



Minimum risk probability for finite horizon semi-Markov decision processes[☆]



Yonghui Huang^a, Xianping Guo^a, Zhongfei Li^{b,*}

^a School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China

^b Business School, Sun Yat-Sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:

Received 3 August 2012

Available online 23 January 2013

Submitted by Vladimir Pozdnyakov

Keywords:

Finite horizon semi-Markov decision processes

Risk probability

Optimal value function

Iteration algorithm

Optimal policy

ABSTRACT

This paper studies the risk probability criteria for finite horizon semi-Markov decision processes. The goal is to find an optimal policy with the minimum risk probability that the total reward produced by a system during a finite horizon does not exceed a reward level, where the optimality is over the class of all randomized historic policies which include states, planning horizons and also reward levels. Under mild conditions, the optimality equation and the existence of optimal policies are established, and in addition, an iteration algorithm for solving optimal policies is developed. Our main results are applied to a manufacturing system.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Finite horizon Markov decision processes (MDPs) form a class of basic and important dynamic programming problems and have many applications in the real world [1,3,8,9,17,20,23,24]. As is well-known, finite horizon optimality for *discrete-time* MDPs (DTMDPs) has been extensively investigated by researchers; see, for instance, [8,20,24] for finite horizon DTMDPs with expected total reward/cost criteria; and [3,4,25,26] for finite horizon DTMDPs with risk probability criteria. In contrast to finite horizon optimality for DTMDPs, however, finite horizon optimality for *continuous-time* models is complex and is difficult to deal with. In the literature, little work has been directly devoted to finite horizon decision models in continuous-time. Indeed, for the finite horizon problems, time is crucial in planning and it often introduces a special structure with additional complexity, especially in the case of continuous-time models. The conventional approach to analyzing finite horizon optimality for continuous-time models is *time-discretization* [22], that is, subdividing the time interval into discrete elements and solving a sequence of DTMDPs concerned. Such an approach, however, seems complicated and difficult to implement in practice; for details, see Buchholz and Schulz [6]. Different from the time-discretization approach, Mamer [17] develops successive approximations for finite horizon semi-Markov decision processes (SMDPs) by introducing a kind of policies depending on the remaining time. This approach affords considerable analytical simplicity over the time-discretization approach and can be used to establish policy structure. Recently, Huang and Guo [9] propose a viewpoint that the finite horizon optimality is *horizon-relevant*, and employ an invariant imbedding technique to introduce a class of so-called horizon-relevant policies which include the usual states and also the planning horizons. This kind of horizon-relevant policies generalize the usual policies for infinite horizon SMDPs with expected discount or average criteria [10,15,20], and in

[☆] This work was supported by NSFC (70825002, 10925107, 11101444), GDUPS, and the High-level Talent Project of Guangdong Province.

* Corresponding author.

E-mail addresses: hyongh5@mail.sysu.edu.cn (Y. Huang), mcsqxp@mail.sysu.edu.cn (X. Guo), lnslzf@mail.sysu.edu.cn (Z. Li).

the same time, extend the policies in Mamer [17] to the randomized historic ones. They consider finite horizon SMDPs with expected cost criteria in the framework of horizon-relevant policies, prove the existence of optimal policies, and establish an effective algorithm for computing optimal policies.

In this paper, we further study finite horizon SMDPs, and we focus on *the risk probability criteria* rather than expected reward/cost criteria in [9,17]. As is known, the risk probability criteria can be used to describe the performance of many systems in real cases, and thus have received increasing attention in the field of control and engineering, especially in reliability and risk analysis [3,4,12,25–27]. Boda et al. [3], Bouakiz and Kebir [4], White [25], and Wu and Lin [26] consider the probability criteria for finite or infinite horizon DTMDPs. Yu, Lin and Yan [27] deal with the probability criteria for random horizon DTMDPs, where the random horizon is a first passage time to some target states of a system. Moreover, Huang, Guo and Song [12] investigate the risk probability criteria in SMDPs with a random horizon. The overview of the literature on the risk probability criteria above shows that the works for DTMDPs are complete in the sense that the cases of finite, infinite and random horizons have been all considered, while the works for SMDPs are concentrated on infinite or random horizons, and the case of finite horizon has not been explored yet. On the other hand, we note that, compared to DTMDPs, SMDPs are a sort of more general stochastic control models in which action choice is allowed at random times whenever the system state changes, and thus can characterize many situations such as queuing control and equipment maintenance [7,14,16,17,21]. Therefore, the study on the risk probability criteria for finite horizon SMDPs is desirable and necessary.

We aim at obtaining an optimal policy which minimizes the risk probability that the total reward produced by a system during a finite horizon does not exceed a reward level in all randomized historic policies. As mentioned above, finite horizon problems in continuous-time are difficult to deal with. In this paper, however, we follow the idea and technique in Huang and Guo [9] for finite horizon SMDPs with expected cost criteria. That is, based on the viewpoint that the finite horizon optimality is horizon-relevant, we introduce a family of horizon-relevant policies specifically for the finite horizon SMDPs with risk probability criteria, in which the planning horizons at decision epochs are taken as part of state information. In this way, we generalize the usual policies for (infinite or random horizon) SMDPs with risk probability criteria [12] and expected reward/cost criteria [10,15,20]; see Remarks 2.2–2.4 for details. In the framework of the horizon-relevant policies, we establish the optimality equation for the finite horizon SMDPs with risk probability criteria, and further obtain the existence of optimal policies with the minimum risk probability. Moreover, we derive an iteration algorithm for solving optimal policies. To demonstrate the applications of our results, we present an example about a manufacturing system, in which we obtain an explicit optimal policy by executing numerical experiments.

Finally, it should be mentioned that since SMDPs are a generalization of DTMDPs, the results developed in this paper can be reduced to the case for DTMDPs, and coincide with those for finite horizon DTMDPs with probability criteria [3,4,25,26]; see Remark 3.4 for details.

The rest of this paper is organized as follows. Section 2 formulates the control model and the optimality problem we are interested in. Our main results on the existence and the computation of optimal policies are stated in Section 3, and illustrated with a manufacturing system in Section 4. Their proofs are postponed to Section 5. We conclude in Section 6 with some comments.

2. The control model

The model of SMDPs we are concerned with is specified by the five objects:

$$\{E, (A(i) \subset A, i \in E), Q(\cdot, \cdot | i, a), r(i, a)\}, \quad (2.1)$$

where E is the *state space*, A is the *action set*, which are both assumed to be denumerable; $A(i)$ denotes the *set of admissible actions* at state $i \in E$, which is assumed to be finite. The transition mechanism of the SMDPs is defined by the *semi-Markov kernel* $Q(\cdot, \cdot | i, a)$ on $\mathbb{R}_+ \times E$ given K , where $\mathbb{R}_+ = [0, +\infty)$, and $K = \{(i, a) \mid i \in E, a \in A(i)\}$ is the set of feasible state-action pairs. It is assumed that: (i) $Q(\cdot, j | i, a)$ (for any fixed $j \in E$ and $(i, a) \in K$) is a nondecreasing, right-continuous real function on \mathbb{R}_+ such that $Q(0, j | i, a) = 0$; (ii) $Q(t, \cdot | i, a)$ (for each fixed $t \in \mathbb{R}_+$) is a sub-stochastic kernel on E given K ; and (iii) $P(\cdot | \cdot, \cdot) := Q(\infty, \cdot | \cdot, \cdot)$ is a stochastic kernel on E given K . If an action $a \in A(i)$ is selected in state i , then $Q(t, j | i, a)$ is the joint probability that the sojourn time in state i is not greater than $t \in \mathbb{R}_+$, and the next state is j (it is possible that $j = i$ with positive probability). Finally, the real function $r : K \rightarrow \mathbb{R}_+$ denotes the *reward rate*.

Remark 2.1. Note that the transition mechanism of SMDPs is different from those of DTMDPs [3,4,25–27]. In fact, SMDPs are a generalization of DTMDPs.

We next describe the evolution of the *finite horizon SMDPs with risk probability criteria*. Suppose that the system occupies state i_0 at the initial decision epoch $s_0 = 0$, and the decision maker has a profit goal (reward level) λ_0 over a planning horizon t_0 (that is, he/she should try to meet the reward level λ_0 within the planning horizon t_0), then he/she chooses an action a_0 according to the current state i_0 , the planning horizon t_0 and the reward level λ_0 . As a consequence of this action choice, the system remains in i_0 until time s_1 , at which point the system state changes to i_1 and the next decision epoch occurs. At time s_1 , a reward $r(i_0, a_0)(s_1 - s_0)$ is earned, and thus there is a remaining profit goal $\lambda_1 := \lambda_0 - r(i_0, a_0)(s_1 - s_0)$ over a remaining planning horizon $t_1 := [t_0 - (s_1 - s_0)]^+$ for the decision maker, where $[x]^+ := \max\{x, 0\}$. According to the current state i_1 , the current planning horizon t_1 and the current profit goal λ_1 as well as the previous ones, the decision

maker chooses an action a_1 and the same sequence of events occur. The decision process evolves in this way, and hence we obtain an admissible horizon-relevant (h-r in short) history h_n of the SMDPs up to the n th decision epoch, i.e.,

$$h_n = (s_0, i_0, t_0, \lambda_0, a_0, \dots, s_{n-1}, i_{n-1}, t_{n-1}, \lambda_{n-1}, a_{n-1}, s_n, i_n, t_n, \lambda_n),$$

where $s_0 = 0$, $s_{m+1} \geq s_m$, $(i_m, a_m) \in K$, $t_0 \in \mathbb{R}_+$, $t_{m+1} := [t_m - (s_{m+1} - s_m)]^+$, $\lambda_0 \in \mathbb{R} := (-\infty, +\infty)$, $\lambda_{m+1} := \lambda_m - r(i_m, a_m)(s_{m+1} - s_m)$ for $m = 0, 1, \dots, n-1$, and $i_n \in E$. Let H_n denote the set of all admissible h-r histories h_n of the system up to the n th decision epoch, where H_n is endowed with a Borel σ -algebra.

