Markov Decision Processes with Applications to Finance

Nicole Bäuerle

KIT

Lunteren, January 2016

Karlsruhe Institute of Technology

Outline

- Markov Decision Processes with Finite Time Horizon
  - Definition
  - Basic Results
  - Financial Applications
- Markov Decision Processes with Infinite Time Horizon
  - Definition
  - Basic Results
  - Financial Applications
- Extensions and Related Problems

Finite Horizon Problem

Markov Decision Processes (MDPs): Motivation

Let (X_n) be a Markov process (in discrete time) with

- state space E,
- transition kernel $Q_n(\cdot|x)$.

Markov Decision Processes (MDPs): Motivation

Let (X_n) be a Markov process (in discrete time) with

- state space E,
- transition kernel $Q_n(\cdot|x)$.

Let (X_n) be a controlled Markov process with

- state space E, action space A,
- admissible state-action pairs $D_n \subset E \times A$,
- transition kernel $Q_n(\cdot|x, a)$.

A decision A_n at time n is in general $\sigma(X_1, \dots, X_n)$-measurable.
However, Markovian structure implies $A_n = f_n(X_n)$ is sufficient.

MDPs: Formal Definition

Definition

A *Markov Decision Model* with planning horizon $N \in \mathbb{N}$ consists of a set of data $(E, A, D_n, Q_n, r_n, g_N)$ with the following meaning for $n = 0, 1, \dots, N-1$:

- E is the *state space*,
- A is the *action space*,
- $D_n \subset E \times A$ admissible state-action combinations at time n,
- $Q_n(\cdot|x, a)$ stochastic transition kernel at time n,
- $r_n : D_n \to \mathbb{R}$ one-stage reward at time n,
- $g_N : E \to \mathbb{R}$ terminal reward at time N.

Policies

- A decision rule at time n is a measurable mapping $f_n : E \to A$ such that $f_n(x) \in D_n(x)$ for all $x \in E$.
- A policy is given by $\pi = (f_0, f_1, \dots, f_{N-1})$ a sequence of decision rules.

Optimization Problem

For $n = 0, 1, \dots, N$, $\pi = (f_0, \dots, f_{N-1})$ define the value functions

$$\begin{aligned} V_{n\pi}(x) &:= \mathbb{E}^{\pi}_{nx}\left[\sum_{k=n}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\right], \\ V_n(x) &:= \sup_{\pi} V_{n\pi}(x), \quad x \in E. \end{aligned}$$

A policy π is called *optimal* if $V_{0\pi}(x) = V_0(x)$ for all $x \in E$.

Optimization Problem

For $n = 0, 1, \dots, N$, $\pi = (f_0, \dots, f_{N-1})$ define the value functions

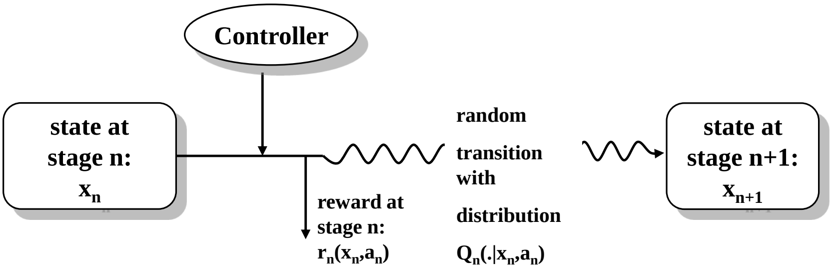
$$\begin{aligned} V_{n\pi}(x) &:= \mathbb{E}_{nx}^{\pi}\left[\sum_{k=n}^{N-1} r_k(X_k, f_k(X_k)) + g_N(X_N)\right], \\ V_n(x) &:= \sup_{\pi} V_{n\pi}(x), \quad x \in E. \end{aligned}$$

A policy π is called *optimal* if $V_{0\pi}(x) = V_0(x)$ for all $x \in E$.

