Markov Decision Processes with Applications

Ulrich Rieder

University of Ulm

Markov Decision Processes with Applications

- Markov Decision Processes

  Basic results, Computational aspects

- Partially Observable Markov Decision Processes

  Hidden Markov models, Filtered MDPs

  Bandit problems, Consumption-Investment problems

- Continuous-Time Markov Decision Processes

  Piecewise deterministic MDPs

  Parallel queueing model

Markov Decision Processes

$\big(E, A, D_n, Q_n, r_n, g_N\big)$ with horizon N

- E state space
- A action space
- $D_n \subset E \times A$ admissible state-action pairs at time n
- Q_n $Q_n\big(\cdot|x,a\big)$ transition law at time n
- $r_n : D_n \to \mathbb{R}$ reward function at time n
- $g_N : E \to \mathbb{R}$ terminal reward function at time N

decision rule at time n $f_n : E \to A$ measurable and $f_n(x) \in D_n(x)$ for all $x \in E$

policy $\pi := \big(f_0, f_1, \ldots, f_{N-1}\big)$

For $n = 0, 1, \dots, N$ define the value functions

$$V_{n\pi}(x) := E_x^\pi\Big[\sum_{k=n}^{N-1} r_k\big(X_k, f_k(X_k)\big) + g_N(X_N)\Big]$$

$$V_n(x) := \sup_\pi V_{n\pi}(x),\ x \in E$$

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

Integrability Assumption (A_N) **:**

For $n = 0, 1, \dots, N$

$$\sup_\pi E_x^\pi\Big[\sum_{k=n}^{N-1} r_k^+\big(X_k, f_k(X_k)\big) + g_N^+(X_N)\Big] < \infty,\ x \in E$$

Bertsekas/Shreve (1978), Hernandez-Lerma/Lasserre (1996)...

Puterman (1994), Feinberg/Schwartz (2002) ...

Bäuerle/Rieder (2011)

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

$$\begin{aligned}
(L_n v)(x,a) &:= r_n(x,a) + \int v(x')Q_n(dx'|x,a), \; (x,a) \in D_n \\
(T_{nf_n} v)(x) &:= (L_n v)(x, f_n(x)) \\
(T_n v)(x) &:= \sup_{a \in D_n(x)} (L_n v)(x,a), \; x \in E
\end{aligned}$$

Note: $T_n v \notin \mathbb{M}(E)$!

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

Reward Iteration: $V_{n\pi} = T_{nf_n} V_{n+1,\pi}$, $V_{N\pi} = g_N$.

Bellman Equation: $V_n = T_n V_{n+1}$, $V_N = g_N$.

Verification Theorem: Let $(v_n) \subset \mathbb{M}(E)$ be a solution of the Bellman equation.

a) $v_n \geqslant V_n$ for $n = 0, 1, \ldots, N$.

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

Structure Assumption (SA_N)**:** There exist sets $\mathbb{M}_n \subset \mathbb{M}(E)$ of measurable functions and sets $\triangle_n$ 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 $f_n \in \triangle_n$.

Structure Theorem:

Assume (SA_N). Then it holds:

a) $V_n \in \mathbb{M}_n$ and (V_n) is a solution of the Bellman equation.

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

c) For $n = 0, 1, \ldots, N-1$ there exists a maximizer f_n of V_{n+1} with $f_n \in \triangle_n$, and every sequence of maximizers f_n^* of V_{n+1} defines an optimal policy $\left(f_0^*, f_1^*, \ldots f_{N-1}^*\right)$ for the N-stage Markov Decision Problem.

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

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

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

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

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

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

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

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

(ii) $|g_N(x)| \le c_g b(x)$.

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

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

(i) $D_n(x)$ is compact and $x \to D_n(x)$ is upper semicontinuous (usc).

(ii) $(x, a) \to \int v(x')Q_n\big(dx'|x, a\big)$ is usc for all usc $v \in \mathbb{B}_b^+$.

(iii) $(x, a) \to r_n(x, a)$ is usc .

(iv) $x \to g_N(x)$ is usc.

Then the sets $\mathbb{M}_n := \big\{v \in \mathbb{B}_b^+ | v \text{ is usc}\big\}$ and $\triangle_n := \big\{f_n \text{ decision rule at time } n\big\}$ satisfy the Structure Assumption $\big(SA_N\big)$, in particular: $V_n \in \mathbb{M}_n$ and there exists an optimal policy $\big(f_0^*, f_1^*, \dots, f_{N-1}^*\big)$ with $f_n^* \in \triangle_n$.