- Remark 2.2.** (a) The h-r history h_n here generalizes the usual ones for SMDPs by the following two ways: on the one hand, the information of the planning horizons at decision epochs is introduced since in the finite horizon problems (rather than infinite horizon cases) the remaining planning time usually affects one's behavior heavily; on the other hand, the information about the reward levels at decision epochs is introduced since we treat the probability criteria (rather than the standard expected discount or average criteria). In fact, when the planning horizons $t_n \equiv \infty$ at every decision epoch (there are always infinite planning horizons at decision epochs in infinite horizon problems), the h-r history here will be reduced to the one for infinite or random horizon SMDPs with risk probability criteria [12]. If, furthermore, the history is independent of the reward levels λ_n , the h-r history here will be reduced to the usual ones for infinite or random horizon SMDPs with expected discount and average criteria [10,15,20].
- (b) It should be emphasized that the parameters t_n denote the remaining planning horizons (rather than the current time) at decision epochs, which are necessary for the decision maker to use and can be regarded as part of state information.
- (c) The case $\lambda_n < 0$ is allowed here since in some decision epochs with the planning horizon $t_n = 0$, the case $\lambda_n < 0$ indicates that the reward goal is achieved while the case $\lambda_n \geq 0$ does not.

To specify decision rules for the decision-makers to select actions, we define policies for the *finite horizon* SMDPs with *risk probability criteria*.

Definition 2.1. An h-r randomized historic policy, or simply an h-r policy, is a sequence $\pi = \{\pi_n, n \geq 0\}$ of stochastic kernels π_n on A given H_n satisfying

$$\pi_n(A(i_n) \mid h_n) = 1 \quad \forall h_n \in H_n, n = 0, 1, 2, \dots$$

- Remark 2.3.** (a) For each $n \geq 0$, the stochastic kernel π_n tells one how to select actions at the n th decision epoch based on the history information h_n .
- (b) It is worthwhile to notice that the h-r policy here generalizes the usual ones for (infinite or random horizon) SMDPs [10,12,15,20]. In fact, by taking the planning horizons $t_n \equiv \infty$ at every decision epoch (since there are always infinite planning horizons at decision epochs in the infinite horizon problems), the h-r policy here will be reduced to the usual ones for (infinite or random horizon) SMDPs with risk probability criteria [12]. If, furthermore, the h-r policy here is independent of the reward levels λ_n , it will be reduced to the usual one for (infinite or random horizon) SMDPs with expected discount and average criteria [10,15,20].

The set of all h-r policies is denoted by Π . To define the subclasses of Π , we introduce the following notation.

Notation. Let Φ represent the set of stochastic kernels φ on A given $E \times \mathbb{R}_+ \times \mathbb{R}$ such that $\varphi(A(i) \mid i, t, \lambda) = 1$ for all $(i, t, \lambda) \in E \times \mathbb{R}_+ \times \mathbb{R}$, and \mathbb{F} the set of measurable functions $f : E \times \mathbb{R}_+ \times \mathbb{R} \rightarrow A$ such that $f(i, t, \lambda)$ is in $A(i)$ for all $(i, t, \lambda) \in E \times \mathbb{R}_+ \times \mathbb{R}$.

- Definition 2.2.** (a) An h-r policy $\pi = \{\pi_n\}$ is said to be h-r randomized Markov if there is a sequence $\{\varphi_n\}$ of stochastic kernels $\varphi_n \in \Phi$ such that $\pi_n(\cdot \mid h_n) = \varphi_n(\cdot \mid i_n, t_n, \lambda_n)$ for every $h_n \in H_n$ and $n \geq 0$. We write such a policy as $\pi = \{\varphi_n\}$.
- (b) An h-r randomized Markov policy $\pi = \{\varphi_n\}$ is said to be h-r randomized stationary if φ_n are independent of n . In this case, we write $\pi = \{\varphi, \varphi, \dots\}$ as φ for simplicity.
- (c) An h-r randomized Markov policy $\pi = \{\varphi_n\}$ is said to be h-r deterministic Markov if there is a sequence $\{f_n\}$ of measurable functions $f_n \in \mathbb{F}$ such that $\varphi_n(\cdot \mid i, t, \lambda)$ is the Dirac measure at $f_n(i, t, \lambda)$ for every $(i, t, \lambda) \in E \times \mathbb{R}_+ \times \mathbb{R}$ and $n \geq 0$. We write such a policy as $\pi = \{f_n\}$.
- (d) An h-r deterministic Markov policy $\pi = \{f_n\}$ is said to be h-r deterministic stationary if f_n are independent of n , and in this case it is simply referred to as an h-r stationary policy. We write $\pi = \{f, f, \dots\}$ as f for simplicity.

For convenience, we denote by Π_{RM} , Π_{RS} , Π_{DM} , and Π_{DS} the families of all h-r randomized Markov, h-r randomized stationary, h-r deterministic Markov, and h-r stationary policies, respectively. Obviously, $\Phi = \Pi_{\text{RS}} \subset \Pi_{\text{RM}} \subset \Pi$, and $\mathbb{F} = \Pi_{\text{DS}} \subset \Pi_{\text{DM}} \subset \Pi$.

Remark 2.4. An h-r stationary policy $f(i, t, \lambda)$ here for the finite horizon SMDPs with risk probability criteria means that the action selection is independent of the decision number n and the current time but may depend on the system state i , the planning horizon t and the reward level λ at decision epochs, which generalizes the usual stationary policies in SMDPs. In fact, if we take the planning horizon $t = \infty$ at every decision epoch (since there are always infinite planning horizons at decision epochs in the infinite horizon problems), the h-r stationary policy $f(i, t, \lambda)$ here will be reduced to the usual one $f(i, \infty, \lambda) = f(i, \lambda)$ for (infinite or random horizon) SMDPs with risk probability criteria [12].

Let (Ω, \mathcal{F}) be the measurable space consisting of the sample space Ω given by

$$\Omega = \left\{ (s_0, i_0, t_0, \lambda_0, a_0, \dots, s_n, i_n, t_n, \lambda_n, a_n, \dots) \mid s_0 = 0, s_{n+1} \geq s_n, (i_n, a_n) \in K, t_0 \in \mathbb{R}_+, \right. \\ \left. t_{n+1} := [t_n - (s_{n+1} - s_n)]^+, \lambda_0 \in \mathbb{R}, \text{ and } \lambda_{n+1} := \lambda_n - r(i_n, a_n)(s_{n+1} - s_n) \text{ for } n = 0, 1, \dots \right\},$$

and the corresponding Borel σ -algebra \mathcal{F} . Then, we define random variables S_n, J_n, T_n, λ_n and $A_n (n = 0, 1, \dots)$ on (Ω, \mathcal{F}) as follows:

For each $\omega = (s_0, i_0, t_0, \lambda_0, a_0, \dots, s_n, i_n, t_n, \lambda_n, a_n, \dots) \in \Omega$, let

$$S_n(\omega) = s_n, \quad J_n(\omega) = i_n, \quad T_n(\omega) = t_n, \quad \lambda_n(\omega) = \lambda_n, \quad A_n(\omega) = a_n,$$

where S_n denotes the n th decision epoch, while $J_n, T_n := [T_{n-1} - (S_n - S_{n-1})]^+, \lambda_n := \lambda_{n-1} - r(J_{n-1}, A_{n-1})(S_n - S_{n-1})$ and A_n denote the system state, the planning horizon, the reward level and the action chosen at the n th decision epoch, respectively. Moreover, for each $(i, t, \lambda) \in E \times \mathbb{R}_+ \times \mathbb{R}$ and $\pi \in \Pi$, by the well-known Tulcea's theorem [8, Proposition C.10], we can construct a probability measure $P_{(i,t,\lambda)}^\pi$ on (Ω, \mathcal{F}) such that, for each $s \in \mathbb{R}_+, j \in E, a \in A$ and $h_n \in H_n, n = 0, 1, \dots$,

$$P_{(i,t,\lambda)}^\pi(S_0 = 0, J_0 = i, T_0 = t, \lambda_0 = \lambda) = 1, \quad (2.2)$$

$$P_{(i,t,\lambda)}^\pi(A_n = a \mid h_n) = \pi_n(a \mid h_n), \quad (2.3)$$

$$P_{(i,t,\lambda)}^\pi(S_{n+1} - S_n \leq s, J_{n+1} = j \mid h_n, a_n) = Q(s, j \mid i_n, a_n), \quad (2.4)$$

$$P_{(i,t,\lambda)}^\pi(T_{n+1} = [t_n - (s_{n+1} - s_n)]^+ \mid h_n, a_n, s_{n+1}) = 1, \quad (2.5)$$

$$P_{(i,t,\lambda)}^\pi(\lambda_{n+1} = \lambda_n - r(i_n, a_n)(s_{n+1} - s_n) \mid h_n, a_n, s_{n+1}) = 1. \quad (2.6)$$

Remark 2.5. Let $X_0 := 0, X_n := S_n - S_{n-1} (n \geq 1)$ denote the sojourn times between the $(n-1)$ th and the n th decision epochs. Then, the stochastic process $\{S_n, J_n, T_n, \lambda_n, A_n, n \geq 0\}$ may be rewritten as the one $\{X_n, J_n, T_n, \lambda_n, A_n, n \geq 0\}$, where $T_n := [T_{n-1} - X_n]^+$, and $\lambda_n := \lambda_{n-1} - r(J_{n-1}, A_{n-1})X_n$.