Integrability Assumption (A_N): For $n = 0, 1, \dots, N$

$$\sup_{\pi} \mathbb{E}_{nx}^{\pi}\left[\sum_{k=n}^{N-1} r_k^+(X_k, f_k(X_k)) + g_N^+(X_N)\right] < \infty, \quad x \in E.$$

General evolution of a Markov Decision Process



VIPs of MDPs

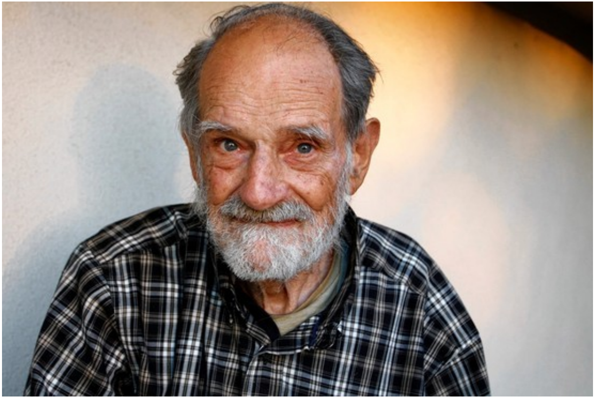


Abbildung: Lloyd Shapley (1923 -)

VIPs of MDPs

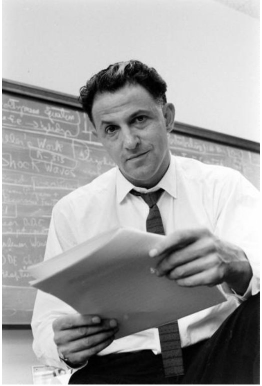


Abbildung: Richard Bellman (1920 - 1984)

VIPs of MDPs

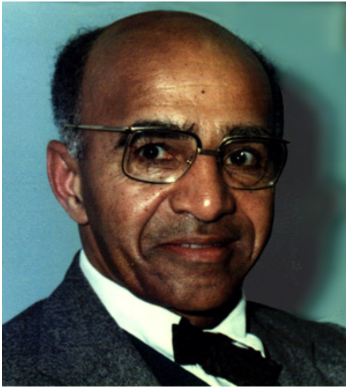


Abbildung: David Blackwell (1912 - 2010)

Literature - Textbooks on MDPs

- Shapley (1953)
- Bellman (1957, Reprint 2003)
- Howard (1960)
- Bertsekas and Shreve (1978)
- Puterman (1994)
- Hernández-Lerma and Lasserre (1996)
- Bertsekas (2001, 2005)
- Feinberg and Shwartz (2002)
- Powell (2007)
- B and Rieder (2011)

Notation

Let $\mathbb{M}(E) := \{v : E \to [-\infty, \infty) \mid v \text{ is measurable}\}$ and define the following operators for $v \in \mathbb{M}(E)$:

Definition

a) $(L_n v)(x,a) := r_n(x,a) + \int v(x')Q_n(dx'|x,a), \ (x,a) \in D_n,$

b) $(T_{nf} v)(x) := (L_n v)(x, f(x)), \ x \in E,$

c) $(T_n v)(x) := \sup_{a \in D_n(x)} (L_n v)(x,a)$. Note $T_n v \notin \mathbb{M}(E)$.

A decision rule f_n is called *maximizer* of v at time n if $T_{nf_n} v = T_n v$.

Theorem (Reward Iteration)

For a policy $\pi = (f_0, \dots, f_{N-1})$ *and* $n = 0, 1, \dots, N-1$:

a) $V_{N\pi} = g_N$ *and* $V_{n\pi} = T_{nf_n} V_{n+1,\pi}$,

b) $V_{n\pi} = T_{nf_n} \dots T_{N-1 f_{N-1}} g_N$.

Theorem (Reward Iteration)

For a policy $\pi = (f_0, \dots, f_{N-1})$ *and* $n = 0, 1, \dots, N-1$*:*

a) $V_{N\pi} = g_N$ *and* $V_{n\pi} = T_{nf_n} V_{n+1,\pi}$,

b) $V_{n\pi} = T_{nf_n} \dots T_{N-1 f_{N-1}} g_N$.