Markov Decision Processes with Infinite Time Horizon

We consider a stationary MDP with $\beta \in \big(0,1\big]$ and $N = \infty$.

$$J_{\infty\pi}(x) := E_x^\pi\Big[\sum_{k=0}^{\infty} \beta^k\, r\big(X_k, f_k(X_k)\big)\Big]$$
$$J_\infty(x) := \sup_\pi J_{\infty\pi}(x),\ x \in E.$$

Integrability Assumption (A):

$$\sup_\pi E_x^\pi\Big[\sum_{k=0}^{\infty}\beta^k\, r^+\big(X_k, f_k(X_k)\big)\Big] < \infty,\ x \in E$$

Convergence Assumption (C):

$$\lim_{n\to\infty} \sup_\pi E_x^\pi\Big[\sum_{k=n}^{\infty}\beta^k\, r^+\big(X_k, f_k(X_k)\big)\Big] = 0,\ x \in E$$

Then it holds: $J_{\infty\pi} = \lim_n J_{n\pi}$

limit value function $J := \lim_n J_n \geqslant J_\infty.$ Note: $J \neq J_\infty$ and $J_\infty \notin \mathbb{M}(E)$!

Verification Theorem: Assume (C). Let $v \in \mathbb{M}(E)$ be a fixed point of T such that $v \geqslant J_\infty$. 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 exist a set $\mathbb{M} \subset \mathbb{M}(E)$ of measurable functions and a set $\triangle$ of decision rules such that:

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

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

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

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

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

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

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

Theorem: Suppose the stationary MDP has an upper bounding function b with $\beta\alpha_b < 1$ and it holds:

(i) $D(x)$ is compact and $x \to D(x)$ is usc.

(ii) $(x, a) \to \int v(x')Q\big(dx'|x, a\big)$ is usc for all usc $v \in \mathbb{B}_b^+$.

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

Then it holds:

(a) $J_\infty \in \mathbb{B}_b^+,\ J_\infty = TJ_\infty$ and $J_\infty = J$ **(value iteration)**.

(b) b is usc $\Longrightarrow J_\infty$ is usc.

(c) $\emptyset \neq LsD_n^*(x) \subset D_\infty^*(x)$ for all $x \in E$ **(policy iteration)**.

(d) There exists a decision rule f^* with $f^*(x) \in LsD_n^*(x)$ for all $x \in E$, and the stationary policy $\big(f^*, f^*, \ldots\big)$ is optimal.

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

Contracting Markov Decision Processes

Structure Theorem: Let b be a bounding function and $\beta\alpha_b < 1$. If there exists a closed subset $\mathbb{M} \subset \mathbb{B}_b$ and a set $\triangle$ of decision rules 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 of v with $f \in \triangle$.

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 $f \in \triangle$ of J_∞, and every maximizer f^* of J_∞ defines an optimal stationary policy $\big(f^*, f^*, \ldots\big)$.

Howard's Policy Improvement Algorithm

Let J_f be the value function of the stationary policy $\big(f, f, \ldots\big)$.

Denote

$$D(x, f) := \big\{a \in D(x) | \big(LJ_f\big)(x, a) > J_f(x)\big\}$$

Let the Markov decision process be contracting.

Then it holds:

a) If for some subset $E_0 \subset E$

$$g(x) \in D(x, f) \text{ for } x \in E_0$$

$$g(x) = f(x) \text{ for } x \notin E_0$$

then $J_g \geqslant J_f$ and $J_g(x) > J_f(x)$ for $x \in E_0$.

In this case the decision rule g is called an **improvement** of f.

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

Remark: $\big(f, f, \ldots\big)$ is optimal $\Longleftrightarrow$ f cannot be improved.