Corresponding to the stochastic process $\{S_n, J_n, A_n, n \geq 0\}$, we define an underlying continuous-time state-action process $\{Z(t), W(t), t \in \mathbb{R}_+\}$ (depending on a given policy π , which is omitted here) by

$$Z(t) = \begin{cases} J_n, & \text{for } S_n \leq t < S_{n+1}, n = 0, 1, \dots, \\ \partial_E, & \text{for } t \geq S_\infty, \end{cases} \\ W(t) = \begin{cases} A_n, & \text{for } S_n \leq t < S_{n+1}, n = 0, 1, \dots, \\ \partial_A, & \text{for } t \geq S_\infty, \end{cases}$$

where ∂_E and ∂_A are the extra state and action joined to E and A , respectively, and S_∞ is the accumulation point of the sequence $\{S_n\}$, i.e., $S_\infty := \lim_{n \rightarrow \infty} S_n$.

Definition 2.3. The stochastic process $\{Z(t), W(t), t \in \mathbb{R}_+\}$ is called a semi-Markov decision process.

To treat the finite horizon optimality, we fix an arbitrary T -horizon (with $T \in \mathbb{R}_+$) without loss of generality. First of all, to make the T -horizon SMDPs sensible, we need a basic assumption below.

Assumption 2.1. For all $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$ and $\pi \in \Pi, P_{(i,t,\lambda)}^\pi(\{S_\infty > T\}) = 1$.

Remark 2.6. Assumption 2.1 above is trivially fulfilled in DTMDPs [3,4,25,26], where one has that $S_\infty = \infty$.

Under Assumption 2.1, the possibility of an infinite number of decision epochs during the interval $[0, T]$ can be avoided. We suppose that Assumption 2.1 holds throughout this paper. Note that, although Assumption 2.1 is natural and mild, it is not easy to verify in practice. For ease of verification of Assumption 2.1, we may use the following fact.

Proposition 2.1. If there exist $\delta > 0$ and $\epsilon > 0$ such that

$$D(\delta \mid i, a) := Q(\delta, E \mid i, a) \leq 1 - \epsilon \quad \forall (i, a) \in K, \quad (2.7)$$

then Assumption 2.1 holds.

Proof. The proof is similar to the one of Proposition 2.1 in Huang and Guo [9]. \square

Remark 2.7. Note that the condition (2.7) above is imposed on the *primitive data* of the model (2.1) and is thus easy to verify. In fact, the condition (2.7) is the standard regular condition widely used in SMDPs; see, for instance, [10,20] for expected discount criteria; [15,20] for expected average criteria; and [11,10] for first passage criteria.

For each $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$, we define the risk probability (risk function) F^π of the SMDP $\{Z(t), W(t)\}$ under a policy $\pi \in \Pi$ by

$$F^\pi(i, t, \lambda) := P_{(i,t,\lambda)}^\pi \left(\int_0^t r(Z(s), W(s)) ds \leq \lambda \right), \quad (2.8)$$

where i is the initial state, t is the planning horizon, and λ is the reward level, respectively. Indeed, F^π measures the risk of the system that the total reward obtained during the interval $[0, t]$ does not exceed level λ when using policy π . Then, the optimization problem we are interested in is to minimize the system's risk F^π over $\pi \in \Pi$. That is, our aim is at finding a policy $\pi^* \in \Pi$ such that

$$F^{\pi^*}(i, t, \lambda) = F^*(i, t, \lambda) \quad \forall (i, t, \lambda) \in E \times [0, T] \times \mathbb{R},$$

where $F^*(i, t, \lambda) := \inf_{\pi \in \Pi} F^\pi(i, t, \lambda)$ is the optimal value (or minimum risk) function. Such a policy π^* , when it exists, is called *optimal*.

Remark 2.8. Compared to the previous study on the risk probability criteria with infinite or random horizon total rewards [12], this paper considers the risk probability criteria with a finite horizon total reward.

Similar to the result in [9, Proposition 2.2], we can also show that the family of all h-r randomized Markov policies is “sufficient” in the class of h-r randomized historic policies, that is, we have

$$F^*(i, t, \lambda) = \inf_{\pi \in \Pi_{\text{RM}}} F^\pi(i, t, \lambda) \quad \forall (i, t, \lambda) \in E \times [0, T] \times \mathbb{R}.$$

Therefore, it suffices to seek optimal policies in Π_{RM} , and hence we limit ourselves to Π_{RM} in the following.

3. Main results

In this section, we show our main results, for which we establish the approximation of the optimal value F^* (Theorem 3.1), the optimality equation (Theorem 3.2), and the existence of optimal policies (Theorem 3.3), respectively.

To characterize the optimal value F^* and optimal policies, we need to introduce some notation. Let \mathcal{F}_m be the set of functions $F : E \times [0, T] \times \mathbb{R} \rightarrow [0, 1]$, such that $F(\cdot, \cdot, \cdot)$ is Borel-measurable on $E \times [0, T] \times \mathbb{R}$ and $F(i, t, \lambda) = 0$ if $\lambda < 0$ for each $(i, t) \in E \times [0, T]$; and \mathcal{F}_r the set of functions $F \in \mathcal{F}_m$ such that $F(i, t, \cdot)$ is monotone nondecreasing and right-continuous on \mathbb{R} for each $(i, t) \in E \times [0, T]$, while $F(i, \cdot, \lambda)$ is monotone non-increasing and left-continuous on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. Also, we define operators H^a, H^φ, H on \mathcal{F}_m as follows: for $F \in \mathcal{F}_m, (i, t) \in E \times [0, T], a \in A(i)$ and $\varphi \in \Phi$, if $\lambda \geq 0$,

$$H^a F(i, t, \lambda) := \mathbb{1}_{[0, \lambda]}(r(i, a)t)(1 - D(t | i, a)) + \sum_{j \in E} \int_0^t Q(du, j | i, a) F(j, t - u, \lambda - r(i, a)u), \quad (3.1)$$

$$H^\varphi F(i, t, \lambda) := \sum_{a \in A(i)} \varphi(a | i, t, \lambda) H^a F(i, t, \lambda), \quad (3.2)$$

$$H F(i, t, \lambda) := \min_{a \in A(i)} H^a F(i, t, \lambda), \quad (3.3)$$

and $H^a F(i, t, \lambda) = H^\varphi F(i, t, \lambda) = H F(i, t, \lambda) := 0$ if $\lambda < 0$, where $\mathbb{1}_C$ is the indicator function on a set C .

Remark 3.1. The definitions of operators H^a, H^φ and H above are mainly based on the semi-Markov kernel $Q(\cdot, \cdot | i, a)$ in the model (2.1). Note that they are different from those for random horizon SMDPs with risk probability criteria [12], and those for finite horizon DTMDPs with risk probability criteria [3,4,25,26]. In fact, these operators are specifically used to characterize the optimal value F^* and optimal policies for finite horizon SMDPs with risk probability criteria; see Theorems 3.1–3.3 below.

To establish algorithms for computing F^π and F^* , we note that for each $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$ and $\pi \in \Pi$,

$$\begin{aligned} F^\pi(i, t, \lambda) &= P_{(i,t,\lambda)}^\pi \left(\int_0^t r(Z(s), W(s)) ds \leq \lambda \right) \\ &= P_{(i,t,\lambda)}^\pi \left(\sum_{m=0}^{\infty} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda \right) \end{aligned}$$

$$\begin{aligned}
&= P_{(i,t,\lambda)}^\pi \left(\bigcap_{n=1}^{\infty} \left\{ \sum_{m=0}^n \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda \right\} \right) \\
&= \lim_{n \rightarrow \infty} P_{(i,t,\lambda)}^\pi \left(\sum_{m=0}^n \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda \right),
\end{aligned} \tag{3.4}$$

where the second equality follows from [Assumption 2.1](#), and the last equality is due to the nonnegativity of the reward rate $r(i, a)$ and the continuity of probability measures. Based on (3.4), we define $F_{-1}^\pi(i, t, \lambda) := \mathbb{1}_{[0, \infty)}(\lambda)$, and

$$F_n^\pi(i, t, \lambda) := \begin{cases} P_{(i,t,\lambda)}^\pi \left(\sum_{m=0}^n \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda \right), & \lambda \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

for every $(i, t) \in E \times [0, T]$ and $n \geq 0$. Clearly, $F_n^\pi \geq F_{n+1}^\pi$ for every $n \geq -1$, and $\lim_{n \rightarrow \infty} F_n^\pi = F^\pi$.