Theorem (Verification Theorem)

Let $(v_n) \subset \mathbb{M}(E)$ *be a solution of the Bellman equation:* $v_n = T_n v_{n+1}, \; v_N = g_N$. *Then it holds:*

a) $v_n \geq V_n$ *for* $n = 0, 1, \dots, N$.

b) *If* f_n^* *is a maximizer of* v_{n+1} *for* $n = 0, 1, \dots, N-1$*, then* $v_n = V_n$ *and* $\pi^* = (f_0^*, f_1^*, \dots, f_{N-1}^*)$ *is optimal.*

Structure Assumption (SA_N):

There exist sets $\mathbb{M}_n \subset \mathbb{M}(E)$ and sets Δ_n of decision rules such that for all $n = 0, 1, \ldots, N-1$:

(i) $g_N \in \mathbb{M}_N$.

(ii) If $v \in \mathbb{M}_{n+1}$ then $T_n v$ is well-defined and $T_n v \in \mathbb{M}_n$.

(iii) For all $v \in \mathbb{M}_{n+1}$ there exists a maximizer f_n of v with $f_n \in \Delta_n$.

Structure Theorem

Theorem

Let (SA_N) be satisfied. Then it holds:

a) $V_n \in \mathbb{M}_n$ *and* (V_n) *satisfies the* Bellman equation.

b) $V_n = T_n T_{n+1} \dots T_{N-1} g_N$.

c) *For* $n = 0, 1, \dots, N-1$ *there exist maximizers* f_n *of* V_{n+1} *with* $f_n \in \Delta_n$*, and every sequence of maximizers* f_n^* *of* V_{n+1} *defines an optimal policy* $(f_0^*, f_1^*, \dots, f_{N-1}^*)$.

Upper Bounding Functions

Definition

$b : E \to \mathbb{R}_+$ is called an *upper bounding function* if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$ such that for $n = 0, 1, \dots, N-1$:

(i) $r_n^+(x, a) \le c_r b(x)$,

(ii) $g_N^+(x) \le c_g b(x)$,

(iii) $\int b(x') Q_n(dx'|x, a) \le \alpha_b b(x)$.

Upper Bounding Functions

Definition

$b : E \to \mathbb{R}_+$ is called an *upper bounding function* if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$ such that for $n = 0, 1, \dots, N-1$:

(i) $r_n^+(x,a) \le c_r b(x)$,

(ii) $g_N^+(x) \le c_g b(x)$,

(iii) $\int b(x') Q_n(dx'|x,a) \le \alpha_b b(x)$.

$\alpha_b := \sup_{(x,a)\in D} \frac{\int b(x')Q(dx'|x,a)}{b(x)}$.

Upper Bounding Functions

Definition

$b : E \to \mathbb{R}_+$ is called an *upper bounding function* if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$ such that for $n = 0, 1, \dots, N-1$:

(i) $r_n^+(x,a) \le c_r b(x)$,

(ii) $g_N^+(x) \le c_g b(x)$,

(iii) $\int b(x') Q_n(dx'|x,a) \le \alpha_b b(x)$.

$\alpha_b := \sup_{(x,a)\in D} \frac{\int b(x')Q(dx'|x,a)}{b(x)}$. Define $\|v\|_b := \sup_{x\in E} \frac{|v(x)|}{b(x)}$.

Upper Bounding Functions

Definition

$b : E \to \mathbb{R}_+$ is called an *upper bounding function* if there exist $c_r, c_g, \alpha_b \in \mathbb{R}_+$ such that for $n = 0, 1, \dots, N-1$:

(i) $r_n^+(x, a) \le c_r b(x)$,

(ii) $g_N^+(x) \le c_g b(x)$,

(iii) $\int b(x') Q_n(dx'|x, a) \le \alpha_b b(x)$.

$\alpha_b := \sup_{(x,a)\in D} \frac{\int b(x')Q(dx'|x,a)}{b(x)}$. Define $\|v\|_b := \sup_{x\in E} \frac{|v(x)|}{b(x)}$.

