} = (\mathbb{R}^{\mathsf{G}}, \leqslant)$

If $q \in \mathbb{R}^{\mathsf{G}}$ and $\sigma(x) \in \operatorname{argmax}_{a \in \Gamma(x)} q(x,a)$ for all $x \in \mathsf{X}$, then $S_\sigma\, q \geqslant S_\tau\, q$ on G for all $\tau \in \Sigma$

Hence σ is q-greedy and $\mathcal{A} \coloneqq (\mathbb{R}^{\mathsf{G}}, \{S_\sigma\})$ is an ADP

Lifetime Value

The objective of dynamic programming is to optimize lifetime value

But what is lifetime value in this abstract context?

In general, for an ADP $(V, \{T_\sigma\})$ and fixed $\sigma \in \Sigma$, we write v_σ for the fixed point of T_σ and call it the **σ-value function**

We interpret it as the lifetime value of policy σ whenever it is uniquely defined

This interpretation was discussed at length for RDPs in Ch. 8 and the situation here is analogous

Example. Let $\mathcal{M}$ be an MDP

Let $\mathcal{A}_{\mathcal{M}}$ be the ADP generated by $\mathcal{M}$

For each σ, the unique fixed point of T_σ is

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

This accords with our interpretation of fixed points of T_σ as lifetime values

Indeed, $(I - \beta P_\sigma)^{-1} r_\sigma$ is precisely the lifetime value of σ under the MDP assumptions

Example. Let $\mathcal{A} = (V, \{T_\sigma\})$ where

- each T_σ is a Koopmans operator on V (Ch. 7)
- $V \subset \mathbb{R}^{\mathsf{X}}$

A fixed point of a Koopmans operator is interpreted as lifetime utility under the preferences it represents

Thus v_σ, when well-defined, is the lifetime value associated with policy σ and the preferences embedded in T_σ

Operators

Let $\mathcal{A} = (V, \{T_\sigma\})$ be an ADP and set

$$T\,v := \bigvee_\sigma T_\sigma\,v \qquad (v \in V)$$

We call T the **Bellman operator** generated by $\mathcal{A}$

By existence of greedy policies, T is a well-defined self-map on V

A function $v \in V$ is said to satisfy the **Bellman equation** if it is a fixed point of T

Example. Consider an RDP $\mathcal{R} = (\Gamma, V, B)$ with Bellman operator

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

We saw in Ch. 8 that T obeys $\bigvee_\sigma T_\sigma v$

Thus, the Bellman operator of the RDP agrees with the Bellman operator T of the corresponding ADP $\mathcal{A}_\mathcal{R}$

It follows that $\mathcal{R}$ and $\mathcal{A}_\mathcal{R}$ have the same Bellman equation

Let $\mathcal{A} = (V, \{T_\sigma\})$ be an ADP with Bellman operator

$$T\, v \coloneqq \bigvee_\sigma T_\sigma\, v$$

Ex. Show that

1. $\sigma \in \Sigma$ is v-greedy if and only if $T_\sigma\, v = T\, v$, and
2. T is order-preserving whenever T_σ is order-preserving for all $\sigma \in \Sigma$

Below we consider Howard policy iteration (HPI) as an algorithm for solving for optimal policies of ADPs

We use precisely the same instruction set as for the RDP case (Ch. 8)

To further clarify the algorithm, we define a map H from V to $\{v_\sigma\}$ via

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

Iterating with H generates the value sequence associated with HPI

Properties

Let $\mathcal{A} \coloneqq (V, \{T_\sigma\}_{\sigma \in \Sigma})$ be an ADP

We call $\mathcal{A}$

- **finite** if Σ is a finite set,
- **order stable** if every policy operator T_σ is order stable on V
- **max-stable** if $\mathcal{A}$ is order stable and T has at least one fixed point in V

Obviously max-stable $\implies$ order stable $\implies$ well-posed

(For the second implication, we understand uniqueness of fixed points as part of the definition of order stability)

Proposition. If $\mathcal{A}$ is order stable and finite, then $\mathcal{A}$ is max-stable

The proposition is proved in the book (Ch. 9)