Remark 3.2. Note that [Assumption 2.1](#) is essential to derive the equality (3.4) above. In fact, we use the approximation $F_n^\pi \rightarrow F^\pi$ to characterize F^π ; see [Lemma 3.1](#) below.

The following lemma is fundamental to our main results.

Lemma 3.1. Let $\pi = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{\text{RM}}$ be arbitrary. For each $n \geq -1$,

- (a) $F_n^\pi \in \mathcal{F}_m$ and $F^\pi \in \mathcal{F}_m$.
- (b) $F_{n+1}^\pi = H^{\varphi_0} F_n^{(1)\pi}$, and $F^\pi = H^{\varphi_0} F^{(1)\pi}$, where $(1)\pi = \{\varphi_1, \varphi_2, \dots\} \in \Pi_{\text{RM}}$. In particular, $F^f = H^f F^f$ for every $f \in \mathbb{F}$.

Proof. See Section 5. \square

Remark 3.3. [Lemma 3.1](#) provides a way of computing the risk function F^f for each policy $f \in \mathbb{F}$, that is, $F^f = \lim_{n \rightarrow \infty} F_n^f$, where F_n^f are recursively defined by $F_{-1}^f(i, t, \lambda) := \mathbb{1}_{[0, \infty)}(\lambda)$, $F_n^f(i, t, \lambda) := H^f F_{n-1}^f(i, t, \lambda)$ for $n \geq 0$.

We now state our first main result, which proposes an iteration algorithm for computing the optimal value function F^* .

Theorem 3.1. Suppose that [Assumption 2.1](#) holds. For each $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$, let

$$F_{-1}^*(i, t, \lambda) := \mathbb{1}_{[0, \infty)}(\lambda), \quad F_{n+1}^*(i, t, \lambda) := H F_n^*(i, t, \lambda), \quad n \geq -1.$$

Then $\lim_{n \rightarrow \infty} F_n^* = F^*$, and $F^* \in \mathcal{F}_r$.

Proof. See Section 5. \square

The second main result below establishes the so-called optimality equation, which is useful for solving optimal policies.

Theorem 3.2. Suppose that [Assumption 2.1](#) holds. Then

- (a) F^* satisfies the optimality equation $F^* = H F^*$.
- (b) There exists an $f \in \mathbb{F}$ such that $F^* = H^f F^*$.

Proof. See Section 5. \square

Remark 3.4. In fact, the finite horizon SMDPs with risk probability criteria here can be reduced to the case of finite horizon DTMDPs with risk probability criteria [3,4,25,26]. Indeed, suppose that the semi-Markov kernel $Q(\cdot, \cdot \mid i, a)$ is of the specific form

$$Q(t, j \mid i, a) = \begin{cases} p(j \mid i, a), & t \geq 1, \\ 0, & \text{otherwise} \end{cases}$$

for every $j \in E$, $(i, a) \in K$, where $\{p(\cdot \mid i, a)\}$ is a stochastic kernel on E given K . Then, the decision epochs $S_m \equiv m$ for every $m \geq 0$. Therefore, for every $(i, n, \lambda) \in E \times \{0, 1, \dots, N\} \times \mathbb{R}$, by (3.4), the risk function $F^\pi(i, n, \lambda)$ defined in (2.8) is reduced to the following form

$$\begin{aligned}
F^\pi(i, n, \lambda) &= P_{(i,n,\lambda)}^\pi \left(\sum_{m=0}^{\infty} \int_{m \wedge n}^{(m+1) \wedge n} r(Z(s), W(s)) ds \leq \lambda \right) \\
&= P_{(i,n,\lambda)}^\pi \left(\sum_{m=0}^{n-1} r(J_m, A_m) \leq \lambda \right),
\end{aligned} \tag{3.5}$$

and moreover, [Theorem 3.2](#) gives the optimality equation

$$F^*(i, n, \lambda) = \min_{a \in A(i)} \left[\sum_{j \in E} p(j | i, a) F^*(j, n-1, \lambda - r(i, a)) \right],$$

which exactly coincides with those for finite horizon DTMDPs with risk probability criteria [[3,4,25,26](#)]. It is worth mentioning that, however, the definition of policies here is different from the usual ones for DTMDPs since in the policies in this paper, the planning horizons at decision epochs are taken as part of the state information rather than as time dimensions in the usual policies for DTMDPs [[3,4,25,26](#)].

In general, a policy $f \in \mathbb{F}$ satisfying $F^* = H^f F^*$ may not be optimal. To ensure such a policy to be optimal, we need an additional condition below.

Assumption 3.1. There exist constants $\delta > 0$ and $\epsilon > 0$ such that

$$D(\delta | i, a) \leq 1 - \epsilon \quad \forall (i, a) \in K.$$

Remark 3.5. (a) Note that [Assumption 3.1](#) is the same as the condition (2.7) in [Proposition 2.1](#), and hence [Assumption 3.1](#) implies [Assumption 2.1](#).

(b) [Assumption 3.1](#) can be fulfilled in many situations; for example, it is obviously satisfied in the model of DTMDPs; furthermore, it can be also verified when the state space E and the action set A are both finite, and the sojourn times are exponential-distributed.

Lemma 3.2. Suppose that [Assumption 3.1](#) holds. Then, for any $f \in \mathbb{F}$, F^f is the unique solution in \mathcal{F}_m to the equation $F = H^f F$.

Proof. See Section 5. \square

Theorem 3.3. Suppose that [Assumption 3.1](#) holds. Then

- (a) Any policy $f \in \mathbb{F}$ such that $F^* = H^f F^*$ is optimal.
- (b) There exists an optimal policy.

Proof. See Section 5. \square

Remark 3.6. In view of [Theorem 3.3](#), we need only [Assumption 3.1](#) to ensure the existence of optimal policies since [Assumption 3.1](#) also implies [Assumption 2.1](#).

Combining [Theorems 3.1–3.3](#), we can derive an algorithm for solving optimal policies. The computing procedure includes two steps as below.

Algorithm for computing optimal policies:

- Step 1.** Compute F^* by the iteration algorithm proposed in [Theorem 3.1](#). The iteration stops in finite steps when $|F_{n+1}^* - F_n^*| < \epsilon$ for some sufficiently small number $\epsilon > 0$. Such a value F_{n+1}^* closely approximates the precise value F^* , and thus, in real computation, the approximate value F_{n+1}^* is usually regarded as the precise value F^* .
- Step 2.** Find a policy $f^* \in \mathbb{F}$ satisfying $H F^* = H^{f^*} F^*$. Then, by [Theorem 3.3](#), the policy f^* obtained is optimal; for details, see [Example 4.1](#) below.

4. Application to manufacturing systems

This section applies our results to manufacturing systems [[5,13,18,19](#)], in which we show how to compute the optimal value and an optimal policy based on the iteration algorithm.

Example 4.1 (*A Manufacturing System*). Consider a failure-prone manufacturing system with three states, say 1, 2 and 3, which represent the good, the medium and the failure ones, respectively. The production rates can be chosen by the decision-makers. When the system occupies state 1, the production rate $r(1, a_{11})$ with action a_{11} or the one $r(1, a_{12})$ with action a_{12} may be taken, while in state 2 the production rate $r(2, a_{21})$ with action a_{21} or the one $r(2, a_{22})$ with action a_{22} may be used. However, the system in state 3 produces at rate $r(3, a_{31}) = 0$ with non-action denoted by a_{31} . We suppose that the transition mechanism of the system obeys the dynamic of an SMDP as follows. Whenever the action a ($a = a_{11}$ or a_{12}) is applied, the system transits to state j with probability $p(j|1, a)$ ($j = 1, 2, 3$) after staying at state 1 for a random time uniformly-distributed in the region $[0, \mu(1, a)]$ with $\mu(1, a) > 0$. On the other hand, if the action a ($a = a_{21}$ or a_{22}) is performed when the system occupies state 2, the system transits to state j with probability $p(j|2, a)$ ($j = 1, 2, 3$) after an exponential-distributed random time with parameter $\mu(2, a) > 0$. Finally, we assume that when it falls in state 3, the system transits to state 3 with probability one after an exponential-distributed random time with parameter $\mu(3, a_{31})$. For this manufacturing system, the decision maker wishes to find an optimal policy with the minimum risk probability over a finite horizon $[0, T]$ with some $T > 0$.

Table 4.1
The data of the model.