$B_b := \{v \in \mathbb{M}(E) \mid \|v\|_b < \infty\},\ B_b^+ := \{v \in \mathbb{M}(E) \mid \|v^+\|_b < \infty\}.$

Bounding Functions

Definition

$b : E \to \mathbb{R}_+$ is called a *bounding function* if there exist $c_r, \alpha_b \in \mathbb{R}_+$ such that

(i) $|r_n(x, a)| \leq c_r b(x)$,

(ii) $|g_N(x)| \leq c_g b(x)$,

(iii) $\int b(x')Q(dx'|x, a) \leq \alpha_b b(x)$.

Example: Consumption-Investment Problem

Financial Market:

- Bond price: $B_n = (1+i)^n$,
- Stock prices: $S_n^k = S_0^k \prod_{m=1}^n Y_m^k, \quad k = 1, \dots, d$.

We denote $Y_n := (Y_n^1, \dots, Y_n^d)$.

Example: Consumption-Investment Problem

Financial Market:

- Bond price: $B_n = (1+i)^n$,
- Stock prices: $S_n^k = S_0^k \prod_{m=1}^n Y_m^k, \quad k = 1, \dots, d.$

We denote $Y_n := (Y_n^1, \dots, Y_n^d)$.

Assumptions:

- $Y_1, \dots, Y_N$ are independent.
- (FM): There are no arbitrage opportunities.

Example: Consumption-Investment Problem

Policies:

- ϕ_n^k = amount of money invested in stock k at time n,
  $\phi_n = (\phi_n^1, \ldots, \phi_n^d) \in \mathbb{R}^d$.
- ϕ_n^0 = amount of money invested in the bond at time n.
- c_n = amount of money consumed at time n, $c_n \geq 0$.

Example: Consumption-Investment Problem

Policies:

- ϕ_n^k = amount of money invested in stock k at time n, $\phi_n = (\phi_n^1, \ldots, \phi_n^d) \in \mathbb{R}^d$.
- ϕ_n^0 = amount of money invested in the bond at time n.
- c_n = amount of money consumed at time n, $c_n \geq 0$.

Wealth process:

$$\begin{aligned} X_{n+1}^{c,\phi} &= (1+i)(X_n^{c,\phi} - c_n) + \phi_n \cdot (Y_{n+1} - (1+i) \cdot e) \\ &= (1+i)(X_n^{c,\phi} - c_n + \phi_n \cdot R_{n+1}) \end{aligned}$$

Optimization Problem

Let $U_c, U_p : \mathbb{R}_+ \to \mathbb{R}_+$ be strictly increasing, strictly concave utility functions.

$$\begin{cases} \mathbb{E}_x \left[\sum_{n=0}^{N-1} U_c(c_n) + U_p(X_N^{c,\phi}) \right] \to \max \\ (c, \phi) = (c_n, \phi_n) \text{ is a consumption-investment strategy with } \\ \qquad X_N^{c,\phi} \geq 0. \end{cases}$$

MDP Formulation

- $E := [0, \infty)$ where $x \in E$ denotes the wealth,
- $A := \mathbb{R}_+ \times \mathbb{R}^d$ where $a \in \mathbb{R}^d$ is amount of money invested in the risky assets, $c \in \mathbb{R}_+$ is amount which is consumed,
- $D_n(x)$ is given by

$$D_n(x) := \Big\{(c, a) \in A \mid 0 \le c \le x \text{ and } (1+i)(x - c + a \cdot R_{n+1}) \in E \ \mathbb{P}\text{-a.s.}\Big\},$$

- $Q_n(\cdot|x, c, a) :=$ distribution of $(1+i)(x - c + a \cdot R_{n+1})$,
- $r_n\big(x, c, a\big) := U_c(c)$,
- $g_N(x) := U_p(x)$.

Structure Result

Note: $b(x) = 1 + x$ is a bounding function for the MDP.

