Dynamic Programming

Chapter 8: Recursive Decision Processes

Thomas J. Sargent and John Stachurski

2024

Topics

1. Recursive Decision Problems (RDPs)
2. RDP optimality
3. Types of RDPs
4. Applications

The MDP framework is powerful, elegant, and broad

At the same time, researchers are pushing past the boundaries of MDPs

Recursive problems that do not fit the MDP framework include

- dynamic programs with nonlinear recursive preferences
- Some models of recursive equilibria in economic geography, production
- some dynamic programming problems with ambiguity or adversarial agents
- etc., etc.

Plan:

1. Construct a dynamic programming framework based on an abstraction of the Bellman equation
2. State optimality results in this framework
3. Connect with applications
4. Further abstract to an operator-theoretic framework (Ch. 9)
5. Use the operator-theoretic framework to complete all proofs (Ch. 9)

(This chapter builds on work by Eric Denardo, Loring Mitten, and Dimitri Bertsekas)

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**)
- v is used to evaluate future states
- $B(x, a, v) =$ total lifetime rewards given current state-action

Definition

The key primitives are

1. the correspondence Γ
2. a set V of "candidate value functions"
3. the function B that forms the r.h.s. of the Bellman equation

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

$$\mathcal{R} = (\Gamma, V, B)$$

where...

1. Γ is a nonempty correspondence from X to A called the **feasible correspondence**

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

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

- and the **feasible policies**

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

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

- A set of candidates for the value function

3. B is a map from $\mathsf{G} \times V$ to $\mathbb{R}$ called the **value aggregator** that satisfies

1. **monotonicity**:

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

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

2. **consistency**:

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

Example. An MDP (Γ, β, r, P) with state space X can be framed as an RDP (Γ, V, B) by setting $V = \mathbb{R}^{\mathsf{X}}$ and

$$B(x, a, v) = r(x, a) + \beta \sum_{x'} v(x') P(x, a, x') \tag{1}$$

- monotonicity and consistency conditions are trivial to check

Inserting (1) into

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

yields 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. An **optimal stopping** problem with

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

becomes an RDP (Γ, V, B) when $V = \mathbb{R}^{\mathsf{X}}$, $\Gamma(x) = \{0, 1\}$ and

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

Inserting B into

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

recovers (2)

Example. An MDP with **state-dependent discounting** and

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

becomes an RDP (Γ, V, B) when $V = \mathbb{R}^{\mathsf{X}}$ and

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

Inserting B into

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

recovers the Bellman equation at the top of the slide

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 **quantile preferences**, so that

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

where

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

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

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

we obtain an RDP with the same Bellman equation

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

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

for nonzero α, γ and $r \geqslant 0$

Setting $V = (0, \infty)^{\mathsf{X}}$ and

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

yields an RDP with the same Bellman equation

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

- $\mathsf{X} =$ the set of vertices
- $E =$ the set of edges
- $c(x, x') =$ cost of traversing edge $(x, x') \in E$
- the direct successors of x denoted by

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

Aim: find the minimum cost path from x to a specified vertex d

No discounting (so cannot use MDP theory)

The Bellman equation is

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

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

Let $\Gamma(x) = \mathcal{O}(x)$ 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)

Recall that the set of feasible policies is

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

choose $\sigma \in \Sigma \implies$

respond to state X_t with action $A_t := \sigma(X_t)$ at **all** $t \geqslant 0$

Ex. Given $v \in V$ and $x \in \mathsf{X}$, show that

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

Policy Operators

Fix $\sigma \in \Sigma$

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

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

Ex. Show that T_σ is an order-preserving self-map on V

Proof: Immediate from monotonicity and consistency

Policy Operators

Fix $\sigma \in \Sigma$

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

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

Ex. Show that 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 risk-sensitive MDP policy operator is

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

Lifetime value

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

If T_σ has a unique fixed point in V, then we

- denote this function by v_σ
- call it the **σ-value function**
- interpret v_σ as the lifetime value of following σ

Example. Let $\mathcal{R}$ be the RDP generated by an MDP, fix $\sigma \in \Sigma$ and recall that

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

Since $|\beta| < 1$, this operator has the unique fixed point

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

In other words,

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

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

Example. We just saw that an RDP generated by an MDP is always well-posed

Example. For the Epstein–Zin RDP,

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

Equivalently, with

$$A^\sigma_{\mathrm{CES}}(x, y) := \{ r(x, \sigma(x))^\alpha + \beta y^\alpha \}^{1/\alpha}$$

and

$$(R^\sigma_\gamma v)(x) := \left[\sum_{x' \in \mathsf{X}} v(x')^\gamma P(x, \sigma(x), x') \right]^{1/\gamma}$$

we have

$$T_\sigma = A^\sigma_{\mathrm{CES}} \circ R^\sigma_\gamma = \text{Epstein–Zin Koopmans operator}$$

In Ch. 7 we gave conditions under which such operators have unique fixed points in $V := (0, \infty)^{\mathsf{X}}$