Corollary. Let $\mathcal{R}$ be an RDP and let $\mathcal{A}_{\mathcal{R}}$ be the ADP generated by $\mathcal{R}$. If $\mathcal{R}$ is globally stable, then $\mathcal{A}_{\mathcal{R}}$ is max-stable

Proof: Let $\mathcal{R}$ and $\mathcal{A}_{\mathcal{R}}$ be as stated

Suppose that $\mathcal{R}$ is globally stable

This implies that each policy operator is order stable (see slide 9)

Hence $\mathcal{A}_{\mathcal{R}}$ is order stable

Since Σ is finite, $\mathcal{A}_{\mathcal{R}}$ is also max-stable

Order stability is central to the optimality results stated below

The next result shows that, at least in simple settings, order stability is necessary for any discussion of optimality

Proposition. Let $\mathcal{A} = (V, \{T_\sigma\})$ be an ADP generated by an RDP $\mathcal{R} = (\Gamma, V, B)$. If V is an order interval in $\mathbb{R}^X$, then the following statements are equivalent:

1. $\mathcal{A}$ is well-posed
2. $\mathcal{A}$ is order stable

Proof: See the book, Ch. 9

Max-Optimality Results

Let $\mathcal{A} = (V, \{T_\sigma\})$ be a well-posed ADP with σ-value functions $\{v_\sigma\}_{\sigma\in\Sigma}$

We define

$$V_\Sigma := \{v_\sigma\}_{\sigma\in\Sigma} \quad \text{and} \quad V_u := \{v \in V : v \preceq Tv\}$$

Ex. Prove that $V_\Sigma \subset V_u$

Proof: For all $v \in V_\Sigma$, we have $v = v_\sigma$ for some σ

Hence $Tv \geqslant T_\sigma\, v = T_\sigma\, v_\sigma = v_\sigma = v$

If V_Σ has a greatest element, then we denote it by v^* and call it the **value function** generated by $\mathcal{A}$

In this setting, a policy $\sigma \in \Sigma$ is called **optimal** for $\mathcal{A}$ if $v_\sigma = v^*$

We say that $\mathcal{A}$ obeys **Bellman's principle of optimality** if

$$\sigma \in \Sigma \text{ is optimal for } \mathcal{A} \quad \Longleftrightarrow \quad \sigma \text{ is } v^*\text{-greedy}$$

These definitions are direct generalizations of of the corresponding definitions for RDPs discussed in Ch. 8

We can now state our main optimality result for ADPs

Theorem. If $\mathcal{A}$ is finite and order stable, then

1. the set of σ-value functions V_Σ has a greatest element v^*
2. v^* is the unique solution to the Bellman equation in V
3. $\mathcal{A}$ obeys Bellman's principle of optimality
4. $\mathcal{A}$ has at least one optimal policy and
5. HPI returns an exact optimal policy in finitely many steps

For a proof see Ch. 9 of the book

This volume focuses on dynamic programming problems with finite states

Here is one high-level result for general state spaces

Proposition. If $\mathcal{A}$ is max-stable, then all results on the previous slide hold

For a proof see Ch. 9

For more results in a general state setting see

1. arXiv paper Completely abstract dynamic programming
2. Vol II — currently being written

Application: Mixed Strategies

Let's consider adding mixed strategies to an RDP

We will need to apply the result on the last slide to discuss optimality (because the set of mixed strategies is not finite)

Let $\mathcal{R} = (\Gamma, V, B)$ be an RDP with finite state space X, finite action space A, policy set Σ and Bellman operator T

A **mixed strategy** for $\mathcal{R}$ is a map φ sending $x \in \mathsf{X}$ into a distribution $\varphi_x \in \mathcal{D}(\mathsf{A})$ supported on $\Gamma(x)$

In other words, for each $x \in \mathsf{X}$,

$$\varphi_x \in \mathcal{D}(\mathsf{A}) \quad \text{and} \quad \sum_{a \in \Gamma(x)} \varphi_x(a) = 1$$

Let Φ be the set of all mixed strategies for $\mathcal{R}$

For each $\varphi \in \Phi$, we introduce

$$(\hat{T}_\varphi\, v)(x) = \sum_{a \in \mathsf{A}} B(x, a, v)\varphi_x(a) \qquad (v \in V,\ x \in \mathsf{X})$$