Partially Observable Markov Decision Processes

- $E_X \times E_Y$ state space x observable state, y unobservable state
- A action space
- $D \subset E_X \times A$ admissible state-action pairs, $D(x) \subset A$
- $Q\big(\cdot|x, y, a\big)$ transition law
- Q_0 initial distribution (prior distribution) of Y_0
- $r\big(x, y, a\big)$ reward function
- $g\big(x, y\big)$ terminal reward function
- $\beta \in \big(0, 1\big]$ discount factor

Examples : Hidden Markov Model (HMM), Bayesian Decision Model

decision rule at time n $f_n\big(x_0, a_0, x_1, \dots, x_n\big) = f_n\big(h_n\big)$

policy $\pi = \big(f_0, f_1, \dots, f_{N-1}\big)$ finite horizon: $N < \infty$

Rieder (1975), Elliott et al. (1995), Bäuerle/Rieder (2011) ...

$$J_{N\pi}(x) := E_x^\pi\Big[\sum_{n=0}^{N-1} \beta^n r\big(X_n, Y_n, f_n(H_n)\big) + \beta^N g\big(X_N, Y_N\big)\Big]$$

$$J_N(x) := \sup_\pi J_{N\pi}(x),\ x \in E_X$$

For $n = 0, 1, \ldots$ and $C \subset E_Y$ define

$$\mu_n\big(C|X_0, A_0, X_1, \ldots, X_n\big) := P_x^\pi\big(Y_n \in C|X_0, A_0, X_1, \ldots, X_n\big)$$

a posteriori-distribution at time n

Filter Equation

$\mu_0 = Q_0$ and $\mu_{n+1}\big(\cdot|H_n, A_n, X_{n+1}\big) = \Phi\big(X_n, \mu_n\big(\cdot|H_n\big), A_n, X_{n+1}\big)$

where

$$\Phi(x, \rho, a, x')(C) := \frac{\int_C \Big[\int q(x', y'|x, y, a)\rho(dy)\Big]\nu(dy')}{\int_{E_Y} \Big[\int q(x', y'|x, y, a)\rho(dy)\Big]\nu(dy')},\ C \subset E_Y,\ \rho \in \mathbb{P}\big(E_Y\big)$$

Bayes-Operator

Filtered Markov Decision Process

- $E' := E_X \times \mathbb{P}\big(E_Y\big) \ni \big(x, \rho\big)$ enlarged state space
- A and $D\big(x, \rho\big) := D\big(x\big)$
- $Q^X\big(B|x, \rho, a\big) := \int Q\big(B \times E_Y|x, y, a\big)\rho(dy),\ B \subset E_X$

  $Q'\big(B \times C|x, \rho, a\big) := \int_B 1_C\big(\Phi(x, \rho, a, x')\big)Q^X\big(dx'|x, \rho, a\big),\ C \subset \mathbb{P}\big(E_Y\big)$
- $r'(x, \rho, a) := \int r\big(x, y, a\big)\rho\big(dy\big)$
- $g'(x, \rho) := \int g\big(x, y\big)\rho\big(dy\big)$

Theorem:

a) $J_{N\pi}(x) = J'_{N\pi}(x, Q_0)$ and $J_N(x) = J'_N(x, Q_0)$.

b) Assume (SA_N). Then the Bellman equation holds, i.e.

$$V'_N(x,\rho) := \beta^N g'(x,\rho)$$

$$V'_n(x,\rho) := \sup_{a\in D(x)} \left\{ r'(x,\rho,a) + \int V'_{n+1}\big(x', \Phi(x,\rho,a,x')\big) Q^X\big(dx'|x,\rho,a\big) \right\}.$$

Let f'_n be a maximizer of V'_{n+1} for $n = 0, \ldots, N-1$. Then the policy $\pi^* := \left(f^*_0, f^*_1, \ldots, f^*_{N-1}\right)$ is optimal for the N-stage POMDP, where

$$f^*_n(h_n) := f'_n\big(x_n, \mu_n(\cdot|h_n)\big),\ h_n = (x_0, a_0, x_1, \ldots, x_n).$$

Note that $V'_n(x,\rho) = \beta^n J'_{N-n}(x,\rho), \quad n = 0, \ldots, N$

Computational aspects Kalman Filter

Sufficient Statistics

Bandit Problems

unknown success probabilities $\theta_1 \in [0,1]$ and $\theta_2 \in [0,1]$

$Q_0 =$ product of two Uniform-distributions of $\big(\theta_1, \theta_2\big)$

Aim: maximize the expected number of successes in a finite or infinite number of trials

- $E' := \mathbb{N}_0^2 \times \mathbb{N}_0^2 \ni \big(m_1, n_1, m_2, n_2\big) = \rho$
- $A = \big\{1, 2\big\}$
- Bayes-Operator $\Phi\big(\rho, a, \big\{\text{success}\big\}\big) = \rho + e_{2a-1}$
- $r'\big(\rho, a\big) := \frac{m_a+1}{m_a+n_a+2}$
- $\beta \in \big(0, 1\big]$.

