AIMS Mathematics, 8(12): 28435–28449.
DOI: 10.3934/math.20231455
Received: 31 May 2023
Revised: 06 September 2023
Accepted: 21 September 2023
Published: 18 October 2023

http://www.aimspress.com/journal/Math

Research article

The optimal probability of the risk for finite horizon partially observable Markov decision processes

Xian Wen, Haifeng Huo* and Jinhua Cui

School of Science, Guangxi University of Science and Technology, Liuzhou 541006, China

* **Correspondence:** Email: xiaohuo08ok@163.com; Tel: +17728090183.

Abstract: This paper investigates the optimality of the risk probability for finite horizon partially observable discrete-time Markov decision processes (POMDPs). The probability of the risk is optimized based on the criterion of total rewards not exceeding the preset goal value, which is different from the optimal problem of expected rewards. Based on the Bayes operator and the filter equations, the optimization problem of risk probability can be equivalently reformulated as filtered Markov decision processes. As an advantage of developing the value iteration technique, the optimality equation satisfied by the value function is established and the existence of the risk probability optimal policy is proven. Finally, an example is given to illustrate the effectiveness of using the value iteration algorithm to compute the value function and optimal policy.

Keywords: partially observable Markov decision processes; risk probability criterion; Bayes operator; the optimal policy
Mathematics Subject Classification: 60J27, 90C40

1. Introduction

Analyzing the risk performance of a stochastic dynamic system is an important optimization control problem. Additionally, both theoretical and applied aspects are observed in relation to financial insurance [1], communication networks [2] and queuing systems [3]. Since the conventional expectation criterion could not effectively reflect the risk performance of the system, the criteria of risk probability were first proposed by Sobel [4], and implemented in Markov decision processes (MDPs). Afterwards, many scholars focused on the research of optimization problems of risk probability in MDPs. According to the characteristics of the sojourn time of the system state, these existing studies can be roughly divided into four categories: (i) Discrete-time Markov decision processes (DTMDPs) [5–8]; (ii) Semi-Markov decision processes (SMDPs) [9–11]; (iii) Continuous-time Markov decision processes (CTMDPs) [12–14]; and (iv) Piecewise deterministic Markov decision processes (PDMDPs) [15]. A common

feature of these existing literatures is that the system state is completely observable. However, in practical applications such as machine maintenance and finance, the traditional models of MDPs cannot effectively depict these practical problems because the information of the decision environment cannot be completely observed or perceived. Therefore, it is necessary to establish partially observable Markov decision processes (POMDPs) to optimize the risk probability of the control system.

Compared with completely observable Markov decision processes (COMDPs), the model of POMDPs is a more extensive stochastic control model with important theoretical significance and practical application values, and is widely used in fields such as industry, computational science, finance, and artificial intelligence. Therefore, many scholars began to focus on the problem of expected optimal for POMDPs. More specifically, Drake [16] established the POMDPs model, which attracted the attention of many experts and scholars. Regarding the expected optimal problem, Hinderer [17] discussed the finite state situation. Rhenius [18] and Hern*á*ndez-Lema [19] discussed a more general state situation. Smallwood and Sondik [20] further expanded the algorithm to calculate the optimal strategies and value function (VF) by employing the dynamic programming method. Sawaki and Ichikawa [21] proposed a successive approximation method to calculate the optimal strategy and VF. White and Scherer [22] solved the infinite discounted optimization problem by modifying the reward function and employing the iterative approximation algorithm. B*ä*uerle and Rieder [1] established the optimality equation by equivalently converting POMDPs into a filter MDPs model and proved the existence of an optimal strategy. Feinberg et al. [23] established some sufficient conditions to assure the existence of optimal strategies and an optimality equation for more general state and action spaces. In addition, many scholars have focused on the research of computational algorithms for the POMDPs [24]. However, these criteria mainly focus on the expected value of the total rewards, which could not effectively describe the risk situation faced by the control system. Therefore, it is necessary to introduce the criteria of the risk probability that can effectively demonstrate the risk performance of the system. An overview of the existing literature indicates that the criterion of the risk probability concerning POMDPs has not been researched thus far. This paper is the first attempt to solve the optimization problem of the risk probability for POMDPs.

The optimization problem intends to minimize the risk probability criterion, that is, the probability value of the system's total rewards does not exceed the the profit goal. Because the reward levels are regarded as the second component of the extended state, it is necessary to redescribe the evolution process of the system and redefine the history-dependent, Markov and stationary policies. Thus, for any given redefined policy, a new probability space must be reconstructed using the well-known Ionescu Tulcea's theorem (see e.g., Proposition 7.45 in [25]), which is based on any initial system state and reward goal. Second, the unobservable state's conditional probability distribution is constructed, by redefining the Bayes operator (including the reward levels) and establishing the filter equations. Then, based on the aforementioned conditional probability distribution, a new filtered risk probability MDPs model is established by expanding the state and action space and modifying the transfer kernel and reward function. Furthermore, we prove that the newly filtered MDPs model can reveal the regular relationship between partially and completely observable optimal problems. On account of risk probability optimality theory for COMDPs, by using the value iteration advanced technique, the optimality equation is established, and the existence of optimal policies is proven. Finally, a machine maintenance example is given to present our main results, which include using the iteration algorithm to calculate the value function and an optimal policy.

The rest of the manuscript is outlined as follows. Section 2 presents a minimization risk probability problem dealing for POMDPs. Section 3 presents the main results, including the existence of the optimality equation and optimal policies. An illustration is given to present the value iteration algorithm for calculating the VF and OP in Section 4.

2. The model of POMDPs

The model of POMDPs consists of the following elements:

$$\{E_X \times E_Y, \{A(x) \subseteq A, x \in E_X\}, Q(\cdot, \cdot \mid x, y, a), r(x, y, a), Q_0\} \tag{2.1}$$

which have the following meanings:

(a) $E_X \times E_Y$ represents a Borel space with a Borel σ-algebra $\mathscr{B}(E_X \times E_Y)$. The element $(x, y) \in E_X \times E_Y$ is the system state, where x denotes the state's observable portion, and y denotes the state's unobservable portion.

(b) A represents the action space of a Borel space with a Borel σ-algebra $\mathscr{A}$. $A(x) \subseteq A$ represents the set of admissible actions in state $x \in E_X$, which is assumed to be finite. Moreover, the set of all composable pairs of state actions is denoted by $\mathbb{K} := \{(x, a) \mid x \in E_X, a \in A(x)\}$.

(c) $Q(\cdot, \cdot \mid x, y, a)$ denotes the probability of the transition from $E_X \times E_Y \times A$ to $E_X \times E_Y$, which is used to describe the transition mechanism in the controlled state process. For simplicity, we introduce Q^X to represent the marginal transition probability $Q^X(B|x, y, a) := Q^X(B \times E_Y|x, y, a)$.

(e) $r(x, y, a)$ denotes a nonnegative real-valued measurable reward function from $\mathbb{K} \times E_Y$ to $R^+ := [0, +\infty)$.

(d) Q_0 denotes the initial probability distribution of the unobservable state.

