Completely Abstract Dynamic Programming

Thomas J. Sargent and John Stachurski

July 21, 2023

Flow

- Motivation 1: uses of dynamic programming
- Motivation 2: the many forms of DP
- Some unifying optimality theory
- Discuss algorithms
- Connected related DPs
- Application: solving an Epstein–Zin problem

Dynamic programming has a **vast** array of applications

- robotics
- artificial intelligence
- computational biology
- management science
- engineering
- finance
- economics

Used daily to

- sequence DNA
- manage inventories
- test products
- control aircraft, route shipping
- optimize database operations
- recommend products, etc., etc.

Example. Nvidia Hopper GPUs hardwired to accelerate dynamic programming

Within economics and finance, dynamic programming is applied to

- unemployment and search
- monetary policy and fiscal policy
- asset pricing and portfolio choice
- firm investment
- firm entry and exit
- wealth dynamics
- commodity pricing
- sovereign default
- economic geography
- dynamic pricing, etc., etc.

Motivation

Consider

$$\max \sum_{t \geqslant 0} \beta^t u(C_t)$$

subject to

$$W_{t+1} = R(W_t - C_t) \quad \text{and} \quad 0 \leqslant C_t \leqslant W_t$$

Standard approach: set up the **Bellman operator**

$$(Tv)(w) = \max_{0 \leqslant c \leqslant w} \{u(c) + \beta v(R(w - c))\}$$

Value function iteration (VFI)

Under some conditions,

1. T is a contraction mapping
2. the unique fixed point of T is the value function $v_\top$
3. $v_\top$ can be approximated via $v_\top = \lim_{k \to \infty} T^k v$ for some v
4. optimal consumption at wealth w can be found by solving

$$c^* \in \operatorname*{argmax}_{0 \leqslant c \leqslant w} \{u(c) + \beta v_\top(R(w-c))\}$$

Howard policy iteration

Alternatively, we can use Howard policy iteration (HPI)

A **feasible policy** is a map $\sigma \colon \mathbb{R}_+ \to \mathbb{R}_+$ with

$$0 \leqslant \sigma(w) \leqslant w \quad \text{for all} \quad w \in \mathbb{R}_+$$

- given current wealth w, choose consumption $c = \sigma(w)$
- $\Sigma :=$ all feasible policies

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

$$\sigma(w) \in \operatorname*{argmax}_{0 \leqslant c \leqslant w} \{u(c) + \beta v(R(w-c))\}$$

Algorithm 1: Howard policy iteration

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

Computing Lifetime Value

The lifetime value v_σ of policy σ is the unique v that solves

$$v(w) = u(\sigma(w)) + \beta v(R(w - \sigma(w)))$$

To compute it we introduce the **policy operator**

$$(T_\sigma \, v)(w) = u(\sigma(w)) + \beta v(R(w - \sigma(w)))$$

Facts:

1. v_σ is the unique fixed point of T_σ
2. $T_\sigma^k v \to v_\sigma$ as $k \to \infty$ for all reasonable v

Under some conditions, HPI converges to an optimal policy

Example. Suppose we discretize wealth and consumption

Then HPI $\to$ an exact optimal policy in finitely many steps

Advantages

1. exact optimality
2. more parallelizable than VFI

(Smaller number of intensive steps)

See `opt_savings.ipynb` in

https://github.com/jstac/sandpit

Complications

What happens if we introduce **state-dependent discounting**?

$$(Tv)(w, z) = \max_{0 \leqslant c \leqslant w} \left\{ u(c) + \beta(z) \sum_{z'} v(R(w - c), z') Q(z, z') \right\}$$

- Is T still a contraction?
- Are the previous optimality results still valid?
- Does HPI converge?

What happens if we switch to the **expected value function**

$$g(w, z, c) := \sum_{z'} v(R(w-c), z') Q(z, z')$$

with "Bellman operator"

$$(Rg)(w, z, c) =$$

$$\sum_{z'} \max_{0 \leqslant c' \leqslant R(w-c)} \left\{ u(c') + \beta(z') g(R(w-c), z', c') \right\} Q(z, z')$$

Does R have the same properties as T?

What are the equivalent algorithms and do they converge?