Theorem

a) *V_n are strictly increasing and strictly concave.*

b) *The value functions can be computed recursively by*

$$\begin{aligned} V_N(x) &= U_p(x), \\ V_n(x) &= \sup_{(c,a)} \Big\{ U_c(c) + \mathbb{E}\, V_{n+1}\Big((1+i)(x - c + a \cdot R_{n+1} \Big) \Big\}. \end{aligned}$$

c) *There exist maximizers $f_n^*(x) = (c_n^*(x), a_n^*(x))$ of V_{n+1} and the strategy $(f_0^*, f_1^*, \dots, f_{N-1}^*)$ is optimal.*

Power Utility

Let us assume $U_c(x) = U_p(x) = \frac{1}{\gamma}x^\gamma$ with $0 < \gamma < 1$.

Theorem

a) *The value functions are given by* $V_n(x) = d_n x^\gamma,\ x \geq 0.$

b) *Optimal consumption is* $c_n^*(x) = x(\gamma d_n)^{-\delta}$ *and the optimal amounts which are invested* ($\delta = (1-\gamma)^{-1}$)

$$a_n^*(x) = x\frac{(\gamma d_n)^\delta - 1}{(\gamma d_n)^\delta}\alpha_n^*, \quad x \geq 0$$

where α_n^* *is the optimal solution of the problem*

$$\sup_{\alpha \in A_n} \mathbb{E}[(1 + \alpha \cdot R_{n+1})^\gamma], \quad A_n = \{\alpha \in \mathbb{R}^d : 1 + \alpha \cdot R_{n+1} \geq 0\}.$$

Semicontinuous MDPs

Theorem

Suppose the MDP has an upper bounding function b and for all $n = 0, 1, \ldots, N-1$ *it holds:*

(i) $D_n(x)$ *is compact and* $x \mapsto D_n(x)$ *is upper semicontinuous (usc),*

(ii) $(x, a) \mapsto \int v(x')Q_n(dx'|x, a)$ *is usc for all usc* $v \in B_b^+$,

(iii) $(x, a) \mapsto r_n(x, a)$ *is usc,*

(iv) $x \mapsto g_N(x)$ *is usc.*

Then $\mathbb{M}_n := \{v \in B_b^+ \mid v \text{ is usc}\}$ *and* $\Delta_n := \{f_n \text{ dec. rule at } n\}$ *satisfy the Structure Assumption* (SA_N). *In particular,* $V_n \in \mathbb{M}_n$ *and there exists an optimal policy* $(f_0^*, \ldots, f_{N-1}^*)$ *with* $f_n^* \in \Delta_n$.

Infinite Horizon Problem

MDPs with Infinite Time Horizon

Consider a stationary MDP with $\beta \in (0,1], g \equiv 0$ and $N = \infty$.

$$\begin{aligned} J_{\infty\pi}(x) &:= \mathbb{E}_x^\pi\left[\sum_{k=0}^\infty \beta^k r(X_k, f_k(X_k))\right], \\ J_\infty(x) &:= \sup_\pi J_{\infty\pi}(x), \quad x \in E. \end{aligned}$$

Integrability Assumption (A):

$$\delta(x) := \sup_\pi \mathbb{E}_x^\pi\left[\sum_{k=0}^\infty \beta^k r^+(X_k, f_k(X_k))\right] < \infty, \quad x \in E.$$

Convergence Assumption (C)

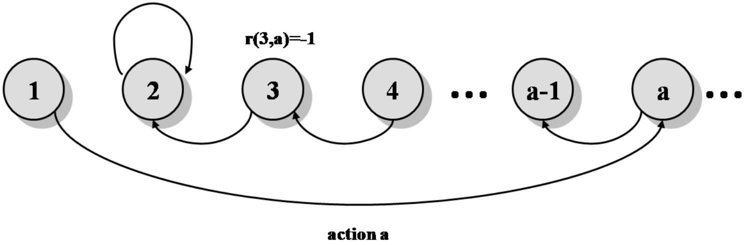
$$\lim_{n\to\infty} \sup_{\pi} \mathbb{E}_x^{\pi} \left[\sum_{k=n}^{\infty} \beta^k r^+ \big(X_k, f_k(X_k)\big) \right] = 0, \quad x \in E.$$

Assumption (C) implies that the following limits exist:

- $\lim_{n\to\infty} J_{n\pi} = J_{\infty\pi}$.
- $\lim_{n\to\infty} J_n =: J \geq J_\infty$.