The evolution of the risk probability POMDP is characterized as follows: At $s_0 = 0$, based on the observed state x_0 and the reward goal (reward level) $\tilde{\lambda}_0 := \lambda_0$, the decision maker could pick an action a_0 from the set of allowed actions $A(x_0)$. Then, the observed state of the system stays until time $s_1 = 1$, at which point, the system state transfers to the state $x_1 \in B_1 \subseteq E_X$ based on the probability of the transition $\int_{B_1} \int_{E_Y} Q(x_1, y_1 \mid x_0, y_0, a_0) Q_0(\mathrm{d}y_0)$. Meanwhile, the unobservable state y_0 also transfers to the next state y_1 with a certain probability, which is constructed by the Bayes operator in the undermentioned (3.3). Moreover, during this period, the control system will generate the rewards $r(x_0, y_0, a_0)$. Then, the goal of the corresponding reward would become $\tilde{\lambda}_1 = \lambda_0 - r(x_0, y_0, a_0)$. At the new decision time $s_1 = 1$, based on the observable information of the system $h_1 = (x_0, \lambda_0, a_0, x_1, \tilde{\lambda}_1)$, the decision maker picks a new action $a_1 \in A(x_1)$. Afterward, the system evolves similarly and produces a so-called observable history up to time $s_k = k$:

$$h_k := (x_0, \lambda_0, a_0, x_1, \tilde{\lambda}_1, a_1, \ldots, x_k, \tilde{\lambda}_k), \tag{2.2}$$

where x_k, y_k denote the system state's observable and unobservable section at the k-th moment of decision, respectively, a_k represents the action chosen by the decision maker at time $s_k = k$, $\tilde{\lambda}_k$ denotes the reward goal, which means that the decision maker will try his/her best to regulate the total rewards not exceeding the goal of the reward, and it conforms to the following relation:

$$\tilde{\lambda}_{k+1} := \tilde{\lambda}_k - r(x_k, y_k, a_k), \tag{2.3}$$

for $k = 0, 1, \ldots$.

The sets of all the histories of observable h_k are represented by $H_0 := E_X \times R$, $H_k := H_{k-1} \times A \times E_X \times R$ for $k \geq 1$. Based on all observable histories, some policies are introduced.

Definition 2.1. (a) A sequence $\pi = \{\pi_k, k \geq 0\}$ is said to be *a randomized history-dependent policy* if a stochastic kernel $\pi_k : H_k \to A$ satisfying the following:

$$\pi_k(A(x_k)|h_k) = 1 \quad \text{for all } h_k \in H_k, k = 0, 1, 2 \ldots.$$

(b) A randomized history-dependent policy is said to be *deterministic* if there exists a sequence $\{g_k\}$ of measurable functions g_k from H_k to A with $g_k(h_k) \in A(x_k)$, and $\pi_k(\cdot|h_k)$ is the Dirac measure at $g_k(h_k)$ for all $h_k \in H_k, k \geq 0$. The sets of all the randomized, deterministic policies are represented by Π, Π_{DH}, respectively.

The risk probability POMDP needs to consider the system state and the reward goal, which is different from the conventional expectation MDP that only considers the system state. Thus, the results of the available classical expectation MDP cannot be applied to the proposed model. First, we need to reconstruct the measurable space $(\Omega, \mathscr{F})$ as follows: $\Omega := \{(x_0, y_0, \lambda_0, a_0, x_1, y_1, \lambda_1, a_1, \ldots, x_k, y_k, \lambda_k\ a_k, \ldots,)|x_0 \in E_X, y_0 \in E_Y, \lambda_0 \in R, a_0 \in A, x_l \in E, y_l \in E_Y, \lambda_l \in R, a_l \in A \text{ with } 1 \leq l \leq k, k \geq 1\}$ denotes the sample space, which is endowed with the Borel σ-algebra $\mathscr{F}$. For any $\omega := (x_0, y_0, \lambda_0, a_0, x_1, y_1, \lambda_1, a_1, \ldots, x_k, y_k, \lambda_k\ a_k, \ldots,) \in \Omega$, some random variables are defined as follows:

$$X_k(\omega) := x_k, Y_k(\omega) := y_k, \Lambda_k(\omega) := \lambda_k, A_k(\omega) := a_k, k \geq 0.$$

The ω will be omitted for convenience.

For any policy $\pi \in \Pi$, $(x, \lambda) \in E_X \times R$, Q_0 of Y_0 on E_Y, based on the Ionescu Tulcea's (e.g., Proposition 7.45 in [25]), the unique probability measure $P^{\pi}_{(x,\lambda)} = \int_{E_Y} P^{\pi}_{(x,\lambda,y)}(\cdot)Q_0(\mathrm{d}y)$ on $(\Omega, \mathscr{F})$ is constructed as follows: for all $B \in \mathscr{B}(E_X), C \in \mathscr{B}(E_Y), G \in \mathscr{B}(R), D \in \mathscr{B}(A), h_k \in H_k, k = 0, 1, \ldots$

$$P^{\pi}_{(x,\lambda,y)}(X_0 = x, \Lambda_0 = \lambda) = 1, \tag{2.4}$$

$$P^{\pi}_{(x,\lambda,y)}(A_k \in D|h_k) = \int_D \pi_k(\mathrm{d}a_k|h_k), \tag{2.5}$$

$$P^{\pi}_{(x,\lambda,y)}(X_{k+1} \in B, Y_{k+1} \in C, \Lambda_{k+1} \in G|h_k, y_k, a_k) = \int_B \int_C \int_G Q(\mathrm{d}x_{k+1}, \mathrm{d}y_{k+1}|x_k, y_k, a_k) \times I_{[\lambda_k - r(x_k, y_k, a_k)]}(\mathrm{d}\lambda_{k+1}). \tag{2.6}$$

The expectation operator corresponding to the probability measure $P^{\pi}_{(x,\lambda)}$ can be expressed as $\mathbb{E}^{\pi}_{(x,\lambda)}$.

For any $(x, \lambda) \in E_X \times R$ and $\pi \in \Pi$, the risk probability criterion of POMDPs is defined as follows:

$$F^{\pi}_N(x, \lambda) = \int_{E_Y} P^{\pi}_{(x,y,\lambda)}\Big(\sum_{n=0}^{N} r(X_n, Y_n, A_n) \leq \lambda\Big)Q_0(\mathrm{d}y). \tag{2.7}$$

Then, $F^*_N(x, \lambda) := \inf_{\pi \in \Pi} F^{\pi}_N(x, \lambda)$ is known as the risk probability value function.

Definition 2.2. A policy $\pi^* \in \Pi$ is called the optimal of the risk probability if

$$F^{\pi^*}_N(x, \lambda) = F^*_N(x, \lambda) \quad \text{for all } (x, \lambda) \in E \times R. \tag{2.8}$$

3. Main result

The general objective of the manuscript is to optimize the criterion of the risk probability and establish both the optimality equation's solution and the existence of the strategy's conditions.
Notation: Let $\mathbb{P}(E_Y)$ be the space of all the probability measures on E_Y.

Assumption 3.1. *Assume that the transition probability has a probability density function q that satisfies $Q(d(x^{'}, y^{'})|x, y, a) = q(x^{'}, y^{'}|x, y, a)\rho(\mathrm{d}x^{'})\nu(\mathrm{d}y^{'})$ for some σ-finite measures ρ and ν.*

The Bayes operator $\Phi : E_X \times R \times A \times E_X \times R \times \mathbb{P}(E_Y) \to \mathbb{P}(E_Y)$ is first defined as follows:

$$\Phi(x, \lambda, a, x^{'}, \lambda^{'}, \mu)(C) := \frac{\int_C \int_{E_Y} q(x^{'}, y^{'}|x, y, a) I_{\{\lambda - r(x,y,a)\}}(\lambda^{'})\mu(\mathrm{d}y)\nu(\mathrm{d}y^{'})}{\int_{E_Y} \int_{E_Y} q(x^{'}, y^{'}|x, y, a)\mu(\mathrm{d}y)\nu(\mathrm{d}y^{'})}, \tag{3.1}$$

where $C \in \mathscr{B}(E_Y)$, μ denotes the distribution of the unobservable state. Furthermore, by using the iterative method, for any $h_k \in H_k, k = 0, 1, \ldots$, the conditional probability distribution μ_n of the unobservable variable Y_n is presented by the following:

$$\mu_0(C|h_0) := Q_0(C), \tag{3.2}$$

$$\mu_{k+1}(C|h_k, a, x^{'}, \lambda^{'}) := \Phi(x_k, \lambda_k, a, x^{'}, \lambda^{'}, \mu_k(\cdot|h_k))(C), \tag{3.3}$$

which are called filter equations.