$N < \infty :$ There exists an optimal policy.

monotonicity results: stay-on-a-winner property

stopping property if θ_2 is known.

$N = \infty$ and $\beta \in \big(0,1\big)$:

For $K \in \mathbb{R}$ let $J(m,n;K)$ be the unique solution of

$v(m,n) = \max\big\{K, \beta\big(p(m,n)v(m+1,n) + \big(1-p(m,n)\big)v(m,n+1)\big)\big\}$

for $(m,n) \in \mathbb{N}_0^2$ and $p(m,n) := \frac{m+1}{m+n+2}$.

Define the **Gittins-Index**

$$I(m,n) := \min\big\{K | J(m,n;K) = K\big\}$$

Then it holds:

The stationary Index-policy $\big(f^*, f^*, \ldots\big)$ is optimal for the infinite-stage Bandit problem where

$$f^*\big(m_1, n_1, m_2, n_2\big) = \begin{cases} 1 \text{ if } I\big(m_1, n_1\big) \geqslant I\big(m_2, n_2\big) \\ 2 \text{ if } I\big(m_1, n_1\big) < I\big(m_2, n_2\big). \end{cases}$$

Gittins (1989), Whittle (1980), (1988)

Cox-Ross-Rubinstein Model

- Bond $\quad B_n = \big(1+i\big)^n$
- Stock $\quad S_n = S_0 \cdot \prod_{k=1}^{n} Y_k \qquad \big(Y_k\big)$ independent and identically distributed

$$P\big(Y_k = \boldsymbol{u}\big) = \theta = 1 - P\big(Y_k = \boldsymbol{d}\big) \quad \text{unknown up-probability} \quad \theta$$

Q_0 =Uniform-distribution of θ

(NA) : $\quad \boldsymbol{d} < 1 + i < \boldsymbol{u}$

π_n = amount of money invested in the stock at time n

Then it holds for the wealth process:

$$X_{n+1}^{\pi} = X_n^{\pi}\big(1+i\big) + \pi_n\big(Y_{n+1} - 1 - i\big),\ X_0^{\pi} = x > 0$$

Utility function $U : \mathbb{R}_+ \longrightarrow \mathbb{R}_+,$ strictly increasing and concave

$$(P) \begin{cases} E_x\Big[U\big(X_N^{\pi}\big)\Big] \longrightarrow \max \\ \quad X_N^{\pi} \geqslant 0 \\ \quad \pi = \big(\pi_n\big) \ \text{portfolio-strategy} \end{cases}$$

- $E' := \mathbb{R}_+ \times \mathbb{N}_0^2 \ni \big(x, (m,n)\big) = (x, \rho)$
- $A = \mathbb{R}, \quad D(x) = \big\{a \in \mathbb{R} \big| (1+i)x + a\big(Y - i - 1\big) \geqslant 0 \text{ a.s.}\big\}$
- Bayes-Operator $\quad \Phi\big(\rho, \boldsymbol{u}\big) = \big(m+1, n\big)$
- $r' \equiv 0,\ g'\big(x,\rho\big) := U\big(x\big)$

$b\big(x,\rho\big) := 1 + x$ is a bounding function for the filtered MDP.

Then it holds:

a) $J_N(x) = J'_N\big(x, Q_0\big)$ is strictly increasing and concave in x.

b) There exists an optimal policy $\big(f_0^*, f_1^*, \ldots, f_{N-1}^*\big)$ for (P).

Application: $U(x) = \frac{1}{\gamma}x^\gamma$ **(power utility)** $\gamma < 1,\ \gamma \neq 0$

(i) $J_N(x,\rho) = J_N(x,m,n) = \frac{1}{\gamma}x^\gamma \cdot d_N(m,n)$.

