Risk probability minimization problems for infinite discounted piecewise deterministic Markov decision processes

Haifeng Huo

School of Science, Guangxi University of Science and Technology, Liuzhou

The 18th Workshop on Markov Processes and Related Topics

2023.7.30-2023.8.2.

Outline

1. Introduction
2. The control model
3. The main results
4. Example

Introduction

- The probability criterion :

$$\text{minimize} \quad \mathbb{F}^{\pi}(x,\lambda) := P^{\pi}_{(x,\lambda)}\Big(\int_0^{+\infty} e^{-\alpha s} r(x_s,\pi_s)\mathrm{d}s \le \lambda\Big).$$

The corresponding standard expectation criterion:

$$V^{\pi}(x) := E^{\pi}_{x}\left(\int_0^{+\infty} e^{-\alpha s} r(x_s,\pi_s)\mathrm{d}s\right).$$

The minimization problem of the risk probability is an important class of topics in the areas of Markov decision processes (MDPs).

- The earlier works : Sobel (1982), White (1993), Wu and Lin (1999), Ohtsubo and Toyonaga (2002).

- The recent works: Huang and Guo (2013), Huang Zou and Guo (2015), Huo Zou and Guo (2017), Huo and Guo (2020), Wen,Huo and Guo (2022)et al.

After reviewing the literature, we find that, as an important case, infinite discounted PDMDPs for the probability criterion with un-bounded transition rates have not been studied. For this case,

- Only using the assumption of non-explosion of the controlled state processes as well as the finiteness of actions available at each state, we not only establish the existence and uniqueness of a solution to the corresponding optimality equation, prove the existence of an optimal policy, but also provide a value iteration algorithm for computing both the value function and an optimal policy.

2 The control model

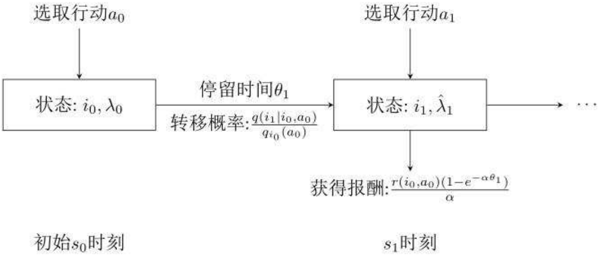
Our PDMDP model is a set of data as below

$$\Big\{E, (A(x) \subset A, x \in E), q(\cdot|x,a), \phi(x,t), r(x,a), \alpha(x)\Big\}, \tag{2.1}$$

consisting of

- E: State space, a Borel state space endowed with a Borel σ-algebra $\mathcal{B}(E)$;
- $A(x)$: Finite sets of actions available at $x \in E$ for the decision marker; Let $K := \{(x,a)|x \in E, a \in A(x)\}$ be the set of pairs of states and actions;
- $q(\cdot|x,a)$: Transition rates are satisfied $0 \leq q(D|x,a) < +\infty$ with $(x,a) \in K, x \notin D \in \mathcal{B}(E)$. The transition rates are conservative (i.e.,$q(E|x,a) = 0$) and stable (i.e., $q^*(x) := \sup_{a \in A(x)} q_x(a) < \infty$), for all $(x,a) \in K$, where $q_x(a) := -q(x|x,a) \geq 0$.

- $\phi(x,t)$: Deterministic flow function from $E \times R^+$ to E, and satisfies (i) $\phi(x,t+s) = \phi(\phi(x,t),s)$; (ii) $\phi(x,\cdot)$ is continuous on $R^+ := [0,+\infty)$ for any $s,t \geq 0$.
- $r(x,a)$: Nonnegative reward/cost function for decision marker under actions $a \in A(x)$ at state x.
- $\alpha(x)$: Discount factor depends on the state $x \in E$.