Lemma 3.1. Under Assumption 3.1, for any $\pi \in \Pi$, $B \in \mathscr{B}(E_Y)$, the following statement holds:

$$P^{\pi}_{(x,\lambda)}(Y_n \in B|X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n) = \mu_n(B|X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n). \tag{3.4}$$

Proof. For each $\pi \in \Pi$ and $x \in E_X, \lambda \in R$, the following result is proven by induction:

$$E^{\pi}_{(x,\lambda)}[V(X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n, Y_n)] = E^{\pi}_{(x,\lambda)}[\hat{V}(X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n)], \tag{3.5}$$

for the bounded and measurable function $V : H_k \times E_Y \to R$ and $\hat{V}(h_n) = \int V(h_n, y_n)\mu_n(\mathrm{d}y_n|h_n)$. Since $\hat{V}(h_0) = \int V(h_0, y)Q_0(\mathrm{d}y)$, Fact (3.5) is true when $n = 0$. For any given $h_{n-1} \in H_{n-1}, n \geq 1$, suppose that the Fact (3.5) holds for $k = n - 1$. Using (2.6), the Bayes operator's definition and Fubini's theorem, we obtain the following:

$$\begin{aligned}
& E^{\pi}_{(x,\lambda)}[\hat{V}(h_{n-1}, A_{n-1}, X_n, \Lambda_n)] \\
= & \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X} \int_A Q^X(\mathrm{d}x_n|x_{n-1}, y_{n-1}, a_{n-1}) \\
& \times \int_R I_{\{\lambda_{n-1} - r(x_{n-1}, y_{n-1}, a_{n-1})\}}(\mathrm{d}\lambda_n)\hat{V}(h_{n-1}, a_{n-1}, x_n, \lambda_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}) \\
= & \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X} \int_A Q^X(\mathrm{d}x_n|x_{n-1}, y_{n-1}, a_{n-1}) \int_R I_{\{\lambda_{n-1} - r(x_{n-1}, y_{n-1}, a_{n-1})\}}(\mathrm{d}\lambda_n) \\
& \times \int_{E_Y} V(h_{n-1}, a_{n-1}, x_n, \lambda_n, y_n)\mu_n(\mathrm{d}y_n|h_{n-1}, a_{n-1}, x_n, \lambda_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1})
\end{aligned}$$

$$
\begin{aligned}
&= \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X}\int_{E_Y}\int_A \rho(\mathrm{d}x_n)\nu(\mathrm{d}y) q(x_n, y|x_{n-1}, y_{n-1}, a_{n-1}) \int_R I_{\{\lambda_{n-1}-r(x_{n-1},y_{n-1},a_{n-1})\}}(\mathrm{d}\lambda_n) \\
&\quad \times \int_{E_Y} V(h_{n-1}, a_{n-1}, x_n, \lambda_n, y_n)\Phi(x_{n-1}, \lambda_{n-1}, a_{n-1}, x_n, \lambda_n, \mu_{n-1})(\mathrm{d}y_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}) \\
&= \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X}\int_{E_Y}\int_A q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1})\rho(\mathrm{d}x_n)\nu(\mathrm{d}y_n) \\
&\quad \times V(h_{n-1}, a_{n-1}, x_n, \lambda_{n-1} - r(x_{n-1}, y_{n-1}, a_{n-1}), y_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}), \qquad (3.6)
\end{aligned}
$$

On the other hand, by induction, we have the following:

$$
\begin{aligned}
&E^{\pi}_{(x,\lambda)}[V(h_{n-1}, A_{n-1}, X_n, \Lambda_n, Y_n)] \\
&= \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X}\int_{E_Y}\int_A Q(\mathrm{d}(x_n, y_n)|x_{n-1}, y_{n-1}, a_{n-1}) \\
&\quad \times \int_R I_{\{\lambda_{n-1}-r(x_{n-1},y_{n-1},a_{n-1})\}}(\mathrm{d}\lambda_n) V(h_{n-1}, a_{n-1}, x_n, \lambda_n, y_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}), \\
&= \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X}\int_{E_Y}\int_A q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1})\rho(\mathrm{d}x_n)\nu(\mathrm{d}y_n) \\
&\quad \times \int_R I_{\{\lambda_{n-1}-r(x_{n-1},y_{n-1},a_{n-1})\}}(\mathrm{d}\lambda_n) V(h_{n-1}, a_{n-1}, x_n, \lambda_n, y_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}), \\
&= \int_{E_Y} \mu_{n-1}(\mathrm{d}y_{n-1}|h_{n-1}) \int_{E_X}\int_{E_Y}\int_A q(x_n, y_n|x_{n-1}, y_{n-1}, a_{n-1})\rho(\mathrm{d}x_n)\nu(\mathrm{d}y_n) \\
&\quad \times V(h_{n-1}, a_{n-1}, x_n, \lambda_{n-1} - r(x_{n-1}, y_{n-1}, a_{n-1}), y_n)\pi_{n-1}(\mathrm{d}a_{n-1}|h_{n-1}), \qquad (3.7)
\end{aligned}
$$

which, together with Eq (3.6), implies the fact that (3.5) is satisfied. Specially, $V(X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n, Y_n) = I_{B\times C}(Y_n, (X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n))$, we obtain the following:

$$
\begin{aligned}
&P^{\pi}_{(x,\lambda)}\Big(Y_n \in B, (X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n) \in C\Big) \\
&= E^{\pi}_{(x,\lambda)}[\mu_n(B|X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n) I_C(X_0, \Lambda_0, A_0, \ldots, X_n, \Lambda_n)],
\end{aligned}
$$

which implies that the Lemma holds.

The partially observable risk probability MDPs can be transformed into the filtered risk probability MDPs by enlarging the state space, modifying the transfer kernel and reward function.

Definition 3.1. The filtered model of POMDPs consists of the following elements $\{E, A, \hat{Q}, \hat{r}\}$, which have the following meanings:

• $E := E_X \times \mathcal{P}(E_Y)$ denotes the state space, and its element is marked as $(x, \mu) \in E$, where x denotes the observable state and μ denotes the unobservable state's conditional probability distribution.

• A denotes the action space. $A(x, \mu) := A(x) \subseteq A$ denotes the class of selectable actions in the state $(x, \mu) \in E$.

