The 9-th Workshop on Markov Processes and Related Topics

Minimizing risk probability in semi-Markov decision processes

XIANPING GUO, YONGHUI HUANG

Sun Yat-Sen University, Guangzhou

6-13 July, 2013, Chengdu

Outline

- Motivation
- Semi-Markov decision processes
- Optimality problem
- Main results
- Numerical example

1 Motivation

Background-1: Reliability engineering

Problem-1:

$$\text{maximize } \mathbb{P}^{\pi}_{i,\lambda}(\tau_B > \lambda) \text{ over } \pi$$

- i is an initial state;
- λ is a reward level;
- π is a policy;
- B is a given target set;
- τ_B is a first passage time to B.

Background-2: Risk analysis

Generalized problem-2:

$$\text{maximize } \mathbb{P}^{\pi}_{i,\lambda}\left(\int_0^{\tau_B} r(x(t),a(t))dt > \lambda\right) \text{ over } \pi$$

The equivalent problem:

$$\inf_{\pi} \mathbb{P}^{\pi}_{i,\lambda}\left(\int_0^{\tau_B} r(x(t),a(t))dt \le \lambda\right),$$

- $r(i,a)$ is the reward function;
- $x(t)$ is the state process;
- $a(t)$ is the action process.

Existing work:

- Bouakiz, Kebir (1995);
- White (1993);
- Ohtsubo, Toyonaga (2002);
- $\cdots\cdots\cdots$

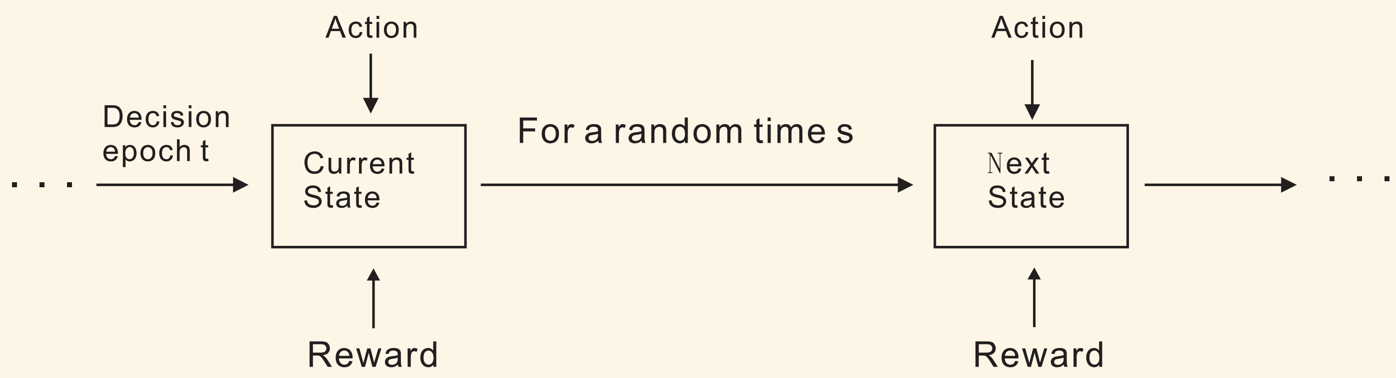
The works are on discrete-time Markov decision processes!

Motivation:

- DTMDP $\Rightarrow$ SMDP ???

2 Semi-Markov decision processes

The description of SMDP:



The difference between SMDP and DTMDP:

- DTMDP: all decisions are made at fixed points $n = 0, 1, \ldots$, and thus the time between successive decisions is a constant, say 1;
- SMDP: all decisions are made at jump points, and the time between successive decisions is a variable, with a distribution $Q(t, j|i, a)$, which depends on the current state i, the action a taken at i, and the jump-in state j from i.

The model of SMDP:

$$\Big\{E, (A(i), i \in E), Q(t, j|i, a), r(i, a)\Big\}$$

where

- E: the state space, a denumerable set;
- $A(i)$: finite set of actions available at $i \in E$;
- $Q(t, j|i, a)$: semi-Markov kernel, $a \in A(i), i, j \in E$;
- $r(i, a)$: the reward rate, $a \in A(i), i \in E$.