- Later we give alternative conditions

Suppose these conditions hold at all $\sigma \in \Sigma$

Then

- the RDP is well-posed and
- the fixed point v_σ of T_σ is understood as lifetime value under Epstein–Zin preferences and the feasible policy σ

Stability

Let $\mathcal{R}$ be an RDP

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

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

Thus, for all σ,

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

Example. Let $\mathcal{R}$ be the optimal stopping RDP from slide 10

Given $\sigma \in \Sigma$, the policy operator is

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

We showed in Ch. 4 that T_σ is globally stable on $\mathbb{R}^{\mathsf{X}}$ for any choice of $\sigma \in \Sigma$

Hence $\mathcal{R}$ is a globally stable RDP

Example. The RDP generated by the MDP model from slide 9 is globally stable by similar reasoning

Recall the risk-sensitive RDP from slide 12, with

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

If

$$A^\sigma_{\mathrm{ADD}}(x, y) := r(x, \sigma(x)) + \beta y$$

and

$$(R^\sigma_\theta v)(x) := \frac{1}{\theta} \ln \left(\sum_{x'} \exp(\theta v(x')) P(x, \sigma(x), x') \right)$$

then

$$T_\sigma = A^\sigma_{\mathrm{ADD}} \circ R^\sigma_\theta = \text{risk-sensitive Koopmans operator}$$

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

Our results from Ch. 7 imply that T_σ is globally stable on $\mathbb{R}^{\mathsf{X}}$ for any choice of $\sigma \in \Sigma$

Hence $\mathcal{R}$ is a globally stable RDP

- v_σ represents lifetime value of $\sigma \in \Sigma$

Optimality

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

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

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

Example. For the Epstein–Zin RDP in slide 14, the Bellman operator is

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

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

Ex. Given RDP $\mathcal{R} = (\Gamma, V, B)$ with policy operators $\{T_\sigma\}$ and Bellman operator T, show that, for each $v \in V$,

$$Tv = \bigvee_{\sigma} T_\sigma v := \bigvee_{\sigma \in \Sigma} (T_\sigma v)$$

Proof: For any $\sigma \in \Sigma$ and $x \in \mathrm{X}$, we have

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

$$= \max_{\sigma \in \Sigma} B(x, \sigma(x), v)$$

$$= \max_{\sigma \in \Sigma} (T_\sigma v)(x)$$

Since x was chosen arbitrarily, we have $Tv = \bigvee_{\sigma \in \Sigma} T_\sigma v$

Ex. Given RDP $\mathcal{R} = (\Gamma, V, B)$ with policy operators $\{T_\sigma\}$ and Bellman operator T, show that, for each $v \in V$,

$$Tv = \bigvee_{\sigma} T_\sigma \, v := \bigvee_{\sigma \in \Sigma} (T_\sigma \, v)$$

Proof: For any $\sigma \in \Sigma$ and $x \in \mathsf{X}$, we have

$$\begin{aligned}(Tv)(x) &= \max_{a \in \Gamma(x)} B(x, a, v) \\ &= \max_{\sigma \in \Sigma} B(x, \sigma(x), v) \\ &= \max_{\sigma \in \Sigma} (T_\sigma \, v)(x)\end{aligned}$$

Since x was chosen arbitrarily, we have $Tv = \bigvee_{\sigma \in \Sigma} T_\sigma \, v$

Ex. In the same setting as the last exercise, show that σ is v-greedy if and only if

$$Tv = T_\sigma v$$

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

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

This is equivalent to

$$B(x, \sigma(x), v) = \max_{a \in \Gamma(x)} B(x, a, v) \quad \text{for all } x \in \mathsf{X}$$

Hence σ is v-greedy if and only if $T_\sigma v = Tv$, as claimed

Optimality

Let $\mathcal{R}$ 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 $\sigma \in \Sigma$

$v_0 \leftarrow v_\sigma$ and $k \leftarrow 0$

repeat

  $\sigma_k \leftarrow$ a v_k-greedy policy

  $v_{k+1} \leftarrow$ the fixed point of T_{σ_k}

  if $v_{k+1} = v_k$ **then** break

  $k \leftarrow k + 1$

return σ_k

Let $\mathcal{R}$ be an RDP

Key question:

What assumptions to we need for optimality?

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

- We cannot maximize lifetime value over policies unless this lifetime value is well-defined

This is the minimum requirement

What else?

The next slide shows that global stability is enough

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

Theorem. If $\mathcal{R}$ 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. Howard policy iteration returns an optimal policy in finitely many steps

Comments on the theorem

Traditional treatments build optimality theory around contractivity

- sufficient (implies global stability of $\mathcal{R}$)
- but not necessary

There are many ways to prove uniqueness and stability of fixed points, including

- Du's theorem
- the spectral methods we used in Ch. 6, etc.