选取行动a_0

状态: i_0, λ_0

停留时间θ_1

转移概率:$\frac{q(i_1|i_0,a_0)}{q_{i_0}(a_0)}$

选取行动a_1

状态: $i_1, \hat{\lambda}_1$

$\cdots$

获得报酬:$\frac{r(i_0,a_0)(1-e^{-\alpha\theta_1})}{\alpha}$

初始s_0时刻

s_1时刻

Figure 1: The evolution of CTMDPs

Based on the observation information (x_0, λ_0) of the system, the control action $a_0 \in A(x_0)$ is chosen by the decision maker. Consequently, the system evolves in two ways:

(i) The change of the system state is based on the flow $\phi(x_0, s)$ $(s \in [s_0, s_1))$ up to s_1. At this point, the system state jumps into a new state $x_1 \in E$.

(ii) During the period of time $[s_0, s_1)$, the decision maker obtains the rewards $\int_0^{s_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0, s), a_0)\mathrm{d}s$. The remaining reward goal becomes

$$\begin{aligned}\hat{\lambda}_1 &= e^{\int_0^{s_1} \alpha(\phi(x_0,t))\mathrm{d}t} \\ &\quad \times(\lambda_0 - \int_0^{s_1} e^{-\int_0^s \alpha(\phi(x_0,t))\mathrm{d}t} r(\phi(x_0, s), a_0)\mathrm{d}s).\end{aligned}$$

Due to the decision maker consider the reward goal as well as the system states when making decisions, so we need to reconstruct a probability space.

The canonical construction is as follows: Let $E_\Delta := E\bigcup\{\Delta\}$ (with some isolated point $\Delta \notin E$), $\Omega^0 := E \times [0, +\infty) \times ((0, +\infty] \times E \times [0, +\infty))^\infty$, $\Omega := \Omega^0 \bigcup\{(x_0, \lambda_0, s_1, x_1, \lambda_1, \ldots, s_k, x_k, \lambda_k, \ldots, \infty, \Delta, \infty, \ldots)|x_0 \in E, \lambda_0 \in [0, +\infty), s_l \in (0, \infty], x_l \in E, \lambda_l \in [0, +\infty)$, for each $1 \leq l \leq k, k \geq 1\}$, and let $\mathcal{F}$ be the Borel σ-algebra on Ω. Then we get the measurable space $(\Omega, \mathcal{F})$.

For each $k \geq 0$, $e := (x_0, \lambda_0, s_1, x_1, \lambda_1, \ \ldots, \ s_k, x_k, \lambda_k, \ldots) \in \Omega$, let $h_0(e) := (x_0, \lambda_0)$, $h_k(e) := (x_0, \lambda_0, s_1, x_1, \lambda_1, \ \ldots, \ s_k, x_k, \lambda_k)$ denote the k-component internal history, and define

$$S_k(e) := s_k, \quad X_k(e) := x_k, \quad \Lambda_k(e) := \lambda_k.$$

In what follows, the argument e is always omitted. Let $S_\infty := \lim_{k\to\infty} S_k$ and define the state process $\{x_s, s \geq 0\}$ by

$$x_s := \sum_{k\geq 0} I_{\{S_k \leq s < S_{k+1}\}} \phi(X_k, s - S_k) + \Delta I_{\{s \geq S_\infty\}}. \tag{2.2}$$

Here I_D stands for the indicator function on any set D.

Definition 1

- History-dependent policy : $\pi = \{f_0, f_1, \ldots\}$ is defined as follows.

$$\begin{aligned}\pi(e,s) &= I_{\{s=0\}} f_0(h_0(e)) + \sum_{k\geq 0} I_{\{S_k < s \leq S_{k+1}\}} f_k(h_k(e)) \\ &\quad + I_{\{s \geq S_\infty\}} \delta_{a_\Delta}(da),\end{aligned} \tag{2.3}$$

where $f_k (k \geq 0)$ is a decision function from Ω onto A_Δ, and $\delta_{a_\Delta}(da)$ denotes the Dirac measure on $A_\Delta (A_\Delta := A \cup \{a_\Delta\})$ at the isolated point a_Δ.

- Markov policy: if $f_k(h_k(e))$ depend only on the current (x_k, λ_k) for all $k \geq 0$ and $e \in \Omega$.
- Stationary policy: if all f_k are equal to one decision function f, and such a stationary policy is denoted as f for simplicity.

We denote by Π the set of all history-dependent policies, by Π_m the set of all Markov policies, and by Π_s the set of all stationary policies.