And what happens if we introduce **Epstein–Zin preferences**?

$$(Tv)(w,z) =$$

$$\max_{0 \leqslant c \leqslant w} \left\{ c^{\alpha} + \beta(z) \left[\sum_{z'} v(R(w-c), z')^{\gamma} Q(z,z') \right]^{\alpha/\gamma} \right\}^{1/\alpha}$$

- Is T still a contraction?
- Are the previous optimality results still valid?
- Does HPI converge?

Or **risk-sensitive preferences**?

$$(Tv)(w,z) =$$

$$\max_{0 \leqslant c \leqslant w} \left\{ u(c) + \frac{\beta(z)}{\theta} \ln \left[\sum_{z'} \mathrm{e}^{\theta v(R(w-c), z')} Q(z, z') \right] \right\}$$

- Is T still a contraction?
- Are the previous optimality results still valid?
- Does HPI converge?

What about if we want to handle

- Q-learning?
- ambiguity?
- Q-learning in an Epstein–Zin framework?
- Q-learning $+$ robust control $+$ state-dependent discounting?
- expected value functions in a risk-sensitive framework?
- expected value functions in a risk-sensitive framework in continuous time?

Is there any unifying theory?

Or are all these problems too diverse?

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 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 and call it the **σ-value function**
- understand v_σ as representing lifetime value of σ

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

$$\max_{(A_t)_{t\geqslant 0}} \mathbb{E} \sum_{t\geqslant 0} \beta^t r(X_t, A_t) \quad \text{subject to} \quad A_t \in \Gamma(X_t)$$

when

- X_t takes values in finite set X (the state space),
- A_t takes values in finite set A (the action space),
- Γ is a correspondence from X to A (feasible correspondence),
- r is a reward function,
- $\beta \in (0, 1)$ is a discount factor, and
- $P(X_t, A_t, \cdot)$ provides transition probabilities

We define the set of **feasible policies** to be

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

Let $\mathbb{R}^{\mathsf{X}} = (\mathbb{R}^{\mathsf{X}}, \leqslant) =$ all $v \colon \mathsf{X} \to \mathbb{R}$ with

$$v \leqslant w \quad \iff \quad v(x) \leqslant w(x) \text{ for all } x \in \mathsf{X}$$

For $\sigma \in \Sigma$ and $v \in \mathbb{R}^{\mathsf{X}}$, let

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

The pair $(\mathbb{R}^{\mathsf{X}}, \{T_\sigma\})$ is an ADP

Let r_σ and P_σ be defined by

$$P_\sigma(x, x') := P(x, \sigma(x), x') \quad \text{and} \quad r_\sigma(x) := r(x, \sigma(x)).$$

The lifetime value of $\sigma \in \Sigma$ given $X_0 = x$ is

$$v_\sigma(x) = \mathbb{E} \sum_{t \geqslant 0} \beta^t r(X_t, \sigma(X_t)), \qquad (X_t)_{t \geqslant 0} \; P_\sigma\text{-Markov}, \; X_0 = x$$

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

Equivalently, v_σ is the unique solution to $v = r_\sigma + \beta P_\sigma \, v$

Equivalently, v_σ is the unique fixed point of $T_\sigma \, v = r_\sigma + \beta P_\sigma \, v$

Example. We can modify to handle **Epstein–Zin** preferences

Set

$$V = \text{ all positive functions in } \mathbb{R}^{\mathsf{X}}$$

and

$$(T_\sigma\, v)(x) = \{r(x, \sigma(x))^\alpha + \beta(x)\, [(Rv)(x, \sigma(x))]^\alpha\}^{1/\alpha}$$

where $r > 0$ and

$$(Rv)(x, a) := \left(\sum_{x'} v(x')^\gamma P(x, a, x') \right)^{1/\gamma}$$

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

What about

- Q-learning?
- ambiguity?
- Q-learning in an Epstein–Zin framework?
- Q-learning $+$ robust control $+$ state-dependent discounting?
- expected value functions in a risk-sensitive framework?
- MDPs in continuous time?