(ii) $f_k^*(x,\rho) = f_k^*(x,m,n) = x \cdot \alpha_k(m,n)$.

monotonicity results: $(m,n) \leqslant (m',n') :\Longleftrightarrow m \leqslant m',\ n \geqslant n'$

(iii) $0 < \gamma < 1: \quad \alpha_k(m,n) \geqslant \alpha_k(\bar{p})$ with $\bar{p} := \frac{m+1}{m+n+2}$

$\gamma < 0: \quad \alpha_k(m,n) \leqslant \alpha_k(\bar{p})$

Piecewise Deterministic Markov Decision Processes

- E state space, $E \subset \mathbb{R}^d$
- $\mathbb{U}$ control space

  $A := \big\{\alpha : \mathbb{R}_+ \longrightarrow \mathbb{U}$ measurable$\big\}$, we write: $\alpha(t) = \alpha_t$
- $\mu(x, u)$ drift between jumps

  $\phi_t^\alpha(x)$ (unique) solution of : $dx_t = \mu\big(x_t, \alpha_t\big)dt,\ x_0 = x$

  deterministic flow between jumps
- $\lambda > 0$ jump rate (here: λ is independent of (x, u))

  $0 := T_0 < T_1 < T_2 < \ldots$ jump time points of a Poisson process with rate λ
- $Q\big(\cdot|x, u\big)$ distribution of jump goals
- $r(x, u)$ reward rate
- $\beta \geq 0$ discount rate

$\pi = \big(\pi_t\big)$ is called a **Markovian policy** (or piecewise open loop policy) if there exists a sequence of measurable functions $f_n : E \longrightarrow A$ such that

$$\pi_t = f_n\big(Z_n\big)\big(t - T_n\big) \text{ for } T_n < t \leqslant T_{n+1}.$$

We write: $\pi = \big(\pi_t\big) = \big(f_n\big)$.

piecewise deterministic Markov process

$X_t = \phi^\pi_{t-T_n}\big(Z_n\big)$ for $T_n \leqslant t < T_{n+1}, \quad Z_n = X_{T_n}$

$$V_\pi(x) := E_x^\pi\Big[\int_0^\infty e^{-\beta t} r\big(X_t, \pi_t\big)dt\Big]$$
$$V_\infty(x) := \sup_\pi V_\pi(x),\ x \in E$$

- Continuous-time stochastic control: Hamilton-Jacobi-Bellman equation
- Solution via discrete-time MDP

Yuskevich (1987), Davis (1993), Schäl et al. (2004)...

Jacobsen (2006), Guo/Hernandez-Lerma (2009): CTMDP

Discrete-time MDP

- E state space (embedded Markov process)
- A action space
- $Q'\big(B|x,\alpha\big) := \lambda \int\limits_0^\infty e^{-(\beta+\lambda)t} Q\big(B|\phi_t^\alpha(x), \alpha_t\big)dt,\ B \subset E$
- $r'(x,\alpha) := \int\limits_0^\infty e^{-(\beta+\lambda)t} r\big(\phi_t^\alpha(x), \alpha_t\big)dt$
- $\beta' = 1$

Note: A is a function space, Q' is substochastic.

$$\big(Tv\big)(x) = \sup_{\alpha\in A}\Big\{\int\limits_0^\infty e^{-(\beta+\lambda)t}\Big[r\big(\phi_t^\alpha(x), \alpha_t\big) + \lambda \int v(z) Q\big(dz|\phi_t^\alpha(x), \alpha_t\big)\Big]dt\Big\}$$

Theorem:

$$V_\pi(x) = E_x^\pi\Big[\sum_{n=0}^\infty r'\Big(Z_n', f_n\big(Z_n'\big)\Big)\Big] =: J_{\infty\pi}(x)$$

$$V_\infty(x) = \sup_\pi J_{\infty\pi}(x) = J_\infty(x),\ x \in E$$

For a proof of the following result we use the set $\mathcal{R} := \{\alpha : \mathbb{R}_+ \longrightarrow \mathbb{P}(\mathbb{U}) \text{ measurable}\}$ of **relaxed controls** (with the Young topology). Since $\mathcal{R} \supset A$, we have to extend the domain of the data Q' and r'. Then it holds:

$$J_\infty^{\text{rel}}(x) \geqslant J_\infty(x) = V_\infty(x),\ x \in E.$$

$b : E \longrightarrow \mathbb{R}_+$ is called an **upper bounding** function for the Piecewise Deterministic Markov Model, if there exist $c_r, c_Q, c_\phi \in \mathbb{R}_+$ such that

(i) $r^+(x,u) \leqslant c_r b(x)$.