For any initial $(x,\lambda)\in E\times R^{+}$and policy $\pi\in\Pi$, by the extension of the well-known Ionescu Tulcea theorem, there exists a unique probability space $(\Omega,\mathcal{F},P^{\pi}_{(x,\lambda)})$.

$E^{\pi}_{(x,\lambda)}$: the expectation operator associated with $P^{\pi}_{(x,\lambda)}$.

Probability criterion: For each $(x,\lambda)\in E\times R^{+}$ and $\pi\in\Pi$,

$$F^{\pi}(x,\lambda):=P^{\pi}_{(x,\lambda)}\Big(\int_0^{+\infty}e^{-\int_0^s\alpha(x_s)\mathrm{d}u}r(x_s,\pi_s)\mathrm{d}s\le\lambda\Big),$$

which measures the risk of the control system.

Definition 2

A policy $\pi^* \in \Pi$ such that

$$F^{\pi^*}(x,\lambda) = \inf_{\pi\in\Pi} F^{\pi}(x,\lambda), \quad (x,\lambda) \in E \times R^+ \tag{2.4}$$

is said to be an *optimal policy*, and the value function is defined as follows:

$$F^{*}(x,\lambda) := \inf_{\pi\in\Pi} F^{\pi}(x,\lambda), \quad (x,\lambda) \in E \times R^+. \tag{2.5}$$

Main goals: Existence and computation of optimal policies.

3 The main results

$\mathcal{G}_m$: The set of function $F : E \times R^+ \to [0,1]$, such that $F(\cdot,\cdot)$ is Borel measurable on $E \times R^+$.

$((M^f)^n F, n \geq 1), (M^n F, n \geq 1)$: The operators on $\mathcal{G}_m$ are defined as follows: For any $(x,\lambda) \in E \times R^+, f \in \Pi_s$ and $a \in A(x)$,

$$\begin{aligned} M^a F(x,\lambda) \quad := \quad & I_{[0,\lambda]}\left(\int_0^{+\infty} e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s\right) \\ & \times e^{-\int_0^{+\infty} q_{\phi(x,s)}(a)\mathrm{d}s} \\ & + \int_0^{+\infty} \int_{E\setminus\{\phi(x,u)\}} F\Big(y, e^{\int_0^u \alpha(\phi(x,t))\mathrm{d}t} \\ & \times (\lambda - \int_0^u e^{-\int_0^s \alpha(\phi(x,t))\mathrm{d}t} r(\phi(x,s),a)\mathrm{d}s)\Big) \\ & \times e^{-\int_0^u q_{\phi(x,s)}(a)\mathrm{d}s} q(\mathrm{d}y|\phi(x,u),a)\mathrm{d}u, \end{aligned}$$

$$
\begin{aligned}
M^f F(x,\lambda) &:= M^{f(x,\lambda)} U(x,\lambda), \\
MF(x,\lambda) &:= \min_{a\in A(x)} M^a F(i,\lambda) \\
(M^f)^{n+1} F &= M^f((M^f)^n F), \\
M^{n+1} F &= M(M^n F), n \geq 1.
\end{aligned}
$$

where $q_x(f) := -q(x|x, f(x,\lambda))$.

In the following, we give some basic results.

Lemma 3

For any $a \in A(x), (x, \lambda) \in E \times R^{+}$, the following assertions hold:

(a) *If $F, G \in \mathcal{G}_m$, and $F \geq G$, then $M^a F(x, \lambda) \geq M^a G(x, \lambda)$, $MF(x, \lambda) \geq MG(x, \lambda)$,*

(b) *If $F \in \mathcal{G}_m$, then $MF \in \mathcal{G}_m$, and there exists an $f \in \Pi_s$ such that $MF(x, \lambda) = M^f F(x, \lambda)$.*

To ensure the states processes $\{x_s, s \geq 0\}$ is non-explosive, we need the following basic assumption and "drift condition".

Assumption 1

For any $\pi \in \Pi$, $P^{\pi}_{(x,\lambda)}(S_\infty = \infty) = 1$.