Useful in settings where contractivity fails

Notice also that assumptions are on the policy operators rather than the Bellman operator — advantageous because the policy operators are simpler (no max)

Proof / algorithms

A direct proof is not hard — but we will prove a more general result in Ch. 9

- embed RDPs in an abstract operator setting
- prove an optimality result in that setting
- show that globally stable RDPs are a special case

For now let's discuss algorithms

The main one we need to discuss is optimistic policy iteration

- includes VFI as a special case $(m = 1)$

Convergence of OPI

Algorithm 1: Optimistic policy iteration for RDPs

input $\sigma \in \Sigma$ and set $v_0 \leftarrow v_\sigma$

input τ, a tolerance level for error

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

$k \leftarrow 0$ and $\varepsilon \leftarrow \tau + 1$

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

$\sigma_k \leftarrow$ a v_k-greedy policy

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

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

$k \leftarrow k + 1$

end

return v_k, σ_k

Theorem. If $\mathcal{R}$ is globally stable, then the sequence (v_k) generated by OPI converges to v^*

Proof: See the book (Ch. 8)

In the theorem,

$$v_k = \text{lifetime value of } k\text{-th greedy policy generated by OPI}$$

Since $v_k \to v^*$, lifetime value of the policy sequence produced by OPI $\to$ the lifetime value of an optimal policy

Topologically conjugate RDPs

Sometimes RDP models can be simplified by transformations

To begin, let $\mathcal{R} = (\Gamma, V, B)$ and $\hat{\mathcal{R}} = (\Gamma, \hat{V}, \hat{B})$ be two RDPs

We consider settings where

$$V = \mathbb{M}^{\mathsf{X}} \quad \text{and} \quad \hat{V} = \hat{\mathbb{M}}^{\mathsf{X}} \quad \text{where } \mathbb{M}, \hat{\mathbb{M}} \subset \mathbb{R},$$

We call $\mathcal{R}$ and $\hat{R}$ **topologically conjugate** under φ if φ is a homeomorphism φ from $\mathbb{M}$ to $\hat{\mathbb{M}}$ and

$$B(x, a, v) = \varphi^{-1}[\hat{B}(x, a, \varphi \circ v)] \quad \text{for all } v \in V \text{ and } (x, a) \in \mathsf{G}$$

Ex. Prove: If φ is a homeomorphism from $\mathbb{M}$ to $\hat{\mathbb{M}}$ and $\Phi v := \varphi \circ v$, then Φ is a homeomorphism from V to $\hat{V}$.

Proposition. If $\mathcal{R}$ and $\hat{\mathcal{R}}$ are topologically conjugate, then $\mathcal{R}$ is globally stable if and only if $\hat{\mathcal{R}}$ is globally stable

Proof: $\Phi v := \varphi \circ v$ is a homeomorphism from V to $\hat{V}$

Moreover, for any $\sigma \in \Sigma$, the respective policy operators T_σ and $\hat{T}_\sigma$ are linked by

$$(T_\sigma\, v)(x) = B(x, \sigma(x), v) = \varphi^{-1}[\hat{B}(x, \sigma(x), \varphi \circ v)]$$

$$= \varphi^{-1}[(\hat{T}_\sigma \varphi \circ v)(x)]$$

This shows that $T_\sigma = \Phi^{-1} \circ \hat{T}_\sigma \circ \Phi$ on V

Hence (V, T_σ) and $(\hat{V}, \hat{T}_\sigma)$ are topologically conjugate dynamical systems

Hence T_σ is globally stable if and only if $\hat{T}_\sigma$ is globally stable

Hence $\mathcal{R}$ is globally stable if and only if $\hat{\mathcal{R}}$ is globally stable

Application: Epstein–Zin RDPs

Consider the Epstein–Zin RDP $\mathcal{R} := (\Gamma, B, V)$ from slide 14, where $V = (0, \infty)^{\mathsf{X}}$ and

$$B(x, a, v) = \left\{ r(x, a) + \beta \left(\sum_{x'} v(x')^{\gamma} P(x, a, x') \right)^{\alpha/\gamma} \right\}^{1/\alpha} \tag{3}$$

Proposition. If $P(x, \sigma(x), x')$ is irreducible for all $\sigma \in \Sigma$ and $r \gg 0$, then $\mathcal{R}$ is globally stable

To prove the proposition, we set up a simpler and more tractable model

Let $\hat{\mathcal{R}} := (\Gamma, \hat{B}, V)$ with

$$\hat{B}(x, a, v) = B\left(x, a, v^{1/\gamma}\right)^{\gamma} \tag{4}$$

Since $\varphi(m) = m^{\gamma}$ is a homeomorphism from V to itself, $\mathcal{R}$ and $\hat{\mathcal{R}}$ are topologically conjugate RDPs

Notice that $\hat{B}$ can also be expressed as

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

where $\theta := \gamma/\alpha$

The value of of introducing $\hat{\mathcal{R}}$ comes from the fact that $\hat{\mathcal{R}}$ is easier to work with than $\mathcal{R}$

Each policy operator $\hat{T}_\sigma$ associated with $\hat{\mathcal{R}}$ takes the form

$$(\hat{T}_\sigma\, v)(x) = \left\{ r(x, \sigma(x)) + \beta \left(\sum_{x'} w(x') P(x, \sigma(x), x') \right)^{1/\theta} \right\}^{\theta}$$

Each such $\hat{T}_\sigma$ is an Epstein–Zin Koopmans operator of the form we have already analyzed

We know it is globally stable under the stated assumptions

Hence $\hat{\mathcal{R}}$ is a globally stable RDP.