• $\hat{r}$ denotes a nonnegative real-valued measurable reward function on $\mathbb{K}$ and satisfies the following:

$$
\hat{r}(x, \mu, a) = \int r(x, y, a)\mu(\mathrm{d}y).
$$

• $\hat{Q}$ denotes the transition law from $E \times R \times A$ to $E \times R$, which is specifically expressed as follows:

$$\begin{aligned}\hat{Q}(B \times C \times D|x, \lambda, \mu, a) &:= \int_B \int_C \int_{E_Y} I_D(\Phi(x, \lambda, a, \hat{x}, \hat{\lambda}, \mu)) I_{\{\lambda - \hat{r}(x,\mu,a)\}}(\mathrm{d}\hat{\lambda}) \\ &\quad \times Q^X(\mathrm{d}\hat{x}|x, \mu, a),\end{aligned}$$

where $Q^X(B|x, \mu, a) = \int_B Q^X(B|x, y, a)\mu(\mathrm{d}y)$ for all $(x, \mu) \in E, \lambda \in R, a \in A(x), B \subset E_X, D \subset \mathcal{P}(E_Y), C \subset R$.

To strictly assure the optimal problems normalization, some notations and the definition of some policies are given in the filtered MDPs. Φ stands for the class of stochastic kernels φ on A provided $E \times R$ with the property $\varphi(A(x)|x, \lambda, \mu) = 1$ for all $(x, \lambda, \mu) \in E \times R$. F stands for the class of all the measurable mappings f from $E \times R$ to A with $f(x, \lambda, \mu) \in A(x)$ for all $(x, \lambda, \mu) \in E \times R$.

Definition 3.2. A *randomized Markov* policy is a sequence $\pi_M = \{\hat{\varphi}_k, k \geq 0\}$ of stochastic kernels $\hat{\varphi}_k \in \Phi$ satisfying $\hat{\varphi}_k(A(x_k)|x_k, \lambda_k, \mu_k) = 1$ for each $\mu_k \in \mathcal{P}(E_Y), k \geq 0$. This randomized Markov policy is represented as $\pi_M = \{\hat{\varphi}_k\}$.

A *randomized Markov policy* $\pi_M = \{\hat{\varphi}_k\}$ is called *a deterministic Markov* if a function sequence $\{f_k, k \geq 0\}$ exists such that $\hat{\varphi}_k(\cdot|x_k, \lambda_k, \mu_k)$ is concentrated at $f_k(x_k, \lambda_k, \mu_k)$ for any $f_k \in F$.

The class of all randomized and deterministic Markov policies are recorded as Π_{RM}, Π_{DM}, respectively. In fact, from the above definition, these randomized Markov policies rely on historical information $h_k, k \geq 0$. Then, for any $\pi_M = \{\hat{\varphi}_0, \hat{\varphi}_1, \ldots\} \in \Pi_M$, we can find a policy $\pi = \{\pi_0, \pi_1, \ldots\} \in \Pi$ that satisfies the following:

$$\pi_0(\mathrm{d}a_0|x_0, y_0, \lambda_0) := \hat{\varphi}_0(\mathrm{d}a_0|x_0, \mu_0(y_0|x_0, \lambda_0), \lambda_0), \tag{3.8}$$

$$\pi_k(\mathrm{d}a_k|h_k) := \hat{\varphi}_k(\mathrm{d}a_k|x_k, \mu_k(y_k|h_k), \lambda_k), \tag{3.9}$$

for $k \geq 0, h_k \in H_k$. Thus, $\Pi_{DM} \subseteq \Pi_{RM} \subseteq \Pi$.

Based on the probability of the transition $\hat{Q}$ and initial distribution μ_0, for any $(x, \lambda, \mu) \in E \times R$ and $\pi \in \Pi$, according to the Ionescu Tulcea theorem (e.g., Proposition 7.45 in [25]), the probability measure $\hat{P}^{\pi}_{(x,\lambda,\mu)}$ can be constructed on $(\Omega, \mathcal{F})$ as follows:

$$\hat{P}^{\pi}_{(x,\lambda,\mu)}(X_0 = x, \Lambda_0 = \lambda) = 1, \tag{3.10}$$

$$\hat{P}^{\pi}_{(x,\lambda,\mu)}(A_k \in G|h_k) = \pi_k(G|h_k), \tag{3.11}$$

$$\begin{aligned}\hat{P}^{\pi}_{(x,\lambda,\mu)}(X_{k+1} \in B, Y_{k+1} \in C, \Lambda_{k+1} \in D|h_k, \mu_k) &= \int_B \int_C \int_D \int_A \\ &\quad \times \hat{Q}(\mathrm{d}x_{k+1}, \mathrm{d}\lambda_{k+1}, \mathrm{d}\mu_{k+1}|x_k, \lambda_k, \mu_k, a_k) \\ &\quad \times \pi_k(\mathrm{d}a_k|h_k),\end{aligned} \tag{3.12}$$

for all $h_k \in H_k, a \in A(x_k), B \subset E_X, C \in E_Y, D \subset R, G \in \mathscr{B}(A)$. The expected operator corresponds to the probability measure $\hat{P}^{\pi}_{(x,\lambda,\mu)}$ and is expressed as $\hat{\mathbb{E}}^{\pi}_{(x,\lambda,\mu)}$.

For any $(x, \lambda, \mu) \in E \times R$ and $\pi \in \Pi$, the value function of the filtered MDP is given by the following:

$$U^{\pi}_N(x, \lambda, \mu) := \hat{P}^{\pi}_{(x,\lambda,\mu)}\Big(\sum_{n=0}^{N} \hat{r}(X_n, \mu_n, A_n) \leq \lambda\Big), \tag{3.13}$$

$$U_N^*(x,\lambda,\mu) := \inf_{\pi\in\Pi} U_N^\pi(x,\lambda,\mu). \tag{3.14}$$

Notation: For any policy $\pi \in \Pi$ and $(x,\lambda,\mu) \in E \times R$, the risk probability of the total rewards U_n^π is defined as follows:

$$U_n^\pi(x,\lambda,\mu) := \hat{P}^\pi_{(x,\lambda,\mu)}\Big(\sum_{k=0}^{n} \hat{r}(X_k,\mu_k,A_k) \le \lambda\Big),$$

with $n = 0, 1, \ldots, N$.

Moreover, the minimal risk probability of the filtered MDPs model is defined by the following:

$$U_n^*(x,\lambda,\mu) := \inf_{\pi\in\Pi} U_n^\pi(x,\lambda,\mu).$$

Let $\mathcal{U}$ be the class of mappings $U : E \times R \to [0,1]$. For any $(x,\lambda,\mu) \in E \times R, U \in \mathcal{U}$, $\varphi \in \phi$, and $a \in A(x)$, the operators $T^\varphi U$ and TU are defined as follows:

$$\begin{aligned} T^aU(x,\lambda,\mu) &:= \int_{E_X}\int_{E_Y} U\Big(\hat{x}, \lambda - \hat{r}(x,\mu,a), \Phi(x,\lambda,a,\hat{x},\lambda-\hat{r}(x,\mu,a),\mu)\Big) \\ &\quad \times Q^X(\mathrm{d}\hat{x}|x,\mu,a), \\ T^\varphi U(x,\lambda,\mu) &:= \int_A T^aU(x,\lambda,\mu)\varphi(\mathrm{d}a|x,\lambda,\mu), \qquad (3.15) \\ TU(x,\lambda,\mu) &:= \min_{a\in A(x)} T^aU(x,\lambda,\mu). \qquad (3.16) \end{aligned}$$

To strictly show the unobservable state's conditional distribution, some characteristics of the filter equation are given.

Lemma 3.2. Under Assumption 3.1, for each $\pi \in \Pi$,$x \in E_X, \lambda \in R$. Then, $F_N^\pi(x,\lambda) = U_N^\pi(x,\lambda,Q_0)$, $F_N^*(x,\lambda) = U_N^*(x,\lambda,Q_0)$.