Lemma 4

There exist a measurable function $W \geq 1$ *on* E *and constant* $c_0 > 0$, $b_0 \geq 0$ *such that*

(a) $\int_E W(\phi(y,s))q(\mathrm{d}y \mid x,a) \leq c_0 W(\phi(x,s)) + b_0$, *for all* $(x,a) \in K, s \geq 0$;

(b) *There is a sequence* $\{E_n, n \geq 1, E_n \subseteq E\}$ *which satisfies* $E_n \uparrow E$, $\lim_{n\to\infty} \inf_{x \notin E_n} W(\phi(x,s)) = \infty$, $\sup_{x \in E_n} q^*(x) < \infty$ *for all* $n \geq 1$ *with* $s \geq 0$, $q^*(x) = \sup_{a \in A(x)} q_x(a)$.

Then Assumption 1 holds.

For each $(x,\lambda) \in E \times R^+$ and $\pi \in \Pi$, since the states processes $\{x_s, s \geq 0\}$ is non-explosion, we can rewrite $F^\pi(x,\lambda)$ as follows:

$$\begin{aligned} F^\pi(x,\lambda) &= P^\pi_{(x,\lambda)}\Big(\int_0^{+\infty} e^{-\int_0^s \alpha(x_t)\mathrm{d}t} r(x_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\ &= P^\pi_{(x,\lambda)}\Big(\sum_{m=0}^{\infty}\int_{S_m}^{S_{m+1}} e^{-\int_0^s \alpha(x_t)\mathrm{d}t} r(x_s,\pi_s)\mathrm{d}s \leq \lambda\Big) \\ &= \lim_{n\to\infty} F^\pi_n(x,\lambda), \end{aligned}$$

where $F^\pi_n(x,\lambda) := P^\pi_{(x,\lambda)}\Big(\sum_{m=0}^{n}\int_{S_m}^{S_{m+1}} e^{-\int_0^s \alpha(x_t)\mathrm{d}t} r(x_s,\pi_s)\mathrm{d}s \leq \lambda\Big)$, $F^\pi_{-1}(x,\lambda) := 1$.

Obviously, $F^\pi_n(x,\lambda) \geq F^\pi_{n+1}(x,\lambda), n \geq -1$.

The following lemma is required to establish the optimality equation.

Lemma 5

Under Assumption 1, for any $(x,\lambda) \in E \times R^{+}$, $n \geq -1$, $\pi = \{f_0, f_1, \ldots\} \in \Pi$, *the following statements hold.*

(a) $F_n^{\pi} \in \mathcal{G}_m$ *and* $F^{\pi} \in \mathcal{G}_m$.

(b) $F_{n+1}^{\pi}(x,\lambda) = M^{f_0} F_n^{^1\pi}(x,\lambda), F^{\pi}(x,\lambda) = M^{f_0} F^{^1\pi}(x,\lambda)$, *where* $^1\pi := (\hat{f}_0, \hat{f}_1, \ldots)$ *being the 1-shift policy of* π, $\hat{f}_k(s_1, x_1, \lambda_1, \cdots, s_{k+1}, x_{k+1}, \lambda_{k+1}) := f_{k+1}(x, \lambda, s_1, x_1, \lambda_1, \cdots, s_{k+1}, x_{k+1}, \lambda_{k+1})$

In particular, for $f \in \Pi_s$, $F^f(x,\lambda) = M^f F^f(x,\lambda)$.

Lemma 6 implies that the iteration technique can be used to compute the value function.

Lemma 6

Under Assumption 1, for any $(x,\lambda) \in E \times R^{+}$*, let* $F^{*}_{-1}(x,\lambda) := 1, F^{*}_{n+1}(x,\lambda) := MF^{*}_{n}(x,\lambda), n \geq -1$*. Then,* $\{F^{*}_{n}(x,\lambda), n \geq -1\}$ *is monotone nondecreasing, and*
$\lim_{n\to\infty} F^{*}_{n}(x,\lambda) = F^{*}(x,\lambda)$*.*

We establish the optimality equation and show the existence of it' solution.

Theorem 7

For $(x, \lambda) \in E \times R^+$. Under Assumption 1, the following assertions hold.

(a) *$F^*(x, \lambda)$ satisfies optimality equation: $F^*(x, \lambda) = MF^*(x, \lambda)$.*

(b) *There exists an $f \in \Pi_s$ such that $F^*(x, \lambda) = M^f F^*(x, \lambda)$.*

To further establish the uniqueness of a solution to the optimality equation, we need the following fact.