By topologically conjugacy, $\mathcal{R}$ is globally stable if and only if $\hat{\mathcal{R}}$ is globally stable

Hence $\mathcal{R}$ is globally stable, as was to be shown

Types of RDPs

The optimality properties require stability

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

Contracting RDPs

We call $\mathcal{R}$ **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 $v, w \in V$

Example. The optimal stopping RDP from 10 is contracting with modulus β

Indeed, an application of the triangle inequality gives

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

$$\leqslant \beta \|v - w\|_\infty$$

Ex. Show that each RDP generated by an MDP is contracting

Proposition. If $\mathcal{R}$ is contracting with modulus β, then T and $\{T_\sigma\}_{\sigma\in\Sigma}$ are all contractions of modulus β on V under $\|\cdot\|_\infty$

If, in addition, V is closed in $\mathbb{R}^{\mathsf{X}}$, then $\mathcal{R}$ is globally stable and all optimality results on slide 36 apply

Proof: Let $\mathcal{R} = (\Gamma, V, B)$ be contracting with modulus β

Fixing $\sigma \in \Sigma$, v, w in V and $x \in \mathsf{X}$,

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

$$\leqslant \beta \|v - w\|_\infty$$

Now max over x

We have shown that

$$\mathcal{R} \text{ contracting w. mod } \beta \implies \text{all } T_\sigma \text{ contracting w. mod } \beta$$

Hence

$$T := \bigvee_\sigma T_\sigma \quad \text{also contracting w. mod } \beta$$

- upper envelopes of contractions are contractions (Ch. 1)

If V is closed, then we can apply Banach's theorem

$$T_\sigma \text{ all globally stable} \implies \mathcal{R} \text{ is globally stable}$$

Error bounds

Since contracting RDPs are globally stable, so VFI converges to v^*

But this result is asymptotic — what happens in finite time?

Proposition. Let $\mathcal{R}$ be a contracting RDP with mod β

Fix $v \in V$ and let $v_k = T^k v$

If σ is v_k-greedy, then

$$\|v^* - v_\sigma\|_\infty \leqslant \frac{2\beta}{1-\beta}\|v_k - v_{k-1}\|_\infty \quad \text{for all } k \in \mathbb{N}$$

A sufficient condition for contractivity

We say (Γ, V, B) satisfies **Blackwell's condition** if

1. $v \in V$ implies $v + \lambda \mathbb{1} \in V$ for all $\lambda \geqslant 0$ and
2. $\exists$ a $\beta \in [0, 1)$ such that

$$B(x, a, v + \lambda) \leqslant B(x, a, v) + \beta\lambda$$

for all $(x, a) \in \mathsf{G}$, $v \in V$ and $\lambda \in \mathbb{R}_+$

Ex. Prove the following: If $\mathcal{R} = (\Gamma, B, V)$ satisfies Blackwell's condition, then $\mathcal{R}$ is contracting with modulus β

Proof: Let $\mathcal{R} = (\Gamma, V, B)$ satisfy Blackwell's condition

Fix $v, w \in V$ and observe that $v = w + v - w \leqslant w + \|v - w\|_\infty$

By monotonicity of B and Blackwell's condition, we have

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

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

Reversing the roles of v and w yields

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

Since $\beta < 1$, the RDP $\mathcal{R}$ is contracting

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 can embed into the RDP framework by taking