Proof. For each $\pi \in \Pi$ and $x \in E_X$, the following result is first proven by induction:

$$F_n^\pi(x,\Lambda) = U_n^\pi(x,\hat{\Lambda},\mu), \tag{3.17}$$

with $n = -1, 0, 1, 2, \ldots$, for the reward goal function $\Lambda : K \times E_Y \to R^+$ and $\hat{\Lambda} = \int \Lambda(x,y,a)\mu(\mathrm{d}y)$.

Based on $F_{-1}^\pi = U_{-1}^\pi = I_{[0,+\infty)}(\lambda)$, for any $\pi \in \Pi$ and $x \in E_X, \lambda \in R$, Eq (3.17) is valid when $n = -1$. Suppose that Fact (3.17) holds for $k = n$; by (2.6),

$$\begin{aligned} & U_{n+1}^\pi(x,\lambda,\mu_0) \\ = \;& \hat{P}^\pi_{(x,\lambda,\mu_0)}\Big(\sum_{k=0}^{n+1} \hat{r}(X_k,\mu_k,A_k) \le \lambda\Big) \\ = \;& \hat{E}^\pi_{(x,\lambda,\mu_0)}[\hat{E}^\pi_{(x,\lambda,\mu_0)}[I_{\{\sum_{k=0}^{n+1}\hat{r}(X_k,\mu_k,A_k)\le\lambda\}}|X_0,\Lambda_0,\mu_0,A_0,X_1,\Lambda_1,\mu_1]] \\ = \;& \int_{E_X}\int_R\int_{\mathcal{P}(E_Y)}\int_A \hat{E}^\pi_{(x,\lambda,\mu)}[I_{\{\sum_{k=0}^{n+1}\hat{r}(X_k,\mu_k,A_k)\le\lambda\}}|X_0=x,\Lambda_0=\lambda,\mu_0=Q_0,A_0=a_0,X_1=x_1, \\ & \times\lambda_1 = \lambda - \hat{r}(x,Q_0,a_0),\mu_1 = \Phi(x,\lambda,a_0,x_1,\lambda_1,Q_0)] \\ & \times\hat{Q}(\mathrm{d}x_1,\mathrm{d}\lambda_1,\mathrm{d}\mu_1|x,\lambda,Q_0,a_0)\pi_0(\mathrm{d}a_0|x,\lambda) \end{aligned}$$

$$
\begin{aligned}
&= \int_{E_X}\int_A\int_{E_Y} \hat{P}^{{}^1\pi}_{(x_1,\lambda-\hat{r}(x,Q_0,a_0),\Phi(x,\lambda,a_0,x_1,\lambda_1,Q_0))}\Big(\sum_{k=0}^{n}\hat{r}(X_k,\mu_k,A_k)\le\lambda_1\Big)\\
&\quad\times Q^X(\mathrm{d}x_1|x,Q_0,a_0)\pi_0(\mathrm{d}a_0|x,\lambda)\\
&= \int_{E_X}\int_A\int_{E_Y} U_n^{{}^1\pi}\Big(x_1,\lambda-\hat{r}(x,Q_0,a_0),\Phi(x,\lambda,a_0,x_1,\lambda_1,Q_0)\Big)\\
&\quad\times Q^X(\mathrm{d}x_1|x,Q_0,a_0)\pi_0(\mathrm{d}a_0|x,\lambda) \qquad (3.18)
\end{aligned}
$$

is obtained, where ${}^1\pi := \{\pi_1,\pi_2,...\}$ represents the 1-shift policy of π.

On the other side, by Eq (3.4), we have the following:

$$
\begin{aligned}
&F^\pi_{n+1}(x,\lambda)\\
&= \int_{E_Y} P^\pi_{(x,y,\lambda)}\Big(\sum_{k=0}^{n+1} r(X_k,Y_k,A_k)\le\lambda\Big)Q_0(\mathrm{d}y)\\
&= \int_{E_Y} E^\pi_{(x,y,\lambda)}[E^\pi_{(x,y,\lambda)}[I_{\{\sum_{k=0}^{n+1} r(X_k,Y_k,A_k)\le\lambda\}}|X_0,\Lambda_0,Y_0,A_0,X_1,\Lambda_1,Y_1]]Q_0(\mathrm{d}y)\\
&= \int_{E_Y}\int_{E_X}\int_R\int_A\int_{E_Y} E^\pi_{(x,y,\lambda)}[E^\pi_{(x,y,\lambda)}[I_{\{\sum_{k=0}^{n+1} r(X_k,Y_k,A_k)\le\lambda\}}|X_0=x,\Lambda_0=\lambda,Y_0=y,A_0=a_0,X_1=x_1,\\
&\quad\Lambda_1=\lambda_1,Y_1=y_1]]\Phi(x,\lambda,a_0,x_1,\lambda_1,Q_0)(\mathrm{d}y_1)Q^X(\mathrm{d}x_1|x,y,a_0)I_{\{\lambda-r(x,y,a)\}}(\mathrm{d}\lambda_1)\pi_0(\mathrm{d}a_0|x,\lambda)Q_0(\mathrm{d}y)\\
&= \int_{E_Y}\int_{E_X}\int_R\int_A\int_{E_Y} P^{{}^1\pi}_{(x_1,y_1,\lambda-r(x,y,a_0))}\Big(\sum_{k=0}^{n} r(X_k,Y_k,A_k)\le\lambda-r(x,y,a_0)\Big)\Phi(x,\lambda,a_0,x_1,\lambda_1,Q_0)(\mathrm{d}y_1)\\
&\quad\times Q^X(\mathrm{d}x_1|x,y,a_0)I_{\{\lambda-r(x,y,a_0)\}}(\mathrm{d}\lambda_1)\pi_0(\mathrm{d}a_0|x,\lambda)Q_0(\mathrm{d}y)\\
&= \int_{E_X}\int_A\int_{E_Y} F_n^{{}^1\pi}(x_1,\lambda-r(x,y,a_0))Q^X(\mathrm{d}x_1|x,y,a_0)\pi_0(\mathrm{d}a_0|x,\lambda)Q_0(\mathrm{d}y),
\end{aligned}
$$

which, together with Eq (3.18) and the inductive hypothesis, can prove Eq (3.17) for $n = 0,1,\ldots,N$, i.e, $F^\pi_n(x,\lambda) = U^\pi_n(x,\lambda,\mu)$.

For $n = N$, $F^\pi_N(x,\lambda) = U^\pi_N(x,\lambda,\mu)$, which yields $F^*_N(x,\lambda) = U^*_N(x,\lambda,Q_0)$ for the arbitrary policy π.

The establishment of the optimality equation requires the following theorem.

Theorem 3.1. Suppose that Assumption 3.1 is satisfied. Then, for any $(x,\lambda,\mu) \in E\times R, \pi = \{\pi_0,\pi_1,\ldots\} \in \Pi, n \ge 0$, the following statement holds: $U^\pi_{n+1}(x,\lambda,\mu) = T^{\pi_0}U_n^{{}^1\pi}(x,\lambda,\mu)$, where $U^\pi_0(x,\lambda,\mu) = I_{[0,+\infty)}(\lambda)$,${}^1\pi := \{\pi_1,\pi_2,...\}$ represents the 1-shift policy of π.