Theorem 8

For all $(x,\lambda) \in E \times R^+$. Suppose that Assumption 1 holds.

(a) *For any $f \in \Pi_s$, if $F(x,\lambda) - G(x,\lambda) \leq M^f(F-G)(x,\lambda)$, then $F(x,\lambda) \leq G(x,\lambda)$.*

(b) *For any $f \in \Pi_s$, F^f is the unique solution in $\mathcal{G}_m$ to the equation $F(x,\lambda) = M^f F(x,\lambda)$.*

We show that the value function is unique solution to the optimality equation.

Theorem 9

Suppose that Assumption 1 holds. Then

(a) *F^* is the unique solution to the equation $F(x,\lambda) = MF(x,\lambda)$.*

(b) *There exists an $f^* \in \Pi_s$ such that $F^*(x,\lambda) = M^{f^*}F^*(x,\lambda)$, and $F^*(x,\lambda) = F^{f^*}(x,\lambda)$.*

(c) *The policy $\pi^* := (\tilde{f}_0, \tilde{f}_1, \ldots, \tilde{f}_k)$ is optimal, where $\tilde{f}_0(x,\lambda) : = f^*(x,\lambda)$, and for $k \geq 1$, $\tilde{f}_k(x,\lambda,s_1,x_1,\lambda_1,\ldots,s_k,x_k,\ \lambda_k) := f^*(x_k,\tilde{\lambda}_k)$.*

4 Example

Example 1 (Optimal production management):

Consider a production management system of a industry corporation where a state $x \in E := [0, +\infty)$ denotes the number of products, the constant $k > 0$ represents the product quantity threshold. When the company has a small amount of products $x \in (0, k)$, the decision maker can use the production plan $a \in \{a_{11}, a_{12}\}$ to expand the production scale and get some rewards at the rate $r(x, a) \geq 0$.

When the company has a lot of products $x \in [k, +\infty)$, the decision maker can choose the production plan $a \in \{a_{11}, a_{12}, a_{21}, a_{22}\}$ to obtain some rewards at the rate $r(x, a) \geq 0$, When the company doesn't have any products $x = 0$, the decision-maker cannot choose any production plan (which is denoted by a_{01}) and will not receive any reward $r(0, a_{01}) = 0$.

We formulate this production management system as a PDMDP with $k = 2$: The state space $E = [0, +\infty)$; the action sets $A(0) = \{a_{01}\}$, $A(x) = \{a_{11}, a_{12}\}$ for any $x \in (0, 2)$, $A(x) = \{a_{11}, a_{12}, a_{21}, a_{22}\}$ for any $x \in [2, +\infty), \phi(x, s) = xe^s$.

Suppose that the transition rates are given as follows: for any $D \in \mathcal{B}(E), q(D|0, a_{01}) = 0$. For $x \in (0,2) \cup (2, +\infty)$,

$$q(D|x, a_{11}) = \begin{cases} 0.056, & D = \{0\}; \\ -0.28, & D = \{x\}; \\ 0.224, & D = \{2\}; \\ 0, & \text{others}. \end{cases} \tag{4.1}$$

$$q(D|x, a_{12}) = \begin{cases} 0.056, & D = \{0\}; \\ -0.08, & D = \{x\}; \\ 0.024, & D = \{2\}; \\ 0, & \text{others}. \end{cases} \tag{4.2}$$

For $x = 2$,

$$q(D|2, a_{11}) = \begin{cases} 0.28, & D = \{0\}; \\ -0.28, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \tag{4.3}$$

$$q(D|2, a_{12}) = \begin{cases} 0.08, & D = \{0\}; \\ -0.08, & D = \{2\}; \\ 0, & \text{others.} \end{cases} \tag{4.4}$$

For $x \in [2, +\infty)$,

$$q(D|x, a_{21}) = \begin{cases} 0.084, & D = \{0\}; \\ 0.056, & D = \{1\}; \\ -0.14, & D = \{x\}; \\ 0, & \text{others}. \end{cases} \tag{4.5}$$

$$q(D|x, a_{22}) = \begin{cases} 0.13, & D = \{0\}; \\ 0.13, & D = \{1\}; \\ -0.26, & D = \{x\}; \\ 0, & \text{others}. \end{cases} \tag{4.6}$$

For any $x \in E$, the reward rates are given by

$$\begin{aligned} r(0, a_{01}) &= 0, \quad r(x, a_{11}) = x, \quad r(x, a_{12}) = 2x, \\ r(x, a_{21}) &= \sqrt{x}, \quad r(x, a_{22}) = 2\sqrt{x}. \end{aligned}$$

Main goals: Seek an optimal policy with the minmum risk probability.