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

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

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

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

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

Proof: We saw in Ch. 7 that R_τ is constant-subadditive

Hence

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

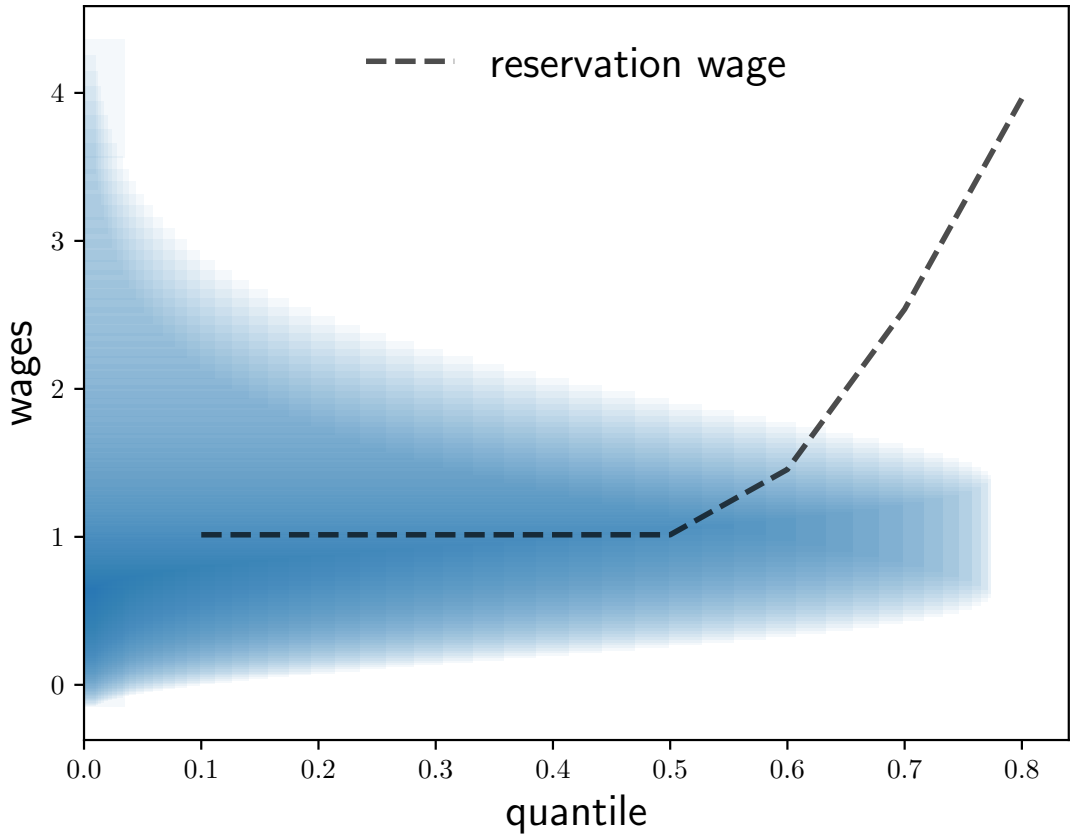
Since V is closed, $\mathcal{R}$ is globally stable

All optimality properties on slide 36 apply

Let's look at the reservation wage for a range of τ values

- computed using optimistic policy iteration
- taking the smallest $w \in \mathsf{W}$ such that $\sigma^*(w) = 1$

We also show the stationary distribution of P, tilted 90 degrees


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

Main message: the reservation wage rises in τ

- higher τ focuses attention on the right tail of the distribution
- increases appetite for risk

Leads to a higher reservation wage

- reluctance to accept a given current offer
- worker prefers to role the dice

Application: Risk-sensitive RDPs

Consider the risk-sensitive preference RDP in slide 12, where $V = \mathbb{R}^{\mathsf{X}}$ and

$$B(x, a, v) = r(x, a) + \beta(R_a v)(x)$$

where

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

Proposition. If $\beta < 1$, then (Γ, V, B) is contracting

Proof. We saw in Ch 7 that the entropic certainty equivalent operator is constant-subadditive

Hence

$$\begin{aligned} B(x, a, v + \lambda) &= r(x, a) + \beta[R_a(v + \lambda)](x) \\ &\leqslant r(x, a) + \beta(R_a v)(x) + \beta\lambda \\ &= B(x, a, v) + \beta\lambda \end{aligned}$$

This shows that Blackwell's condition holds, so (Γ, V, B) is contracting

Risk-sensitive job search

We consider a job search problem where future wage outcomes are evaluated via risk-sensitive expectations

The associated Bellman operator is

$$(Tv)(w) = \max\left\{\frac{w}{1-\beta},\, c + \frac{\beta}{\theta}\ln\left[\sum_{w'}\exp(\theta v(w'))P(w,w')\right]\right\}$$

The parameter θ controls attitude to risk

We can represent the problem as an RDP with

1. feasible correspondence $\Gamma(w) = \{0, 1\}$
2. value space $V := \mathbb{R}^{\mathsf{W}}$
3. value aggregator

$$B(w, a, v) =$$

$$a\frac{w}{1-\beta} + (1-a)\left\{c + \frac{\beta}{\theta}\ln\left[\sum_{w'}\exp(\theta v(w'))P(w, w')\right]\right\}$$