Proof. For any $(x,\lambda,\mu)\in E\times R, \pi=\{\pi_0,\pi_1,\ldots\}\in\Pi, n = 0,1,\ldots,N-1$, by (3.8) and the properties of conditional expectation, we can obtain the following:

$$
\begin{aligned}
&U^\pi_{n+1}(x,\lambda,\mu)\\
&= \hat{P}^\pi_{(x,\lambda,\mu)}\Big(\sum_{k=0}^{n+1}\hat{r}(X_k,\mu_k,A_k)\le\lambda\Big)\\
&= \hat{E}^\pi_{(x,\lambda,\mu)}[\hat{E}^\pi_{(x,\lambda,\mu)}[I_{\{\sum_{k=0}^{n+1}\hat{r}(X_k,\mu_k,A_k)\le\lambda\}}|X_0,\Lambda_0,\mu_0,A_0,X_1,\Lambda_1,\mu_1]]
\end{aligned}
$$

$$
\begin{aligned}
&= \int_A \int_{E_X} \int_R \int_{\mathcal{P}(E_Y)} \hat{P}^{\pi}_{(x,\lambda,\mu)}\Big(\sum_{k=0}^{n+1} \hat{r}(X_k, \mu_k, A_k) \le \lambda | X_0 = x, \Lambda_0 = \lambda, \mu_0 = Q_0, A_0 = a, X_1 = \hat{x}, \\
&\quad \Lambda_1 = \hat{\lambda}, \mu_1 = \hat{\mu}\Big) \hat{Q}(\mathrm{d}\hat{x}, \mathrm{d}\hat{\lambda}, \mathrm{d}\hat{\mu} | x, \lambda, \mu, a) \pi_0(da | x, \lambda) \\
&= \int_{E_X} \int_R \int_A \int_{E_Y} \hat{P}^{\pi}_{(x,\lambda,\mu)}\Big(\sum_{k=0}^{n+1} \hat{r}(X_k, \mu_k, A_k) \le \lambda | X_0 = x, \Lambda_0 = \lambda, \mu_0 = Q_0, A_0 = a, X_1 = \hat{x}, \\
&\quad \Lambda_1 = \hat{\lambda}, \mu_1 = \hat{\mu}\Big) Q^X(\mathrm{d}\hat{x} | x, \mu, a) I_{\{\lambda - \hat{r}(x,\mu,a)\}}(\mathrm{d}\hat{\lambda}) \pi_0(\mathrm{d}a | x, \lambda) \\
&= \int_{E_X} \int_A \int_{E_Y} \hat{P}^{1\pi}_{(\hat{x}, \lambda - \hat{r}(x,\mu,a), \Phi(x,\lambda,a,\hat{x},\lambda - \hat{r}(x,\mu,a),\mu))}\Big(\sum_{k=0}^{n} \hat{r}(X_k, \mu_k, A_k) \le \lambda - \hat{r}(x,\mu,a)\Big) \\
&\quad \times Q^X(\mathrm{d}\hat{x} | x, \mu, a) \pi_0(\mathrm{d}a | x, \lambda) \\
&= \int_{E_X} \int_A \int_{E_Y} U_n^{1\pi}\Big(\hat{x}, \lambda - \hat{r}(x,\mu,a), \Phi(x,\lambda,a,\hat{x},\lambda - \hat{r}(x,\mu,a),\mu)\Big) Q^X(\mathrm{d}\hat{x} | x, \mu, a) \pi_0(\mathrm{d}a | x, \lambda) \\
&:= T^{\pi_0} U_n^{1\pi}(x, \lambda, \mu).
\end{aligned}
$$

The proof of this conclusion has been completed.

Theorem 3.2. Suppose that Assumption 3.1 holds. For each $(x, \lambda, \mu) \in E \times R$, then:

(a) $\{U_n^*, n = 0, 1, \ldots, N-1\}$ satisfies the corresponding optimality equation:

$$U_0^*(x, \lambda, \mu) := I_{[0,\infty)}(\lambda), \ U_{n+1}^*(x, \lambda, \mu) := T U_n^*(x, \lambda, \mu).$$

(b) There exists a policy $g_n \in \Pi_{DM}$ such that $U_{n+1}^* = T^{g_n} U_n^*$ for $n = 0, 1, \ldots, N-1$. Then, the policy $\pi^* := \{f_0^*, f_1^*, \ldots, f_{N-1}\} \in \Pi_{DH}$ is optimal, where $f_n^*(h_n) := g_{N-1-n}(x_n, \lambda_n, \mu_n)$ for each $h_n \in H_n$, $n = 0, 1, \ldots, N-1$.

Proof. (a) According to Theorem 3.1 and (3.16), for each $(x, \lambda, \mu) \in E \times R, \pi = \{\pi_0, \pi_1, \ldots\} \in \Pi$, we have the following:

$$U_{n+1}^{\pi}(x, \lambda, \mu) = T^{\pi_0} U_n^{1\pi}(x, \lambda, \mu) \ge T^{\pi_0} U_n^*(x, \lambda, \mu) \ge T U_n^*(x, \lambda, \mu). \tag{3.19}$$

Since π is arbitrary, this implies $U_{n+1}^*(x, \lambda, \mu) \ge T U_n^*(x, \lambda, \mu)$.

To prove the reverse condition, the following fact is needed to be proven: for any $(x, \lambda, \mu) \in E \times R$ and $n \ge -1$, there is a policy $\eta \in \Pi_{DM}$ which satisfies $U_n^*(x, \lambda, \mu) = U_n^{\eta}(x, \lambda, \mu)$. Since $U_{-1}^*(x, \lambda, \mu) = U_{-1}^{\pi}(x, \lambda, \mu) = I_{[0,+\infty)}(\lambda)$ for any $\pi \in \Pi_M$, this fact trivially holds for $n = -1$. Assume that there exists a policy $\zeta \in \Pi_{DM}$ that satisfies $U_k^*(x, \lambda, \mu) = U_k^{\zeta}(x, \lambda, \mu)$ for $n = k \ge -1$. On the other hand, since the set of actions is finite, there exists a policy $f \in \Pi_s$ that satisfies $T^f U_k^*(x, \lambda, \mu) = T U_k^*(x, \lambda, \mu)$. Then, let $\theta = \{f, \zeta\} \in \Pi_{DM}$, we know that

$$U_{k+1}^*(x, \lambda, \mu) \le U_{k+1}^{\theta}(x, \lambda, \mu) = T^f U_k^{\zeta}(x, \lambda, \mu) = T^f U_k^*(x, \lambda, \mu) = T U_k^*(x, \lambda, \mu), \tag{3.20}$$

where the first equality is obtained by Lemma 3.1, and the second equality follows from the induction hypothesis. Combining (3.19) and (3.20), we have $T U_k^* = U_{k+1}^*$, which implies the induction hypothesis is satisfied and the result is proven.

(b) For each $(x, \lambda, \mu) \in E \times R$, the existence of a policy $g_n^* \in \Pi_{DM}$ satisfying $U_{n+1}^* = T^{g_n^*} U_n^*$, is determined by the finiteness of the action set for $n = 0, 1, \ldots, N-1$. Letting $\pi = \pi(n) := \{g_n, g_{n-1}, \ldots, g_0\} \in \Pi$, when $n = 0$, by Lemma 3.1 (b), $U_1^\pi = T^{g_0} U_0^{1\pi} = T^{g_0} U_0^* = T U_0^* = U_1^*$. Assuming that $U_k^* = U_k^{1\pi^*}$ for $n = k$, by Lemma 3.1 (b) and part (a),

$$U_{k+1}^\pi = T^{g_k^*} U_k^{1\pi^*} = T^{g_k} U_k^* = T U_k^* = U_{k+1}^*.$$

Thus, the induction hypothesis is established and $U_N^{\pi^*} = U_N^*$ for $\pi^* := \{f_0^*, f_1^*, \ldots, f_{N-1}\}$ with $f_n^*(h_n) := g_{N-1-n}(x_n, \lambda_n, \mu_n(\cdot|h_n)), n = 0, 1, \ldots, N-1$. Then, the policy π^* is optimal.