Intuitively, the right hand side is the expected lifetime value from current state x, when

1. the current action is drawn from φ_x and
2. future states are evaluated via v

Fix $v \in V$ and suppose $\varphi \in \Phi$

Ex. Prove: If for each $x \in \mathsf{X}$,

$$\varphi_x \text{ is supported on } \operatorname*{argmax}_{a \in \Gamma(x)} B(x, a, v)$$

then $\hat{T}_\varphi v \geqslant \hat{T}_\psi v$ for all $\psi \in \Phi$

Ex. Show that, given $v \in V$ and $x \in \mathsf{X}$ we have

$$\max_{\varphi \in \Phi} (\hat{T}_\varphi v)(x) = \max_{a \in \Gamma(x)} B(x, a, v)$$

It follows from the discussion above that

- $\mathcal{A}_M := (V, \{\hat{T}_\varphi\}_{\varphi\in\Phi})$ is an ADP
- the Bellman operator $\hat{T}$ associated with $\mathcal{A}_M$ is given by

$$(\hat{T}v)(x) = \max_{a\in\Gamma(x)} B(x,a,v) = (Tv)(x) \qquad (v \in V,\ x \in \mathsf{X}).$$

Let us assume for simplicity that

1. $\mathcal{R}$ is contracting with modulus of contraction $\beta \in (0,1)$
2. V is closed in $\mathbb{R}^{\mathsf{X}}$

As a result, the value function v^* for $\mathcal{R}$ exists in V and is the unique fixed point of T in V

Ex. Show that, under the assumptions stated above, $\{\hat{T}_\varphi\}_{\varphi\in\Phi}$ and $\hat{T}$ are all contraction mappings

By the last exercise, the ADP $\mathcal{A}_M$ is max-stable

- globally stable operators are order stable
- the Bellman operator $\hat{T}$ has a fixed point).

Hence, by the result on slide 29,

- the value function $\hat{v}^*$ for $\mathcal{A}_M$ exists in V and
- is the unique fixed point of $\hat{T}$ in V.

But $\hat{T}$ and T agree on V

Hence $\hat{v}^* = v^*$

Thus, for RDPs, although the set of mixed strategies is larger than the set of pure strategies, the maximal lifetime value from each state is the same

Min-Optimality

Until now, our ADP theory has focused on maximization of lifetime values

Now we turn to minimization

Let $\mathcal{A} = (V, \{T_\sigma\})$ be a well-posed ADP and let $V_\Sigma := \{v_\sigma\}$ be the set of σ-value functions

We call $\sigma \in \Sigma$ **min-optimal** for $\mathcal{A}$ if v_σ is a least element of V_Σ

When V_Σ has a least element we denote it by $v^*_\downarrow$ and call it the **min-value function** generated by $\mathcal{A}$

A policy is called v-**min-greedy** if $T_\sigma\, v \preceq T_\tau\, v$ for all $\tau \in \Sigma$

Existence of a v-min-greedy policy for each $v \in V$ is guaranteed by the definition of ADPs

We say that $\mathcal{A}$ obeys **Bellman's principle of min-optimality** if

$$\sigma \in \Sigma \text{ is min-optimal for } \mathcal{A} \quad \Longleftrightarrow \quad \sigma \text{ is } v^*_\downarrow\text{-min-greedy}$$

Results analogous to the maximization theorem on slide 28 hold for the minimization case

Theorem. If $\mathcal{A}$ is min-stable, then

1. the min-value function $v^*_\downarrow$ generated by $\mathcal{A}$ exists in V
2. $v^*_\downarrow$ is the unique solution to the Bellman min-equation in V
3. $\mathcal{A}$ obeys Bellman's principle of min-optimality and
4. $\mathcal{A}$ has at least one min-optimal policy

If, in addition, Σ is finite, then min-HPI converges to $v^*_\downarrow$ in finitely many steps

An easy proof using order duality is given in Ch. 9

Additional Topics

Ch. 9 also discussed the following topics

- Isomorphic dynamic programs
- Subordinate dynamic programs
- Applications of these ideas

These advanced topics help us deduce optimality in one dynamic program from another closely related one

Detals are omitted from the lecture slides but provided in the book