J is called *limit value function*. Note: $J \neq J_\infty, J_\infty \notin \mathbb{M}(E)$.

Example: $J \neq J_\infty$ $(\beta = 1)$



We obtain:

$$J_\infty(1) \quad = \quad -1 < 0 = J(1).$$

Verification Theorem

$$Tv(x) = \sup_{a \in D(x)} \left\{ r(x,a) + \beta \int v(x')Q(dx'|x,a) \right\}$$

Theorem

Assume (C) and let $v \in \mathbb{M}(E)$, $v \leq \delta$ be a fixed point of T such that $v \geq J_\infty$. If f^* is a maximizer of v, then $v = J_\infty$ and the stationary policy $(f^*, f^*, \ldots)$ is optimal for the infinite-stage Markov Decision Problem.

Structure Assumption (SA)

There exists a set $\mathbb{M} \subset \mathbb{M}(E)$ and a set of decision rules Δ such that:

(i) $0 \in \mathbb{M}$.

(ii) If $v \in \mathbb{M}$ then $Tv(x)$ is well-defined and $Tv \in \mathbb{M}$.

(iii) For all $v \in \mathbb{M}$ there exists a maximizer $f \in \Delta$ of v.

(iv) $J \in \mathbb{M}$ and $J = TJ$.

Structure Theorem

Theorem

Let (C) and (SA) be satisfied. Then it holds:

a) $J_\infty \in \mathbb{M}$, $J_\infty = TJ_\infty$ *and* $J_\infty = J = \lim_{n\to\infty} J_n$.

b) *There exists a maximizer* $f \in \Delta$ *of* J_∞, *and every maximizer* f^* *of* J_∞ *defines an optimal stationary policy* $(f^*, f^*, \ldots)$.

Example: Dividend Pay-Out

Let X_n be the risk reserve of an insurance company at time n. We assume that

- $Z_n =$ difference between premia and claim sizes in n-th time interval,
- $Z_1, Z_2, \dots$ are iid, $Z_n \in \mathbb{Z}$ and $\mathbb{P}(Z_1 = k) = q_k, k \in \mathbb{Z}$.
- $\mathbb{P}(Z_1 < 0) > 0$ and $\mathbb{E}\, Z^+ < \infty$.

Control: We can pay-out a dividend at each time-point.

$$X_{n+1} = X_n - f_n(X_n) + Z_{n+1}.$$

Let $\tau := \inf\{n \in \mathbb{N} : X_n < 0\}$ be the ruin time point.
Aim: Maximize the expected disc. dividend pay-out until τ.

Formulation as an MDP

- $E := \mathbb{Z}$ where $x \in E$ denotes the risk reserve,
- $A := \mathbb{N}_0$ where $a \in A$ is the dividend pay-out,
- $D(x) := \{0, 1, \dots, x\}, x \geq 0$, and $D(x) := \{0\}, x < 0$,
- $Q(\{y\}|x, a) := q_{y-x+a}$ if $x \geq 0$, else $Q(\{y\}|x, a) = \delta_{xy}$,
- $r(x, a) := a$,
- $\beta \in (0, 1)$.

Then for a policy $\pi = (f_0, f_1, \dots)$ we have

$$J_{\infty\pi}(x) = \mathbb{E}_x^\pi \left[\sum_{k=0}^{\tau-1} \beta^k f_k(X_k) \right].$$

First Results

Corollary

a) *The function* $b(x) = 1 + x, x \geq 0$ *and* $b(x) = 0, x < 0$ *is a bounding function. (A) is satisfied.*

b) *(C) is satisfied.*

c) *It holds for* $x \geq 0$ *that*

$$x + \frac{\beta \, \mathbb{E} \, Z^+}{1 - \beta q_+} \leq J_\infty(x) \leq x + \frac{\beta \, \mathbb{E} \, Z^+}{1 - \beta}$$

where $q_+ := \mathbb{P}(Z_1 \geq 0)$*.*

In particular (SA) is satisfied with $\mathbb{M} := B_b$.