Based on Theorem 3.2, the value iteration algorithm is established as follows:
The value iteration algorithm

Step 1. Let $U_0^*(x, \lambda, \mu) := I_{[0,+\infty)}(\lambda)$, for $(x, \lambda, \mu) \in E \times R$.

Step 2. The computation of the value function U_n^* is as follows for $n = 0, 1, \ldots, N-1$:

$$\begin{aligned} T^a U_n^*(x, \lambda, \mu) &= \int_{E_X} \int_{E_Y} U_n^*\Big(\hat{x}, \lambda - \hat{r}(x, \mu, a), \Phi(x, \lambda, a, \hat{x}, \lambda - \hat{r}(x, \mu, a), \mu)\Big) \\ &\quad \times Q^X(\mathrm{d}\hat{x}|x, \mu, a). \\ U_{n+1}^*(x, \lambda, \mu) &= \min_{a \in A(x)} \{T^a U_n^*(x, \lambda, \mu)\}. \end{aligned}$$

Step 3. Find a policy g_{N-1} that satisfies $U_N^* = T^{g_{N-1}} U_{N-1}^*$. Then, by Theorem 3.2, the policy π^* is optimal.

4. Illustration

An illustration is provided to show how both the VF and OP are calculated and illustrates the effectiveness and feasibility of the value iteration algorithm.

Example 4.1. Consider a machine production process with two types of observable product quality states (i.e., nonconforming product 0 and qualified products 1), and two types of unobservable machine operation states (i.e. poor state 1 and good state 2). According to the product quality situation $x = 1$ and the reward goal λ, at the initial time $n = 0$, when the production process is in the state $y \in \{1, 2\}$, the decision-maker can either select an ordinary maintenance action a_{11} or an advanced maintenance action a_{12} with a reward $r(x, y, a)$. If the product quality situation is $x = 0$, the decision maker must select an action of the advanced maintenance a_{01}. When the action of the maintenance a is applied, the system transits to the state (x', y') with probability $Q(\cdot, \cdot|x, y, a)$ at the next moment. The general objective of the decision maker is to select the optimal action to ensure that the minimum probability value of the total rewards does not exceed the target λ from 0 to $N = 15$.

This evolution process can be formulated as a discrete-time POMDP with the state space $E_X \times E_Y = \{0, 1\} \times \{1, 2\}$; the admissible class of actions $A(0) = \{a_{01}\}, A(1) = \{a_{11}, a_{12}\}$. Assume that the probabilities of the transition are given by $Q(\cdot, \cdot|x, y, a) = Q^X(\cdot|x, y, a)p(\cdot|y)$, in which the probabilities of the transition $Q^X(\cdot|x, y, a)$ are given by the following:

$$Q^X(0|0, 1, a_{01}) = 1, \quad Q^X(1|0, 1, a_{01}) = 0, \quad Q^X(0|0, 2, a_{01}) = 1, \quad Q^X(1|0, 2, a_{01}) = 0,$$

$$Q^X(0|1,1,a_{11}) = 0.5, \quad Q^X(1|1,1,a_{11}) = 0.5, \quad Q^X(0|1,1,a_{12}) = 0.3, \quad Q^X(0|1,1,a_{12}) = 0.7, \quad (4.1)$$
$$Q^X(0|1,2,a_{11}) = 0.4, \quad Q^X(1|1,2,a_{11}) = 0.6, \quad Q^X(0|1,2,a_{12}) = 0.2, \quad Q^X(0|1,2,a_{12}) = 0.8,$$

The transition probabilities of the unobserved state are given by $p(2 \mid 2) = 1 - p(1 \mid 2) = 0.7$, $p(1 \mid 1) = 1$. The reward rates are given as follows:

$$r(0,1,a_{01}) = r(0,2,a_{01}) = 0, \ r(1,1,a_{11}) = 2,$$
$$r(1,1,a_{12}) = 4, \ r(1,2,a_{11}) = 1, \ r(1,2,a_{22}) = 3.$$

Our main goal is to use the value iteration algorithm to compute the value function and the optimal policies.

First, according to (3.13), since $r(0,1,a_{01}) = r(0,2,a_{01}) = 0$, it is known that $U^*(0,\lambda,\mu) = I_{[0,+\infty)}(\lambda)$. Based on the value iteration algorithm (Algorithm 1) and Matlab software, the curves of functions $T^{a_{11}}U^*(1,\lambda,\mu)$, $T^{a_{12}}U^*(1,\lambda,\mu)$ and the approximated value function $U^*(1,\lambda,\mu)$ are plotted (see Figures 1 and 2). By observing the figures, the following conclusions are attained:

(*a*) As seen in Figure 1, when $x = 1$, if $\lambda \in (0,4)$, the value $T^{a_{12}}U_N^*(1,\lambda,\mu)$ is less than $T^{a_{11}}U_N^*(1,\lambda,\mu)$. Otherwise, if $\lambda \in [4,+\infty)$, the value $T^{a_{11}}U_N^*(1,\lambda,\mu)$ is less than $T^{a_{12}}U_N^*(1,\lambda,\mu)$. As shown above, the observable state is $x = 1$, $\lambda \in (0,4)$, the decision maker should choose the low risk action a_{12}. Conversely, if $\lambda \in [4,+\infty)$, the decision maker should choose the low risk action a_{11} instead of the action a_{12}.

(*b*) Based on Figure 1, the risk probability optimal policy for POMDPs at time $n = 0,1,\ldots,N$ is given by the following:

$$f^*(1,\lambda) = \begin{cases} a_{12}, & 0 \le \lambda < 4; \\ a_{11}, & \lambda \ge 4. \end{cases} \quad (4.2)$$

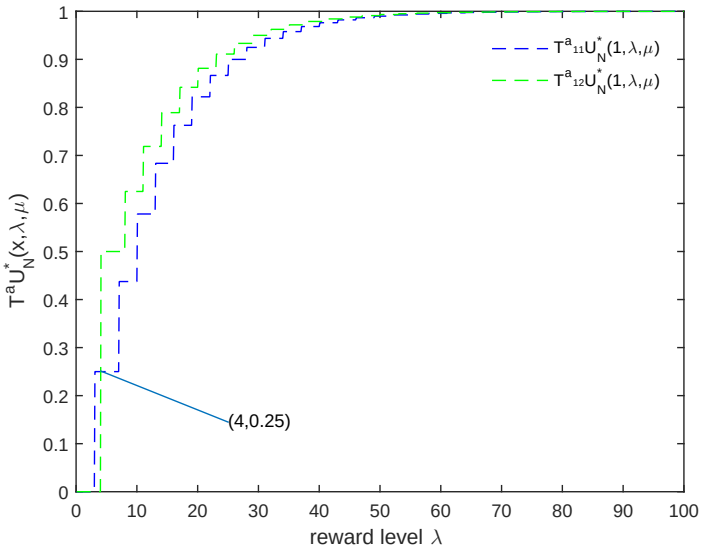

Figure 1. The function $T^aU_N^*(1,\lambda,\mu)$.