State i	Action a	Sojourn time $\mu(i, a)$	Transition probability $p(j i, a)$		Reward rate $r(i, a)$	Horizon T
			$j = 2$	$j = 3$		
1	a_{11}	25	0.9	0.1	4	15
	a_{12}	20	0.7	0.3	5	
2	a_{21}	0.12	0.6	0.4	2	
	a_{22}	0.15	0.3	0.7	3	
3	a_{31}	0.3	0	1	0	

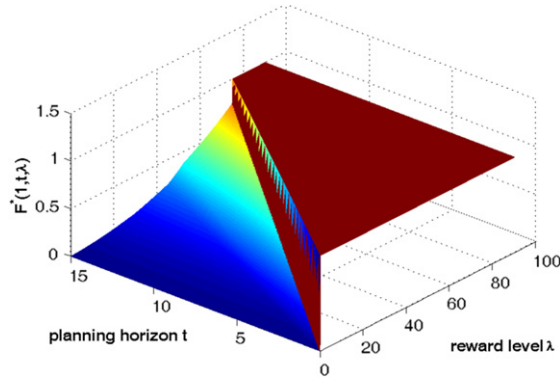


Fig. 4.1a. The value function $F^*(1, t, \lambda)$.

Remark 4.1. In the manufacturing system, action choice usually occurs at random times, and thus it is more natural to model the system as an SMDP rather than as a DTMDP. Here we use uniformly and exponentially distributed sojourn times as examples to show the generality of SMDPs that arbitrary distributions are allowed for the sojourn times. In fact, other suitable distributions can be chosen for the sojourn times according to the actual situations.

First, we compute the optimal value F^* in MATLAB software by using the value iteration proposed in Theorem 3.1. To do so, we suppose that the data of the model are specifically given as follows.

Remark 4.2. From the data in Table 4.1 above, we see that actions a_{11} and a_{21} cause slower production rates but lead to longer working durations and less probabilities of failure, while actions a_{12} and a_{22} lead to faster production rates but cause shorter working durations and more probabilities of failure. In the following, we will show that which actions are better depends on the planning horizons and reward levels, and more importantly, we will give the exact planning horizons and reward levels determining when to take the faster production rates or the slower ones.

Under the data above, Assumption 3.1 holds in this example with $\delta = 1$ and $\epsilon = \min\{24/25, 19/20, e^{-0.12}, e^{-0.15}, e^{-0.3}\} > 0$, and thus Assumption 2.1 is also fulfilled. Therefore, the value iteration in Theorem 3.1 is valid, and an optimal policy is ensured by Theorem 3.3.

Note that, since $r(3, a_{31}) = 0$ and state 3 is absorbing, $F^*(3, t, \lambda) = \mathbb{1}_{[0, \infty)}(\lambda)$ for all $(t, \lambda) \in [0, 15] \times \mathbb{R}$. Hence, we focus on computing the functions $F^*(1, t, \lambda)$ and $F^*(2, t, \lambda)$, for which the computational results are shown in Fig. 4.1.

From Fig. 4.1, it is clear that the optimal value function (i.e., the minimum risk) $F^*(i, t, \lambda)$ is monotone nondecreasing and right-continuous in λ for each $(i, t) \in \{1, 2\} \times [0, 15]$, while $F^*(i, t, \lambda)$ is monotone non-increasing and left-continuous in t for each $(i, \lambda) \in \{1, 2\} \times [0, 90]$.

To obtain an optimal policy with the minimum risk probability, we should compare the data $H^a F^*(i, t, \lambda)$ under admissible actions a for every $(i, t, \lambda) \in \{1, 2\} \times [0, 15] \times \mathbb{R}$. To be specific, we analyze the data $H^a F^*(i, 10, \lambda)$ and $H^a F^*(i, 15, \lambda)$ as examples, which are shown in Fig. 4.2.

In view of Fig. 4.2, we have the following observations.

- When there is a planning horizon $t = 10$ at some decision epoch, a_{11} is with lower risk than a_{12} if the reward level $\lambda \in (0, 34.45)$, but action a_{12} is with lower risk than action a_{11} if the reward level $\lambda \in (34.45, 90)$. On the other hand, when there is a planning horizon $t = 15$ at some decision epoch, a_{11} is with lower risk than a_{12} if the reward level $\lambda \in (0, 52.25)$, but action a_{12} is with lower risk than action a_{11} if the reward level $\lambda \in (52.25, 90)$. Similar conclusions can be drawn for the actions a_{21} and a_{22} .
- The facts above implies that for a fixed planning horizon, the slower production rates may more likely reach a smaller reward goal, while the faster production rates may more likely reach a larger reward goal. Moreover, for a fixed reward

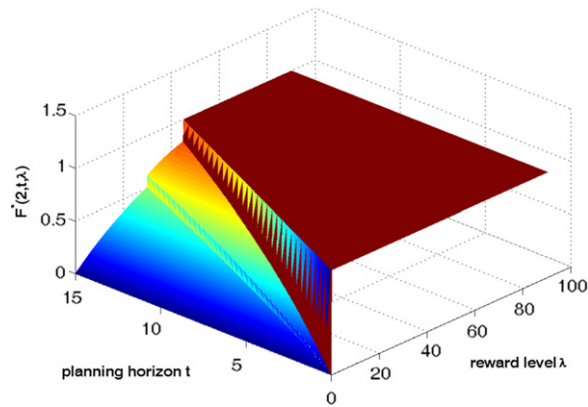


Fig. 4.1b. The value function $F^*(2, t, \lambda)$.

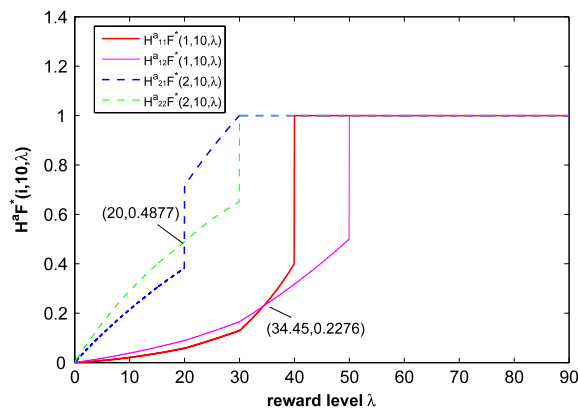


Fig. 4.2a. The function $H^a F^*(i, 10, \lambda)$.

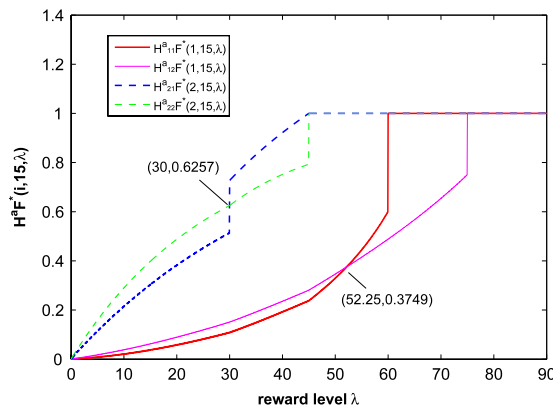


Fig. 4.2b. The function $H^a F^*(i, 15, \lambda)$.

goal λ , the faster production rates may more likely reach the goal when the planning horizon is small, while the slower production rates may more likely reach the goal when the planning horizon is large. To summarize, the faster production rates may lead to lower risk as the reward goal becomes large, whereas the slower production rates may lead to lower risk as the planning horizon increases.

As exhibited in Fig. 4.2 above, to obtain an explicit optimal policy, it is of importance to find critical points $\lambda^*(i, t)$ (depending on the states and planning horizons) from stored values of $H^a F^*$ for each $(i, t) \in \{1, 2\} \times [0, 15]$, for which action a_{i1} is with lower risk if $\lambda < \lambda^*(i, t)$ and a_{i2} is with lower risk if $\lambda \geq \lambda^*(i, t)$. Critical points $\lambda^*(i, t)$ are graphed below.

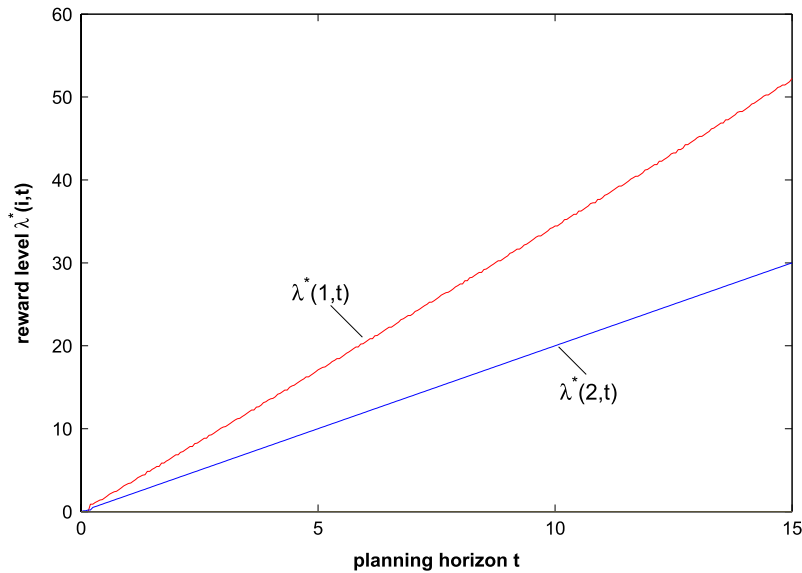


Fig. 4.3. The function $\lambda^*(i, t)$.