Notation:

- Policy π: a sequence $\pi = \{\pi_n, n = 0, 1, \ldots\}$ of stochastic kernels π_n on the action space A given H_n satisfying

$$\pi_n(A(i_n)|0, i_0, \lambda_0, a_0, \ldots, t_{n-1}, i_{n-1}, \lambda_{n-1}, a_{n-1}, t_n, i_n) = 1;$$

- Stationary policy: measurable f, $f(i, \lambda) \in A(i)$ for all (i, λ);
- $\mathbb{P}^{\pi}_{(i,\lambda)}$: probability measure on $(E \times [0, \infty)^2 \times (\cup_{i \in S} A(i)))^{\infty}$;
- i_n, a_n: n-th the state variable, action variable, respectively;
- T_n: n-th decision epoch.

Semi-Markov decision process $\{(x(t), a(t), t \geq 0\}$:

$$x(t) = i_n, a(t) = a_n, \ \text{ for } \ T_n \leq t < T_{n+1}, t \geq 0.$$

Let

$$T_\infty := \lim_{n\to\infty} T_n.$$

Assumption A. There exist $\delta > 0$ and $\epsilon > 0$ such that

$$\sum_{j\in E} Q(\delta, j|i,a) \leq 1-\epsilon, \text{ for all } i \in E, a \in A(i).$$

Assumption A $\Rightarrow \mathbb{P}^{\pi}_{(i,\lambda)}(\{T_\infty = \infty\}) = 1$

The first passage time into B, is defied by

$$\tau_B := \inf\{t \geq 0 \mid x(t) \in B\}, \ \ (\text{with } \inf \emptyset := \infty).$$

3 The optimality problem

The risk probability (of policy π):

$$p^{\pi}(i,\lambda) := \mathbb{P}^{\pi}_{(i,\lambda)}\Big(\int_0^{\tau_B} r(x(t),a(t))dt \le \lambda\Big)$$

The optimal value:

$$p^{*}(i,\lambda) := \inf_{\pi\in\Pi} \mathbb{P}^{\pi}(i,\lambda),$$

Definition 1. A policy $\pi^* \in \Pi$ is called optimal if

$$p^{\pi^*}(i,\lambda) = p^{*}(i,\lambda) \ \ \forall\,(i,\lambda) \in E \times R.$$

- Existence and computation of optimal policies ???

4 Main results

Notation:

For $i \in B^c$, $a \in A(i)$, and $\lambda \geq 0$, let

$$T^a u(i,\lambda) := Q(\lambda/r(i,a), B|i,a) + \sum_{j \in B^c} \int_0^{\lambda/r(i,a)} Q(dt, j|i,a) u(j, \lambda - r(x,a)t),$$

with $u \in \mathcal{F}_{[0,1]}$ (the set of measurable functions $u : B^c \times R \to [0,1]$),

$$Q(\lambda/r(i,a), B|i,a) := \sum_{j \in B} Q(\lambda/r(i,a), j|i,a), \quad T^a u(i,\lambda) := 0 \text{ for } \lambda < 0.$$

Then, define operators T and T^f:

$$Tu(i,\lambda) := \min_{a \in A(i)} T^a u(i,\lambda); \quad T^f u(i,\lambda) := T^{f(i,\lambda)} u(i,\lambda),$$

for each stationary policy f.

Theorem 1. Under Assumption A, we have

(a) $p^f = \lim_{n\to\infty} u_n^f$, where $u_n^f := T^f u_{n-1}, u_0^f := 1$;

(b) p^f satisfies the equation, $u = T^f u$, for each stationary policy f.

Remark 1.

- Theorem 1 gives an approximation to the risk probability p^f.