All these and more can be framed as ADPs

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

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

Example. In the MDP example we have

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

so σ is v-greedy iff

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

Bellman equation

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

We define the **Bellman operator** via

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

(if it exists)

Equivalently,

$$T_\top v = T_\sigma \, v \text{ when } \sigma \text{ is } v\text{-greedy}$$

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

Example. For the MDP,

$$(T_{\top} v)(x) = (T_\sigma \, v)(x) \text{ when } \sigma \text{ is } v\text{-greedy}$$

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

Hence the ADP Bellman equation is

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

And this is the same as the MDP Bellman equation

Example. In the Epstein–Zin case,

$$(T_\top v)(x) = \max_{\sigma \in \Sigma} \left\{ r(x, \sigma(x))^\alpha + \beta(x) \left[(Rv)(x, \sigma(x)) \right]^\alpha \right\}^{1/\alpha}$$

$$= \max_{a \in \Gamma(x)} \left\{ r(x, a)^\alpha + \beta(x) \left[(Rv)(x, a) \right]^\alpha \right\}^{1/\alpha}$$

Hence the ADP Bellman equation is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a)^\alpha + \beta(x) \left[(Rv)(x, a) \right]^\alpha \right\}^{1/\alpha}$$

And this is the standard Bellman equation for the EZ problem

Properties

We say that $\mathcal{A} = (V, \{T_\sigma\})$ 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 $\mathcal{A}$ is order stable, each $v \in V$ has at least one greedy policy, and $T_\top$ has at least one fixed point in V

Note: order stability is a regularity property — see the paper

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

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

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

We set $v_\top := \bigvee_\sigma v_\sigma$ and call $v_\top$ 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 $\mathcal{A}$ is max-stable, then

1. $v_\top$ exists in V
2. $v_\top$ is the unique solution to the Bellman equation in V
3. a policy is optimal if and only if it is $v_\top$-greedy
4. at least one optimal policy exists

If, in addition, Σ is finite, then HPI $\to v_\top$ in finitely many steps

Min-Optimality

Analogous results exist for minimization

The proof follows easily from

1. the max case
2. order duality

Subordinate ADPs

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

We say that $\hat{\mathcal{A}}$ is **subordinate** to $\mathcal{A}$ 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 $G_\top = \bigvee_\sigma G_\sigma$

Theorem. If

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

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

$$T_\top = G_\top \circ F \quad \text{and} \quad \hat{T}_\top = F \circ G_\top$$

while the value functions are related by

$$v_\top = G_\top \, \hat{v}_\top \quad \text{and} \quad \hat{v}_\top = F \, v_\top$$

Moreover,

1. if σ is optimal for $\mathcal{A}$, then σ is optimal for $\hat{\mathcal{A}}$, and
2. if $G_\sigma \, \hat{v}_\top = G_\top \, \hat{v}_\top$, then σ is optimal for $\mathcal{A}$

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

(Details in paper)

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

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

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 2: Solving $\mathcal{A}$ via $\mathcal{B}$

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

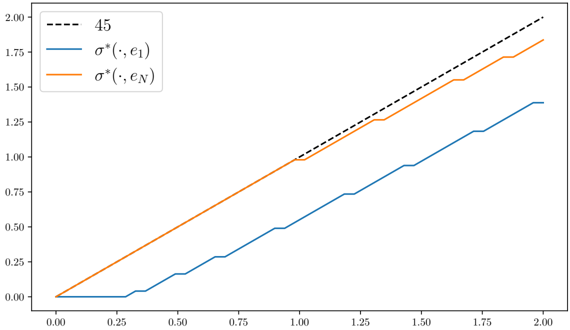

Figure: Optimal savings policy with Epstein–Zin preference

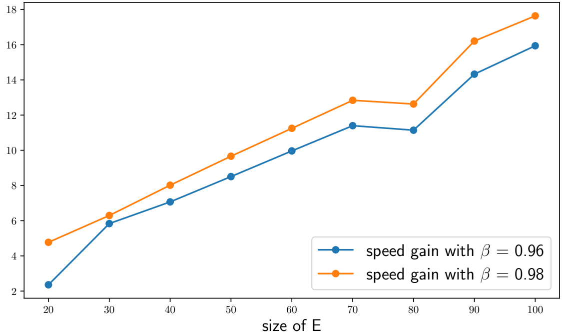

Figure: Speed gain from replacing $\mathcal{A}$ with subordinate model $\mathcal{B}$

For details of computations see

https://github.com/jstac/adps_public