In the light of Fig. 4.3 and the analysis above, we can define a policy f^* by

$$f^*(1, t, \lambda) = \begin{cases} a_{11}, & \lambda < \lambda^*(1, t), \\ a_{12}, & \lambda \geq \lambda^*(1, t); \end{cases} \quad f^*(2, t, \lambda) = \begin{cases} a_{21}, & \lambda < \lambda^*(2, t), \\ a_{22}, & \lambda \geq \lambda^*(2, t) \end{cases}$$

such that $F^*(i, t, \lambda) = H^{f^*} F^*(i, t, \lambda)$ for every $(i, t, \lambda) \in \{1, 2\} \times [0, 15] \times \mathbb{R}$; for example, when $t = 15$, we have

$$f^*(1, 15, \lambda) = \begin{cases} a_{11}, & \lambda < 52.25, \\ a_{12}, & \lambda \geq 52.25, \end{cases} \quad f^*(2, 15, \lambda) = \begin{cases} a_{21}, & \lambda < 30, \\ a_{22}, & \lambda \geq 30; \end{cases}$$

and when $t = 10$, we have

$$f^*(1, 10, \lambda) = \begin{cases} a_{11}, & \lambda < 34.45, \\ a_{12}, & \lambda \geq 34.45, \end{cases} \quad f^*(2, 10, \lambda) = \begin{cases} a_{21}, & \lambda < 20, \\ a_{22}, & \lambda \geq 20. \end{cases}$$

Then, (by Theorem 3.3) such a policy f^* is optimal with the minimum risk probability.

• *Suggestions from the optimal policy f^* :* For a fixed planning horizon t , it is optimal to perform the slower production rates with less probability of failure when the reward goal is small, and the faster production rates with more probability of failure when the reward goal is large. On the other hand, for a fixed reward goal λ , it is optimal to use the faster production rates when the planning horizon is small, and the slower production rates when the planning horizon is large.

5. Proofs of main results

This part is devoted to the proofs of our main results.

5.1. Proofs of Lemma 3.1 and Theorems 3.1 and 3.2

To begin with, we give some properties about the operators H^a and H .

Lemma 5.1. Suppose that Assumption 2.1 holds.

- (a) If $G \in \mathcal{F}_m$, then $HG \in \mathcal{F}_m$, and there exists an $f \in \mathbb{F}$ such that $HG = H^f G$.
- (b) If $G \in \mathcal{F}_r$, then both $H^a G$ and HG are in \mathcal{F}_r for any $a \in A$.
- (c) If $G_n \in \mathcal{F}_r$ and $G_n \geq G_{n+1}$ for each $n \geq 0$, then $\lim_{n \rightarrow \infty} G_n \in \mathcal{F}_r$.

Proof. (a) Under Assumption 2.1, the measurable selection theorem (see [2, Proposition 7.33, p. 153]) ensures the existence of an $f \in \mathbb{F}$ such that

$$H^f G(i, t, \lambda) = \min_{a \in A(i)} \{H^a G(i, t, \lambda)\} = HG(i, t, \lambda)$$

for each $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$, and thus (a) follows.

(b) Obviously, it follows from the definition of H^a and $G \in \mathcal{F}_r$ that $H^a G \in \mathcal{F}_m$, and furthermore, $H^a G(i, t, \cdot)$ is monotone nondecreasing and right-continuous on \mathbb{R} for each $(i, t) \in E \times [0, T]$, and on the other hand, $H^a G(i, \cdot, \lambda)$ is left-continuous on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. To prove that $H^a G \in \mathcal{F}_r$, we need only show that $H^a G(i, \cdot, \lambda)$ is monotone non-increasing on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. Indeed, for fixed $(i, \lambda) \in E \times \mathbb{R}$ and $a \in A(i)$, if $t_2 > t_1 > \lambda/r(i, a)$, we see that

$$\begin{aligned} H^a G(i, t_2, \lambda) - H^a G(i, t_1, \lambda) &= \sum_{j \in E} \int_0^{\lambda/r(i, a)} Q(du, j|i, a) G(j, t_2 - u, \lambda - r(i, a)u) \\ &\quad - \sum_{j \in E} \int_0^{\lambda/r(i, a)} Q(du, j|i, a) G(j, t_1 - u, \lambda - r(i, a)u) \leq 0, \end{aligned}$$

and if $t_1 < t_2 \leq \lambda/r(i, a)$, we find that

$$\begin{aligned} H^a G(i, t_2, \lambda) - H^a G(i, t_1, \lambda) &= [1 - D(t_2 | i, a)] + \sum_{j \in E} \int_0^{t_2} Q(du, j|i, a) G(j, t_2 - u, \lambda - r(i, a)u) \\ &\quad - [1 - D(t_1 | i, a)] - \sum_{j \in E} \int_0^{t_1} Q(du, j|i, a) G(j, t_1 - u, \lambda - r(i, a)u) \\ &= [D(t_1 | i, a) - D(t_2 | i, a)] + \sum_{j \in E} \int_{t_1}^{t_2} Q(du, j|i, a) \\ &\quad \times G(j, t_2 - u, \lambda - r(i, a)u) + \sum_{j \in E} \int_0^{t_1} Q(du, j|i, a) \\ &\quad \times [G(j, t_2 - u, \lambda - r(i, a)u) - G(j, t_1 - u, \lambda - r(i, a)u)] \\ &\leq [D(t_1 | i, a) - D(t_2 | i, a)] + \sum_{j \in E} \int_{t_1}^{t_2} Q(du, j|i, a) \times 1 + 0 \\ &\leq 0, \end{aligned}$$

which implies that $H^a G(i, \cdot, \lambda)$ is monotone non-increasing on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$.

Using this fact $H^a G \in \mathcal{F}_r$ together with part (a), it is clear that $HG \in \mathcal{F}_m$, and moreover, $HG(i, t, \cdot)$ is monotone nondecreasing on \mathbb{R} for each $(i, t) \in E \times [0, T]$, and on the other hand, $HG(i, \cdot, \lambda)$ is monotone non-increasing on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. Hence, to prove that HG is in \mathcal{F}_r , it need only to show that $HG(i, t, \cdot)$ is right-continuous on \mathbb{R} for each $(i, t) \in E \times [0, T]$, and on the other hand, $HG(i, \cdot, \lambda)$ is left-continuous on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. Indeed, for every fixed $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$, take an arbitrary sequence $\{\lambda_k\}$ in \mathbb{R} such that $\lambda_k \downarrow \lambda$. Then, we have $\lim_{\lambda_k \downarrow \lambda} H^a G(i, t, \lambda_k) = H^a G(i, t, \lambda)$ for any $a \in A(i)$. On the one hand, since $HG(i, t, \lambda_k) \leq H^a G(i, t, \lambda_k)$, we obtain that

$$\limsup_{\lambda_k \downarrow \lambda} HG(i, t, \lambda_k) \leq \limsup_{\lambda_k \downarrow \lambda} H^a G(i, t, \lambda_k) = H^a G(i, t, \lambda),$$

which together with the arbitrariness of a yields that

$$\limsup_{\lambda_k \downarrow \lambda} HG(i, t, \lambda_k) \leq HG(i, t, \lambda). \quad (5.1)$$

On the other hand, noting that $HG(i, t, \cdot)$ is nondecreasing on \mathbb{R} , we have

$$HG(i, t, \lambda_k) \geq HG(i, t, \lambda),$$

which implies that

$$\liminf_{\lambda_k \downarrow \lambda} HG(i, t, \lambda_k) \geq HG(i, t, \lambda). \quad (5.2)$$

Combining (5.1) with (5.2) gives that $\lim_{\lambda_k \downarrow \lambda} HG(i, t, \lambda_k) = HG(i, t, \lambda)$. Similarly, we can show that $HG(i, \cdot, \lambda)$ is left-continuous on $[0, T]$ for each $(i, \lambda) \in E \times \mathbb{R}$. Therefore, the proof of part (b) is achieved.