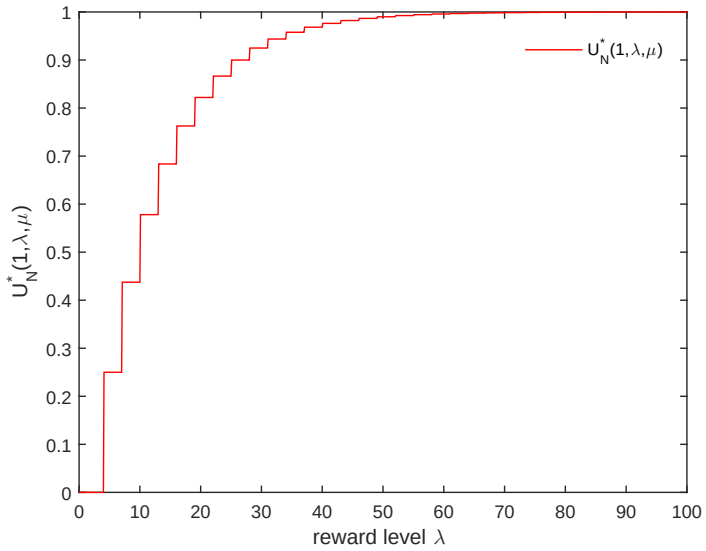

Figure 2. The value function $U_N^*(1, \lambda, \mu)$.

5. Conclusions

In this paper, we studied the problem of minimizing the risk probability criterion for finite horizon partially observable discrete-time Markov decision processes (POMDPs). Different from the classical expectation criterion, which are regarded as a component of an extended state according to the reward levels, we redefined a history-dependent policy, and reconstructed a new probability measure. Based on the Bayes operator and the filter equations we constructed, the optimization problem of risk probability can be equivalently reformulated as filtered Markov decision processes. We proposed a value iteration algorithm to establish the existence of a solution to the optimality equation, and a risk probability optimal policy.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by Guangxi science and technology base and talent project (Grant No. AD21159005); Foundation of Guangxi Educational Committee (Grant No. KY2022KY0342); National Natural Science Foundation of China (Grant No. 11961005,12361091); Guangxi Natural Science Foundation Program (Grant No. 2020GXNSFAA297196); The Doctoral Foundation of Guangxi University of Science and Technology (Grant No. 18Z06).

Conflict of interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. N. Bauerle, U. Rieder, *Markov decision processes with applications to finance*, Heidelberg: Springer, 2011. https://doi.org/10.1007/978-3-642-18324-9
2. J. Janssen, R. Manca, *Semi-Markov risk models for finance, insurance and reliability*, New York: Springer, 2006. https://doi.org/10.1007/0-387-70730-1
3. X. P. Guo, O. Hernández-Lerma, *Continuous-time Markov decision processes: Theorey and applications*, Berlin: Springer-Verlag, 2009. https://doi.org/10.1007/978-3-642-02547-1
4. M. J. Sobel, The variance of discounted Markov decision processes, *J. Appl. Probab.*, **19** (1982), 794–802. https://doi.org/10.1017/s0021900200023123
5. Y. Ohtsubo, K. Toyonaga, Optimal policy for minimizing risk models in Markov decision processes, *J. Math. Anal. Appl.*, **271** (2002), 66–81. https://doi.org/10.1016/s0022-247x(02)00097-5
6. D. J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.*, **173** (1993), 634–646. https://doi.org/10.1006/jmaa.1993.1093
7. C. B. Wu, Y. L. Lin, Minimizing risk models in Markov decision processes with policies depending on target values, *J. Math. Anal. Appl.*, **231** (1999), 47–67. https://doi.org/10.1006/jmaa.1998.6203
8. X. Wu, X. P. Guo, First passage optimality and variance minimization of Markov decision processes with varying discount factors, *J. Appl. Probab.*, **52** (2015), 441–456. https://doi.org/10.1017/S0021900200012560
9. Y. H. Huang, X. P. Guo, Optimal risk probability for first passage models in Semi-Markov processes, *J. Math. Anal. Appl.*, **359** (2009), 404–420. https://doi.org/10.1016/j.jmaa.2009.05.058
10. Y. H. Huang, X. P. Guo, Z. F. Li, Minimum risk probability for finite horizon semi-Markov decision processes, *J. Math. Anal. Appl.*, **402** (2013), 378–391. https://doi.org/10.1016/j.jmaa.2013.01.021
11. X. X. Huang, X. L. Zou, X. P. Guo, A minimization problem of the risk probability in first passage semi-Markov decision processes with loss rates, *Sci. China Math.*, **58** (2015), 1923–1938. https://doi.org/10.1007/s11425-015-5029-x
12. H. F. Huo, X. L. Zou, X. P. Guo, The risk probability criterion for discounted continuous-time Markov decision processes, *Discrete Event Dyn. syst.*, **27** (2017), 675–699. https://doi.org/10.1007/s10626-017-0257-6
13. H. F. Huo, X. Wen, First passage risk probability optimality for continuous time Markov decision processes, *Kybernetika*, **55** (2019), 114–133. https://doi.org/10.14736/kyb-2019-1-0114
14. H. F. Huo, X. P. Guo, Risk probability minimization problems for continuous time Markov decision processes on finite horizon, *IEEE T. Automat. Contr.*, **65** (2020), 3199–3206. https://doi.org/10.1109/tac.2019.2947654

15. X. Wen, H. F. Huo, X. P. Guo, First passage risk probability minimization for piecewise deterministic Markov decision processes, *Acta Math. Appl. Sin. Engl. Ser.*, **38** (2022), 549–567. https://doi.org/10.1007/s10255-022-1098-0

16. A. Drake, *Observation of a Markov process through a noisy channel*, Massachusetts Institute of Technology, 1962.

17. K. Hinderer, *Foundations of non-stationary dynamic programming with discrete time parameter*, Berlin: Springer-Verlag, 1970.

18. D. Rhenius, Incomplete information in Markovian decision models, *Ann. Statist.*, **26** (1974), 1327–1334. https://doi.org/10.1214/aos/1176342886

19. O. Hernández-Lerma, *Adaptive Markov control processes*, New York: Springer-Verlag, 1989. https://doi.org/10.1007/978-1-4419-8714-3

20. R. D. Smallwood, E. J. Sondik, The optimal control of partially observable Markov processes over a finite horizon, *Oper. Res.*, **21** (1973), 1071–1088. https://doi.org/10.1287/opre.21.5.1071

21. K. Sawaki, A. Ichikawa, Optimal control for partially observable Markov decision processes over an infinite horizon, *J. Oper. Res. Soc. JPN*, **21** (1978), 1–16. https://doi.org/10.15807/jorsj.21.1

22. C. C.White, W. T. Scherer, Finite memory suboptimal design for partially observed Markov decision processes, *Oper. Res.*, **42** (1994), 439–455. https://doi.org/10.1287/opre.42.3.439

23. E. A. Feinberg, P. O. Kasyanov, M. Z. Zgurovsky, Partially observable total cost Markov decision processes with weakly continuous transition probabilities, *Math. Oper. Res.*, **41** (2016), 656–681. https://doi.org/10.1287/moor.2015.0746

24. M. Haklidir, H. Temeltas, Guided soft actor critic: A guided deep reinforcement learning approach for partially observable Markov decision processes, *IEEE Access*, **9** (2021), 159672–159683. https://doi.org/10.1109/access.2021.3131772

25. D. Bertsekas, S. Shreve, *Stochastic optimal control: The discrete-time case*, Athena Scientific, 1996.

AIMS Press