Bellman Equation

The Structure Theorem yields that

- $\lim_{n\to\infty} J_n = J_\infty$,
- Bellman equation

$$J_\infty(x) = \max_{a\in\{0,1,\dots,x\}} \left\{ a + \beta \sum_{k=a-x}^{\infty} J_\infty(x-a+k) q_k \right\},$$

- Every maximizer of J_∞ (which obviously exists) defines an optimal stationary policy $(f^*, f^*, \dots)$.

Let f^* be the largest maximizer of J_∞.

Further Properties of J_∞ and f^*

Theorem

a) *The value function* $J_\infty(x)$ *is increasing.*

b) *It holds that*

$$J_\infty(x) - J_\infty(y) \geq x - y, \quad x \geq y \geq 0.$$

c) *For* $x \geq 0$ *it holds that* $f^*\big(x - f^*(x)\big) = 0.$

Band and Barrier Policies

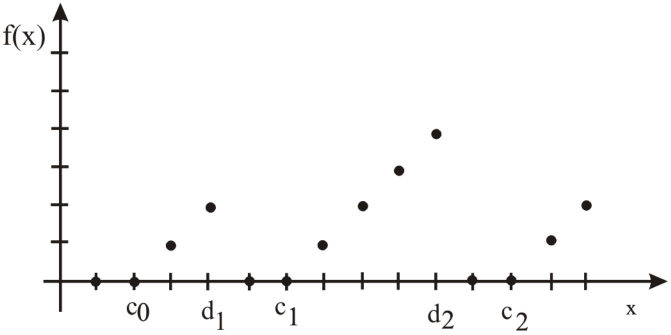
Definition

a) A stationary policy $(f, f, \dots)$ is called *band-policy*, if $\exists$ $n \in \mathbb{N}_0$ and $c_0, \dots c_n, d_1, \dots d_n \in \mathbb{N}_0$ s.t. $d_k - c_{k-1} \geq 2$, $0 \leq c_0 < d_1 \leq c_1 < \dots < d_n \leq c_n$ and

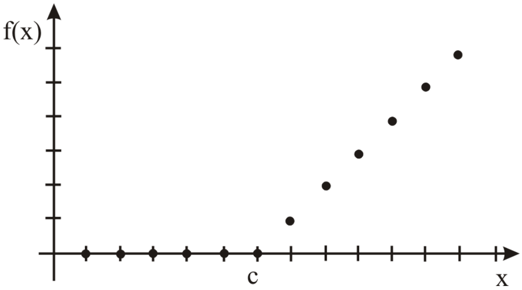
$$f(x) = \begin{cases} 0, & \text{if } x \leq c_0 \\ x - c_k, & \text{if } c_k < x < d_{k+1} \\ 0, & \text{if } d_k \leq x \leq c_k \\ x - c_n, & \text{if } x > c_n. \end{cases}$$

b) A stationary policy $(f, f, \dots)$ is called *barrier-policy* if it is a band-policy and $c_0 = c_n$.

Band Policies



Barrier Policy



Main Results

Lemma

Let $\xi := \sup\{x \in \mathbb{N}_0 \mid f^*(x) = 0\}$. *Then* $\xi < \infty$ *and*

$$f^*(x) = x - \xi \quad \text{for all } x \geq \xi.$$

Theorem

The stationary policy $(f^*, f^*, \dots)$ *is optimal and a band-policy.*

When is the Band a Barrier?

Known Condition: $\mathbb{P}(Z_1 \geq -1) = 1$.