Ex. Show that $\mathcal{R} := (\Gamma, V, B)$ is globally stable

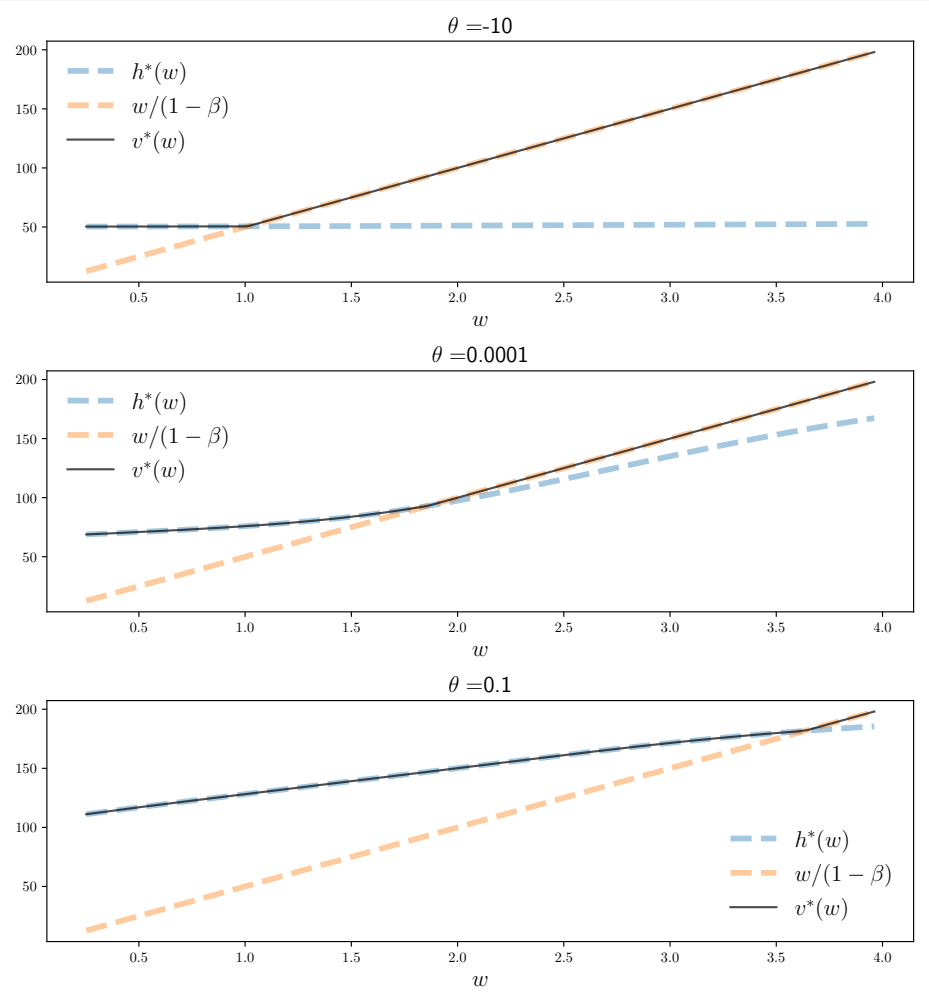


Figure: Job search with risk-sensitive preferences

Eventually Contracting RDPs

We call $\mathcal{R}$ **eventually contracting** if $\exists$ a map L from G $\times$ X to $\mathbb{R}_+$ with

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

for all $(x,a) \in$ G and $v, w \in V$, and moreover,

$$\sigma \in \Sigma \implies \rho(L_\sigma) < 1 \quad \text{where} \quad L_\sigma(x,x') := L(x,\sigma(x),x')$$

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

Proof: Let $\mathcal{R}$ be as stated and fix $\sigma \in \Sigma$

We need to show that T_σ is globally stable on V

Given $v, w \in V$ and $x \in$ X, we have

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

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

In vector notation, $|T_\sigma\, v - T_\sigma\, w| \leqslant L_\sigma\, |v - w|$

Since $\rho(L_\sigma) < 1$ and V is closed, T_σ is globally stable (see Ch. 6)

In Ch. 6 we left the main optimality proof for MDPs with state-dependent discounting till later

Now is that later

We work with the RDP $\mathcal{R} = (\Gamma, V, B)$ on slide 11, where $V = \mathbb{R}^{\mathsf{X}}$ and

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

Given $\sigma \in \Sigma$, we set

$$L_\sigma(x, x') := \beta(x, \sigma(x), x')P(x, \sigma(x), x')$$

Recall Assumption SD from Ch. 6: $\rho(L_\sigma) < 1$ for all $\sigma \in \Sigma$

In view of the theorem on slide 67, to prove that $\mathcal{R}$ obeys the main optimality results from Ch. 6, it suffices to show that

Proposition. If Assumption SD holds, then $\mathcal{R}$ is eventually contracting

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

$$L(x, a, x') := \beta(x, a, x')P(x, a, x')$$

For given $v, w \in V$, is is easy to check that

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

Recall that

$$L_\sigma(x, x') = \beta(x, \sigma(x), x')P(x, \sigma(x), x')$$

Since V is closed, we only need to show that $\rho(L_\sigma) < 1$ for all σ

But this is imposed by SD

Hence $\mathcal{R}$ is an eventually contracting RDP

Concave RDPs

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

1. $V = [v_1, v_2]$
2. $v \mapsto B(x, a, v)$ is concave for all $(x, a) \in \mathsf{G}$
3. there exists a $\delta > 0$ such that

$$B(x, a, v_1) \geqslant v_1(x) + \delta[v_2(x) - v_1(x)] \text{ for all } (x, a) \in \mathsf{G} \quad (5)$$

For (3) it suffices that

$$B(x, a, v_1) > v_1(x) \text{ for all } (x, a) \in \mathsf{G}$$

(Why?)