First, by $r(0, a_{01}) = 0$, we know that the state 0 is absorbing and $F^*(0, \lambda) = 1$ for $\lambda \geq 0$.

Next, we will calculate the value function $F^*(x, \lambda)$ by value iteration algorithm.

Step 1: Let $n = -1$, and $F^*_{-1}(x, \lambda) = 1$, for $\lambda \in [0, +\infty)$.

Step 2: By Lemma 6, we compute the function.

$$F^*_{n+1}(x, \lambda) \quad = \quad \min_{a \in A(x)} \{M^a F^*_n(x, \lambda)\},$$

Step 3: For any $\lambda \geq 0$, give any sufficiently small ε, if $|F^*_{n+1}(x,\lambda) - F^*_n(x,\lambda)| < \varepsilon$, the iteration stops. Then, go to step 4, the approximate value F^*_{n+1} is usually received as the value F^*; otherwise, return to step 2 and by replacing n with $n+1$.

Step 4: For any $\lambda \geq 0$, plot out the graphs of these functions $M^a F^*(x,\lambda), F^*(x,\lambda)$, see Fig.2-3.

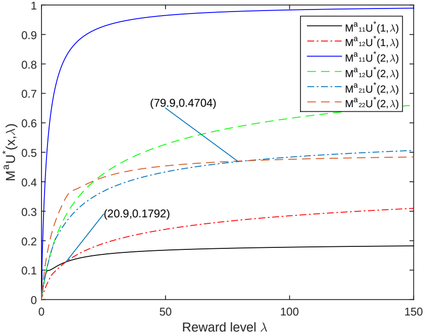


Figure 2: The function $M^aF^*(x,\lambda)$

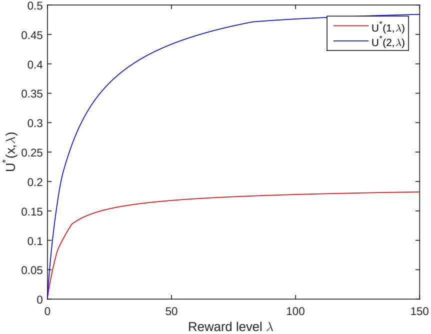


Figure 3: The value function $F^*(x, \lambda)$

From Fig.2-3, we obtain the following conclusions.

(a) In state 1, $M^{a_{12}}F^*(1,\lambda)$ is below $M^{a_{11}}F^*(1,\lambda)$ when $\lambda \in [0,20.9)$, and $M^{a_{11}}F^*(1,\lambda)$ is below $M^{a_{12}}F^*(1,\lambda)$ when $\lambda \in [20.9,+\infty)$. It means that the decision maker should take the action a_{12} with lower risk rather than the action a_{11} if the reward level $\lambda \in [0,20.9)$. While he/she should take the action a_{11} with lower risk rather than the action a_{12} when the reward level $\lambda \in [20.9,+\infty)$.

Similarly, we know that in state 2, the action a_{21} is with lower risk than the action a_{22} when $\lambda \in [0,79.9)$, but the action a_{22} is with lower risk than the action a_{21} when $\lambda \in [79.9,+\infty)$.

(b) When $s_0 = 0$, we obtain an optimal action f^* as follows:

$$f^*(1,\lambda) = \begin{cases} a_{12}, & 0 \le \lambda < 20.9; \\ a_{11}, & \lambda \ge 20.9, \end{cases} f^*(2,\lambda) = \begin{cases} a_{21}, & 0 \le \lambda < 79.9; \\ a_{22}, & \lambda \ge 79.9. \end{cases}$$

Thanks a lot for your attentions!