Theorem 2. Under Assumption A, we have

(a) $\lim_{n\to\infty} p_n^* = p^*$, where

$$p_0^*(i,\lambda) := 1,\ p_{n+1}^*(i,\lambda) := Tp_n^*(i,\lambda),\ \ n \geq 0;$$

(b) p^* satisfies the optimality equation: $p^* = Tp^*$;

(c) p^* is the maximal fixed point of T in $\mathcal{F}_{[0,1]}$.

Remark 2.

- Theorem 2(a) gives a value iteration algorithm for computing the optimal value p^*.
- Theorem 2(b) establishes the optimality equation.

To ensure the existence of optimal policies, we need the following condition.

Assumption B. For every $(i, \lambda) \in B^c \times R$ and f,

$$\mathbb{P}^f_{(i,\lambda)}(\tau_B < \infty) = 1.$$

To verify Assumption B, we have a fact below:

Proposition 3. If there exists a constant $\alpha > 0$ such that

$$\sum_{j \in B} Q(\infty, j|i, a) \geq \alpha \ \text{ for all } i \in B^c, a \in A(i),$$

then Assumption B holds.

Theorem 3. Under Assumptions A and B, we have

(a) p^f and p^* are the unique solution in $\mathcal{F}_{[0,1]}$ to equations $u = T^f u$ and $u = Tu$, respectively;

(b) any f, such that $p^* = T^f p^*$, is optimal;

(c) there exists a stationary policy f^* satisfying the optimality equation:

$$p^* = Tp^* = T^{f^*} p^*,$$

and such a policy f^* is optimal.

Remark 2.

- Theorem 3(c) shows the existence of an optimal policy, and moreover, provides a way of finding an optimal policy.

5 Numerical example

Example 5.1. Let $E = \{1, 2, 3\}$, B=$\{3\}$, where

- state 1: the good state;
- state 2: the medium state;
- state 3: the failure state.

Let $A(1) = \{a_{11}, a_{12}\}$, $A(2) = \{a_{21}, a_{22}\}$, $A(3) = \{a_{31}\}$.

The reward rates are as below:
$r(1, a_{11}) = 1$, $r(1, a_{12}) = 2$, $r(2, a_{21}) = 0.5$, and $r(2, a_{22}) = 0.8$.

The semi-Markov kernel is of the form:

$$Q(t, j \mid i, a) = G(t \mid i, a)p(j \mid i, a)$$

where

- $G(t \mid i, a)$: the distribution functions of the sojourn time
- $p(j \mid i, a)$: the transition probabilities.

Let $G(t \mid i, a)$ be of the form:

$$G(t|1, a_{11}) = \begin{cases} 1/25, & t \in [0, 25], \\ 1, & t > 25; \end{cases} \quad G(t|1, a_{12}) = 1 - e^{-0.16t}, \ t \in R_+;$$
$$G(t|2, a_{21}) = \begin{cases} 1/40, & t \in [0, 40], \\ 1, & t > 40; \end{cases} \quad G(t|2, a_{22}) = 1 - e^{-0.08t}, \ t \in R_+;$$
$$G(t|3, a_{31}) = 1 - e^{-0.2t}, \ t \in R_+;$$

and $p(j \mid i, a)$ is given by

$$p(1|1, a_{11}) = 0, \ \ p(2|1, a_{11}) = 0.7, \ \ p(3|1, a_{11}) = 0.3;$$
$$p(1|1, a_{12}) = 0, \ \ p(2|1, a_{12}) = 0.6, \ \ p(3|1, a_{12}) = 0.4;$$
$$p(1|2, a_{21}) = 0.2, \ \ p(2|2, a_{21}) = 0, \ \ p(3|2, a_{21}) = 0.8;$$
$$p(1|2, a_{22}) = 0.1, \ \ p(2|2, a_{22}) = 0, \ \ p(3|2, a_{22}) = 0.9; \ \ p(3|3, a_{31}) = 1.$$

In this Example, Assumptions A and B are fulfilled.

Using the value iteration algorithm in Theorem 2, we obtain some computational results as in Figure 1 and Figure 2.