(ii) $\int b(x')Q(dx'|x,u) \leqslant c_Q b(x)$.

(iii) $\lambda \int_0^\infty e^{-(\lambda+\beta)t} b\big(\phi_t^\alpha(x)\big)\,dt \leqslant c_\phi b(x)$.

If r is bounded from above, then $b \equiv 1$ is an upper bounding function and $c_Q = 1$ and $c_\phi = \frac{\lambda}{\lambda+\beta}$.

If b is an upper bounding function, then b is an upper bounding function for the MDP' (with and without relaxed controls) and $\alpha_b \leqslant c_Q c_\phi$.

Theorem: Suppose the Piecewise Deterministic Markov Model has a continuous upper bounding function b with $\alpha_b < 1$ and it holds:

(i) $\mathbb{U}$ is compact.

(ii) $(t, x, \alpha) \longrightarrow \phi_t^\alpha(x)$ is continuous.

(iii) $(x, u) \longrightarrow \int v(z) Q\big(dz|x, u\big)$ is usc for all usc $v \in \mathbb{B}_b^+$

(iv) $(x, u) \longrightarrow r(x, u)$ is usc.

Then it holds:

a) J_∞^{rel} is upper semi-continuous and $J_\infty^{\text{rel}} = T J_\infty^{\text{rel}}$.

b) There exists an optimal relaxed policy $\pi^* = \big(\pi_t^*\big)$, i.e. π_t^* takes values in $\mathbb{P}(\mathbb{U})$.

c) If $\phi_t^\alpha(x)$ is independent of α or if $\mathbb{U}$ is convex, $\mu(x, u)$ is linear in u and $u \longrightarrow \Big[r(x, u) + \lambda \int J_\infty^{\text{rel}}(z) Q\big(dz|x, u\big)\Big]$ is concave on $\mathbb{U}$, then there exists an optimal **nonrelaxed** policy $\pi^* = \big(\pi_t^*\big)$ such that

$$\pi_t^* = f\big(X_{T_n}^{\pi^*}\big)\big(t - T_n\big),\ T_n < t \leqslant T_{n+1} \text{ for a decision rule } f : E \to A.$$

In particular, π_t^* takes values in $\mathbb{U}$ and $J_\infty^{\text{rel}} = J_\infty = V_\infty$.

Continuous-Time Markov Decision Processes

- CTMDP (X_t) with contable state space E_X and intensities $q_{ij}(u)$

  uniformized CTMDP : $\lambda \geqslant \sum_{j \neq i} q_{ij}(u) = -q_{ii}(u),\ i \in E_X,\ u \in \mathbb{U}$

- Partially Observable CTMDP:

  intensities depend on CTMC (Y_t) with finite state space E_Y, i.e.

  $q_{ij}(y,u)$ if $Y_t = y \in E_Y,\ y$ **unobservable state**

  Q_0 initial distribution of Y_0

  $$\lambda \geqslant \sum_{j \neq i} q_{ij}(y,u) = -q_{ii}(y,u),\ i \in E_X,\ y \in E_Y,\ u \in \mathbb{U}$$

  $$V_\pi(i) := E_i^\pi\Big[\int_0^\infty e^{-\beta t} r\big(X_t^\pi, Y_t, \pi_t\big)dt\Big]$$
  $$V_\infty(i) := \sup_\pi V_\pi(i),\ i \in E_X$$

Filter Equation $\mu_t := P_i^\pi\big(Y_t = \cdot|\mathcal{F}_t^X\big) \in \mathbb{P}\big(E_Y\big)$

$$d\mu_t = b\big(X_t, \mu_t, \pi_t\big)dt + H\left(X_{t-}, \mu_{t-}, X_t, \pi_{t-}\right),\ \mu_0 = Q_0$$

Reformulation as filtered PDMDP:

$\big(X_t, \mu_t\big)$ piecewise deterministic MDP with state space $E_X \times \mathbb{P}\big(E_Y\big)$.