- de Finetti (1957)
- Shubik and Thomson (1959)
- Miyasawa (1962)
- Gerber (1969)
- Reinhard (1981)
- Schmidli (2008)
- Asmussen and Albrecher (2010)

Semicontinuous MDPs

Theorem

Suppose there exists an upper bounding function b, (C) is satisfied and

(i) $D(x)$ *is compact for all* $x \in E$ *and* $x \mapsto D(x)$ *is usc,*

(ii) $(x, a) \mapsto \int v(x')Q(dx'|x, a)$ *is usc for all usc* $v \in B_b^+$,

(iii) $(x, a) \mapsto r(x, a)$ *is usc.*

Then it holds:

a) $J_\infty \in B_b^+$, $J_\infty = TJ_\infty$ *and* $J_\infty = J$ **(Value Iteration)**.

b) $\emptyset \neq LsD_n^*(x) \subset D_\infty^*(x)$ *for all* $x \in E$ **(Policy Iteration)**.

c) *There exists an* $f^* \in F$ *with* $f^*(x) \in LsD_n^*(x)$ *for all* $x \in E$, *and the stationary policy* $(f^*, f^*, \dots)$ *is optimal.*

Contracting MDP

Theorem

Let b be a bounding function and $\beta\alpha_b < 1$. *If there exists a closed subset* $\mathbb{M} \subset B_b$ *and a set* Δ *such that*

(i) $0 \in \mathbb{M}$,

(ii) $T : \mathbb{M} \to \mathbb{M}$,

(iii) *for all* $v \in \mathbb{M}$ *there exists a maximizer* $f \in \Delta$ *of* v,

then it holds:

a) $J_\infty \in \mathbb{M}$, $J_\infty = TJ_\infty$ *and* $J_\infty = J$.

b) J_∞ *is the unique fixed point of* T *in* $\mathbb{M}$.

c) *There exists a maximizer* $f \in \Delta$ *of* J_∞, *and every maximizer* f^* *of* J_∞ *defines an optimal stationary policy* $(f^*, f^*, \ldots)$.

Howard's Policy Improvement Algorithm

Let J_f be the value function of the stationary policy $(f, f, \dots)$.
Denote $D(x, f) := \{a \in D(x) \mid LJ_f(x, a) > J_f(x)\}, \quad x \in E$.

Theorem

Suppose the MDP is contracting. Then it holds:

a) *If for some subset $E_0 \subset E$ we define a decision rule h by*

$$h(x) \in D(x, f) \text{ for } x \in E_0, \qquad h(x) := f(x) \text{ for } x \notin E_0,$$

then $J_h \geq J_f$ and $J_h(x) > J_f(x)$ for $x \in E_0$. In this case the decision rule h is called an improvement of f.

b) *If $D(x, f) = \emptyset$ for all $x \in E$, then the stationary policy $(f, f, \dots)$ is optimal.*

Extensions and Related Problems

- Stopping Problems
- Partially Observable Markov Decision Processes
- Piecewise Deterministic Markov Decision Processes
- Problems with Average Reward
- Games

- Bäuerle, N., Rieder, U. (2011) : Markov Decision Processes with Applications to Finance. Springer.
- Bellman, R. (1957, 2003): Dynamic Programming. Princeton University Press, NJ.
- Bertsekas, D.P. and Shreve, S.E. (1978) : Stochastic optimal control. Academic Press, New York.
- Bertsekas, D.P. (2001,2005) : Dynamic programming and optimal control. Vol. I, II. Athena Scientific, Belmont, MA.
- Feinberg, E.A. and Shwartz, A. (2002): Handbook of Markov decision processes. Kluwer Academic Publishers, Boston, MA.
- Hernández-Lerma, O. and Lasserre, J.B. (1996): Discrete-time Markov control processes. Springer-Verlag, New York.

Howard, R. (1960) : Dynamic programming and Markov processes. The Technology Press of M.I.T., Cambridge, Mass.

Powell, W.B. (2007): Approximate dynamic programming. Wiley-Interscience, Hoboken, NJ.

Puterman, M.L. (1994): Markov decision processes: discrete stochastic dynamic programming, John Wiley & Sons, New York.

Shapley, L. S. (1953): Stochastic games, Proc. Nat. Acad. Sci., pp. 1095–1100.

Thank you very much
for your attention!