(c) Note that $G_n \geq G_{n+1} \geq 0$ for each $n \geq 0$, and so the limit $G := \lim_{n \rightarrow \infty} G_n$ exists. We easily see that $G(i, t, \cdot)$ is nondecreasing on \mathbb{R} , and $G(i, \cdot, \lambda)$ is non-increasing on $[0, T]$. To prove $G \in \mathcal{F}_r$, it suffices to show that $G(i, t, \cdot)$ is right-continuous on \mathbb{R} , and $G(i, \cdot, \lambda)$ is left-continuous on $[0, T]$. We next show that $G(i, \cdot, \lambda)$ is left-continuous on $[0, T]$, while the fact that $G(i, t, \cdot)$ is right-continuous on \mathbb{R} can be similarly proved. For every $t \in [0, T]$ and any sequence $\{t_k\}$ in $[0, T]$ such that $t_k \uparrow t$, we have $G(i, t_k, \lambda) \leq G_n(i, t_k, \lambda)$ for any $n, k \geq 0$. Hence,

$$\limsup_{t_k \uparrow t} G(i, t_k, \lambda) \leq \limsup_{t_k \uparrow t} G_n(i, t_k, \lambda) = G_n(i, t, \lambda)$$

for any $n \geq 0$, which implies that $\limsup_{t_k \uparrow t} G(i, t_k, \lambda) \leq G(i, t, \lambda)$. On the other hand, since $G(i, t_k, \lambda) \geq G(i, t, \lambda)$, we obtain $\liminf_{t_k \uparrow t} G(i, t_k, \lambda) \geq G(i, t, \lambda)$, which together with the previous inequality yields $\lim_{t_k \uparrow t} G(i, t_k, \lambda) = G(i, t, \lambda)$. Therefore, $G \in \mathcal{F}_r$ and thus the proof is complete. \square

Proof of Lemma 3.1. (a) To show that $F_n^\pi \in \mathcal{F}_m$, it suffices to prove that $F_n^\pi(i, \cdot, \cdot)$ is Borel-measurable on $[0, T] \times \mathbb{R}$ for each $i \in E$. We establish this by induction. When $n = -1$, it is obviously true. Now assume $F_n^\pi(i, \cdot, \cdot)$ is Borel-measurable for some $n \geq -1$ and every $\pi \in \Pi_{RM}$. It then follows that for any $\pi = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{RM}$,

$$\begin{aligned} H^{\varphi_0} F_n^{(1)\pi}(i, t, \lambda) &= \sum_{i \in A(i)} \varphi_0(a|i, t, \lambda) \left[\mathbb{1}_{[0, \lambda]}(r(i, a)t)(1 - D(t | i, a)) \right. \\ &\quad \left. + \sum_{j \in E} \int_0^t Q(du, j|i, a) F_n^{(1)\pi}(j, t - u, \lambda - r(i, a)u) \right] \end{aligned}$$

is well defined and measurable in (t, λ) for each $i \in E$, where $^{(1)}\pi = \{\varphi_1, \varphi_2, \dots\} \in \Pi_{RM}$. On the other hand, for $\lambda < 0$, $F_{n+1}^\pi(i, t, \lambda) = H^{\varphi_0} F_n^{(1)\pi}(i, t, \lambda) = 0$, and for $\lambda \geq 0$, we have

$$\begin{aligned} F_{n+1}^\pi(i, t, \lambda) &= P_{(i, t, \lambda)}^\pi \left(\sum_{m=0}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda \right) \\ &= P_{(i, t, \lambda)}^\pi \left(\sum_{m=0}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda, S_1 > t \right) \\ &\quad + P_{(i, t, \lambda)}^\pi \left(\sum_{m=0}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda, S_1 \leq t \right) \\ &= P_{(i, t, \lambda)}^\pi \left(\int_0^t r(J_0, A_0) ds \leq \lambda, S_1 > t \right) \\ &\quad + P_{(i, t, \lambda)}^\pi \left(\sum_{m=1}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda - r(J_0, A_0)X_1, S_1 \leq t \right) \\ &= E_{(i, t, \lambda)}^\pi \left[P_{(i, t, \lambda)}^\pi(r(J_0, A_0)t \leq \lambda, S_1 > t | S_0, J_0, T_0, \lambda_0, A_0) \right] \\ &\quad + E_{(i, t, \lambda)}^\pi \left[P_{(i, t, \lambda)}^\pi \left(\sum_{m=1}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda - r(J_0, A_0)X_1, S_1 \leq t | S_0, J_0, T_0, \right. \right. \\ &\quad \left. \left. \lambda_0, A_0, S_1, J_1, T_1 := [T_0 - X_1]^+, \lambda_1 := \lambda_0 - r(J_0, A_0)X_1 \right) \right] \\ &= E_{(i, t, \lambda)}^\pi \left[\mathbb{1}_{[0, \lambda]}(r(J_0, A_0)t) P_{(i, t, \lambda)}^\pi(S_1 > t | S_0, J_0, T_0, \lambda_0, A_0) \right] \\ &\quad + E_{(i, t, \lambda)}^\pi \left[\mathbb{1}_{[S_1 \leq t]} P_{(i, t, \lambda)}^\pi \left(\sum_{m=1}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda - r(J_0, A_0)X_1 | S_0, J_0, T_0, \right. \right. \\ &\quad \left. \left. \lambda_0, A_0, S_1, J_1, T_1 := [T_0 - X_1]^+, \lambda_1 := \lambda_0 - r(J_0, A_0)X_1 \right) \right] \\ &= \sum_{i \in A(i)} \varphi_0(a|i, t, \lambda) \left[\mathbb{1}_{[0, \lambda]}(r(i, a)t)(1 - D(t | i, a)) \right] + \sum_{i \in A(i)} \varphi_0(a|i, t, \lambda) \sum_{j \in E} \int_0^t Q(du, j|i, a) \\ &\quad \times P_{(i, t, \lambda)}^\pi \left(\sum_{m=1}^{n+1} \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda - r(J_0, A_0)X_1 | S_0 = 0, J_0 = i, T_0 = t, \right. \\ &\quad \left. \lambda_0 = \lambda, A_0 = a, S_1 = u, J_1 = j, T_1 := [t - u]^+, \lambda_1 := \lambda_0 - r(i, a)u \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i \in A(i)} \varphi_0(a|i, t, \lambda) \left[\mathbb{1}_{[0, \lambda]}(r(i, a)t)(1 - D(t | i, a)) + \sum_{j \in E} \int_0^t Q(du, j|i, a) \right. \\
&\quad \times P_{(j, t-u, \lambda_0-r(i, a)u)}^{(1)\pi} \left(\sum_{m=0}^n \int_{S_m \wedge t}^{S_{m+1} \wedge t} r(Z(s), W(s)) ds \leq \lambda - r(i, a)u \right) \Big] \\
&= \sum_{i \in A(i)} \varphi_0(a|i, t, \lambda) \left[\mathbb{1}_{[0, \lambda]}(r(i, a)t)(1 - D(t | i, a)) + \sum_{j \in E} \int_0^t Q(du, j|i, a) \right. \\
&\quad \times F^{(1)\pi}(j, t-u, \lambda_0-r(i, a)u) \Big],
\end{aligned}$$

where the sixth equality follows from the properties (2.2)–(2.6), and the seventh equality is due to the Markov property of policy π and the properties (2.2)–(2.6) again. Hence,

$$F_{n+1}^\pi(i, t, \lambda) = H^{\varphi_0} F_n^{(1)\pi}(i, t, \lambda) \quad \forall (i, t, \lambda) \in E \times [0, T] \times \mathbb{R},$$

and thus $F_{n+1}^\pi(i, \cdot, \cdot)$ is measurable in (t, λ) for each $i \in E$. Therefore, by induction, $F_n^\pi(i, \cdot, \cdot)$ is measurable for every $n \geq -1$. Furthermore, since a (pointwise) limit of measurable functions is still measurable, we have $F^\pi = \lim_{n \rightarrow \infty} F_n^\pi \in \mathcal{F}_m$.

(b) From the proof of (a), we have $F_{n+1}^\pi = H^{\varphi_0} F_n^{(1)\pi}$. Letting $n \rightarrow \infty$, by the dominated convergence theorem we obtain $F^\pi = H^{\varphi_0} F^{(1)\pi}$.

The last statement is obvious. \square

Proof of Theorem 3.1. Obviously, it follows from Lemma 5.1(b) that $F_n^* \in \mathcal{F}_r$ (hence F_n^* is in \mathcal{F}_m), and thus HF_n^* is well defined for every $n \geq -1$. Further, it is easy to see that $F_n^* \geq F_{n+1}^*$ for all $n \geq -1$, and so $\lim_{n \rightarrow \infty} F_n^*$ exists and moreover (by Lemma 5.1(c)) $\lim_{n \rightarrow \infty} F_n^* \in \mathcal{F}_r$. To complete the proof, it remains to prove that $\lim_{n \rightarrow \infty} F_n^* \leq F^*$ and $\lim_{n \rightarrow \infty} F_n^* \geq F^*$.

(i) To prove that $\lim_{n \rightarrow \infty} F_n^* \leq F^*$, it suffices to show that $F_n^* \leq F_n^\pi$ for all $\pi \in \Pi_{\text{RM}}$ and $n \geq -1$. Indeed, it is obviously true when $n = -1$. Now assume that $F_n^* \leq F_n^\pi$ for all $\pi \in \Pi_{\text{RM}}$ and some $n \geq -1$. Then, for any $\eta = \{\eta_0, \eta_1, \dots\} \in \Pi_{\text{RM}}$, we have

$$F_{n+1}^* := HF_n^* \leq HF_n^{(1)\eta} \leq H^{\eta_0} F_n^{(1)\eta} = F_{n+1}^\eta, \quad (5.3)$$

where the first inequality follows from the induction hypothesis, and the last equality is due to Lemma 3.1(b). Therefore, (by induction) we get $F_n^* \leq F_n^\pi$ for all $\pi \in \Pi_{\text{RM}}$ and $n \geq -1$. Hence, $\lim_{n \rightarrow \infty} F_n^* \leq F^\pi$ for all $\pi \in \Pi_{\text{RM}}$, and so (by the arbitrariness of π) $\lim_{n \rightarrow \infty} F_n^* \leq F^*$.