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

Proof: Given $\sigma \in \Sigma$,

- T_σ is order-preserving (by the monotonicity property of RDPs)
- T_σ is concave (by prop. 2)
- $T_\sigma\, v_2 \leqslant v_2$ (because T_σ maps V to itself)
- $T_\sigma\, v_1 \geqslant v_1 + \delta(v_2 - v_1)$ (by prop. 3)

Hence, by Du's theorem, T_σ is globally stable on V

Hence $\mathcal{R}$ is a globally stable RDP — and all optimality properties on slide 36 apply

Adversarial Agents

Some problems in economics, AI etc. assume decisions emerge from a dynamic two-person zero sum game in which preferences are misaligned

This can lead to a DP where the Bellman equation takes the form

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

Here

- $B(x, a, d, v) =$ lifetime value for the controller given her current action a and her adversary's action d
- The controller chooses $a \in \Gamma(x)$ knowing her adversary will then choose $d \in D(x, a)$ to minimize her lifetime value

We introduce the following assumptions:

(a) If $v, w \in \mathbb{R}^{\mathsf{X}}$ with $v \leqslant w$, then

$$B(x, a, d, v) \leqslant B(x, a, d, w) \quad \text{for all } x, a, d$$

(b) There exists a $v_1 \in \mathbb{R}^{\mathsf{X}}$ and $\varepsilon > 0$ such that

$$v_1(x) + \varepsilon \leqslant B(x, a, d, v_1) \quad \text{for all } x, a, d$$

(c) There exists a $v_2 \in \mathbb{R}^{\mathsf{X}}$ such that $v_1 \leqslant v_2$ and

$$B(x, a, d, v_2) \leqslant v_2(x) \quad \text{for all } x, a, d$$

(d) If $\lambda \in [0, 1]$ and $v, w \in \mathbb{R}^{\mathsf{X}}$, then

$$v \mapsto B(x, a, d, v) \text{ is concave} \quad \text{for all } x, a, d$$

To analyze the decision maker's problem, we set

$$V := [v_1, v_2]$$

and

$$\hat{B}(x, a, v) := \inf_{d \in D(x,a)} B(x, a, d, v)$$

We consider $\mathcal{R} = (\Gamma, V, \hat{B})$

Proposition. If conditions (a)–(d) hold, then $\mathcal{R}$ is a concave RDP

To prove concavity of $\mathcal{R}$, we must show that, for fixed $\lambda \in [0,1]$ and $v, w \in V$,

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

This holds because, given $(x,a) \in \mathsf{G}$, $\lambda \in [0,1]$ and $v, w \in V$,

$$\begin{aligned}
&\hat{B}(x,a,\lambda v + (1-\lambda)w) \\
&\quad = \inf_{d \in D(x,a)} B(x,a,d,\lambda v + (1-\lambda)w) \\
&\quad \geqslant \inf_{d \in D(x,a)} [\lambda B(x,a,d,v) + (1-\lambda)B(x,a,d,w)] \\
&\quad \geqslant \lambda \inf_{d \in D(x,a)} B(x,a,d,v) + (1-\lambda) \inf_{d \in D(x,a)} B(x,a,d,w)
\end{aligned}$$

Minor remaining details of the proof can be found in the book

We conclude that, under (a)–(d), the decision maker's problem is a globally stable RDP

Hence the fundamental optimality properties in slide 36 all hold

- Bellman's principle of optimality holds
- An optimal policy exists
- HPI and OPI converge
- etc.

A Perturbed MDP Problem

As an example of the preceding result, consider a perturbed MDP with Bellman equation

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

- the choice $d \in D(x,a)$ is made an the adversary
- $P(x,a,d,\cdot)$ is a distribution over X for each feasible (x,a,d)
- $\Gamma(x)$ and $D(x,a)$ are nonempty for all $(x,a) \in$ G

We set

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

To construct V we let $r_1 := \min r$ and $r_2 := \max r$, and set

$$V := [v_1, v_2]$$

where

$$v_1 := \frac{r_1 - \varepsilon}{1-\beta} \quad \text{and} \quad v_2 := \frac{r_2}{1-\beta}$$

Ex. Prove: For v_1, v_2 as above, conditions (b)–(c) on slide 75 hold

Proof: Regarding (b), note that

$$v_1 = \frac{r_1 - \varepsilon}{1 - \beta}$$

is constant

Hence, at fixed $(x, a) \in$ G and $d \in D(x, a)$, we have

$$B(x, a, d, v_1) = r(x, a) + \beta \frac{r_1 - \varepsilon}{1 - \beta} \geqslant r_1 + \beta \frac{r_1 - \varepsilon}{1 - \beta}$$

$$= \frac{r_1 - \beta r_1 + \beta r_1 - \beta \varepsilon}{1 - \beta} = v_1 + \varepsilon$$

Hence (b) is confirmed

Checking (c) is left to the reader

Proposition. The perturbed MDP model $\mathcal{R} := (\Gamma, V, \hat{B})$ is a concave RDP

Proof: It suffices to show that $\mathcal{R}$ obeys conditions (a)–(d) on slide 75

Conditions (b) and (c) were established in the preceding exercise

Condition (a) and (d) are elementary in this setting

Hence $\mathcal{R}$ is concave

It follows that $\mathcal{R}$ is globally stable and all optimality results on slide 36 apply

Ambiguity and Robustness

Until now we have considered agents facing decision problems where outcomes are uncertain but probabilities are known