Theorem.

a) $V_\pi(i) = J_{\infty\pi}\big(i, Q_0\big)$ and $V_\infty(i) = J_\infty(i, Q_0\big),\ i \in E_X$.

b) Assumptions! Then the Bellman equation holds, i.e.

$$J_\infty(i,\rho) = \sup_{\alpha \in A}\Big\{\int_0^\infty e^{-(\beta+\lambda)t}\big(LJ_\infty\big)\big(i, \phi_t^\alpha(i,\rho), \alpha_t\big)dt\Big\}$$

where

$$\big(LJ_\infty\big)(i,\rho,u) := r(i,\rho,u) + \sum_{j\neq i}\Big(J_\infty(j, \rho + H(i,\rho,j,u)) - J_\infty(i,\rho)\Big)q_{ij}(\rho,u) + \lambda J_\infty(i,\rho)$$

Application: Parallel Queueing Model

two parallel queues and one server

Aim: minimize the expected number of waiting customers

- complete information: μc-rule is optimal
- partial information: $Y_t \equiv Y \in \{\mu_1, \nu_1\} \times \{\mu_2, \nu_2\}$

  e.g. two types of customers are in the system and the server can not differ which group is waiting in which queue.

There exists an optimal nonrelaxed policy $f^*(i, \rho) \in \{1, 2\}$.

symmetric case: $Y \in \{(\mu_1, \mu_2),\ (\mu_2, \mu_1)\},\ \mu_1 < \mu_2$

$\mu_t = P_i^\pi\big(Y = (\mu_1, \mu_2) | \mathcal{F}_t^X\big) \Longrightarrow d\mu_t = (\mu_2 - \mu_1)(2\pi_t - 1)\mu_t(1 - \mu_t)dt + \triangle \mu_t$

$$\triangle \mu_t = \begin{cases} H_1(\mu_{t-}) & \text{if } X_t^1 = X_{t-}^1 - 1 \\ H_2(\mu_{t-}) & \text{if } X_t^2 = X_{t-}^2 - 1 \end{cases} \qquad \text{where}$$

$H_1(\rho) := \frac{\mu_1\rho}{\mu_1\rho+\mu_2(1-\rho)} - \rho,\ H_2(\rho) := \frac{\mu_2\rho}{\mu_2\rho+\mu_1(1-\rho)} - \rho,\ \rho \in [0,1]$

It holds: $H_1(\rho) \leq 0,\ H_2(\rho) \geq 0.$

The stationary policy $(f^*, f^*, \ldots)$ is optimal with

$$f^*(i_1, i_2, \rho) = \begin{cases} 1 & i_2 = 0 \\ 2 & i_1 = 0 \\ 1 & \rho \leq \frac{1}{2},\ (i_1, i_2) \in \mathbb{N} \times \mathbb{N} \\ 2 & \rho > \frac{1}{2},\ (i_1, i_2) \in \mathbb{N} \times \mathbb{N} \end{cases}$$

$\bar{\mu}_1 := \mu_1\rho + \mu_2(1-\rho),\ \bar{\mu}_2 := \mu_2\rho + \mu_1(1-\rho)$

$\bar{\mu}_1 \geqslant \bar{\mu}_2 \Longleftrightarrow \rho \leq \frac{1}{2}$

certainty equivalence principle for the μc-rule holds (if $c_1 = c_2$)!

Rieder/Winter (2009), Bäuerle/Rieder (2009)

References

Bäuerle/Rieder (2011) : Markov decision processes with applications to finance.

Bertsekas/Shreve (1978): Stochastic optimal control.

Feinberg/Schwartz (2002): Handbook of Markov decision processes.

Hernandez-Lerma/Lasserre (1996): Discrete-time Markov control processes.

Hinderer(1970): Foundations of non-stationary dynamic programming.

Puterman (1994): Markov decision processes.

Davis (1993): Markov models and optimization.

Elliott/Aggoun/Moore (1995): Hidden Markov models.

Gittins (1989): Bandit processes and dynamic allocation processes.

Guo/Hernandez-Lerma (2009): Continuous-time Markov decision processes.

Jacobsen (2006): Point process theory and applications.

Jeanblanc/Yor/Chesney (2009): Mathematical methods for financial markets.

Rieder/Winter (2009): Optimal control of Markovian jump processes with partial information and applications to a parallel queueing model. Math. Meth. Operat. Res.70, 567-596.

Bäuerle/Rieder (2009): MDP algorithms for portfolio optimization problems in pure jump markets. Finance Stoch. 13, 591-611.