(ii) To show $\lim_{n \rightarrow \infty} F_n^* \geq F^*$, we need the following fact: For each $n \geq -1$, there exists $\pi \in \Pi_{\text{DM}}$ (depending on n) such that $F_n^* = F_n^\pi$. Indeed, suppose that this fact is true. Then, we have $F_n^* = F_n^\pi \geq F^\pi \geq F^*$, and thus $\lim_{n \rightarrow \infty} F_n^* \geq F^*$. We now prove this fact by induction. When $n = -1$, it is clear that $F_{-1}^* = \mathbb{1}_{[0, \infty)} = F_{-1}^\pi$ for any policy $\pi \in \Pi_{\text{DM}}$. Now assume that the fact is true for some $n \geq -1$, that is, there exists a policy $\theta \in \Pi_{\text{DM}}$ such that $F_n^* = F_n^\theta$. On the other hand, it follows from Lemma 5.1(a) and $F_n^* \in \mathcal{F}_r$ that there exists an $f \in \mathbb{F}$ such that $HF_n^* = H^f F_n^*$. Therefore, for $\eta = \{f, \theta\} \in \Pi_{\text{DM}}$, we have

$$F_{n+1}^* := HF_n^* = H^f F_n^* = H^f F_n^\theta = F_{n+1}^\eta, \quad (5.4)$$

where the last equality follows from Lemma 3.1(b). By induction, this fact is proved.

Combining (i) with (ii) yields that $\lim_{n \rightarrow \infty} F_n^* = F^*$, and thus the proof is achieved. \square

Proof of Theorem 3.2. (a) It follows from Lemma 3.1(b) that

$$F^\pi = H^{\varphi_0} F^{(1)\pi} \geq H^{\varphi_0} F^* \geq HF^* \quad \forall \pi = \{\varphi_0, \varphi_1, \dots\} \in \Pi_{\text{RM}}.$$

Since π is arbitrary, we have $F^* \geq HF^*$. We now show the reverse inequality. For each $(i, t, \lambda) \in E \times [0, T] \times \mathbb{R}$, (by Theorem 3.1) it is clear that $F_{n+1}^*(i, t, \lambda) = HF_n^*(i, t, \lambda) \leq H^a F_n^*(i, t, \lambda)$ for any $a \in A(i)$. By Theorem 3.1 and the dominated convergence theorem, we have $F^*(i, t, \lambda) \leq H^a F^*(i, t, \lambda)$ for any $a \in A(i)$, and so $F^*(i, t, \lambda) \leq HF^*(i, t, \lambda)$. Therefore we obtain $F^* = HF^*$.

(b) It is a straightforward result of part (a) and Lemma 5.1(a). \square

5.2. Proof of Lemma 3.2 and Theorem 3.3

Proof of Lemma 3.2. It is easy to see from Lemma 3.1(b) that F^f satisfies $F^f = H^f F^f$. Now let $F \in \mathcal{F}_m$ be another solution to the equation $F = H^f F$, and $G := |F^f - F|$. Then, we have $0 \leq G(i, t, \lambda) \leq \hat{H}^f G(i, t, \lambda)$, where

$$\hat{H}^f G(i, t, \lambda) := \sum_{j \in E} \int_0^t Q(du, j|i, f(i, t, \lambda)) G(j, t-u, \lambda-r(i, f(i, t, \lambda))u).$$

Hence, for any n , $G \leq (\hat{H}^f)^n G$ because the operator \hat{H}^f is monotonic. Moreover, since $0 \leq G \leq 1$, using [Assumption 3.1](#) and the technique as in the proof of [\[17, Theorem 1\]](#), we can show that

$$(\hat{H}^f)^n G(i, t, \lambda) \leq (1 - \epsilon^k)^{\lfloor n/k \rfloor} \quad \forall (i, t, \lambda) \in E \times [0, T] \times \mathbb{R}, \quad n > k, \quad (5.5)$$

where k is an arbitrary fixed nonnegative integer satisfying $k > T/\delta$, and $\lfloor n/k \rfloor$ denotes the largest integer not bigger than n/k .

Therefore, noting that $0 < (1 - \epsilon^k) < 1$ and letting $n \rightarrow \infty$ on inequality (5.5), we have $G = 0$, and thus this lemma follows. \square

Proof of Theorem 3.3. Obviously, part (a) is an immediate result of [Lemma 3.2](#), whereas part (b) follows from [Theorem 3.2\(b\)](#) and part (a). \square

6. Concluding remarks

This paper is the first attempt to investigate finite horizon SMDPs with risk probability criteria. As explained in previous sections, finite horizon optimality for continuous-time models is difficult to deal with. The key (new) point of this paper lies in that we take the planning horizons at decision epochs as part of state information, and introduce horizon-relevant policies. This fact makes it convenient to establish the optimality equation for the finite horizon SMDPs with risk probability criteria, and further develop an iteration algorithm for computing optimal policies with the minimum risk probability. As a kind of basic and important dynamic programming problems, finite horizon SMDPs with risk probability criteria have great potential applications. Besides manufacturing systems stated in this paper, further applications of the results in the paper to practical situations may be explored.

References

- [1] N. Bauerle, U. Rieder, Markov Decision Processes with Applications to Finance, in: Universitext, Springer, Heidelberg, 2011.
- [2] D.P. Bertsekas, S.E. Shreve, Stochastic Optimal Control: The Discrete-Time Case, Athena Scientific, Belmont, Massachusetts, 1996.
- [3] K. Boda, J.A. Filar, Y.L. Lin, L. Spanjers, Stochastic target hitting time and the problem of early retirement, IEEE Trans. Automat. Control 49 (2004) 409–419.
- [4] M. Bouakiz, Y. Kebir, Target-level criterion in Markov decision processes, J. Optim. Theory Appl. 86 (1995) 1–15.
- [5] E.K. Boukas, Q. Zhu, Q. Zhang, Piecewise deterministic Markov process model for flexible manufacturing systems with preventive maintenance, J. Optim. Theory Appl. 81 (1994) 259–275.
- [6] P. Buchholz, I. Schulz, Numerical analysis of continuous time Markov decision processes over finite horizons, Comput. Oper. Res. 38 (2011) 651–659.
- [7] B. Cekay, S. Ozekici, Mean time to failure and availability of semi-Markov missions with maximal repair, European J. Oper. Res. 207 (2010) 1442–1454.
- [8] O. Hernández-Lerma, J.B. Lasserre, Discrete-Time Markov Control Processes, Springer-Verlag, New York, 1996.
- [9] Y.H. Huang, X.P. Guo, Finite horizon semi-Markov decision processes with application to maintenance systems, European J. Oper. Res. 212 (2011) 131–140.
- [10] Y.H. Huang, X.P. Guo, First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs, Acta Math. Appl. Sin. 27 (2011) 177–190.
- [11] Y.H. Huang, X.P. Guo, Optimal risk probability for first passage models in semi-Markov decision processes, J. Math. Anal. Appl. 359 (2009) 404–420.
- [12] Y.H. Huang, X.P. Guo, X.Y. Song, Performance analysis for controlled semi-Markov systems with application to maintenance, J. Optim. Theory Appl. 150 (2011) 395–415.
- [13] K. Lazarski, L. Stettner, Average cost per unit time control of discrete time unreliable manufacturing systems with Markov demand, Math. Methods Oper. Res. 49 (1999) 457–473.
- [14] N. Limnios, J. Oprisan, Semi-Markov Processes and Reliability, Birkhäuser, Boston, 2001.
- [15] J. Liu, X. Zhao, On average reward semi-Markov decision processes with a general multichain structure, Math. Oper. Res. 29 (2004) 339–352.
- [16] C.E. Love, Z.G. Zhang, M.A. Zitron, R. Guo, A discrete semi-Markov decision model to determine the optimal repair/replacement policy under general repairs, European J. Oper. Res. 125 (2000) 398–409.
- [17] J.W. Mamer, Successive approximations for finite horizon semi-Markov decision processes with application to asset liquidation, Oper. Res. 34 (1986) 638–644.
- [18] F. Martinelli, Manufacturing systems with a production dependent failure rate: structure of optimality, IEEE Trans. Automat. Control 55 (2010) 2401–2406.
- [19] S. Pradhan, P. Damodaran, K. Srihari, Predicting performance measures for Markovian type of manufacturing systems with product failures, European J. Oper. Res. 184 (2008) 725–744.
- [20] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons Inc., New York, 1994.
- [21] S.S. Singh, V.B. Tadic, A. Doucet, A policy gradient method for semi-Markov decision processes with application to call admission control, European J. Oper. Res. 178 (2007) 808–818.
- [22] N.M. van Dijk, On the finite horizon Bellman equation for controlled Markov jump models with unbounded characteristics: existence and approximation, Stochastic Process. Appl. 28 (1988) 141–157.
- [23] C.C. White, Procedures for the solution of a finite-horizon, partially observed, semi-Markov optimization problem, Oper. Res. 24 (1976) 348–358.
- [24] D.J. White, Finite horizon Markov decision processes with uncertain terminal payoffs, Oper. Res. 43 (1995) 862–869.
- [25] D.J. White, Minimizing a threshold probability in discounted Markov decision processes, J. Math. Anal. Appl. 173 (1993) 634–646.
- [26] 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.
- [27] S.X. Yu, Y.L. Lin, P.F. Yan, Optimization models for the first arrival target distribution function in discrete time, J. Math. Anal. Appl. 225 (1998) 193–223.