Example. In job search, the worker

- does not know the next period wage offer
- but does know its distribution

Typically, the assumption that the decision maker knows all probability distributions that govern outcomes under different actions is debatable

Now we study lifetime valuations in settings of **Knightian uncertainty**

- outcome distributions are themselves unknown
- Knightian uncertainty also called **ambiguity**

We start in an MDP setting with a controller who distrusts her specification of the stochastic kernel P

She knows only that P belongs to some class of stochastic kernels from G $\times$ X to X

- a way of modeling Knightian uncertainty

This can lead to aggregators of the form

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

As usual, r maps G to $\mathbb{R}$ and $\beta \in (0, 1)$

Using this aggregator, the decision maker can construct a policy that is robust to her distrust of P

The set $\mathcal{P}$ of stochastic kernels is entirely arbitrary

We take V is as defined in slide 80 and set $\mathcal{R} = (\Gamma, V, B)$

The next result shows that $\mathcal{R}$ obeys the fundamental optimality results

Proposition. $\mathcal{R}$ is a concave RDP

Proof: We rewrite B as

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

It is now clear that $\mathcal{R}$ is a special case of the perturbed MDP model from slide 80

Concavity follows from the result on slide 82

Example. Consider the simple job search problem from Ch. 1

- worker is currently unemployed
- wage offer sequence is IID

Suppose now that the worker

- does not know the exact offer distribution
- believes that it lies in some subset $\mathcal{P}$ of $\mathcal{D}(\mathsf{W})$

She can seek a decision rule that is robust to worst-case beliefs by optimizing with value aggregator

$$B(w, a, v) = a\frac{w}{1-\beta} + (1-a)\inf_{\varphi \in \mathcal{P}} \sum_{w'} v(x')\varphi(w')$$

Robustness and Adversarial Agents

A more general way to implement robustness is via the aggregator

$$B(x, a, v) = r(x, a) +$$

$$\beta \inf_{P \in \mathcal{P}(x,a)} \left\{ \sum_{x'} v(x') P(x, a, x') + d(P(x, a, \cdot), \bar{P}(x, a, \cdot)) \right\} \quad (6)$$

In this set up, $\mathcal{P}(x, a)$ is often large, weakening the constraint on P

At the same time, the term $d(P(x, a, \cdot), \bar{P}(x, a, \cdot))$ penalizes deviation between some baseline specification $\bar{P}$

One interpretation: the decision maker

- begins with a baseline specification of dynamics
- but lacks confidence in its accuracy

In her desire to choose a robust policy, she imagines herself playing against an adversarial agent

Her adversary

- can choose transition kernels that deviate from the baseline
- but the presence of the penalty term means that extreme deviations are curbed

Suppose we define

$$\hat{r}(x,a) = r(x,a) + d(P(x,a,\cdot), \bar{P}(x,a,\cdot)),$$

Now (6) can be expressed as

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

This is a special case of the perturbed MDP model

Hence the same optimality theory applies

Connection to Risk-Sensitive Preferences

One common specification of deviation between probability distributions is **Kullback–Liebler divergence** (KL divergence), which is defined by

$$d_{KL}(q \,|\, p) := \sum_x q(x) \ln \left(\frac{q(x)}{p(x)} \right) \quad \text{for } q, p \in \mathcal{D}(\mathsf{X})$$

It is assumed here that $q \prec_{\text{ac}} p$, which means that $q(x) = 0$ whenever $p(x) = 0$

We note for future reference that d_{KL} obeys the **duality formula for variational inference** given by

$$\ln \sum_x \exp(h(x)) p(x) = \sup_{q \prec_{\text{ac}} p} \left\{ \sum_x h(x) q(x) - d_{KL}(q \,|\, p) \right\}$$

Under KL divergence, there is a tight relationship between robust control and risk-sensitive preferences

To illustrate, we fix $\theta < 0$ and set $d_\theta := -(1/\theta) d_{KL}$, so that d_θ is a simple positive rescaling of the Kullback–Leibler divergence

Using d_θ in (6) leads to

$$B(x, a, v) = r(x, a) + \beta \inf_{P \in \mathcal{P}(x,a)} \left\{ \sum_{x'} v(x') P(x, a, x') + d_\theta(P(x, a, \cdot) \,|\, \bar{P}(x, a, \cdot)) \right\}$$

The constraint set $\mathcal{P}(x, a)$ is all $P \in \mathcal{M}(\mathbb{R}^{\mathsf{X}})$ such that $P(x, a, \cdot) \prec_{\text{ac}} \bar{P}(x, a, \cdot)$

If we multiply both sides of the variational formula by $(1/\theta)$ and set $h = \theta v$ we get

$$\frac{1}{\theta} \ln \sum_x \exp(\theta v(x)) p(x) = \inf_{q \prec_{\mathrm{ac}} p} \left\{ \sum_x v(x) q(x) - \frac{1}{\theta} d_{KL}(q \,|\, p) \right\}$$

This allows us to rewrite B as

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

Hence, for this choice of deviation, the robust control aggregator (6) reduces to the risk-sensitive aggregator under the baseline transition kernel