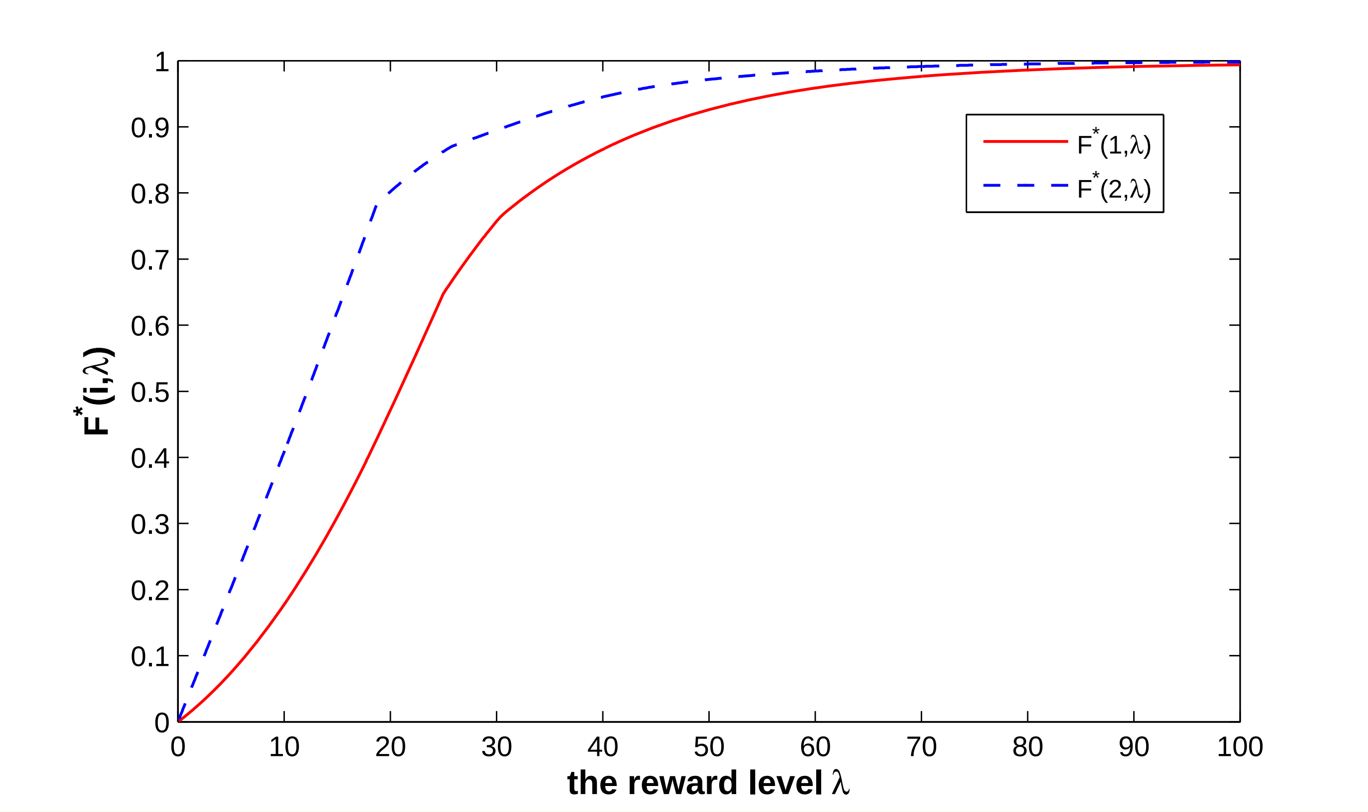


Figure 1. The value function $p^*(i, \lambda)$

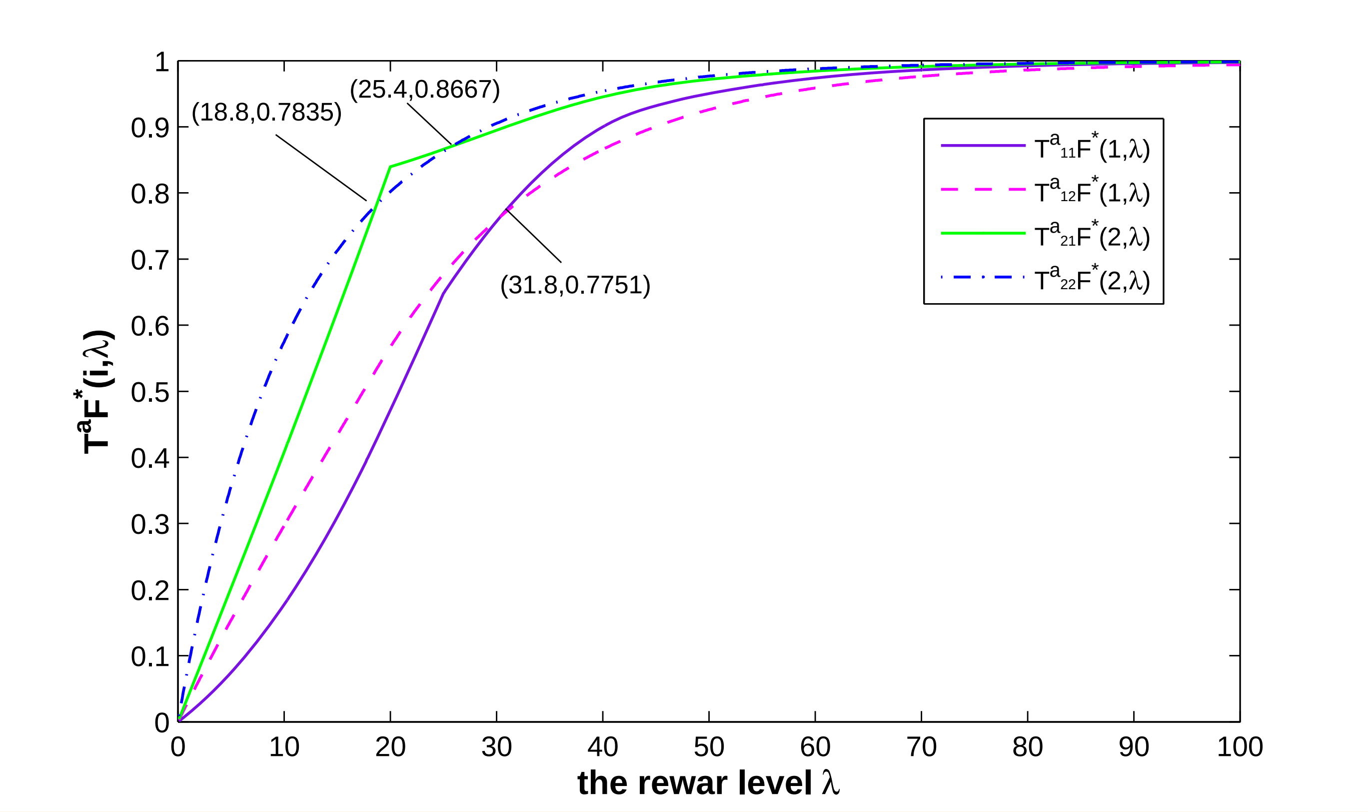


Figure 2. The functions $T^a p^*(i, \lambda)$

Define a policy f^* by

$$f^*(1,\lambda) = \begin{cases} a_{11}, & 0 \le \lambda \le 31.8, \\ a_{12}, & 31.8 < \lambda \le 100, \\ a_{11}, & \lambda > 100, \end{cases} \quad f^*(2,\lambda) = \begin{cases} a_{21}, & 0 \le \lambda \le 18.8, \\ a_{22}, & 18.8 < \lambda \le 25.4, \\ a_{21}, & 25.4 < \lambda \le 100, \\ a_{22}, & \lambda > 100. \end{cases}$$

Then, we have

- $p^*(i,\lambda) = T^{f^*} p^*(i,\lambda)$, for $i = 1, 2$ and all $\lambda \ge 0$;
- f^* is an optimal stationary policy.

Many Thanks!